Contents

AI in Medicine

Towards Real-Time Simulations Of Induced Electric Fields During Brain Stimulation Using Conditioned Transformers

Fabian Greifeneder
Research Department Medical Informatics
RISC Software GmbH, Hagenberg, Austria

Dominik Freinberger
Research Department Medical Informatics
RISC Software GmbH, Hagenberg, Austria

Philipp Moser
Research Department Medical Informatics
RISC Software GmbH, Hagenberg, Austria
philipp.moser@risc-software.at

Abstract

Real-time simulations of the induced electric fields during transcranial magnetic stimulation play an important role in guiding and optimizing the coil positioning. In this paper, we present our ongoing work on a deep learning-based surrogate model designed to rapidly predict the induced electric field distribution across the entire cortex, offering a much faster alternative to traditional numerical solvers. Leveraging (conditioned) transformer architectures, our approach operates directly on mesh-based head geometries, achieving highly accurate simulations in just 0.08 seconds on consumer hardware. While we continue to improve the neural surrogate, its current accuracy and efficiency have already enabled integration into an augmented reality platform, demonstrating a promising foundation for live electric field-guided brain stimulation applications.

1 Introduction

Transcranial magnetic stimulation (TMS) is a non-invasive medical procedure where a coil placed on a subject's scalp induces intracortical electric currents thereby enhancing or inhibiting neuronal activity. TMS is primarily used to treat neuropsychiatric disorders but is also being explored in preoperative functional mapping in tumor patients to minimize the risk of post-surgical deficits. Simulating the induced electric fields can help in efficiently and effectively positioning the coil during treatment to optimize targeted brain stimulation [1]. For live visualizations of the electric field in neuronavigation systems (e.g., via augmented reality) while moving the coil, these simulations necessitate (near) real-time computations.

In modern deep learning (DL), neural surrogates have emerged as an attractive alternative to traditional numerical solvers for approximating the solutions of complex partial differential equations (PDEs) that govern the behavior of physical systems. While numerical solvers offer high accuracy, they are often computationally prohibitive for real-time applications, especially with nonlinear physical models and high-resolution geometries. In contrast, neural PDE surrogates learn input-output relationships from precomputed numerical data, enabling both fast and accurate predictions. Although such surrogate approaches have been applied to e-field modeling in TMS [2], they primarily used convolution-based neural networks. This requires converting the 3D brain mesh to a regular grid, applying the surrogate model, and then mapping the predicted e-fields back onto the mesh, which is cumbersome, increases latency, and is inefficient in terms of memory.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

To overcome this detour, this paper presents our ongoing work on rapid e-field simulations using neural PDE surrogates that operate directly on mesh geometries. We leverage recent (diffusion) transformer architectures to jointly process mesh-based subject geometries with positional information of the TMS coil, achieving both highly accurate and fast e-field predictions, even on consumer hardware.

2 Methods

2.1 Neural PDE Surrogate: Model Architecture and Training Details

Neural operators are a popular concept to approximate solutions to PDEs by learning a mapping $\mathbf{G}$ between an input space $\mathbf{U}$ and an output space $\mathbf{V}$ as $\mathbf{G} : \mathbf{U} \rightarrow \mathbf{V}$. Our work builds on *Universal Physics Transformers* (UPT) [3], a learning approach that approximates $\mathbf{G}$ via three maps: $\mathbf{G} \approx \hat{\mathbf{G}} := \mathbf{D} \circ \mathbf{A} \circ \mathbf{E}$, with encoder $\mathbf{E}$, approximator $\mathbf{A}$, and decoder $\mathbf{D}$. In a supervised learning fashion, the network's parameters are optimized during training by repeatedly evaluating the N input-output pairs $(u_i, v_i) = (u_i, \mathbf{G}(u_i)), i = 1..., N$.

In the encoder $\mathbf{E}$, the k mesh coordinates were first normalized using global min-max scaling and multiplied by 300, following [3], and then embedded into the latent space using multi-scale sine-cosine positional encoding. Further, each coordinate had four input features—tissue conductivity and three Euclidean distances to the coil center and both loop centers. These features were linearly projected via a multi-layer perceptron to the latent dimension and added to the latent positional input. Next, a message-passing layer aggregated information at $n_s \ll k$ randomly sampled supernodes from within a radius r_{sn} around each supernode. The TMS coil position (x, y, z) and orientation (3×3 rotation matrix) were linearly projected to the latent dimension and injected as modulation/conditioning [3, 4] in the encoder's diffusion transformer and perceiver blocks.

Due to our stationary simulation scenario in this work, the approximator $\mathbf{A}$ consists solely of regular transformer blocks and employs no temporal evolution as in the original UPT [3]. The decoder $\mathbf{D}$ also starts with conditioned diffusion transformer blocks (using a separate multi-layer perceptron projecting the 12 coil positioning values to the latent dimension). Perceiver-like cross-attention layers with queries based on positional-encoded coordinates finally generate the induced electric field across the entire cortex, see Figure 1.

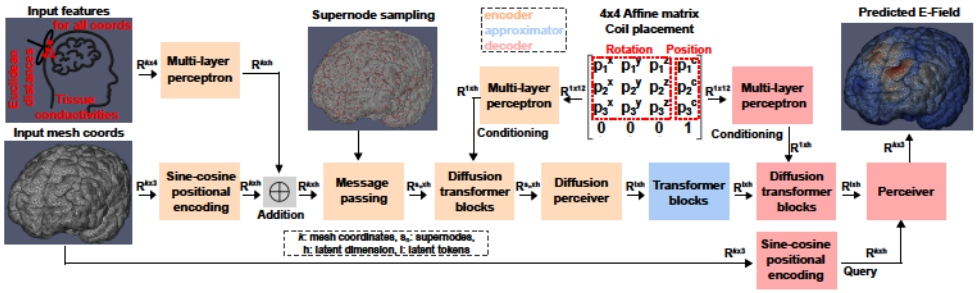


Figure 1: Schematics of the proposed transformer-based neural PDE surrogate.

The following summarizes the selected model hyperparameters: cortical coordinates k: 150000, supernodes n_s: 8000, supernode aggregation radius r_{sn}: 8, latent tokens: 160, latent dimension: 160, attention heads in all transformer blocks: 4, drop path: 0.25, total trainable parameters: 5.36 million. A learning rate of 5×10^{-5}, a batch size of 16, Lion optimizers with weight decay of 0.25, and a weighted mean squared error (MSE) loss (low/high weights for small/large target e-field values) were chosen. All trainings and evaluations were performed on a local workstation (Intel i7-13700 and NVIDIA GeForce RTX 4090 24GB) in PyTorch 2.2.0.

2.2 Generation of training/testing data: Numerical simulations of induced electric fields

A comprehensive dataset for training and evaluation of our proposed neural surrogate was precomputed using SimNIBS v4.01 [5], a popular toolkit for finite element-based simulations of non-invasive brain stimulation. First, realistic head models for 16 randomly selected subjects of the publicly

available WU-Minn Human Connectome Project were constructed: Based on 3T T_1w- and T_2w-MR images, SimNIBS's preprocessing pipeline *charm* automatically performed subject-specific whole-head multi-tissue segmentations, assigned conductivities to various tissues types and created computational brain meshes. Second, we randomly selected 24 rotations of a MagVenture-MCF-B65 butterfly coil on the 21 central positions of the EEG-10-10 system, totaling 8064 samples. For each coil placement, SimNIBS was used to simulate the induced electric field on the gray matter cortex, see Figure 2. We used a random train-test split of $75\%/25\%$.

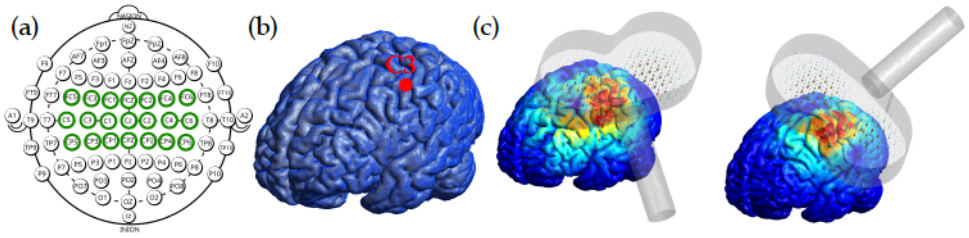


Figure 2: (a) EEG-10-10 system with central positions in green. (b) Head model of one subject with indicated cortical position C3. (c) Two exemplary coil rotations on the C3 position and the corresponding induced electric field magnitude on the cortical gray matter.

3 Results

Evaluating the trained surrogate model demonstrated highly accurate and smooth predictions, characterized by low MSE values and strong similarities of the predicted and ground truth electric field distributions. We achieved mean MSEs of $(1.36 \pm 0.34) \times 10^{-3}$ V/m and $(1.39 \pm 0.37) \times 10^{-3}$ V/m averaged over all train and test samples, respectively (see Figure 3)a. For four test set samples, the exemplary field distributions induced by different coil placements are illustrated in Figure 3b. The training of the neural PDE surrogate with 180 epochs took 82 hours on our consumer hardware.

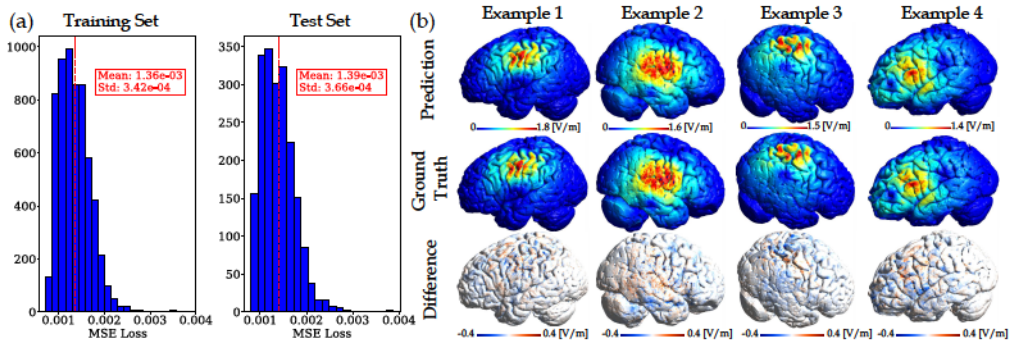


Figure 3: (a) Histogram of the train/test MSE losses. (b) Exemplary e-field magnitude predictions.

3.1 Integration into an Augmented Reality application

For a selected subject from our dataset, we implemented an augmented reality (AR) application in Unity for real-time, electric field-guided TMS, running on a Meta Quest 3. As illustrated in Figure 4, a virtual TMS coil is moved over a virtual, semi-transparent scalp, with the induced electric fields visualized on the gray matter below. Coil orientation and position are tracked by the AR headset and streamed to a Python backend, where a trained surrogate model predicts the corresponding e-fields in real time (0.075 s per simulation including latency, corresponding to the overall coil repositioning time; i.e., 130 times faster compared to the 10 seconds of SimNIBS's default conjugate gradient solver).

4 Discussion and Work-in-Progress

Simulating the induced electric field during brain stimulation has been described as beneficial for accurate, fast, and reproducible coil handling [1, 5]. To this end, our proposed neural surrogate

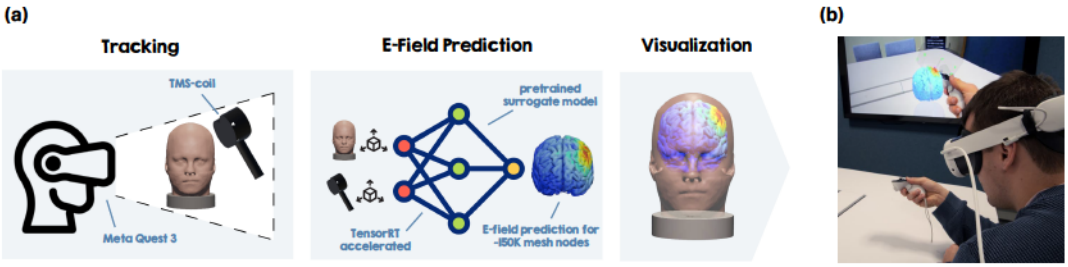


Figure 4: (a) Schematic overview of the AR application and its (b) live use.

leverages transformer-based techniques to rapidly generate highly precise cortical field distributions based on subject-specific head geometries. A key advancement is that we directly operate on mesh geometries, avoiding the need for convolutional architectures that require cumbersome (re-)sampling onto regular grids. Also, diffusion transformers [4] allowed to elegantly merge the TMS coil positional information with the mesh-based geometrical data via model conditioning.

While the model was successfully trained using diverse coil positions across the scalp, careful considerations were necessary to stabilize the training process due to repeated and significant loss spikes (as also mentioned in the original UPT paper [3]). We could control these by employing Lion optimizers with weight decay, a low learning rate, and float32 precision.

In summary, our neural PDE surrogate has already demonstrated promising results and has been successfully integrated into an AR application for electric field-guided TMS, providing real-time insights into brain stimulation. As a neural surrogate, the model is limited by the quality and diversity of the numerical simulations it is trained on. Hence, our ongoing efforts to include a broader range of subjects and cortical positions will ensure robust cross-subject generalizability. Additionally, feedback from medical professionals will help refine the AR application and support its future integration into a clinical TMS study. These advancements pave the way for more personalized, effective, and real-time electric field-guided TMS workflows.

Acknowledgments and Disclosure of Funding

This project is financed by research subsidies granted by the government of Upper Austria (MIMAS.ai, mimas.ng) and by the FFG (nARvibrain, grant no. 894756). www.ffg.at. "ICT of the Future" programme – an initiative of the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation and Technology (BMK). RISC Software GmbH is a member of UAR (Upper Austrian Research) Innovation Network.

References

[1] Opitz, A., Zafar, N., Bockermann, V., Rohde, V. & Paulus, W. (2014) Validating computationally predicted TMS stimulation areas using direct electrical stimulation in patients with brain tumors near precentral regions. *NeuroImage: Clinical* **4**:500–507.

[2] Park, T.Y., Franke, L., Pieper, S., Haehn, D. & Ning, L. (2024) A review of algorithms and software for real-time electric field modeling techniques for transcranial magnetic stimulation. *Biomedical Engineering Letters* **14**(3):393–405.

[3] Alkin, B., Fürst, A., Schmid, S., Gruber, L., Holzleitner, M. & Brandstetter, J. (2024) Universal Physics Transformers: A Framework For Efficiently Scaling Neural Operators. In *Proceedings of the 37th Annual International Conference of the Advances in Neural Information Processing Systems (NeurIPS)*, pp. 25152-25194

[4] Peebles, W. & Xie, S. (2023) Scalable Diffusion Models with Transformers. In *Proceedings of 2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172-4182

[5] Thielscher, A., Antunes, A. & Saturnino, G.B. (2015) Field modeling for transcranial magnetic stimulation: A useful tool to understand the physiological effects of TMS? In *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 222–225.

Generating Realistic and Accurate SMPL Body Shapes from Anthropometric Measurements

Maja Katharina Nikolic[1,2], Sophie Kaltenleithner[1], Ulrich Bodenhofer[2], and Michael Giretzlehner[1]

[1]RISC Software GmbH
[2]University of Applied Sciences Upper Austria
[1]{maja.nikolic, sophie.kaltenleithner, michael.giretzlehner}@risc-software.at
[1]ulrich.bodenhofer@fh-hagenberg.at

Abstract

Accurate human 3D avatars are essential for various medical applications, such as pain visualization and burn size estimation. Generating these avatars from simple anthropometric measurements offers a cheaper and more practical alternative to conventional 3D body scanners and image-based reconstruction methods, especially when patients have limited mobility. In this work, we investigate the reliability of predicting Skinned Multi-Person Linear model (SMPL) shape parameters from anthropometric measurements in the presence of real-world noise. We further introduce β-likelihood to quantify the anatomical plausibility of generated shapes against a learned distribution. Multiple regression models are evaluated on two external datasets, revealing a clear trade-off between metric accuracy and shape plausibility. The results indicate that regularized regression models are best suited to balance this trade-off when dealing with real-world measurement noise.

1 Introduction

Human 3D avatars are gaining more relevance in medical contexts for various applications. For instance, in burn care, human 3D models can be used to visualize burn injuries and calculate the percentage of total body surface area (TBSA) burned, which is a key metric for determining appropriate burn treatment [8]. Another application is pain visualization, where patient-specific 3D body models can help patients to communicate the location and extent of their pain to medical professionals. Since pain is still often documented on generic, frequently male, 2D templates [12], patient-specific 3D avatars can improve anatomical correspondence of pain markings, especially when perceived pain depth relates to tissue volume.

As 3D body scanners are costly, image-based reconstruction methods have been proposed [4, 11, 13]. However, patient photos may be impractical in clinical settings, due to patient's limited mobility or privacy concerns. In such cases, fitting a parametric body model from anthropometric measurements alone can serve as a practical alternative. While Ludwig et al. [6] demonstrate that machine learning can predict body shape from such measurements, their evaluation does not consider measurement errors, which are common in manual anthropometry due to variation in landmark placement, soft-tissue compression and posture. To better understand the reliability of anthropometric measurements the the authors of The Virtual Caliper [10] further investigate which body measurements can be measured most reliably. Their results show that individual measurements can be accurately reconstructed on a 3D model but as the number of approximated measurements increases, the generated body shapes

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

tend to appear less realistic. However, these observations are based solely on perceptual evaluation, leaving the quantitative assessment of shape plausibility unexplored.

To address this gap, this study investigates the reliability of Skinned Multi-Person Linear Model (SMPL) shape parameters predicted from anthropometric measurements under real-world measurement noise. Instead of relying on perceptual judgment to assess plausibility, we propose β-likelihood as an objective metric that measures how likely the predicted shape parameters are under the learned shape distribution. To investigate the trade-off between measurement accuracy and anatomical plausibility across model classes, multiple regression models are trained to predict SMPL shape parameters from a set of anthropometric measurements. The models are then evaluated on external validation datasets containing measurement noise with respect to both metric accuracy and shape plausibility.

2 Data Preparation

Skinned Multi-Person Linear Model (SMPL) [5] is the state-of-the-art parametric human body model, developed and published by the Max Planck Institute for Intelligent Systems. The body shape is controlled by β-coefficients, which are referred to as **shape parameters** throughout this paper. These coefficients encode the principal components of variation derived from a large dataset of aligned 3D human meshes, and capture differences in height, limb proportions, torso shape and overall body composition. Each β-coefficient represents a distinct dimension of shape variation, for example controlling leg length, hip circumference or other morphological features. SMPL provides gender-specific models, which capture typical shape differences between male and female bodies, as well as a gender-neutral model.

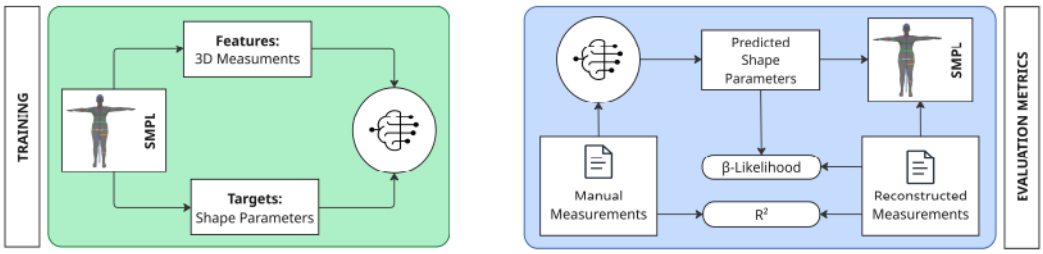
Anthropometric measurements of SMPL body models are calculated using the SMPL-Anthropometry repository [1]. The repository defines two types of measurements, lengths and circumferences, that are computed differently. Both approaches use predefined landmarks, whose exact spatial positions are determined using the SMPL shape parameters. Length measurements are computed by summing the Euclidean distances between defined landmarks, while circumference measurements are obtained by slicing the mesh at a defined landmark and summing the distances along the convex hull of the resulting intersection points.

Training Dataset This study utilizes the Avatars in Geography Optimized for Regression Analysis (AGORA) dataset [9], which provides SMPL shape parameters based on a neutral body model, fitted to high-quality 3D human scans. The dataset includes 3187 subjects and captures a diverse range of body shapes, each represented by 10 shape parameters. Using the SMPL shape from the AGORA dataset, 3D body models are generated, and anthropometric measurements are computed. The resulting dataset, consisting of SMPL shape parameters paired with their derived anthropometric measurements, serves as the foundation for model training (see Figure 1a). As the measurements are derived directly from the parametric model, they are free from manual measurement noise, ensuring a consistent and deterministic relationship between measurements and shape parameters.

Validation For external validation, the anthropometric datasets Anthropometric Survey of US Army Personnel (ANSUR) [3] and Study of Health in Pomerania (SHIP) [2] are used. ANSUR contains 93 anthropometric measurements collected from over 6000 adult US military personnel, collected under real-world measurement conditions, including potential measurement errors. As most participants are subject to military recruitment standards, the dataset does not fully represent the diversity of body types in the general population. The SHIP dataset originates from a study in which 2313 full-body scans were collected and processed to derive an anthropometric dataset representing the German working-age population, containing 39 anthropometric measurements.

For the present study a set of 16 commonly used anthropometric measurements was identified as a suitable starting point. Future work will focus on determining which subset of measurements is truly essential and most effective for this task. Since the real-world anthropometric data available in SHIP and ANSUR datasets do not perfectly match the predefined set in terms of landmark positions, closely related measurements were selected where necessary. Accordingly, the landmarks used for measurement computation on the SMPL model were adapted to match the ANSUR measurement definitions. Note that the SHIP dataset includes only 11 of the 16 measurements, as corresponding measurements for the remaining five are not available.

3 Experimental Setup and Evaluation



(a) The regression models are trained using SMPL body measurements as input features and the corresponding SMPL shape parameters as target outputs.

(b) The predicted shape parameters are directly used to assess plausibility, while for metric accuracy they are first used to reconstruct the corresponding body measurements, from which R^2 is computed.

Figure 1: Overview of the model training and evaluation process, showing how SMPL body measurements are used to predict shape parameters and how these predictions are evaluated in terms of plausibility and metric accuracy.

Evaluation Metrics Model performance is evaluated using two complementary metrics. The coefficient of determination (R^2) measures the accuracy of reconstructed body measurements, while β-likelihood quantifies the plausibility of generated body shapes. As depicted in Figure 1b, R^2 is computed by reconstructing body measurements from predicted shape parameters and calculating a weighted average across measurements, with weights proportional to each measurement's variance. More precisely, if R_j^2 denotes the coefficient of determination for measurement j and σ_j^2 its variance, the reported score is computed as

$$R_w^2 = \frac{\sum_{j=1}^{M} \sigma_j^2 R_j^2}{\sum_{j=1}^{M} \sigma_j^2},$$

where M is the number of evaluated measurements. This ensures that errors in measurements with greater natural variability, such as waist circumference, contribute more to the overall metric than measurements with lower variability, like wrist circumference. To address the problem of generating unrealistic body shapes, we propose using β-likelihood to quantify the plausibility of predicted shapes. Using the SMPL shape parameters in the AGORA dataset, a multivariate distribution is estimated by computing the mean vector and covariance matrix of the β-coefficients. The likelihood of a given shape is computed relative to this distribution using multivariate normal log-density. For a predicted shape-parameter vector $\hat{\beta}$ we report the log-likelihood

$$\log p(\hat{\beta}) = -\frac{1}{2}\left[(\hat{\beta}-\mu)^\top \Sigma^{-1}(\hat{\beta}-\mu) + d\log(2\pi) + \log|\Sigma|\right],$$

where μ and Σ are the mean vector and covariance matrix estimated from AGORA, and d is the dimensionality of the shape space.

Model Training Multiple regression models, namely Linear Regression, Ridge Regression, Elastic Net, K-Nearest-Neighbors (KNN), Random Forest and a Multilayer Perceptron (MLP), were trained on the AGORA dataset, using anthropometric measurements as input features and the corresponding SMPL shape parameters as target outputs (see Figure 1a). Hyperparameter tuning was performed to jointly optimize both the coefficient of determination (R^2) and log-β-likelihood on the validation datasets. The hyperparameters were selected from predefined search spaces, using the Optuna Multi-Objective TPE Sampler [7]. For each model class, optimization was performed over 20 trials. As the measurement sets of SHIP and ANSUR differ, hyperparameters were optimized separately for each dataset.

4 Results and Conclusion

Results A Pareto front analysis was conducted to compare the different model classes, based on their performance on the SHIP and ANSUR validation datasets. In Figure 2, Pareto-optimal points

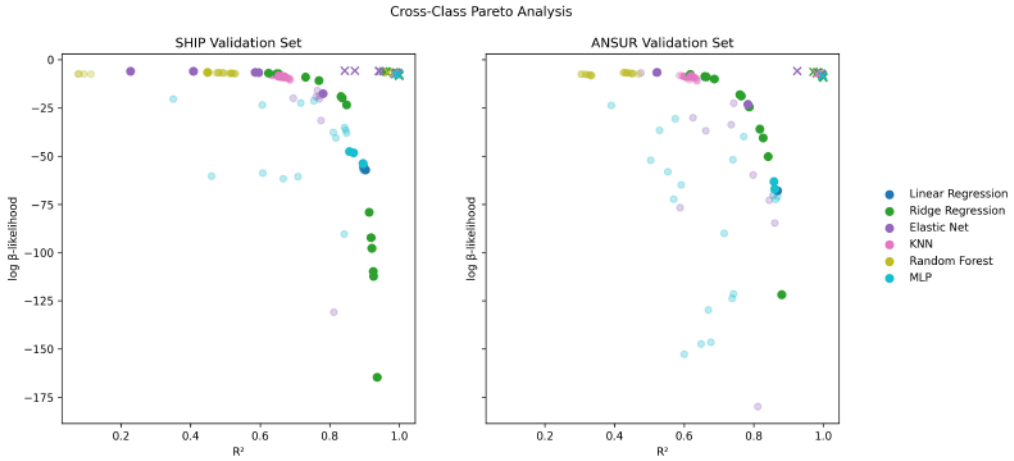


Figure 2: Pareto front analysis of different model classes, evaluated on the SHIP and ANSUR validation datasets. Each point corresponds to one hyperparameter configuration of the respective model class. Pareto-optimal points are highlighted, while the remaining validation results are shown in faded colors. Crosses indicate the corresponding results on a validation split of the AGORA training dataset. The figure illustrates the trade-off between R^2 and β-likelihood on the validation datasets under real-world measurement conditions.

are highlighted, while the remaining validation results are shown in faded colors. Crosses indicate the corresponding results on an AGORA validation split. The analysis reveals a clear trade-off between the two evaluation metrics for the validation datasets. Higher metric accuracy is associated with less plausible body shapes, and vice versa. However, this pattern is not observed for the validation split of the AGORA training dataset, suggesting that it is mainly caused by dataset shift between deterministic SMPL-derived measurements and real-world anthropometric measurements. KNN and Random Forest models produce shapes with better β-likelihood, indicating greater plausibility, but at the expense of metric accuracy. In contrast, linear regression-based models like Ridge Regression and Elastic Net achieve the highest metric accuracy but tend to generate less plausible body shapes. Notably, Ridge Regression demonstrates that the trade-off between metric accuracy and shape plausibility can be controlled by tuning the regularization parameter, which pulls the predicted shape parameters closer to zero.

Conclusion In this work, the AGORA dataset, which provides SMPL shape parameters, was used to generate a dataset consisting of paired SMPL shape parameters and corresponding measurements. This dataset served to train multiple regression models, where anthropometric measurements were used as input features and the associated SMPL shape parameters were predicted as outputs. In addition, the dataset was used to estimate a multivariate distribution that forms the basis of the proposed β-likelihood metric for quantifying the plausibility of generated body shapes. For validation, additional anthropometric datasets were incorporated to assess model performance under real-world measurement conditions including potential measurement errors. While both high metric accuracy and plausible body shapes can be achieved on a validation split of the training dataset, evaluation on the external validation datasets reveals a clear trade-off between metric accuracy and the plausibility of the generated body shapes. In this context, linear regression models with regularization are particularly suitable, as the trade-off can be controlled by tuning the regularization parameter. Future work will focus on training models whose loss function optimizes the error of reconstructed measurements, rather than the shape parameters, addressing a key limitation of the current study. To simplify the measurement process in a clinical setting, essential measurements will be identified to define a minimal set of necessary inputs, and strategies will be explored to handle missing measurements.

Acknowledgments This work was funded by the FFG (Austrian Research Promotion Agency) under the grant 913010 (Embodied Perceptions). This project is also financed by research subsidies granted by the government of Upper Austria. RISC Software GmbH is Member of UAR (Upper Austrian Research) Innovation Network. This work is part of a Master's thesis at the University of Applied Sciences Upper Austria.

References

[1] David Bojanić. SMPL-Anthropometry. https://github.com/DavidBoja/SMPL-Anthropometry, February 2026.

[2] Dominik Bonin, Alexander Ackermann, Dörte Radke, Markus Peters, and Sascha Wischniewski. Anthropometric dataset for the German working-age population using 3D body scans from a regional epidemiological health study and a weighting algorithm. *Ergonomics*, 66(8):1057–1071, August 2023.

[3] C. C. Gordon, C. L. Blackwell, B. Bradtmiller, Joseph L. Parham, Patricia Barrientos, S. Paquette, B. Corner, J. Carson, Joseph Venezia, Belva M. Rockwell, M. Mucher, and S. Kristensen. 2012 Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics, December 2014.

[4] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. VIBE: Video Inference for Human Body Pose and Shape Estimation, April 2020.

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A Skinned Multi-Person Linear Model. In Mary C. Whitton, editor, *Seminal Graphics Papers: Pushing the Boundaries, Volume 2*, pages 851–866. ACM, New York, NY, USA, 1 edition, August 2023.

[6] Katja Ludwig, Julian Lorenz, Daniel Kienzle, Tuan Bui, and Rainer Lienhart. Leveraging Anthropometric Measurements to Improve Human Mesh Estimation and Ensure Consistent Body Shapes, April 2025.

[7] Yoshihiko Ozaki, Yuki Tanigaki, Shuhei Watanabe, Masahiro Nomura, and Masaki Onishi. Multiobjective Tree-Structured Parzen Estimator. *Journal of Artificial Intelligence Research*, 73:1209–1250, April 2022.

[8] Daryousch Parvizi, Lars-Peter Kamolz, Michael Giretzlehner, Herbert L. Haller, Maria Trop, Harald Selig, Peter Nagele, and David B. Lumenta. The potential impact of wrong TBSA estimations on fluid resuscitation in patients suffering from burns: Things to keep in mind. *Burns*, 40(2):241–245, March 2014.

[9] Priyanka Patel, Chun-Hao P. Huang, Joachim Tesch, David T. Hoffmann, Shashank Tripathi, and Michael J. Black. AGORA: Avatars in Geography Optimized for Regression Analysis, April 2021.

[10] Sergi Pujades, Betty Mohler, Anne Thaler, Joachim Tesch, Naureen Mahmood, Nikolas Hesse, Heinrich H. Bülthoff, and Michael J. Black. The Virtual Caliper: Rapid Creation of Metrically Accurate Avatars from 3D Measurements. *IEEE Transactions on Visualization and Computer Graphics*, 25(5):1887–1897, May 2019.

[11] Akash Sengupta, Ignas Budvytis, and Roberto Cipolla. Probabilistic Estimation of 3D Human Shape and Pose with a Semantic Local Parametric Model, November 2021.

[12] K.V. Stein, T.E. Dorner, W. Ilias, and A. Rieder. Schmerzpatienten und ihre Erwartungen an die ärztliche Versorgung. *Der Schmerz*, 24(5):468–473, September 2010.

[13] Moyu Wang and Qingping Yang. From prediction to measurement, an efficient method for digital human model obtainment. *International Journal of Metrology and Quality Engineering*, 15:1, 2024.

Flow Matching for Conditional MRI-CT and CBCT-CT Image Synthesis

Arnela Hadzic, Simon Johannes Joham, Martin Urschler
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz
{arnela.hadzic, martin.urschler}@medunigraz.at

Abstract

Generating synthetic CT (sCT) from MRI or CBCT plays a crucial role in enabling MRI-only and CBCT-based adaptive radiotherapy, improving treatment precision while reducing patient radiation exposure. To address this task, we adopt a fully 3D Flow Matching (FM) framework, motivated by recent work demonstrating FM's efficiency in producing high-quality images. In our approach, a Gaussian noise volume is transformed into an sCT image by integrating a learned FM velocity field, conditioned on features extracted from the input MRI or CBCT using a lightweight 3D encoder. We evaluated the method on the SynthRAD2025 Challenge benchmark, training separate models for MRI → sCT and CBCT → sCT across three anatomical regions: abdomen, head and neck, and thorax. Validation and testing were performed through the challenge submission system. The results indicate that the method accurately reconstructs global anatomical structures; however, preservation of fine details was limited, primarily due to the relatively low training resolution imposed by memory and runtime constraints. Future work will explore patch-based training and latent-space flow models to improve resolution and local structural fidelity.

1 Introduction

Accurate dose calculation in radiotherapy (RT) requires quantitative tissue information typically obtained from computed tomography (CT). However, CT exposes patients to ionizing radiation, offers limited soft-tissue contrast, and is often unavailable in treatment rooms. Magnetic resonance imaging (MRI) provides superior soft-tissue contrast without radiation but lacks attenuation information, while cone-beam CT (CBCT) offers in-room imaging and lower radiation doses but suffers from artifacts.

Converting MRI or CBCT to synthetic CT (sCT) can enable MRI-only workflows, adaptive MRI-based RT, and CBCT-based adaptive RT, improving precision while reducing radiation exposure. Despite rapid methodological progress, the lack of public datasets and standardized evaluation of CT synthesis approaches hampers objective comparison. The SynthRAD2025 Challenge addresses this gap by providing a public benchmark for MRI → sCT and CBCT → sCT methods with unified data, metrics, and evaluation protocols. Compared to the SynthRAD2023 Challenge [1], which focused on a smaller brain and pelvis dataset, the SynthRAD2025 Challenge includes over 2,300 cases across abdomen (AB), head and neck (HN), and thorax (TH) regions from multiple centers, providing a larger and more diverse benchmark for evaluating sCT generation methods.

To tackle the sCT generation problem for the SynthRAD2025 Challenge, we employ a Flow Matching (FM) generative model [2, 3]. FM offers an efficient way to model complex data distributions by learning continuous probability flows from Gaussian noise to target images. Compared to generative adversarial networks traditionally used in medical image synthesis [4, 5], FM provides more stable training and distribution coverage. Moreover, it requires fewer integration steps than diffusion models

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

[6–8], thereby achieving faster inference while maintaining high image quality [9–12].
In this work, we adapt the FM formulation [2] to explicitly condition the model on the given MRI/CBCT volumes using a lightweight 3D encoder. To the best of our knowledge, this is the first application of Flow Matching to conditional synthetic CT generation.

2 Materials and Methods

2.1 Dataset

The dataset provided by the SynthRAD2025 Challenge [13] contains paired MR-CT images for Task 1 (MRI $\rightarrow$ sCT) and CBCT-CT images for Task 2 (CBCT $\rightarrow$ sCT), acquired from multiple centers with varying sizes. For Task 1, image dimensions range from 238-608 $\times$ 239-553 $\times$ 42-164, while for Task 2, image dimensions range from 265-560 $\times$ 225-560 $\times$ 49-139 voxels across anatomical regions. All images have an anisotropic spacing of $1 \times 1 \times 3$ mm^3.

The Task 1 training set consisted of 578 paired MRI-CT volumes, comprising 175 AB, 221 HN, and 182 TH cases. An additional 89 unpaired MRI volumes were provided for validation. For Task 2, the training set included 1,472 paired CBCT-CT volumes (309 AB, 325 HN, and 321 TH), and 148 unpaired CBCT volumes were provided for validation. The test phases for both tasks were conducted via the challenge platform on 223 MRI volumes (Task 1) and 369 CBCT volumes (Task 2). Since the ground-truth CT images for the validation and test sets were not accessible to participants, we performed an internal data split using 75% of the training data for training and 25% for validation. The setup that achieved the lowest MSE on this internal validation set was chosen and the final models were trained from scratch on all available training images.

2.2 Methodology

We performed CT synthesis from MRI or CBCT, respectively, using a conditional Flow Matching framework [2, 3]. In our formulation, the target CT image $x_1 \sim p_1$ is connected to a base Gaussian noise sample $x_0 \sim p_0 = \mathcal{N}(0, I)$ through a linear interpolation path

$$x_t = t\, x_1 + \sigma_t\, \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad \sigma_t = 1 - (1 - \sigma_{\min})\, t,$$

with $\sigma_{\min} = 10^{-5}$ [2]. For this path, the corresponding probability flow velocity field is given by

$$u_t(x_t, t \mid x_1) = \frac{x_1 - (1 - \sigma_{\min})\, x_t}{1 - (1 - \sigma_{\min})\, t}.$$

We train the Flow Matching model on paired MRI/CT or CBCT/CT images. Specifically, we draw $t \sim \mathcal{U}(0, 1)$, construct x_t from the target CT, and train the network $v_\theta(x_t, t \mid c)$ to approximate u_t, where c is the corresponding MRI/CBCT conditioning image. The conditioning volume is processed by a lightweight 3D convolutional encoder, and its features are concatenated with x_t before being passed to a 3D U-Net that predicts a velocity field. The objective combines L1 and MSE losses:

$$\mathcal{L}(\theta) = \lambda_{\text{L1}} \| v_\theta(x_t, t \mid c) - u_t \|_1 + \lambda_{\text{MSE}} \| v_\theta(x_t, t \mid c) - u_t \|_2^2,$$

with $\lambda_{\text{L1}} = \lambda_{\text{MSE}} = 1$.

At inference time, we sample $x_0 \sim \mathcal{N}(0, I)$ and solve the probability-flow ODE

$$\dot{x}_t = v_\theta(x_t, t \mid c), \quad t : 0 \rightarrow 1,$$

where v_θ is the learned velocity field. Integration is performed with a 4th-order Runge-Kutta (RK4) solver using 32 steps. Starting from random noise x_0, the sample is gradually transformed along the learned flow to produce the final synthetic CT image x_1.

For each task and anatomical region, separate models were trained using the corresponding paired MRI-CT or CBCT-CT training images (see Section 2.1).

3 Experimental Setup

3.1 Implementation Details

3.1.1 Data preprocessing

All images were resampled to a uniform spatial resolution of $128 \times 128 \times 128$ voxels with an isotropic spacing of $1 \times 1 \times 1$ mm^3. Intensity normalization was applied independently to each modality: MR images underwent z-score normalization followed by clipping to the range $[-3,\ 3]$, while CBCT and CT images were clipped to the Hounsfield Unit (HU) range $[-1024,\ 3071]$ and scaled by dividing by 1000, resulting in a normalized range of $[-1024,\ 3071]$.

3.1.2 Architectural and training details

To predict velocity fields, we employed a 3D conditional U-Net architecture from [7]. We adapted the network to first process the conditioning MR or CBCT image through two $3 \times 3 \times 3$ convolutional layers with ReLU activations, producing feature maps with 64 channels. These features were then concatenated with the input CT image and passed to the main 3D U-Net [7]. The U-Net consisted of four resolution levels, each containing one residual block. The number of convolutional channels at each level was determined by a set of channel multipliers $(1, 1, 2, 3, 4)$, scaled from a base of 64 channels. This resulted in 64, 64, 128, 192, and 256 channels across the levels. Self-attention mechanisms were applied at resolutions corresponding to $16 \times 16 \times 16$ and $8 \times 8 \times 8$ feature maps. All residual blocks used dropout with a probability of 0.05.

Training data was augmented with random 3D translations ($\pm$ 5 voxels) and rotations (± 0.1 radians). The model was optimized using AdamW with a learning rate of 10^{-4} and a weight decay of 10^{-5}, with an effective batch size of 2. This configuration was used for each task (MRI $\rightarrow$ sCT, CBCT $\rightarrow$ sCT) and anatomical region (AB, HN, TH), resulting in six models in total.

MRI $\rightarrow$ sCT models were trained for 100,000 steps, whereas CBCT $\rightarrow$ sCT models were trained for 50,000 steps. Implementation was done in PyTorch. The ODE solver (i.e., RK4) was implemented using the torchdiffeq library. Additional packages included SimpleITK, NumPy, and scikit-image. Training was performed on an 80 GB NVIDIA A100 and required approximately 3 days and 15 hours for MRI $\rightarrow$ sCT (HN), and 2 days and 9 hours for CBCT $\rightarrow$ sCT (HN). Inference per scan took approximately 2 minutes on a 24 GB NVIDIA GeForce RTX 3090 GPU. The total number of trainable parameters was 41,237,057.

3.1.3 Data postprocessing

Synthesized CT images were resampled back to the original spatial resolution of the input MR or CBCT image using nearest-neighbor interpolation. Additionally, the original image spacing and origin were reapplied and intensity values were denormalized by multiplying by 1000 to recover values in the HU range $[-1024,\ 3071]$.

3.2 Evaluation

Quantitative scores for our models on the validation set provided by the SynthRAD2025 Challenge were obtained via the submission system. The evaluation metrics used assessed both image similarity and geometric consistency. Image similarity between synthetic CT and target CT within the provided body mask was assessed using mean absolute error (MAE), peak signal-to-noise ratio (PSNR), and multi-scale structural similarity index (MS-SSIM). Geometric consistency was evaluated through automated segmentation (using TotalSegmentator [14] and nnUNet [15]), with performance quantified using the multiclass Dice coefficient (mDice) and the 95th percentile Hausdorff distance (HD95), averaged across all anatomical structures (abdomen, head and neck, and thorax).

4 Results

Table 1 presents the quantitative results for the MRI $\rightarrow$ sCT model on 89 validation images and for the CBCT $\rightarrow$ sCT model on 148 CBCT validation images. The results are reported using the metrics described in Section 3.2, with both the mean and standard deviation (mean $\pm$ std) for each metric.

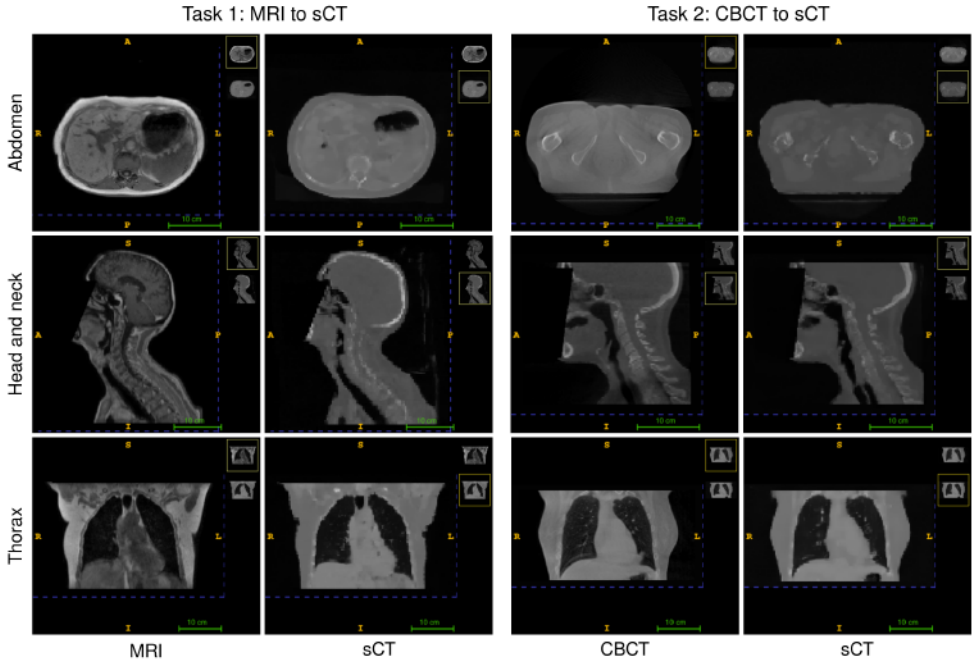


Figure 1: Examples of synthetic CT images generated by our MRI → sCT (left column) and CBCT → sCT models (right column) for different anatomical regions.

Qualitative results are shown in Figure 1, where examples of synthetic CT images for both tasks and all three anatomical regions are displayed.

We observed no artifacts in the generated images, but they appear blurry due to the relatively low training resolution. While the global anatomy is captured well, the preservation of local details could be improved. This is also reflected in the quantitative scores reported in Table 1.

Overall, our approach was ranked 12th in the final SynthRAD2025 Challenge leaderboard. While our quantitative performance was lower than the top-ranked methods, such as the winning team which achieved an MAE of approximately 65 HU for Task 1, it is important to note the fundamental difference in methodology. The majority of competing participants, including the highest-ranking teams, utilized established supervised learning frameworks (e.g., highly optimized 3D U-Nets or ensembled CNNs) designed to learn the synthesis mapping directly in a discriminative manner. In contrast, our work follows a different approach by employing a generative probability flow framework. Although this generative approach faced challenges with local structural fidelity due to current resolution constraints stemming from GPU memory limitations, it introduces a generative modeling alternative that can be further scaled to capture complex data distributions.

Table 1: Quantitative results obtained on the corresponding validation sets for both tasks.

Model	MAE	PSNR	MS-SSIM	DICE	HD95
MRI → sCT	146.17 ± 27.91	24.62 ± 1.43	0.82 ± 0.08	0.44 ± 0.14	18.54 ± 10.70
CBCT → sCT	114.74 ± 23.12	26.30 ± 1.61	0.88 ± 0.04	0.58 ± 0.14	12.88 ± 7.73

5 Discussion and Conclusions

In this work, we addressed the MR/CBCT → sCT task with a Flow Matching generative model conditioned on MR/CBCT inputs. The method consistently reproduced global anatomy (see Figure 1), but preservation of fine structures remains suboptimal. We attribute this primarily to the training resolution ($128 \times 128 \times 128$) being lower than the original resolution of many scans. Scaling to higher resolutions was limited by memory and runtime constraints. Importantly, the same architecture and training setup were used for all models, without task- or anatomy-specific tuning.

In future work, we plan to adopt 3D patch-based training and inference, and explore latent-space flow models to enable higher effective resolution and improved local detail.

Acknowledgments and Disclosure of Funding

This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/PAT1748423.

References

[1] Evi MC Huijben, Maarten L Terpstra, Suraj Pai, Adrian Thummerer, Peter Koopmans, Manya Afonso, Maureen Van Eijnatten, Oliver Gurney-Champion, Zeli Chen, Yiwen Zhang, et al. Generating synthetic computed tomography for radiotherapy: SynthRAD2023 challenge report. *Medical Image Analysis*, 97:103276, 2024.

[2] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

[3] Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

[4] Dong Nie, Roger Trullo, Jun Lian, Caroline Petitjean, Su Ruan, Qian Wang, and Dinggang Shen. Medical image synthesis with context-aware generative adversarial networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 417–425. Springer, 2017.

[5] Thomas Neff, Christian Payer, Darko Štern, and Martin Urschler. Generative adversarial network based synthesis for supervised medical image segmentation. In *Proceedings of the OAGM&ARW Joint Workshop 2017: Vision, Automation and Robotics*, pages 140–145, 2017.

[6] Gustav Müller-Franzes, Jan Moritz Niehues, Firas Khader, Soroosh Tayebi Arasteh, Christoph Haarburger, Christiane Kuhl, Tianci Wang, Tianyu Han, Teresa Nolte, Sven Nebelung, et al. A multimodal comparison of latent denoising diffusion probabilistic models and generative adversarial networks for medical image synthesis. *Scientific Reports*, 13(1):12098, 2023.

[7] Zolnamar Dorjsembe, Hsing-Kuo Pao, Sodtavilan Odonchimed, and Furen Xiao. Conditional diffusion models for semantic 3D brain MRI synthesis. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4084–4093, 2024.

[8] Arnela Hadzic, Lea Bogensperger, Simon Johannes Joham, and Martin Urschler. Synthetic Augmentation for Anatomical Landmark Localization Using DDPMs. In *International Workshop on Simulation and Synthesis in Medical Imaging (SASHIMI)*, pages 1–12. Springer, 2024.

[9] Yasi Zhang, Peiyu Yu, Yaxuan Zhu, Yingshan Chang, Feng Gao, Ying Nian Wu, and Oscar Leong. Flow priors for linear inverse problems via iterative corrupted trajectory matching. *Advances in Neural Information Processing Systems (NeurIPS)*, 37:57389–57417, 2024.

[10] Daikun Zhang, Qiuyi Han, Yuzhu Xiong, and Hongwei Du. Mutli-modal straight flow matching for accelerated MR imaging. *Computers in Biology and Medicine*, 178:108668, 2024.

[11] Lea Bogensperger, Dominik Narnhofer, Alexander Falk, Konrad Schindler, and Thomas Pock. FlowSDF: Flow matching for medical image segmentation using distance transforms. *International Journal of Computer Vision*, pages 1–13, 2025.

[12] Arnela Hadzic, Lea Bogensperger, Andrea Berghold, and Martin Urschler. Flow matching-based data synthesis for robust anatomical landmark localization. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[13] Adrian Thummerer, Erik van der Bijl, Arthur Jr Galapon, Florian Kamp, Mark Savenije, Christina Muijs, Shafak Aluwini, Roel JHM Steenbakkers, Stephanie Beuel, Martijn PW Intven, et al. SynthRAD2025 Grand Challenge dataset: Generating synthetic CTs for radiotherapy from head to abdomen. *Medical Physics*, 52(7):e17981, 2025.

[14] Jakob Wasserthal, Hanns-Christian Breit, Manfred T Meyer, Maurice Pradella, Daniel Hinck, Alexander W Sauter, Tobias Heye, Daniel T Boll, Joshy Cyriac, Shan Yang, et al. TotalSegmentator: robust segmentation of 104 anatomic structures in CT images. *Radiology: Artificial Intelligence*, 5(5):e230024, 2023.

[15] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021.

Evidential Deep Learning for Missing Boundary Detection in Topologically Constrained OCT Layer Segmentation

Botond Fazekas, Hrvoje Bogunović
Christian Doppler Laboratory for Artificial Intelligence in Retina
Institute of Artificial Intelligence, Center for Medical Data Science
Medical University of Vienna
Vienna, Austria
botond.fazekas@meduniwien.ac.at

Abstract

Optical coherence tomography (OCT) layer boundary regression methods provide sub-pixel precision and topological guarantees but fundamentally assume that every layer exists across all A-scans. This mathematical constraint fails in severe pathologies such as Geographic Atrophy (GA), where specific retinal layers disappear. We extend the topologically constrained SD-RetinaNet framework to jointly perform boundary regression and explicitly detect missing layers using uncertainty quantification. We introduce a Gaussian Negative Log-Likelihood (NLL) formulation to calibrate aleatoric uncertainty, capturing spatial boundary errors. Concurrently, we employ an Evidential Deep Learning (EDL) module to model epistemic uncertainty directly from the network outputs, allowing the network to detect regions with zero structural evidence for a layer. Our framework addresses the largely overlooked challenge of anatomical absence in boundary regression, combining sub-pixel localization with direct atrophy segmentation.

1 Introduction

Optical Coherence Tomography (OCT) is the gold standard for managing sight-threatening diseases like age-related macular degeneration (AMD) [1, 2]. Extracting fine retinal layer thicknesses is crucial for tracking disease progression. State-of-the-art layer segmentation increasingly utilizes boundary regression [3, 4], which achieves sub-pixel accuracy and implicitly prevents multi-surface topological violations by computing the expected vertical position from a column-wise probability mass function.

However, this formulation presents a fundamental limitation: it assumes every layer is present across the entire scan. Because the column-wise softmax must sum to one, the network falsely predicts boundary coordinates even when structural evidence is entirely absent. Detecting missing layers is highly relevant clinically, most notably in Geographic Atrophy (GA), where the localized loss of the photoreceptors and retinal pigment epithelium (RPE) serves as the primary endpoint for novel therapeutics [5].

Historically, automated atrophy detection follows three paradigms. *Direct lesion segmentation* identifies GA footprints via classification on *en face* projections [6, 7] or cross-sectional B-scans [8], but ignores the 3D layer geometry needed to monitor early structural thinning. *Pixel-wise layer segmentation* can theoretically capture atrophy by omitting the RPE prediction locally, yet lacks anatomical coherence, risking topological violations such as reversed layers or implausible holes.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

Conversely, *boundary regression* guarantees strict topology but its continuous-surface constraint prevents it from accurately omitting missing layers.

To address this gap within the boundary regression paradigm, we introduce an uncertainty-aware extension to the topologically constrained SD-RetinaNet. We explicitly quantify two sources of uncertainty directly from the network's spatial probability maps. First, we model *aleatoric uncertainty* using a Gaussian Negative Log-Likelihood (NLL) loss to calibrate the spatial variance of the boundary predictions, capturing ambiguity from speckle noise, low contrast, and shadowing from overlying hyperreflective areas. Second, we integrate epistemic uncertainty via Evidential Deep Learning (EDL) [9]. By placing a Dirichlet prior over the 1D spatial column, the network estimates the total structural evidence and identifies out-of-distribution scenarios where layers are missing. To our knowledge, this is the first evidential boundary regression framework to detect anatomical absence, bridging high-precision topology with robust atrophy detection.

2 Related Work

Retinal Biomarker Segmentation and Atrophy: Recent deep learning methods for OCT predominantly address layer and fluid lesion segmentation [10, 11, 4]. To overcome topological violations common in pixel-wise U-Nets, He et al. [3] introduced layer boundary regression (LBRM). While SD-RetinaNet [4] successfully integrated lesion constraints into this framework, neither method accounts for layers that vanish due to GA. Detection of GA has mostly been handled via direct regional segmentation [7], disconnected from sub-pixel layer morphology. Adapting deep regression models to correctly omit layers remains an unsolved challenge.

Uncertainty Quantification in Deep Learning: Uncertainty in deep learning is categorized into aleatoric and epistemic uncertainty [12]. While Monte Carlo Dropout (MCD) captures epistemic uncertainty, it requires expensive multiple forward passes. Evidential Deep Learning (EDL) instead formulates learning as an evidence acquisition process [9, 13], quantifying epistemic uncertainty in a single forward pass. By adapting EDL to place Dirichlet priors over spatial column-wise probability maps, we leverage it to act as an anomaly detector for anatomical absence.

3 Methodology

Base Architecture: Our framework builds upon the SD-RetinaNet backbone [14], which employs a U-Net architecture equipped with an EfficientNet [15] encoder. To extract anatomical boundaries, the network utilizes an anatomy module that projects high-dimensional feature maps into 1D column-wise boundary representations over the image height H. Instead of enforcing a strict softmax normalization that forces probability maps to sum to one, we reframe the spatial localization task as a hybrid discrete-to-continuous evidential learning problem.

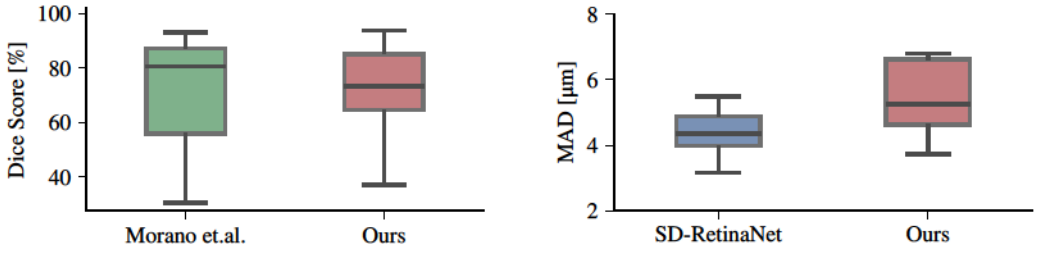
Discrete-to-Continuous Evidential Distribution: While standard boundary regression directly predicts a continuous value, our model treats the H vertical depth bins of each A-scan as a discrete probability space. To model epistemic uncertainty, we place a Dirichlet prior over these spatial bins. The network outputs non-negative evidence $e_h \geq 0$ for each bin $h \in H$, forming the Dirichlet parameters $\alpha_h = e_h + 1$. The total structural evidence is $S = \sum_{h=1}^{H} \alpha_h$, and the expected spatial probability mass function (PMF) is $p_h = \alpha_h / S$.

Uncertainty Calibration and Missing Layer Detection: We bridge this discrete evidential distribution to continuous boundary regression via spatial expectation. The sub-pixel boundary position is computed as $\mu = \sum_{h=1}^{H} p_h \cdot h$, and its aleatoric spatial variance is derived as $\sigma^2 = \sum_{h=1}^{H} p_h \cdot (h-\mu)^2$. This variance naturally expands in regions obscured by speckle noise or shadowing from hyperreflective areas. *Epistemic uncertainty* is dictated by the total evidence S. We leverage this to formulate our missing layer decision rule: if the predicted evidence S falls below a predefined threshold τ (empirically determined on a held-out validation set of 110 volumes) the layer boundary is classified as structurally missing. Otherwise, the network applies strict topological corrections to ensure non-crossing constraints among the valid layers.

Joint Optimization: The network is optimized using a composite loss function conditioned by a binary ground-truth mask indicating layer presence. In regions where the layer exists, we minimize the Gaussian Negative Log-Likelihood (NLL) utilizing the derived continuous spatial moments μ

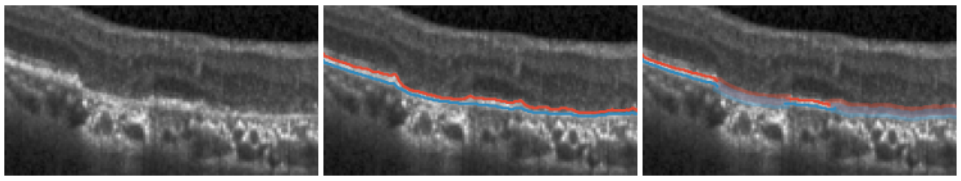
and σ^2 against the target coordinate. This calibrates the spatial variance directly to the prediction error. We augment this with an evidence-maximizing objective that explicitly encourages the network to predict high total structural evidence when the layer is present. Conversely, in regions where the layer is anatomically missing, we apply an annealed Kullback-Leibler (KL) divergence penalty. This forces the predicted Dirichlet parameters toward a uniform, zero-evidence distribution ($KL(\alpha||1)$), effectively pushing the total evidence S to zero and training the network to confidently predict pathological absence.

4 Experimental Setup and Results

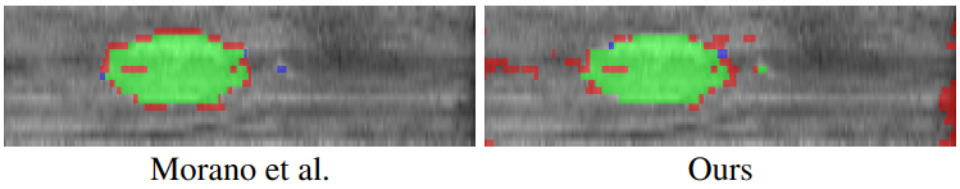


(a) GA Segmentation Performance (OCTAVE)

(b) Layer Segmentation (AROI)



(c) Boundary and atrophy detection visualization



(d) En-face GA footprint comparison

Figure 1: Comparative performance analysis. (a) Segmentation performance of the 2D projected atrophy area (GA footprint) against specialized baselines. (b) Maintenance of sub-pixel boundary localization accuracy compared to standard topological regression. (c) Qualitative B-scan highlighting the proposed method's ability to identify missing layers (indicated by the transparent colored overlay). (d) En-face visualization of GA segmentation compared to ground truth (Green: correct, Red: over-segmentation, Blue: under-segmentation).

Dataset and Training Setup: We trained the proposed evidential model, the baseline SD-RetinaNet [4], and a state-of-the-art specialized 3D-to-2D GA segmentation model (Morano et al. [8]) on a shared private cohort consisting of 1100 OCT volumes with manual layer and lesion annotations. Of these, 55 volumes exhibited GA. The scans were acquired using Heidelberg Spectralis, Zeiss Cirrus, and Heidelberg HighRes devices, ensuring multi-vendor robustness.

Experiment 1: Geographic Atrophy Footprint Detection: To evaluate the network's ability to detect missing layers, we tested the atrophy en-face footprint extraction performance on the independent public OCTAVE dataset [16], comprising 198 Spectralis OCT volumes (3762 B-scans), 9 of which contain GA. Evaluated specifically on these 9 GA-positive volumes, our evidential boundary regression achieved highly competitive GA footprint segmentation performance with a Dice score of 0.6959 ± 0.1878 and a PR-AUC of 0.8746. This is directly comparable to the specialized direct-segmentation method by Morano et al. [8], which yielded a Dice of 0.7102 ± 0.2101 and a

PR-AUC of 0.9297. This demonstrates that our 1D epistemic uncertainty mechanism successfully bridges boundary regression and region-based GA footprint detection without requiring a separate 3D pixel-wise classifier or the multiple inference passes required by Monte Carlo Dropout (MCD) (see Figure 1a).

Experiment 2: Layer Segmentation Maintenance: To confirm that adding uncertainty constraints does not severely degrade standard topological layer regression, we evaluated boundary localization on the public AROI dataset [17], consisting of 1136 annotated B-scans from 24 AMD patients featuring three manually annotated retinal layers. The proposed method demonstrated a Mean Absolute Distance (MAD) of 5.85 ± 2.82 μm, compared to the original SD-RetinaNet baseline performance of 4.68 ± 2.29 μm. These results demonstrate that while the introduction of evidential constraints yields a slight increase in localization error (from 4.68 μm to 5.85 μm), this represents an acceptable and necessary trade-off; sub-pixel precision remains highly robust while enabling the critical clinical capability of detecting missing pathological layers (see Figure 1b).

Experiment 3: Uncertainty Calibration: A core contribution of our Gaussian NLL formulation is superior spatial error calibration. We evaluated the aleatoric calibration on the AROI dataset by computing the Expected Calibration Error (ECE) and the Pearson correlation (r) between the predicted standard deviation (σ) and the absolute boundary error (MAD). Our evidential framework achieved an ECE of 4.1 μm and a strong correlation of $r = 0.4798$ ($p < 0.001$), vastly outperforming the uncalibrated softmax entropy of the baseline SD-RetinaNet (ECE: 6.2 μm, $r = 0.3812$, $p < 0.001$). This indicates that our predicted aleatoric variance serves as a highly reliable proxy for boundary ambiguity caused by speckle noise and shadowing.

Qualitative Evaluation: Visually (Figure 1c), the baseline SD-RetinaNet incorrectly interpolates continuous boundaries across atrophic regions, maintaining false confidence despite the lack of structural evidence. In contrast, our evidential framework correctly drops boundary predictions within GA lesions. Furthermore, the predicted aleatoric variance (σ) visibly widens at lesion margins and areas of poor signal, faithfully representing structural ambiguity. In the en-face projections (Figure 1d), while our method accurately captures the core GA footprint, the observed over-segmentation primarily stems from detecting missing layers near the peripheral borders of the B-scans. These false positives correlate strongly with signal roll-off and scanning artifacts rather than true anatomical atrophy.

5 Conclusion

In this work, we introduced a novel uncertainty-aware framework that successfully bridges the gap between spatial evidential deep learning and 1D topological boundary regression. By explicitly modeling aleatoric spatial variance via a Gaussian NLL formulation and quantifying epistemic structural evidence through Dirichlet priors, our model overcomes the critical limitation of previous boundary regression techniques: the mathematical obligation to hallucinate layers in pathological regions. The proposed method reliably detects structural layer absence, preventing erroneous boundary predictions in areas of complete tissue loss like Geographic Atrophy, while providing highly calibrated error estimates in regions obscured by shadowing. Ultimately, this framework combines high-precision sub-pixel layer tracking with robust 2D atrophy footprint extraction in a single pass. While this extended abstract presents an ongoing work in progress, it establishes a reliable and mathematically sound paradigm for automated pathological OCT analysis. Current efforts are actively focused on further improving both the layer segmentation accuracy and the GA footprint detection performance. Future work will evaluate the framework on larger GA-positive cohorts to strengthen clinical claims, and develop spatial priors to counteract false positive absence detections near peripheral scan borders.

Acknowledgements

The financial support by the Christian Doppler Research Association, Austrian Federal Ministry of Economy, Energy and Tourism, the National Foundation for Research, Technology and Development, and Heidelberg Engineering is gratefully acknowledged.

References

[1] Neil M. Bressler. Age-Related Macular Degeneration Is the Leading Cause of Blindness... *Jama*, 291(15):1900–1901, 2004.

[2] U. Schmidt-Erfurth, S. Klimscha, S. M. Waldstein, and H. Bogunovic. A view of the current and future role of optical coherence tomography in the management of age-related macular degeneration. *Eye*, 31(1):26–44, January 2017. ISSN 1476-5454. doi: 10.1038/eye.2016.227.

[3] Yufan He, Aaron Carass, Yihao Liu, Bruno M. Jedynak, Sharon D. Solomon, Shiv Saidha, Peter A. Calabresi, and Jerry L. Prince. Structured layer surface segmentation for retina OCT using fully convolutional regression networks. *Medical Image Analysis*, 68:101856, February 2021. ISSN 1361-8415. doi: 10.1016/j.media.2020.101856.

[4] Botond Fazekas, Guilherme Aresta, Dmitrii Lachinov, Sophie Riedl, Julia Mai, Ursula Schmidt-Erfurth, and Hrvoje Bogunovic. SD-LayerNet: Semi-supervised Retinal Layer Segmentation in OCT Using Disentangled Representation with Anatomical Priors. In Linwei Wang, Qi Dou, P. Thomas Fletcher, Stefanie Speidel, and Shuo Li, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*, Lecture Notes in Computer Science, pages 320–329, Cham, 2022. Springer Nature Switzerland. ISBN 9783031164521. doi: 10.1007/978-3-031-16452-1_31.

[5] Spencer C. Cleland, Sri Meghana Konda, Ronald P. Danis, Yijun Huang, Dawn J. Myers, Barbara A. Blodi, and Amitha Domalpally. Quantification of Geographic Atrophy Using Spectral Domain OCT in Age-Related Macular Degeneration. *Ophthalmology Retina*, 5(1):41–48, January 2021. ISSN 2468-7219, 2468-6530. doi: 10.1016/j.oret.2020.07.006.

[6] Dmitrii Lachinov, Philipp Seeböck, Julia Mai, Felix Goldbach, Ursula Schmidt-Erfurth, and Hrvoje Bogunovic. Projective Skip-Connections for Segmentation Along a Subset of Dimensions in Retinal OCT. In Marleen de Bruijne, Philippe C. Cattin, Stéphane Cotin, Nicolas Padoy, Stefanie Speidel, Yefeng Zheng, and Caroline Essert, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*, pages 431–441, Cham, 2021. Springer International Publishing. ISBN 9783030871932. doi: 10.1007/978-3-030-87193-2_41.

[7] Bart Liefers, Paul Taylor, Abdulrahman Alsaedi, Clare Bailey, Konstantinos Balaskas, Narendra Dhingra, Catherine A. Egan, Filipa Gomes Rodrigues, Cristina González Gonzalo, Tjebo F. C. Heeren, Andrew Lotery, Philipp L. Müller, Abraham Olvera-Barrios, Bobby Paul, Roy Schwartz, Darren S. Thomas, Alasdair N. Warwick, Adnan Tufail, and Clara I. Sánchez. Quantification of Key Retinal Features in Early and Late Age-Related Macular Degeneration Using Deep Learning. *American Journal of Ophthalmology*, 226:1–12, June 2021. ISSN 0002-9394. doi: 10.1016/j.ajo.2020.12.034.

[8] José Morano, Guilherme Aresta, Dmitrii Lachinov, Julia Mai, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. Self-supervised learning via inter-modal reconstruction and feature projection networks for label-efficient 3d-to-2d segmentation. In Hayit Greenspan, Anant Madabhushi, Parvin Mousavi, Septimiu Salcudean, James Duncan, Tanveer Syeda-Mahmood, and Russell Taylor, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, pages 589–599, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-43901-8.

[9] Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential Deep Learning to Quantify Classification Uncertainty. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/a981f2b708044d6fb4a71a1463242520-Abstract.html.

[10] Abhijit Guha Roy, Sailesh Conjeti, Sri Phani Krishna Karri, Debdoot Sheet, Amin Katouzian, Christian Wachinger, and Nassir Navab. ReLayNet: retinal layer and fluid segmentation of macular optical coherence tomography using fully convolutional networks. *Biomedical Optics Express*, 8(8):3627–3642, August 2017. ISSN 2156-7085. doi: 10.1364/BOE.8.003627.

[11] Hrvoje Bogunovic, Freerk Venhuizen, Sophie Klimscha, Stefanos Apostolopoulos, Alireza Bab-Hadiashar, Ulas Bagci, Mirza Faisal Beg, Loza Bekalo, Qiang Chen, Carlos Ciller, Karthik Gopinath, Amirali K. Gostar, Kiwan Jeon, Zexuan Ji, Sung Ho Kang, Dara D. Koozekanani,

Donghuan Lu, Dustin Morley, Keshab K. Parhi, Hyoung Suk Park, Abdolreza Rashno, Marinko Sarunic, Saad Shaikh, Jayanthi Sivaswamy, Ruwan Tennakoon, Shivin Yadav, Sandro De Zanet, Sebastian M. Waldstein, Bianca S. Gerendas, Caroline Klaver, Clara I. Sánchez, and Ursula Schmidt-Erfurth. RETOUCH: The Retinal OCT Fluid Detection and Segmentation Benchmark and Challenge. *IEEE Transactions on Medical Imaging*, 38(8):1858–1874, August 2019. ISSN 1558-254X. doi: 10.1109/TMI.2019.2901398.

[12] Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/2650d6089a6d640c5e85b2b88265dc2b-Abstract.html.

[13] Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep Evidential Regression. In *Advances in Neural Information Processing Systems*, volume 33, pages 14927–14937. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/aab085461de182608ee9f607f3f7d18f-Abstract.html.

[14] Botond Fazekas, Guilherme Aresta, Philipp Seeböck, Julia Mai, Ursula Schmidt-Erfurth, and Hrvoje Bogunović. SD-RetinaNet: Topologically Constrained Semi-Supervised Retinal Lesion and Layer Segmentation in OCT. *IEEE Transactions on Medical Imaging*, pages 1–1, 2025. ISSN 1558-254X. doi: 10.1109/TMI.2025.3615240. URL https://ieeexplore.ieee.org/document/11186218.

[15] Mingxing Tan and Quoc Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning*, pages 6105–6114. PMLR, May 2019.

[16] Daniel S. Kermany, Wesley Poon, Anaya Bawiskar, Natasha Nehra, Orhun Davarci, Glori Das, Matthew Vasquez, Shlomit Schaal, Raksha Raghunathan, and Stephen T. C. Wong. Identifying Retinal Features Using a Self-Configuring CNN for Clinical Intervention. *Investigative Ophthalmology & Visual Science*, 66(6):55, June 2025. ISSN 1552-5783. doi: 10.1167/iovs.66.6.55.

[17] M. Melinščak, M. Radmilovič, Z. Vatavuk, and S. Lončarić. Aroi: Annotated retinal oct images database. In *2021 44th International Convention on Information, Communication and Electronic Technology (MIPRO)*, pages 371–376, 2021. doi: 10.23919/MIPRO52101.2021.9596934.

Evaluation of Anatomical Shape Priors in Deep Learning-Based Cardiac Multi-Compartment Segmentation

Michael Hudler, Franz Thaler, Martin Urschler
Institute for Medical Informatics, Statistics and Documentation
Medical University of Graz
{michael.hudler, martin.urschler}@medunigraz.at

Abstract

Whole-heart multi-compartment CT segmentation is clinically important, but standard CNNs do not explicitly enforce anatomical plausibility. Based on statistics derived from the training data, we evaluate whether lightweight explicit shape priors, implemented as shape-aware losses and spatial label distribution heatmap-guided U-Net variants, improve 3D cardiac segmentation on MM-WHS CT and WHS++. Across all experiments, a standard 3D U-Net surprisingly remained a very strong baseline, with handcrafted priors yielding at best marginal and inconsistent changes and often degrading performance. These results suggest that the baseline already captures substantial implicit anatomical regularities and that future gains will likely require more expressive learned priors rather than simple handcrafted anatomical shape constraints.

1 Introduction

Multi-compartment whole-heart segmentation from cardiac CT [1] is a core task in medical image analysis because it supports quantitative assessment of clinical parameters like ejection fraction, builds the foundation for treatment planning or simulation [2], and enables image-guided interventions. Deep learning [3], especially 3D U-Net variants [4, 5], has become the dominant approach for this problem due to its strong multiscale feature extraction capabilities and its support for accurate localization [6]. Their success is explained by the clever combination of encoder-decoder feature extraction, multiscale context modeling, and skip connections that preserve spatial detail. However, these models are primarily appearance-driven and do not explicitly encode anatomical shape knowledge as was prominently done in the pre-deep learning era via statistical shape models [7, 8]. This gap motivates the study of shape priors in deep learning-based segmentation [9].

This work evaluates whether explicit shape priors improve whole-heart multi-compartment CT segmentation beyond a strong 3D U-Net baseline. Rather than tediously building a full statistical shape-modeling pipeline, our study tests lightweight priors that can be incorporated directly into the training objective or 3D network design. The central question is whether such priors provide measurable benefit on modern deep learning baselines for seven-class cardiac CT segmentation, involving ventricles, both atria, myocardium, and the great vessels. The main finding is negative but clear: in the studied setting, explicitly designed shape priors did not consistently improve performance over a well-trained 3D U-Net baseline.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

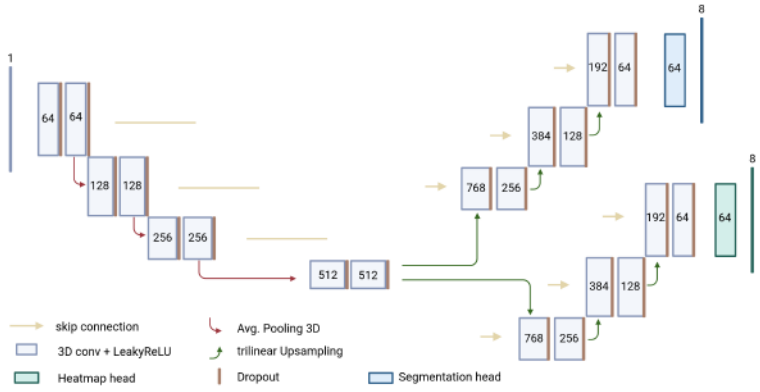


Figure 1: Exemplary architecture of one of our proposed networks incorporating label distribution heatmaps: The 2-Decoder network, with separate decoders for label segmentation and heatmap predictions, thus forcing the encoder to extract features supporting both predictions.

2 Methods

Dataset and preprocessing. Experiments are based primarily on the CT subset of the MM-WHS challenge dataset [1], comprising seven foreground classes: left ventricle, right ventricle, left atrium, right atrium, myocardium, ascending aorta, and pulmonary artery. The benchmark provides 20 annotated CT scans for training and 40 CT test scans evaluated with the official hidden-label evaluation script used for the Challenge. To complement this with accessible ground truth, evaluation was extended using the second half of the WHS++ training set (20 CT cases), which corresponds to the publicly released extension of the MM-WHS CT dataset. Images and labels were reoriented to a common anatomical convention, resampled isotropically, centered using label centroids, and embedded into a standardized field of view. A Procrustes-based alignment [10] of training labels was further used to derive population heatmaps representing average spatial label distributions in the registered space.

Baseline model. The reference model is a standard 3D U-Net [4, 5] with single-channel CT input and eight output classes including background. It uses a conventional encoder-decoder design with skip connections with 64 base channels and doubling the number of channels at each downsampling step. It uses LeakyReLU activations [11] and is trained with a combined Generalized Dice [12] and Cross-Entropy loss. This baseline serves as the main point of comparison throughout the study.

Shape-aware losses. Three families of explicit regularizers were evaluated in combination with the baseline loss. *Volume regularization* penalizes deviations from expected compartment volumes estimated from the training set via the label-specific volume means and standard deviations. Moment-based *shape regularization* compares soft first- and second-order spatial moments of predictions to reference shape moments (centroids, ellipsoids) from the training set via L2 distance. *Anatomical relation* loss constrains pairwise distances and angular relations between class centroids via reference angle statistics derived from the training data. All losses aim to inject coarse anatomical prior knowledge without changing the overall segmentation backbone.

Architectural priors. In addition to loss-based priors, we also investigated population-level multi-class probability heatmaps derived from aligned labels of the training dataset. These shape priors were integrated into several U-Net variants: a model with an auxiliary heatmap prediction head attached to the last decoder layer, a multilayer deep-supervision version of the latter architecture (*HM multilayer*), a two-decoder network with separate segmentation and heatmap branches (*2-Decoder*, see exemplary architecture depicted in Fig. 1), a dual-encoder network that processes image and heatmap inputs in parallel (*2-Encoder*), and a cascaded three-U-Net architecture for coarse prediction and refinement (*Cascaded*).

Experimental setup. Models were trained on cropped regions of interest at 64^3 and 128^3 input resolution using the same extensive geometric and intensity data augmentation for all architectures. Evaluation used Dice, Jaccard, Hausdorff distance (HD), and Average Symmetric Surface

Distance (ASSD), as overlap- and boundary-based metrics, respectively. Qualitative comparisons were additionally assessed on WHS++, since the MM-WHS test set did not provide ground truth segmentations.

3 Results

Table 1 summarizes the main findings. More results, additional descriptions of methods and implementation details can be found in [13]. On MM-WHS at 64^3, the baseline achieved 90.85% Dice, 83.63% Jaccard, 7.64 mm HD, and 1.03 mm ASSD. Volume and moment regularization were essentially tied with the baseline (90.85% and 90.84% Dice), whereas the anatomical relation loss reduced performance to 88.98% Dice. Thus, simple handcrafted losses did not harm but also did not improve the already strong baseline.

Table 1: Main quantitative results. MM-WHS values report Dice, Jaccard, HD, and ASSD; WHS++ reports Dice only, as summarized in the thesis. Best values per block are in bold.

Setting Method	Dice (%)	Jaccard (%)	HD (mm)	ASSD (mm)
MM-WHS CT, 64^3, shape-aware losses				
Baseline	**90.85**	**83.63**	**7.64**	**1.03**
Volume regularization	**90.85**	83.62	7.70	1.04
Moment regularization	90.84	83.60	7.67	**1.03**
Anatomical relation	88.98	80.65	8.23	1.27
MM-WHS CT, 64^3, selected architectural priors				
Baseline	**90.85**	**83.63**	7.64	**1.03**
HM multilayer	90.60	83.23	7.78	1.06
2-Decoder	90.73	83.43	7.58	1.06
Cascaded	90.32	82.74	**7.55**	1.08
MM-WHS CT, 128^3, selected architectural priors				
Baseline	**92.05**	**85.78**	7.35	**0.88**
HM multilayer	91.80	85.38	7.28	0.90
2-Encoder	92.02	85.70	7.40	0.89
Cascaded	92.04	85.70	**7.26**	0.89
WHS++ CT, 64^3, shape-aware losses				
Baseline	88.93	81.01	18.47	1.68
Volume regularization	89.09	81.26	**17.91**	1.63
Moment regularization	**89.16**	**81.47**	18.31	**1.61**
Anatomical relation	88.66	80.67	18.13	1.70
WHS++ CT, 64^3, selected architectural priors				
Baseline	**88.93**	**81.01**	**18.47**	**1.68**
HM multilayer	88.72	80.63	20.88	1.73
2-Encoder	87.75	79.43	18.52	1.77
Cascaded	86.13	77.05	21.27	2.09

At the architectural level, heatmap-guided models remained competitive but did not clearly surpass the reference U-Net. At 64^3, the 2-Decoder variant reached 90.73% Dice and the cascaded model produced the best HD (7.55 mm), suggesting slightly improved boundary refinement, but overall overlap remained below baseline. At 128^3, the baseline improved to 92.05% Dice, while the closest competitors, 2-Encoder and Cascaded, achieved 92.02% and 92.04%, respectively. Hence, higher resolution improved most models, but not our main finding.

Evaluation on WHS++ confirmed the same trend. The baseline achieved 88.93% Dice, while volume and moment regularization yielded only marginal changes, reaching 89.09% and 89.16% Dice, respectively. Moment regularization produced the best Dice, Jaccard, and ASSD, whereas volume regularization achieved the lowest HD (17.91 mm). However, these improvements were small and inconsistent. Architectural prior-based models did not outperform the baseline: HM multilayer and 2-Encoder showed slightly lower overlap, and the cascaded architecture degraded performance across all metrics. Overall, the main finding remained consistent across datasets and model families.

Figure 2: Representative qualitative comparison on WHS++ (subject 2014, coronal slice 76). The baseline U-Net, the best loss-based prior (Mean-Shape), and the best architecture-based prior (2-Decoder) all capture the global anatomy well. Differences are mainly confined to boundaries and smaller structures.

Qualitatively, all models reproduced the overall cardiac configuration and inter-structure arrangement well (Fig. 2). The shape-aware variants appeared visually very similar to the baseline, with differences concentrated at boundaries and in thinner structures rather than in gross anatomical localization. This visual pattern is consistent with the quantitative results: the baseline already learned strong global anatomical regularities, leaving limited room for coarse handcrafted priors to add useful information.

4 Discussion and Conclusions

The central result of this study is that, surprisingly, explicit handcrafted shape priors did not consistently outperform a strong 3D U-Net baseline for whole-heart CT segmentation, which is performing in-line with the winner of the MM-WHS Challenge (see [6, 1]) as well as the participants at the Challenge associated with WHS++ [14]. This is a relevant negative result. It suggests that on MM-WHS and WHS++, the baseline model already learns substantial implicit anatomical regularities directly from the image data, and that coarse constraints such as expected volumes, low-order moments, centroid relations, or average label distribution heatmaps add little information beyond that baseline.

Several factors likely explain this outcome. First, the tested priors describe anatomy only at a coarse level and cannot capture the complex nonlinear variability of cardiac shape. Second, the remaining errors are mainly boundary-related, whereas the priors regularize global structure more strongly than local boundary detail. Third, because baseline performance is already high, measurable improvements are inherently limited. The experiments also show that greater architectural complexity does not automatically improve segmentation: models such as 2-Encoder and Cascaded remained competitive, and Cascaded slightly improved HD, but none clearly surpassed the simpler baseline in overall Dice or boundary overlap.

In summary, this work provides a focused evaluation of explicit shape priors in deep learning-based whole-heart segmentation and shows that simple handcrafted priors are insufficient to reliably improve a strong 3D U-Net. Future work should therefore move toward more expressive learned anatomical priors, such as generative diffusion-based [15] or flow matching-based [16] models trained on segmentation masks, which may better represent the distribution of plausible cardiac anatomy.

Acknowledgments and Disclosure of Funding

This research was funded in whole or in part by the Austrian Science Fund (FWF) 10.55776/PAT1748423.

References

[1] Xiahai Zhuang, Lei Li, Christian Payer, Darko Štern, Martin Urschler, Mattias P Heinrich, Julien Oster, Chunliang Wang, Örjan Smedby, Cheng Bian, Xin Yang, Pheng-Ann Heng, Aliasghar Mortazi, Ulas Bagci, et al. Evaluation of algorithms for multi-modality whole heart segmentation: An open-access grand challenge. *Medical Image Analysis*, 58:101537, December 2019.

[2] Elena Zappon, Luca Azzolin, Matthias A F Gsell, Franz Thaler, Anton J Prassl, Robert Arnold, Karli Gillette, Mohammadreza Kariman, Martin Manninger, Daniel Scherr, Aurel Neic, Martin

Urschler, Christoph M Augustin, Edward J Vigmond, and Gernot Plank. An efficient end-to-end computational framework for the generation of ECG calibrated volumetric models of human atrial electrophysiology. *Medical Image Analysis*, 107(Pt B):103822, October 2025.

[3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer International Publishing, 2015.

[5] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3D U-net: Learning dense volumetric segmentation from sparse annotation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016*, Lecture notes in computer science, pages 424–432. Springer International Publishing, Cham, 2016.

[6] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Multi-label whole heart segmentation using CNNs and anatomical label configurations. In *Statistical Atlases and Computational Models of the Heart. ACDC and MMWHS Challenges*, pages 190–198. Springer International Publishing, 2018.

[7] Tim F Cootes, Chris J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer Vision and Image Understanding*, 61(1):38–59, January 1995.

[8] Tobias Heimann and Hans-Peter Meinzer. Statistical shape models for 3D medical image segmentation: a review. *Medical Image Analysis*, 13(4):543–563, August 2009.

[9] Simon Bohlender, Ilkay Oksuz, and Anirban Mukhopadhyay. A survey on shape-constraint deep learning for medical image segmentation. *IEEE Reviews in Biomedical Engineering*, 16:225–240, January 2023.

[10] Peter H Schoenemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31:1–10, 1966.

[11] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning (ICML)*, volume 30, 2013.

[12] Carole H Sudre, Wenqi Li, Tom Vercauteren, Sebastien Ourselin, and M Jorge Cardoso. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support - DLMIA ML-CDS 2017*, volume 10553 of *Lecture Notes in Computer Science*, pages 240–248. Springer, September 2017.

[13] Michael Hudler. *Evaluation of Shape Models for Deep Learning Based Cardiac Image Segmentation.* Master's Thesis, Graz University of Technology, Graz, Austria, 2026.

[14] Franz Thaler, Darko Štern, Gernot Plank, and Martin Urschler. Augmentation-based domain generalization and joint training from multiple source domains for whole heart segmentation. In *Comprehensive Analysis and Computing of Real-World Medical Images. CARE 2024*, volume 15548 of *Lecture notes in computer science*, pages 168–179. Springer Nature Switzerland, Cham, 2025.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020.

[16] Arnela Hadzic, Lea Bogensperger, Andrea Berghold, and Martin Urschler. Flow matching-based data synthesis for robust anatomical landmark localization. *IEEE Journal of Biomedical and Health Informatics*, 2025.

Forecasting individual survival in irregularly sampled patient trajectories

Daniel Sobotka[1,2,3], Nino Bogveradze[2], Lucian Beer[2], Philipp Seeböck[1,2,3], Helmut Prosch[2] and Georg Langs[1,2,3]

[1] Computational Imaging Research Lab,
Department of Biomedical Imaging and Image-guided Therapy,
Medical University of Vienna, Vienna, Austria
[2] Christian Doppler Laboratory for Machine Learning Driven Precision Imaging,
Department of Biomedical Imaging and Image-guided Therapy,
Medical University of Vienna, Vienna, Austria
[3] Comprehensive Center for Artificial Intelligence in Medicine,
Medical University of Vienna, Vienna, Austria

Abstract

Time series forecasting of patient trajectories plays a critical role in the clinical environment by enabling the prediction of possibly treatment relevant patient events . Clinical data such as imaging studies, surgical records, laboratory measurements or tumor staging provide rich longitudinal information reflecting the progression of disease or treatment response. Modeling these data involves several challenges such as integrating multi-modal data, handling irregularly sampling over time, or managing missing values. Many existing forecasting approaches rely on regularly sampled data and perform poorly when facing irregularly sampled clinical data. Here, we evaluate three different deep learning models for predicting individual six month survival from irregularly sampled lung cancer patient trajectories. Results show that state-of-the-art models can integrate sparse clinical data and benefit from multi-modality, improving forecasting of clinical outcomes despite irregular sampling patterns.

1 Introduction

The prediction of future time series from observed partial time series is a widely studied problem across many academic fields, such as climate modeling [6] or biological sciences [8]. In medical research, the prediction of future patient trajectories is critical for improving clinical decision-making and patient outcomes. Time-series models are increasingly being employed to predict key health events such as the onset of heart failure [2] or patient mortality [10]. By modeling these temporal health trajectories, clinicians can anticipate adverse events, optimize treatment plans, and provide more personalized care. Incorporating clinical relevance in time series forecasting not only enhances predictive accuracy of events but also directly impacts patient treatment planing. Time-series forecasting in the medical domain presents unique challenges that distinguish it from other fields. Clinical data are often *multi-modal*, containing a variety of information from imaging data, laboratory results to surgery information. Furthermore, clinical events are typically *irregularly sampled*, with visits occurring at non-uniform intervals based on patient conditions. Finally, prediction models need to cope with *missing data* due to clinical decisions determining the modalities acquired during an examination. Compared to conventional statistical models such as auto-regressive [11] models, machine- and deep learning models offer promising results in addressing time series forecasting challenges, such as multi-objective or multi-modality forecasting [5]. Deep learning models can

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

be grouped into 1) encoder-decoder-based, 2) transformer-based and 3) GAN-based architectures, where recent research focused on transformer based deep learning time forecasting approaches [4]. Transformer-based models leverage self-attention and multi-head attention mechanisms. They have emerged as one of the most effective architectures for capturing semantic dependencies in long input sequences [9]. Beside that classification other deep learning models are focusing on forecasting, such as *DeepSurv*, a multi-layer perceptron implementing the Cox proportional hazards model [3]. *Recurrent neural networks (RNNs)* are also considered suitable for sequence modeling and can solve time series related tasks. [4]. [1] introduced *GRU-D*, an extension of the gated recurrent unit (GRU) that incorporates decay mechanisms to model multivariate time series data with missing values.

2 Methods

We evaluate a patch-based time series transformer (PatchTST) [7], GRU-D [1] and DeepSurv [3] for using multi-modal patient trajectories as basis of forecasting for survival prediction within the 6 months (24 weeks) following the last observed time point. The networks take look-back windows $L_n \mapsto (v_1, \dots v_n)$ as input and predict survival probabilities at 24 future weekly time points $t = 1, \dots, 24$. PatchTST consists of an transformer encoder that maps input patches into a latent space using a trainable linear projection and positional encoding, which encodes the temporal position of each patch in the input sequence, followed by multi-head self-attention layers. The resulting latent representations are flattened and passed through a linear layer with ReLu activation to obtain predictions $\hat{y}_{b,t}$ $(t = 1, \dots, 24)$ for each sample b in the batch. During training we ignore empty patches and use a value mask for missing values inside each patch. GRU-D is a recurrent neural network based on gated recurrent units and decay mechanisms to model past observation fading over time and how missing values should be imputed dynamically. Similar to the transformer, missing input values are masked. To account for imbalances in the survival data, a binary cross-entropy (BCE) with logits loss weighted by a sample-wise weight based on the number of non-survival weeks per sample is proposed. The sample-wise weight for batch b is defined as

$$w_b = 1 + \sum_{t=1}^{T} (1 - y_{b,t}), \tag{1}$$

where T denotes the number of 24 forecasting weeks and $y_{i,t}$ the corresponding ground truth target. The final loss is computed as

$$\mathcal{L} = \frac{1}{BT} \sum_{b=1}^{B} \sum_{t=1}^{T} w_b \cdot BCEWithLogits(\hat{y}_{b,t}, y_{b,t}) \tag{2}$$

where B denotes the batch size. In contrast to PatchTST and GRU-D, which are designed for time-series representation learning and prediction, DeepSurv optimizes the negative log partial likelihood of the Cox proportional hazards model to learn a risk score that reflects the relative hazard of an event while explicitly handling censored observations.

3 Data

In this study, a fixed temporal window of 3072 days was defined for each of the 4642 patients, corresponding to the longest observation period recorded within the clinical cohort. Each patient's temporal sequence was aligned such that the final day of the window (day 3072) represented the patient's most recent clinical encounter, which could include one or more modalities such as computed tomography (CT) imaging extracted lung tumor volume, surgical interventions, laboratory assessments or ICD diagnosis. For patients with shorter observation periods or sparse visit histories (e.g., visits limited to a single month), the earlier portion of the window was zero-padded, while the final segment contained the available clinical data. This alignment strategy ensured temporal consistency across patients and facilitated comparability independent of follow-up duration. Across the cohort, patients underwent a total of 10747 CT scans (average 2.32 per patient) and 1981 surgical procedures (0.43 per patient). Laboratory measurements included 126572 CRP tests (27.27 per patient), 110880 albumin tests (23.89 per patient), 20359 leukocyte counts (4.39 per patient), 42019 hemoglobin measurements (9.05 per patient), 42027 hematocrit measurements (9.05 per patient), and 112892 creatinine measurements (24.32 per patient). Diagnoses included 31241 C34* ICD-10 codes (6.73 per patient), and tumor

staging was recorded 237 times (0.05 per patient). From each CT scan we extracted with a pretrained 3D U-Net lung tumor annotations and used that information as tumor load input channel. We used a train/val/test ratio of $0.7/0.15/0.15$ resulting in 3249 patients for training, 696 for validation and 697 for testing. A total of 2584 patients (55.67%) are non survival during the next 24 week prediction horizon, where 2058 patients (44.33%) survive the 24 week prediction horizon. To prevent the model from overfitting to full observation windows, we employ a random endpoint strategy, where for each patient up to five additional endpoints are randomly selected from the whole patient time series. The time series is then truncated at this randomly chosen endpoint and left-padded with zeros to obtain a fixed-length input window. Survival time is recalculated relative to the new endpoint, and a discrete survival target is constructed over the prediction horizon. This strategy reduces bias toward specific temporal positions and encourages the model to learn time-robust representations. After augmentation, the dataset distribution shifted, with the proportion of non-survivors decreasing from 55.67% to 31.03%.

4 Experiments

We evaluated if multi-modal time series models can be used for future survival forecasting with irregularly sampled and missing values input time series data. We reported standard classification metrices, precision-recall (PR) curves, as well as receiver operating characteristic (ROC) analysis for the 24 weeks forecasting. We assessed prediction of survival after 6 and 24 weeks.

5 Results

All models were trained on the same combined dataset comprising both original and randomly augmented samples, while evaluation is reported separately for full windows (FW) and random windows (RW). Quantitative results for prediction week 6 and 24 are shown in Table 1.

Table 1: Per-week performance metrics for full lookback windows (FW) and random created windows (RW) for weeks 6 and 24.

Week	Metric	PatchTST		GRU-D		DeepSurv	
		FW	RW	FW	RW	FW	RW
6	Sensitivity	0.714	0.519	0.862	0.726	0.000	0.000
	Specificity	0.789	0.842	0.765	0.833	0.998	1.000
	PPV	0.689	0.272	0.706	0.331	0.000	0.000
	NPV	0.808	0.939	0.894	0.964	0.603	0.898
	MAE	0.800	0.842	0.319	0.294	0.425	0.230
	AP (PR-AUC)	0.862	0.966	0.926	0.980	0.763	0.927
	ROC-AUC	0.810	0.778	0.888	0.861	0.692	0.593
24	Sensitivity	0.963	0.891	0.990	0.955	0.016	0.004
	Specificity	0.300	0.390	0.174	0.278	0.994	0.996
	PPV	0.622	0.328	0.589	0.307	0.750	0.214
	NPV	0.872	0.914	0.932	0.948	0.457	0.749
	MAE	1.982	1.575	0.357	0.515	0.508	0.407
	AP (PR-AUC)	0.778	0.883	0.846	0.918	0.612	0.803
	ROC-AUC	0.805	0.744	0.869	0.807	0.672	0.587

5.1 Full lookback windows

This result represents the prediction at the last time point available for each patient in the data set. PatchTST and GRU-D showed highest sensitivity and NPV across week 6 and week 24, indicating strong performance in identifying patients who experience non survival within the prediction horizon. GRU-D achieved the highest sensitivity (0.862 at week 6, 0.990 at week 24), while PatchTST had slightly higher specificity at week 6 (0.789) and week 24 (0.300). DeepSurv generally failed to correctly predicted non survival patients (sensitivity of < 0.02) and therefore achieved only high

specificity in predicting survival patients. MAE values indicated that GRU-D consistently produced more accurate survival predictions than PatchTST or DeepSurv. For the PR curves similar results are visible. At week 6, all models demonstrated strong precision–recall performance when evaluated separately on the original subsets, despite being trained on the combined dataset. GRU-D achieved the highest average precision on the original data (AP = 0.926), followed by PatchTST (AP = 0.862) and DeepSurv (AP = 0.763). ROC analysis showed that GRU-D consistently achieved the highest discriminative performance across both prediction horizons (AUC of 0.888 at week 6), outperforming PatchTST (AUC = 0.810) and DeepSurv (AUC = 0.692).

5.2 Random generated windows

This result represents the prediction at randomly chosen time points in the patient trajectory data. For the random generated windows at week 6 GRU-D achieved the highest NPV (0.964), sensitivity (0.726) and specificity slightly lower (0.833) than PatchTST. At week 24 GRU-D reached a higher sensitivity (0.990), but a lower specificity (0.174) compared to week 6. Overall, FW results showed higher sensitivity and lower specificity compared to RW results. For the PR curves with GRU-D reaching an AP of 0.980, PatchTST 0.966, and DeepSurv 0.927. A similar pattern was observed at week 24, although overall performance decreased compared to week 6, reflecting the increased difficulty of longer-term prediction. For ROC analysis, GRU-D reached an AUC of 0.861, PatchTST 0.778, and DeepSurv 0.593 at week 6. A similar trend was evident at week 24, with GRU-D achieving 0.807, PatchTST 0.744, and DeepSurv 0.587.

6 Discussion

We evaluated three different deep learning models for predicting individual six month survival from irregularly sampled lung patient trajectories. We propose a balanced loss function not only for multiple forecasting time points. Using more than one forecasting time point enhances model training, since more information can be learned from each prediction sample. We enhanced our training with random generated time windows and for training we used 10 different input channels. Results showed that the recurrent neural network yields best accuracy in predicting survival at 6 and 24 weeks, compared to the transformer model. State-of-the-art deep learning models offer a viable route towards handling highly irregularly sampled time series data, and integrate multiple modalities.

Acknowledgement

This work has been partially funded by the Vienna Science and Technology Fund (WWTF, PREDICTOME) [10.47379/LS20065], European Union's Horizon Europe research and innovation programme under grant agreement No.101100633—EUCAIM and No.101080302 AI-POD, and the Austrian Science Fund (FWF, P35189 ONSET). It has been carried out within an Inter-University Cluster Project jointly funded by the University of Vienna and the Medical University of Vienna (AICARD). The financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

References

[1] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent neural networks for multivariate time series with missing values. *Scientific reports*, 8(1):6085, 2018.

[2] Edward Choi, Andy Schuetz, Walter F Stewart, and Jimeng Sun. Using recurrent neural network models for early detection of heart failure onset. *Journal of the American Medical Informatics Association*, 24(2):361–370, 2017.

[3] Jared L Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC medical research methodology*, 18(1):24, 2018.

[4] Xiangjie Kong, Zhenghao Chen, Weiyao Liu, Kaili Ning, Lechao Zhang, Syauqie Muhammad Marier, Yichen Liu, Yuhao Chen, and Feng Xia. Deep learning for time series forecasting: a survey. *International Journal of Machine Learning and Cybernetics*, 16(7):5079–5112, 2025.

[5] Wenxiang Li and KL Eddie Law. Deep learning models for time series forecasting: A review. *IEEE Access*, 12:92306–92327, 2024.

[6] Manfred Mudelsee. Trend analysis of climate time series: A review of methods. *Earth-science reviews*, 190:310–322, 2019.

[7] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*, 2022.

[8] David S Stoffer and Hernando Ombao. Special issue on time series analysis in the biological sciences, 2012.

[9] Liyilei Su, Xumin Zuo, Rui Li, Xin Wang, Heng Zhao, and Bingding Huang. A systematic review for transformer-based long-term series forecasting. *Artificial Intelligence Review*, 58(3):80, 2025.

[10] Ruoxi Yu, Yali Zheng, Ruikai Zhang, Yuqi Jiang, and Carmen CY Poon. Using a multi-task recurrent neural network with attention mechanisms to predict hospital mortality of patients. *IEEE journal of biomedical and health informatics*, 24(2):486–492, 2019.

[11] George Udny Yule. Vii. on a method of investigating periodicities disturbed series, with special reference to wolfer's sunspot numbers. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 226(636-646):267–298, 1927.

Multimodal Contrastive Learning for Alzheimer's Disease Prediction in Imaging Genetics

Jonas Fallmann
Institute for Machine Learning
Johannes Kepler University
Linz
jonas.fallmann@jku.at

Erich Kobler
Institute for Machine Learning,
LIT AI Lab,
Department of Virutal Morphology,
Clinical Research Institute Medical AI
Johannes Kepler University
Linz
erich.kobler@jku.at

Abstract

Alzheimer's disease (AD) is an inherently multimodal pathology driven by complex genetic and phenotypic interactions, making reliable early detection a critical challenge. Existing multimodal approaches often struggle to effectively align static baseline genetic risk with longitudinal physical changes. In this work, we introduce a novel two-stage contrastive learning framework integrating Single Nucleotide Polymorphisms (SNPs) and structural MRI volumes. To overcome the bottleneck of single-timepoint genetic measurements, we propose an age-conditioned augmentation strategy that generates time-aware genetic embeddings for longitudinal contrastive pairing. Utilizing a dynamic Gated Fusion mechanism for downstream classification, our approach effectively weights modality contributions. Evaluated on the ADNI database, our framework consistently outperforms strong classical baselines and state-of-the-art generative models, demonstrating particularly significant improvements in early-stage cognitive decline detection.

1 Introduction

Alzheimer's Disease (AD) is a progressive neurodegenerative disorder and the leading global cause of dementia [11]. Because current treatments primarily slow progression rather than cure it [16], early detection is critical. Capturing the complete disease trajectory requires integrating upstream genetic susceptibility with downstream phenotypic consequences.

Existing multimodal AD research predominantly fuses imaging modalities such as MRI and PET using standard supervised architectures [1, 13, 14, 10]. However, these methods often struggle to learn robust representations when labeled data is scarce. While contrastive learning has emerged as a powerful alternative for extracting features from medical imaging [8, 9, 5] and imaging-genetics cohorts [15, 12, 17], its potential to align high-dimensional SNPs with MRI brain volumes for AD remains underexplored. A primary hurdle is data asymmetry: effectively combining a single static genetic measurement with longitudinal imaging data requires complex setups that standard fusion approaches fail to address [6].

In this work, we introduce a multimodal contrastive learning architecture to extract meaningful, modality-invariant latent representations from MRI brain volumes and SNPs. To overcome the bottleneck of relying on a single genetic measurement per subject, we introduce time-aware genetic embeddings. By augmenting static genetic profiles with the patient's age at the time of the scan, we enable the generation of rich, multi-view positive pairs. We then evaluate the expressivity of

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

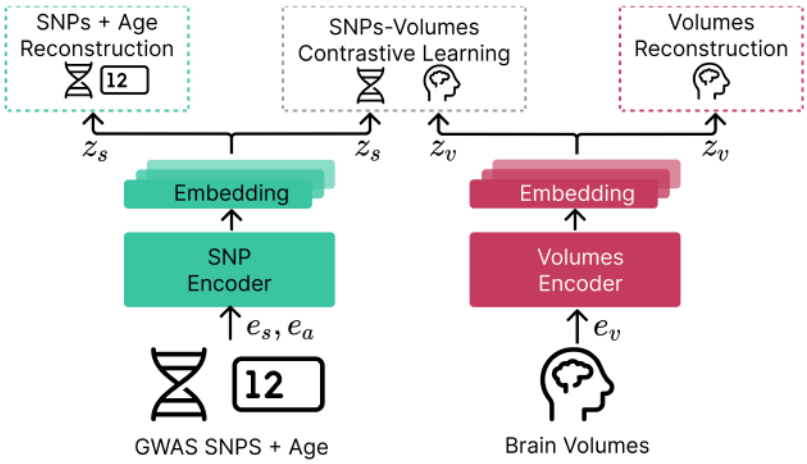


Figure 1: Contrastive learning setup featuring dual-branch encoders for input modalities and distinct decoders for feature-preserving reconstruction. Inter-modal contrastive loss aligns patient-specific embeddings across modalities.

these learned embeddings by training a downstream classifier on the frozen encoders to predict AD diagnosis and cognitive decline.

2 Methodology

To address the problem of Alzheimer's disease classification, we adopt a two-stage pipeline that decouples representation learning from downstream classification. In the first stage, a contrastive encoder module is trained to produce dense latent representations that capture the most salient and modality-invariant features of the input data. In the second stage, a fusion model combines the latent representation produced by the frozen contrastive encoding module. Classification of the disease class and regression of the Mini-Mental State Examination (MMSE) [4] score are performed.

2.1 Contrastive encoding

At a conceptual level, the contrastive encoding module maps different data modalities into a shared latent space (Figure 2). Using contrastive learning, representations of the same subject at a specific time point are pulled together, while representations of different entities are pushed apart, promoting modality-invariant embeddings.

Following the SimCLR framework [2], we define the contrastive objective using the Normalized Temperature-scaled Cross Entropy (NT-Xent) loss. Let e_s denote the SNP features, e_a the patient age, and e_v the standardized MRI volumes. We define the modality-specific latent representations as $z_s = f_\theta(e_s, e_a)$ and $z_v = g_\phi(e_v)$, where f_θ and g_ϕ are the genetic and volume encoders, respectively. For a batch of size N, the loss for a genetic anchor $z_s^{(i)}$ against the volume representations $\mathcal{Z}_v$ is formulated as:

$$\ell(z_s^{(i)}, \mathcal{Z}_v) = -\log \frac{\exp(\text{sim}(z_s^{(i)}, z_v^{(i)})/\tau)}{\sum_{k\neq i}^{N} \exp(\text{sim}(z_s^{(i)}, z_v^{(k)})/\tau)}$$

where $\text{sim}(\cdot,\cdot)$ denotes the cosine similarity and τ represents the temperature. To ensure bidirectional alignment, we symmetrically compute the loss for the volume anchor against the genetic representations, yielding the total contrastive objective $\mathcal{L}_{\text{cont}} = \frac{1}{2N}\sum_{i=1}^{N}[\ell(z_s^{(i)}, \mathcal{Z}_v) + \ell(z_v^{(i)}, \mathcal{Z}_s)]$. In addition, we concurrently apply a mean squared error (MSE) reconstruction objective to preserve features critical for downstream classification thereby enhancing the overall representational quality. The total loss is a weighted sum of the individual loss terms.

To form contrastive pairs, we match a subject's static genetic profile with brain volumes from their longitudinal MRI scans. Because genetic profiles are static and measured only once, we face a critical bottleneck in generating the abundant, distinct multi-view positive pairs required for

effective contrastive learning. To overcome this, we augment the static genetic profile by directly concatenating the subject's scalar age at the time of MRI acquisition. Empirical evaluations showed no significant performance difference between simple concatenation and mapping age through a learned embedding, so we opted for the computationally simpler strategy. This simple augmentation extracts age-conditioned representations, transforming a single genetic profile into multiple distinct views corresponding to each longitudinal scan. Biologically, this strategy is highly plausible, as the phenotypic effects of genetic variants on brain morphology are strongly age-dependent.

2.2 Gated fusion classification

To fuse the projected genetic (z_s) and structural (z_v) latent vectors, we employ a learned modality-level gating mechanism. Scalar modality-specific attention weights are computed via a linear projection and softmax activation:

$$[\alpha_s, \alpha_v] = \text{Softmax}(W_g[z_s, z_v] + b_g)$$

where $[\cdot, \cdot]$ denotes concatenation and $\alpha_s, \alpha_v \in \mathbb{R}$. The final representation is the weighted sum $z_{\text{fused}} = \alpha_s z_s + \alpha_v z_v$, enabling adaptive downweighting of noisy or missing inputs. Let e_d represent the demographic embedding encompassing sex and education years. The fused vector is concatenated to form the complete patient representation $h = [z_{\text{fused}}, e_d]$. This representation is then routed to parallel MLP heads to yield the final predictions $\hat{y} = \text{MLP}(h)$ for diagnostic classification and continuous MMSE score regression, optimized simultaneously via a weighted compound loss.

3 Experimental setup

Our dataset comprised 844 subjects from the ADNI database, yielding 3,784 multimodal pairs. Data curation involved dropping missing records and applying a 90-day soft-alignment threshold to match imaging assessments with the closest temporal diagnostic labels. Across all longitudinal visits, the distribution was 1,354 Cognitively Normal (CN), 1,921 Mild Cognitive Impairment (MCI), and 509 Alzheimer's Disease (AD) assessments. The final-visit classification subset comprised 299 CN, 349 MCI, and 196 AD subjects. We evaluated the framework across three binary classification tasks, a multiclass task (CN vs. MCI vs. AD), and continuous regression of the Mini-Mental State Examination (MMSE) score.The MMSE is a 30-point scale where a score of 30 indicates no cognitive impairment, which we min-max normalized to a [0, 1] range for our prediction task. For the imaging modality, structural MRI volumes were extracted via Freesurfer [3] and directly z-score standardized to preserve absolute volume information, rather than normalizing by intracranial volume. For the genetic modality, SNPs derived from microarray data were ranked by p-value using the Alzheimer's Disease Variant Portal (ADVP) [7]—an aggregated database of multiple studies. These selected SNPs were mapped from discrete categorical values into continuous representation embeddings using the population-based probability transformation matrix introduced by Ko et al. [6]. Both continuous modality-specific embeddings were subsequently fed into our contrastive encoding architecture.

To prevent longitudinal data leakage, we enforced a strict subject-wise split that remained identical across both training stages. Within each cross-validation fold, the contrastive encoder was pretrained using all available timepoints of the training subjects only. Subsequently, the downstream Gated Fusion classifier was trained and evaluated exclusively on the final timepoint of the respective train and test subjects. Performance was assessed using five-fold nested stratified cross-validation. We benchmarked our framework against classical machine learning algorithms trained on concatenated features—among which Random Forest proved to be the strongest performing—as well as a state-of-the-art multimodal generative-adversarial attention framework [6]. Hyperparameters were rigorously optimized within the inner cross-validation loop using exhaustive grid search for classical models and Optuna Bayesian optimization for the neural architectures.

4 Results

Our complete framework, integrating contrastive pretraining with downstream Gated Fusion, demonstrated strong performance across diagnostic classification tasks, particularly when compared to robust classical baselines (Table 1). While we included a recent generative VAE+GAN model as a

baseline, its adversarial training proved highly unstable in practice, frequently underperforming a standard Random Forest Classifier (RFC). Consequently, we treat the classical RFC as our primary competitive benchmark.

Our method achieved the most significant gains in the challenging task of early cognitive decline recognition (CN vs. MCI). The framework reached an Area Under the Curve (AUC) of 0.707, providing a definitive improvement over both the classical RFC (AUC 0.665) and the VAE+GAN (AUC 0.624). However, this classification advantage comes with a clear trade-off in continuous Mini-Mental State Examination (MMSE) regression. The Random Forest Regressor (RFR) achieved lower Root Mean Square Errors (RMSE) than our model. This highlights a limitation of our multi-task compound loss: while joint optimization benefits categorical disease staging, it slightly compromises the precision of continuous cognitive score prediction compared to a dedicated, single-objective regressor.

Ablation experiments confirmed that structural MRI volumes primarily drive predictions in advanced disease stages (AUC of 0.928 for CN vs. AD using MRIa alone), whereas genetic features (SNPs) supply complementary variance for early-stage detection. Crucially, relying solely on either the reconstruction or contrastive objectives yielded suboptimal performance (AUCs of 0.916 and 0.882 for CN vs. AD, respectively). This demonstrates that neither objective alone fully captures multi-modal complexity; rather, their synergy is essential for aligning distinct views in the latent space, noticeably boosting classification across all categories.

Table 1: Comparison of model performance across diagnostic groups. Best results are in bold (highest ROC-AUC, lowest RMSE) and second-best results are underlined.

Model	Metric	CN/AD	CN/MCI	MCI/AD	CN/MCI/AD
Only Reconstruction	AUC	0.916 ± 0.024	0.668 ± 0.032	0.817 ± 0.035	0.760 ± 0.019
	RMSE	0.133 ± 0.015	$\underline{0.073} \pm 0.006$	0.142 ± 0.010	0.134 ± 0.016
Only Contrastive	AUC	0.882 ± 0.027	0.659 ± 0.046	0.805 ± 0.025	0.739 ± 0.024
	RMSE	0.132 ± 0.009	0.074 ± 0.011	0.140 ± 0.013	0.125 ± 0.009
Single modal SNPs	AUC	0.638 ± 0.069	0.626 ± 0.038	0.534 ± 0.045	0.590 ± 0.028
	RMSE	0.174 ± 0.018	0.103 ± 0.039	0.167 ± 0.024	0.148 ± 0.018
Single modal MRI	AUC	$\underline{0.928} \pm 0.028$	$\underline{0.699} \pm 0.024$	$\underline{0.819} \pm 0.029$	$\underline{0.774} \pm 0.013$
	RMSE	0.128 ± 0.007	0.156 ± 0.011	$\underline{0.136} \pm 0.006$	0.132 ± 0.011
RFC / RFR	AUC	0.893 ± 0.030	0.665 ± 0.037	0.812 ± 0.041	0.747 ± 0.028
	RMSE	$\mathbf{0.125} \pm 0.010$	$\mathbf{0.062} \pm 0.006$	0.137 ± 0.006	$\mathbf{0.114} \pm 0.006$
VAE + GAN Fusion [6]	AUC	0.861 ± 0.023	0.624 ± 0.042	0.768 ± 0.027	0.691 ± 0.022
	RMSE	0.145 ± 0.013	0.077 ± 0.005	$\mathbf{0.133} \pm 0.011$	$\underline{0.123} \pm 0.010$
Contrastive + Fusion (ours)	AUC	$\mathbf{0.936} \pm 0.017$	$\mathbf{0.707} \pm 0.035$	$\mathbf{0.831} \pm 0.019$	$\mathbf{0.778} \pm 0.014$
	RMSE	$\underline{0.127} \pm 0.015$	0.080 ± 0.014	0.144 ± 0.012	0.127 ± 0.006

5 Conclusion

We presented a two-stage multimodal contrastive learning framework for AD detection. By integrating SNPs with longitudinal MRI via an age-conditioned augmentation strategy, our approach mitigates data asymmetry and captures complex disease trajectories. Dynamic Gated Fusion adaptively weights these modalities, outperforming classical baselines particularly in early-stage detection (CN vs. MCI). Ablations confirm that combining contrastive and reconstruction objectives preserves critical pathological details. Future work will integrate PET imaging and fluid biomarkers, validating the framework on independent cohorts.

Acknowledgments and Disclosure of Funding

The authors acknowledge support from the DFG within the SPP2298 under project number 543939932 and from the Austrian Science Fund (FWF) project number 10.55776/COE12.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

[1] Giovanna Castellano, Andrea Esposito, Eufemia Lella, Graziano Montanaro, and Gennaro Vessio. Automated detection of Alzheimer's disease: A multi-modal approach with 3D MRI and amyloid PET. *Scientific Reports*, 14(1):5210, 2024. doi: 10.1038/s41598-024-56001-9.

[2] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference on Machine Learning*, volume 119, pages 1597–1607. PMLR, 2020.

[3] Bruce Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012. doi: 10.1016/j.neuroimage.2012.01.021.

[4] Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. "mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, 1975. doi: 10.1016/0022-3956(75)90026-6.

[5] Paul Hager, Martin J. Menten, and Daniel Rueckert. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23924–23935, 2023.

[6] Wonjun Ko, Wonsik Jung, Eunjin Jeon, and Heung-Il Suk. A deep generative-discriminative learning for multimodal representation in imaging genetics. *IEEE Transactions on Medical Imaging*, 41(9):2348–2359, 2022. doi: 10.1109/TMI.2022.3162870.

[7] Pavel P Kuksa, Chia-Lun Liu, Wei Fu, Liming Qu, Yi Zhao, Zivadin Katanic, Kaylyn Clark, Amanda B Kuzma, Pei-Chuan Ho, Kai-Teh Tzeng, et al. Alzheimer's disease variant portal: A catalog of genetic findings for Alzheimer's disease. *Journal of Alzheimer's Disease*, 86(1):461–477, 2022. doi: 10.3233/JAD-215055.

[8] Min Gu Kwak, Yi Su, Kewei Chen, David Weidman, Teresa Wu, Fleming Lure, Jing Li, and Alzheimer's Disease Neuroimaging Initiative. Self-supervised contrastive learning to predict the progression of Alzheimer's disease with 3D amyloid-PET. *Bioengineering*, 10(10):1141, 2023. doi: 10.3390/bioengineering10101141.

[9] Jiaguang Li, Ying Wei, Chuyuan Wang, Qian Hu, Yue Liu, and Long Xu. 3D CNN-based multi-channel contrastive learning for Alzheimer's disease automatic diagnosis. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022. doi: 10.1109/TIM.2022.3162265.

[10] Modupe Odusami, Rytis Maskeliūnas, Robertas Damaševičius, and Sanjay Misra. Explainable deep-learning-based diagnosis of Alzheimer's disease using multimodal input fusion of PET and MRI images. *Journal of Medical and Biological Engineering*, 43(3):291–302, 2023. doi: 10.1007/s40846-023-00801-3.

[11] Saeid Safiri, Amir Ghaffari Jolfayi, Asra Fazlollahi, Soroush Morsali, Aila Sarkesh, Amin Daei Sorkhabi, Behnam Golabi, Reza Aletaha, Kimia Motlagh Asghari, Sana Hamidi, et al. Alzheimer's disease: A comprehensive review of epidemiology, risk factors, symptoms diagnosis, management, caregiving, advanced treatments and associated challenges. *Frontiers in Medicine*, 11:1474043, 2024. doi: 10.3389/fmed.2024.1474043.

[12] Daniel Sens, Liubov Shilova, Adrian V Dalca, Julia Schnabel, and Francesco Paolo Casale. GEMCONT: Genetics-based multimodal contrastive learning for disease-focused imaging genetics. In *Medical Imaging with Deep Learning*, 2026. URL https://openreview.net/forum?id=Y8gkT7s44N.

[13] Juan Song, Jian Zheng, Ping Li, Xiaoyuan Lu, Guangming Zhu, and Peiyi Shen. An effective multimodal image fusion method using MRI and PET for Alzheimer's disease diagnosis. *Frontiers in Digital Health*, 3, 2021. doi: 10.3389/fdgth.2021.637386.

[14] Jinju Sun, Chao Cong, Xinpeng Li, Weicheng Zhou, Renxiang Xia, Huan Liu, Yi Wang, Zhiqiang Xu, and Xiao Chen. Identification of Parkinson's disease and multiple system atrophy using multimodal PET/MRI radiomics. *European Radiology*, 34(1):662–672, 2024. doi: 10.1007/s00330-023-10003-9.

[15] Aiham Taleb, Matthias Kirchler, Remo Monti, and Christoph Lippert. ContIG: Self-supervised multimodal contrastive learning for medical imaging with genetics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20921, 2022.

[16] Sebastian Walsh, Richard Merrick, Edo Richard, Shirley Nurock, and Carol Brayne. Lecanemab for Alzheimer's disease. *BMJ*, 379:o3010, 2022. doi: 10.1136/bmj.o3010.

[17] Rong Zhou, Houliang Zhou, Li Shen, Brian Y. Chen, Yu Zhang, and Lifang He. Integrating multimodal contrastive learning and cross-modal attention for Alzheimer's disease prediction in brain imaging genetics. In *Proceedings of the IEEE International Conference on Bioinformatics and Biomedicine*, pages 1806–1811, 2023. doi: 10.1109/BIBM58861.2023.10385864.

xLSTM for Irregular Multivariate Clinical Time-Series Forecasting

Laura Legat
Institute for Machine Learning
Johannes Kepler University Linz
laura.legat@jku.at

Erich Kobler
Institute for Machine Learning
LIT AI Lab
Department of Virtual Morphology
Clinical Research Institute Medical AI
Johannes Kepler University Linz
erich.kobler@jku.at

Abstract

Intensive care units (ICUs) provide lifesaving treatments to patients with severe medical conditions, producing large amounts of clinical time-series data that reflect patient health trajectories. Forecasting future trajectory changes helps clinicians anticipate adverse events. While prior work addresses the challenges of missing values and irregularities in clinical time-series, designing effective forecasting architectures for such data remains an open research area. At the same time, limitations of Transformer-based models are prompting a renewed interest in recurrent architectures for processing time-series. Among them, the recently proposed xLSTM demonstrates strong forecasting capabilities across several domains, yet its potential for clinical use-cases remains largely unexplored. In this work, we address this gap by extending xLSTM to forecast irregular multivariate clinical time-series with missing values. To this end, we replace the temporal and cross-channel modeling components of an established forecasting architecture with xLSTM blocks. Our models achieve competitive predictive performance compared to several baselines on a subset of MIMIC-III, highlighting xLSTM's potential as a powerful backbone for clinical time-series forecasting.

1 Introduction

Continuous monitoring in the ICU is employed to improve clinical outcomes and provide insight into individual health journeys [20], generating large volumes of patient data organized as electronic health records (EHR) [6]. A substantial portion of these records constitute clinical time-series in the form of sequential, time-indexed, multivariate observations [13]. We refer to them as multivariate clinical time-series (MCTS), which capture a patient's physiological trajectory over time. Anticipating future trajectory changes early-on is crucial, as it helps to avoid adverse events and prolonged hospital stays [19]. This can be framed as a time-series forecasting task [14] and remains a challenging endeavor, since MCTS often contain a multitude of channels [22], all sampled at different frequencies, with different levels of missingness, outliers, and artifacts [22, 29].

Transformer-based architectures [21] remain a popular choice for this task [29, 17], yet they face limitations, such as large training data requirements as well as the quadratic complexity of the attention mechanism with respect to sequence length and number of channels [12, 1]. This results in a renewed interest in linear-complexity recurrent models for time-series forecasting [12]. Beck et al. [2] introduce such an architecture with the Extended Long Short-Term Memory (xLSTM), a successor to the LSTM, which shows strong forecasting results across several non-medical domains, such as weather, solar, or electricity [12, 1]. However, the potential of xLSTM for MCTS forecasting remains unexplored.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

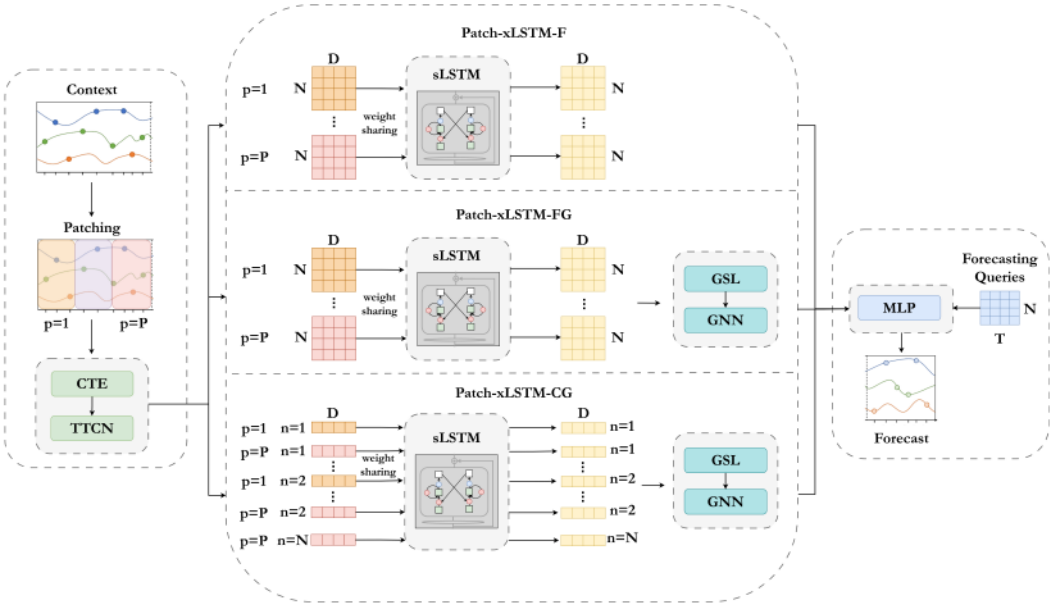


Figure 1: Overview of our proposed approaches. On the left, each patient's time-series are first split into fixed-length patches and encoded into patch embeddings. These are processed either jointly (full-patch), or for each channel separately (per-channel). Given the final representations and future forecasting queries, a MLP produces the irregular MCTS forecast.

Motivated by this limitation, we propose three xLSTM-based model variants for forecasting irregular MCTS with missing values. Our approach integrates xLSTM into the established forecasting architecture T-PatchGNN [29] by replacing components responsible for temporal and cross-channel modeling with xLSTM blocks. We assess whether xLSTM can function as an effective backbone for irregular MCTS forecasting by evaluating our models on a subset of the MIMIC-III dataset, and comparing their performance with various baselines. The findings show that our xLSTM-based variants achieve competitive performance, indicating their suitability for forecasting irregular MCTS.

2 Method

In our work, we investigate xLSTM-based architectures for forecasting irregularly-sampled MCTS with missing values. For this, we integrate xLSTM blocks in the established forecasting model T-PatchGNN [29], which we choose due to its open-source implementation and strong forecasting performance.

Problem Let $\mathbf{X} = \{\{(t_i^n, x_i^n, m_i^n)\}_{i=1}^{T_n}\}_{n=1}^{N}$ denote irregular MCTS data with missing values, where N is the number of channels and the n-th channel contains T_n observations. An observation at time t for channel n consists of a chronological time delta $t_i^n \in \mathbb{R}^+$ as the minutes elapsed since the measurement start, a value $x_i^n \in \mathbb{R}$, and a binary missingness mask $m_i^n \in \{0, 1\}$. Specifically, $m_i^n = 1$ if the measurement exists at t, otherwise $m_i^n = 0$. Irregular forecasting then aims to predict future values at specific requested timestamps along a continuous time axis, which are given through forecasting queries [29]. We denote such queries by $\mathbf{Q} = \{\{q_j^n\}_{j=1}^{Q_n}\}_{n=1}^{N}$, where q_j^n is the j-th forecasting request for channel n. Then, given historical MCTS data $\mathbf{X}$ and queries $\mathbf{Q}$, we want to forecast $\hat{\mathbf{X}} = \{\{\hat{x}_j^n\}_{j=1}^{Q_n}\}_{n=1}^{N}$ according to $\mathcal{F}_\theta(\mathbf{X}, \mathbf{Q}) \to \hat{\mathbf{X}}$, where $\mathcal{F}(\cdot)_\theta$ denotes a forecasting function parameterized by θ whose parameters are learned from data.

Models For forecasting models to be applicable in clinical settings, they should naturally handle irregularly-sampled data and missing values. To achieve this, we propose three forecaster variants $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3$ parameterized by $\theta_1, \theta_2, \theta_3$, and illustrated in Figure 1. The raw MCTS are first partitioned into P non-overlapping consecutive patches, each covering fixed time intervals, resulting in a unified

temporal resolution despite varying observation counts in the patches. Following prior work [29], the continuous time-embedding (CTE) and the transformable time-aware convolution network (TTCN) encode the patches into a sequence of patch embeddings $\mathbf{Z} = \{\mathbf{z}_p\}_{p=1}^{P}$, $\mathbf{z}_p \in \mathbb{R}^{N \times D}$ that are concatenated with the missingness masks, where N is the number of channels and D the embedding dimension. We distinguish two input processing settings. In the full-patch setting, each sLSTM time-step comprises a multivariate patch embedding, allowing the capture of cross-channel dependencies. In the per-channel setting, the time-step is a univariate patch embedding, meaning all channels are processed independently in parallel. In both cases, weights are shared across time-steps and patients. Originally, T-PatchGNN [29] employs a Transformer for temporal processing, followed by graph structure learning (GSL) and a graph neural network (GNN) for cross-channel modeling. Our first model $\mathcal{F}_1$, abbreviated as Patch-xLSTM-F, replaces both with a 2-layer sLSTM, a variant of xLSTM, for jointly modeling channel and temporal dynamics. It processes the full-patch input representation, producing patch representations of dimensions $N \times D$. These are then passed to a multilayer perceptron (MLP) that generates forecasts at given query timestamps $\mathbf{Q}$. The second full-patch variant is $\mathcal{F}_2$, or Patch-xLSTM-FG, which, in contrast to $\mathcal{F}_1$, separates temporal and cross-channel modeling between a 2-layer sLSTM and the GNN before producing forecasts. Finally, $\mathcal{F}_3$ is Patch-xLSTM-CG, where the C denotes the per-channel setting. The temporal dependencies between the patches are processed for each channel independently by a 2-layer sLSTM, before employing GSL and GNN for cross-channel processing.

Data For reproducibility, we evaluate our models on a publicly-available subset of the MIMIC-III dataset, which we further refer to as MIMIC-III-TP.[1] MIMIC-III is a large single-center ICU dataset containing EHR data of patients admitted to Beth Israel Deaconness Medical Center in Boston [9]. After pre-processing, we obtain irregularly-sampled MCTS from 23.457 unique patients, each covering the first 48 hours after ICU admission across 96 different channels. The channels exhibit varying but strong levels of missingness, as well as irregularity, and include laboratory measurements, medication administrations, fluid inputs like insulin, and fluid outputs, such as urine. Each observation is concatenated to a corresponding binary missingness mask indicating whether or not a value is observed at some timestamp.

3 Experiments and results

We evaluate all models on the task of forecasting irregular MCTS over a 24-hour prediction window given the previous 24-hour context, together with a set of forecasting queries. The data is split admission-wise into training, validation and test sets according to a 60-20-20% ratio. Channel values and timestamps are minimum-maximum normalized based on training set statistics, before each admission is partitioned into non-overlapping context and prediction windows. Finally, the sequences are zero-padded to equal length, constituting the input to the model. For this, we follow the protocol established by Zhang et al. [29].

Baselines We compare our proposed variants to 18 baselines covering both regular and irregular forecasting architectures: DLinear [27], TimesNet [23], PatchTST [15], Crossformer [31], CrossGNN [8], Graph WaveNet [24], MTGNN [25], FourierGNN [26], StemGNN [4], SeFT [7], Latent-ODE [17], Neural Flows [3], mTAN [17], CRU [18], GRU-D [5], RAINDROP [30], Warpformer [28], and the original T-PatchGNN [29].

Training and evaluation All models use the Adam optimizer [10] with a learning rate of 0.001, no weight decay, and betas (0.9, 0.999). We set the batch size to 32 and apply early stopping based on validation performance with a patience of 10 epochs. The hidden dimension is fixed to 64 for all models, and hyperparameters unrelated to sLSTM are set as in the original work [29]. The models are optimized using the masked mean squared error (MSE) loss between the predicted and ground truth values at the query timestamps. Model performance is evaluated using the masked MSE and masked mean absolute error (MAE), following prior protocols [29, 11]. We repeat all experiments five times with different random seeds and report the mean and standard deviations of the metrics in Table 1. Separate 100-iteration random hyperparameter searches are conducted over the sLSTM-related parameters, selecting the final parameters based on the lowest validation masked MSE.

[1] https://physionet.org/content/mimic-iii-ext-tpatchgnn/1.0.0/

Table 1: Performance comparison on MIMIC-III-TP for forecasting 24 hours from a 24-hour context window. Lower is better. **Bold** shows the best performance, underlined shows the second-best. † indicates that this result is reported by Zhang et al. [29].

Algorithm	MSE $\times 10^{-2}$	MAE $\times 10^{-2}$
TimesNet† [23]	5.88 ± 0.08	13.62 ± 0.07
DLinear† [27]	4.90 ± 0.00	16.29 ± 0.05
PatchTST† [15]	3.78 ± 0.03	12.43 ± 0.10
CrossGNN† [8]	2.95 ± 0.16	10.82 ± 0.21
Graph WaveNet† [24]	2.93 ± 0.09	10.50 ± 0.15
MTGNN† [25]	2.71 ± 0.23	9.55 ± 0.65
Crossformer† [31]	2.65 ± 0.19	9.56 ± 0.29
FourierGNN† [26]	2.55 ± 0.03	10.22 ± 0.08
RAINDROP† [30]	1.99 ± 0.03	8.27 ± 0.07
Latent-ODE† [16]	1.89 ± 0.11	8.11 ± 0.52
Neural Flows† [3]	1.87 ± 0.05	8.03 ± 0.06
SeFT† [7]	1.87 ± 0.01	7.84 ± 0.08
mTAN† [17]	1.85 ± 0.06	7.73 ± 0.13
CRU† [18]	1.81 ± 0.05	8.06 ± 0.07
GRU-D† [5]	1.76 ± 0.03	7.53 ± 0.09
Warpformer† [28]	1.73 ± 0.04	7.58 ± 0.13
StemGNN† [4]	1.73 ± 0.02	7.71 ± 0.11
T-PatchGNN [29]	$\underline{1.71 \pm 0.03}$	$\underline{7.33 \pm 0.10}$
Patch-xLSTM-FG (ours)	1.76 ± 0.05	7.63 ± 0.2
Patch-xLSTM-F (ours)	1.74 ± 0.06	7.51 ± 0.03
Patch-xLSTM-CG (ours)	$\mathbf{1.68 \pm 0.02}$	$\mathbf{7.29 \pm 0.09}$

Results The results of our experiments are summarized in Table 1. Patch-xLSTM-CG achieves the lowest MSE and MAE among the evaluated models, performing on par with other strong baselines, while the original T-PatchGNN [29] ranks second. These findings suggest that sLSTM's powerful state tracking capabilities, as well as capturing temporal dynamics separately for each channel before cross-channel aggregation can be beneficial for learning MCTS. Accordingly, joint modeling temporal and cross-channel dependencies with a recurrent component instead of separating these responsibilities can limit the model's ability to learn channel-specific dynamics, as indicated by Patch-xLSTM-F's results. Patch-xLSTM-FG, performs the worst out of our proposed models, suggesting that mixing channel and temporal information early on leads to the loss of channel-specific information which cannot be recovered by subsequent components. In summary, our results indicate that both per-channel temporal modeling, as well as separate modules for temporal and cross-channel modeling, are beneficial for effective MCTS learning.

4 Discussion and Conclusion

To investigate the potential of xLSTM for forecasting irregular MCTS with missing values, we integrate it as a temporal modeling backbone into an established forecasting architecture [29]. We compare the three resulting model variants to a range of regular and irregular forecasting approaches on MIMIC-III-TP. Our results indicate that employing separate architectural components for temporal and cross-channel modeling, as well as capturing temporal dependencies per-channel, are beneficial for MCTS learning. Looking the beyond predictive performance, sLSTM offers linear complexity and a sequential inductive bias [2], both favorable attributes for clinical time-series forecasting. Overall, our proposed models achieve competitive forecasting performance to strong baselines, positioning xLSTM as a promising temporal backbone for irregular MCTS forecasting.

Acknowledgments and Disclosure of Funding

The authors acknowledge support from the DFG within the SPP2298 under project number 543939932 and from the Austrian Science Fund (FWF) project number 10.55776/COE12.

References

[1] Musleh Alharthi and Ausif Mahmood. xlstmtime: Long-term time series forecasting with xlstm. *AI*, 5(3): 1482–1495, 2024.

[2] Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xlstm: Extended long short-term memory. In *Advances in Neural Information Processing Systems*, volume 37, pages 107547–107603, Red Hook, NY, USA, 2024. Curran Associates Inc.

[3] Marin Biloš, Johanna Sommer, Syama Sundar Rangapuram, Tim Januschowski, and Stephan Günnemann. Neural flows: Efficient alternative to neural odes. In *Advances in Neural Information Processing Systems*, volume 34, pages 21325–21337, Red Hook, NY, USA, 2021. Curran Associates Inc.

[4] Defu Cao, Yujing Wang, Juanyong Duan, Ce Zhang, Xia Zhu, Conguri Huang, Yunhai Tong, Bixiong Xu, Jing Bai, Jie Tong, and Qi Zhang. Spectral temporal graph neural network for multivariate time-series forecasting. In *Advances in Neural Information Processing Systems*, volume 33, Red Hook, NY, USA, 2020. Curran Associates Inc.

[5] Zhengping Che, Sanjay Purushotham, Kyunghyun Cho, David Sontag, and Yan Liu. Recurrent Neural Networks for Multivariate Time Series with Missing Values. *Scientific Reports*, 8(1):6085, 2018.

[6] Chris Giordano, Meghan Brennan, Basma Mohamed, Parisa Rashidi, François Modave, and Patrick Tighe. Accessing Artificial Intelligence for Clinical Decision-Making. *Frontiers in Digital Health*, 3:645232, 2021.

[7] Max Horn, Michael Moor, Christian Bock, Bastian Rieck, and Karsten Borgwardt. Set functions for time series. In *Proceedings of the International Conference on Machine Learning*, pages 4353–4363. JMLR.org, 2020.

[8] Qihe Huang, Lei Shen, Ruixin Zhang, Shouhong Ding, Binwu Wang, Zhengyang Zhou, and Yang Wang. Crossgnn: Confronting noisy multivariate time series via cross interaction refinement. In *Advances in Neural Information Processing Systems*, volume 36, pages 46885–46902, Red Hook, NY, USA, 2023. Curran Associates Inc.

[9] Alistair E.W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3(1):160035, 2016.

[10] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[11] Christian Klötergens, Tim Dernedde, Lars Schmidt-Thieme, and Vijaya Krishna Yalavarthi. Mixing it up: Exploring mixer networks for irregular multivariate time series forecasting, 2026.

[12] Maurice Kraus, Felix Divo, Devendra Singh Dhami, and Kristian Kersting. xlstm-mixer: Multivariate time series forecasting by mixing via scalar memories, 2025.

[13] Jeong Min Lee and Milos Hauskrecht. Modeling multivariate clinical event time-series with recurrent temporal mechanisms. *Artificial Intelligence in Medicine*, 112:102021, 2021.

[14] Bryan Lim and Stefan Zohren. Time Series Forecasting With Deep Learning: A Survey. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 379(2194): 20200209, 2021.

[15] Yuqi Nie, Nam H. Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *International Conference on Learning Representations*, 2023.

[16] Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent odes for irregularly-sampled time series. In *Advances in Neural Information Processing Systems*, volume 32, Red Hook, NY, USA, 2019. Curran Associates Inc.

[17] Satya Narayan Shukla and Benjamin Marlin. Multi-time attention networks for irregularly sampled time series. In *International Conference on Learning Representations*, 2021.

[18] Sunghyun Sim, Dohee Kim, and Hyerim Bae. Correlation recurrent units: A novel neural architecture for improving the predictive performance of time-series data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(12):14266–14283, 2023.

[19] Mahanazuddin Syed, Shorabuddin Syed, Kevin Sexton, Hafsa Bareen Syeda, Maryam Garza, Meredith Zozus, Farhanuddin Syed, Salma Begum, Abdullah Usama Syed, Joseph Sanford, and Fred Prior. Application of machine learning in intensive care unit (icu) settings using mimic dataset: Systematic review. *Informatics*, 8(1), 2021.

[20] Davy Van De Sande, Michel E. Van Genderen, Joost Huiskens, Diederik Gommers, and Jasper Van Bommel. Moving from bytes to bedside: a systematic review on the use of artificial intelligence in the intensive care unit. *Intensive Care Medicine*, 47(7):750–760, 2021.

[21] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[22] Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark, 2025.

[23] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.

[24] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, and Chengqi Zhang. Graph wavenet for deep spatial-temporal graph modeling. In *Proceedings of the International Joint Conference on Artificial Intelligence*, page 1907–1913. AAAI Press, 2019.

[25] Zonghan Wu, Shirui Pan, Guodong Long, Jing Jiang, Xiaojun Chang, and Chengqi Zhang. Connecting the dots: Multivariate time series forecasting with graph neural networks. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 753–763, New York, NY, USA, 2020. Association for Computing Machinery.

[26] Kun Yi, Qi Zhang, Wei Fan, Hui He, Liang Hu, Pengyang Wang, Ning An, Longbing Cao, and Zhendong Niu. Fouriergnn: Rethinking multivariate time series forecasting from a pure graph perspective. In *Advances in Neural Information Processing Systems*, volume 36, Red Hook, NY, USA, 2023. Curran Associates Inc.

[27] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11121–11128, 2023.

[28] Jiawen Zhang, Shun Zheng, Wei Cao, Jiang Bian, and Jia Li. Warpformer: A multi-scale modeling approach for irregular clinical time series. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, page 3273–3285. Association for Computing Machinery, 2023.

[29] Weijia Zhang, Chenlong Yin, Hao Liu, Xiaofang Zhou, and Hui Xiong. Irregular multivariate time series forecasting: A transformable patching graph neural networks approach. In *International Conference on Machine Learning*, 2024.

[30] Xiang Zhang, Marko Zeman, Theodoros Tsiligkaridis, and Marinka Zitnik. Graph-guided network for irregularly sampled multivariate time series. In *International Conference on Learning Representations*, 2022.

[31] Yunhao Zhang and Junchi Yan. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations*, 2023.

Applied Vision

Obstacle Detection Pipeline using Monocular Depth Estimation in Mobile Robotics

Christian Schweighofer
University of Applied Sciences Upper Austria
Roseggerstraße 15, Wels, 4600, Austria
S2410564020@fhooe.at

Michael Zauner
University of Applied Sciences Upper Austria
Roseggerstraße 15, Wels, 4600, Austria
Michael.Zauner@fh-wels.at

Abstract

Autonomous mobile robots must navigate dynamic environments safely, yet high-end depth sensors are often expensive or impractical. Monocular cameras are widely available, but estimating metric depth and detecting obstacles in real time remain challenging. We address this by implementing a pipeline that combines monocular depth estimation with metric scale calibration, 3D back-projection, filtering, and clustering. Our marker-based calibration achieves a depth RMSE as low as 13 mm, while the proposed pipeline successfully detects all 8 obstacles in our evaluation. With OpenVINO optimizations, the model achieves an inference rate of up to 17 FPS, establishing a foundation for real-time processing. Overall, the pipeline demonstrates promising results for safe navigation using only monocular cameras on resource-constrained robots, evaluated in the context of the international robotic contest Eurobot.

1 Introduction

In autonomous mobile robotics, reliable obstacle detection in dynamic environments is essential for avoiding collisions. While high-end platforms often employ sensors such as LiDAR or stereo vision to perceive the environment in 3D, these solutions impose significant hardware costs and spatial constraints. This is particularly relevant in competitive settings such as Eurobot, where robots must navigate dynamic environments while adhering to strict size limits [1, 10, 12].

Eurobot is an international robotic contest that takes place in Europe, with teams participating from around the world. The goal is to build a robot that performs tasks based on the yearly changing set of rules against another robot. The one scoring more points wins the match. A situation that illustrates the need for an obstacle detection system is shown in Figure 1.

Most Eurobot platforms already use monocular cameras and embedded compute for task-specific computer vision, such as ArUco [7] marker localization. This paper proposes a method that leverages monocular vision for obstacle detection. By integrating a foundation model (Depth Anything V2 [4, 18, 19]) with polynomial regression for metric calibration, we transform 2D imagery into metric 3D point clouds. The resulting workflow enables the detection of obstacles (e.g., Eurobot game elements) and thereby supports safer navigation without requiring additional sensors.

This work demonstrates a lightweight, ArUco-based calibration strategy for converting foundation-model depth into metric 3D point clouds and shows the feasibility of real-time obstacle detection on resource-constrained mobile robots.

The remainder of this paper is structured as follows: Section 2 details the processing pipeline from image to clustered obstacles; Section 3 evaluates the system using game elements from the Eurobot 2024, 2025, and 2026 seasons; Section 4 summarizes the results and outlines future work.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

2 Methods

To detect obstacles and assess whether they may lead to a collision, information such as object size and distance is required. These parameters are difficult to extract directly from a 2D image. The following steps describe how an RGB image is converted into a 3D point cloud of the environment, which is then filtered and clustered to obtain obstacle hypotheses.

2.1 Monocular depth estimation

Recovering depth information from a single RGB camera is challenging [15]. In recent years, considerable research has been directed toward monocular depth estimation foundation models [3, 8, 13, 17]. These models take an image as input and predict per-pixel relative (and sometimes metric) depth. Their outputs are often proportional to either depth $\hat{y}$ or disparity $\hat{d}$ (inverse depth). Given the predicted relative depth $\hat{y} = \frac{1}{\hat{d}}$, the metric depth Z can be recovered through a linear transformation:

$$Z = a_1 \hat{y} + a_0. \tag{1}$$

The unknown regression coefficients a_0, a_1 can be estimated via a calibration procedure. Specifically, an optimization algorithm can be used to find parameters that minimize the error between ground-truth and predicted depth over a set of pixels. Results from [2] indicate that a second-order polynomial regressor ($Z = a_2\hat{y}^2 + a_1\hat{y} + a_0$) can further reduce the metric-depth error. The regression coefficients a_0, a_1, a_2 are identified from a sequence of frames containing an ArUco marker moving along the camera optical axis, where the absolute depth is obtained by solving the perspective-n-point (PnP) problem [11]. The regressor is tuned to minimize the error between the marker center depth estimated via PnP and the predicted depth at the corresponding pixel. This approach avoids the need for additional sensors during calibration.

2.2 2D-to-3D back projection with pinhole model and intrinsics

The standard pinhole model allows the conversion from 3D-space into the 2D-image plane via

$$s \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K} \begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix}. \tag{2}$$

This assumes that the intrinsic camera matrix $\mathbf{K}$ is known and that an undistorted image is used [20]. Given metric depth $Z_c = s$ for each pixel along the Z-axis, the formula can be rearranged to

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \end{bmatrix} = Z_c \mathbf{K}^{-1} \begin{bmatrix} u \\ v \\ 1 \end{bmatrix}. \tag{3}$$

This transformation projects the 2D depth map into a 3D point cloud. Initially, the points are expressed in the camera coordinate frame; other frames can be obtained by applying an extrinsic transformation.

2.3 Point cloud filtering

Once the environment is represented as a 3D point cloud, points that do not correspond to obstacles are filtered out. We remove points beyond a maximum distance by thresholding the Euclidean distance to the origin. In our experimental setup, most remaining points correspond to the planar floor; thus, we remove floor points by applying a height threshold parallel to the floor plane [14]. While RANSAC-based plane fitting or Hough-transform-based methods offer robust plane estimation [6, 9], a fixed-height pass-through filter was chosen for computational efficiency.

2.4 Object Clustering

The final step is to cluster the remaining points into obstacle candidates. For this purpose, we apply DBSCAN, a density-based clustering algorithm [5, 16]. Once clusters are obtained, properties such

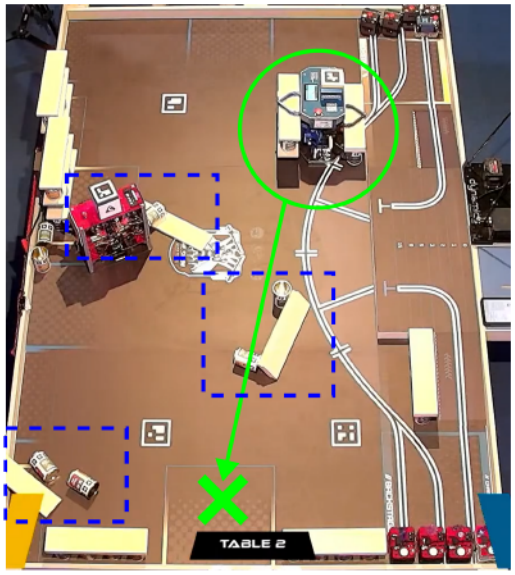


Figure 1: Match situation illustrating a potential obstacle encounter in Eurobot season 2025.

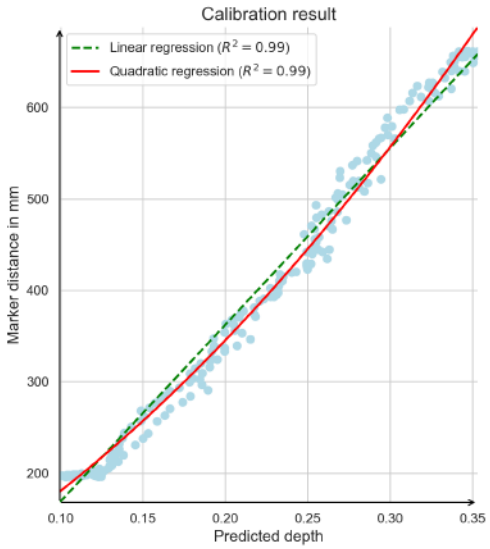


Figure 2: Calibration results of first and second order regression.

as obstacle size and distance can be estimated, enabling downstream decisions for path planning and collision avoidance.

3 Results

This section evaluates the results of the proposed obstacle detection workflow. The situations in Figure 3a, 3f, and 3k are used as test scenarios. The tests were conducted on a Surface Pro 7 with an Intel Core i5-1035G4 CPU and 8GB of RAM. The Depth Anything V2 - base model with an input image size of $(256, 256)$ pixels was used for inference.

The basis of the proposed workflow lies in an accurate conversion from 2D-image points into a 3D-spatial representation. The ground truth distance was obtained via PnP-estimation of an ArUco marker. The results of our calibration approach, which utilizes ground truth distance to find regressor parameters for conversion from relative to absolute depth, can be seen in Figure 2.

Consistent with the findings of [2], both the first-order and second-order polynomial regressors approximate the ground truth well. The first-order approximation achieves an RMSE of 18 mm, while the second-order polynomial achieves an RMSE of 13 mm. In the following experiments, we use the second-order polynomial.

After converting each pixel from 2D to 3D using Eq. 3, the resulting point clouds for the test scenarios are shown in Figures 3c, 3h, and 3m.

After removing points beyond 1000 mm and points below a height threshold of 30 mm above the floor, the filtered point clouds in Figures 3d, 3i, and 3n remain. DBSCAN ($\epsilon = 0.008$, min. points $= 20$) then segments the scene into discrete clusters and successfully detects all 8 obstacles in our evaluation (Figures 3e, 3j, and 3o).

While the above implementation shows promising results, its real-time performance is currently insufficient. Using the calibration frames, we measured inference time and calibration error on different models and optimizations, as shown in Table 1. Non-inference steps—including point cloud generation, filtering, and clustering—currently account for approximately 0.2 s per frame in the unoptimized Python implementation. These steps can be further accelerated using standard techniques such as voxelization, more efficient algorithms, and compiled code. Overall, these results indicate that the proposed pipeline is feasible for real-time obstacle detection on resource-constrained mobile robots.

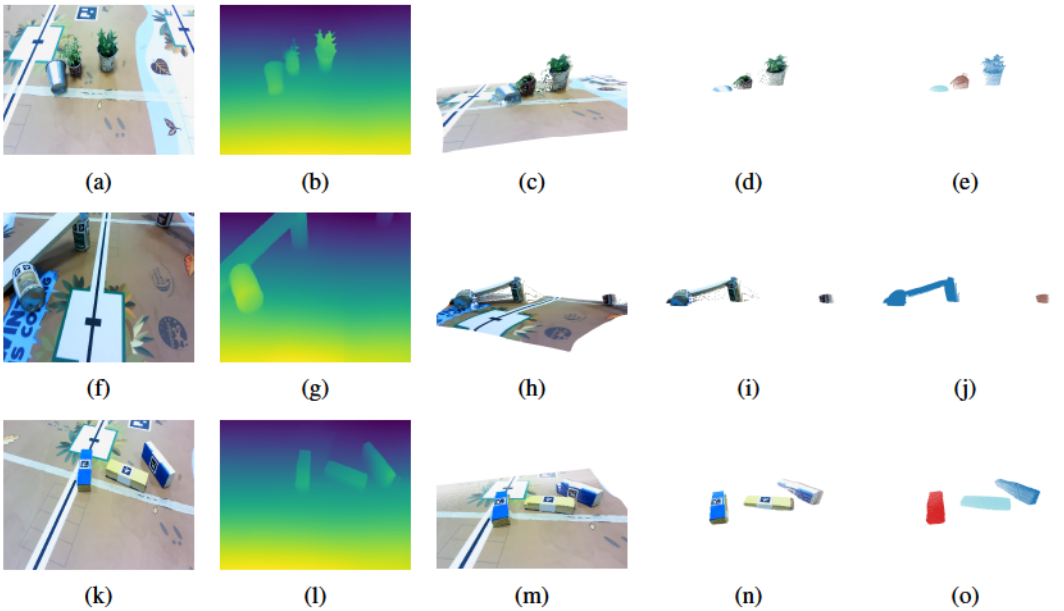


Figure 3: Sample obstacle-detection results for three seasons of Eurobot (2024–2026). Rows correspond to the different seasons, while columns illustrate the processing pipeline from left to right: undistorted RGB input, estimated depth map, 3D point-cloud projection, filtered point cloud, and final clustered obstacles.

Table 1: Inference time and calibration error (RMSE in mm) for selected depth models at 256×256 px, except OpenVINO (OV) at 252×252 px.

Model	Time [ms]	RMSE 1st order [mm]	RMSE 2nd order [mm]
DepthAnythingV2 Base	840	18	13
DepthAnythingV2 Small	325	36	35
DPT Hybrid [13]	1019	29	24
DepthAnythingV2 Small (OV)	57	25	24

4 Conclusion

In conclusion, we show that monocular depth estimation models can be used for obstacle detection in the Eurobot setting. Using an ArUco-based calibration procedure, we map relative predictions to a metric 3D point cloud representation and successfully detect representative game elements from recent seasons. However, metric accuracy of obstacle dimensions and distances beyond the calibration setup remains unverified, as the inferred scale was not validated on an independent dataset. While the current implementation is neither fully optimized nor evaluated on the target platform, the results indicate that the proposed pipeline is a viable alternative when dedicated depth sensors are impractical.

Future work should validate and improve the metric scale conversion on independent test data and under changes in camera pose, scene composition, and illumination. In addition, further engineering efforts could benchmark and optimize the pipeline on embedded platforms.

Acknowledgments and Disclosure of Funding

The authors used AI tools from OpenAI (ChatGPT) and Google (Gemini) to assist with writing and coding. All methodological choices and experimental results were independently designed and verified by the authors. The authors gratefully acknowledge the financial support of the University of Applied Sciences Upper Austria for this project.

References

[1] Seongmin Ahn, Yunjin Kyung, Seunguk Choi, Dongyoung Choi, and Dongil Choi. Monocular vision-based obstacle height estimation for mobile robot. *Applied Sciences*, 15(23), 2025. doi: 10.3390/app152312711.

[2] Soofiyan Atar, Yuheng Zhi, Florian Richter, and Michael Yip. Kinedepth: Utilizing robot kinematics for online metric depth estimation. 2025. doi: 10.48550/arXiv.2409.19490.

[3] Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. ZoeDepth: Zero-shot Transfer by Combining Relative and Metric Depth. 2023. doi: 10.48550/arXiv.2302.12288.

[4] Xingshuai Dong, Matthew A. Garratt, Sreenatha G. Anavatti, and Hussein A. Abbass. Towards real-time monocular depth estimation for robotics: A survey. 2021. doi: 10.48550/arXiv.2111.08600.

[5] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[6] Martin A. Fischler and Robert C. Bolles. Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. In Martin A. Fischler and Oscar Firschein, editors, *Readings in Computer Vision*, pages 726–740. Morgan Kaufmann, San Francisco (CA), 1987.

[7] S. Garrido-Jurado, R. Muñoz-Salinas, F.J. Madrid-Cuevas, and M.J. Marín-Jiménez. Automatic generation and detection of highly reliable fiducial markers under occlusion. *Pattern Recognition*, 47(6):2280–2292, 2014.

[8] Clément Godard, Oisin Mac Aodha, Michael Firman, and Gabriel Brostow. Digging Into Self-Supervised Monocular Depth Estimation. 2019. doi: 10.1109/ICCV.2019.00393.

[9] Rostislav Hulik, Michal Spanel, Pavel Smrz, and Zdenek Materna. Continuous plane detection in point-cloud data based on 3D Hough Transform. *Journal of Visual Communication and Image Representation*, 25(1):86–97, 2014. doi: 10.1016/j.jvcir.2013.04.001.

[10] Kornél Katona, Husam A. Neamah, and Péter Korondi. Obstacle avoidance and path planning methods for autonomous navigation of mobile robot. *Sensors*, 24(11), 2024. doi: 10.3390/s24113573.

[11] Vincent Lepetit, Francesc Moreno-Noguer, and Pascal Fua. EPnP: An accurate O(n) solution to the PnP problem. *International Journal of Computer Vision*, 81, 2009. doi: 10.1007/s11263-008-0152-6.

[12] Yu Liu, Shuting Wang, Yuanlong Xie, Tifan Xiong, and Mingyuan Wu. A review of sensing technologies for indoor autonomous mobile robots. *Sensors*, 24(4), 2024. doi: 10.3390/s24041222.

[13] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision Transformers for Dense Prediction. 2021. doi: 10.48550/arXiv.2103.13413.

[14] Radu Bogdan Rusu and Steve Cousins. 3D is here: Point Cloud Library (PCL). In *2011 IEEE International Conference on Robotics and Automation*, pages 1–4, 2011. doi: 10.1109/ICRA.2011.5980567.

[15] Ashutosh Saxena, Sung Chung, and Andrew Ng. Learning depth from single monocular images. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

[16] Erich Schubert, Jörg Sander, Martin Ester, Hans Peter Kriegel, and Xiaowei Xu. DBSCAN Revisited, Revisited: Why and How You Should (Still) Use DBSCAN. *ACM Trans. Database Syst.*, 42(3), 2017. doi: 10.1145/3068335.

[17] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data. 2024. doi: 10.1109/CVPR52733.2024.00987.

[18] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2, 2024. arXiv preprint arXiv:2406.09414.

[19] Jiuling Zhang. Survey on monocular metric depth estimation. 2025. doi: 10.48550/arXiv.2501.11841.

[20] Zhengyou Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22:1330–1334, 2000. doi: 10.1109/34.888718.

AI-Based Optimization of Roadside Mowing Operations in Austria

Roland Perko, Stefanie Onsori-Wechtitsch, Helmut Neuschmied,
Peter Schallauer, Katharina Hofer-Schmitz
JOANNEUM RESEARCH, Graz, Austria, {firstname.lastname}@joanneum.at

Michaela Stolz
biohelp, Vienna, Austria, {firstname.lastname}@biohelp.at

Abstract

Roadside vegetation management is vital for traffic safety, efficiency, and biodiversity. Conventional mowing relies on routine schedules and manual inspections, limiting route optimization and adaptation to changing vegetation growth. To address the challenges of roadside maintenance, we developed MeadowLevelSeg, a deep learning approach that employs Mask2Former to map meadow heights into precise 5 cm classes. Around 800 high-resolution roadside images were recorded and annotated. Performance is evaluated using a novel Distance-Aware Accuracy metric, which takes the ordinal nature of height classes into account. Initial results demonstrate that the model effectively identifies different meadow heights and high-growth zones, achieving a mean absolute error of less than 7 cm using monocular images. This provides a robust basis for automated maintenance scheduling.

1 Introduction

Roadside vegetation management is essential for traffic safety, operational efficiency, and biodiversity along public roads. Traditionally, roadside mowing is labor-intensive, guided by weekly lists and routine checks, with limited capacity to optimize routes or adapt to dynamic growth patterns. This necessitates balancing economic and environmental goals, characterizing roadside maintenance as a complex multi-objective optimization problem [8]. This work aims to leverage AI-based computer vision for estimating meadow height to optimize mowing schedules and routes while balancing road safety and ecological goals. The visibility of safety equipment close to the road (e.g., the reflectors of guiding posts) is given priority. Furthermore, overgrown pathways that pose risks to children and cyclists are systematically identified and reported at an early stage. The research is also motivated by the *Federal Law on the Procurement and Use of Clean Road Maintenance Vehicles* [1], which requires road maintenance authorities to transition to emission-free fleets by 2030. Since those vehicles are limited in their maximum range, optimized route planning is mandatory. Although computer vision is well-established in agriculture, its application to measuring the height of roadside meadows remains largely unexplored. While existing Digital Twin approaches (e.g., [9]) address the safety risks of vegetation overgrowth, they primarily focus on detecting overhanging trees rather than providing fine-grained classification of meadow heights. A similar approach was followed by [7], who assess ground conditions, such as soil moisture and trafficability, directly on a mower. This work aspires to provide road maintenance authorities with actionable insights into when and where mowing is needed, improving transparency, reducing unnecessary interventions, and supporting sustainable management. Compared to costly LiDAR systems, our camera-based MeadowLevelSeg offers a more cost-effective, easily integratable alternative. This study presents initial results on the semantic segmentation of meadow height classes, which serves as a prerequisite for GIS-based mowing optimization. For illustration, Figure 1 (left) shows a mowing vehicle in operation.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

Figure 1: Road maintenance service vehicle from the province of Styria during mowing activity (left). Statistical breakdown of the annotated dataset (right). The histograms show a comparison between total annotated area (top) and polygon counts (bottom) across 5 cm height increments.

2 Data Collection

A sensor platform was constructed which combines high-end imaging (dual Basler acA4096-30uc industrial cameras), low-cost monocular imaging (GoPro Hero 13), edge AI processing (Intel i7, NVIDIA RTX 3070 GPU), high-accuracy EGNSS positioning, and synchronized custom hardware triggers. The system was installed in a road maintenance vehicle and its components were water-proofed and thoroughly calibrated to ensure radiometric fidelity. From April to October 2025, the sensor platform captured roadside imagery along predefined routes around Feldbach in Styria. A GIS-based planning tool was used to select routes and features, with the aim of targeting different vegetation conditions, slopes, bushes, trees, and pollinator habitats. Annotations were performed using the *Computer Vision Annotation Tool (CVAT)* [4], focusing on meadow height classes from 5 to 80 cm with increments of 5 cm. All relevant roadside strips (50 cm wide) next to roads and paths were annotated for a given dataset, which currently comprises approximately 800 images. Figure 1 (right) shows the distribution of meadow classes in the annotated data, with the percentage share of the whole annotated area in image space on the top and the annotated polygons on the bottom. While the distribution of annotated pixels suggests a relatively balanced dataset across almost all classes, a comparison with the number of annotated polygons reveals a significant under-representation of the classes 30 cm to 80 cm. This highlights the need for targeted data collection during the next vegetation period, in order to improve the diversity of the training data.

3 Methodology

The primary goal of the proposed MeadowLevelSeg framework is to estimate meadow heights along roadside strips. For this specific semantic segmentation task, we adopted Mask2Former [3] architecture, as its mask-classification approach outperforms traditional per-pixel models like UPerNet [10] in capturing the fine, irregular boundaries of vegetation. Furthermore, Mask2Former provides an optimal balance between inference speed and accuracy, which is crucial for the high-resolution processing demanded by roadside maintenance tasks. The dataset consists of 833 high-resolution roadside monocular images ($4,112 \times 2,176$ pixels) which were annotated in CVAT and exported as segmentation masks. This set was randomly split into 617 training, 152 validation, and 64 test images. The architecture was implemented within the OpenMMLab framework [2], using a Swin Transformer backbone [5] initialized with pretrained ADE20K weights [11] before being fine-tuned on our domain-specific roadside data. To better align the model with the specific application requirements, the training strategy was adapted by modifying the loss formulation and optimization process (also cf. [6]). Even minor height class deviations have a significant impact on the standard metric, which is why distance-aware accuracy (DistAwareAcc) was introduced as a more representative performance metric. This semantic segmentation metric measures accuracy while accounting for the semantic distance between classes. Instead of treating all misclassifications equally, it penalizes predictions proportionally to how different the predicted height class is from the ground-truth class.

Distance-Aware Accuracy. Let K be the total number of classes, where classes 1 to $K-1$ represent height classes and class 0 represents the background class. $C \in \mathbb{N}^{K\times K}$ be the confusion matrix, where C_{ij} denotes the number of pixels of ground-truth class i predicted as class j. We define a class-similarity matrix $W \in [0,1]^{K\times K}$ that encodes the semantic proximity between categories. The weights W_{ij} are determined by the ordinal distance between class indices for meadow, while ensuring no tolerance for background misclassifications:

$$W_{ij} = \begin{cases} 1.0 & \text{if } i = j \\ \max(0, 1.0 - 0.1 \cdot |i-j|) & \text{if both } i, j \text{ are height classes and } i \neq j \\ 0.0 & \text{if either } i \text{ or } j \text{ is the background class and } i \neq j \end{cases} \tag{1}$$

By setting $W_{ij} = 0.0$ for any mismatch involving the background, the model maintains a strict boundary between meadow and non-meadow areas. For meadow classes, the weight decreases in steps of 0.1 for every discrete height interval (5cm steps). The per-class accuracy is calculated as:

$$\text{DistAwareAcc}_i = \frac{\sum_{j=1}^{K} C_{ij} W_{ij}}{\sum_{j=1}^{K} C_{ij}} \tag{2}$$

The overall DistAwareAcc is the mean across all classes and is used within the loss function to reduce penalization of semantically similar height predictions during training.

4 Results

Intersection over Union (IoU) was utilized as the primary metric for validating the semantic segmentation of roadside meadow height regions. For the test dataset, the model achieved an average IoU of 17.3% across all classes. However, it should be noted that a perfect overlap is intrinsically limited by two factors. Firstly, there is the inherent ambiguity in annotating non-uniform meadow heights. Secondly, there are the perspective-related challenges of defining polygon boundaries at a fine level. Given the inherent uncertainties associated with annotation, the results indicate a high level of reliability for operational use. As demonstrated in Table 1, a comprehensive analysis of the performance of each class is provided. The low scores, particularly the score of 0 for classes 55 cm and 80 cm, can be attributed to the strong under-representation of these classes in both the training and test datasets. With a limited number of ground-truth pixels available, the model predicts very few instances, resulting in minimal to zero overlap. A similar lack of intersection is observed for the 35 cm, 60 cm, and 70 cm classes. In the case of the 80 cm class specifically, the model failed to detect any corresponding areas. As there were no positive predictions (true or false), the precision remains undefined (NaN). The DistAwareAcc metric, which is more relevant for the absolute meadow height estimation, yields significantly higher values. It indicates that while the model is not always capable of accurately identifying the exact height, it frequently assigns a neighboring height class instead of the exact class. This is further demonstrated by the qualitative results, presented in Figure 3. The detected classes are overlaid on the input image. Compared to manual annotations, which often group larger areas into a single height, the model often delivers more precise results with several different height classes. It captures details that were either simplified or overlooked during manual annotations.

Mean Absolute Error (MAE). To quantify the prediction quality, the MAE was calculated across all meadow pixels on the test set. By converting the class indices back to the corresponding heights in centimeters, it was determined that the model's predictions deviate from the actual height by only 6.57 cm on average. This minor discrepancy in spatial height confirms that, although the exact 5 cm category is not achieved, the anticipated trend for operational planning remains highly precise.

Table 1: Segmentation results per meadow height class. While IoU, Precision, and Recall reflect the difficulty of exact classification, the DistAwareAcc shows the capability for operational planning.

Class [cm]	5	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80
IoU [%]	39.0	11.5	23.2	18.0	27.9	15.0	0.5	11.8	5.6	18.4	0.0	5.0	13.8	4.7	8.5	0.0
Precision [%]	56.5	13.7	33.5	28.0	38.4	21.4	5.4	14.8	12.8	30.4	0.0	14.4	17.9	5.9	47.0	NaN
Recall [%]	55.7	42.4	42.9	33.5	50.6	33.2	0.5	36.9	9.1	31.7	0.0	7.2	37.4	19.5	9.3	0.0
DistAwareAcc [%]	82.9	84.1	87.6	87.4	85.0	81.9	76.7	81.7	77.9	82.3	67.0	81.1	70.4	69.9	60.6	56.4

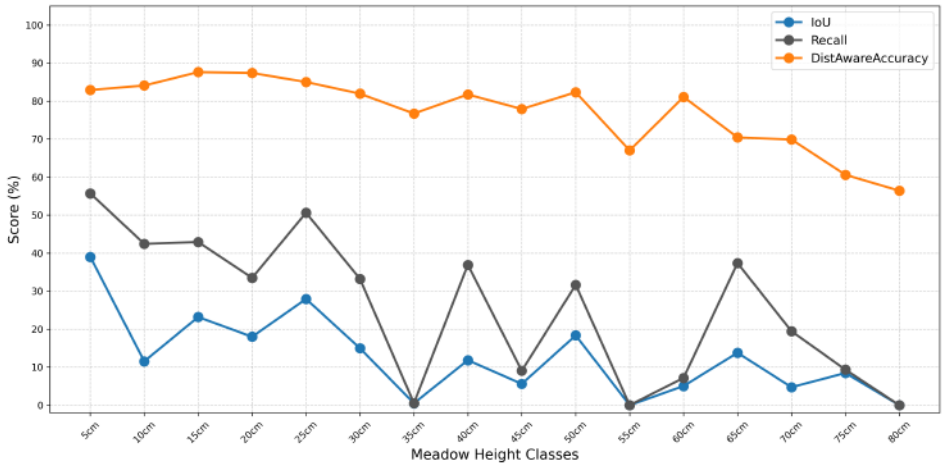


Figure 2: Detection performance of Meadow Heights when considering neighboring classes and their similarities.

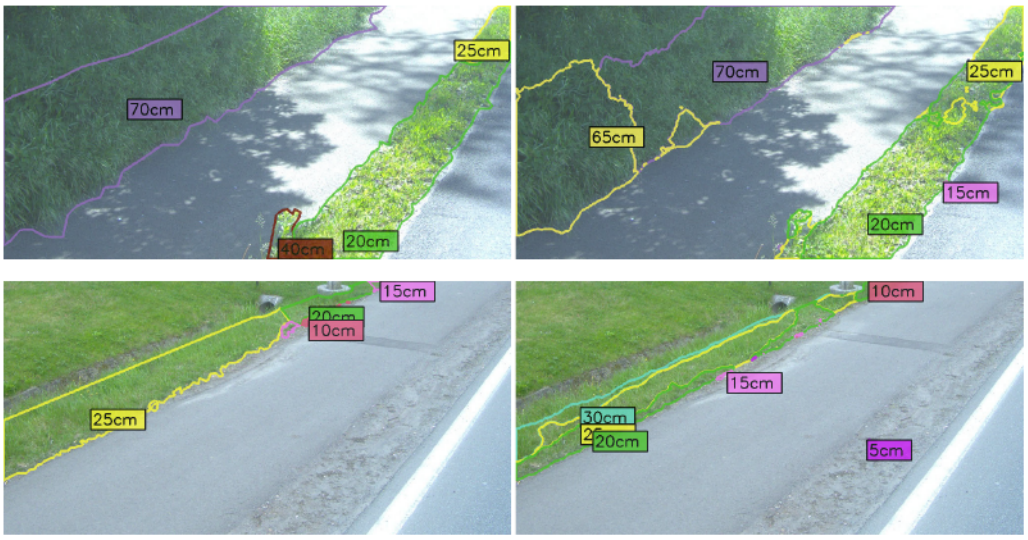


Figure 3: Exemplary results of semantic segmentation into meadow height classes – ground-truth annotated meadow height classes (left), predicted meadow height classes (right). Both examples show cycling paths directly adjacent to the road.

5 Conclusions

The presented work demonstrates feasibility for scalable AI-guided meadow management on Austrian roads. The results indicate promising generalization behavior and reasonable agreement between predictions and ground truth, where the error of the meadow height estimation was below 7 cm. However, under-representation of high meadow classes suggests areas for improvement in future annotation efforts and model refinement, which are planned for the upcoming vegetation period.
The next steps will focus on (1) increasing annotation density across all height classes and diverse geographic routes to improve model generalization, (2) develop an automated mowing-planner for optimized routing based on meadow height classification of an entire road network, and (3) design and implement a proof-of-concept demonstrator, allowing the end users to evaluate the whole system. The groundwork is laid for a collaborative, evidence-based management strategy that balances safety requirements, cost-efficiency, and ecological biodiversity goals.

Acknowledgments and Disclosure of Funding

This research was partly funded by the AI for Green program through the project SmartMowAI (FFG project number 910271).

References

[1] Bundesministerium für Justiz. BGBl. I Nr. 163/2021: Bundesgesetz über die Beschaffung und den Einsatz sauberer Straßenfahrzeuge - Federal Act on the Procurement and Use of Clean Road Vehicles (Road Vehicle Procurement Act). https://www.ris.bka.gv.at/eli/bgbl/I/2021/163, 2021.

[2] Kai Chen, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jiarui Xu, Zheng Zhang, Dazhi Cheng, Chenchen Zhu, Tianheng Cheng, Qijie Zhao Zhao, Buyu Li, Xin Lu, Rui Zhu, Yue Wu, Jifeng Dai, Jingdong Wang, Jianping Shi, Wanli Ouyang, Chen Change Loy, and Dahua Lin. MMDetection: Open MMLab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*, 2019.

[3] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[4] CVAT Contributors. Computer Vision Annotation Tool (cvat). https://github.com/cvat-ai/cvat, 2024. Version 2.48.0.

[5] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *International Conference on Computer Vision*, pages 10012–10022, 2021.

[6] Jun Ma, Jianan Chen, Matthew Ng, Rui Huang, Yu Li, Chen Li, Xiaoping Yang, and Anne L Martel. Loss odyssey in medical image segmentation. *Medical Image Analysis*, 71:102035, 2021.

[7] Christoph Manss, Viktor Martel, and Roman Weisgerber. Assessment of ground conditions in grassland on a mower with artificial intelligence. In *GIL-Jahrestagung, Biodiversität fördern durch digitale Landwirtschaft*, pages 335–340, 2024.

[8] Adriana-Simona Mihaita, Brunelle Marche, Mauricio Camargo, Iman Rahimi, and Christophe Bachmann. Multi-objective modelling of a roadside mowing problem: A case study in France. In *IEEE International Conference on Engineering, Technology and Innovation & International Association For Management of Technology, Joint Conference*, pages 1–8, 2022.

[9] Varun Kumar Reja, Diana Davletshina, Mengtian Yin, Ran Wei, Quentin Felix Adam, Ioannis Brilakis, and Federico Perrotta. A digital twin based approach to control overgrowth of roadside vegetation. In *Proceedings of the International Symposium on Automation and Robotics in Construction*, pages 1–8, 2024.

[10] Tete Xiao, Yingcheng Liu, Bolei Zhou, Yuning Jiang, and Jian Sun. Unified perceptual parsing for scene understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 36–51, 2018.

[11] Bolei Zhou, Hang Zhao, Xavier Puig, Tete Xiao, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Semantic understanding of scenes through the ADE20K dataset. *International Journal of Computer Vision*, 127(3):302–321, 2019.

Synthetic Skeletal Pose Pre-training to Mitigate Data Scarcity in In-Cabin 2D-to-3D Pose Lifting

Thummanoon Kunanuntakij
Computer Vision Lab
TU Wien
Vienna, 1040
thummanoon.kunanuntakij@tuwien.ac.at

Dominik Schörkhuber
Computer Vision Lab
TU Wien
Vienna, 1040
dominik.schoerkhuber@tuwien.ac.at

Margrit Gelautz
Computer Vision Lab
TU Wien
Vienna, 1040
margrit.gelautz@tuwien.ac.at

Abstract

Driver-related factors contribute to nearly 90% of traffic accidents. Estimating 3D driver poses can help track risky behaviors. However, the scarcity of annotated 3D pose data, together with the complexity and high cost of 3D annotation, limits the training of domain-specific estimators. We address this challenge by pre-training 2D-to-3D pose lifting models using synthetic 3D poses from a simulated dataset. In experiments on the Drive&Act dataset, we compare training from scratch with synthetic pre-training while gradually increasing the amount of real-world data. For example, when only 5% of training data is available, MPJPE is reduced from 90.0 mm to 70.9 mm for the GraFormer model. Our results demonstrate that synthetic pre-training consistently reduces estimation errors, particularly when real-world data are limited. Furthermore, synthetic pre-training improves the best fine-tuned results across different models from 48.1 mm to 46.0 mm in our tests.

1 Introduction and Related Work

Driver-related factors such as fatigue and distraction account for an estimated 87.7% of road traffic accidents [7]. To address these risks, driver monitoring systems (DMS) are employed to track risky behaviors and, under the European Union's General Safety Regulation, have been mandatory in new vehicles since 2024 [26]. A core component of many DMS is human pose estimation. Although multi-camera setups can provide more accurate 3D estimates, single-camera systems are simpler and more cost-effective for in-cabin deployment.

Recent deep learning methods achieve highly accurate monocular 2D human pose estimation (2D-HPE), whereas monocular 3D human pose estimation (3D-HPE) remains more challenging due to depth ambiguity introduced by camera projection, requiring learned priors for recovery [4, 33]. This gap has motivated a two-stage formulation that first estimates 2D keypoints (i.e., body joint coordinates) and then lifts them to 3D keypoints using 2D-to-3D pose lifters [19]. Subsequent work has proposed graph-based [30], transformer-based [17], and hybrid architectures [14, 31]. To reduce pose ambiguity, particularly under self-occlusion, temporal models exploit pose sequences [5, 34, 11, 16], while other approaches incorporate image features alongside pose representations [29]. The resulting 3D poses can further be used for various downstream tasks, such as action recognition [13, 20] and pose anomaly detection [8].

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

The scarcity of annotated 3D poses is a major factor contributing to the performance gap. Unlike 2D keypoints, which can be labeled directly from single images, 3D pose annotation requires calibrated multi-view systems for triangulation [28]. Synthetic data has proven to be a scalable alternative for reducing annotation cost and has been successfully applied across domains such as medical imaging [3], autonomous driving [1], hand pose estimation [32], and action recognition [27]. In the context of pose estimation, synthetic datasets can provide large-scale, domain-specific human poses with exact annotations. They have recently also gained attention in in-cabin driver monitoring scenarios [6, 24, 12]. While most prior studies focus on synthetic pre-training for image-based models, synthetic-to-real transfer for pose-based lifting networks remains relatively underexplored, with examples including multi-view 2D-to-3D pose lifters trained on synthetic data [9] and pose-level augmentation via interpolation [10].

In this work, we investigate the benefits of using synthetic poses to pre-train 2D-to-3D pose lifters. Unlike image-based inputs, skeletal poses provide a compact and structured representation that is less sensitive to variations in texture, background, and lighting. By leveraging off-the-shelf pre-trained 2D pose estimators, we focus exclusively on training the lifting stage, enabling rapid development of 3D-HPE models for domain-specific scenarios such as in-cabin driver monitoring.

In our experiments, we test whether pre-training on synthetic pose data enables the lifting model to achieve lower estimation error after fine-tuning with the same amount of real data. Because generating synthetic 3D poses is substantially less costly than annotating them from real images, this approach has the potential to reduce real-data requirements while maintaining comparable performance. We validate our assumptions on Drive&Act [18], an in-cabin driver monitoring dataset.

2 Method

Our setup follows a 2D-to-3D pose lifting scheme [19], in which 2D keypoints are extracted using off-the-shelf estimators and mapped to 3D coordinates via a dedicated lifting network. As illustrated in Figure 1, the pipeline consists of three stages: driver detection (Faster R-CNN [23]), 2D pose estimation (HRNet [25]), and 2D-to-3D lifting. Following a top-down strategy [4, 28, 33], the detector first localizes the driver, 2D keypoints are then estimated within the detected region, and the resulting 2D poses are fed into the lifter to reconstruct 3D joints. Faster R-CNN[1] and HRNet[2] are COCO [15]-pretrained models from MMDetection [2] and MMPose [21], respectively. Leveraging these pre-trained 2D models allows us to focus exclusively on the lifting stage, improving training efficiency by operating on compact skeletal representations. We trained two versions of 2D-to-3D pose lifters on progressively larger subsets of real data, roughly doubling the dataset size at each step. The first set of models was trained from scratch, while the second set was pre-trained on synthetic poses before fine-tuning. We then compared the estimation errors between the two versions.

Training of 2D-3D pose lifters We selected five different 2D-to-3D pose lifters which represent common choice of deep learning architectures, SimpleBL [19], SemGCN [30], Graformer [31], GraphMLP[14], and JointFormer [17], using the official source codes if available. The training procedure for the 2D-to-3D pose lifters was identical for both pre-training and fine-tuning. Models were trained for up to 200 epochs using the Adam optimizer, with mean squared error (MSE) as the loss function and mean per-joint position error (MPJPE) [19] as the evaluation metric. Validation MPJPE was assessed every five epochs, and the model with the lowest validation MPJPE was retained, with early stopping applied if no improvement was observed for three consecutive evaluations. Training was performed with a batch size of 64, an initial learning rate of 0.001 decayed by a factor of 0.96 per epoch, and gradient clipping with a maximum norm of 1.0 to ensure stable training [22]. The model achieving the best validation MPJPE was selected for testing. For Drive&Act, unannotated occluded keypoints were excluded from both loss computation and evaluation.

3 Datasets

The proposed approach is evaluated on real and synthetic driver pose datasets. We adopt the COCO[15]-style skeletal pose representation. Due to the confined cabin space, only the 13 upper-

[1]MMDetection Config: `faster-rcnn_r50-caffe_fpn_ms-1x_coco-person`
[2]MMPose Config: `td-hm_hrnet-w48_udp-8xb32-210e_coco-384x288`

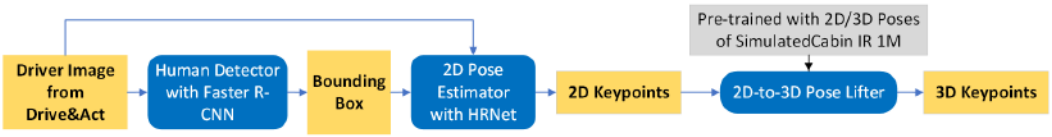


Figure 1: The setup of the 3D Driver Pose Estimation pipeline. The blue rounded boxes denote estimation models, while the yellow boxes represent data.

body keypoints (from the hips upward) are retained. Both 2D and 3D keypoints are centered at the neck, defined as the midpoint between the left and right shoulders. The 2D poses are normalized by the maximum horizontal and vertical extents. Real poses are extracted from Drive&Act [18], while synthetic poses are obtained from SimulatedCabin IR 1M [24].

Drive&Act (Figure 2a) is a publicly available multi-modal driver monitoring dataset [18] containing recordings of 15 subjects performing 34 in-vehicle activities. The provided 3D joint coordinates were obtained by triangulating OpenPose-based 2D detections from three views, with joints occluded in at least two views left unannotated. For our experiments, only the near-infrared (NIR) recordings from the *center_mirror* view were used. The data are split by subject: subject 10 for validation, subjects 11–14 for testing, and the remaining subjects for training.

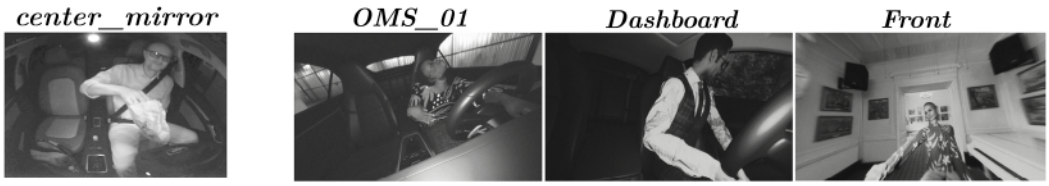
SimulatedCabin IR 1M (Figure 2b) is a simulated data set that contains one million driver poses, extending Sagmeister et al. [24]. It provides simulated NIR images of 3D human models seated in a vehicle under diverse movements, interiors, lighting conditions, and 10 camera viewpoints. Poses are randomly generated using inverse kinematics, enabling the exact computation of corresponding 2D and 3D keypoint annotations, which serve as training data for synthetic pre-training of the lifting model. As no exact counterpart to the Drive&Act *center_mirror* view exists, we use three front-facing views, *OMS_01* (which refers to the *O*ccupant *M*onitoring *S*ystem mounted on top of the dashboard), *Dashboard*, and *Front*, for pre-training, as shown in Figure 2.

4 Experiments, Results and Conclusion

Experiment Setup We compared models trained from scratch with those using pre-training by defining a progressive training scheme with increasing amounts of real data. Specifically, each model was trained on eight different training subsets, with each step approximately doubling the size of the previous one. The subsets consisted of randomly selected 5%, 10%, 25%, and 50% of the data from a single subject; the full data from one subject (100%); random subsets of two and four subjects; and finally the complete training set containing all eight subjects. For reference, 5% of one subject corresponds to approximately 102 frames. For the single-subject settings (5%–100%), training was repeated eight times, once per subject. To maintain a comparable number of runs, we also generated eight random subsets for the two- and four-subject settings. The full eight-subject training set was trained only once. This procedure was applied to both from-scratch and pre-trained models. All models were evaluated on the same test set, and the average test MPJPE for each subset size was reported across runs. As in the training, occluded keypoints were excluded from evaluation.

Results The results in Table 1 compare models trained from scratch with synthetic pre-trained variants (*PT*). Pre-training reduces MPJPE in most settings, but the gap narrows as more real data are used. For instance, GraFormer improves by 19.1 mm, 12.8 mm, 3.6 mm, and shows no improvement when trained on 5%, 50% of one subject, two subjects, and all eight subjects, respectively. Figure 3 illustrates this trend with an example from GraFormer. SimpleBL shows strong improvements in low-data settings, with reductions of 316.3 mm at the 5% setting and 44.4 mm at 50%. In contrast, the hybrid models GraphMLP and GraFormer show smaller improvements, typically under 15 mm.

Conclusion Our results show that synthetic pre-training generally reduces MPJPE for 2D-to-3D pose lifting in our driver monitoring scenarios, particularly when real data are limited, in line with prior findings [24]. Although hybrid models benefit less from pre-training, they achieve low MPJPE across settings, suggesting better data efficiency. A more detailed investigation of the relationship between model architectures and synthetic pre-training could be a topic for future work.



(a) Drive&Act (b) SimulatedCabin IR 1M

Figure 2: Example frames from Drive&Act and SimulatedCabin IR 1M. The name of each respective camera view is indicated at the top. The three most similar views from SimulatedCabin IR 1M to the *center_mirror* view of Drive&Act were chosen.

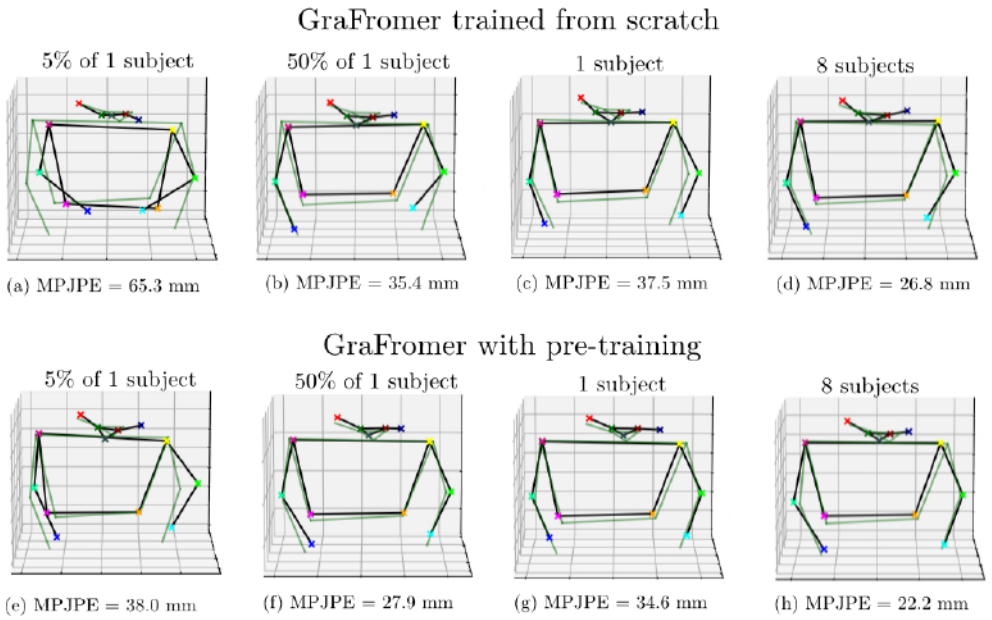


Figure 3: Example 3D pose estimations from GraFormer trained with different amounts of Drive&Act training data. Green indicates ground truth; black indicates the estimate. The input frame is from Subject 12.

	Average Test Set MPJPE[mm]							
	Amount of fine-tuning data							
Model	5% of a subject	10% of a subject	25% of a subject	50% of a subject	Single subject	Two subjects	Four subjects	Eight subjects
SimpleBL[19]	404.3	477.8	251.8	117.3	75.2	62.5	55.2	48.6
***PT* SimpleBL**	88.0	76.7	72.6	72.9	73.4	66.7	52.9	**46.0**
SemGCN[30]	115.9	93.9	82.2	81.1	79.1	69.3	60.6	54.1
***PT* SemGCN**	83.5	74.8	69.7	69.2	67.6	58.4	54.2	48.6
GraphMLP[14]	82.4	77.2	73.3	69.7	65.7	60.8	57.0	53.7
***PT* GraphMLP**	**67.4**	69.6	64.9	61.9	61.0	55.3	57.9	57.2
GraFormer[31]	90.0	78.3	72.9	72.0	66.6	58.5	57.1	48.1
***PT* GraFormer**	70.9	**65.1**	66.2	**59.2**	**59.2**	**54.9**	52.8	48.1
JointFormer[17]	109.9	105.3	99.1	91.7	85.4	65.9	54.2	50.2
***PT* JointFormer**	69.6	66.5	**61.3**	66.3	66.0	55.1	**50.0**	**46.0**

Table 1: Average test results on the Drive&Act *center_mirror* view (see section 4). *PT* denotes synthetic pre-training. Bold indicates the best model per training size; the overall best result is underlined.

Acknowledgments and Disclosure of Funding

This work was funded by FFG (BMK) in parts under the SyntheticCabin (No. 884336) project and in parts under the UNISCOPE-3D (No. 911019/923852/937153) project. Synthetic datasets were provided by emotion3D[3].

References

[1] Chen, C., Seff, A., Kornhauser, A., and Xiao, J. (2015). Deepdriving: Learning affordance for direct perception in autonomous driving. In *IEEE International Conference on Computer Vision*, pages 2722–2730.

[2] Chen, K., Wang, J., Pang, J., Cao, Y., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Xu, J., Zhang, Z., Cheng, D., Zhu, C., Cheng, T., Zhao, Q., Li, B., Lu, X., Zhu, R., Wu, Y., Dai, J., Wang, J., Shi, J., Ouyang, W., Loy, C. C., and Lin, D. (2019). MMDetection: Open MMLab detection toolbox and benchmark. *arXiv preprint arXiv:1906.07155*.

[3] Chen, R. J., Lu, M. Y., Chen, Y, T., Williamson, D. F. K., and Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature biomedical engineering*, 5(6):493–497.

[4] Chen, Y., Tian, Y., and He, M. (2020). Monocular human pose estimation: A survey of deep learning-based methods. *Computer Vision and Image Understanding*, 192(102897).

[5] Cheng, Y., Yang, B., Wang, B., and Tan, R. T. (2020). 3d human pose estimation using spatio-temporal networks with explicit occlusion training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):10631–10638.

[6] Da Cruz, S. D., Wasenmuller, O., Beise, H.-P., Stifter, T., and Stricker, D. (2020). SVIRO: Synthetic vehicle interior rear seat occupancy dataset and benchmark. In *IEEE Winter Conference on Applications of Computer Vision*, page 962–971.

[7] Dingus, T. A., Guo, F., Lee, S., Antin, J. F., Perez, M., Buchanan-King, M., and Hankey, J. (2016). Driver crash risk factors and prevalence evaluationusing naturalistic driving data. *Proceedings of the National Academy of Sciences*, 113(10):2636–2641.

[8] Fusek, R., Sojka, E., and Gaura, J. (2026). Driver anomaly detection using 3d human pose estimation. In *Computer Information Systems and Industrial Management*, page 3–14.

[9] Ghasemzadeh, S. A., Alahi, A., and De Vleeschouwer, C. (2025). Rumpl: Ray-based transformers for universal multi-view 2d to 3d human pose lifting. *arXiv preprint arXiv:2512.15488*.

[10] Gong, K., Zhang, J., and Feng, J. (2021). PoseAug: A differentiable pose augmentation framework for 3D human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8575–8584.

[11] Huo, R., Zhang, Y., Guo, Y., Ju, Z., and Gao, Q. (2023). GTFormer: 3D driver body pose estimation in video with graph convolution network and transformer. *IEEE Transactions on Intelligent Vehicles*, page 1–12.

[12] Ko, K.-L., Yoo, J.-S., Han, C.-W., Kim, J., and Jung, S.-W. (2024). Pose and shape estimation of humans in vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 25(1):402–416.

[13] Lea, C., Flynn, M. D., Vidal, R., Reiter, A., and Hager, G. D. (2017). Temporal convolutional networks for action segmentation and detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1003–1012.

[14] Li, W., Liu, H., Guo, T., Tang, H., and Ding, R. (2025). GraphMLP: A graph MLP-Like architecture for 3D human pose estimation. *Pattern Recognition*, 158:110925.

[15] Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., and Dollár, P. (2015). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision*, page 740–755.

[16] Liu, J., Rojas, J., Li, Y., Liang, Z., Guan, Y., Xi, N., and Zhu, H. (2021). A graph attention spatio-temporal convolutional network for 3d human pose estimation in video. In *IEEE International Conference on Robotics and Automation*, page 3374–3380.

[3] https://www.indie.inc/perception-software/

[17] Lutz, S., Blythman, R., Ghostal, K., Moynihan, M., Simms, C., and Smolic, A. (2022). JointFormer: Single-frame lifting transformer with error prediction and refinement for 3D human pose estimation. In *International Conference on Pattern Recognition*, pages 1156–1163.

[18] Martin, M., Roitberg, A., Haurilet, M., Horne, M., Reiß, S., Voit, M., and Stiefelhagen, R. (2019). Drive&Act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles. In *IEEE/CVF International Conference on Computer Vision*, pages 2801–2810.

[19] Martinez, J., Hossain, R., Romero, J., and Little, J. J. (2017). A simple yet effective baseline for 3D human pose estimation. In *IEEE/CVF International Conference on Computer Vision*, pages 2659–2668.

[20] Mazzia, V., Angarano, S., Salvetti, F., Angelini, F., and Chiaberge, M. (2022). Action Transformer: A self-attention model for short-time pose-based human action recognition. *Pattern Recognition*, 124:108487.

[21] MMPose Contributors (2020). OpenMMLab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose. Accessed: 2023-07-18.

[22] Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *International Conference on International Conference on Machine Learning*, pages 1310–1318.

[23] Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99.

[24] Sagmeister, D., Schörkhuber, D., Nezveda, M., Stiedl, F., Schimkowitsch, M., and Gelautz, M. (2023). Transfer learning for driver pose estimation from synthetic data. In *IEEE Intelligent Vehicles Symposium*, pages 1–7.

[25] Sun, K., Xiao, B., Liu, D., and Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5686–5696.

[26] The European Commission (2023). Supplementing regulation (eu) 2019/2144 of the european parliament and of the council. https://eur-lex.europa.eu/eli/reg/2019/2144/oj/eng. Accessed: 2025-04-30.

[27] Varol, G., Laptev, I., Schmid, C., and Zisserman, A. (2021). Synthetic humans for action recognition from unseen viewpoints. *International Journal of Computer Vision*, 129(7):2264–2287.

[28] Wang, J., Tan, S., Zhen, X., Xu, S., Zheng, F., He, Z., and Shao, L. (2021). Deep 3d human pose estimation: A review. *Computer Vision and Image Understanding*, 210:103225.

[29] Yao, Z., Liu, Y., Ji, Z., Sun, Q., Lasang, P., and Shen, S. (2019). 3D driver pose estimation based on joint 2D-3D network. In *IEEE International Conference on Image Processing*, page 2546–2550.

[30] Zhao, L., Peng, X., Tian, Y., Kapadia, M., and Metaxas, D. N. (2019). Semantic graph convolutional networks for 3D human pose regression. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3425–3435.

[31] Zhao, W., Wang, W., and Tian, Y. (2022). GraFormer: Graph-oriented transformer for 3D pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20406–20415.

[32] Zhao, Z., Yang, L., Sun, P., Hui, P., and Yao, A. (2025). Analyzing the synthetic-to-real domain gap in 3d hand pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12255–12265.

[33] Zheng, C., Wu, W., Chen, C., Yang, T., Zhu, S., Shen, J., Kehtarnavaz, N., and Shah, M. (2023). Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, 56(1):1–37.

[34] Zheng, C., Zhu, S., Mendieta, M., Yang, T., Chen, C., and Ding, Z. (2021). 3d human pose estimation with spatial and temporal transformers. In *IEEE International Conference on Computer Vision*, pages 11636–11645.

Organ Level Representation Learning for Region Based Medical Image Retrieval

Donghwan Lee
Department of Industrial Engineering
Yonsei University
Seoul, Republic of Korea
dhlee.ie@yonsei.ac.kr

Wooju Kim*
Department of Industrial Engineering
Yonsei University
Seoul, Republic of Korea
wkim@yonsei.ac.kr

Abstract

As medical image databases expand, precise Content-Based Medical Image Retrieval (CBMIR) techniques are increasingly required to support case-based reasoning, clinical education, and data-driven decision-making. Recent deep learning–based CBMIR approaches typically rely on global embeddings to enhance retrieval performance. However, such image-level representations often dilute localized anatomical features and fail to capture clinically relevant organ-specific details. To address this limitation, we propose a region-based CBMIR framework that integrates organ-level information into both representation learning and retrieval. The ROI Embedding Selector extracts patch-level embeddings from user-specified regions of interest (ROIs). The Region-aware Organ Attention (ROA) module then learns structured organ representations through cross-attention between image patches and dedicated organ tokens. During inference, a visibility-weighted aggregation strategy guided by Organ Visibility Recognition incorporates query-relevant organs, enabling anatomically targeted and clinically meaningful retrieval. Experiments on the TotalSegmentator dataset demonstrate that the proposed framework consistently outperforms global embedding–based vision foundation models, particularly in region query settings.

1 Introduction

Medical images play an important role in clinical practice, including diagnosis, treatment planning, and prognosis prediction. Advances in imaging modalities such as computed tomography (CT), positron emission tomography (PET), magnetic resonance imaging (MRI), and X-ray imaging have enabled more precise assessment of disease severity and progression. These developments contribute to improved patient outcomes and reduced healthcare costs, and provide essential evidence for clinical decision-making (1; 2). Meanwhile, the digitization of healthcare systems and ongoing improvements in imaging devices have accelerated the large-scale accumulation of medical data (3), increasing the need for more precise and reliable methods to organize, manage, and utilize these resources (4).

Medical image retrieval (MIR) has been studied as a technology to support case-based reasoning and clinical education, and to assist disease prediction by artificial intelligence models (5). In particular, Content-Based Medical Image Retrieval (CBMIR), which retrieves images based on visual similarity, has received attention as an approach for large-scale medical image databases. CBMIR enables clinicians to identify visually similar prior cases, supporting clinical decision-making (6).

Over the past decade, advances in deep learning–based representation learning have significantly improved the performance of medical image retrieval. For example, CNN and Vision Transformer–based

*Corresponding author

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

models have achieved substantial performance gains by learning high-level semantic representations (7; 8; 9; 10). Nevertheless, most existing CBMIR methods still rely on global image-level representations, which do not explicitly capture clinically meaningful anatomical structures. As a result, they remain limited in performing fine-grained, region-specific retrieval.

In medical images, diagnosis depends on the anatomical context of lesions and their associated organs. Therefore, global representations alone are insufficient to identify clinically relevant similar cases. Even when a Region of Interest (ROI) is available, global embedding–based methods have limited ability to emphasize local regions or to perform organ-aware, weighted retrieval. To address these limitations, we propose a region-aware, organ-centric CBMIR framework that incorporates anatomical structure into representation learning and retrieval. First, the ROI Embedding Selector extracts patch embeddings corresponding to user-specified regions to enhance local feature representation. Next, the Region-aware Organ Attention (ROA) module refines interactions between image patches and organ tokens to learn structured organ-level representations. During inference, we introduce an Organ Visibility Recognition–based visibility-weighted aggregation strategy to emphasize query-relevant organs and support region-focused retrieval. Through the use of organ-aware representation learning and a visibility-weighted aggregation strategy, the proposed framework mitigates the limitations of global embeddings and performs region-based CBMIR. In addition, experiments on the TotalSegmentator dataset demonstrate that the proposed framework achieves competitive performance in region-based retrieval compared with global embedding–based methods. The main contributions of this work are summarized as follows:

- We propose a region-aware, organ-centric CBMIR framework that integrates anatomical structure into representation learning and retrieval.
- We design an organ-aware representation learning scheme with a visibility-weighted aggregation strategy to enable precise region-based retrieval.
- We validate the proposed framework on the TotalSegmentator dataset and demonstrate strong performance in region-based retrieval compared with global embedding–based methods.

2 Related Work

2.1 Content-Based Medical Image Retrieval

Content-Based Medical Image Retrieval (CBMIR) extracts feature representations from medical images and retrieves similar cases based on feature similarity, where representation quality directly affects retrieval performance. Early methods relied on hand-crafted features and similarity computation in feature space (11; 12). With the advancement of deep learning, convolutional neural networks (CNNs) (13) enabled hierarchical semantic representation learning and improved retrieval performance. Recent studies adopt CNN-based foundation models such as Inception V3 (14), VGG19 (15), and ResNet (16) for feature extraction (7; 8; 9). Lo et al. (2025) (17) further introduced a multi-level CNN framework to model modality–organ–disease relationships. More recently, Transformer-based models (10) have been applied to capture global contextual relationships. Vision Transformers with contrastive learning (18) and attention-based fusion frameworks such as DaRF (5) have been proposed to enhance representation learning. In addition, pretrained vision foundation models, including BiomedCLIP (19), have demonstrated strong retrieval performance without extensive task-specific fine-tuning. Despite these advances, most CBMIR methods rely primarily on global image representations and do not explicitly model anatomical structures for region-specific retrieval.

2.2 Representation Learning using Query Token

Following Vaswani et al. (2017) (10), an increasing number of studies have employed learnable tokens or queries to encode information from data distributions into structured token representations. DETR (20) introduced learnable object queries for object detection, enabling the decoder to perform set prediction by decoding objects from image features through cross-attention. Perceiver (21) and Perceiver IO (22) utilized learnable latent tokens as a bottleneck to project large-scale multimodal inputs into fixed-size latent representations and extended to diverse output formats through query-based readout. VCT (23) learned disentangled concept tokens from image tokens for scene decomposition and representation disentanglement. BoQ (24) employed learnable global queries to aggregate place-specific attributes into a unified representation, improving visual place recognition and retrieval

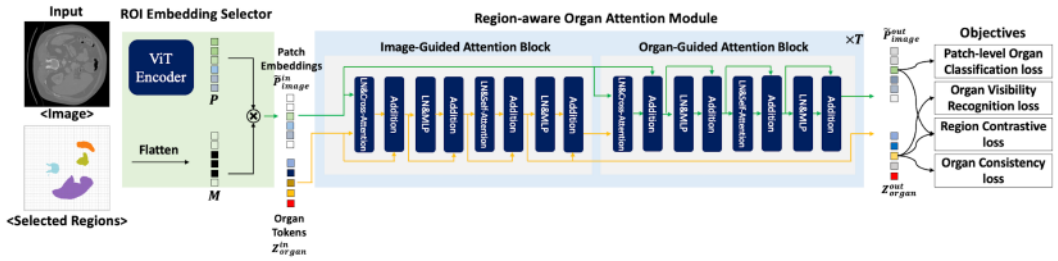


Figure 1: Overall framework of the proposed region-aware organ representation learning framework.

performance. In the medical domain, OWT (25) introduced organ-wise tokenization, decomposing holistic representations into organ-specific token groups and enabling organ-level representation learning through selective token group combinations. However, existing token-based approaches do not explicitly model anatomically meaningful structures or address region-specific retrieval in medical images.

3 Methodology

3.1 ROI Embedding Selector

The ROI Embedding Selector applies a binary mask to retain patch embeddings corresponding to a user-specified region of interest (ROI), ensuring anatomically localized feature extraction (Figure 1). Given patch embeddings $P = p_1, \ldots, p_N$ extracted by a Vision Transformer (ViT) encoder, a binary mask $M \in 0, 1^N$ is defined such that $M_i = 1$ if the i-th patch belongs to the ROI and $M_i = 0$ otherwise. The masked embeddings are computed as $\tilde{p}_i = M_i \cdot p_i$, allowing only ROI patches to contribute to subsequent representation learning. During training, organ-level ROIs are constructed from ground-truth segmentation masks with random region selection to prevent overfitting to specific anatomical structures and promote generalization. During inference, users can manually specify the ROI or select from pre-segmented regions, enabling interactive and region-focused retrieval.

3.2 Region-aware Organ Attention Module

The Region-aware Organ Attention (ROA) module models organ regions in medical images and learns organ-level representations that reflect structural and semantic characteristics. As illustrated in Figure 1, the module takes the masked patch embeddings $\tilde{P}^{\text{in}}_{\text{image}} \in \mathbb{R}^{N \times d}$ obtained from the ROI Embedding Selector and predefined organ token embeddings $Z^{\text{in}}_{\text{organ}} \in \mathbb{R}^{C \times d}$ as inputs. The ROA consists of an Image-Guided Attention (IGA) block and an Organ-Guided Attention (OGA) block, which are alternately applied for T iterations to iteratively refine image and organ representations. Through this interaction, organ tokens aggregate structural cues from masked patch embeddings, while patch representations incorporate organ-level semantic context.

In the IGA block, cross-attention is performed with organ tokens as queries and image patch embeddings as keys and values. A Bin Token is introduced to absorb patch information that does not clearly correspond to a specific organ, reducing interference among organ tokens. Self-attention is then applied among organ tokens to model inter-organ relationships and maintain contextual consistency. In the OGA block, cross-attention is performed with image patch embeddings as queries and organ tokens as keys and values. A Background Token is introduced to separately model background information and stabilize organ-centric representation learning. Self-attention is subsequently applied to image patch embeddings to refine inter-patch relationships and improve structural coherence of the learned representations.

3.3 Objectives

We define a unified training objective consisting of four complementary loss terms for organ-aware representation learning, jointly optimizing spatial accuracy, organ visibility modeling, representation discriminability, and semantic stability.

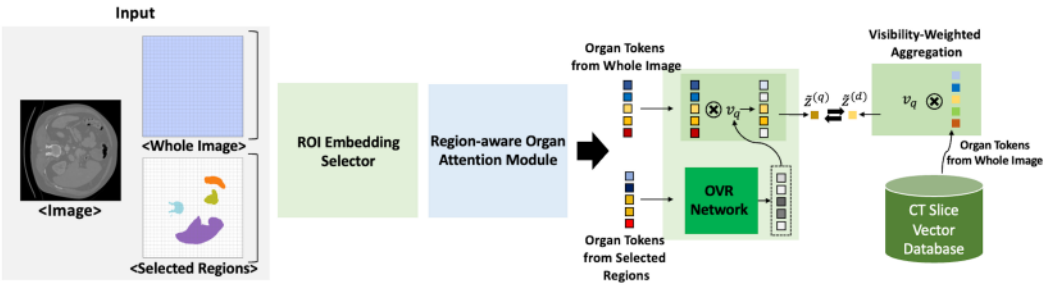


Figure 2: Region-based organ-level retrieval pipeline. Organ tokens are precomputed for database images and matched with query representations based on user-specified regions.

Patch Classification Loss. To provide fine-grained spatial supervision, we introduce a Patch-Level Organ Classification (POC) network that consists of a two-layer MLP. Given the output patch embedding $\tilde{p}_i^{\text{out}}$, the predicted organ label is computed as $\hat{y}_i = \text{POC}(\tilde{p}_i^{\text{out}})$. The supervision is formulated using cross-entropy, defined as $\mathcal{L}_{\text{PC}} = \frac{1}{N}\sum_{i=1}^{N} \text{CE}(y_i, \hat{y}_i)$.

Organ Visibility Loss. To explicitly model organ presence, we introduce an Organ Visibility Recognition (OVR) network, which consists of a two-layer MLP. Given the refined organ token z_c^{out}, the predicted visibility score is computed as $\hat{v}_c = \text{OVR}(z_c^{\text{out}})$. The visibility supervision is formulated using binary cross-entropy, defined as $\mathcal{L}_{\text{OV}} = \frac{1}{C}\sum_{c=1}^{C} \text{BCE}(v_c, \hat{v}_c)$. As shown in Figure 2, the OVR network is used during the image retrieval process.

Region Contrastive Loss. To align organ token embeddings with their corresponding region representations, we introduce a region-level contrastive learning objective. For each organ c, the region embedding r_c is obtained by average pooling the refined patch embeddings $\tilde{p}_i^{\text{out}}$ within the corresponding organ region. The contrastive objective is defined using an InfoNCE-based loss as $\mathcal{L}_{\text{RC}} = -\frac{1}{C}\sum_{c=1}^{C} \log \frac{\exp(\text{sim}(z_c^{\text{out}}, r_c)/\tau)}{\sum_{k=1}^{K} \exp(\text{sim}(z_c^{\text{out}}, r_k)/\tau)}$, where $\{r_k\}_{k=1}^{K}$ denotes region embeddings sampled from all organs within the current mini-batch.

Organ Consistency Loss. To prevent unnecessary updates when relevant evidence is limited, we regularize refined organ embeddings toward their initial organ queries. This keeps unsupported tokens close to their initial states. Given the refined organ embedding z_c^{out} and the corresponding initial query z_c^{in}, the consistency supervision is formulated using an ℓ_2 regression loss. To prevent gradients from flowing into the initial queries, we apply a stop-gradient operator $\text{sg}(\cdot)$, and define the loss as $\mathcal{L}_{\text{OC}} = \frac{1}{C}\sum_{c=1}^{C} \|z_c^{\text{out}} - \text{sg}(z_c^{\text{in}})\|_2^2$.

The overall objective is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{PC}} + \lambda_2 \mathcal{L}_{\text{OV}} + \lambda_3 \mathcal{L}_{\text{RC}} + \lambda_4 \mathcal{L}_{\text{OC}}. \tag{1}$$

In all experiments, we set all $\lambda_i = 1$, assigning equal importance to each loss term. This configuration yielded stable optimization without additional hyperparameter tuning.

3.4 Region-based Organ-level Retrieval

We propose a region-based organ-level retrieval framework that operates on precomputed organ token representations, as illustrated in Figure 2. During database construction, all predefined regions are selected by the ROI Embedding Selector, and the trained ROA module generates organ token embeddings for each database image. For a database image d, we obtain an organ-wise representation set $Z_d = \{z_{d,c}^{\text{out}}\}_{c=1}^{C}, \quad z_{d,c}^{\text{out}} \in \mathbb{R}^D$. These representations are stored in a vector database and used for similarity computation. For a query image, the same feature extraction procedure is applied to obtain $Z_q = \{z_{q,c}^{\text{out}}\}_{c=1}^{C}$. When a region of interest (ROI) is specified, the ROI Embedding Selector extracts masked patch embeddings from the selected region, which are processed by the ROA module to produce organ-wise representations. To model organ presence, we introduce an Organ Visibility Recognition (OVR) network during training. The OVR network estimates the visibility probability of

each organ, $\hat{v}_c \in [0, 1]$, and these probabilities are used to construct a visibility-weighted aggregated representation:

$$\tilde{z}_x = \frac{1}{C} \sum_{c=1}^{C} \hat{v}_c z_{x,c}^{\text{out}}, \tag{2}$$

where $x \in \{q, d\}$. The visibility weights are predicted from the query image and applied to both the query and database representations.

The similarity between a query image I_q and a database image I_d is computed as the cosine similarity between their corresponding visibility-weighted representations, $\tilde{z}_q$ and $\tilde{z}_d$. Database images are ranked in descending order of this similarity score. Compared with global embedding-based retrieval, this query-driven weighting scheme focuses on anatomically observable organs and supports region-specific retrieval.

4 Experiments

4.1 Datasets

We construct the training and evaluation datasets using the TotalSegmentator (TS) dataset (Wasserthal et al., 2023, Version 2). The dataset consists of 1,228 volumes and 317,863 slices in total. The data split follows the official train/test partition defined in Vista3D (26). The training set includes 980 volumes (252,468 slices), and the test set comprises 248 volumes (65,395 slices). For retrieval evaluation, we build a slice-level dataset. The database set is created by uniformly sampling slices from the training volumes at intervals of 10 slices, in order to control slice density and reduce redundancy. This results in 25,696 database slices. The query set is constructed by randomly selecting five slices from each test volume, yielding a total of 1,241 query slices. To analyze the effect of region configuration on retrieval performance, we define four query region types: (1) a single randomly selected region, (2) two randomly selected regions, (3) three randomly selected regions, and (4) whole-image queries. All query types use the same number of queries (1,241) and the same database set (25,696 slices), enabling comparison under consistent evaluation conditions.

4.2 Evaluation metric

We adopt a region-centric evaluation protocol inspired by Jush et al. (27), which evaluates retrieval performance for each anatomical region independently. Given our slice-level retrieval setting, we adopt organ-level Precision@K to incorporate ranking information. This metric measures the proportion of retrieved slices within the top-K results that contain the target organ. For a given organ i, Precision$_i$@K is defined as

$$\text{Precision}_i@K = \frac{1}{|Q_i|} \sum_{q \in Q_i} \frac{|\{s \in R_q@K \mid i \in \mathcal{L}(s)\}|}{K}, \tag{3}$$

where Q_i denotes the set of query slices containing organ i, $R_q@K$ represents the set of top-K retrieved slices for query q, s denotes a retrieved slice, and $\mathcal{L}(s)$ is the set of organ labels annotated for slice s. The notation $|\cdot|$ indicates set cardinality. Note that Precision$_i$@K is not necessarily monotonically increasing with respect to K.

The overall performance is computed by averaging across all N evaluated organs:

$$\text{Precision}@K = \frac{1}{N} \sum_{i=1}^{N} \text{Precision}_i@K. \tag{4}$$

4.3 Comparison with Baselines

Recent studies (9; 28) have shown that pretrained Transformer-based vision foundation models trained with self-supervised learning achieve competitive performance in medical image retrieval.

Table 1: Retrieval performance (P@K) under different region configurations.

Region Configuration	Method	P@1	P@3	P@5	P@10	P@20
1 Region	Random	0.169	0.162	0.161	0.156	0.160
	ViT	0.150	0.163	0.162	0.164	0.164
	DINOv3	0.158	0.150	0.148	0.151	0.149
	DreamSim	0.117	0.113	0.108	0.109	0.110
	Proposed Model	**0.943**	**0.942**	**0.944**	**0.945**	**0.942**
2 Regions	Random	0.169	0.161	0.161	0.158	0.159
	ViT	0.336	0.323	0.314	0.314	0.308
	DINOv3	0.307	0.305	0.303	0.298	0.295
	DreamSim	0.347	0.342	0.341	0.331	0.320
	Proposed Model	**0.935**	**0.931**	**0.932**	**0.930**	**0.928**
3 Regions	Random	0.165	0.161	0.160	0.160	0.158
	ViT	0.425	0.417	0.408	0.398	0.388
	DINOv3	0.413	0.406	0.398	0.391	0.380
	DreamSim	0.489	0.475	0.462	0.450	0.434
	Proposed Model	**0.954**	**0.953**	**0.951**	**0.947**	**0.943**
Whole Image	Random	0.165	0.158	0.159	0.158	0.160
	ViT	0.704	0.690	0.678	0.663	0.645
	DINOv3	0.715	0.699	0.688	0.672	0.653
	DreamSim	0.785	0.773	0.764	0.751	0.740
	Proposed Model	**0.891**	**0.885**	**0.881**	**0.874**	**0.865**

Motivated by these findings, we adopt the pretrained baselines used in Jush et al. (28), including Vision Transformer (ViT) (29), DINOv3 (30), and DreamSim (31). For ROI-based retrieval, the target organ is cropped using a bounding box, and region embeddings are extracted and compared with global database embeddings to compute similarity scores. We also include a random ranking baseline without learned features. This setup enables quantitative evaluation of the limitations of global-embedding-based retrieval and assessment of the proposed region-based strategy.

Table 1 reports retrieval performance under varying region configurations. The proposed model achieves the highest scores across all settings, outperforming ViT, DINOv3, DreamSim, and the random baseline across all configurations. The performance gap is particularly pronounced in single- and multi-region configurations, indicating that organ-wise representations are effective for region-focused retrieval. Although global embedding methods improve as more regions are selected, their performance remains lower than that of the proposed model. In the Whole Image setting, the performance gap decreases, yet our method still yields the best results. Overall, organ-level representation learning with visibility-weighted aggregation consistently enhances region-focused retrieval performance.

Figure 3 provides a qualitative comparison under different region configurations. In the Whole Image setting (Fig. 3(a)), retrieval is based on the entire slice, leading to selection driven by global abdominal appearance; consequently, localized structures such as the colon are not consistently preserved among top results. In contrast, the Single Region setting (Fig. 3(b)) uses the colon as the ROI, resulting in more consistent localization and morphology alignment with the query region. In contrast, baseline methods relying on simple bounding-box cropping exhibit less stable structural alignment due to limited contextual information. DINOv3 and DreamSim occasionally retrieve magnified regions, while ViT retrieves colon-containing slices, but with less consistent localization and morphological alignment compared to the proposed method.

4.4 Ablation Study

Table 2 presents the ablation results analyzing the contribution of each training component in the proposed model. The baseline configuration (RC+VR) achieves stable performance across region settings. When Patch-level Classification (PC) is introduced, consistent improvements are observed under all configurations. This indicates that patch-level supervision contributes to more spatially

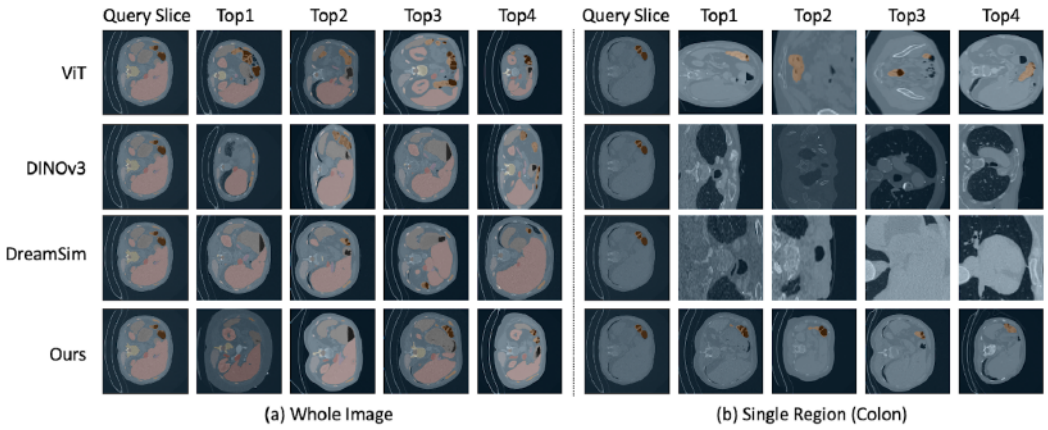


Figure 3: Qualitative comparison of retrieval results under different region configurations. The organ of interest in the query slice is highlighted with a colored segmentation mask. In the retrieved images, the corresponding organ is marked with the same color to facilitate visual comparison.

Table 2: Ablation results under different region configurations (Precision@K).

Region Configuration	Setting	P@1	P@3	P@5	P@10	P@20
1 Region	RC+VR	0.928	0.925	0.923	0.923	**0.983**
	RC+VR+PC	**0.946**	**0.947**	**0.948**	**0.946**	0.943
	RC+VR+PC+OC	0.943	0.942	0.944	0.945	0.942
2 Regions	RC+VR	0.915	0.909	0.904	0.902	0.898
	RC+VR+PC	**0.936**	0.929	0.928	0.924	0.921
	RC+VR+PC+OC	0.935	**0.931**	**0.932**	**0.930**	**0.928**
3 Regions	RC+VR	0.933	0.930	0.926	0.921	0.911
	RC+VR+PC	0.947	0.948	0.947	0.943	0.937
	RC+VR+PC+OC	**0.954**	**0.953**	**0.951**	**0.947**	**0.943**
Whole Image	RC+VR	0.876	0.869	0.863	0.854	0.845
	RC+VR+PC	0.889	0.885	**0.990**	0.874	**0.866**
	RC+VR+PC+OC	**0.891**	**0.885**	0.881	**0.874**	0.865

precise organ token representations, leading to improved region-based retrieval performance. With the addition of Organ Consistency (OC), the model achieves the highest performance under multi-region settings. In particular, under the 3 Region configuration, the model attains the highest P@1 score of 0.954. This result shows that enforcing consistency among organ token representations becomes increasingly beneficial when multiple organs are jointly considered. In contrast, under the Whole Image setting, the performance gains from OC are relatively modest, indicating that retrieval based on global input depends more heavily on the base representation learned by RC and VR. Overall, RC and VR establish the core representation, while PC and OC progressively refine it, resulting in incremental improvements in organ-centric retrieval performance. Although the relative contribution of each component may vary across region configurations, these variations reflect inherent differences between retrieval scenarios rather than instability in the model. Nevertheless, the full configuration (RC+VR+PC+OC) consistently achieves the best or near-best performance across all settings, indicating that a single unified configuration can be effectively adopted without query-dependent tuning.

4.5 Visualization and Interpretability Analysis

In this study, we adopt the pairwise similarity visualization method proposed by Black et al. (2022) (32) to analyze the spatial alignment between query and retrieved images. Figure 4 presents pairwise similarity maps between the query slice and the top-ranked retrieval results (top1–top5),

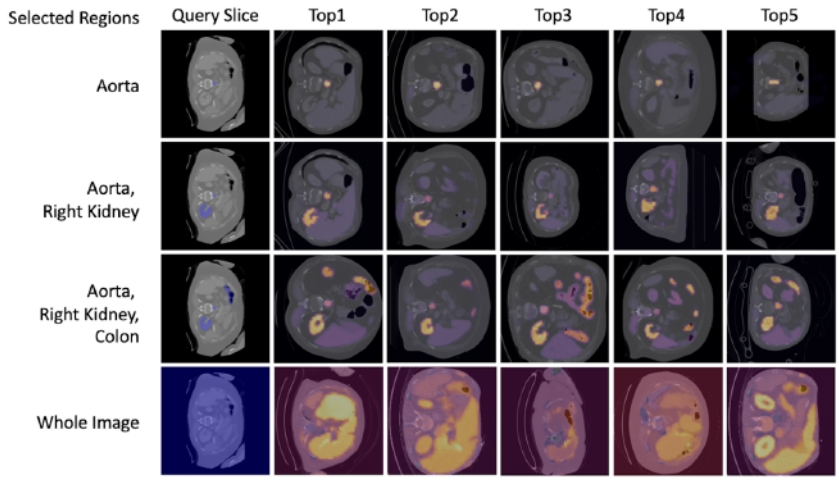


Figure 4: Pairwise rollout-based similarity maps under different region configurations. Each row shows the query slice and the top-ranked retrieval results (top1–top5). Regions contributing strongly to the similarity score are overlaid using a heatmap.

with each row corresponding to a different region configuration. When a single organ is selected as the query region, the similarity response is concentrated on the corresponding organ area in the retrieved images. When multiple organs are selected, the similarity maps jointly reflect the selected organs. As the number of selected regions increases, adjacent organ structures are progressively reflected in the similarity response. In contrast, under the whole image setting, retrieval is based on the entire slice, and similarity responses are primarily driven by global morphological resemblance. the highlighted regions are less consistent across retrieved images in terms of the areas referenced by the model, with relatively large organs such as the liver and kidneys showing stronger responses.

4.6 Limitations and Future works

This study demonstrates the potential of organ-level representation learning for region-focused medical image retrieval; however, several limitations remain. First, the proposed model is trained using precise organ segmentation annotations from the TotalSegmentator dataset, which assumes the availability of large-scale ground-truth segmentation labels. Second, the evaluation is primarily conducted at the slice level and thus does not fully capture organ relationships or spatial continuity at the 3D volume level. Third, comparative experiments are limited to ViT-based vision foundation models, and further validation across a broader range of retrieval models is necessary. Future work will include robustness evaluation under settings that rely on automatic segmentation models, such as Vista3D (26), and modeling organ relationships at the 3D volume level. In addition, we will analyze the impact of model parameters within the proposed architecture and perform systematic validation across diverse retrieval architectures to comprehensively evaluate the scalability and clinical applicability of the proposed approach.

4.7 Conclusions

In this study, we propose an organ-aware, region-centric representation learning framework to address the limitations of global similarity–based retrieval. The ROI Embedding Selector filters patch embeddings corresponding to the region of interest, while the Region-aware Organ Attention (ROA) module learns interactions between image patches and organ tokens to construct representations that precisely capture organ-specific structural and semantic information. In addition, during inference, we introduce a visibility-weighted aggregation strategy based on Organ Visibility Recognition, which prioritizes anatomically observable organs and enables retrieval focused on clinically meaningful regions of interest. By moving beyond simple global similarity computation and explicitly incorporating anatomical context, the proposed method supports more precise retrieval. This framework offers both methodological significance and practical potential for precise case retrieval and clinical decision support in real-world settings.

Acknowledgments

This work was supported by the Technology Innovation Program(or Industrial Strategic Technology Development Program-ATC+)(20023280, Development of life cycle management platform by providing artificial intelligence-based real-time model observability and explanability) funded By the Ministry of Trade, Industry and Resources(MOTIR, Korea)

References

[1] L. Cui and M. Liu, "An intelligent deep hash coding network for content-based medical image retrieval for healthcare applications," *Egyptian Informatics Journal*, vol. 27, p. 100499, 2024.

[2] J. Wang, H. Zhu, S.-H. Wang, and Y.-D. Zhang, "A review of deep learning on medical image analysis," *Mobile Networks and Applications*, vol. 26, no. 1, pp. 351–380, 2021.

[3] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing medical imaging data for machine learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[4] J. Choe, H. J. Hwang, J. B. Seo, S. M. Lee, J. Yun, M. J. Kim, J. Jeong, Y. Lee, K. Jin, R. Park, J. Kim, H. Jeon, N. Kim, J. Yi, D. Yu, and B. Kim, "Content-based image retrieval by using deep learning for interstitial lung disease diagnosis with chest ct," *Radiology*, vol. 302, no. 1, pp. 187–197, 2022.

[5] Y. Nan, H. Zhou, X. Xing, G. Papanastasiou, L. Zhu, Z. Gao, A. F. Frangi, and G. Yang, "Revisiting medical image retrieval via knowledge consolidation," *Medical Image Analysis*, vol. 102, p. 103553, 2025.

[6] S. Agrawal, A. Chowdhary, S. Agarwala, V. Mayya, and S. Kamath S, "Content-based medical image retrieval system for lung diseases using deep cnns," *International Journal of Information Technology*, vol. 14, no. 7, pp. 3619–3627, 2022.

[7] R. Shetty, V. S. Bhat, and J. Pujari, "Content-based medical image retrieval using deep learning-based features and hybrid meta-heuristic optimization," *Biomedical Signal Processing and Control*, vol. 92, p. 106069, 2024.

[8] I. Issaoui, M. A. Alohali, W. Mtouaa, F. A. Alotaibi, A. Mahmud, and M. Assiri, "Archimedes optimization algorithm with deep learning assisted content-based image retrieval in healthcare sector," *IEEE Access*, vol. 12, pp. 29768–29777, 2024.

[9] A. Mahbod, N. Saeidi, S. Hatamikia, and R. Woitek, "Evaluating pre-trained convolutional neural networks and foundation models as feature extractors for content-based medical image retrieval," *Engineering Applications of Artificial Intelligence*, vol. 150, p. 110571, 2025.

[10] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017.

[11] P. Das and A. Neelima, "An overview of approaches for content-based medical image retrieval," *International Journal of Multimedia Information Retrieval*, vol. 6, no. 4, pp. 271–280, 2017.

[12] A. Kumar, J. Kim, W. Cai, M. Fulham, and D. Feng, "Content-based medical image retrieval: A survey of applications to multidimensional and multimodality data," *Journal of Digital Imaging*, vol. 26, no. 6, pp. 1025–1039, 2013.

[13] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[14] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2015.

[15] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

[17] C. M. Lo and C. Y. Hsieh, "Large-scale hierarchical medical image retrieval based on a multilevel convolutional neural network," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 9, no. 4, pp. 2782–2792, 2025.

[18] A. S. Susmitha and V. P. Namboodiri, "Analysis of transformers for medical image retrieval," 2024.

[19] S. Denner, D. Zimmerer, D. Bounias, M. Bujotzek, S. Xiao, R. Stock, L. Kausch, P. Schader, T. Penzkofer, P. F. Jäger, and K. Maier-Hein, "Leveraging foundation models for content-based image retrieval in radiology," *Computers in Biology and Medicine*, vol. 196, p. 110640, 2025.

[20] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020.

[21] A. Jaegle, F. Gimeno, A. Brock, O. Vinyals, A. Zisserman, and J. Carreira, "Perceiver: General perception with iterative attention," in *The 38th International Conference on Machine Learning (ICML)*, pp. 4651–4664, PMLR, 2021.

[22] A. Jaegle, S. Borgeaud, J. B. Alayrac, C. Doersch, C. Ionescu, D. Ding, S. Koppula, D. Zoran, A. Brock, E. Shelhamer, O. J. Henaff, M. Botvinick, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver io: A general architecture for structured inputs and outputs," *The 10th International Conference on Learning Representations (ICLR)*, 2022.

[23] T. Yang, Y. Wang, Y. Lu, and N. Zheng, "Visual concepts tokenization," in *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, p. Article 2289, Curran Associates Inc., 2022.

[24] A. Ali-bey, B. Chaib-draa, and P. Giguère, "Boq: A place is worth a bag of learnable queries," in *The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 17794–17803, 2024.

[25] S. Song, S. Yoon, P. Jin, S. Kim, M. Tivnan, Y. Oh, R. Meng, L. Chen, Z. Lyu, and D. Wu, "Owt: A foundational organ-wise tokenization framework for medical imaging," *arXiv preprint arXiv:2505.04899*, 2025.

[26] Y. He, P. Guo, Y. Tang, A. Myronenko, V. Nath, Z. Xu, D. Yang, C. Zhao, B. Simon, and M. Belue, "Vista3d: A unified segmentation foundation model for 3d medical imaging," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 20863–20873, 2024.

[27] F. Khun Jush, S. Vogler, and M. Lenga, "Content-based 3d image retrieval and a colbert-inspired re-ranking for tumor flagging and staging," *Journal of Imaging Informatics in Medicine*, 2025.

[28] F. K. Jush, S. Vogler, T. Truong, and M. Lenga, "Content-based image retrieval for multi-class volumetric radiology images: A benchmark study," *IEEE Access*, 2025.

[29] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[30] O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, *et al.*, "Dinov3," *arXiv preprint arXiv:2508.10104*, 2025.

[31] S. Fu, N. Y. Tamir, S. Sundaram, L. Chai, R. Zhang, T. Dekel, and P. Isola, "Dreamsim: learning new dimensions of human visual similarity using synthetic data," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, (Red Hook, NY, USA), Curran Associates Inc., 2023.

[32] S. Black, A. Stylianou, R. Pless, and R. Souvenir, "Visualizing paired image similarity in transformer networks," in *The IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1534–1543, 2022.

Diffusion Edge Detection Of Texture-less Objects

Matvey Ivanov
Automation and Control Institute (ACIN)
TU Wien
Wien, 1040
e11775774@student.tuwien.ac.at

Peter Hönig
Automation and Control Institute (ACIN)
TU Wien
Wien, 1040
hoenig@acin.tuwien.ac.at

Markus Vinzce
Automation and Control Institute (ACIN)
TU Wien
Wien, 1040
vincze@acin.tuwien.ac.at

Abstract

Edge detection of complex, smooth, transparent, reflective and texture-less objects is an unsolved problem in computer vision. In this work, an existing approach using diffusion in the image space is adapted to enable fast and accurate edge detection. The method is applied to texture-less industrial objects from the T-LESS and XYZIBD datasets. The models are trained on datasets, generated synthetically using BlenderProc. Three training datasets are created using T-LESS objects to evaluate the impact of edge type and object texturing on prediction quality. Two more datasets are generated using XYZIBD objects to investigate the influence of the crease angle used in edge rendering. The diffusion models are evaluated using the NMSE, SSIM, DICE, and CRISP metrics, to assess accuracy, structural fidelity, and perceptual sharpness. Experiments show that our approach achieves competitive edge prediction quality and consistently outperforms existing diffusion based methods in computational efficiency at a lower resolution, while offering overall better prediction fidelity compared to the Canny edge detector. With a runtime of 95ms per image on an NVIDIA RTX3090, the approach demonstrates suitability for deployment in robotic vision systems. A quantitative edge prediction quality evaluation is conducted on real-world test sets which are extended with the edge ground-truth.

1 Introduction

Edge detection in images remains a fundamental problem in computer vision [Huang and Huang, 2025], forming the basis for a wide range of tasks such as object detection, segmentation and scene understanding. Traditional techniques, such as Sobel [Kittler, 1983] and Canny [Canny, 1986], have been studied extensively [Kanopoulos et al., 1988], [Luo and Duraiswami, 2008], [Wang et al., 2021], but despite their efficiency, these methods are limited by their use of gradient-based features, which are sensitive to noise and variations in scene luminosity. These methods fall short in open-world robotics, where lightning conditions change constantly [Pulli et al., 2024].

To overcome the limitations of classical edge detectors, machine learning-based methods are introduced, as surveyed by [Hu, 2025]. Among these, diffusion models [Sohl-Dickstein et al., 2015] produce perceptually coherent and semantically rich edge maps by iteratively refining structural information, even under challenging conditions such as noise, occlusion, or low contrast. However, this denoising process introduces substantial computational overhead, hindering real-time applica-

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

bility. A recent approach by [Ye et al., 2024] employs a diffusion network with a U-Net [Ronneberger et al., 2015] backbone to achieve high precision, yet inference time remains a significant bottleneck.

Learning-based methods depend on reliable ground-truth annotations. In 2D imagery however, the absence of depth information often causes annotated edges to deviate from their true 3D locations. This limitation is mitigated by deriving edge ground truth directly from the corresponding 3D meshes. Numerous open-source 3D model repositories support such workflows. Datasets such as [Hodan et al., 2017], [Drost et al., 2017], [Kaskman et al., 2019], and [Xiang et al., 2018] primarily target industrial object scenarios, whereas [Morrison et al., 2020] provides a diverse set of geometries specifically designed for evaluating robotic grasping and manipulation.

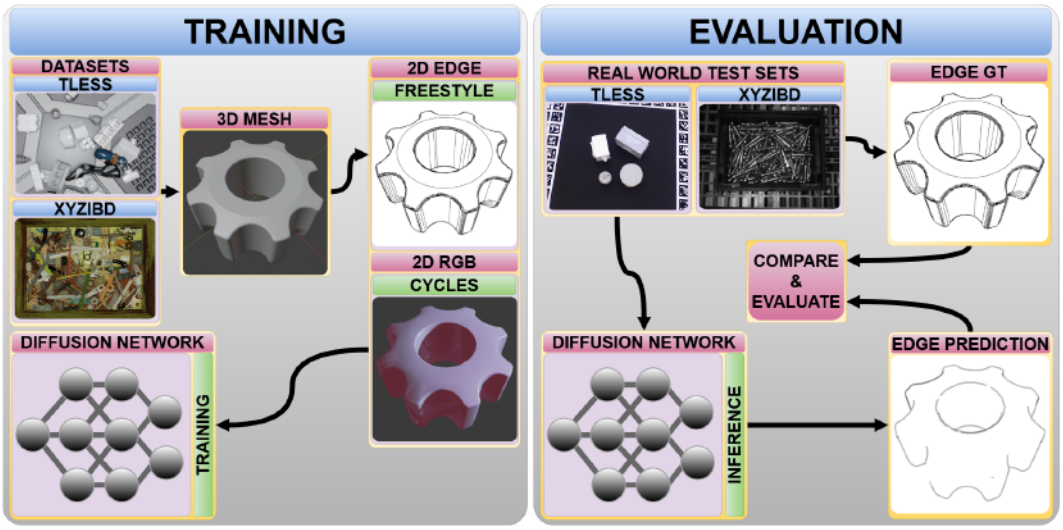


Figure 1: Complete Training-Evaluation Pipeline Block Diagram

This work presents a training and evaluation pipeline (see Figure 1) for diffusion based edge detection that maintains prediction quality while significantly reducing inference time. The objective is to facilitate the deployment of fast diffusion-based edge detectors on resource-constrained robotic platforms, where computational efficiency and reaction speed are critical.

We summarize our contributions as follows. We:

1. Extend the Blenderproc [Denninger et al., 2020] pipeline to allow precise, procedural edge rendering based on 3D-meshes and train a diffusion network to detect object edges in 2D, based on RGB images
2. Empirically prove that texture randomization improves prediction results [Hönig et al., 2025], when added to conventional domain randomization [Tobin et al., 2017]

2 Related Work

To highlight the constraints of existing methods related to edge detection, modern representative approaches and their trade-offs are reviewed. Works such as [Yu et al., 2017] and [Liu et al., 2017] integrate multi scale feature representations and semantic awareness to detect edges with high precision. Their performance is particularly strong in natural images. However, their reliance on deep backbones leads to high computational costs and long inference times. As a result, these methods are less suitable for resource constrained robotic systems [Ren et al., 2023]. A method targeting pose estimation refinement for transparent objects in laboratory settings via detection silhouette is proposed in [Weibel et al., 2026]. The approach remains constrained by the fidelity and confidence of silhouette extraction and by limited texture cues. It can be enhanced by training with large-scale synthetic datasets that provide precise per-object edge ground truth. Such datasets supply robust edge and silhouette supervision to support pose refinement. In addition, the fully synthetic TGF-Net

dataset for transparent objects [Yu et al., 2023] would similarly benefit from the inclusion of edge annotations, thereby strengthening its geometric supervision and improving downstream 6D pose estimation.

A key inspiration for this work is DiffusionEdge [Ye et al., 2024], which uses a diffusion probabilistic model with adaptive Fourier filters and a U-Net backbone to produce accurate edges in complex scenes, without requiring post-processing. Its effectiveness has been demonstrated in tasks such as 6D pose estimation of metallic objects [Leimeister, 2025]. However, the method suffers from long inference times on the order of magnitude of multiple seconds, making it impractical for real-time applications.

To address the run-time limitations imposed by prior methods, while retaining their robust edge detection performance in domain-specific contexts without reliance on manual annotation, the existing diffusion model architecture is trained on a set of synthetically generated datasets at a lower input resolution.

3 Synthetic Dataset Generation

Training a model to perform edge detection on real-world images requires the network to generalize well from its training set. To achieve this, domain randomization is used in all generated synthetic scenes, following the Benchmark for 6D Object Pose Estimation (BOP) convention. Different training datasets are rendered without textures and while applying object texture randomization, to further increase edge detection accuracy. Each scene is rendered from multiple camera viewpoints using BlenderProc [Denninger et al., 2020], which outputs RGB images, visibility masks, and edge maps for all selected target objects. Two types of datasets are generated, each based on different object collections and scene configurations:

1. T-LESS Floor Scenes With Distractors - CAD models from the T-LESS dataset [Hodan et al., 2017] are randomly placed on textured planes using physics simulation to ensure realistic spatial distribution. Distractor objects are sampled from HB [Kaskman et al., 2019], YCB-V [Xiang et al., 2018], and ITODD [Drost et al., 2017], representing industrial domain variability. Figure 2 illustrates an example of a T-LESS based scene without texture randomization, showcasing visible Freestyle edges and object masks.
2. XYZIBD Box Scenes - Multiple instances of a small subset of objects from the XYZIBD dataset [Huang et al., 2025] are dropped into a box using physics simulation. This setup reflects bin-picking scenarios with high object density and occlusion.

Figure 2: A Sample of a T-LESS Training Scene with Overlaid Target Object Edges on the Left and Visibility Masks on the Right

To obtain object-specific edge maps, the Blender Freestyle tool is integrated into the Blenderproc rendering pipeline [1]. Due to its CPU-bound nature, mesh simplification is performed to reduce computational overhead during edge rendering. This involves dissolving geometry below a specified angular threshold and segmenting non-planar faces to reduce mesh complexity. Additionally, loose vertices, edges, and faces are removed to eliminate redundant geometry using built-in operations available in the Blender Python module [2].

[1] https://github.com/DLR-RM/BlenderProc/pull/1203
[2] https://pypi.org/project/bpy/

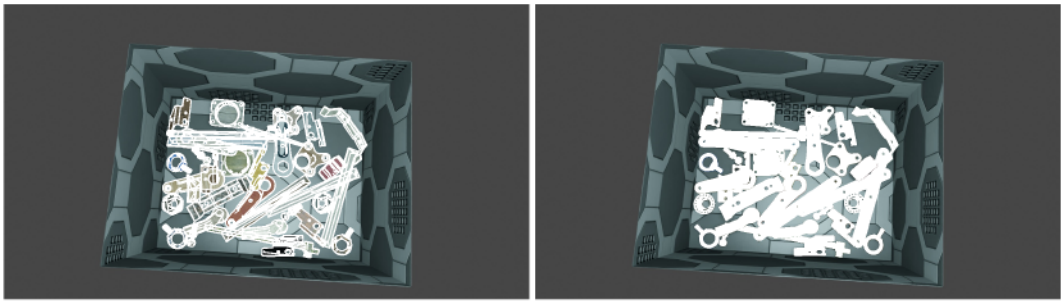
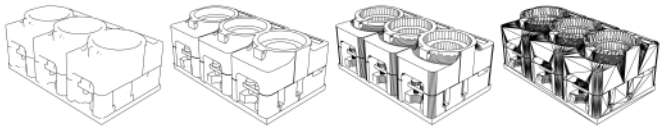


Figure 3: A Sample of a XYZIBD Training Scene with Overlaid Target Object Edges on the Left and Visibility Masks on the Right

Rendering performance is further improved by replacing the Cycles engine with EEVEE, which prioritizes rendering speed over photorealistic accuracy, while maintaining sufficient edge fidelity for the intended application. Freestyle provides a wide range of configuration parameters that influence the appearance and structure of the generated edge maps. In configuring Freestyle for line rendering, several parameters are enabled to ensure consistent and perceptually coherent edge extraction. Edge chaining is activated to guarantee that adjacent edge segments are properly connected, thereby producing continuous line structures. To preserve essential geometric and perceptual features, silhouette, crease and contour are enabled. These collectively ensure that major object outlines and sharp angular transitions are retained in the final render. In contrast, border edges are omitted to exclude outer object perimeters, while the visibility is set to only render visible edges. The angle between two adjacent faces, referred to as crease angle [3], plays a critical role in determining which mesh edges are rendered, as illustrated in Figure 4. A higher crease angle includes more shallow edges. A crease angle value of 180° (see Figure 4d) results in all mesh edges being drawn, whereas lower values like 175° (see Figure 4c), 160° (see Figure 4b) and 60° (see Figure 4a) exclude increasingly steep contours.



(a) 60° (b) 160° (c) 175° (d) 180°

Figure 4: Crease Angle Variation on a T-LESS Object

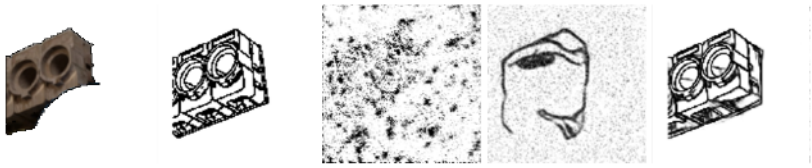
To evaluate the edge prediction quality of the trained models on real world images from T-LESS and XYZIBD, the test datasets are supplemented by the edge-ground truth. To achieve this, the scenes from the test sets are recreated in BlenderProc using available scene parameters and the ground-truth edge maps of each object in the scene are rendered using a crease angle of 160°. Predicted edges are later compared against the edge ground-truth in the test sets. Metrics aggregated in this manner enable a quantitative, real-world evaluation of the models trained on the fully synthetic training datasets. In this work, gray edges describe the grayscale edge maps provided by default by the Freestyle pipeline. These edge maps in the uint8 data format, contain values between 0 and 255 and represent the strength of an edge in each pixel. Clamped edge maps are created by setting all non-zero pixel values in the gray edge maps to 255. This results in all non-zero pixels being defined as an edge pixel, without regards to the strength of the edge. To evaluate the impact of edge type, object textures and crease angle on the edge prediction quality, multiple distinct training and test datasets are created. The same diffusion network architecture is trained separately on each dataset. In the clamped T-LESS dataset, texture-less objects are paired with clamped edge maps. The gray T-LESS dataset contains texture-less objects with grayscale edge maps. In the random texture T-LESS dataset, gray edge maps are used, while randomized textures are applied to the object surfaces. The XYZIBD training datasets are created using only gray edges, but with variation in texture, resulting

[3] https://docs.blender.org/manual/en/latest/render/freestyle/view_layer/line_style/modifiers/color/crease_angle.html

in two training sets. The XYZIBD test datasets are rendered using gray edges at six different crease angles.

4 Diffusion Model Training

In image generation tasks, U-Net commonly serves as the backbone denoising network [Rombach et al., 2022], learning to reverse the noise injection process iteratively. This property is especially advantageous for edge detection, where preserving fine-grained spatial features and their relations is critical. In this work, the U-Net operates at each denoising timestep to refine edge representations from noisy image samples. The training data generated by the BlenderProc pipeline provides synthetic RGB images along with corresponding object edges, masks, and visibility masks. The visibility masks are used to crop each image around the visible object, expand the bounding box by x1.5, and rescale it to a resolution of 128x128 pixels. Each training sample consists of three RGB channels and one grayscale edge channel, concatenated as input for the diffusion model. Each model is trained for approximately 120 hours on a single NVIDIA RTX 3090, using one of three dataset configurations: texture-less objects with clamped edges, texture-less objects with grayscale edges, and randomly textured objects with grayscale edges. During training, randomly selected training dataset samples are used as an estimate of the models edge detection ability. To achieve this, the model is switched to inference mode periodically and the edge predictions of the randomly selected samples are saved. The model architecture is based on the example from [4].



(a) (b) (c) (d) (e)

Figure 5: Running Validation During Model Training On Clamped Edges

Figure 5a shows the RGB and Figure 5b the edge ground-truth of a T-LESS validation sample. In early training stages (see Figure 5c), predictions exhibit high noise levels, with diffuse white regions loosely corresponding to the objects location. After approximately 2000 steps (see Figure 5d), the model begins to delineate object contours, though it continues to struggle with occlusion and complex geometries. By the end of training (see Figure 5e) the model consistently produces coherent and structurally plausible edge predictions. The models are deemed as sufficiently trained, when the training and validation loss converge to a minimum. For T-LESS-trained models 20 epochs are enough to achieve this. Models trained on XYZIBD training sets take 30 epochs to reach a comparable prediction quality.

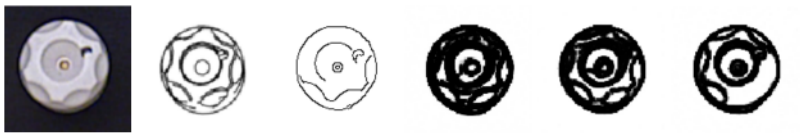
5 Model Quality Evaluation

The network predicts a single-channel edge map, which should ideally converge towards the edge ground-truth, independent of occlusion. However, when occlusion is severe and only a small portion of the object is visible, the network lacks sufficient context to infer object identity, orientation, or shape. In such cases, it tends to produce diffuse or spatially inconsistent edge responses, often concentrated around the region where the object is partially visible. These predictions reflect the model's uncertainty and its attempt to approximate the most likely object configuration under limited visual evidence.

Figure 6 provides a visual method comparison for a single sample from the T-LESS testset. The RGB image of the object from which all edge predictions are computed is shown on the left Figure 6a. It is followed by the edge ground-truth in Figure 6b acquired via Freestyle renderer. Figure 6c shows the Canny detector prediction with thin, but structurally incomplete edges. The remaining edges in Figure 6d, Figure 6e and Figure 6f stem from the models trained on clamped edges and texture-less

[4] https://huggingface.co/docs/diffusers/tutorials/basic_training#create-a-unet2dmodel

objects, gray edges and texture-less objects, as well as gray edges and randomly textured objects respectively. The model trained on clamped edges showcases thick, but mostly coherent edges. On the other hand the model trained on random textures does not fully capture the objects edges, but offers well defined edge contours. The gray edge, texture-less object model sits in between the two others, with relatively thin edges and a mostly complete edge structure.



(a) (b) (c) (d) (e) (f)

Figure 6: Edge Detection Method Comparison on a Single T-LESS Object

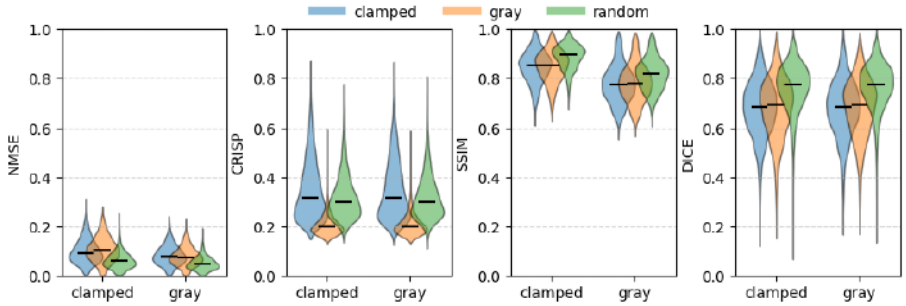


Figure 7: Edge prediction quality metrics NMSE↓, SSIM↑, DICE↑, CRISP↑ of models trained on clamped, gray, and random texture training datasets and evaluated on clamped and gray test datasets

To assess the overall performance of the trained models quantitatively, an evaluation is conducted using real world images from the T-LESS and XYZIDB test datasets, which are extended by the edge ground-truth via scene replication and the established Freestyle edge rendering approach. Predicted edge maps are compared against ground-truth edge maps using several established metrics, each capturing distinct aspects of edge quality, ranging from pixel-level accuracy to perceptual sharpness and structural fidelity. During evaluation, samples with a visibility mask containing fewer than 1% non-zero pixels when compared to the total number of pixels, are discarded. In these cases the target objects are deemed as too severely occluded.

Normalized Mean Square Error (NMSE) ([0,1] $\rightarrow$ lower is better $\downarrow$) provides a direct pixel-wise comparison between the predicted and ground-truth edge maps. This metric is sensitive to absolute intensity differences and does not consider spatial structure. Normalization is performed by scaling pixel intensities to the [0,1] range, enabling consistent comparison across images with varying brightness levels. Structural Similarity Index (SSIM) ([0,1] $\rightarrow$ higher is better $\uparrow$) [Wang et al., 2004] evaluates perceptual similarity between predicted and ground-truth edge maps by comparing local patterns of pixel intensities. It incorporates luminance, contrast, and structural alignment, and is computed over sliding windows, making it sensitive to localized distortions and edge deformation. In this work, SSIM is computed using a fixed intensity range of 255, consistent with 8-bit grayscale images. The Dice Similarity Coefficient (DICE/F-Score) ([0,1] $\rightarrow$ higher is better $\uparrow$) quantifies the spatial overlap between binary segmentation maps. Both predicted and ground-truth edge maps are thresholded to binary masks with values 0 or 255 and converted to boolean arrays. The Dice coefficient, equivalent to the F1 score in binary classification, measures the degree of overlap between the two masks, reflecting edge localization accuracy. The Crispness (CRISP) [Ye et al., 2023] ([0,1] $\rightarrow$ higher is better $\uparrow$) factor quantifies edge sharpness by computing the ratio of total pixel intensity after Non-Maximum Suppression (NMS) to that before NMS, using grayscale edge maps. A higher value indicates that NMS effectively preserves strong, localized edges while suppressing diffuse or thick contours. This metric penalizes blurred boundaries and favors well-defined edge structures.

Figure 7 presents violin plots of the discussed metrics for models trained on clamped edges and texture-less objects, gray edges and texture-less objects, and gray edges combined with randomly

textured objects. All objects stem from the T-LESS dataset. All three models are evaluated on two variations of the T-LESS test set. The first test set contains clamped and the second test set gray ground-truth edges. The models exhibit consistent performance across evaluation domains, indicating robustness to edge type, except in the case of SSIM, where the clamped test set provides a sharper edge ground-truth. The texture type used during training significantly influences prediction quality [Hönig et al., 2025].

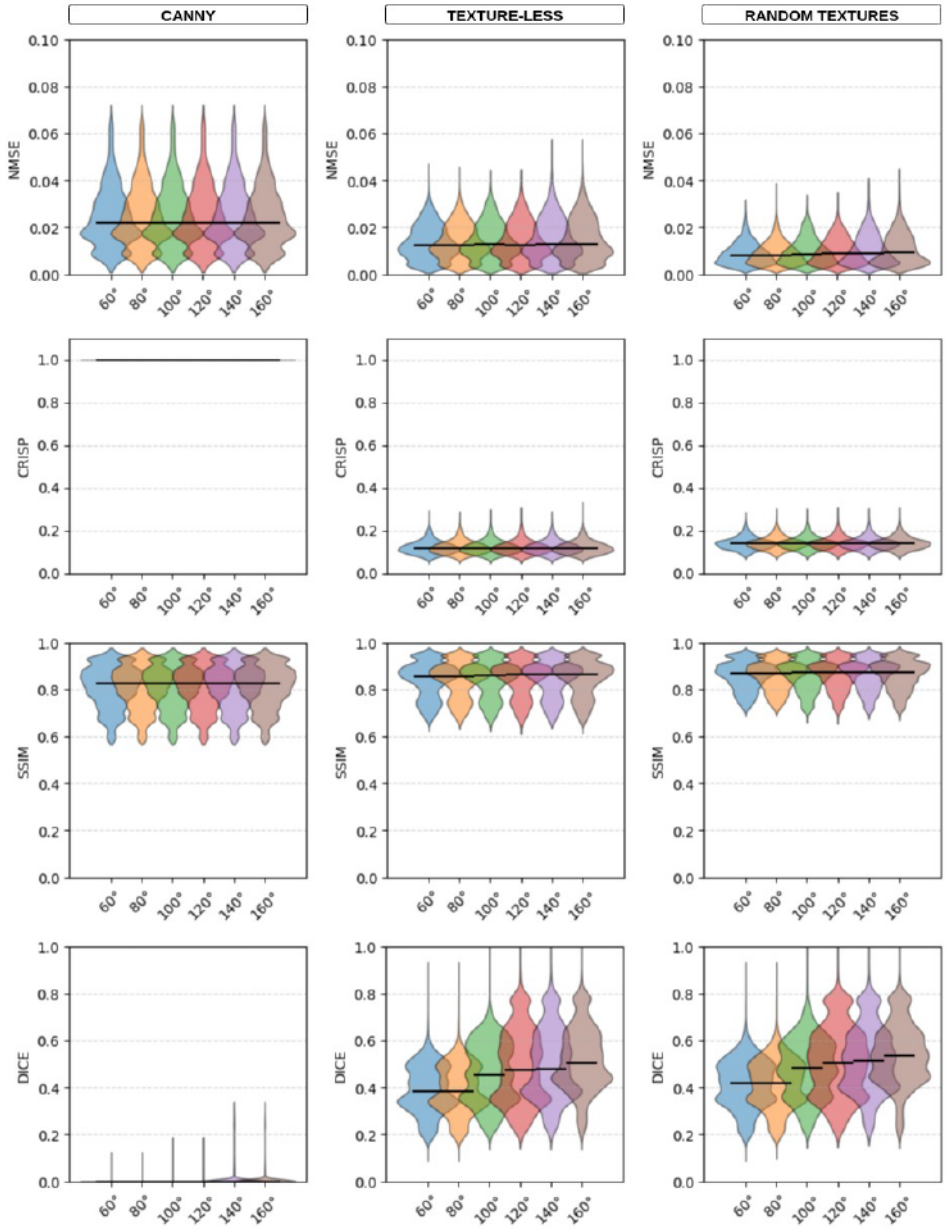


Figure 8: Edge Prediction Quality Metrics NMSE↓, SSIM↑, DICE↑, CRISP↑ for the canny edge detector, a texturelessly trained diffusion model and a diffusion model trained on objects with random textures

The model trained using gray edges and random textures consistently outperforms the clamped and gray models which are trained on texture-less objects across all metrics and test datasets, with the exception of the CRISP score. It yields the lowest NMSE, indicating better pixel-level accuracy, and achieves the highest SSIM, suggesting superior perceptual and structural similarity. It also yields the highest DICE coefficient, denoting improved segmentation overlap. Its CRISP score is slightly

worse than that of the clamped texture-less model, indicating sharp and well localized edges post-NMS. These results support the conclusion that the random texture model offers the most effective edge prediction performance for texture-less objects, aligning with the findings of As visualized in Figure 4, the crease angle has a large impact on the rendered edges. To evaluate its importance quantitatively, two diffusion models are trained on XYZ training datasets with a crease angle of 160° and gray edges. The first model is trained on a texture-less variant, while the textures in the other dataset are randomized. Both models are evaluated on the same XYZ test set with ground-truth edges rendered at crease angles: 60°, 80°, 100°, 120°, 140° and 160°.

The resulting metrics of both diffusion models are compared against the Canny edge detector [5] in Figure 8. The model trained on random textures performs better than the texture-less model across all metrics, while also being superior to Canny in NMSE, SSIM and DICE. Canny, by definition, exceeds in the CRISP metric, since NMS is part of its edge detection pipeline. This results in the CRISP score for Canny always being 1. On the other hand, Canny struggles with precise edge localization, especially at low crease angles, which results in a low DICE score across the validation dataset.

The trained models are evaluated on a system equipped with an AMD Ryzen 7 5800X 8-Core Processor and an NVIDIA RTX 3090 24GB GPU. Inference with diffusion models proceeds through a series of denoising steps, where the network iteratively refines a noisy input towards a clean edge prediction. Each step incrementally improves structural coherence and reduces noise, making the number of inference steps a key factor in both prediction quality and computational cost. For a 128×128 image, inference with 5 denoising steps takes approximately 95ms with a batch size of 1. In contrast, DiffusionEdge processes full scenes in a around 3.2 seconds. The step-wise nature of diffusion inference allows for fine-grained control over the quality–speed trade-off. To quantify the impact of inference step count on runtime, the clamped model is evaluated on the clamped test set using a batch size of 32. Increasing the number of inference steps from 1 to 2 results in a 1.3× increase in computation time, while increasing from 5 to 10 steps leads to a 1.7× increase. Empirically, 5 inference steps offer a favorable balance, where they allow for sufficient noise suppression and coherent edge maps production, while keeping inference time low. Increasing to 10 steps yields only marginal qualitative improvements.

6 Conclusion

This work presents a diffusion-based model for edge prediction of industrial, texture-less objects, trained on procedurally generated synthetic data and evaluated on real-world scenes. The resulting model demonstrates strong qualitative performance, particularly in handling complex geometries and partial occlusions, where it produces coherent edge maps even under limited visual evidence. Compared to existing model-based methods, the proposed approach offers improved computational efficiency at lower resolutions, making it suitable for deployment on resource-constrained platforms. Future work will focus on extending the method to multi-object and cluttered scenes, validating performance on broader real-world datasets, and integrating the model into robotic perception pipelines.

Declaration of AI-assisted Tools Used in Manuscript Preparation

Generative AI tools, including ChatGPT and Microsoft Copilot, were used to improve the clarity and readability of the manuscript. They were also employed to assist in writing Python code for implementation and experiment visualization. All generated material was reviewed and edited by the authors. The authors take full responsibility for the final content of the article.

Acknowledgments and Disclosure of Funding

This project is funded by the FFG, project GemSort, project number FO999923008 (www.ffg.at), the Austrian Science Fund (FWF), under project No. I 6114, project iChores, and by the EU program EC Horizon 2020 for Research and Innovation.

[5] https://docs.opencv.org/3.4/da/d22/tutorial_py_canny.html

References

John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986.

Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Dmitry Olefir, Tomas Hodan, Youssef Zidan, Mohamad Elbadrawy, Markus Knauer, Harinandan Katam, and Ahsan Lodhi. Blenderproc: Reducing the reality gap with photorealistic rendering. In *16th Robotics: Science and Systems (RSS), Workshops*, 2020.

Bertram Drost, Markus Ulrich, Paul Bergmann, Philipp Hartinger, and Carsten Steger. Introducing mvtec itodd - a dataset for 3d object recognition in industry. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops*, 2017.

Tomáš Hodan, Pavel Haluza, Štěpán Obdržálek, Jiří Matas, Manolis Lourakis, and Xenophon Zabulis. T-less: An rgb-d dataset for 6d pose estimation of texture-less objects. *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2017.

Gang Hu. A mathematical survey of image deep edge detection algorithms: From convolution to attention. *Mathematics*, 13(15), 2025. ISSN 2227-7390.

Junwen Huang, Jizhong Liang, Jiaqi Hu, Martin Sundermeyer, Peter KT Yu, Nassir Navab, and Benjamin Busam. Xyz-ibd: High-precision bin-picking dataset for object 6d pose estimation capturing real-world industrial complexity, 2025.

Qinyuan Huang and Jiaxiong Huang. Comprehensive review of edge and contour detection: from traditional methods to recent advances. In *Neural Computing and Applications*, pages 2175–2209, 2025.

Peter Hönig, Stefan Thalhammer, Jean-Baptiste Weibel, Matthias Hirschmanner, and Markus Vincze. Shape-biased texture agnostic representations for improved textureless and metallic object detection and 6d pose estimation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8806–8815, 2025.

N. Kanopoulos, N. Vasanthavada, and R.L. Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of Solid-State Circuits*, 23(2):358–367, 1988.

Roman Kaskman, Sergey Zakharov, Ivan Shugurov, and Slobodan Ilic. Homebreweddb: Rgb-d dataset for 6d pose estimation of 3d objects. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, 2019.

J. Kittler. On the accuracy of the sobel edge detector. *Image and Vision Computing*, 1(1):37–42, 1983.

Lukas Leimeister. Electriceye: Metallic object pose estimation. Diploma thesis, Technische Universität Wien, Vienna, Austria, 2025.

Yun Liu, Ming-Ming Cheng, Xiaowei Hu, Kai Wang, and Xiang Bai. Richer convolutional features for edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Yuancheng Luo and Ramani Duraiswami. Canny edge detection on nvidia cuda. In *2008 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2008.

Douglas Morrison, Peter Corke, and Jürgen Leitner. Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation. *IEEE Robotics and Automation Letters*, 5 (3):4368–4375, 2020.

Hönig Pulli, Hirschmanner Thalhammer, and Vincze. Enhancing transparent object pose estimation: A fusion of gdr-net and edge detection. In *Proceedings of Austrian Symposium on AI, Robotics, and Vision 2024*, pages 355–363, 2024.

Wei-Qing Ren, Yu-Ben Qu, Chao Dong, Yu-Qian Jing, Hao Sun, Qi-Hui Wu, and Song Guo. A survey on collaborative dnn inference for edge intelligence. *Machine Intelligence Research*, 20 (3):370–395, 2023.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer, 2015.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

Josh Tobin, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba, and Pieter Abbeel. Domain randomization for transferring deep neural networks from simulation to the real world. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 23–30, 2017.

Shigang Wang, Xianghua Liao, and Guoqiang Wu. Infrared image edge detection based on improved canny algorithm. In *2021 IEEE 3rd Eurasia Conference on IOT, Communication and Engineering (ECICE)*, pages 280–284, 2021.

Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

Jean-Baptiste Weibel, Clemence Dubois, Negar Layegh Khavidaki, Saifeddine Aloui, Mathieu Grossard, Markus Vincze, and Andreas Holzinger. Silref: Joint visual silhouette and tactile pose optimization for transparent object manipulation. *IEEE Robotics and Automation Letters*, 11(3): 2490–2497, 2026.

Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *Robotics: Science and Systems Proceedings*, 2018.

Yunfan Ye, Renjiao Yi, Zhirui Gao, Zhiping Cai, and Kai Xu. Delving into crispness: Guided label refinement for crisp edge detection. *IEEE Transactions on Image Processing*, 32:4199–4211, 2023.

Yunfan Ye, Kai Xu, Yuhang Huang, Renjiao Yi, and Zhiping Cai. Diffusionedge: Diffusion probabilistic model for crisp edge detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):6675–6683, 2024.

Haixin Yu, Shoujie Li, Houde Liu, Chongkun Xia, Wenbo Ding, and Bin Liang. Tgf-net: Sim2real transparent object 6d pose estimation based on geometric fusion. *IEEE Robotics and Automation Letters*, 8(6):3868–3875, 2023.

Zhiding Yu, Chen Feng, Ming-Yu Liu, and Srikumar Ramalingam. Casenet: Deep category-aware semantic edge detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

Intelligent Augmentation Methods for Training Defect Detection on Circuit Boards

Olaf Kähler **Werner Bailer** **Georg Thallinger**
JOANNEUM RESEARCH Forschungsgesellschaft mbH
DIGITAL – Institut für Digitale Technologien
Steyrergasse 17, 8010 Graz, Austria
{olaf.kaehler,werner.bailer,georg.thallinger}@joanneum.at

Abstract

We discuss intelligent data augmentation strategies to help train object detection models from low-volume datasets. In particular, many industrial inspection tasks suffer from a lack of samples showing defects in the training data, and furthermore the failure cases are typically heterogeneous, leaving only a handful of samples for each of them. For our application scenario of printed circuit board (PCB) inspection, we propose and evaluate a strategy for synthesizing defects, as well as a strategy to copy-paste difficult, challenging, or otherwise rare cases into the training images. Maintaining this library of challenging or rare cases offers an easy way to update the model and integrate feedback after deployment. We evaluate the benefits of the augmentation strategies in experiments and present a reliable and accurate PCB inspection model trained with only 25 images.

1 Introduction

While dataset sizes for vision-language or foundation models are ever increasing, low-volume datasets are still a common challenge in many specialized real-world applications of machine learning. In particular, training samples of defects are typically rare in industrial inspection tasks and often show a heterogeneous range of defects with even fewer samples for any given failure mode. In this paper, we present and evaluate two data augmentation techniques to deal with low-volume datasets in an application to printed circuit board (PCB) inspection. The first approach automatically synthesizes defects on annotated instances of intact objects, the second is a variant of copy-paste augmentation, which enables continuous improvement of the model with little effort.

In the intended application scenario, PCBs are to be disassembled at the end of their useful life to recover reusable components from the boards [19] while discarding defective parts. Closely related inspection tasks also arise when assessing PCB repairability, e.g. by replacing individual defective components, or even as a quality control measure in PCB assembly. All of these scenarios basically need a system for identification and classification of PCB components, which is a classical object detection problem, except that there are only very few samples of defective parts available for training.

The approaches we propose for dealing with this lack of training data are widely applicable to a range of similar scenarios – it is very common that samples of real defects are rare when developing and training a model. In our particular case, we synthesize two kinds of typical defects using elementary image processing operations. This dramatically reduces the class imbalance and hence boosts classification performance. As an additional direction, we copy and paste instances from a library of challenging examples to random places into the training images, similar to [3, 4]. However, we propose manually maintaining and extending this library of challenging samples as an avenue for incorporating feedback using targeted adaptation and retraining.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

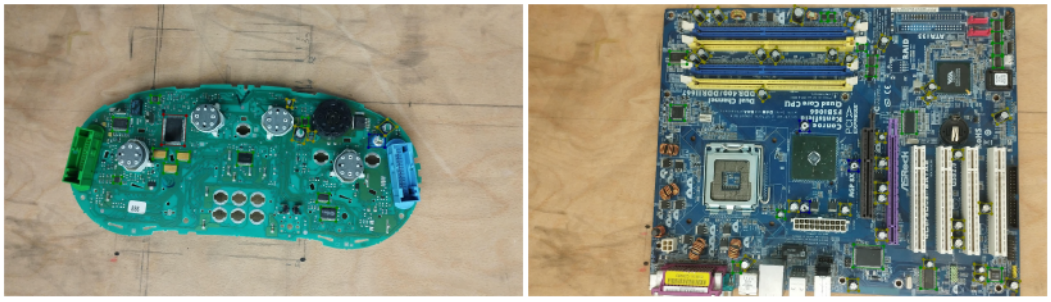


Figure 1: Samples of the images and annotations in the PCB dataset. Defect and intact ICs as well as defect and intact capacitors are annotated, but we focus on ICs in this work. On average, more than 7 ICs appear in each image and they typically appear relatively small. The recording perspective and illumination are constant within the entire dataset.

In the following, we will briefly revisit the relevant literature in Section 2. Continuing in Sections 3 and 4, we will give details of our proposed data augmentation techniques. Finally, we will evaluate our methods in Section 5 and summarize our findings in Section 6.

2 Literature Overview

Automated optical inspection of PCBs during manufacturing is receiving a lot of attention. A recent review of methods focusing on defects in PCB traces is given in [1], and an overview of methods for inspecting mounted components in [20]. In a similar vein, the detection of PCB components and the checking of completeness and correctness have been covered before with publicly available benchmark datasets. For example, the PCB-DSLR dataset [16] focuses on integrated circuits (ICs) and additional components are also considered in the PCB-METAL [12], WACV-PCB [8] and FICS-PCB [10] datasets. In contrast, end-of-life treatment and defects, that occur during use of components, have received little attention so far. Consequently, to the best of our knowledge, no datasets covering such defects have been published.

Synthesizing defects is a common strategy in industrial inspection tasks [13]. Samples of works that use synthetic defects in surface inspection include [6], and a recent work includes a sophisticated combination of a rendering pipeline and style transfer to maximize realism [17]. In addition, in the context of PCB inspection during assembly [18] and for X-ray CT scans of ICs [14], synthetic defects have been evaluated before. Relevant defects are often specific for the respective application scenario, and, to our knowledge, no pipeline has been proposed for synthesizing end-of-life defects from the real use of electronic components so far. Accordingly, we consider the details of our synthetization pipeline novel and different from the state-of-the-art.

In contrast, we see our copy-paste augmentation as a more generic strategy, which follows the idea of [4]. Although existing works evaluated the benefits of considering context when mixing images [5], we focus more on the library of source patches that such strategies can paste into new images. Although our experiments can only verify this in a limited scope, we believe that model training can be guided by deliberately selecting rare or novel out-of-distribution samples for the source library. To our knowledge, this aspect has also been neglected in the existing literature.

3 Synthesizing Defects

In our work, we consider a dataset for PCB inspection as shown in Figure 1, which is a subset of the dataset in [15]. In total, it consists of 101 images and shows 652 instances of intact ICs and additionally 78 instances of defect ICs. The defects are not further subdivided but contain a set of burned ICs with visible black spots on the surface and a set of ICs with broken and merged pins. Samples of these defects are shown in Figure 2. It is a common scenario that defects are rare and heterogeneous, such as in our dataset. To alleviate this class imbalance, we aim to introduce synthetic defects to the known and annotated ICs in our data.

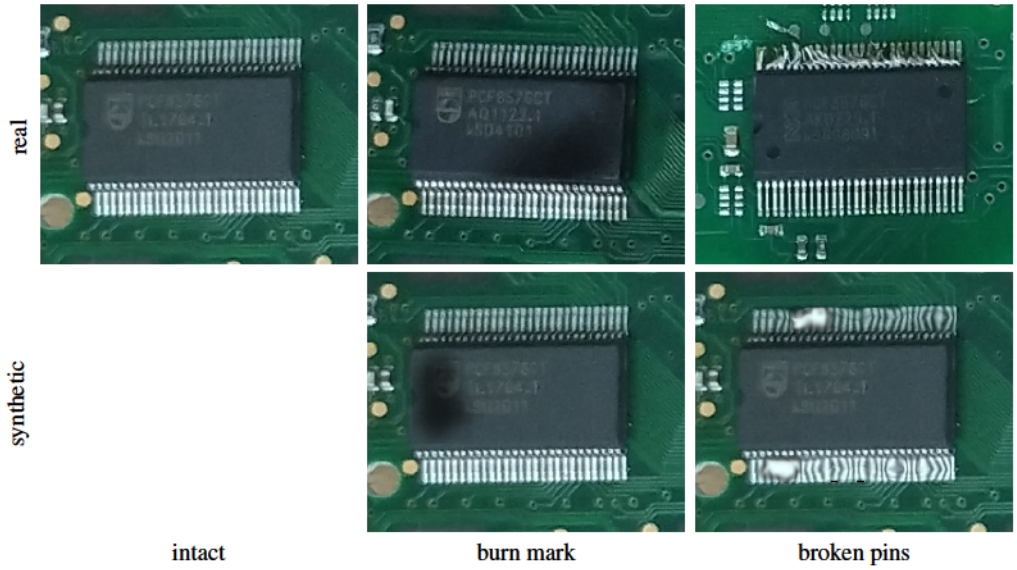


Figure 2: Samples of real defects in the training data in the top row as well as corresponding synthesized additional defects at the bottom. The synthetic defects are all added to the same reference image of an intact IC shown on the top left.

Since the perspective and illumination of images are fairly well constrained in our scenario, we skip any 3D modeling and rendering. Realistic and varied training data can be generated with 2D image processing in our case. In the following, we propose two different approaches to cater for the two modes of defects we observe in the given real sample data, namely dark spots on the IC package and pin deformations.

The first kind of synthetic defect introduces dark burn marks onto the ICs. To this end, we reduce the annotated bounding box of an intact IC by 10%, generate a random polyline chain within these bounds, apply some heavy Gaussian smoothing on its rendering, and alpha-blend the result onto the original input image. This simple but effective strategy produces realistic dark spots as shown in Figure 2.

For pin deformations, we perform three steps to a) identify image areas covered by pins, b) apply a random deformation to the relevant area, and c) add a random white spot on top resembling a solder blob. To identify the pin area, we compute image gradients near the image edges and identify blobs of high gradients. For deformations, a coarse mesh of source and target coordinates is generated randomly and bilinearly interpolated. This will ensure continuous and somewhat realistic looking outputs. In addition, we apply a similar procedure as for burn marks to add white spots onto the pins to simulate solder blobs. Again, a sample of the final result is shown in Figure 2.

The aforementioned transformations are applied to a random subset of all intact ICs in an image, and of course the result is an image with additional realistic-looking defective ICs. The procedures involved are lightweight and robust. They can hence be applied not only offline to generate a synthetic training dataset but also online as an additional data augmentation step.

4 Copy-Paste Augmentation

The aforementioned augmentation strategy significantly increases the variety of data, but does not offer a way of introducing novel variants of ICs or defects and adapting to distribution shifts. Annotating entirely new images for retraining purposes is an option here, but each image will easily contain dozens of objects, imposing a large annotation effort for each new sample.

We therefore argue that a much smaller library of challenging samples or rare cases can be extracted from additional training images. This merely requires cropping e.g. a single IC from a novel image, annotating it, and saving it to the library. Furthermore, with pre-cropped content, the annotation can

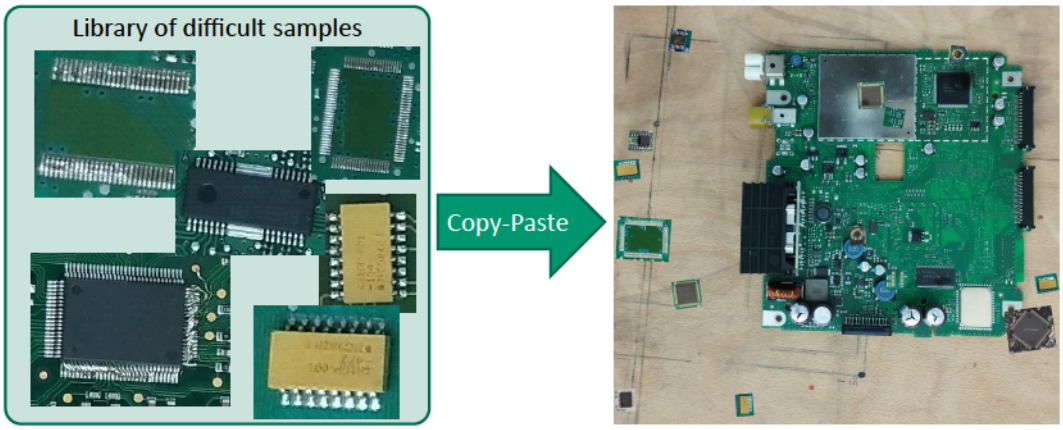


Figure 3: Illustration of pasting challenging samples from a library into an existing training image. The inserted objects appear at random locations and with minor variations in scale and orientation. Note that we do not perform any smoothing at the edges.

Table 1: Libraries of additional training instances used for weak and strong copy-paste augmentation.

	Weak Augmentation			Strong Augmentation		
Source	**intact**	**defect**	**background**	**intact**	**defect**	**background**
Training	7	0	16	7	0	16
Extra	5	3	0	26	5	4
WACV-PCB [8]	6	0	1	37	0	4

be done efficiently with models such as SegmentAnything [7], in most cases with a single inference run. A sample of such a library of rare cases is illustrated in Figure 3.

These crops are then pasted to random locations in the training images, where no other object is already present. This simplifies handling of overlapping boxes or annotation masks and is sufficiently realistic for our images of flat PCBs with components mounted on the top. To further increase variability, we apply random scaling by $\pm 10\%$, flipping, rotations by multiples of $90°$, and additional rotations by up to $\pm 10°$. We do not apply any smoothing to the edges of the pasted crops, as this does not generate any additional benefit. This is in line with reports, e.g. in [4]. If outlines of the PCBs are known, this procedure can be further refined so that the novel samples are only pasted to the foreground regions. With unknown PCB outlines, simply pasting larger numbers of samples in both the foreground and background is also feasible. Pasting of multiple novel objects can be handled sequentially.

Maintaining the library of challenging or rare samples provides an easy way to incorporate feedback into the training, since expanding this library is a quick and efficient process. For a real application, we envisage using this process, whenever a misdetection is observed by an operator. For reproducible results in our experiments, we only incorporate a fixed set of additional training samples. In particular, these samples are extracted using images from the WACV-PCB dataset [8], as well as 7 additional images, that are not part of any other dataset. We also include crops of challenging parts from the original training data to ensure that these have a good chance of being properly learned. To evaluate whether just repeating known parts of the training data improves performance or whether the actually novel out-of-distribution samples help with generalization, we consider weak and strong copy-paste libraries in our experiments, with detailed statistics given in Table 1.

Note that a majority of the ICs in these libraries are in fact intact. Defects therefore remain rare as the WACV-PCB dataset, like other relevant datasets, does not contain any defects. Hence, a combination with the aforementioned methods of synthesizing defects seems to be the most promising to boost performance. In particular, there are no defective samples of the rare yellow ICs shown as part of our library in Figure 3, but with the combination of our proposed techniques, these are also represented in the training process. Examples of augmentation results for the original images from Figure 1 are shown in Figure 4.

Table 2: Evaluation of several object detectors with different augmentation schemes.

Synthesis	Copy-Paste	Mask2Former [2]		Dino [21]		ConvNeXt [9]		RTMDet [11]	
		mAP @0.5	mAP @0.75	mAP @0.5	mAP @0.75	mAP @0.5	mAP @0.75	mAP @0.5	mAP @0.75
-	-	78.4	67.2	83.7	60.1	88.1	78.9	62.2	33.8
-	weak	80.3	69.8	86.1	62.7	89.3	81.0	67.3	46.7
-	strong	80.7	71.1	87.1	63.3	88.8	80.6	69.4	46.6
active	-	87.6	**81.0**	90.8	68.1	91.8	**84.6**	67.6	48.6
active	weak	88.3	79.0	92.3	**70.1**	92.6	82.9	**69.7**	**49.1**
active	strong	**89.1**	80.5	**92.4**	70.0	**93.1**	84.2	68.6	47.7

5 Evaluation

In order to verify the efficiency of the proposed data augmentation schemes, we first perform a qualitative visual analysis and then a thorough quantitative analysis of their effect on the training results.

For visual analysis, we inspect several of the augmented images, such as shown in Figure 4. In these images, a random number of up to 20 patches from the copy-paste library are added, and up to 70% of the intact ICs are modified by adding synthetic defects. Closeups of real defects, synthetic defects, and synthetic defects on pasted ICs from the library are also shown in Figure 5. Although a human observer can spot the difference between real and synthetic defects and the pasted image patches do not line up with the rest of the PCBs, the overall appearance of the modified training data is realistic enough upon casual observation. In any case, we consider it worthwhile to run experiments to validate their impact on the training of machine learning models.

To this end, we performed several experiments on the presented dataset of 101 images. We randomly split this dataset into a fixed training set of 25 images, containing 196 intact and 15 defect ICs respectively, and a validation set of 76 images, showing 456 intact and 63 defect ICs. The training set is deliberately selected as small as 25 images to show the efficiency of our data augmentation strategies for low-volume datasets. Experiments with a split of 75 training images and 26 validation images showed similar trends, but at higher variance due to the small validation dataset.

We train various object detection and instance segmentation models, in particular Mask2Former [2], Dino [21], Cascade Mask-RCNN with a ConvNeXt backbone [9] and RTMDet [11], to ensure that our results are generalizable and transferable. For training, we use stochastic gradient descent running for 1,000 epochs, while we observe that all models achieve convergence and stabilize their performance way before the training ends. The model weights are initialized from pretraining on the COCO dataset. A cosine annealing strategy is applied to reduce the learning rate, and we switch from a strong data augmentation pipeline for the first 600 epochs to a lighter one for the remaining final epochs. If enabled, the first steps in data augmentation are our proposed copy-paste augmentation and the strategies for synthesizing defects on annotated ICs. This is followed by a strong set of standard data augmentation routines of mosaicing, resizing, cropping, HSV and photometric distortions, random horizontal and/or vertical flipping, mixup and random rotations of up to 15°. With this strong set of state-of-the-art augmentation methods, we observe little to no signs of overfitting in any of our experiments.

To evaluate the effects of our novel data augmentation schemes, we first train a baseline model using only the state-of-the-art standard augmentation methods. We then selectively apply and combine our proposed copy-paste augmentation and the pipeline to generate synthetic defects. The resulting values for mAP@0.5 and mAP@0.75 obtained at the bounding box level in our validation set are listed in Table 2.

From these results, it is obvious that additional data augmentation improves the performance in all cases, which is not surprising given the small training dataset. A significant boost is observed whenever the pipeline for synthesizing defects is enabled, indicating that 2D image processing routines provide an adequate way to generate realistic images of damage. The copy-paste augmentation shows some weak gains, even if mainly the challenging samples from the training set are repeated. However,

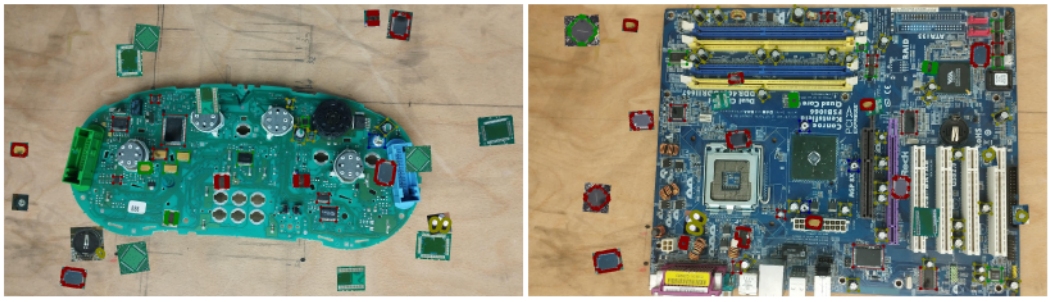


Figure 4: Samples of typical PCBs with several components that were modified using our proposed data augmentation steps. Both our copy-and-paste augmentation and the defect synthesis are enabled and significantly increase the available training samples of defects.

with additional out-of-distribution samples pasted into existing training images, additional gains are consistently achieved, indicating that this is indeed a viable way of extending the training data via online feedback. With the already strong pipeline for synthesizing defects enabled, the additional gains of the copy-paste augmentation only manifest at the mAP@0.5 level. Note that the library of source patches in our copy-paste augmentation already focuses on typical mistakes and challenging samples, that we identified in prior experiments. Some examples of success stories illustrating the benefits of our augmentation schemes are also given in Figure 6.

Comparing the different models, ConvNeXt [9] appears to perform marginally better than Dino [21] and Mask2Former [2] in both the mAP@0.5 and mAP@0.75 metrics, while RTMDet [11] being optimized for speed is trailing behind by a larger margin. In all cases, the proposed data augmentation strategies improve the mAP scores up to a top value of 93.1 for mAP@0.5, thus making the automatic PCB inspection system a viable approach in the intended use case. Some remaining failure cases are shown in Figure 7. We believe some of these could be addressed by collecting additional samples and injecting them into the copy-paste library. However, these are hard to find in our small training set of 25 images. It should also be noted that the presented detectors already generated valuable feedback on annotation errors for the domain experts who prepared the dataset and perform at least on par with human annotators.

6 Conclusions

In this paper, we consider techniques for applying machine learning in use cases with low-volume datasets. Furthermore, the data in our use cases have very specific characteristics, and approaches such as pre-training on web images have little chance of covering the specific defects we are interested in. This is a typical scenario in many industrial vision applications, where datasets show only few defects, that are also heterogeneous. Since each failure case is represented only by very few samples, conventional training methods show only limited success.

We proposed two data augmentation strategies to counter the data scarcity in the application of PCB component detection and inspection. As a first strategy, defects can be synthesized on the intact samples by image transformations. As the second strategy, we propose to create a library of challenging samples and paste them into the training images. This provides a viable approach for incorporating feedback and out-of-distribution samples into training with little manual effort.

Overall, these techniques significantly improve the performance of a detector on the given PCB dataset. They can also be used as a blueprint for similar applications with low-volume datasets.

Acknowledgments and Disclosure of Funding

This work has been supported by the project European Lighthouse to Manifest Trustworthy and Green AI (ENFIELD), which has received funding from the European Union's Horizon Europe research and innovation program under grant agreement No. 101120657.

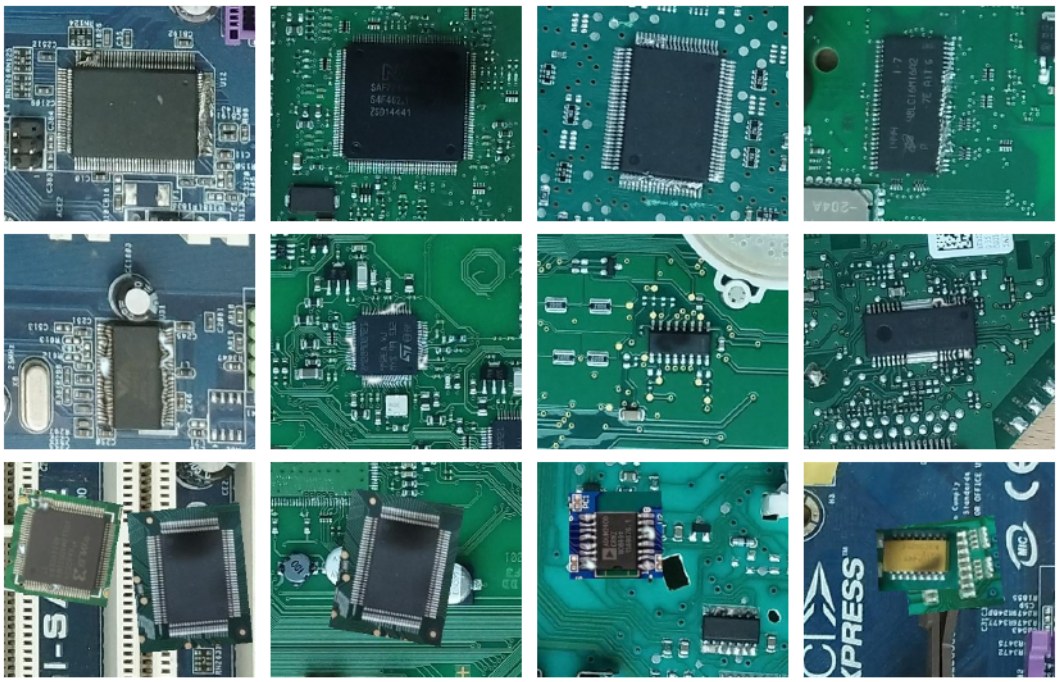


Figure 5: Closeup illustration of real defects (top row), synthesized defects (middle row) and synthesized defects on copy-and-paste augmented samples (bottom row) generated with our proposed data augmentation schemes. While close observation will reveal the synthetic nature of some of the images, they appear convincing and well suited to extend the very limited training set.

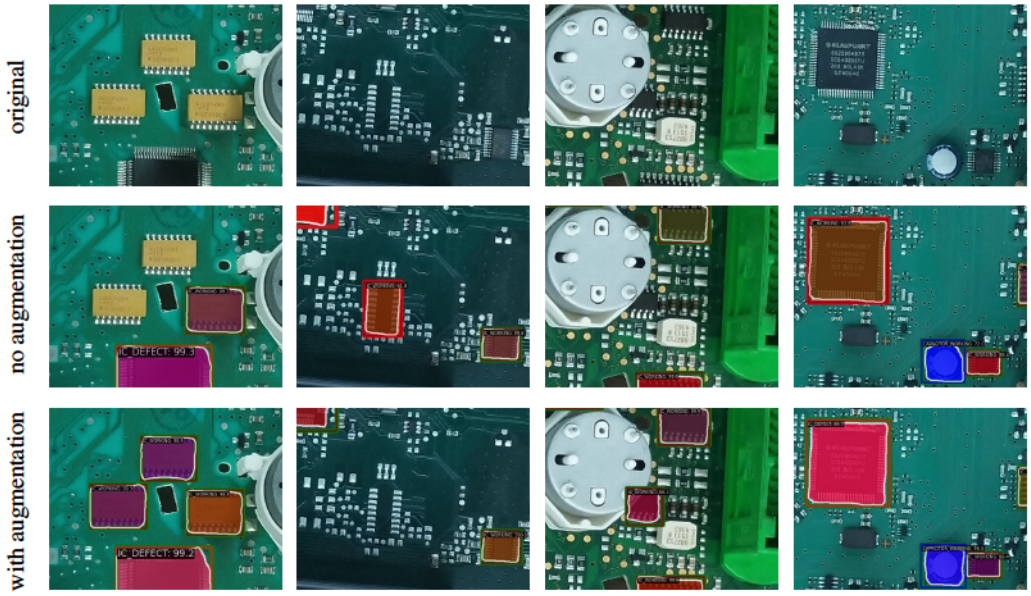


Figure 6: Illustrations of success stories, where the data augmentation schemes improved detection results. Left to right: Additional samples of yellow ICs in the copy-paste library improve their detection, samples of unpopulated IC areas reduce false positives, samples of partly occluded ICs are also included in the copy-paste library and example of a slightly damaged IC, that is correctly classified with our augmentation scheme.

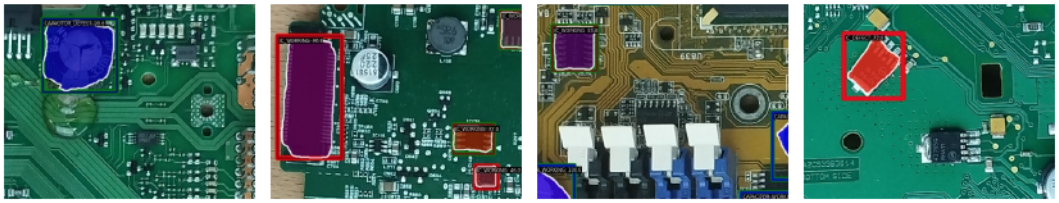


Figure 7: Samples of remaining failure cases: ICs of odd shapes are occasionally missed, in some cases, connectors are mistaken for ICs, not all occlusions can be handled and often, small MOSFETs or other irrelevant components are identified as ICs.

References

[1] Chen, X., Wu, Y., He, X., and Ming, W. (2023). A comprehensive review of deep learning-based pcb defect detection. *IEEE Access*, 11:139017–139038.

[2] Cheng, B., Misra, I., Schwing, A. G., Kirillov, A., and Girdhar, R. (2022). Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1290–1299.

[3] Dwibedi, D., Misra, I., and Hebert, M. (2017). Cut, paste and learn: Surprisingly easy synthesis for instance detection. In *Proceedings of the IEEE international conference on computer vision*, pages 1301–1310.

[4] Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., and Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR*, pages 2918–2928.

[5] Guo, Q., Wang, S., Chang, C., and Rambach, J. (2025). Cacp: Context-aware copy-paste to enrich image content for data augmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR) Workshops*, pages 5177–5186.

[6] Jain, S., Seth, G., Paruthi, A., Soni, U., and Kumar, G. (2022). Synthetic data augmentation for surface defect detection and classification using deep learning. *Journal of Intelligent Manufacturing*, 33(4):1007–1020.

[7] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. (2023). Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026.

[8] Kuo, C.-W., Ashmore, J. D., Huggins, D., and Kira, Z. (2019). Data-efficient graph embedding learning for pcb component detection. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 551–560. IEEE.

[9] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., and Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11976–11986.

[10] Lu, H., Mehta, D., Paradis, O., Asadizanjani, N., Tehranipoor, M., and Woodard, D. L. (2020). FICS-PCB: A multi-modal image dataset for automated printed circuit board visual inspection. Cryptology ePrint Archive, Paper 2020/366.

[11] Lyu, C., Zhang, W., Huang, H., Zhou, Y., Wang, Y., Liu, Y., Zhang, S., and Chen, K. (2022). Rtmdet: An empirical study of designing real-time object detectors.

[12] Mahalingam, G., Gay, K. M., and Ricanek, K. (2019). Pcb-metal: A pcb image dataset for advanced computer vision machine learning component analysis. In *16th International Conference on Machine Vision Applications (MVA)*, pages 1–5. IEEE.

[13] Nikolenko, S. I. (2021). *Synthetic data for deep learning*, volume 174 of *Springer Optimization and Its Applications*. Springer.

[14] Phoulady, A., Suleiman, Y., Choi, H., Moore, T., May, N., Shahbazmohamadi, S., and Tavousi, P. (2023). Synthetic data augmentation to enhance manual and automated defect detection in microelectronics. *Microelectronics Reliability*, 150:115220. Special issue of 34th European Symposium on Reliability of Electron Devices, Failure Physics and Analysis (ESREF 2023).

[15] Pižurica, N., Milović, N., Jovancevic, I., Nasirimajd, A., and Quadrini, W. (2025). Epid: The enfield pcb inspection dataset for visual defect detection. Zenodo, https://doi.org/10.5281/zenodo.16811808.

[16] Pramerdorfer, C. and Kampel, M. (2015). A dataset for computer-vision-based pcb analysis. In *14th IAPR international conference on machine vision applications (MVA)*, pages 378–381. IEEE.

[17] Ren, W., Song, K., Chen, C.-y., Chen, Y., Hong, J., Fan, M., Ouyang, X., Zhu, Y., and Xiao, J. (2025). Dd-aug: A knowledge-to-image synthetic data augmentation pipeline for industrial defect detection. *IEEE Transactions on Industrial Informatics*, 21(3):2284–2293.

[18] Saif, S. S., Aras, K., and Giuseppi, A. (2022). Automated optical inspection for printed circuit board assembly manufacturing with transfer learning and synthetic data generation. In *2022 30th Mediterranean conference on control and automation (MED)*, pages 318–323. IEEE.

[19] Simaei, E. and Rahimifard, S. (2024). Ai-based decision support system for enhancing end-of-life value recovery from e-wastes. *International Journal of Sustainable Engineering*, 17(1):80–96.

[20] Singh, K., Kharche, S., Chauhan, A., and Salvi, P. (2024). Pcb defect detection methods: A review of existing methods and potential enhancements. *Journal of Engineering Science & Technology Review*, 17(1).

[21] Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., Ni, L., and Shum, H.-Y. (2023). DINO: DETR with improved denoising anchor boxes for end-to-end object detection. In *The Eleventh International Conference on Learning Representations*.

Assessing Compressive Strength of Reclaimed Clay Bricks Using SWIR Hyperspectral Imaging and Deep Learning

Jean-Philippe Andreu, Maria Jernej
JOANNEUM RESEARCH
Graz, Austria
{firstname.lastname}@joanneum.at

Maximilian Klammer, Benjamin Kromoser
Universität für Bodenkultur Wien
Vienna, Austria
{firstname.lastname}@boku.ac.at

Abstract

A non-destructive approach is proposed to assess the compressive strength of reclaimed bricks using short-wave infrared (SWIR) hyperspectral imaging (HSI) and a spectral–spatial 1D-Convolutional Neural Network (CNN). Hyperspectral images of 60 bricks, capturing both outer (weathered) and inner (pristine) surfaces, were analyzed. Regression reached $R^2 = 0.625$, while a three class (*low*, *medium*, *high*) compressive strength classification achieved 83 % pixel level accuracy. At the brick level, aggregating predictions with a majority-vote scheme attained an accuracy of 91 % for outer and 98 % for inner surfaces. Score-CAM identified key wavelengths around 1200–1400 nm (moisture) and 2300–2500 nm (clay minerals) as driving the predictions. The results demonstrate that SWIR HSI can capture mineral- and moisture-related signatures relevant to compressive strength, offering a rapid, non-destructive screening tool for reclaimed bricks.

1 Introduction

Construction and demolition waste accounts for over a third of total waste in the European Union, highlighting the need for effective material recovery strategies. Within the circular economy framework, the European Commission promotes approaches that retain material value while minimizing additional processing. Clay bricks are particularly suitable for reuse, as their material value is preserved and energy-intensive reprocessing avoided [2]. Before reuse, however, material performance must be verified by assessing properties such as compressive strength, frost resistance, and water absorption [10]. Standard compressive strength testing involves crushing representative bricks in a hydraulic press, with the average strength determining the strength class. As this method is destructive and time-consuming, it is unsuitable for rapid screening of large quantities of reclaimed bricks.

Therefore, non-destructive techniques that enable rapid assessment of brick quality are of increasing interest. One promising approach is SWIR HSI, which provides information about both mineral composition and moisture content. Clay bricks are primarily composed of clay minerals such as kaolinite and montmorillonite, along with silicate minerals. These clay minerals exhibit characteristic absorption features in the SWIR region, particularly between 2000 nm and 2500 nm due to vibrational overtones and combination bands of hydroxyl (OH), metal–OH, and H_2O molecular bonds [13]. These spectral features allow the mineral composition to be inferred from hyperspectral data, which is closely linked to firing behavior, porosity, microstructure, and thus indirectly to compressive strength. SWIR wavelengths are also highly sensitive to water, with strong absorption bands near 1400 nm and 1900 nm [7, 9]. In porous ceramic materials such as bricks, moisture uptake is largely governed by open porosity, which in turn is again closely related to compressive strength [4]. Relationships between SWIR spectral features and compressive strength have been reported for several geomaterials. SWIR absorption bands related to clay minerals and alteration phases show negative correlations

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

with compressive strength of granite [12]. Similar correlations have also been observed in soils [8], carbonates [1], volcanic rocks [6] and water-bearing sandstones [14].

In this work, a spectral–spatial 1D-CNN framework, initially introduced by Hsieh and Kiang [5] on remote sensing hyperspectral data, was applied in a regression setting to investigate whether compressive strength can be predicted directly from SWIR hyperspectral images of demolition bricks. The model was subsequently repurposed for a three-class classification task, reflecting common practice in the construction industry where bricks are typically evaluated against compressive strength thresholds. In addition, post-hoc explainability using Score-CAM was applied to identify the wavelengths most relevant to the model's predictions.

2 Data Collection

A total of sixty bricks were extracted from different walls of a more than 100 years old building scheduled for demolition in Vienna, Austria. Buildings of this age typically contain bricks from different production batches and factories, as indicated by the different manufacturer stamps found on the samples. However, as all bricks originate from a single building, including samples from multiple demolition sites would further increase dataset diversity and should be considered in future work.

After extraction, the bricks were split in half immediately. One half of each brick was sealed to preserve its original moisture content, and the other half was stored to dry. Only the dried halves were used in the present experiments to reduce variability. In practical applications, bricks would likely be analyzed immediately after extraction. Therefore, the influence of site-specific moisture needs to be investigated in the future. Two surfaces of each selected half were imaged: (i) the largest outer face, which may contain dust, mortar residues, or other surface coatings, and (ii) the freshly exposed inner surface created by splitting the brick, representing uncontaminated brick material.

All images were acquired using a Specim SWIR hyperspectral push-broom camera, covering the spectral range of 1000 nm to 2500 nm with a spectral resolution of 12 nm for a total of 288 bands. A two-sided halogen lamp illumination was supplemented by a quartz glass rod to extend the spectral coverage toward the thermal range. The data were radiometrically calibrated to obtain relative reflectance, eliminating the influence of illumination and sensor response. Dark and white references were recorded using a black cold cloth and a Spectralon target, and each pixel spectrum was normalized against these references. After data acquisition, all bricks were subjected to an irreversible uniaxial compressive strength test at the University of Natural Resources and Life Sciences Vienna (BOKU). The resulting compressive strength values (in MPa) serve as ground-truth measurements.

3 Methodology

Instead of using transfer learning with an established backbone like ResNet [3] (which would require an early frequency selection / dimensional reduction in order to fit the required 3 input channels) we opted to follow Hsieh and Kiang [5], exploiting a 1D-CNN with spectral-spatial input features (rather than single spectra) to take into account the heterogeneous brick surfaces. For each target pixel, a 3×3 neighborhood was extracted and the nine spectral signatures were stacked together, yielding an input tensor of dimension $9 \times F$ (with $F = 288$ SWIR spectral bands). This representation encodes local texture and reduces sensitivity to single-pixel noise, relevant for heterogeneous brick surfaces.

A 1D-CNN regression model was designed with the following sequential architecture:

- 1D Convolutional layer: 40 filters of size 5×1, applied across the spectral dimension
- 1D Max pooling layer: factor-of-2 down-sampling followed by a ReLU activation function
- 1D Convolutional layer: 80 filters of size 5×1
- 1D Max pooling layer: factor-of-5 down-sampling followed by a ReLU activation function
- Intermediate fully connected linear layer with output size of 100
- Output fully connected linear layer for scalar regression (compressive strength in MPa)

Batch normalization was used after each convolutional layer to stabilize the learning process.

For classification, the same convolutional backbone was retained and a softmax function applied to the output of the final layer in order to get the probability distribution over the three compressive strength classes (*low*, *medium* and *high*). For this experiment, the boundaries of the three compressive

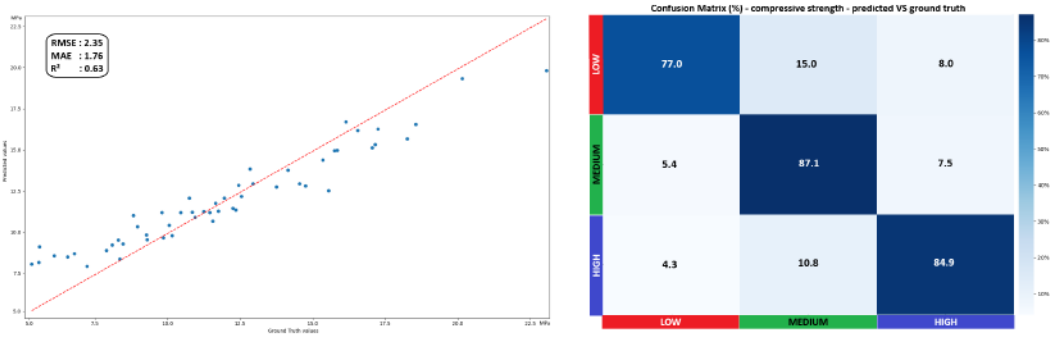
strength classes were automatically selected as the 1/3 and 2/3 percentiles of the compressive strength values of the dataset: < 9.95 MPa (*low*), 9.95 MPa – 13 MPa (*medium*) and > 13 MPa (*high*).

The network was implemented in PyTorch and trained using the Adam optimizer with a learning rate of 10^{-3} and a mean-squared-error loss for regression (cross-entropy for classification). The training was run over 2000 epochs with a confidence-based early exit strategy.

For each brick sample, 2000 spectral-spatial patches were extracted at random, resulting in a 120,000 sample dataset. It was then partitioned into 60 % training, 20 % validation and 20 % test sets using stratified sampling to balance the representation of outer and inner surfaces and strength ranges across splits.

4 Results

The performance of the regression model on the test set yielded the following results: a RMSE of 2.35 MPa, a MAE of 1.76 MPa and a R^2 of 0.625 (Fig. 1a). Approximately 63 % of the variance in compressive strength is explained by the model spectral-spatial input alone. The MAE of 1.76 MPa is operationally meaningful given typical quality-class boundaries of 5 MPa to 10 MPa in brick standards.



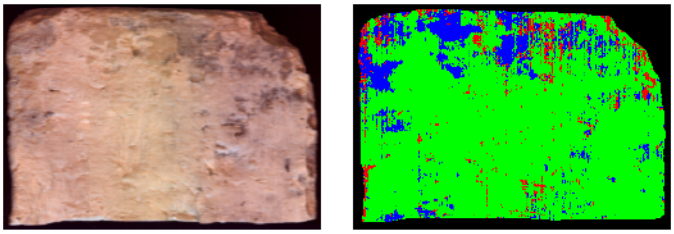
(a) Regression - Correlation Plot

(b) Classification - Confusion Matrix

Figure 1: Regression and classification results for the compressive strength of the reclaimed bricks.

When the CNN backbone was applied as a three-class classifier (*low*, *medium* and *high* compressive strength), pixel-level predictions on the test set achieved an overall accuracy of 83 % (Fig. 1b). Misclassifications were primarily caused by surface contamination (dust, coatings or weathering crusts) and small-scale material variability, rather than systematic differences between class boundaries. A comparison between the rendered (with false colors from the HSI data) image of a sample outer-face (Fig. 2a) and its pixelwise classification (Fig. 2b) shows the influence on the classification of localized surface contamination. We noticed that inner surfaces, due to their cleaner and more homogeneous composition, generally produced lower prediction variance on classification results. To estimate the overall strength of each brick in a way that reflects conventional compressive testing, a majority-vote scheme was applied across all pixels. At the brick-level we reached an accuracy of 91 % for the outer-face classification and 98 % for the inner-surface classification. This aggregation improved brick-level classification accuracy, and for the few bricks that remained misclassified after majority voting, errors were mostly concentrated at the boundaries between adjacent classes, consistent with the continuous and overlapping nature of compressive-strength distributions.

To gain insight into the spectral regions driving the three-class classifier, Score-CAM [11] was applied as a post-hoc explainability tool. Score-CAM generates gradient-free saliency maps by weighting each activation map according to its forward-passing score on the target class, thus identifying which part of the input data (i.e. spectral channels) contribute most to a given prediction. A strong influence of the spectral regions ranging from 1200 nm to 1400 nm and from 2300 nm to 2500 nm was observed for all compressive strength classes (high confidence on the jet color scale of Fig. 3). Still, the contribution of that last range was more pronounced in the *high* compressive strength class. Furthermore, we identified an influence of the wavelengths around 1900 nm on the predictions for the *medium* class (Fig. 3). These findings can be physically interpreted: the regions around 1400 nm and 1900 nm correspond to known water absorption bands, where elevated moisture content, indicative



(a) Rendered RGB (false colors) (b) Classification

Figure 2: Pixelwise classification (low | medium | high compressive strength) of a outer face sample with a measured (destructively) high compressive strength value of 11.6 MPa

of higher open porosity, is associated with weaker bricks [7]. The prominence of the 2300 nm to 2500 nm range corresponds to the characteristic absorption features of clay minerals [13, 12], the abundance of which indicates the composition of the brick. Together, these results confirm that the CNN has learned physically meaningful features rather than spurious correlations.

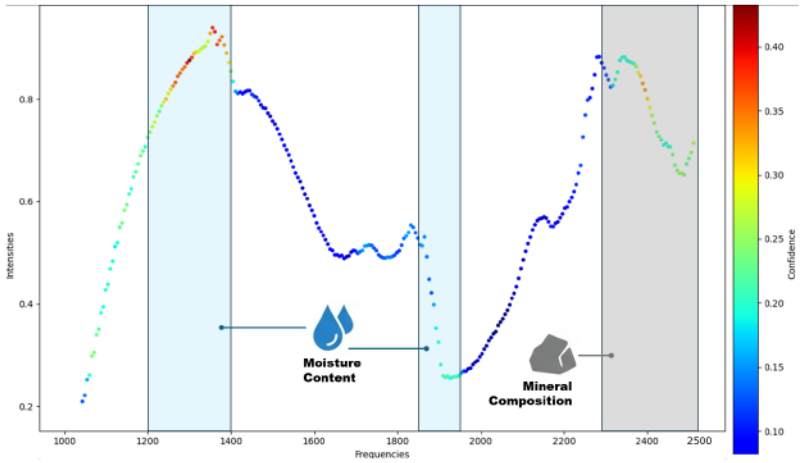


Figure 3: Average Score-CAM results for 100 spectra of the *medium* compressive strength class.

5 Conclusion and Outlook

We have demonstrated that SWIR HSI can capture mineralogical and moisture-related features and, combined with a spectral-spatial 1D-CNN, it enables the estimation of compressive strength in reclaimed bricks (regression: MAE 1.76 MPa, $R^2 = 0.625$). For classification into three compressive strength classes, a pixel-level accuracy of 83 % was reached, increasing to 91 % (outer) and 98 % (inner) at the brick level when using majority voting.

Compressive strength measurements provide a single bulk value for each brick. For this reason, majority voting was used to aggregate pixel-level predictions into a brick-level estimate that is comparable to the measured value. However, the measured compressive strength is also influenced by local instabilities, inclusions, and cracks. Such structural information is lost when reducing spatially resolved predictions to a single value through majority voting. Future work should therefore investigate how structural effects influence the overall compressive strength and develop more advanced classification schemes that integrate pixel-wise predictions into a structural assessment of the brick, providing an overall strength score. In addition, HSI is a surface inspection method and cannot detect structural defects located inside the brick. It therefore remains unclear whether surface information alone is sufficient to reliably predict compressive strength or whether complementary bulk measurements are required. This should be further investigated using techniques such as X-ray or CT imaging and acoustic testing to capture internal structural features.

Acknowledgments and Disclosure of Funding

The present work was funded by the Austrian Research Promotion Agency (FFG) through project KRAISBAU (48302986). The authors would furthermore like to thank University of Natural Resources and Life Sciences Vienna (BOKU) for extracting the brick samples and for carrying out the compressive strength measurements.

References

[1] D. Bakun-Mazor, Y. Ben-Ari, G. Notesko, S. Marco, and E. Ben-Dor. Measuring carbonate rock strength using spectroscopy across the optical and thermal region. *IOP Conference Series: Earth and Environmental Science*, 833(1):012025, August 2021.

[2] J. Cristobal Garcia, D. Caro, G. Foster, G. Pristera, F. Gallo, and D. Tonini. Techno-economic and environmental assessment of construction and demolition waste management in the European Union, 2024.

[3] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] M. He, Z. Zhang, J. Zhu, and N. Li. Correlation between the constant m_i of Hoek–Brown criterion and porosity of intact rock. *Rock Mechanics and Rock Engineering*, 55(2):923–936, February 2022.

[5] T.-H. Hsieh and J.-F. Kiang. Comparison of CNN algorithms on hyperspectral image classification in agricultural lands. *Sensors*, 20(6):1734, March 2020.

[6] G. Kereszturi, M. Heap, L. N. Schaefer, H. Darmawan, F. M. Deegan, B. Kennedy, J.-C. Komorowski, S. Mead, M. Rosas-Carbajal, A. Ryan, V. R. Troll, M. Villeneuve, and T. R. Walter. Porosity, strength, and alteration – Towards a new volcano stability assessment tool using VNIR-SWIR reflectance spectroscopy. *Earth and Planetary Science Letters*, 602:117929, January 2023.

[7] B. Koirala and P. Scheunders. An efficient method for water content estimation of building materials from spectral reflectance. *NDT and E International*, 147:103214, October 2024.

[8] F. Mousavi, E. Abdi, P. Fatehi, A. Ghalandarzadeh, H. A. Bahrami, B. Majnounian, and N. Ziadi. Rapid determination of soil unconfined compressive strength using reflectance spectroscopy. *Bulletin of Engineering Geology and the Environment*, 80(5):3923–3938, March 2021.

[9] M. Sadeghi, S. B. Jones, and W. D. Philpot. A linear physically-based model for remote sensing of soil moisture using short wave infrared bands. *Remote Sensing of Environment*, 164:66–76, July 2015.

[10] P. Stępień, E. Spychał, and K. Skowera. A comparative study on hygric properties and compressive strength of ceramic bricks. *Materials*, 15(21):7820, November 2022.

[11] H. Wang, Z. Wang, M. Du, F. Yang, Z. Zhang, S. Ding, P. Mardziel, and X. Hu. Score-CAM: score-weighted visual explanations for convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 111–119, 2020.

[12] E. C. Wellman, D. Riley, A. Hughes, N. Risso, M. Momayez, and J. Kemeny. A proposed concept for classifying uniaxial compressive strength (UCS) from SWIR hyperspectral data. *Engineering Geology*, 356:108300, September 2025.

[13] F. A. Yitagesu, F. van der Meer, H. van der Werff, and W. Zigterman. Quantifying engineering parameters of expansive soils from their reflectance spectra. *Engineering Geology*, 105(3–4):151–160, May 2009.

[14] X.-L. Zhang, F. Zhang, Y.-Z. Wang, Z.-G. Tao, and X.-Y. Zhang. Strength prediction model for water-bearing sandstone based on near-infrared spectroscopy. *Journal of Mountain Science*, 20(8):2388–2404, August 2023.

GraspGen+HSR: Adapting Simulation-Trained 6-DoF Grasping to Real Service Robots Without Retraining

Alexander Dvorak[*], Michael Nowak[*], Tessa Pulli, Markus Vincze

All authors are with the Automation and Control Institute, TU Wien, Vienna, Austria.
Emails: {e11912029, e12002155}@student.tuwien.ac.at; {pulli, vincze}@acin.tuwien.ac.at

Abstract

Recent diffusion-based 6-DoF grasp generation methods like GraspGen achieve state-of-the-art performance in simulation but face significant challenges when deployed on real robotic platforms. We present a unified adaptation pipeline for the Toyota Human Support Robot (HSR) that bridges these gaps without retraining the foundation model. Our approach combines symmetry-based point cloud completion to mitigate self-occlusion artifacts, three geometric feasibility filters that reduce motion planning failures from 66 % to 16 %, and a kinematic compensation for the HSR's arc-shaped gripper trajectory. We show in our experiments, that our pipeline achieves an overall success rate of 85 % which is competitive with simulation of GraspGen while outperforming baselines M2T2 (56 %) and AnyGrasp (70 %) by up to 29 percentage points. Ablation studies confirm the necessity of each component: symmetry completion improves success by +13 percentage points, while geometric filtering enables 4× more grasp candidates to reach execution. These results demonstrate that post-hoc adaptations can unlock the real-world potential of simulation-trained grasping foundation models on diverse hardware platforms. The code and repository are available at: https://github.com/Ziegenschmuggler/GraspGenforHSR

1 Introduction

Recent advances in 6-DoF grasp generation have enabled robots to predict diverse, stable grasps directly from single-view RGB-D observations [1], [2], [3], [4]. Diffusion-based methods such as GraspGen [1] achieve outstanding results in simulation. However, transferring these simulation-trained models to real-world robotic platforms remains challenging due to three primary factors: (1) hardware-specific kinematics that deviate from the parallel-jaw grippers assumed during training [5], (2) incomplete point clouds from single-view perception that cause grasp predictions on artificial boundaries [6], and (3) execution constraints that prevent motion planners from reaching many grasps that are geometrically feasible [7].

We address these challenges on the Toyota Human Support Robot (HSR) [8], a representative service robot platform with non-standard arc-shaped gripper kinematics. Our study shows that the simulation-trained GraspGen foundation model can be adapted to real hardware without retraining through a unified pipeline that implements symmetry-based point cloud completion to mitigate single-view self-occlusion, three geometric feasibility filters (plane distance, approach-from-below, approach-from-behind) to prune non-executable grasps before motion planning, and analytical kinematic compensation mapping GraspGen's parallel-jaw poses to the HSR's arc trajectory. These results demonstrate that a simulation-trained foundation model, combined with geometric constraints and platform-specific adaptations, achieves competitive real-world performance without requiring

[*]These authors contributed equally to the work.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

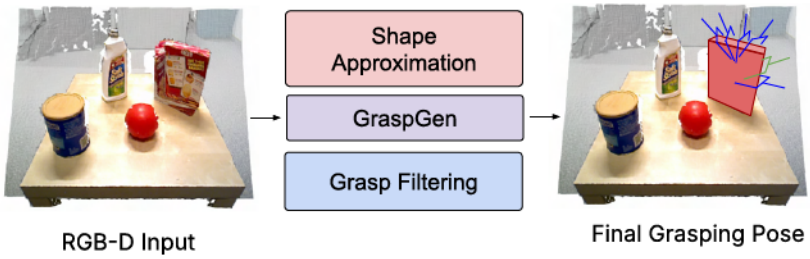


Figure 1: Overview of GraspGen+HSR: Instead of retraining, we combine GraspGen [1] with a symmetric shape approximation and grasp filtering to advance grasping performance.

retraining or collecting real-world data. The consistent success across isolated objects, cluttered scenes, and shelf grasps highlights the robustness of our unified adaptation strategy.

In summary, our contributions are the following:

- Demonstration that a simulation-trained foundation model, combined with geometric constraints and platform-specific adaptations, achieves competitive real-world performance without retraining.
- A ROS-based [9] integration of the GraspGen framework that bridges the gap between learned grasp generation and real-world execution on the HSR platform
- A symmetry-based shape completion method that creates pseudo-volumetric object representations from single-view point clouds, preventing edge grasps while preserving concave regions.
- Three lightweight filters that prune non-executable grasps prior to motion planning, reducing planning failures from two-thirds to 16 %.
- A kinematic compensation for the HSR's arc-shaped gripper trajectory, enabling direct use of GraspGen poses trained for parallel-jaw grippers.

In the following, we review related work on diffusion-based 6-DoF grasp generation and shape completion (Section 2), present our unified adaptation pipeline (Section 3), and evaluate its real-world performance on the Toyota HSR (Section 4).

2 Related Work

In this section, we review diffusion-based 6-DoF grasp generation and shape completion methods for robotic manipulation.

2.1 Diffusion-based 6-DoF grasp generation

Modern 6-DoF grasping is typically framed as generating and scoring grasp poses directly in SE(3) from 3D observations such as point clouds or depth data [10], [11], [1]. Earlier work explored autoregressive models [12] and variational autoencoders [10] to sample grasp candidates and then rank them with a learned critic, resulting in substantial improvements in diversity and success rates compared to purely analytical approaches. More recent methods introduce diffusion-based generators and combine them with discriminators that evaluate sampled poses, which have proven effective for cluttered scenes and across different object shapes [1], [2], [3], [4].

GraspGen [1], follows this line of work and combines a diffusion transformer with an on-generator discriminator, trained entirely in simulation on a large multi-gripper dataset, to achieve strong 6-DoF grasping performance across different embodiments, levels of observability, and scene complexity.

While such models provide high-quality grasps in simulation and for standard parallel-jaw grippers, they typically assume ideal point clouds and do not explicitly encode platform-specific execution constraints [1], [13]. In our work, we adopt GraspGen as the fixed generative backbone and focus on

adapting its predictions to the Toyota HSR without any retraining by adding perception, kinematic, and execution layers that are tailored to the real robot.

2.2 Perception: single-view point clouds and completion

Most 6-DoF grasp networks operate on 3D information derived either from voxel grids, implicit surfaces, or point clouds [1], [14], [15]. In practical setups, especially for mobile manipulators, grasping must often be performed from single-view RGB-D observations, which leads to partial and self-occluded object point clouds [6]. This incompleteness causes grasp generators to place contact points in inefficient locations, leading to grasping failures. To mitigate these issues, shape completion methods have been proposed that reconstruct full meshes or volumetric occupancy from partial inputs before grasp planning [16]. These methods can significantly improve planning robustness but usually require additional training data and heavy inference, which complicates deployment on resource-constrained platforms [17]. As an alternative, we introduce light-weight geometric priors to approximate the object shape. A Symmetry-based point cloud augmentation around the object centroid creates a pseudo-volumetric shell that fills in occluded geometry without requiring a learned completion network. Building on this insight, our pipeline uses a symmetry-based point cloud augmentation that is explicitly designed as a lightweight front-end to GraspGen, preserving graspable concavities while avoiding the cost of full shape completion.

3 Method

Fig. 2 illustrates our unified pipeline that adapts the pre-trained GraspGen [1] model to the HSR [8] without retraining. The system processes single-view RGB-D observations through four sequential stages: perception augmentation, grasp generation, geometric feasibility filtering, and final pose selection with kinematic compensation, followed by MoveIt! [18] motion planning.

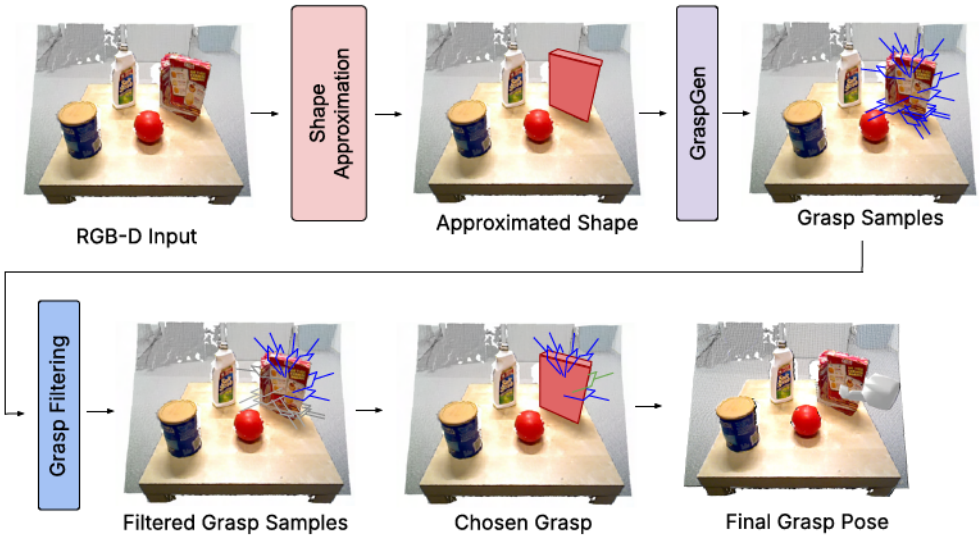
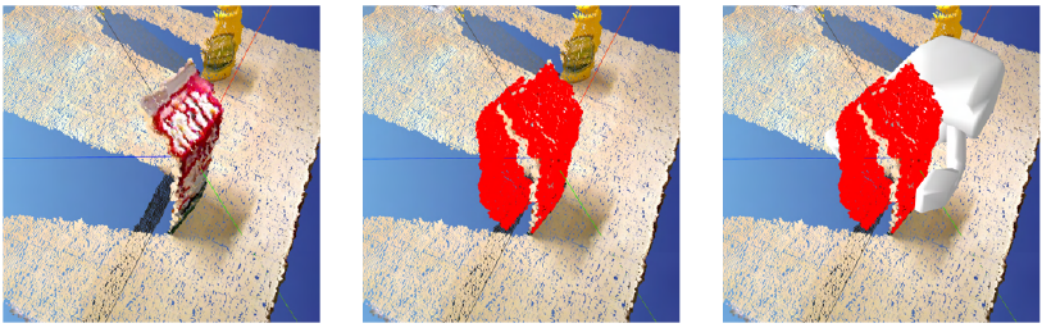


Figure 2: Single-view RGB-D input is processed through symmetry-based shape approximation to create complete object geometry for GraspGen inference. Generated 6-DoF grasp candidates undergo geometric feasibility filtering to remove non-executable poses. The highest-confidence filtered grasp receives HSR-specific kinematic compensation before MoveIt! execution.

3.1 Symmetry-Based Point Cloud Augmentation

Single-view RGB-D data produces incomplete point clouds due to self-occlusion. We create a pseudo-complete object representation by reflecting visible points around their centroid and shifting the reflected points behind the visible surface along the ray directions of the original points, see Fig. 3.

This shape approximation forms a volumetric shell that prevents GraspGen from predicting grasps on artificial depth discontinuities while preserving concave regions suitable for stable grasping.



(a) The original point cloud in real coloring with the blue ray being the view ray direction.

(b) The fully augmented point cloud in red with the symmetry-based added part.

(c) The chosen grasp based on the augmented point cloud.

Figure 3: Visualization of the symmetry-based point cloud augmentation process for creating pseudo-volumetric shells. The mirrored points are shifted by the semi-heuristic parameters s along the ray directions of each corresponding point p in the original point cloud. The view ray of the camera is the blue line visible in all images.

Formally, given a partial point cloud $P_{obs} = \{p_1, \ldots, p_n\}$ captured from a single viewpoint, we compute its centroid as:

$$c = \frac{1}{n} \sum_{i=1}^{n} p_i \tag{1}$$

We generate an augmented cloud P_{aug} by reflecting P_{obs} across its centroid c. To ensure the completed hull does not violate free-space constraints or overlap with the visible surface, we apply shifts s along the ray direction of each point. The augmented points $p' \in P_{aug}$ are defined as:

$$P_{aug} = \{p'_i \mid p'_i = 2c - p_i + s_i, \forall p_i \in P_{obs}\} \tag{2}$$

where s_i are computed as:

$$s_i = \frac{p_i}{||p_i||} \cdot \left(d + \frac{1}{n} \sum_{i=1}^{n} ||p_i - c|| \right), \tag{3}$$

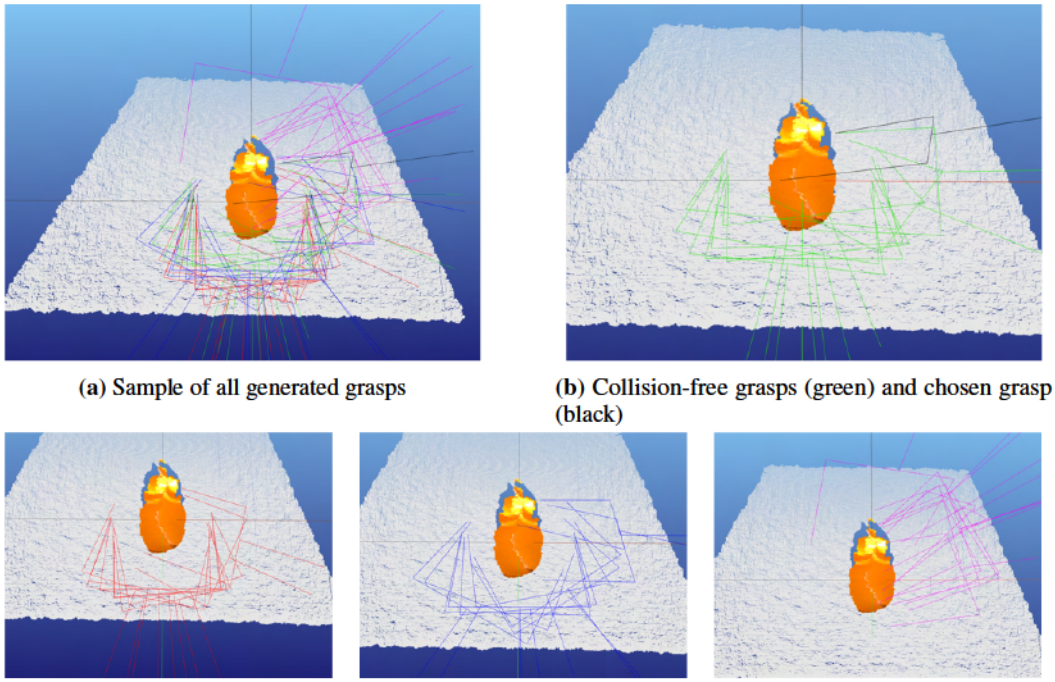
with d being a heuristic safety offset set to 5 mm ensuring that P_{aug} lies strictly behind the visible surface P_{obs}, effectively creating a pseudo-volumetric representation for more stable grasp prediction.

As symmetry assumptions are applied, the augmentation process is only reliable if the object of interest is also mostly symmetrical. Asymmetry to a certain degree is acceptable, as shown in Fig. 3a and Fig. 3b. The milk box has an asymmetric gable-top design where the cap is located. This part is mirrored to the backside bottom of the pseudo-volumetric shell and is therefore irrelevant for grasping as grippers can not get that close to the supporting surface.

3.2 Geometric Grasp Filtering

From the set of collision-free grasp candidates $G = \{g_i\}$ generated by GraspGen (Fig. 4a), we apply three lightweight geometric filters in the camera frame to discard poses that are likely to fail during execution. Each grasp g_i is represented by its position $\vec{t}_i$ and an approach vector $\vec{a}_i$ (the gripper $-z$-axis) derived from the grasp's rotation matrix $\mathbf{R_i}$. The support surface is expected to be the dominant plane and its normal $\vec{p}$ and normal distance d are estimated from the scene point cloud using RANSAC [19]. The effect of each filter is visualized in Fig. 4c–4e.

1) Distance to table: Grasps positioned in proximity to the supporting plane often result in gripper-table collisions. As all points $\vec{r}$ in the dominant plane fulfill the plane equation $\vec{r} \cdot \vec{p} = d$, the normal



(a) Sample of all generated grasps

(b) Collision-free grasps (green) and chosen grasp (black)

(c) Filtered grasps in proximity to the table

(d) Filtered grasps approaching from below

(e) Filtered grasps approaching from behind

Figure 4: Visualization of geometric grasp filtering.

distance δ between grasp position vectors and the plane can be calculated as:

$$\delta = \vec{t} \cdot \vec{p} - d \tag{4}$$

If the distance δ is lower than a given threshold Δ, the grasp is filtered out as shown in Fig. 4c. Typical threshold values Δ would be in the range of a few centimeters, depending on the size and form of the gripper.

2) Approach from below: Grasps with an upward approach vector require the gripper-joint to be positioned lower than the gripper itself, which might cause collisions with the support surface. The gripper's approach vector is defined as $\vec{a} = \mathbf{R}[0:3, 2]$. To quantify this orientation, we compute the alignment:

$$\gamma = \vec{a} \cdot \vec{p} \tag{5}$$

As shown in Fig. 4d, assuming the plane normal points upwards, candidates are pruned if they are above a given threshold $\gamma > \Gamma$, eliminating the risk of table collisions. As γ lies in the range $[-1, 1]$, a threshold of $\Gamma = 0$ would effectively remove all grasps pointing upwards. Slightly higher thresholds may be applicable if the gripper has a slim design. The lower the threshold, the more rigorous is also the exclusion of downward approaching grasps.

3) Approach from behind: To avoid trajectories that require the robot to move around the table or approach the object from the far side, we filter grasps that come from behind relative to the robot base (Fig. 4e). The robot approach direction is calculated as $\vec{r}_a = -\vec{e}_x \times \vec{p}$, assuming $\vec{e}_x$ points to the right in the camera frame. All grasps coming from behind fulfill the condition below and are filtered out, as shown in Fig. 4e.

$$\vec{r}_a \cdot \vec{a} < 0 \tag{6}$$

Applying these three constraints yields the geometrically feasible subset $G_{\text{feas}} \subset G$, which still covers diverse approach directions but significantly reduces non-executable candidates. The final selection among the remaining collision-free, feasible grasps is illustrated in Fig. 4b.

3.3 Kinematic Compensation

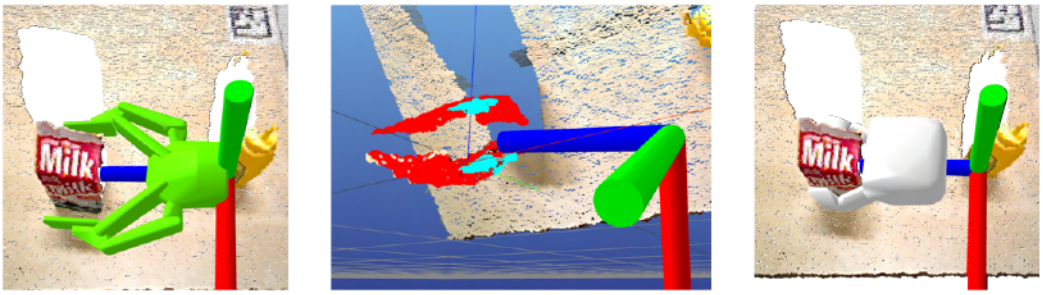
The GraspGen model predicts a grasp pose $T_{grasp} \in SE(3)$ assuming a fixed Tool Center Point (TCP). However, the HSR's arc-shaped fingers cause the physical contact point to shift along the local z-axis as a function of the gripper aperture w. To align the predicted pose with the physical hardware, we define the corrected execution pose T_{exec} as:

$$T_{exec} = T_{grasp} \cdot \text{Trans}(0, 0, \Delta z(w)) \tag{7}$$

The compensation value $\Delta z(w)$ is derived from the kinematic linkage of the HSR gripper using trigonometric relations:

$$\Delta z(w) = L \cdot \left(1 - \sqrt{1 - \left(\frac{w}{2L}\right)^2}\right) \tag{8}$$

where L represents the distance between the gripper-base and the fingertips. This transformation ensures that the finger pads align precisely with the object surface, regardless of its width, preventing collisions or shallow grasps caused by the arc-shaped closing trajectory. A comparison of the gripper kinematics and a visualization of the object's width estimation is shown in Fig. 5.



(a) Robotiq 2F-140 parallel-jaw kinematics assumed by GraspGen.

(b) Width estimation using the augmented point cloud.

(c) HSR gripper kinematics with an arc-shaped closing trajectory.

Figure 5: Gripper compensation: As the deployed gripper on the HSR has an arc-shaped closing trajectory, the TCP needs to be adjusted based on the estimated width, which defines the gripper's closing aperture w.

4 Experiments

We evaluate our integrated pipeline on a Toyota HSR in real-world grasping experiments. The experiments are designed to assess the impact of kinematic compensation, symmetry-based point cloud completion, and geometric grasp filtering across different scene configurations.

4.1 Experimental Setup

The HSR is equipped with a head-mounted RGB-D sensor used for object perception. We employ Grounded SAM [20] for instance segmentation of target objects and MoveIt! [18] for motion planning. Our test suite consists of 8 diverse household objects (e.g., bottles, cans, boxes, and small tools) spanning different geometric classes, including cylindrical, box-like, and thin elongated shapes. A *successful grasp* is defined as the robot closing the gripper on the target object, lifting it from the support surface, and maintaining a stable hold during a short transport motion.

4.2 Real Robot Experiments

We compare our full GraspGen+HSR pipeline against the unmodified GraspGen model, M2T2 [3], and AnyGrasp [4]. For the baselines, we use the official weights and default configurations as released by the authors.

We evaluate three settings that progressively increase task difficulty: grasping isolated objects on a clear tabletop, grasping a defined object from cluttered tabletop scenes, and grasping an object

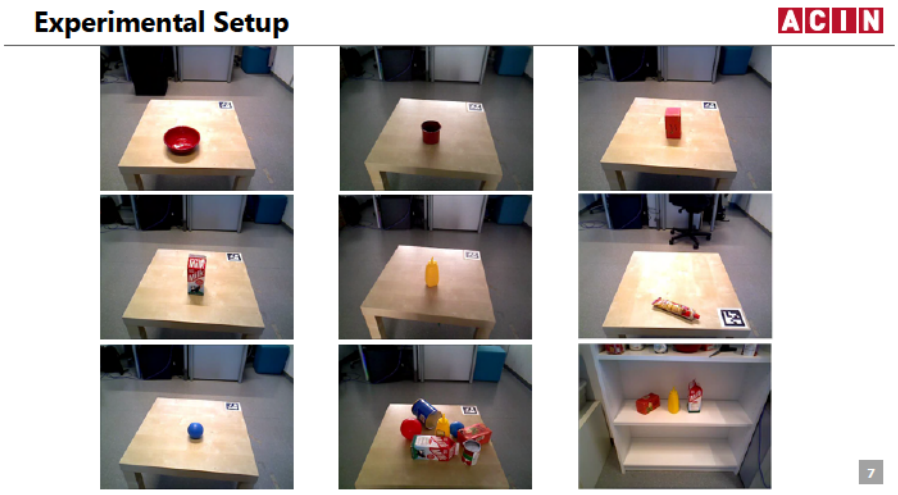


Figure 6: Experimental Setup for some single standing objects, the cluttered scene and the shelf test. The objects are from left to right and top to bottom labeled as: Bowl, Mug, Tea Box, Milk, Mustard, Tomato Tube, Blue Ball. The Small Cylinder can be seen in the cluttered scene (bottom middle) at the bottom right.

standing in a shelf compartment (Fig. 6). All isolated objects are grasped 10 times, resulting in a total of 80 trials. In the cluttered scenes and the shelf tests, the mustard bottle is always the object of interest, again being grasped 10 times in both settings. Experimental results are summarized in Tab. 1. While GraspGen slightly edges out our method on isolated objects (91 % vs. 86 %), our full pipeline clearly excels in clutter (90 % vs. 83 %) due to symmetry-based shape completion and in shelf scenarios (80 % vs. 72 %) thanks to geometric feasibility filtering.

These results show that a simulation-trained foundation model, combined with geometric constraints (approach filtering, plane distance) and platform-specific adaptations (kinematic compensation, symmetry completion), can achieve competitive real-world performance without the need for retraining or real-world data collection. The consistent success across diverse scenarios highlights the robustness of our unified adaptation strategy.

Table 1: Comparison with recent 6-DoF grasping methods across three scenarios. Our GraspGen+HSR pipeline significantly outperforms baselines, particularly in clutter and constrained shelf environments. Experimental results for GraspGen, M2T2 and AnyGrasp have been reported in [1].

Method	**Isolated**	**Cluttered Table**	**Shelf**
GraspGen+HSR (Ours)	86 %	**90 %**	**80 %**
GraspGen [1]	**91 %**	83 %	72 %
M2T2 [3]	81 %	75 %	14 %
AnyGrasp [4]	86 %	83 %	43 %

4.3 Ablation Study: Geometric Filtering

To quantify the effect of the geometric feasibility filters, we compare our full pipeline against a baseline that uses GraspGen generation and scene collision checking only, without the additional plane-distance and approach-direction constraints. In the baseline, the highest-confidence grasp is rejected by the motion planner as infeasible in 66% of trials, leading to long planning times and frequent failures. In Fig. 7, this metric is reported as the planning failure rate, defined as the fraction of highest-scoring grasps rejected by the planner. With our geometric filtering in place, this planning failure rate is reduced to 16 % (see Fig. 7), which significantly decreases the latency between perception and execution and increases the number of trials that reach the execution phase.

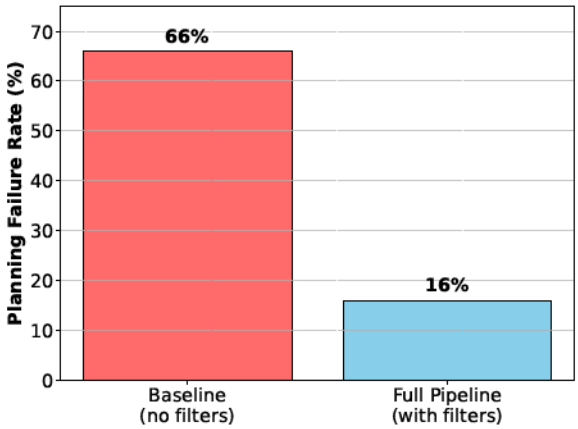


Figure 7: Effect of Geometric Filtering on Motion Planning

4.4 Grasping Performance

We conducted 100 grasp attempts (isolated objects + cluttered scenes + shelf tests) that systematically test each component of our pipeline. In isolated-object scenarios on clear tabletops, we evaluate the accuracy of our kinematic compensation by measuring grasp stability and lift success for single objects without occlusions. Cluttered scenes assess the effectiveness of symmetry-based point cloud completion under partial self-occlusion, where GraspGen would otherwise predict grasps on incomplete object boundaries. Shelf grasping scenarios challenge the geometric approach-direction filters by requiring precise top-down trajectories in vertically constrained environments.

The complete pipeline achieves an overall success rate of 86 % across all conditions. Table 2 reports per-category performance, revealing that cubic objects (Milk) benefit most from our symmetry completion. Transparent and metallic items remain challenging due to inherent perception limitations. Object detection from RGB-D sensor data with Grounded SAM [20] yielded poor point clouds, which inevitably led to poor grasping results and frequent planning failures. This is why they are not included in Tab. 2. Fig. 7 visualizes the reduction in motion planning failures enabled by our geometric filters.

Object Category	w/o Symmetry Expansion	w/ Symmetry Expansion
Bowl	80 %	90 %
Milk	50 %	90 %
Blue Ball	80 %	100 %
Mug	70 %	80 %
Mustard	100 %	90 %
Small Cylinder	80 %	80 %
Tea Box	80 %	90 %
Tomato Tube	90 %	70 %
Cluttered Scene	20 %	90 %
Shelf Tests	80 %	80 %
Total	**73 %**	**86 %**

Table 2: Grasping Success Rate Across Different Object Categories. Details about setup, shape, and size can be seen in Fig. 6

5 Conclusion

We presented a unified pipeline adapting simulation-trained GraspGen to the Toyota HSR without retraining. Our experiments demonstrate that post-hoc adaptations can unlock foundation grasping models for diverse service robots, avoiding costly retraining. Future work will extend the pipeline to a language-guided zero-shot pick-and-place method.

ACKNOWLEDGMENT

We gratefully acknowledge the support of the EU-program EC Horizon 2020 for Research and Innovation under project No. I 6114, project iChores and the EU-program EC Horizon 2020 for Research and Innovation under grant agreement No. 101017089, project TraceBot.

Use of LLMs

During the preparation of this work, the authors used ChatGPT, Google Gemini, and Perplexity to improve the language and readability of the manuscript and to assist in writing code for visualizing experimental results. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

References

[1] A. Murali et al., *GraspGen: A Diffusion-based Framework for 6-DoF Grasping with On-Generator Training*, 2025. arXiv: 2507.13097 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2507.13097.

[2] J. Urain, N. Funk, J. Peters, and G. Chalvatzaki, *SE(3)-DiffusionFields: Learning smooth cost functions for joint grasp and motion optimization through diffusion*, 2023. arXiv: 2209.03855 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2209.03855.

[3] W. Yuan, A. Murali, A. Mousavian, and D. Fox, "M2T2: Multi-Task Masked Transformer for Object-centric Pick and Place," in *7th Conference on Robot Learning*, vol. 229, PMLR, 2023, pp. 3619–3630. [Online]. Available: https://proceedings.mlr.press/v229/yuan23a.html.

[4] H.-S. Fang et al., "AnyGrasp: Robust and Efficient Grasp Perception in Spatial and Temporal Domains," *IEEE Transactions on Robotics*, vol. 39, no. 5, pp. 3929–3945, 2023. DOI: 10.1109/TRO.2023.3281153.

[5] C. M. Kim, M. Danielczuk, I. Huang, and K. Goldberg, "IPC-GraspSim: Reducing the Sim2Real Gap for Parallel-Jaw Grasping with the Incremental Potential Contact Model," in *ICRA*, IEEE, May 2022, pp. 6180–6187. DOI: 10.1109/ICRA46639.2022.9811777.

[6] Y.-K. Wang, C. Xing, Y.-L. Wei, X.-M. Wu, and W.-S. Zheng, "Single-View Scene Point Cloud Human Grasp Generation," in *IEEE/CVF CVPR*, 2024, pp. 831–841. DOI: 10.1109/CVPR52733.2024.00085.

[7] A. Murali, A. Mousavian, C. Eppner, C. Paxton, and D. Fox, *6-DoF Grasping for Target-driven Object Manipulation in Clutter*, 2020. arXiv: 1912.03628 [cs.RO]. [Online]. Available: https://arxiv.org/abs/1912.03628.

[8] T. Yamamoto, K. Terada, A. Ochiai, F. Saito, Y. Asahara, and K. Murase, "Development of Human Support Robot as the Research Platform of a Domestic Mobile Manipulator," in *IEEE/RSJ IROS*, 2018, pp. 1–9. DOI: 10.1109/IROS.2018.8594344.

[9] M. Quigley et al., "ROS: An open-source Robot Operating System," in *ICRA Workshop on Open Source Software*, vol. 3, Jan. 2009.

[10] A. Mousavian, C. Eppner, and D. Fox, *6-DoF GraspNet: Variational Grasp Generation for Object Manipulation*, 2019. arXiv: 1905.10520 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1905.10520.

[11] H. Liang et al., "PointNetGPD: Detecting Grasp Configurations from Point Sets," in *2019 ICRA*, IEEE, May 2019, pp. 3629–3635. DOI: 10.1109/icra.2019.8794435. [Online]. Available: http://dx.doi.org/10.1109/ICRA.2019.8794435.

[12] J. Tobin et al., "Domain Randomization and Generative Models for Robotic Grasping," in *IEEE/RSJ IROS*, 2018, pp. 3482–3489. DOI: 10.1109/IROS.2018.8593933.

[13] B. Han, M. Parakh, D. Geng, J. A. Defay, G. Luyang, and J. Deng, *FetchBench: A Simulation Benchmark for Robot Fetching*, 2024. arXiv: 2406.11793 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2406.11793.

[14] T. G. W. Lum et al., *Get a Grip: Multi-Finger Grasp Evaluation at Scale Enables Robust Sim-to-Real Transfer*, 2024. arXiv: 2410.23701 [cs.RO]. [Online]. Available: https://arxiv.org/abs/2410.23701.

[15] M. Breyer, J. J. Chung, L. Ott, R. Siegwart, and J. Nieto, "Volumetric Grasping Network: Real-time 6 DoF Grasp Detection in Clutter," in *Conference on Robot Learning*, PMLR, 2021, pp. 1602–1611.

[16] J. Varley, C. DeChant, A. Richardson, J. Ruales, and P. Allen, "Shape Completion Enabled Robotic Grasping," in *IROS*, IEEE, 2017, pp. 2442–2447.

[17] S. S. Mohammadi et al., "3DSGrasp: 3D Shape-Completion for Robotic Grasp," in *ICRA*, 2023, pp. 3815–3822. DOI: 10.1109/ICRA48891.2023.10160350.

[18] M. Görner, R. Haschke, H. Ritter, and J. Zhang, "MoveIt! Task Constructor for Task-Level Motion Planning," in *ICRA*, 2019, pp. 190–196. DOI: 10.1109/ICRA.2019.8793898.

[19] M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, no. 6, pp. 381–395, Jun. 1981. DOI: 10.1145/358669.358692.

[20] T. Ren et al., *Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks*, 2024. arXiv: 2401.14159 [cs.CV].

Fourier contrast optimization for occluded motion estimation

Ido Akov[1,2] **Roman Pflugfelder**[1,2]

Daniel Cremers[1]
[1]Technical University of Munich (TUM) [2]Austrian Institute of Technology (AIT)
{ido.akov, roman.pflugfelder, cremers}@tum.de

Abstract

Fragmented occlusion, as encountered in through-foliage observation, makes monocular motion estimation difficult because the target is visible only through sparse, discontinuous image fragments. We estimate motion by warping frames under a parametric model and maximizing the contrast of their integrated image. Although effective for 2DoF translation, this objective becomes ill-conditioned for 4DoF similarity motion. To analyze this, we derive a Fourier-domain reformulation that exposes the optimization structure and shows that static occlusion biases the objective toward zero motion. This motivates a decoupled 4DoF pipeline in which rotation and scale are estimated separately from translation. On synthetic videos with controlled fragmented occlusion, the Fourier formulation matches the spatial baseline at low-to-mid occlusion while converging faster, and the decoupled pipeline restores reliable translation recovery where joint 4DoF optimization fails.

1 Introduction

Fragmented occlusion [1] arises when an object is visible only through disconnected foreground gaps, as in observation through foliage, fences, or clutter. The visible evidence is sparse and spatially discontinuous, making correspondence-based motion estimation unreliable: classical optical flow and local registration methods [2–4] rely on brightness constancy and locally coherent support, both of which become fragile when the target never appears as a large connected region. We therefore consider monocular motion estimation from image sequences in which the object is never fully visible in any single frame, and motion must be inferred without reliable local matches.

An alternative is to estimate motion by warping and integrating observations under a low-dimensional parametric model. Variants of this principle appear in direct parametric alignment [5], synthetic-aperture reconstruction [6], and contrast-maximization methods for event cameras [7]. These approaches avoid explicit correspondences, but they do not directly address similarity motion under persistent fragmented occlusion in a monocular setting with a single global motion model.

Here we revisit contrast-based motion estimation under fragmented occlusion. The formulation works well for 2DoF translation, but becomes ill-conditioned for 4DoF similarity motion because repeated warp composition couples translation, rotation, and scale. We derive a Fourier-domain reformulation that makes the objective interpretable as a competition between moving-object alignment and static occlusion, and use this perspective to explain the instability of joint 4DoF optimization. Guided by this structure, we decouple the estimation: rotation and scale are first recovered by maximizing the same contrast objective in log-polar Fourier coordinates, and translation is then estimated separately using the same objective.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

2 Problem formulation

2.1 Spatial domain

Let $I_0, \ldots, I_{T-1}$ be a grayscale video sequence with $I_t \in \mathbb{R}^{M \times N}$, and let $W(\cdot, \theta)$ denote a parametric warp with motion parameters θ. We align each frame before temporal integration:

$$\bar{I}(\theta) = \frac{1}{T} \sum_{t=0}^{T-1} W^t(I_t, \theta), \tag{1}$$

where W^t denotes the t-fold application of the motion model. We then define

$$f_{\text{opt}}(I_{[0\ldots T-1]}, \theta) := \text{Var}\big(\bar{I}(\theta)\big), \qquad \theta^\star = \arg\max_\theta f_{\text{opt}}(I_{[0\ldots T-1]}, \theta). \tag{2}$$

Correct motion compensation sharpens the target in $\bar{I}(\theta)$, while static occluders and background become blurred by averaging. In the 2DoF setting, $\theta = [\tau_x, \tau_y]^\top$ is a coherent translation. In the 4DoF setting, we use a similarity model with translation, rotation, and isotropic scale.

2.2 Fourier domain

We first consider the translation case. By the Fourier shift theorem and linearity, the integrated image becomes

$$\mathcal{F}\{\bar{I}(\theta)\} = \frac{1}{T} \sum_{t=0}^{T-1} e^{-j\omega t \phi(\theta)}\, \mathcal{F}\{I_t\}. \tag{3}$$

where

$$\phi(\theta) := \frac{\tau_x}{M} + \frac{\tau_y}{N}. \tag{4}$$

We model each frame as the sum of a moving component and a static occlusion component,

$$I_t = W^t(I_{\text{mov}}, \theta^\star) + I_{\text{occ}}, \tag{5}$$

where I_{mov} denotes the target appearance, I_{occ} the static occluder, and $\theta^\star$ the true motion. Substituting into (3) yields

$$\mathcal{F}\{\bar{I}(\theta)\} = H(\Delta\phi)\, \mathcal{F}\{I_{\text{mov}}\} + H(\phi)\, \mathcal{F}\{I_{\text{occ}}\}, \tag{6}$$

with

$$\Delta\phi := \phi(\theta) - \phi(\theta^\star), \qquad H(\psi) := \frac{1 - e^{-j\omega T \psi}}{T(1 - e^{-j\omega\psi})}. \tag{7}$$

Since the variance objective equals non-DC Fourier energy, we obtain

$$f_{\text{opt}}(I_{[0\ldots T-1]}, \theta) \overset{\text{DFT}}{\leftrightarrow} \sum_{(m,n)\neq(0,0)} \left| H(\Delta\phi)\, \mathcal{F}\{I_{\text{mov}}\} + H(\phi)\, \mathcal{F}\{I_{\text{occ}}\} \right|^2. \tag{8}$$

This makes the optimization structure explicit: the moving-object term is maximized at the true motion, whereas the static-occlusion term is maximized at zero motion and therefore biases the objective accordingly. In higher-DoF settings, additional motion parameters further weaken the coherence of the moving term, exacerbating this competition and destabilizing joint optimization.

3 Motion decoupling for 4DoF

The Fourier formulation suggests a natural way to address this instability. In similarity motion, translation is encoded in Fourier phase, while rotation and scale are encoded in Fourier magnitude; after log-polar remapping, the latter become translations in log-polar coordinates [8, 9]. This yields a two-stage pipeline. In phase 1, each frame is mapped to Fourier magnitude, remapped to log-polar coordinates, and rotation and scale are estimated by maximizing the same contrast objective over the integrated representations. In phase 2, frames are aligned using the recovered rotation and scale, and the residual translation is estimated by maximizing f_{opt} in the image or Fourier domain. This separation prevents rotation and scale errors from being absorbed as translation drift and stabilizes optimization.

4 Experiments

4.1 Experimental setup

We evaluate the proposed methods on synthetic grayscale videos of simple high-contrast geometric shapes undergoing coherent motion. Each sequence contains $T = 8$ frames of size 128×128. This setup isolates the effect of fragmented occlusion under controlled variation in occlusion density. Fragmented occlusion is simulated by static structured masks, and occlusion density is defined as the fraction of pixels within the motion support covered by the mask. All methods are optimized with Adam, learning rate 0.1, and a fixed budget of 200 iterations for 2DoF and for each phase of the 4DoF pipeline.

In the 2DoF setting, motion is restricted to coherent translation and we compare the original spatial-domain contrast objective with its Fourier-domain reformulation. In the 4DoF setting, we compare direct joint optimization against the proposed decoupled pipeline.

4.2 Evaluation metrics

We use translation endpoint error (EPE) as the primary metric:

$$\mathrm{EPE} = \|\hat{\boldsymbol{\tau}} - \boldsymbol{\tau}^\star\|_2. \tag{9}$$

We treat optimization on a single video-motion instance as successful once the recovered translation satisfies $\mathrm{EPE} < 0.5$. We additionally report median time-to-threshold (TTT), i.e. the number of optimization steps required to first satisfy this criterion. In the 4DoF setting, our primary goal is not precise estimation of rotation and scale in isolation, but successful motion decoupling: we evaluate whether the recovered rotation and scale are sufficient to remove drift and restore accurate translation recovery.

4.3 Results

Figure 1 summarizes the main empirical findings. In the 2DoF setting, the Fourier-domain objective closely matches the spatial objective at low occlusion densities, confirming that the reformulation preserves the behavior of the original contrast objective in the translation regime. Over this low-to-mid density range, the Fourier formulation reaches the EPE threshold in fewer iterations and exhibits a flatter time-to-threshold profile. At moderate and heavy occlusion, however, the spatial objective is consistently more robust in success rate, and both methods eventually break down.

The 4DoF experiment reveals a qualitatively different phenomenon. Direct joint optimization of translation, rotation, and scale fails already in the unoccluded case, indicating that the difficulty is not caused by occlusion alone but by the structure of the joint objective itself. This is consistent with the analysis in the previous sections: repeated warp composition couples the motion parameters, so that errors in rotation and scale are absorbed as translation drift. In contrast, the proposed decoupled pipeline succeeds reliably at low occlusion and degrades gradually as density increases. The gap is therefore structural rather than purely quantitative.

5 Conclusion

We introduced a Fourier-domain reformulation of a contrast-maximization objective for motion estimation under fragmented occlusion. The analysis reveals a competition between moving-object alignment and static occlusion, which increasingly biases the objective and leads to instability in joint 4DoF optimization.

This perspective leads to a decoupled estimation strategy in which rotation and scale are recovered separately from translation. Experiments confirm that, at low-to-mid occlusion densities, the Fourier formulation retains the behavior of the spatial formulation in the translation regime while improving convergence, and that decoupling resolves the failure of joint optimization in 4DoF settings.

Future work will extend the analysis to real through-foliage and other naturally occluded sequences to assess robustness beyond synthetic settings.

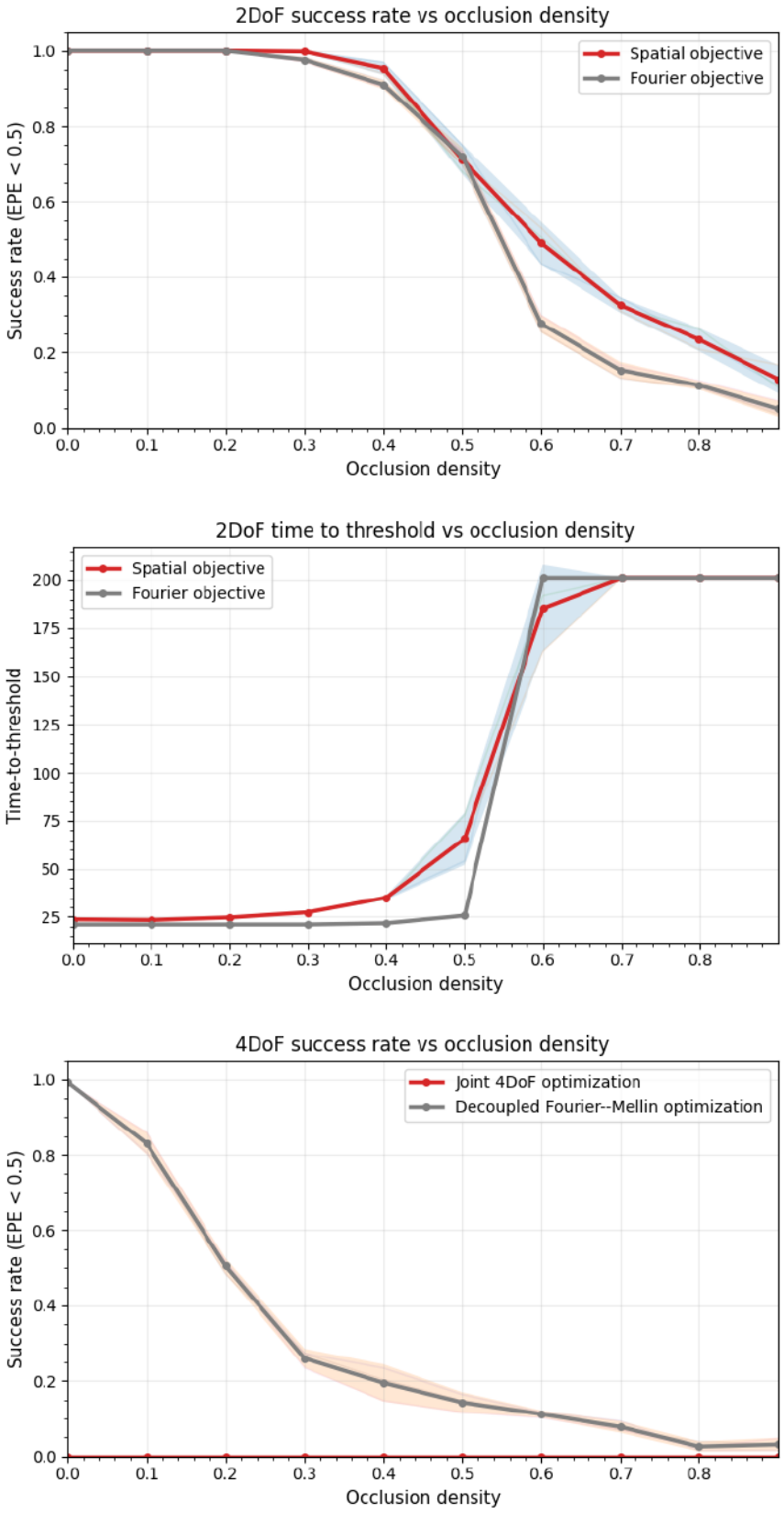


Figure 1: Optimization under fragmented occlusion. Top: 2DoF translation success rate versus occlusion density for spatial and Fourier objectives. Middle: corresponding median time-to-threshold (TTT). Bottom: 4DoF similarity-motion success rate versus occlusion density for joint optimization and the proposed decoupled pipeline.

Acknowledgements

This work received funding from the European Defence Fund under grant agreement EDF-2022-101121405-STORE.

References

[1] Julian Pegoraro and Roman Pflugfelder. The problem of fragmented occlusion in object detection, 2020. URL https://arxiv.org/abs/2004.13076.

[2] Bruce Lucas and Takeo Kanade. An iterative image registration technique with an application to stereo vision (ijcai). volume 81, 04 1981.

[3] Berthold K. P. Horn and Brian G. Schunck. Determining optical flow. *Artificial Intelligence*, 17 (1–3):185–203, 1981.

[4] Michael J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow fields. *Computer Vision and Image Understanding*, 63(1):75–104, 1996.

[5] Michal Irani, B. Rousso, and Shmuel Peleg. Detecting and tracking multiple moving objects using temporal integration. *European Conference on Computer Vision*, pages 282–287, 01 1992.

[6] Vaibhav Vaish, Richard Szeliski, C. Lawrence Zitnick, Sing Bing Kang, and Marc Levoy. Reconstructing occluded surfaces using synthetic apertures: Stereo, focus and robust measures. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2006.

[7] Guillermo Gallego, Henri Rebecq, and Davide Scaramuzza. A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, page 3867–3876. IEEE, June 2018. doi: 10.1109/cvpr.2018.00407. URL http://dx.doi.org/10.1109/CVPR.2018.00407.

[8] B. S. Reddy and B. N. Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *IEEE Transactions on Image Processing*, 5(8):1266–1271, 1996.

[9] Santosh Thoduka, Frederik Hegger, Gerhard K. Kraetzschmar, and Paul G. Plöger. Motion detection in the presence of egomotion using the fourier-mellin transform. In *RoboCup 2017: Robot World Cup XXI*, page 252–264, Berlin, Heidelberg, 2017. Springer-Verlag. ISBN 978-3-030-00307-4. doi: 10.1007/978-3-030-00308-1_21. URL https://doi.org/10.1007/978-3-030-00308-1_21.

Effect of polarization filters on hand vein sample image quality

Christof Kauba and Andreas Uhl
Department of Artifical Intelligence and Human Interfaces
University of Salzburg
5020 Salzburg, AUSTRIA
{ckauba,uhl}@cs.sbg.ac.at

Abstract

This is work about using (polarization) filters in hand-vein biometric recognition. Experiments clearly demonstrate that the respective application of linear polarization, circular polarization, and band pass filters on the capturing lens improve hand-vein sample image quality across a considerable range of specific vascular image quality metrics. In case the illumination source is additionally equipped with a linear polarization filter (in relative perpendicular direction), further quality improvement could not be demonstrated.

1 Introduction

Biometric authentication systems are well established today as they exhibit many advantages over traditional password and token based ones. The most prominent examples are fingerprint and face recognition systems. In recent times, authentication based on finger- and hand-veins gains more attention as it provides advantages over the well established fingerprint techniques (Uhl et al. [2019]). Over the past few years, pioneered by Fujitsu's PalmSecure palm vein authentication technology, major technology companies including Tencent (i.e. Tencent PalmAI – Weixin/WeChat Pay) and Amazon (i.e. Amazon One) have released palm vein scanning-based payment systems, which have been applied to various scenarios such as grocery shops, subway stations, and sports venues, reaching tens of millions of registered users (Kuang et al. [2024]).

Hand-vein recognition utilizes the pattern of the blood vessels inside the hand of a human, which is captured using near-infrared (NIR) illumination. The vein patterns are neither susceptible to abrasion nor to skin surface conditions. However, the vein images have low contrast and quality in general and the vein structure may be influenced by temperature, physical activity and certain injuries and diseases (Kirchgasser et al. [2019]). NIR illumination is the key to finger- and hand-vein recognition. The positioning of the light source with respect to the camera and the subject's finger or hand plays an important role. We distinguish between reflected light, where the light source and the camera are placed on the same side of the hand and the light is reflected from the skin surface and deeper tissue layers and trans-illumination, where the light source and the camera are located on the opposite side of the hand and the light passes through tissue.

One way to deal with the low quality and low contrast of hand-vein imagery is to consider polarized NIR light to better cope with reflections and scattering light. In two keynote talks given at the IEEE/IAPR International Joint Conferences on Biometrics (IJCB) in 2023 and 2024, the Amazon One team postulated the usage of polarized NIR in their system, while hardly any details have been disclosed. On the other hand, the WeChat Pay group published work in this direction: In (Sun et al. [2023]) they report on the usage of polarization in a traditional hand-vein imaging system (however, only providing visual examples on the effects of doing so) while in (Kuang et al. [2024]) they suggest a polarization-selective metalens, a highly specialized imaging system, for hand-vein acquisition.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

In this work, we extend a traditional hand-vein acquisition system (Kauba and Uhl [2018]) by employing a set of filters including circular and linear polarization filters. Instead of only considering visual examples as a "proof" of improved sample quality (as done in (Sun et al. [2023])), we apply a set of vascular image quality metrics to the data, which have been introduced to predict the suitability of sample data in a subsequent recognition process. In Section 2, we describe the experiments conducted in terms of the acquisition setup and the vascular image quality metrics employed. In Section 3, we present the results of applying the image quality metrics to the acquired data while providing conclusions and an outlook to further work in this direction in the Conclusion.

2 Experimental Setup

2.1 Hand-Vein Capturing Device

The hand-vein samples for the experiments were acquired using a custom built capturing device introduced earlier (Kauba and Uhl [2018]), with modified reflected light emitters. This capturing devices uses a NIR enhanced industrial camera, IDS Imaging UI-1240ML-NIR with a maximum resolution of 1280x1024 pixels together with a 9 mm wide-angle-lens. The used polarization filters can be screwed on the front of the lens. The device has two illumination sources, located at the top left and top right side wall, each targeting the wooden inside wall rather then directly targeting the hand surface. Each illumination source consists of two rows of 8 NIR-LEDs, one row with 850 nm LEDs, the other one with 950 nm LEDs – the former was used in data acquisition for the experiments. The brightness of the LEDs can be controlled to achieve an optimal contrast.

2.2 Polarization Filters

Several different filters, including one circular and three linear polarization filters, have been applied to the lens:

Lens without any filter: No filter was applied to the lens.
BN850: narrow band-pass filter with a center frequency of 850 nm.
PR032: linear polarisation filter from MIDOPT with a wavelength band of 400 - 700 nm.
PR120: linear polarisation filter (high contrast version) from MIDOPT with a wavelength band of 400 - 750 nm.
PR1000: linear polarisation filter (high contrast version) from MIDOPT with a wavelength band of 400 - 2000 nm.
PC052: circular polarisation filter from MIDOPT with a wavelength band of 450 - 650 nm.

In addition, the following filter has been applied to the illumination sources optionally:
PS1000: linear polarizing film from MIDOPT with a wavelength band of 400 - 2000 nm. In case of its application, the polarization direction is perpendicular to that of the lens mounted linear polarization filters. Fig. 1 illustrates the illumination and capturing principle with linear polarisation filters. Details and datasheets for the filters can be found at: https://midopt.com/filters/polarizing/

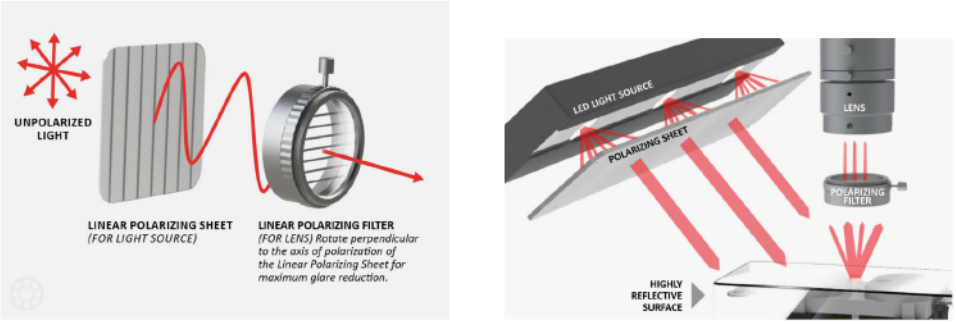


Figure 1: Linear polarisation filter applied to illumination source and camera lens. Pictures from: https://midopt.com/filters/polarizing/

2.3 Capturing Protocol

Hand-vein sample data from a single subject have been captured in different illumination and polarization filter settings, respectively:

- Outdoor (with direct sunlight or with cloudy sky)
- Indoor (with artificial ceiling light or with light shining through window only)
- Frontal lid closed or frontal lid open

The frontal lid of the capturing device is meant to protect the imaging process from incident light. Per default it is closed, but to study the effect of even more challenging acquisition, we have added settings with this frontal lid being removed (i.e. "frontal lid open"). For each of those settings, all of the aforementioned polarization filters have been applied. In addition, the polarization filter on the illumination source has been applied as well in selected settings. For each configuration, 5 dorsal and 5 palmar hand-vein samples have been acquired, and their quality values have been averaged for the experimental results provided. Figure 2 displays three dorsal example images as captured for the experiments, overall, roughly 1600 samples have been taken.

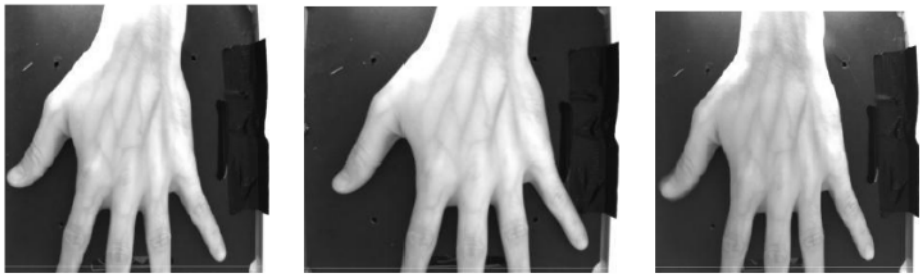


Figure 2: Example dorsal sample images captured with sensors PC052, PR032 and PR1000 (from left to right): Capturing in direct sunlight with lid of capturing box open.

2.4 Image Quality Metrics

To evaluate the impact of the polarization filters, the following hand-vein specific (all but the last) and general image quality metrics (NIQE) were employed (see Kirchgasser et al. [2025] for an overview):
Wang (Wang et al. [2017]): This palm-vein specific quality metric is based on clarity and brightness uniformity.
GCF (Matkovic et al. [2005]): The Global Contrast Factor (GCF) should correspond closer to the human perception of contrast. GCF uses contrasts at various resolution levels to compute the overall contrast.
HSNR (Ma et al. [2013]): The HSNR (signal to noise ratio based on the human visual system) finger-vein image quality evaluation index is based on both, the human visual characteristics and finger-vein image characteristics, by combining several features: contrast score, effective area score, finger shifting score and a noise level.
GLES (Yang et al. [2013]): The grey level entropy score is based on a discrete image entropy calculation.
ES (Yang et al. [2013]): The entropy score is also based on a discrete image entropy calculation, with only slight variation relative to GLES.
NIQE (Saad and Bovik [2012]): Is a general purpose no-reference opinion-unaware and distortion unaware image quality metric.
For all metrics but NIQE (where the opposite is true), larger values indicate higher sample image quality.

3 Experimental Results

Results in Tables 1 and 2 are sample-averaged quality metric values. In the "No Filter" row, samples are acquired by the sensor without any filter being mounted. For all other rows, filters denoted in the

first column are screwed on the lens. The first metric value is obtained with plain illumination source, the second value originates from the additional application of polarizing film to the illumination source. In Table 1, we display the results for the most controlled acquisition environment, i.e. indoor capturing with artificial light from the ceiling and closed frontal lid.

Table 1: Controlled environment: Indoor capturing (artificial ceiling light) with closed frontal lid.

	Wang	GCF	GLES	ES	HSNR	NIQE
No Filter	0.32	1.54	0.94	5.92	86.4	8.02
BN850	0.47/0.52	2.23/1.51	1.00/0.98	7.66/6.49	87.8/95.6	8.22/10.2
PC052	0.34/0.41	1.68/1.65	0.98/0.98	6.31/6.81	87.9/88.9	10.7/8.56
PR032	0.55/0.42	1.63/0.98	0.98/0.99	6.08/7.24	90.2/91.9	13.0/7.98
PR1000	0.57/0.35	1.50/2.00	0.98/0.99	6.12/6.80	93.5/93.0	12.6/11.0
PR120	0.50/0.47	2.83/2.00	0.99/0.99	6.98/6.82	94.3/92.8	11.0/8.98

In green, we depict the overall best metric value per image quality metric. We notice that this value is inconsistently attained: With and without polarization film on illumination source and for different filters (only the circular polarization filter PC052 values never show up in green). In any case, for all quality metrics specific to vascular data, the metric values observed with filters on the lens are superior to those without filter being applied ("No Filter"; only for the GCF metric we observe one inferior value indicated in red). For the NIQE quality metric, all but a single value are worse in case of filters being applied, i.e. indicating better quality for the "No Filter" setting. In Table 2, we display the results for the most uncontrolled acquisition environment considered, outdoor capturing with cloudy sky and open frontal lid.

Table 2: Uncontrolled environment: Outdoor capturing (cloudy sky) with open frontal lid

	Wang	GCF	GLES	ES	HSNR	NIQE
No Filter	0.32	1.21	0.81	5.68	84.4	8.05
BN850	0.36/0.47	2.16/2.00	0.97/0.99	7.39/6.82	86.6/92.8	8.10/8.98
PC052	0.31/0.42	1.54/0.98	0.94/0.99	6.11/7.24	85.9/91.9	9.45/7.98
PR032	0.49/0.35	1.49/2.00	0.87/0.99	6.12/5.80	88.2/93.0	10.2/11.0
PR1000	0.42/0.41	1.20/1.65	0.84/0.98	6.12/6.81	90.5/88.9	10.4/8.56
PR120	0.41/0.52	2.63/1.51	0.87/0.98	6.18/6.49	91.3/95.6	9.18/10.2

Considering the "No Filter" setting, the quality metric values for the uncontrolled environment are consistently inferior as compared to the controlled environment (as it is expected; only for the Wang metric, the values are identical and for NIQE, the value for uncontrolled setting is slightly superior). In green, we again depict the overall best metric value per image quality metric. We notice that this value is still inconsistently attained, but not entirely random: In four out of six quality metrics green values are observed with polarization film on illumination source and the values of the linear polarization filter PR120 show up three times in green. In any case, again for all but two quality metrics specific to vascular data, the metric values observed with filters on the lens are superior to those without filter being applied ("No Filter"; exceptions where we observe inferior values are indicated in red: Wang & GCF for one filter and negligible decrease). Again, for the NIQE quality metric, all but a single value are worse in case of filters being applied.

4 Conclusion

Experimental results clearly indicate that applying polarization (and other filters) to the lens, improves hand-vein sample image quality across the considered range of specific vascular quality metrics. However, we have found no empirical evidence that equipping illumination source additionally with a linear polarization foil (in relative perpendicular direction to the lens filters if applicable) further enhances those quality values. Also, linear polarization does not turn out to be superior to circular polarization and band-pass filtering. NIQE behaves antithetic to the vascular sample quality metrics, which is not too surprising, as it is designed to work on natural images captured under visible light.

In future work we aim to confirm these findings in actual recognition experiments, confirming results in (Kuang et al. [2024]) for a standard hand-vein acquisition system.

Acknowledgments and Disclosure of Funding

We thank Latif Muhammad Ummar who has captured the analyzed hand vein sample data in the context of a Multimedia Technologies Seminar at the University of Salzburg (master's program on Applied Geoinformatics).

References

Andreas Uhl, Christoph Busch, Sebastien Marcel, and Raymond Veldhuis. *Handbook of Vascular Biometrics*. Advances in Computer Vision and Pattern Recognition. Springer Nature Switzerland AG, Cham, Switzerland, 2019. ISBN 978-3-030-27731-4. doi: 10.1007/978-3-030-27731-4.

Ying Kuang, Shuai Wang, Bincheng Mo, Shiyou Sun, Kai Xia, and Yuanmu Yang. Palm vein imaging using a polarization-selective metalens with wide field-of-view and extended depth-of-field. *npj Nanophotonics*, 1, 07 2024. doi: 10.1038/s44310-024-00027-4.

Simon Kirchgasser, Christof Kauba, and Andreas Uhl. Towards understanding acquisition conditions influencing finger-vein recognition. In Andreas Uhl, Christoph Busch, Sebastien Marcel, and Raymond Veldhuis, editors, *Handbook of Vascular Biometrics*, chapter 7, pages 177–199. Springer Nature Switzerland AG, Cham, Switzerland, 2019. ISBN 978-3-030-27731-4. doi: 10.1007/978-3-030-27731-4_7.

Shiyou Sun, Zheng'ao Wang, Fanglin Chen, and Kai Xia. Palm vein imaging enhancement in highly reflective hand scenarios. In *2023 8th International Conference on Image, Vision and Computing (ICIVC)*, pages 438–442, 2023. doi: 10.1109/ICIVC58118.2023.10270232.

Christof Kauba and Andreas Uhl. Shedding light on the veins - reflected light or transillumination in hand-vein recognition. In *Proceedings of the 11th IAPR/IEEE International Conference on Biometrics (ICB'18)*, pages 1–8, Gold Coast, Queensland, Australia, 2018. doi: 10.1109/ICB2018.2018.00050. URL https://doi.org/10.1109/ICB2018.2018.00050.

Simon Kirchgasser, Christof Kauba, Georg Wimmer, and Andreas Uhl. Advanced image quality assessment for hand- and finger-vein biometrics. *IET Biometrics*, 2025(1):8869140, 2025. doi: https://doi.org/10.1049/bme2/8869140.

Chunyi Wang, Xinhua Zeng, Xiongwei Sun, Wengong Dong, and Zede Zhu. Quality assessment on near infrared palm vein image. In *2017 32nd Youth academic annual conference of Chinese association of automation (YAC)*, pages 1127–1130. IEEE, 2017.

Kresimir Matkovic, László Neumann, Attila Neumann, Thomas Psik, Werner Purgathofer, et al. Global contrast factor-a new approach to image contrast. In *Computational Aesthetics*, pages 159–167, 2005.

Hui Ma, Feng Peng Cui, and Popoola Oluwatoyin P. A non-contact finger vein image quality assessment method. *Applied Mechanics and Materials*, 239:986–989, 2013.

Lu Yang, Gongping Yang, Yilong Yin, and Rongyang Xiao. Finger vein image quality evaluation using support vector machines. *Optical engineering*, 52(2):027003–027003, 2013.

Michele A Saad and Alan C Bovik. Blind quality assessment of videos using a model of natural scene statistics and motion coherency. In *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*, pages 332–336. IEEE, 2012.

Physics-informed Machine Learning

Introducing Monge-GPs: A new class of physics-informed Gaussian Processes (Extended abstract)

J. Moser[1], **C. Albert**[1], **S. Ranftl**[2,3]
(1) ITPCP, Graz University of Technology, Austria
(2) Courant Institute, New York University, New York, USA
(3) Division of Applied Mathematics, Brown University, USA

1 Introduction

Hybrid approaches combining differential equations and machine learning, commonly referred to as physics-informed machine learning, have gained significant attention in recent years. Prominent examples include Physics-Informed Neural Networks (PINNs) [1] and Physics-Informed Gaussian Processes (PIGPs), the latter naturally providing uncertainty quantification. PIGPs encode differential constraints directly in the covariance kernel, and existing approaches can be roughly grouped into two schools of thought. Operator-based constructions apply differential operators to a base kernel, yielding systematic and algorithmic methods, but are often restricted to controllable systems [2] or specific classes of differential equations [3] or may require many auxiliary outputs [4] and a relatively large amount of data. In contrast, Mercer-type constructions build kernels from problem-specific solution components such as Green's functions [5] or fundamental solutions [6]; while typically data efficient, they rely on analytical insight and substantial manual derivation. We propose Monge-GPs, a hybrid construction based on Monge parametrization that unifies operator-based kernels as in [2] and Mercer kernels as in [6]. By parametrizing the controllable dynamics algorithmically and restricting problem-specific design to a low-dimensional autonomous component, the approach substantially reduces the need for manual kernel design, and stays data efficient while lifting the restriction to controllable systems.

2 We present: Monge GPs - The theory

Let's suppose we have a system of linear ODEs or PDEs, represented by a full row-rank operator matrix R and unknown solutions η of the form

$$R\eta = 0. \tag{1}$$

This system is called *controllable*, if all possible solutions in a function space (for example C^∞) can be written in the parametrized form $\eta = B\zeta$ for any $\zeta \in C^\infty$, where B is an operator matrix that forms the null-space of R, i.e. $RB = 0$. Gaussian processes are closed under linear operators, so we can construct a parametrized GP by transforming the kernel with this parametrization matrix such as in [2]. We need a base kernel k_0, whose realizations are dense in C^∞, in order to be able to realize the ζ in our solution - the standard RBF kernel is the natural choice here. The parametrized kernel is then simply $K(\mathrm{x}, \mathrm{x}') = B(\mathrm{x})B^T(\mathrm{x}')k_0(\mathrm{x}, \mathrm{x}')$. We see that this parametrization approach depends crucially on$B$ being non-trivial and expressive enough, and therefore per definition on controllability.

If the system is not fully controllable however, we propose to instead decompose the solution using the so-called Monge parametrization as used in [7], leading to our Monge-GP. The basic idea can be formalized as follows: We can split the system matrix R into its controllable part R' and its autonomous part represented by R'', such that $R = R''R'$. Since R' is controllable, it has

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

a parametrization matrix B (i.e. $R'B = 0$), as well as a right-inverse matrix $(R')^{-1}$. Defining $R'\eta =: \xi$, we define a new, but equivalent, system of equations

$$R\eta = 0 \Leftrightarrow \big(R''\xi = 0 \wedge R'\eta = \xi\big). \tag{2}$$

This means we first need to find the solutions of the autonomous problem $R''\xi = 0$. This corresponds to a Mercer-type construction applied to the reduced autonomous subsystem, which is typically much simpler than the original system since the parametrizable component has already been separated. Given our found solution ξ (for $n \geq 1$ independent vector-valued solutions, ξ will be a matrix with n columns), we can go ahead and solve the inhomogeneous equation $R'\eta = \xi$, whose general solution can be split into a homogeneous and a particular part $\eta = \eta_h + \eta_p$. With the parametrization matrix B we can write the general homogeneous solution as $\eta_h = B\zeta$ for any $\zeta \in C^\infty$. Furthermore, by inverting R', we can find a particular solution to the inhomogeneous part of our problem, thus writing the general solution as

$$\eta = B\zeta + ((R')^{-1}\xi)\vec{C} \quad \forall \zeta \in C^\infty, \vec{C} \in \mathbb{R}^n. \tag{3}$$

How can we translate this into our kernel function? The homogeneous part of the solution is controllable and can therefore be translated into the kernel through parametrization just as discussed earlier. The autonomous, i.e. particular part of the solution in contrast has more constraints and can therefore in general be expressed through its fundamental solutions $\phi = (R')^{-1}\xi$ as a Mercer kernel

$$k_p(\mathbf{x}, \mathbf{x}') = \sum_{ij} \phi_i(\mathbf{x})\Sigma_{ij}\phi_j(\mathbf{x}')^T \qquad \phi_i, \phi_j \in \phi. \tag{4}$$

with a positive semi-definite, and in general diagonal, covariance Σ. Since GPs are closed under addition, we can then form a full GP that is able to realize all controllable and autonomous solutions of our system by adding the parametrized and autonomous kernel. The full solution kernel is then simply

$$K(\mathbf{x}, \mathbf{x}') = B(\mathbf{x})B^T(\mathbf{x}')k_0(\mathbf{x}, \mathbf{x}') + k_p(\mathbf{x}, \mathbf{x}'). \tag{5}$$

3 Example: Bipendulum

Let's consider the Bipendulum ODE system also discussed in [3, 8]. It consists of two pendula, of lengths $\ell_1 = \ell_2 = 1$ that are connected on top by a rod that moves according to its acceleration $u(t)$. The linearized system is represented by

$$R\eta = \begin{pmatrix} \partial_t^2 + 1 & 0 & -1 \\ 0 & \partial_t^2 + 1 & -1 \end{pmatrix} \begin{pmatrix} \theta_1(t) \\ \theta_2(t) \\ u(t) \end{pmatrix} = 0. \tag{6}$$

setting the gravitational acceleration to $g = 1$ for simplicity. The parametrizable part of our system R', the corresponding parametrization matrix B and right-inverse matrix $(R')^{-1}$ are

$$R' = \begin{pmatrix} 1 & -1 & 0 \\ 0 & \partial_t^2 + 1 & -1 \end{pmatrix} \qquad B = \begin{pmatrix} 1 \\ 1 \\ \partial_t^2 + 1 \end{pmatrix} \qquad (R')^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & -1. \end{pmatrix} \tag{7}$$

Since $R' \neq R$, we need to define the autonomous subsystem represented by R'' by factorizing R, yielding

$$\begin{aligned} R''\xi &= \begin{pmatrix} \partial^2 + 1 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \xi_0 \\ \xi_1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \\ &\Rightarrow \xi_1 = 0,\ \xi_0 \in \text{span}\{\cos(t), \sin(t)\}. \end{aligned} \tag{8}$$

This leads to the autonomous solution

$$\eta_p = (R')^{-1}\xi = \begin{pmatrix} 1 & 0 \\ 0 & 0 \\ 0 & -1 \end{pmatrix} \begin{pmatrix} \cos(t) & \sin(t) \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \cos(t) & \sin(t) \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \tag{9}$$

and therefore the full solution therefore reads

$$\eta = \begin{pmatrix} 1 \\ 1 \\ \partial^2 + 1 \end{pmatrix} \zeta + \begin{pmatrix} \cos(t) & \sin(t) \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \vec{C} \qquad \zeta \in C^\infty, \vec{C} \in \mathbb{R}^2. \tag{10}$$

All that is left to do is to translate this into a kernel function. Since the autonomous solution can be represented by the tuple $\phi(t) = \text{span}\{\begin{pmatrix} \cos(t) \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sin(t) \\ 0 \\ 0 \end{pmatrix}\}$, the corresponding kernel is a Mercer kernel as in eq. 4. Choosing $\Sigma = \mathbb{I}$ and exploiting trigonometric identities, we get the autonomous kernel

$$k_p(t, t') = \begin{pmatrix} \cos(t - t') & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{11}$$

and therefore the total Monge-GP kernel reads

$$\begin{aligned} k(t, t') &= B(t) k_0(t, t') B^T(t') + k_p(t, t') \\ &= \begin{pmatrix} k_0(t, t') + \cos(t - t') & k_0(t, t') & (\partial_{t'}^2 + 1) k_0(t, t') \\ k_0(t, t') & k_0(t, t') & (\partial_{t'}^2 + 1) k_0(t, t') \\ (\partial_t^2 + 1) k_0(t, t') & (\partial_t^2 + 1) k_0(t, t') & (\partial_t^2 + 1)(\partial_{t'}^2 + 1) k_0(t, t') \end{pmatrix}. \end{aligned} \tag{12}$$

We can now implement this kernel in python using the python package `PCGP`[1] [9] and use it to either solve the differential equation for any initial conditions, or to solve the inverse problem of estimating ℓ from available data, as discussed for example in [9]. In Fig. 1 we show a solution to the bipendulum problem for a sinusoidal input $u(t) = \sin(2t)$ with randomly chosen boundary conditions and additive Gaussian noise of $\sigma = 0.03$ on the input. The analytical solution for the particular boundary conditions is marked with a red dotted line. Even with few conditioning points, the posterior closely matches the analytical solution.

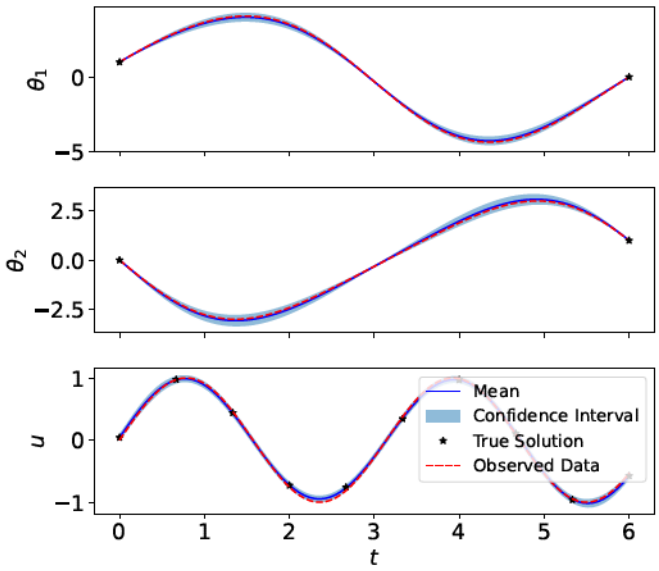


Figure 1: Example of the solution of a forward problem using Monge-GPs. The mean is depicted in dark blue, the 2σ confidence interval is light blue, the analytical solution is marked in red, and the training data are shown as black stars.

[1] freely available on github: https://github.com/moserjo/PCGP

4 Ongoing work

We are currently comparing this approach to others [4, 3, 6] in performance and expect it to work well, especially in the low data regime. It is a general approach that is specifically not restricted to controllable systems, and can be used both in ODE and PDE settings with constant coefficients. We are currently looking into well-definedness for systems with non-constant coefficients using for example Weyl algebras (allowing for polynomial and rational coefficients) as a suitable operator ring. We are also currently looking into finding the autonomous solution algorithmically, where grade filtration [10] may be a promising candidate. If successful, we will implement an automatic construction of the Monge-kernel as an extension in the python package `PCGP` [9].

5 Acknowledgements

The authors thank the reviewers for their helpful feedback. J.M. would also like to thank A. Quadrat for the introduction to homological algebra and grade filtration and fruitful discussion thereof. J.M. and C.A. were financially supported by the Austrian Research Promotion Agency (FFG) project VENTUS within the AI for Green funding program, Grant no. 910263. S.R. was financially supported by the Austrian Science Fund (FWF), Grant DOI: 10.55776/J4774.

References

[1] George Em Karniadakis, Ioannis G Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3(6):422–440, 2021.

[2] Markus Lange-Hegermann. Algorithmic Linearly Constrained Gaussian Processes. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/hash/68b1fbe7f16e4ae3024973f12f3cb313-Abstract.html.

[3] Andreas Besginow and Markus Lange-Hegermann. Constraining Gaussian processes to systems of linear ordinary differential equations. *Advances in Neural Information Processing Systems*, 35: 29386–29399, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/bcef27c5825d1ed8757290f237b2d851-Abstract-Conference.html.

[4] Maziar Raissi, Paris Perdikaris, and George Em Karniadakis. Machine learning of linear differential equations using Gaussian processes. *Journal of Computational Physics*, 348:683–693, November 2017. ISSN 0021-9991. doi: 10.1016/j.jcp.2017.07.050. URL https://www.sciencedirect.com/science/article/pii/S0021999117305582.

[5] Mauricio A. Álvarez, David Luengo, and Neil D. Lawrence. Linear Latent Force Models Using Gaussian Processes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2693–2705, November 2013. ISSN 1939-3539. doi: 10.1109/TPAMI.2013.86. URL https://ieeexplore.ieee.org/abstract/document/6514873.

[6] Christopher G. Albert. Gaussian Processes for Data Fulfilling Linear Differential Equations. *Proceedings*, 33(1):5, 2019. ISSN 2504-3900. doi: 10.3390/proceedings2019033005. URL https://www.mdpi.com/2504-3900/33/1/5. Number: 1 Publisher: Multidisciplinary Digital Publishing Institute.

[7] A Quadrat and D Robertz. Parametrizing all solutions of uncontrollable multidimensional linear systems. *IFAC Proceedings Volumes*, 38(1):49–54, 2005.

[8] Jean-Francois Pommaret. *Partial Differential Control Theory: Mathematical tools*, volume 530. Springer Science & Business Media, 2001.

[9] Johanna Moser, Christopher Albert, and Sascha Ranftl. Parameter learning with physics-consistent gaussian processes. *In preparation*, 2026.

[10] Alban Quadrat. Grade filtration of linear functional systems. *Acta Applicandae Mathematicae*, 127(1): 27–86, 2013.

Joint Bayesian Inference on Lagrangian Physics and Trajectories

Michael Obermayr
Institute for Machine Learning and Neural Computation
Technical University Graz
Graz, Austria
michael.obermayr@tugraz.at

Robert Peharz
Institute for Machine Learning and Neural Computation
Graz, Austria
robert.peharz@tugraz.at

Abstract

Numerical integration and ODE discovery are two sides of the same coin—converse problems of finding trajectories from known physics versus inferring physics from observed trajectories. Although these problems have been extensively studied in isolation, they can be unified through the minimization of a common quantity: the Euler–Lagrange residual. In this paper, we build on this insight and introduce the Integrated Squared Action Residual (ISAR), which enables both tasks to be performed simultaneously. We formulate numerical integration and model discovery as a joint Bayesian inference problem, allowing for the systematic incorporation of physical prior knowledge and domain constraints in settings with sparse and noisy observations, where traditional approaches typically fail. While we demonstrate the performance on two mechanical toy problems, it can be readily extended towards multiphysics systems including dissipative dynamics.

1 Introduction

Physicists are primarily concerned with two activities: discovering models that describe nature and using those models to predict future events. The former is an active field with recent powerful data-driven approaches (Brunton et al., 2016; Raissi et al., 2017; M. Raissi, 2019; Greydanus et al., 2019), while the latter relies mainly on classical numerical integrators with notable learning-based extensions (Chen et al., 2018). Interestingly, these two activities are rarely recognized as two sides of the same coin.

Both can be seen as algorithms minimizing the residual of a differential equation evaluated over the observed or proposed trajectory. While equation discovery does this very explicitly, numerical integrators inherently minimize discretization error, which can also be seen as a physical residual. Variational integrators, for instance, are designed to minimize the local residual of the Euler-Lagrange equations, which vanishes only if the trajectory is perfectly explained by the underlying physics.

We propose using the global Euler-Lagrange residual as a unified loss function for both numerical integration (the forward problem) and equation discovery (the backward problem). While this loss function can be used for each problem in isolation, it is now possible to tackle both simultaneously. This becomes particularly useful when dealing with systems with partially known dynamics and sparse, noisy observations of trajectories. For such systems, numerical integration is inapplicable due

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

to incomplete physical equations, and standard discovery methods struggle because derivatives cannot be reliably estimated via finite differences. Moreover, solutions may be highly underdetermined, especially when the applied algorithm relies solely on observations and not all available partial knowledge of the dynamics is utilized.

To address these challenges, we employ a fully Bayesian treatment to infer the complete solution distribution rather than a single point estimate. We define a probabilistic model of the form Physics $\phi \rightarrow$ Trajectory $\theta \rightarrow$ Data, where ϕ and θ are linked by a physics likelihood derived from the global Euler-Lagrange residual. This framework allows for data-efficient inference over the full posterior of plausible physics and trajectories, given sparse noisy observations and weak priors, which are derived from partial system knowledge.

While preliminary results on mechanical toy problems are promising, the remaining challenge consist of reliably estimating the normalizing constant in the proposed physics likelihood to counteract the observed posterior distortion.

2 Background

2.1 Principle of Stationary Action

Our methodology is based on the Principle of Stationary Action, which is fundamental to all reversible non-dissipative systems. It states that for a physical trajectory $x(t)$ connecting two fixed points $x(t_0)$ and $x(t_1)$, the action functional $S[x(t)]$ must be stationary under infinitesimal variations of the path. It provides a rigorous method for solving boundary-value problems in classical mechanics.

Using variational calculus, one can derive the Euler-Lagrange Equations of the first kind. For a system governed by a Lagrangian function L, subject to m constraints f_j, weighted by multipliers λ_j these equations read:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x} = \sum_{j=1}^{m} \lambda_j \frac{\partial f_j}{\partial x} \tag{1}$$

This transformation of the Stationary Action Principle into a system of differential equations allows to address both boundary-value and initial-value problems.

In a "Natural Lagrangian System," where coordinates x are holonomic and the Lagrangian has no explicit time dependence, the right-hand side of the equation vanishes. For simplicity, we limit our examples to such systems, although the framework is readily extendable.

2.2 ISAR: Integrated-Squared-Action-Residual

We now propose to measure the global violation of the Principle of Stationary Action by integrating the squared residual of the Euler-Lagrange Equations, which we define in Eq. 2. It is a strong form residual loss for Lagrangian systems and similar losses are established in literature (Lutter et al., 2019; Cranmer et al., 2020; Kharazmi et al., 2019). The integral, which we will call **ISAR** is non-negative and only zero, if a trajectory $x(t)$ perfectly satisfies the equations of motion, derived from the Lagrangian function, or vice versa.

$$ISAR := \int_{t_0}^{t_1} \left(\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) - \frac{\partial L}{\partial x}\right)^2 dt \tag{2}$$

$$= \int_{t_0}^{t_1} \left(\ddot{x} \cdot \frac{\partial^2 L}{\partial \dot{x}^2} + \dot{x} \cdot \frac{\partial^2 L}{\partial \dot{x} \partial x} - \frac{\partial L}{\partial x}\right)^2 dt \tag{3}$$

Evaluation of *ISAR* is straightforward with modern autodifferentiation libraries, where differential operators can be readily evaluated on trajectories. This leaves us with a one-dimensional integral for each state-space dimension, which can be approximated numerically.

The unit of the resulting integral is $\frac{kg^2m^2}{s^3} = N^2s$, we can interpret it as the integrated squared discrepancy in predicted vs. observed force. Note, that Lagrangian dynamics are invariant under scaling and the addition of a gauge, leading to identifiability issues. To avoid trivial solutions (such as $L = 0$), we need either strong physical priors or measures to prevent a collpase of the phase space

volume, such as normalizing *ISAR* with the generalized mass-matrix $\frac{\partial^2 L}{\partial \dot{x} \partial \dot{x}}$. This mass-normalized *ISAR* can be interpreted as the integrated discrepancy in acceleration (unit $\frac{m^2}{s^3}$) and is the same training objective used by Cranmer et al. (2020) to learn Lagrangians with neural networks.

The integral *ISAR* is differentiable with respect to both the trajectory parameters and the Lagrangian parameters. Thus, one may apply gradient-based optimization methods. If the Lagrangian function and boundary/initial conditions are given, one can minimizing *ISAR* with respect to trajectory parameters to find the unknown trajectory. This essentially corresponds to solving a boundary-value/initial-value problem via numerical integration. Conversely, if a trajectory is given, one can minimize with respect to the Lagrangian parameters to arrive at the governing Lagrangian, essentially performing equation discovery.

3 Methodology

3.1 Probabilistic Model

To infer jointly the trajectory parameters θ and physics parameters ϕ, we propose a physics likelihood based on *ISAR*:

$$p(\theta|\phi) = p(\theta) \cdot \exp(-\lambda \cdot ISAR(\phi, \theta)) \cdot \frac{1}{Z_\phi} \tag{4}$$

Here $p(\theta)$ is the prior over trajectories which may encourage smoothness or limit the class of possible functions. Z_ϕ is the normalizing constant, which is nontrivial to estimate and remains the main hurdle in the current work.

Together with a physics prior $p(\phi)$ and a gaussian observation likelihood $p(Data|\theta)$ we propose the following posterior distribution for trajectory parameters θ and Lagrangian parameters ϕ given the noisy observations $Data$.

$$p(\phi, \theta|Data) = \frac{p(\phi)p(\theta|\phi)p(Data|\theta)}{p(Data)} \tag{5}$$

The model hierarchy $\phi \rightarrow \theta \rightarrow Data$ resembles the conventional causal direction in physics: physical equations ϕ determine trajectories θ, whose measurement produces observed $Data$. By defining $p(\theta|\phi)$ we ensure that measurements remain conditionally independent of the underlying physics if the trajectory is given, while establishing a connection between the random variables ϕ and θ.

3.2 Parametrization and Inference

We parametrize the trajectories using third-order Hermite splines, as they are simple, and are continuously and analytically differentiable. For the Lagrangian, we begin with a physically motivated ansatz and learn its physical parameters. In scenarios where physical knowledge is minimal, this framework remains flexible: the Lagrangian may instead be represented by a structured neural network, similar to the approach by Cranmer et al. (2020).

For inference, we draw samples from the log-posterior, using the No-U-Turn Sampler (NUTS), a common implementation of Hamiltonian Monte Carlo (HMC). Additionally, we compute a MAP estimate from a single starting-point using the L-BFGS-B algorithm. This serves both as a quick sanity check and to highlight the loss of information when favoring a single point estimate over the full posterior distribution.

We can neglect the evidence term $log(p(Data))$ in the log-posterior, as it is constant. In our preliminary results, we further omit the normalizing constant of the physics likelihood Z_ϕ, due to difficulties in estimation. This results in an unnormalized physics likelihood, which distorts the posterior, as we observe in our results.

3.3 Experiments

3.3.1 Harmonic Oscillator

As an introductory example, we begin by studying the harmonic oscillator. First we simulate a single trajectory $q(t)$ with mass $m = 1.5$ and a spring constant $k = 5.0$, which we observe at 8 equally

spaced points with a noise level of 1 %. We parametrize the Lagrangian function with the correct harmonic oscillator ansatz in Eq. 6 with learnable parameters $\phi = \{m, k\}$ and the trajectory with 125 3rd order Hermite splines, to allow for high-quality approximations of high-order frequencies. Our goal is to recover the posterior distribution of physics parameters ϕ and trajectory parameters θ from the noisy observations as well as broad priors over ϕ and a smoothness prior on the trajectory parameters θ.

$$L(q, \dot{q}) = \frac{1}{2} m \dot{q}^2 - \frac{1}{2} k q^2 \tag{6}$$

The solutions to the harmonic oscillator are sinusoidal functions with the frequency $\omega = \sqrt{\frac{k}{m}}$. Note, that the observations are far too sparse to estimate reliable derivatives from finite differences, many established Physics Discovery methods are doomed to fail. Also note, that there is an infinite set of frequencies and thus trajectories satisfying the Lagrangian for the given observations and only by utilizing further assumptions the infinite set collapses to a limited set of posterior modes. The key advantage of our method is to define these assumptions explicitly and systematically with the priors on $p(\phi)$ and $p(\theta)$.

Although the MAP Estimate can only identify one mode, posterior samples obtained via NUTS reveal a multitude of plausible frequencies, as can be seen in Fig. 1. Many solutions might be missed, if one does not perform proper Bayesian Inference.

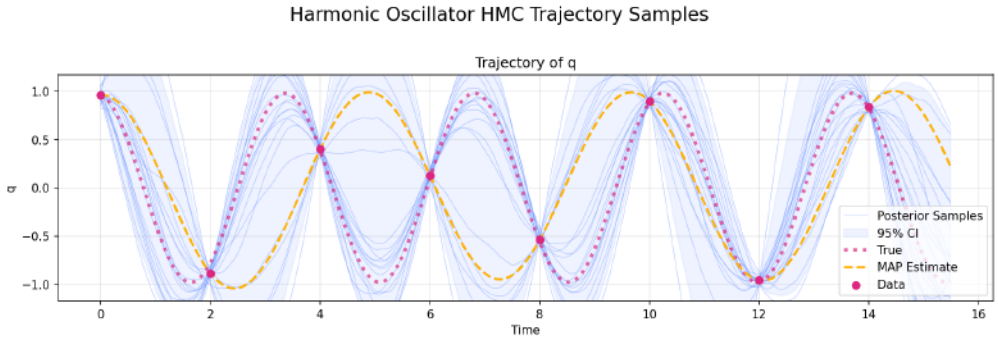


Figure 1: Trajectories generated from marginal posterior samples for the harmonic oscillator system

3.3.2 Double Pendulum

Next we address the chaotic double pendulum system, defined by the Lagrangian function in Eq. 7. Again, we provide this Lagrangian as the model ansatz with physical parameters $\phi = \{m_1, m_2, l_1, l_2\}$ and parametrize the trajectories with 301 3rd order Hermite splines. From noisy observations of a single simulated trajectory and priors on ϕ and θ we infer the joint posterior distribution over physical and trajectory parameters via NUTS.

$$\begin{aligned} \mathcal{L} =& \frac{1}{2}(m_1 + m_2) l_1^2 \dot{\theta}_1^2 + \frac{1}{2} m_2 l_2^2 \dot{\theta}_2^2 + m_2 l_1 l_2 \dot{\theta}_1 \dot{\theta}_2 \cos(\theta_1 - \theta_2) \\ &+ (m_1 + m_2) g l_1 \cos\theta_1 + m_2 g l_2 \cos\theta_2 \end{aligned} \tag{7}$$

In Fig. 2 we plot the marginal posterior distribution for all physical parameters $\phi = \{m_1, m_2, l_1, l_2\}$, together with the initial prior distributions. The ground-truth parameters used for simulation are shown in dark magenta, the MAP estimate is shown in orange. We observe that the pendulum length parameters l_1, l_2 collapse precisely around the true values. On the other hand the posterior distribution over m_1 and m_2 barely contract, showcasing the scale invariance of the Lagrangian. Strikingly, we observe a distinct shift in the mass posteriors away from the prior, which was centered at the true value. Further experiments confirm that this bias stems from neglecting the normalizing constant of the physics likelihood. Therefor it is necessary to estimate Z_ϕ reliably and subsequently correct the log posterior.

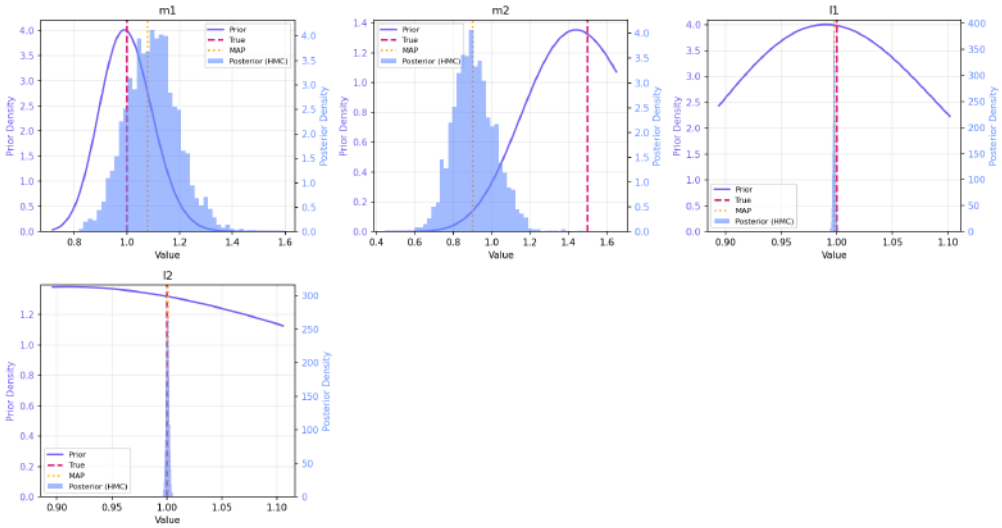


Figure 2: Marginal posteriors of physics parameters for the double pendulum system

4 Conclusion

The joint inference over Lagrangian physics and trajectories allows to infer distributions of plausible solutions in scenarios with sparse noisy data and limited physical knowledge. In such scenarios, finite-difference methods typically fail and deterministic approaches are inadequate as the inverse problem is heavily underdetermined.

Preliminary results on mechanical toy problems look promising, although the found posteriors remain distorted as an unnormalized physics likelihood is used. Therefor, the focus of our ongoing work is on reliably estimating the normalizing constant for correcting the log posterior, to obtain proper and unbiased posterior samples.

References

Brunton, S. L., Proctor, J. L., and Kutz, J. N. Discovering governing equations from data by sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, March 2016. ISSN 1091-6490. doi: 10.1073/pnas.1517384113. URL http://dx.doi.org/10.1073/pnas.1517384113.

Chen, R. T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural ordinary differential equations. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/69386f6bb1dfed68692a24c8686939b9-Paper.pdf.

Cranmer, M., Greydanus, S., Hoyer, S., Battaglia, P., Spergel, D., and Ho, S. Lagrangian Neural Networks, July 2020. URL http://arxiv.org/abs/2003.04630. arXiv:2003.04630 [cs].

Greydanus, S., Dzamba, M., and Yosinski, J. Hamiltonian neural networks. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/26cd8ecadce0d4efd6cc8a8725cbd1f8-Paper.pdf.

Kharazmi, E., Zhang, Z., and Karniadakis, G. E. Variational Physics-Informed Neural Networks For Solving Partial Differential Equations. *arXiv e-prints*, art. arXiv:1912.00873, November 2019. doi: 10.48550/arXiv.1912.00873.

Lutter, M., Ritter, C., and Peters, J. Deep lagrangian networks: Using physics as model prior for deep learning. In *International Conference on Learning Representations*, 2019. URL https://ui.adsabs.harvard.edu/abs/2019arXiv190704490L.

M. Raissi, P. Perdikaris, G. K. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2018.10.045. URL https://www.sciencedirect.com/science/article/pii/S0021999118307125.

Raissi, M., Perdikaris, P., and Karniadakis, G. E. Machine learning of linear differential equations using gaussian processes. *Journal of Computational Physics*, 348:683–693, 2017. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2017.07.050. URL https://www.sciencedirect.com/science/article/pii/S0021999117305582.

Stabilizing PINNs: A regularization scheme for PINN training to avoid unstable fixed points of dynamical systems

Miloš Babić
Christian Doppler Laboratory for Physics-driven Machine Learning in Industrial Applications, Graz, Austria
Institute of Thermodynamics and Sustainable Propulsion Systems, Graz University of Technology, Graz, Austria
Know Center Research GmbH, Graz, Austria

Franz M. Rohrhofer
Know Center Research GmbH, Graz, Austria

Bernhard C. Geiger
Signal Processing and Speech Communication Laboratory, Graz University of Technology, Graz, Austria
Know Center Research GmbH, Graz, Austria

Abstract

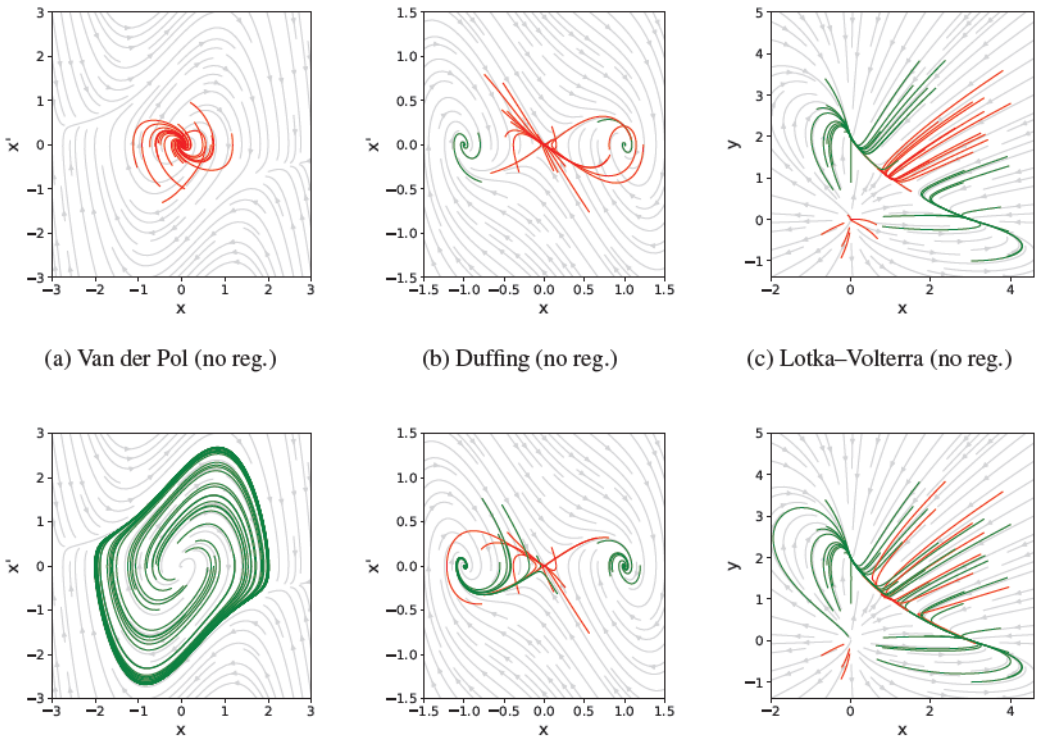
It was recently shown that the loss function used for training physics-informed neural networks (PINNs) exhibits local minima at solutions corresponding to fixed points of dynamical systems. In the forward setting, where the PINN is trained to solve initial value problems, these local minima can interfere with training and potentially lead to physically incorrect solutions. Building on stability theory, this paper proposes a regularization scheme that penalizes solutions corresponding to unstable fixed points. Experimental results on four dynamical systems, including the Lotka-Volterra model and the van der Pol oscillator, show that our scheme helps avoiding physically incorrect solutions and substantially improves the training success rate of PINNs.

1 Introduction

Physics-informed neural networks (PINNs, [10]) are among the most prominent instantiations of physics-informed machine learning. Capable of including systems of differential equations during training, they have been proposed for solving boundary and initial value problems , for inferring parameters of a differential equation from data and for estimating unobservable scalar or vector fields from measurements . Especially for the forward problem – i.e., for learning the solution to a system of differential equations given boundary and initial conditions – PINNs experience training difficulties for all but the most trivial problem settings.

There is a substantial body of literature that investigates the root causes of and offers remedies for these training difficulties. For example, for stiff problems in which the gradient of the solution function varies strongly across the computational domain, it was shown that reducing the weight of points close to gradient maxima leads to a more well-behaved loss and, hence, successful training [11]. Other studies attribute challenges in PINN training to the use of large computational domains, where techniques such as domain decomposition [5, 8] and sinusoidal feature mappings [15] have

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).



(a) Van der Pol (no reg.) (b) Duffing (no reg.) (c) Lotka–Volterra (no reg.)

(d) Van der Pol (reg.) (e) Duffing (reg.) (f) Lotka–Volterra (reg.)

Figure 1: Phase portraits comparing standard PINN training (top row) and regularized training with derivative penalty at unstable fixed points (bottom row) for three dynamical systems: Van der Pol, Duffing, and Lotka–Volterra.

been shown to improve training success. More generally, PINN training is often susceptible to convergence toward the trivial all-zero solution, as standard neural network initializations tend to bias the model toward this outcome. To mitigate this, several strategies have been proposed, including specialized initialization schemes [15], ensemble methods [4], and adaptive collocation point weighting techniques [2, 13].

Interestingly, the all-zero solution is a valid general solution to a large class of differential equations, cf. [15, Prop. 1]. In addition to this all-zero solution, the authors of [12] showed that other, non-zero solutions can become minima of the training loss despite corresponding to unphysical behavior. More specifically, in the context of ordinary differential equations (ODEs), it was shown that solutions at fixed points are always global optima of the physics loss, regardless whether these fixed points are stable or not. While convergence to the all-zero solution can be prevented by some of the remedies mentioned above, convergence to unstable, but non-zero fixed points has not been addressed so far.

In this work, we fill this gap by proposing a regularization scheme that avoids training convergence to unstable fixed points, as illustrated in Fig. 1. More specifically, given a system of ODEs, we linearize the candidate solution around each collocation point and characterize its stability via the eigenvalues of the resulting Jacobian matrix (Section 3). For collocation points for which the candidate solution exhibits unstable behavior, the regularization term is large. Although the approach appears simple, we show in Section 5 that it substantially improves training success for several (low-dimensional) ODEs (Section 4).

2 Background

We focus on autonomous dynamical systems that can be described by ODEs of the general form:

$$x^{(n)} = f(x, x', x'', ..., x^{(n-1)}), \tag{1}$$

where f is a nonlinear function of time t and the unknown solution function $x(t)$ and where x', x'', and $x^{(n)}$ denote the first-, second-, and nth-order derivative with respect to time t. Since any higher-order ODE can be rewritten as a system of first-order ODEs, we consider the multi-dimensional first-order system $x' = f(x)$, where $f = (f_1, f_2, ..., f_n)$ and $x \in \mathbb{R}^n$. For this system, a *fixed point* $x^* \in \mathbb{R}^n$ satisfies $f(x^*) = 0$.The local stability around x^* can be determined by linearizing the system through Taylor expansion, and then analyzing the resulting Jacobian matrix $J(x^*) = \partial \vec{f} / \partial \vec{x}|_{x^*}$. If all eigenvalues $\{\lambda_1, ..., \lambda_n\} \subset \mathbb{C}$ of J have negative real parts, the fixed point x^* is asymptotically stable [1]. If at least one eigenvalue has a positive real part, the fixed point is unstable. An unstable fixed point is considered a saddle point if it is stable in some directions and unstable in others.

We use PINNs to approximate the unknown solution $x(t)$ to (1) with a fully connected neural network $x_\theta(t)$, where θ denotes the network's weights and biases. The PINN is trained on both labeled and unlabeled data. Labeled data is used to encode the initial condition (IC) $x(0)$ via the IC loss $\mathcal{L}_{\text{IC}} = \|x_\theta(0) - x(0)\|^2$. Unlabeled data (in the context of PINNs also called *collocation points*) is used to enforce the system dynamics (1) via the physics loss

$$\mathcal{L}_f = \frac{1}{N_{\text{col}}} \sum_{i=1}^{N_{\text{col}}} \|x'_\theta(t_i) - f\left(x_\theta(t_i)\right)\|^2, \tag{2}$$

where the collocation points $\{t_i\}_{i=1}^{N_{\text{col}}}$ are (randomly) sampled from the temporal domain $0 < t \leq T$, where T is the *simulation time*. Both losses are combined by linear scalarization, i.e. $\mathcal{L} = \mathcal{L}_{\text{IC}} + \mathcal{L}_f$, where an additional regularization loss, central to our study, will be introduced in Section 3.

As discussed in [12], fixed points of ODEs correspond to global optima of the physics loss, characterized by non-trivial basins of attraction. This arises from the fact that the fixed point condition $f(x^*) = 0$ leads to inherently small physics residuals in the vicinity of fixed points. Specifically, the physics loss (2) becomes small when $f(x^\star_\theta) \approx 0$, i.e. when the network output $x^\star_\theta$ is close to a fixed point x^*. As a consequence, regardless of whether a fixed point is stable or unstable, PINN solutions tend to be attracted to fixed points, often resulting in trajectories that converge to the nearest one (as shown in the Fig. 1). Closely related to convergence issues caused by fixed points is the prominent influence of the trivial zero solution, $x^* = 0$, which is a fixed point in many dynamical systems. For instance, Wong et al. [15] propose learning in sinusoidal space, introducing sinusoidal mappings to initialize PINNs with an appropriate input gradient distribution to overcome trainability issues.

3 Method

To prevent the PINN from converging to unstable fixed points, or at least reduce the likelihood of such convergence, we introduce an additional regularization term into the physics loss, designed to modify the loss landscape such that it penalizes predictions near or at unstable fixed points. The proposed regularization term consists of three factors: i) a factor $\mathcal{R}_{\text{LS}}$ that penalizes solutions x_θ for which the Jacobian indicates local instability at a given collocation point t_i; ii) a factor $\mathcal{R}_{\text{SE}}$ that ensures that this regularization is only active at collocation points at which the system has converged to a fixed point, i.e., for which $x'_\theta(t_i) = 0$, and iii) a decaying term C that gradually reduces the influence of regularization.

Stability can be easily determined by first converting higher-order ODEs into systems of first-order ODEs and then computing the eigenvalues of the Jacobian matrix obtained through a Taylor expansion around the fixed point. To penalize a candidate solution x_θ that is unstable at a collocation point t^*, we propose the regularization term

$$\mathcal{R}_{\text{LS}}(t^\star) = \sum_{\lambda \in \sigma(J(x_\theta(t^\star)))} \max\left(\text{Re}(\lambda), 0\right), \tag{3}$$

where $\sigma(J(x_\theta(t^\star))) = \{\lambda^\star_1, ..., \lambda^\star_n\}$ is the spectrum of the Jacobian matrix evaluated at $x_\theta(t^\star)$ and Re is the real part.

Since stability is a statement about fixed points, this regularization is not meaningful unless $x'_\theta(t^\star) = 0$. We thus multiply $\mathcal{R}_{\text{LS}}(t^\star)$ by a factor $\mathcal{R}_{\text{SE}}(t^\star)$ that is obtained by applying a Gaussian kernel to the time derivatives of the ODE:

$$\mathcal{R}_{\text{SE}}(t^\star) = \exp\left(-\|x'_\theta(t^\star)\|^2 / \varepsilon\right), \tag{4}$$

where ε is a hyperparameter that adjusts the sensitivity of the regularization around the fixed point. $\mathcal{R}_{\text{SE}}(t^\star)$ attains its maximum value of one if and only if all time derivatives vanish simultaneously, i.e., if $x_\theta(t^\star)$ is precisely at a fixed point. Candidate solutions converging to unstable fixed points are thus regularized by the term

$$\mathcal{R} = \frac{1}{N}\sum_{i=1}^{N} \left[\mathcal{R}_{\text{SE}}(t_i) \times \mathcal{R}_{\text{LS}}(t_i)\right], \tag{5}$$

where the regularization function is evaluated at the collocation points $\{t_i\}_{i=1}^N$ used to compute the physics loss.

Recognizing the importance of initialization in PINN training, cf. [15], it may suffice to steer away the candidate solution x_θ from unstable fixed points early during training. We thus propose a linear decaying scheme that turns off regularization after a certain percentage $\gamma \in (0, 1)$ of epochs:

$$C = \max\left(C_0 \left(\gamma - \frac{\text{epoch}}{N_{\text{epochs}}}\right), 0\right), \tag{6}$$

where C_0 denotes the initial regularization coefficient and N_{epochs} the total number of epochs. The regularization scheme is finally used in the total loss function via:

$$\mathcal{L} = \mathcal{L}_{\text{IC}} + \mathcal{L}_f + C \times \mathcal{R}. \tag{7}$$

4 Experimental Setup

In our analysis, we aim to demonstrate the impact of our regularization on the PINN training performance across four distinct dynamical systems, governed by first- or second-order ODEs in one- or two-dimensional settings. These systems include the pitchfork bifurcation, unforced Duffing oscillator, van der Pol oscillator, and Lotka-Volterra model.

PINN Settings. We use fully connected PINNs with four hidden layers, each containing 50 units, and employ the Swish activation function. Training is performed using the Adam optimizer with a learning rate of 0.001 and a decay rate of 1.0. During each training epoch, 1024 collocation points are randomly sampled to compute the physics loss (2), and the total number of epochs is set to $25k$. In the case of first-order systems, we use hard constraints for the initial conditions.

Evaluation Procedure. We train PINNs with and without regularization, for different ICs and simulation times, as convergence to local optima is highly sensitive to these settings, cf. [12]. To account for the inherent randomness in neural network initialization, we train 10 different PINNs for each configuration.

Our primary goal is to investigate a failure mode where the system converges to a local optimum with fundamentally different dynamics. Thus, rather than focusing on the accuracy of $x_\theta(t)$, we evaluate training success rates based on whether training converged to an incorrect local optimum or to the true solution $x_{\text{ref}}(t)$. The solution $x_{\text{ref}}(t)$ is obtained using the Runge–Kutta method; for the pitchfork bifurcation system an analytical solution is known (found in [12]). Training is considered successful if the L_2 relative error from the reference solution remains below 0.15.

Hyperparameters. We set the hyperparameter ε to 0.01, the initial regularization coefficient C_0 to 1.0 and the fraction of epochs γ to 0.5. We study the sensitivity of these hyperparameters in Section 5.1.

Candidate Systems: The **pitchfork bifurcation** system can be used to model population dynamics with negative growth effects at both low and high population levels. It is described by the first-order ODE $x' = x - x^3$ and has one unstable fixed point at $x^* = 0$ and two (asymptotically) stable fixed points at $x^* = 1$ and $x^* = -1$. As shown in [12], the unstable fixed point can cause severe convergence issues for PINNs when the simulation time is long and/or the ICs are close to the unstable point. The **unforced damped Duffing oscillator** is described by the second-order ODE $x'' + x' - x + x^3 = 0$, which gives an unstable saddle point at $(x, x')^* = (0, 0)$ and two (asymptotically) stable fixed points at $(x, x')^* = (\pm 1, 0)$. The **Van der Pol oscillator** [9] is a dynamical system described by the second-order ODE of the form $x'' - (1 - x^2)x' + x = 0$. The system exhibits stable oscillations, referred to in the literature as relaxation oscillations, and possesses a single unstable fixed point at $(x, x')^* = (0, 0)$. The **Lotka–Volterra equations** are a well-known system of first-order ODEs used to describe predator–prey dynamics in biological systems. While the

Table 1: Success rates for unmodified (left number) and regularized (right number) PINN training for different dynamical systems across different simulation times (T) and ICs. Bold numbers indicate the best success rate for the considered setting.

	$T = 11$	$T = 12$	$T = 13$	$T = 14$	$T = 15$
$x(0), x'(0)$	**Unforced Duffing Oscillator**				
(0.1, 0)	**0.5** / 0.4	0.0 / **0.5**	0.0 / **0.7**	0.0 / **0.5**	0.0 / **0.6**
(0.2, 0)	**0.9** / **0.9**	0.1 / **0.6**	0.0 / **0.3**	0.0 / **0.5**	0.0 / **0.7**
(0.3, 0)	0.9 / **1.0**	0.2 / **0.7**	0.0 / **0.3**	0.0 / **0.2**	0.0 / **0.6**
(0.4, 0)	**1.0** / **1.0**	0.2 / **0.6**	0.0 / **0.5**	0.0 / **0.5**	0.0 / **0.4**
(0.5, 0)	0.9 / **1.0**	0.7 / **0.8**	0.0 / **0.4**	0.0 / **0.8**	0.0 / **0.7**
$x(0), x'(0)$	**Van der Pol Oscillator**				
(0.1, 0)	0.0 / **1.0**	0.0 / **0.2**	0.0 / **0.6**	0.0 / **0.3**	0.0 / **0.1**
(0.2, 0)	0.6 / **1.0**	0.0 / **0.9**	0.0 / **0.6**	0.0 / **0.5**	0.0 / **0.2**
(0.3, 0)	0.9 / **1.0**	0.2 / **1.0**	0.0 / **1.0**	0.0 / **1.0**	0.0 / **0.4**
(0.4, 0)	**1.0** / **1.0**	0.8 / **1.0**	0.1 / **0.9**	0.0 / **1.0**	0.0 / **0.6**
(0.5, 0)	0.9 / **1.0**	0.6 / **1.0**	0.0 / **1.0**	0.0 / **0.9**	0.0 / **0.5**
$x(0), y(0)$	**Lotka–Volterra Model**				
(0.0, 0.1)	0.0 / **1.0**	0.0 / **1.0**	0.0 / **1.0**	0.0 / **1.0**	0.0 / **1.0**
(0.1, 0.0)	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0	0.0 / 0.0
(0.0, 0.5)	0.0 / **1.0**	0.0 / **1.0**	0.0 / **0.9**	0.0 / **0.8**	0.0 / **0.5**
(0.5, 0.0)	0.0 / **1.0**	0.0 / **1.0**	0.0 / **1.0**	0.0 / **0.7**	0.0 / **0.6**
(1.1, 1.1)	**1.0** / **1.0**	0.3 / **1.0**	0.0 / **1.0**	0.0 / **1.0**	0.0 / **1.0**

system's behavior can be complex depending on its specific formulation, in our work we consider a generalized version of the Lotka-Volterra equations:

$$x' = x(3 - x - 2y) \qquad y' = y(2 - x - y) \tag{8}$$

This system exhibits two (asymptotically) stable fixed points at $(x, y)^* = (0, 2)$ and $(x, y)^* = (3, 0)$, and two unstable fixed points at $(x, y)^* = (0, 0)$ and $(x, y)^* = (1, 1)$, with the latter being a saddle point.

5 Results

Training Success Rates. Table 1 shows the quantitative results of our experiments for unmodified PINN training and for our proposed regularization scheme (7) with default hyperparameters as reported in Section 4. We deliberately selected ICs and simulation times such that unmodified PINN training exhibits convergence problems (e.g., the simulation times were chosen longer for the pitchfork bifurcation system than in [12]). Furthermore, we did not apply training modifications proposed previously, such as loss weighting, adaptive collocation point sampling, etc., as they are orthogonal to our regularization scheme. In all of the unsuccessful training runs, training converged to the local optimum corresponding to the unstable fixed point of the dynamical system, confirming the results of [12] that unstable fixed points play a major role in PINN training. As the table shows, regularization substantially improves success rates for all considered systems and problem parameterizations, validating the effectiveness of our proposal even for longer simulation times.

Unforced Duffing and van der Pol Oscillators. For the unforced Duffing and the van der Pol oscillator we present an additional evaluation. Specifically, we fix the simulation time at $T = 12.5$ and $T = 15$ respectively and randomly sample 20 ICs from a Gaussian distribution centered at the unstable fixed point $(x, x')^* = (0, 0)$, with a covariance matrix $\Sigma = 0.25\mathbb{I}$. For the van der Pol system, we increased the number of epochs to $50k$. The simulated trajectories in the phase portrait are shown in Fig. 1. Consistent with the results in Table 1, the unmodified PINN frequently fails by converging to the unstable fixed point, whereas the regularized PINN exhibits much better performance. Indeed, the

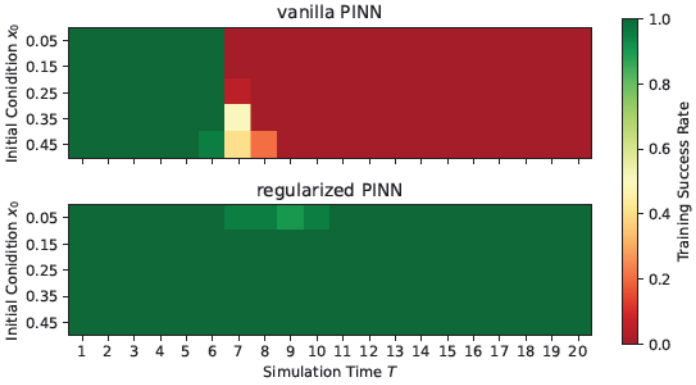


Figure 2: Success rates of unmodified (top) and regularized (bottom) training runs over different initial conditions x_0 and simulation times T.

training success rates are up from 15% to 70% for the Duffing and from 0% to 100% for the van der Pol oscillator.

Pitchfork Bifurcation. For the pitchfork bifurcation system, experiments were performed over a broader range of simulation times, $T \in \{1, \ldots, 20\}$, in order to better illustrate the combined effect of simulation time and initial condition on the training results. The results are visualized as a heatmap in Fig. 2. From the figure, we can observe that while the unmodified PINN training begins to fail at approximately $T = 7$, the regularized PINN continues to achieve successful training for all of the tested simulation times. Only for simulation times in the range $7 \geq T \geq 10$ with an IC of $x_0 = 0.05$, the regularized PINN exhibits a slight performance degradation, which we believe to be an artifact of the specific experimental settings and potentially resolvable through more optimized hyperparameter tuning.

Lotka-Volterra Model. For the Lotka–Volterra model, we uniformly sample 50 ICs from the interval $(-1, 4)^2$ along both x and y axis, and use a fixed simulation time of $T = 12.5$. The phase portrait for this system is presented in Fig. 1. While the proposed regularization is not as effective as in the two second-order ODEs, it still yields a notable improvement in training success rate from 62% to 78%.

5.1 Hyperparameter Sensitivity Analysis

We analyze the sensitivity of our regularization scheme w.r.t. the three hyperparameters C_0, ε and λ. Rather than varying all three hyperparameters simultaneously, which would be computationally expensive, in this experiment we keep two parameters fixed and vary the third.
We conduct the experiment on the Duffing oscillator using the IC $(x_0, x_0') = (0.01, 0.0)$ and simulation time of $T = 15$. While we believe that the optimal hyperparameters still depend on the considered system, its ICs, and the required simulation time, the results presented in Table 2 suggests that the performance of our regularization scheme depends only mildly on even large variations of the hyperparameters, and always improves upon unmodified PINN training (which in this case has a success rate of 0% for the considered 20 runs).

5.2 Ablation Study

Penalizing candidate solutions that converge to static equilibria via (4) is, when combined with the coefficient decay (6), a valid regularization scheme on its own – at least for systems that do not have (asymptotically) stable fixed points. In this section, we thus investigate the performance of PINN training according to (7), both with and without including the local stability term (3) in the rgularization loss.

Table 2: Hyperparameter sensitivity analysis for the proposed regularization scheme. Success rates are shown for 20 training runs for the Duffing oscillator with IC $(x_0, x_0') = (0.01, 0.0)$ and simulation time $T = 15$. Unmodified PINN training fails for this setting.

ϵ	γ	C	0.001	0.01	0.1	1.0	10.0	100.0	1000.0
0.01	0.5	**Success rate**	0	0.3	0.4	0.35	0.6	0.6	0.6

C	γ	ϵ	0.0001	0.001	0.01	0.1	1.0	10.0	100
1.0	0.5	**Success rate**	0.5	0.4	0.45	0.45	0.35	0.6	0.4

ϵ	C	γ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.01	1.0	**Success rate**	0.5	0.7	0.45	0.4	0.7	0.75	0.6	0.6	0.4

To this effect, we compare the two schemes on the Duffing oscillator, as its physical trajectories converges to one of the asymptotically stable fixed points, a property for which taking stability into account seems as the most sensible approach. We also compared the two schemes for the Van der Pol oscillator, a system for which the regularization scheme shown in (4) is, in theory, a valid scheme on its own as the system does not exhibit any stable fixed point. We perform the experiment on a single simulation time and IC, but optimized the PINN performance over all three hyperparameters in a grid search for each regularization scheme on each system separately. We chose a simulation time of $T = 15$ for both systems and ICs $(x_0, x_0') = (0.01, 0.0)$ for the Duffing oscillator and $(x_0, x_0') = (0.1, 0.0)$ for the Van der Pol oscillator. These settings were deliberately chosen to avoid perfect success rates for either approach, thereby creating challenging conditions that highlight meaningful differences between the two methods. Table 3 reports the highest success rates achieved during hyperparameter optimization. As shown, the best performance achieved by the combined loss function (7) is either comparable to (in the case of the van der Pol oscillator) or superior to (in the case of the unforced Duffing oscillator) that achieved by the static equilibrium regularization (4) alone. (There may not be an ordering between the two regularization schemes for fixed sets of hyperparameters, however.) Thus, we conclude that regularizing against unstable fixed points is necessary for systems that have asymptotically stable fixed points, and does not deteriorate performance for systems that do not.

Table 3: Comparing regularization with and without taking local stability $\mathcal{L}_{\text{LS}}$ into account. The success rate computed over 20 training runs, maximized over all hyperparameter combinations, is shown.

Unforced Duffing Oscillator ($T = 15$, $x_0 = 0.01$)		**Van der Pol Oscillator** ($T = 15$, $x_0 = 0.1$)	
Regularization	**Success**	**Regularization**	**Success**
–	0	–	0
$\mathcal{R}_{\text{SE}}$	0.6	$\mathcal{R}_{\text{SE}}$	0.7
$\mathcal{R}_{\text{SE}} \times \mathcal{R}_{\text{LS}}$	0.7	$\mathcal{R}_{\text{SE}} \times \mathcal{R}_{\text{LS}}$	0.7

6 Discussion, Limitation & Outlook

Training Difficulties due to Fixed Points. In [12], it was shown that fixed points—regardless of their local stability—correspond to local minima in the physics loss function, which can lead to severe convergence issues in PINNs, particularly for long simulation times and/or when ICs are close to these fixed points. Closely related are common training difficulties in PINNs caused by the zero solution $x^* = 0$, which is a trivial solution and fixed point for many dynamical systems. This solution is especially attractive due to common network initialization schemes that tend to bias the model toward it. Our experiments on the Lotka–Volterra model demonstrated that fixed points may cause convergence issues not only when the IC is close to them, but also when trajectories pass near them—even if the IC is set far away. This specific scenario is illustrated in the case of

Lotka-Volterra system in Fig. 1, where simulated trajectories are drawn toward the unstable fixed point at $(x, y)^* = (1, 1)$, despite originating from regions that are initially far away.

Overall Performance. Our experiments consistently highlight the prominent role of fixed points in training PINNs, demonstrating that the vanilla, unmodified PINN framework struggles to simulate dynamical systems in which fixed points interfere—either because trajectories originate near them or pass close to them during the simulation. With our proposed regularization scheme, introduced in Section 3, we specifically aim to mitigate these training difficulties by leveraging both the fixed point condition, which states that $f(x^*) = 0$, and the local stability information provided by the eigenvalues $\{\lambda_i\}$ of the Jacobian $J(x^*)$. Our results on various dynamical systems demonstrate that incorporating the regularization scheme into the overall loss function substantially improves training success rates. This improvement was shown quantitatively in Table 1 and qualitatively in Figures 1 and 2.

Limitations due to Saddle Points. Although the training success rate improved across many tested settings, our results also indicate that the proposed regularization scheme does not fully eliminate training difficulties. Manual inspection of the training process (not shown in the manuscript) revealed that many unsuccessful training outcomes—despite the proposed regularization being active—can be traced to the presence of saddle points, which constitute a specific class of unstable fixed points. Saddle points are characterized by the coexistence of stable and unstable manifolds: along certain directions (corresponding to negative eigenvalues of the Jacobian), trajectories are attracted toward the fixed point, while along others (positive eigenvalues), they are repelled (cf. Fig 1). This mixed stability behavior leads to a critical sensitivity to ICs. In particular, the stable manifold of a saddle point effectively acts as a *separatrix*, dividing the phase space into regions of qualitatively different dynamics. If the candidate solution $x_\theta(t)$ at network initialization lies on the "wrong side" of this separatrix, the regularization term—which is designed to repel trajectories from fixed points—may inadvertently drive the solution further into the repelling region. Once the regularization decays or is deactivated, the dynamics governed solely by the physical loss may guide the trajectory back toward the saddle point, ultimately resulting in convergence failure. In contrast, if the initial candidate solution lies on the "correct side" of the stable manifold, the regularization successfully steers the trajectory away from the saddle point, allowing it to follow the correct streamlines toward the physically meaningful solution. This limitation highlights the importance of network initialization in the presence of saddle-type fixed points, where the delicate geometry of the phase space can significantly affect training outcomes.

Outlook. Recognizing this limitation of the proposed regularization scheme motivates the discussion of combining it with existing techniques from the literature. For instance, sequence-to-sequence learning [6] or causality-respecting methods [14, 7, 3] can be naturally integrated with our approach to progressively extend the region of the simulation. These methods allow the training process to begin in a narrow temporal window around the IC and gradually expand toward the full simulation domain as training progresses. This approach could potentially circumvent the limitations introduced by saddle points, as their separatrices would not significantly influence the optimization during the initial training phase.

Finally, we would like to comment on the role of the static equilibrium regularization term (4) within the overall regularization framework. This component is designed to prevent PINN training from becoming trapped in trivial solutions that remain constant over time. Although its effectiveness was demonstrated only for systems governed by ODEs, we believe that it is also applicable to dynamical systems described by partial differential equations. Given that PINN training has also been shown to be adversely affected by steady-state solutions in PDE-based systems [12], the static equilibrium regularization may help push the solution away from such steady states during the early stages of training, thereby facilitating convergence toward transient, physically meaningful dynamics.

7 Conclusion

In this work, we introduced a regularization scheme for avoiding unstable fixed points when training PINNs on dynamical systems governed by ODEs. This approach proves especially useful in challenging scenarios, such as those involving long simulation times or initial conditions near unstable fixed points, where training would otherwise be difficult or infeasible. While this method does not entirely

eliminate convergence to local optima, it can be effectively combined with other strategies designed to mitigate such failure modes.

Although ODEs remain relevant in many fields, including control theory, systems biology, and mechanical engineering, the extension of this approach to partial differential equations (PDEs) is a compelling direction for future work. This is not a trivial step, as analyzing the stability of steady-state solutions in PDEs is significantly more complex. Nonetheless, our results demonstrate that explicitly addressing unstable fixed points can substantially improve PINN training, providing a promising foundation for broader applications.

References

[1] W.E. Boyce and R.C. DiPrima. *Elementary differential equations and boundary value problems*. 8th. Wiley New York, 2004.

[2] Arka Daw et al. "Mitigating Propagation Failures in Physics-informed Neural Networks using Retain-Resample-Release (R3) Sampling". In: *Proc. Int. Conf. on Machine Learning (ICML)*. Honolulu, Hawaii, USA, July 2023, pp. 7264–7302.

[3] Jia Guo, Haifeng Wang, and Chenping Hou. "A novel adaptive causal sampling method for physics-informed neural networks". In: *arXiv preprint arXiv:2210.12914* (2022).

[4] Katsiaryna Haitsiukevich and Alexander Ilin. "Improved training of physics-informed neural networks with model ensembles". In: *Proc. Int. Joint Conf. on Neural Networks (IJCNN)*. 2023, pp. 1–8.

[5] Ameya D Jagtap and George Em Karniadakis. "Extended physics-informed neural networks (XPINNs): A generalized space-time domain decomposition based deep learning framework for nonlinear partial differential equations". In: *Communications in Computational Physics* 28.5 (2020).

[6] Aditi Krishnapriyan et al. "Characterizing possible failure modes in physics-informed neural networks". In: *Advances in neural information processing systems* 34 (2021), pp. 26548–26560.

[7] Rambod Mojgani, Maciej Balajewicz, and Pedram Hassanzadeh. "Lagrangian PINNs: A causality-conforming solution to failure modes of physics-informed neural networks". In: *arXiv preprint arXiv:2205.02902* (2022).

[8] Ben Moseley, Andrew Markham, and Tarje Nissen-Meyer. "Finite basis physics-informed neural networks (FBPINNs): a scalable domain decomposition approach for solving differential equations". In: *Advances in Computational Mathematics* 49.4 (2023), p. 62.

[9] Balth. van der Pol Jun. "LXXXVIII. On "relaxation-oscillations"". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1926), pp. 978–992.

[10] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. "Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations". In: *Journal of Computational Physics* 378 (2019), pp. 686–707.

[11] Franz M. Rohrhofer et al. "Approximating Families of Sharp Solutions to Fisher's Equation with Physics-Informed Neural Networks". In: *Computer Physics Communications* 307 (2025). open-access: `arXiv:2402.08313 [cs.LG]`, p. 109422. DOI: 10.1016/j.cpc.2024.109422.

[12] Franz M. Rohrhofer et al. "On the Role of Fixed Points of Dynamical Systems in Training Physics-Informed Neural Networks". In: *Trans. Machine Learning Research* 1 (2023). Open-access.

[13] Sifan Wang, Shyam Sankaran, and Paris Perdikaris. "Respecting causality for training physics-informed neural networks". In: *Computer Methods in Applied Mechanics and Engineering* 421 (2024), p. 116813. DOI: 10.1016/j.cma.2024.116813.

[14] Sifan Wang, Shyam Sankaran, and Paris Perdikaris. "Respecting causality for training physics-informed neural networks". In: *Computer Methods in Applied Mechanics and Engineering* 421 (2024), p. 116813.

[15] Jian Cheng Wong et al. "Learning in Sinusoidal Spaces With Physics-Informed Neural Networks". In: *IEEE Transactions on Artificial Intelligence* 5.3 (2024), pp. 985–1000. DOI: 10.1109/TAI.2022.3192362.

Derivative-Enhanced Training for Data-efficient Surrogate Modeling

Paul Horvath[1,2], **Marian Staggl**[1,2], **Stefan Posch**[1,2]
[1] CD Laboratory for Physics-driven Machine Learning in Industrial Applications, Graz, Austria
[2] Institute of Thermodynamics and Sustainable Propulsion Systems,
Graz University of Technology, Austria
Corresponding author: Paul Horvath (paul.horvath@tugraz.at)

Abstract

Accurate surrogate modeling in engineering is often constrained by the high computational cost of generating training data from large scale numerical simulations. In many industrial applications, only a limited number of simulations can be afforded, which severely restricts the achievable surrogate accuracy, particularly in high dimensional parametric spaces. A promising approach to mitigate this curse of dimensionality is the incorporation of derivative information into surrogate training, which can be obtained efficiently via graph based implementations or adjoint calculations. This additional information captures local function structure, offering the potential to significantly improve data efficiency. In this work, we quantify the potential gains of derivative-enhanced training both theoretically and numerically, using a representative linear elasticity problem and an analytical benchmark. The findings provide guidance on the efficiency improvements achievable and the order of derivatives that yields the greatest benefit.

1 Introduction

The use of surrogate models is a state-of-the-art approach in modern engineering to avoid repeated evaluations of computationally expensive numerical simulations in applications such as optimization or uncertainty quantification [1]. However, the quantity of data required to train purely data-based surrogates, such as neural networks or Gaussian processes, can still be considerably high, leading to a large number of numerical simulations a priori. To reduce these data demands, several measures have been proposed to enhance the training of machine-learning-based surrogate models by injecting structure or auxiliary information. Among these, physics-informed neural networks (PINNs) [2] are particularly popular due to their relatively simple formulation and implementation. By using automatic differentiation (AD), PINNs introduce the residual of the governing differential equations as an additional loss term. In this way, PINNs can be used to identify unknown parameters of differential equations from data and to learn or even solve differentiable models by enforcing initial and boundary conditions through soft or hard constraints. PINN-based surrogates have been successfully applied to Bayesian inverse problems [3] and design optimization [4]. Complementary to physics-based regularization, another approach to improve data efficiency is to exploit derivative information available from simulations or adjoint solvers. Gradient-enhanced training, often referred to as Sobolev training [5], augments standard data fitting by incorporating sensitivity information, which can sharpen surrogate accuracy and generalization in low-data regimes [6]. In case of Gaussian processes (GP) for surrogate modeling, since differentiation is a linear operator, the derivative of a GP is another GP [7]. Thus, the use of derivative information to train GPs can have a significant influence on the data quantity requirements. Semler and Weiser [8] demonstrated the benefits of gradient-enhanced Gaussian process surrogates for inverse problems compared with GPs that rely solely on function values. They employed an adaptive selection of evaluation locations and tolerances

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

using a greedy heuristic, and assessed performance on both an analytical test case and a numerical example based on the finite element method. Their parameter reconstruction results consistently indicated that incorporating gradient information improves the quality of the inferred parameters, emphasizing the value of derivative data in surrogate-based inversion. Bouhlel and Martins [9] motivate gradient-enhanced Kriging with the availability of efficient sensitivities, e.g. from adjoint methods, but their engineering demonstrations rely on benchmark engineering functions rather than on gradients generated by a high-fidelity adjoint solver. Further prior work has shown that gradient information from high-fidelity solvers can substantially improve surrogate quality, both for neural surrogates and Gaussian-process-based models. In particular, gradient histories from adjoint-based optimization have been used to train multi-fidelity neural surrogates [10], while adjoint-based sensitivities have also been incorporated into gradient-enhanced deep GP aerospace-related applications [11]. More recent works extend this idea also to operator learning [12].

There are relevant works that use automatically differentiable simulators or differentiable physics for optimization and for training neural components [13], but explicit studies that use AD-generated gradients from PDE solvers as supervised signals for surrogate training remain comparatively scarce. This appears to be especially true for Gaussian-process-based surrogates, where gradient-enhanced formulations are well established, but the gradients are often not explicitly attributed to AD. Building on this perspective, we investigate gradient-enhanced strategies that combine differentiable physics solvers and AD to supply derivative information systematically. Two representative applications, namely the Rosenbrock function and a structural mechanics case, are considered to show how gradient augmentation affects sample-efficiency scaling in theory. Our results indicate that incorporating first-order derivatives provides a substantial and robust performance gain, whereas the value of higher-order derivatives is more problem-specific and depends on factors such as model smoothness.

1.1 Motivation

In many engineering applications, generating training data via numerical simulations represents the primary computational bottleneck, often taking significantly longer than the subsequent training of surrogate models. Standard data driven approaches require sufficiently large datasets to accurately capture complex physical behaviors, severely limiting their scalability. Consequently, our primary objective in applying derivative-enhanced modeling is to fundamentally reduce this data generation effort. By extracting more information from each individual simulation run, we aim to maximise data efficiency by reducing the high computational cost of acquiring training sets.

2 Methodology

To systematically evaluate this data efficiency, we introduce a theoretical scaling framework designed to weigh the information gain of higher order derivatives against the increased cost of their generation. Derivative-enhanced modeling enriches the surrogate by incorporating derivatives of the simulation output with respect to the input parameters. Because the underlying simulation model is formulated in a differentiable manner, these derivatives can be obtained efficiently via AD at a computational cost comparable to a single additional forward run. The following sections detail how this derivative information is incorporated into the respective training pipelines of neural network and Gaussian Process regression surrogates.

2.1 Sobolev training of neural networks

When learning a function u using classic neural network training, we typically aim to minimize a loss using function values $u(x_i)$ for training points x_i. Having access to first- or higher order derivative information, we extend this idea to a Sobolev space formulation by training not only with function values, but also matching first- or higher order derivatives of the target function [5]. Given a surrogate model $u_\theta(\mathbf{x})$ and reference data $u(\mathbf{x})$, the general Sobolev loss can be written as

$$\mathcal{L}_{\text{Sob}} = \sum_{k=0}^{r} \lambda_k \left\| \nabla^k (u_\theta(\mathbf{x}) - u(\mathbf{x})) \right\|_{L_p}, \tag{1}$$

where λ_k controls the relative weighting of the derivative information. Incorporating derivative terms constrains the local shape of the learned function and may reduce the number of required training samples, or increase model accuracy, as explained in section 2.3.

2.2 Gradient-enhanced GP regression

GP models naturally incorporate derivative observations by extending their covariance structure [7]. Because differentiation is a linear operator, the joint prior's covariance can be expanded to include not only function values but also their full gradient vectors by simply differentiating the base kernel $k(\mathbf{x}, \mathbf{x}')$. For an input space of dimension d, the complete extended joint covariance matrix combining function values and gradients can be expressed as [9]:

$$K_{\text{GE}}(\mathbf{x}, \mathbf{x}') = \begin{bmatrix} k(\mathbf{x}, \mathbf{x}') & (\nabla_{\mathbf{x}'} k(\mathbf{x}, \mathbf{x}'))^\top \\ \nabla_{\mathbf{x}} k(\mathbf{x}, \mathbf{x}') & \nabla_{\mathbf{x}} \nabla_{\mathbf{x}'}^\top k(\mathbf{x}, \mathbf{x}') \end{bmatrix}$$

Here, the off diagonal blocks represent the cross covariance column and row vectors between a function value and its gradients. The bottom right block captures the $d \times d$ covariance matrix between two gradient vectors via the mixed second derivatives. By augmenting the training data with these gradient observations, the GP posterior becomes more informative, leading to improved predictive accuracy and reduced uncertainty under limited data conditions.

2.3 Theoretical efficiency scaling

Evaluating the data efficiency of derivative-enhanced models requires weighing the information gained per calculation against its computational cost. This section theoretically assesses the potential for reducing the computational effort required to generate training data by establishing an equivalent information assumption. This approach suggests that each piece of extracted data, whether a function value or a partial derivative of an arbitrary degree, provides equally valuable information. Such an assumption holds perfectly, for instance, when fitting an n-coefficient polynomial surrogate to a multivariate polynomial system, which simply requires n pieces of information regardless of their source. Under this premise, the total information $I(d, r)$ extracted from one calculation is determined by the problem's dimensionality d and the maximum derivative degree r:

$$I(d, r) = \binom{d+r}{r} = \frac{(d+r)!}{d!r!}$$

A standard forward calculation ($r = 0$) yields exactly one piece of information. Including the gradient ($r = 1$) yields $1 + d$ pieces, and so forth. For a fair comparison, the escalating computational cost of higher-order derivatives must be accounted for. Assuming a single forward calculation has a normalized computational cost of $C = 1$ (requiring time Δt), evaluating the full gradient via AD also takes $\approx \Delta t$, resulting in a combined cost of $C = 2$. Computing the Hessian and higher-order derivatives requires differentiating the gradient entries, leading to cost scaling $\propto d^{r-1}$. The normalized computational cost $C(d, r)$ relative to a standard forward calculation is thus:

$$C(d, r) = 1 + \binom{d+r-1}{r-1} = 1 + \frac{r}{d+r} \frac{(d+r)!}{d!r!}$$

To estimate the computational savings of incorporating derivatives, we define $R(d, r)$ as the ratio of computational cost to the amount of information gained per sample:

$$R(d, r) = \frac{C(d, r)}{I(d, r)} = \frac{d!r!}{(d+r)!} + \frac{r}{d+r}$$

Effectively, $R(d, r)$ quantifies the data generation cost efficiency of using a derivative-enhanced method compared to relying solely on function values. Notably, for both gradients ($r = 1$) and Hessians ($r = 2$), this ratio is identical ($R = \frac{2}{d+1}$). However, for higher-order derivatives, the cost advantage deteriorates. Thus, only first and second derivatives appear theoretically beneficial for reducing simulation calls, though the vulnerability to numerical errors often reduces the practicality of evaluating the Hessian. As illustrated in Figure 1, the relative generation cost $R(d, r)$ reaches its minimum for first and second derivatives but worsens for higher orders. Furthermore, the efficiency advantage of the derivative-enhanced approach grows significantly as the dimensionality of the parameter space d increases.

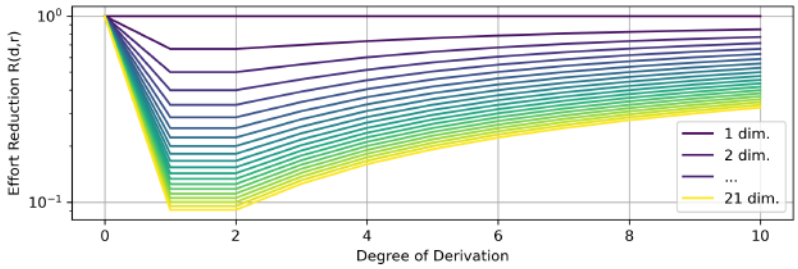


Figure 1: Reduction of computational effort following the assumption of equivalent information content with respect to different dimensions and derivative degrees.

3 Numerical studies

In this section, we assess how incorporating derivative information into surrogate model training impacts overall data demand. We evaluate the proposed approach using both a representative engineering application and an analytical benchmark. We investigate the effects of Sobolev training and gradient-enhanced Gaussian Process Regression (GE-GPR) on a structural example utilizing a differentiable finite element (FE) solver. This demonstrates the practical utility of derivative-enhanced learning within simulation environments, highlighting its potential to scale to industrial applications where finite element methods are already well established. Additionally, we use the Rosenbrock function as a benchmark to analyze how these methods perform on functions characterized by strong nonlinear characteristics. In order to evaluate how well the different models scale for higher dimensional parametric spaces, every model is trained for an increasing number of datasets while the accuracy of the trained models is evaluated (see Figure 2). The accuracy of each model is measured in terms of a R2 score, defined as

$$score = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}, \tag{2}$$

where $\hat{y}_i$ denotes the predicted values, y_i the reference solution, and $\bar{y}$ the mean of the reference data. A value of $score = 1$ corresponds to a perfect prediction. Increasing the number of training

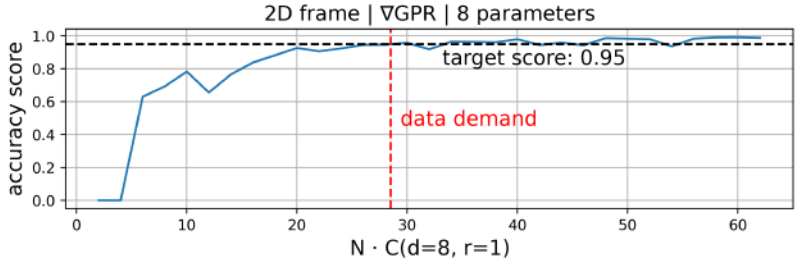


Figure 2: Computational effort of data generation for the eight parameter FE case using the gradient-enhanced GPR. The plotted computational cost represents the number of training samples multiplied by the evaluation cost per sample.

samples N generally improves model accuracy, however, it also increases the computational cost of data generation. We define total computational effort as the number of samples N multiplied by the relative cost per evaluation. Using a standard forward simulation as the baseline cost of 1, the total effort for standard data generation is simply N. Because extracting first order derivatives via AD requires roughly one additional pass through the computational graph, the effort per sample doubles, resulting in a total computational effort of $2N$. The figures below therefore present the obtained accuracy as a function of this total computational effort. From each study, the amount of data required

to reach a specified target accuracy is extracted, enabling the underlying data generation effort to be analyzed as a function of the parameter space dimensionality (see Figures 5 and 6).

3.1 Experimental Setup

Data Generation and Experimental Design. Datasets containing analytical derivatives are generated offline for both a structural FE problem (detailed in Section 3.2) and an analytical Rosenbrock benchmark. For the structural case, we use torchFEM, a differentiable FE library, utilizing mesh morphing to map a parameterized base mesh to target geometries. This system is defined by up to eight parameters: two representing the applied loads and six defining the geometry. For the analytical benchmark, a standard PyTorch implementation of the Rosenbrock function is utilized to evaluate scaling across varying dimensionalities. In both scenarios, input configurations are generated via uniform random sampling within predefined bounds. Following the forward evaluation of the target function, exact gradients and Hessians are extracted via AD. The target function is defined as the displacement magnitude at the load position for the FE model, and as the scalar output for the Rosenbrock case. As a preprocessing step, all input parameters and target function values are normalized via min-max scaling, and the corresponding gradients and Hessians are scaled accordingly. From a total generated dataset of 600 offline samples, 20% (120 samples) are reserved as a fixed test set to evaluate final model accuracy score and 10% (60 samples) are used for validation. Training subsets, ranging from 1 to 300 samples, are then randomly drawn from the remaining pool of 420 samples to systematically evaluate data demand and scaling behavior. Because the validation set is used exclusively for function value early stopping (requiring no gradients), it represents a constant across all NN models. Therefore, visualizations in the subsequent sections focus strictly on the computational effort required to generate the active training data.

Mesh Morphing. In order to deform the mesh of the parameterized FE example, a control-mesh is defined and wrapped over the computational mesh. This control mesh is discretized into triangular elements and uses the minimal number of nodes required to accurately define the geometric contours. During the setup of the parameterized model, each node of the FE mesh is mapped to a specific control element. When the control nodes are displaced, the internal FE nodes are updated via a piecewise linear transformation (see Fig. 3). Following this spatial deformation, the shape function derivatives and Jacobian determinants of the FE model are recomputed to reflect the new geometry.

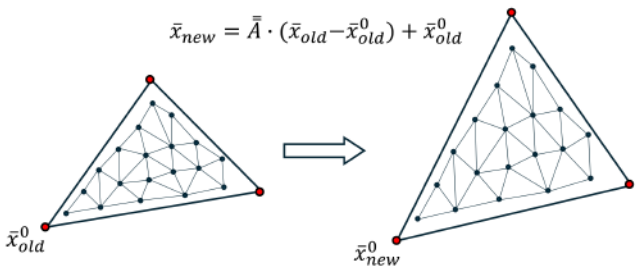


Figure 3: Mesh deformation methodology

GP Surrogate. The Gaussian Process regression models are implemented using GPyTorch. The model utilizes a constant mean and a Radial Basis Function (RBF) base kernel with varying length scales per dimension together with a scale kernel. The length scales are constrained to the interval [0.01, 1000]. A Gaussian likelihood is used, with the observation noise fixed at $1e-7$ and excluded from gradient updates to reflect the deterministic nature of the finite element solver. Model hyperparameters are optimized by maximizing the Exact Marginal Log-Likelihood using the Adam optimizer with a learning rate of 0.1 for 200 iterations.

NN Surrogate. The neural network surrogate is a compact Multilayer Perceptron (MLP) mapping input parameters to the target displacement. The architecture consists of two hidden layers with Tanh activation functions, utilizing four neurons per layer for the FE example and 16 for the Rosenbrock function. Training is executed with a learning rate of 0.005 and an L2 weight decay of 0.001. The Sobolev training loss function applies equal weighting (1.0) to the mean squared errors of the function values, gradients, and Hessians. The maximum number of training epochs is set to 100000, with an early stopping criterion triggered either upon reaching a validation score of 0.997 or after 10000 consecutive epochs without improvement.

3.2 FE example

As a structural reference model, a parameterized two dimensional frame structure is considered, as illustrated in Figure 4. The frame is fixed at its base, and an external force is applied at the upper right node. This system is governed by an eight dimensional parameter space: six geometric parameters control the individual bar thicknesses as well as the overall height and width of the frame, while two load parameters define the horizontal and vertical magnitudes of the applied force. The parameter bounds used for data generation can be found in Table 1. As the target function, the displacement magnitude at the load position is extracted. To evaluate surrogate performance, standard function only approaches are compared against the derivative-enhanced GE-GPR and Sobolev models across these respective training pipelines (see Sections 2.1 and 2.2).

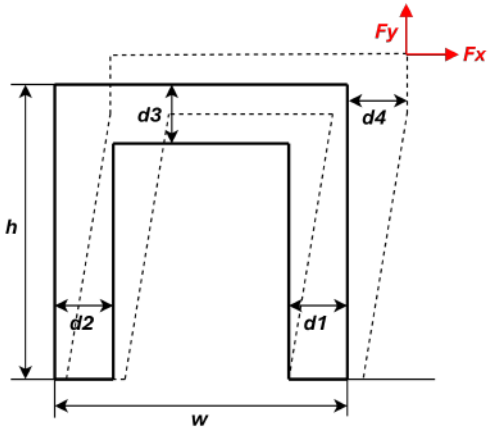


Figure 4: Illustration of the frame model with six geometric and two load parameters. The Young's modulus is $E = 1000$ MPa.

Table 1: Parameter bounds

Parameter	Min	Max	Unit
w	5.0	7.0	[m]
h	5.0	7.0	[m]
d1	0.5	1.5	[m]
d2	0.5	1.5	[m]
d3	0.5	1.5	[m]
d4	-0.5	0.5	[m]
Fx	0.1	1.0	[N]
Fy	0.1	1.0	[N]

Figure 5 presents the results for the GPR (left) and neural network (right) surrogates at target scores of 0.9 and 0.99. The blue and orange curves correspond to the classical GPR and the GE-GPR models, respectively. Similarly, the green and red curves represent the neural network surrogates trained with the standard loss formulation and the first order Sobolev loss. As expected, data demand increases exponentially with the number of parameters, and to some extent, this growth accelerates when higher target accuracies are required. The derivative-enhanced surrogates outperform their standard counterparts, requiring less computational effort in data generation to achieve the target scores. Moreover, the data demand based on theoretical efficiency scaling, introduced in Section 2.3, is plotted for both models (black dotted line). This line represents the theoretical reduction in computational effort based on the equivalent information assumption. The observed data demand for the neural network trained with first order Sobolev loss follows this theoretical prediction, demonstrating an approximate $2/(1+d)$ reduction in computational effort compared to the classical NN model.

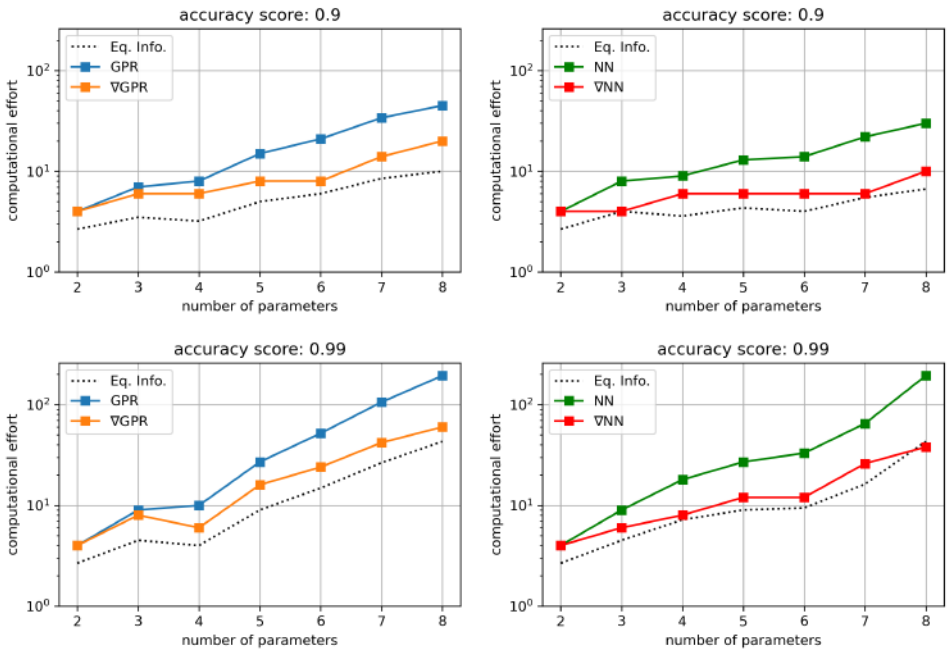


Figure 5: Computational-effort-corrected performance of surrogate models on the parameterized FE example for target accuracies of 0.9 and 0.99 (left: GPR, right: neural networks). Incorporating first order derivatives successfully reduces data generation effort. Furthermore, the neural network trained with first order sobolev loss exhibits similar scaling behavior to the theoretical baseline of the theoretical equivalent information assumption (black dotted line).

3.3 Rosenbrock function

As a second example, we consider the n-dimensional Rosenbrock function evaluated over the domain $[-1, 1]^n$. Although the computational effort required for data generation is negligible in this case, the problem provides a convenient benchmark for analyzing the proposed methods on a smooth yet highly nonlinear function. Evaluating the same surrogate models for target scores of 0.9 and 0.99 (Figure 6), we observe consistent trends: the first order derivative-enhanced surrogates (see red and orange curves) outperform their standard counterparts, lowering the required data generation effort. Furthermore, much like the neural network in the FE example, the GE-GPR closely follows the theoretical efficiency scaling (black dotted line) at the higher target accuracy. Besides first order derivatives, we also investigated the influence of incorporating Hessian information via a second order Sobolev loss for training the neural network. While the Hessian-informed model (purple curve) required more computational effort than the standard model at lower target accuracies and parameter dimensions, its data efficiency improved in higher dimensional spaces and at increased accuracy thresholds. However, it did not surpass the overall data efficiency of the first order Sobolev model. Exploring even higher dimensional parameter spaces may eventually reveal a comparative advantage for this second order approach.

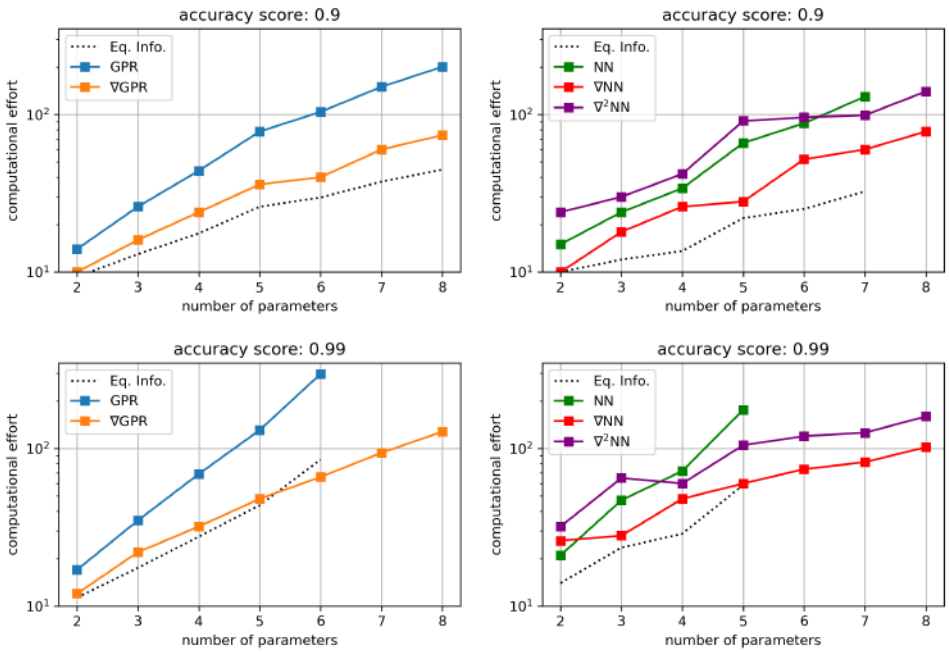


Figure 6: Computational-effort-corrected performance on the Rosenbrock function for target accuracies of 0.9 and 0.99 (left: GPR, right: neural networks). Consistent with the FE example, first order derivative-enhanced models outperform standard counterparts at higher dimensions. Here, the GE-GPR model follows the overall trend of the theoretical efficiency scaling (black dotted line). The Hessian-informed neural network demonstrates improved efficiency at higher accuracies but does not surpass the first order model.

4 Discussion and conclusion

This study investigates the benefits of enhancing surrogate models with derivative information obtained from differentiable numerical simulations to reduce data demand. By formulating simulation models in a differentiable manner, derivatives of the simulation output with respect to input parameters can be obtained efficiently via AD. A critical challenge, however, is determining the optimal derivative order to incorporate and balancing the reduced data demand against the computational cost of generating this data. To systematically evaluate these trade offs, a theoretical model is introduced in Section 2.3 and two benchmark problems are investigated: an FE-based structural example in Section 3.2 and the Rosenbrock function in Section 3.3. To evaluate surrogate performance, we measure the computational effort of generating data required to reach a specified target score for problems with up to eight parameters. The incorporation of higher order derivative information into surrogate training is, however, not straightforward. For example, the computational cost of fitting a GPR model scales with approximately $\mathcal{O}(N^3)$, where N denotes the total number of observations. For a first order GE-GPR model, this includes all function evaluations and gradients, causing the data vector to scale linearly with the parameter dimension. Extending this to second order methods incorporates the Hessian matrix, causing the observation count per sample to scale quadratically with the number of parameters, severely bottlenecking the training process. While Sobolev training of neural networks scales more efficiently with higher order derivatives, balancing the corresponding loss terms remains a significant challenge. In the Rosenbrock example, we observed that incorporating Hessian information yielded a benefit over the standard model at high target accuracies, however, it did not surpass the overall efficiency of the first order models. The successful application of higher order Sobolev training to structural examples remains an open subject for future investigation. Overall, our numerical studies show that using first order derivatives in surrogate training can reduce the computational effort of generating data compared to the standard training methods. The results therefore highlight the potential of such approaches to scale well to industrial sized problems, particularly for applications constrained by the high cost of simulation data.

Acknowledgments and Disclosure of Funding

The financial support by the Austrian Federal Ministry of Economy, Energy and Tourism, and the Christian Doppler Research Association is gratefully acknowledged.

References

[1] G.I. Schuëller. On the treatment of uncertainties in structural mechanics and analysis. *Computers & Structures*, 85(5):235–243, 2007. Computational Stochastic Mechanics.

[2] M. Raissi, P. Perdikaris, and G.E. Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378:686–707, 2019.

[3] Yongchao Li, Yanyan Wang, and Liang Yan. Surrogate modeling for bayesian inverse problems based on physics-informed neural networks. *Journal of Computational Physics*, 475:111841, 2023.

[4] Yubiao Sun, Ushnish Sengupta, and Matthew Juniper. Physics-informed deep learning for simultaneous surrogate modeling and pde-constrained optimization of an airfoil geometry. *Computer Methods in Applied Mechanics and Engineering*, 411:116042, 2023.

[5] Wojciech M Czarnecki, Simon Osindero, Max Jaderberg, Grzegorz Swirszcz, and Razvan Pascanu. Sobolev training for neural networks. *Advances in neural information processing systems*, 30, 2017.

[6] Rhys Newbury, Jack Collins, Kerry He, Jiahe Pan, Ingmar Posner, David Howard, and Akansel Cosgun. A review of differentiable simulators. *IEEE Access*, 12:97581–97604, 2024.

[7] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[8] Phillip Semler and Martin Weiser. Adaptive gradient-enhanced gaussian process surrogates for inverse problems. *Mathematical Optimization for Machine Learning: Proceedings of the MATH+ Thematic Einstein Semester 2023*, page 59, 2025.

[9] Mohamed A Bouhlel and Joaquim RRA Martins. Gradient-enhanced kriging for high-dimensional problems. *Engineering with Computers*, 35(1):157–173, 2019.

[10] Tao Zhang, Mark Woodgate, George Barakos, and Yu Luo. A multi-start aerodynamic shape optimisation approach via multi-fidelity neural networks. *Aerospace Science and Technology*, 168:111006, 2026.

[11] Viv Bone, Chris van der Heide, Kieran Mackle, Ingo Jahn, Peter M. Dower, and Chris Manzie. Gradient-enhanced deep gaussian processes for multifidelity modeling. *Journal of Computational Physics*, 520:113474, 2025.

[12] Chanik Kang, Joonhyuk Seo, Ikbeom Jang, and Haejun Chung. Adjoint method-based fourier neural operator surrogate solver for wavefront shaping in tunable metasurfaces. *iScience*, 28(1):111545, 2025.

[13] Philipp Andelfinger. Differentiable agent-based simulation for gradient-guided simulation-based optimization. In *Proceedings of the 2021 ACM SIGSIM Conference on Principles of Advanced Discrete Simulation*, pages 27–38, 2021.

Towards a PIRL framework for efficient airflow diffuser design

Alfredo López
SCCH Software Competence Center Hagenberg
Softwarepark 32a, 4232 Hagenberg, Austria
alfredo.lopez@scch.at

Florian Sobieczky
SCCH Software Competence Center Hagenberg
Softwarepark 32a, 4232 Hagenberg, Austria
florian.sobieczky@scch.at

Christopher Lackner
CERBSim GmbH
Taubstummengasse 11. 1040 Vienna, Austria
clackner@cerbsim.com

Matthias Hochsteger
CERBSim GmbH
Taubstummengasse 11. 1040 Vienna, Austria
mhochsteger@cerbsim.com

Bernhard Scheichl
Institute of Fluid Mechanics and Heat Transfer
TU Wien, BA Tower/E322, 7th floor, Getreidemarkt 9, 1060 Vienna, Austria
bernhard.scheichl@tuwien.ac.at

Helmuth Sobieczky
Institute of Fluid Mechanics and Heat Transfer
TU Wien, BA Tower/E322, 7th floor, Getreidemarkt 9, 1060 Vienna, Austria
hesobi@gmail.com

Christoph Feichtinger
Windpuls Gmb
Wiener Strasse 131, 4020 Linz, Austria
c.feichtinger@windpuls.com

Abstract

This extended abstract presents a physics-informed reinforcement learning framework for optimal diffuser design to improve airflow homogeneity upstream of a heat exchanger. This approach addresses key challenges in simulation-based optimization, including high-dimensional design spaces, expensive CFD evaluations, and the lack of gradient information. Physics-based flow features related to early pressure loss occurrence and eddy formation were employed as low-cost proxies

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

for the target homogeneity objective. The problem is formulated as a partially observable Markov decision process in which the agent sequentially selects the geometries to be evaluated. Using an expected improvement reward function, the method adaptively balances exploration and exploitation. The approach is demonstrated on a synthetic one-dimensional example, and a two-dimensional diffuser optimization problem is presented.

1 Introduction

In recent years, there has been a growing interest in integrating reinforcement learning (RL) into optimization algorithms. Rather than adhering to a predefined set of optimization rules, the agent learns the optimization parameters and search strategy through interactions with the objective function [6, 18, 7, 16, 8]. On the other hand, the incorporation of physical information into RL has helped bridge the gap between simulated environments and real-world applications by providing a physics-based mathematical framework for addressing key RL challenges, such as exploration in high-dimensional search spaces and the design of meaningful reward functions [4, 9].

The aim of the research project [1] is to utilize physics-informed reinforcement learning (PIRL) to design an optimal channel geometry that directs the (incompressible) airflow produced by a ventilator towards a heat exchanger, with the goal of maximizing the homogeneity of the flow before it reaches the exchanger. This extended abstract formulates the PIRL optimization problem for an airflow diffuser shape design as a Partially Observable Markov Decision Process (POMDP). As each CFD run is computationally expensive (ranging from approximately 3 minutes in the present 2D example to several hours in the ongoing industrial application), each new shape iteration must be selected carefully. Within this framework, the design of new shapes is treated as a sequential decision-making process in which each CFD evaluation provides information that can either reduce the uncertainty or improve the objective function. The proposed methodology is illustrated using a synthetic one-dimensional example, and an ongoing application to diffuser shape optimization is outlined.

2 Problem formulation

2.1 Physics-informed black-box optimization

Black-box optimization (BBO) [3, 2] optimizes objective functions whose gradients and internal structures are either unavailable or too complex to handle. In this study, we address the black-box optimization problem

$$\max_{x \in X} f_{\text{obj}}(x), \tag{1}$$

where $X \subset \mathbb{R}^d$ is a finite search space, and $f_{\text{obj}} : \mathbb{R}^d \to \mathbb{R}$ is a continuous objective function. The derivatives and internal structure of f_{obj} are inaccessible, and each evaluation $x \mapsto f_{\text{obj}}(x)$ requires a costly computer simulation. We also consider a physics-informed feature mapping $f_{\text{feat}} : \mathbb{R}^d \to \mathbb{R}^p$, which extracts low-dimensional features $f_{\text{feat}}(x)$ that capture physically meaningful characteristics that influence the objective function. We also assume f_{feat} is computationally inexpensive; therefore, an additional evaluation of this mapping does not constitute a significant overhead. We then define the vector-valued function

$$f(x) = \big(f_{\text{obj}}(x), f_{\text{feat}}(x)\big), \tag{2}$$

which jointly represents costly objective values and inexpensive physics-informed features. Our aim is to develop a computationally efficient iterative optimization strategy that, after n iterations, generates a sequence of query points $X_n := (x_t)_{t=0}^n \subseteq X$, such that X_n leads to a near-optimal design, that is, $\max_{x \in X_n} f_{\text{obj}}(x) \approx \max_{x \in X} f_{\text{obj}}(x)$.

2.2 POMDP

POMDPs provide a mathematical framework for RL problems in which the environment state is hidden from the agent and can only be observed indirectly [17]. A POMDP is defined as a tuple (S, A, Ω, T, O, R), where S, A, and Ω are the state, action, and observation spaces, respectively. The dynamics are specified by a transition model T, an observation model O and a reward function R.

The agent-environment interaction generates a random trajectory $s_0, a_1, o_1, r_1, s_1, a_2, o_2, r_2, s_2, \ldots$, where the initial state is drawn from an initial belief distribution $s_0 \sim b_0$. At each time step $t = 1, 2, \ldots$, the agent selects an action $a_t \sim \pi(h_t)$ according to a policy π that depends on the agent's history $h_t = (a_1, o_1, \ldots, a_{t-1}, o_{t-1})$. The environment then transitions to a new state according to the distribution $s_t \sim T(s_{t-1}, a_t)$, after which the agent receives an observation $o_t \sim O(s_t, a_t)$ and a reward $r_t = R(s_{t-1}, a_t)$. The agent then updates its history as $h_{t+1} = (h_t, a_t, o_t)$ and maintains a belief state $b_t(s) = P(s_t = s \mid h_t)$, and the process is repeated. The objective is to find a policy π that maximizes the expected discounted return $V_\pi(h) = \mathbb{E}\left[\sum_{k=t}^{t+m-1} \gamma^{k-t} r_k \,\middle|\, h_t = h\right]$, where $\gamma \in (0, 1]$ is the discount factor and $m \geq 1$ is a (possibly infinite) time horizon.

2.3 Black-box optimization as a POMDP

The optimization process is modeled as a partially observable Markov decision process (POMDP), where the latent state includes the unknown function f given in eq. (2). Because f cannot be observed directly, the agent only has access to noisy evaluations. At each time step $t = 1, 2, \ldots$, the agent selects a design point $x_t \in X$ and observes a noisy evaluation $(y_t, z_t) = f(x_t) + \varepsilon_t$. The reward r_t encourages improvements over the best objective value observed so far. More formally, the induced POMDP is defined as follows:

Initial belief $b_0 = \mathcal{GP}(0, k)$, a multi-output Gaussian process prior over the unknown vector-valued response function $f = (f_{\text{feat}}, f_{\text{obj}})$ and k is a kernel function.

State space $S = C(\mathbb{R}^{d+1}) \times X$, where $s_t = (f, x_t^*)$ consists of the latent response function f and the current best design x_t^* among the points where the objective has been evaluated.

Action space $A = X$, the finite set of candidate query points to be sampled.

Observation space $\Omega = \mathbb{R}^{p+1}$ corresponds to (possibly noisy) vector function evaluations.

Transition model The latent function remains unchanged, and the best design is updated $T(f, x_{t-1}^*, x_t) = (f, x_t^*)$, where $x_t^* = \arg\max_{x \in \{x_{t-1}^*, x_t\}} f_{\text{obj}}(x)$.

Observation model After selecting x_t, the agent observes an eventually noisy evaluation of the form $o_t = (y_t, z_t) = f(x_t) + \sigma\varepsilon_t$ with $\varepsilon_t \sim \mathcal{N}(0, \Sigma)$.

Reward function The reward corresponds to the improvement over the previous best value, $R(f, x_{t-1}^*, x_t) = \max\big(0, f_{\text{obj}}(x_t) - f_{\text{obj}}(x_{t-1}^*)\big)$.

The following proposition shows that, under a single-step horizon and noiseless observations, our POMDP formulation coincides with standard Bayesian optimization using the expected improvement acquisition function [10, 5]. For conciseness, we omit the proof.

Proposition 1 (Bayesian Optimization). *If $m = 1$, $\gamma = 1$ and $\sigma = 0$, then the optimal policy is $\pi^*(h_t) = \arg\max_{x \in X} q(x_{t-1}^*, x)$ where*

$$q(x_{t-1}^*, x) = \mathbb{E}\left[\max\big(0, f_{\text{obj}}(x) - f_{\text{obj}}(x_{t-1}^*)\big) \,\middle|\, \{(x_k, y_k, z_k)\}_{k=0}^{t-1}\right],$$

The expectation is taken under the Gaussian process posterior distribution of f_{obj} conditioned on the past evaluations $\{(x_k, y_k, z_k)\}_{k=0}^{t-1}$.

3 Experiments and analysis

In this section, we first demonstrate and evaluate the proposed method using a one-dimensional synthetic example. We then present an ongoing real-world application of diffuser shape optimization.

3.1 Illustrative 1D-example

We apply the proposed method to maximize the objective function $f_{\text{obj}}(x) = -\sin(x) - x^2 + 0.7x$ over the discrete domain X, consisting of an equidistant grid of $n = 101$ points between -1 and 2 (Figure 1). As a feature function, we consider $f_{\text{feat}}(x) = f_{\text{obj}}(x) + x$, which serves as a simple, highly correlated proxy for auxiliary information. In practical CFD settings, such features would typically not follow an explicit functional relationship with the objective but instead arise from physically correlated quantities. . To highlight the potential benefits of incorporating additional

physical information, we compared the proposed framework in two settings: without additional features ($f = f_{\text{obj}}$) and physics-informed ($f = (f_{\text{obj}}, f_{\text{feat}})$). For illustrative purposes, we assume noiseless observations ($\sigma = 0$) and consider a one-step time horizon ($m = 1$).

The process is initialized with a set of six exploratory actions $X_6 = \{-1, 0.41, 0.8, 1.01, 1.61, 1.91\}$, leading to the associated initial observations $Y_6 = \{f_{\text{obj}}(x) : x \in X_6\}$ and $Z_6 = \{f_{\text{feat}}(x) : x \in X_6\}$. The current best design is $x_6^* = 1.01$. In practice, such initial points can be selected based on expert knowledge.

Under this setting, we solve our POMDP problem using the Partially Observable Monte Carlo Planning (POMCP) method introduced in [15] and implemented in the Python library `pomdp_py` [19]. Figure 1 compares the results of the featureless (left column) and physics-informed (right column) cases. The upper row shows the objective function f_{obj} (orange dashed line), sampled objective data (crosses), sampled feature data (circles), and the expected value (blue line) and confidence interval (light blue area) of the posterior distribution $b_t(f_{\text{obj}}) = \mathbb{E}[f_{\text{obj}} \mid X_6, Y_6]$. The second row shows the action-value function $x \mapsto q(x_6^*, x)$ defined in Proposition 1. The vertical lines in the second row correspond to the next point to be queried, x_7, as determined by the algorithm, which coincides with $\arg\max_{x \in X} q(x_6^*, x)$ as stated in Proposition 1. Observe that, in the featureless case, the next queried point x_7 corresponds to a local minimizer of f_{obj}. In contrast, when additional physical information is incorporated, the selected point x_7 coincides with the global minimizer of f_{obj}.

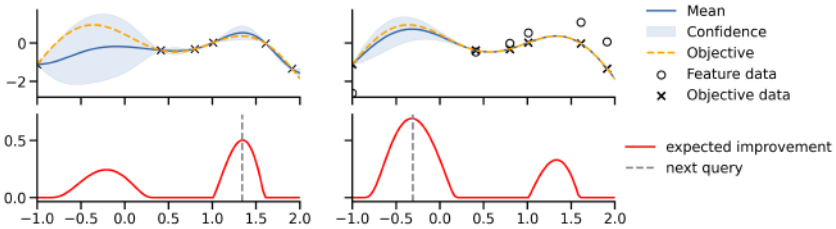


Figure 1: POMDP black-box optimization applied to a 1D example for the cases without features (left column) and physics-informed features (right column).

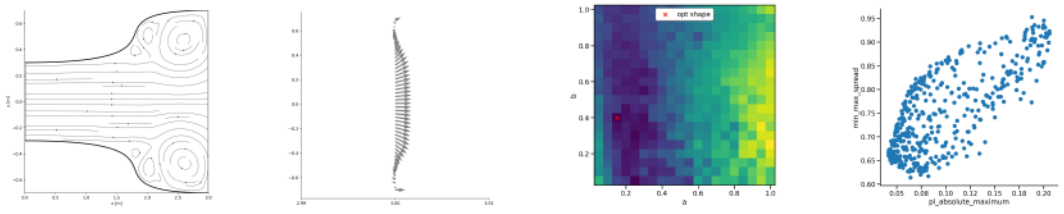
3.2 Towards diffuser shape optimization

This section outlines the application of the proposed framework to diffuser shape optimization. As this work is presented as an extended abstract, we limit the description to the main ideas and omit the detailed technical aspects.

A two-dimensional symmetric diffuser contour shape is parameterized by a vector $x \in [0, 1]^2$, which determines the position of the Breezier control point. The resulting flow field was evaluated through computational fluid dynamics (CFD) simulations based on a mass-conserving mixed stress formulation for time-dependent implicit large eddy simulations [12, 11]. Figure 2a illustrates the diffuser shape for $x = (1, 0.25)$, along with a snapshot of the corresponding velocity field. Each CFD evaluation is computationally expensive, and gradient information is unavailable, therefore this problem naturally falls within the scope of black-box optimization. We formulate the shape optimization task as in (1), where f_{obj} corresponds to a suitable homogeneity metric of the outlet velocity field, and $X \subset [0, 1]^2$ is a 20×20 regular grid of candidate designs. The heatmap in 2c shows $f_{\text{obj}}(x)$ for $x \in X$ and the optimal shape (red cross).

Modeling flow fields is challenging due to the high dimensionality of CFD data and their strong variability across geometries. Therefore, an important step in physics-informed machine learning is to extract low-dimensional, physics-based features from CFD data, guided by domain expertise [14, 13]. In our case of airflow channels, such features are selected based on their ability to capture flow regimes of enhanced mixing, which correlates with the objective of homogeneity at the outlet. More precisely, we assesed the pressure drop of the ideal pressure recovery at the outlet (estimated for a virtually steady inviscid and irrotational flow using Bernoulli's equation) and the actual one found by CFD (finite-elements) simulations, including RANS/temporal averaging. A samller pressure loss indicates a more developed turbulent regime, which enhance momentum redistribution and results in a more homogeneous RANS-averaged flow at the outlet.

In this context, the feature map f_{feat} is defined by a vector derived from the early pressure loss and vortex formation features. These features were identified by combining expert knowledge and exploratory data analysis. We obtained a set of more than ten candidate features that exhibited a strong correlation with the target and were therefore included in feature mapping. For instance, Figure 2d shows a scatter plot of $f_{\text{obj}}(x)$ against the maximum absolute value of the pressure loss associated with x for all $x \in X$. Now having all the building blocks required to apply our POMDP framework to diffuser shape optimization, we can define an appropriate covariance kernel structure for the function $f = (f_{\text{obj}}, f_{\text{feat}})$ so as to extend the computational framework from a one-dimensional search space to a multi-dimensional setting.



(a) (b) (c) (d)

Figure 2: (a) Diffuserhannel geometry and a snapshot of the velocity field for a representative shape; (b) velocity field at the diffuser outlet; (c) heatmap of the objective function over the shape search space; (d) scatter plot of the objective value versus a low-cost physical feature.

4 Conclusion and Outlook

We defined a POMDP-based black-box optimization framework that is suitable for airflow channel design optimization problems. The proposed method leads to an iterative process in which each evaluated new shape provides information by reducing uncertainty or improving the objective function, thereby fulfilling the exploration–exploitation paradigm. Our roadmap consists of the following tasks:

- Extend the one-dimensional search space case to the multidimensional setting and evaluate the method on the application described in Section 3.2.
- Extend the action space to pairs of the form $a_t = (x_t, d_t)$, where d_t is a binary decision variable that determines the evaluation fidelity: if $d_t = 0$, only the inexpensive feature map $f_{\text{feat}}(x_t)$ is evaluated, whereas if $d_t = 1$, both $f_{\text{feat}}(x_t)$ $f_{\text{obj}}(x_t)$ are evaluated.
- Apply the method to a real industrial diffuser shape optimization problem, where geometries are described by more than 30 parameters and each CFD evaluation requires several hours of computation on high-performance computing systems.

Finally, we note that the proposed framework can also be applied to optimization problems beyond shape design that may benefit from the incorporation of physics-informed features.

Acknowledgments and Disclosure of Funding

The main contributor to this research was the FFG (Austrian Applied Research Fund) project "AIrFoil" (FFG-No. 915010). The research reported in this paper has also been partly funded by the Federal Ministry for Innovation, Mobility and Infrastructure (BMIMI), the Federal Ministry for Economy, Energy and Tourism (BMWET), and the State of Upper Austria in the frame of the SCCH competence center INTEGRATE [(FFG grant no. 892418)] in the COMET - Competence Centers for Excellent Technologies Programme managed by Austrian Research Promotion Agency FFG.

References

[1] AIrFoil. Airfoil — ffg-project no. 915010 (austrian applied research foundation). https://projekte.ffg.at/projekt/5125991, 2024. FFG Projektdatenbank - Effizienzsteigerung und Lärmminderung von Luftwärmepumpen.

[2] Stéphane Alarie, Charles Audet, Aïmen E. Gheribi, Michael Kokkolaras, and Sébastien Le Digabel. Two decades of blackbox optimization applications. *EURO Journal on Computational Optimization*, 9:100011, 2021.

[3] Charles Audet and Warren Hare. *Derivative-Free and Blackbox Optimization.* Springer Series in Operations Research and Financial Engineering. Springer International Publishing, Cham, 2017.

[4] Chayan Banerjee, Kien Nguyen, Clinton Fookes, and Maziar Raissi. A survey on physics informed reinforcement learning: Review and open problems. *Expert Systems with Applications*, 287:128166, 2025.

[5] Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *ArXiv*, abs/1012.2599, 2010.

[6] Camille Castera and Peter Ochs. From learning to optimize to learning optimization algorithms, 2024.

[7] Tianlong Chen, Xiaohan Chen, Wuyang Chen, Howard Heaton, Jialin Liu, Zhangyang Wang, and Wotao Yin. Learning to optimize: A primer and a benchmark. In *AutoML Conference 2023 (Journal Track)*, 2023.

[8] Mujin Cheon, Jay H. Lee, Dong-Yeun Koh, and Calvin Tsay. EARL-BO: Reinforcement Learning for Multi-Step Lookahead, High-Dimensional Bayesian Optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 10182–10198. PMLR, 2025.

[9] Thomas P. Dussauge and Chong-Kun Sung. A reinforcement learning approach to airfoil shape optimization. *Scientific Reports*, 13(1):9753, Jun 2023.

[10] Roman Garnett. *Bayesian Optimization*. Cambridge University Press, 2023.

[11] Jay Gopalakrishnan, Philip L. Lederer, and Joachim Schöberl. A mass conserving mixed stress formulation for the stokes equations. *IMA Journal of Numerical Analysis*, 40(3):1838–1874, 2020.

[12] Philip L. Lederer, Xaver Mooslechner, and Joachim Schöberl. High-order projection-based upwind method for implicit large eddy simulation. *Journal of Computational Physics*, 493: 112492, 2023.

[13] Riccardo Margheritti, Onofrio Semeraro, Maurizio Quadrio, and Giacomo Boracchi. Feature Extraction from Flow Fields: Physics-Based Clustering and Morphing with Applications. *Applied Sciences*, 15(23):12421, November 2025.

[14] Pushan Sharma, Wai Tong Chung, Bassem Akoush, and Matthias Ihme. A review of physics-informed machine learning in fluid mechanics. *Energies*, 16(5), 2023.

[15] David Silver and Joel Veness. Monte-carlo planning in large pomdps. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[16] Lei Song, Chen-Xiao Gao, et al. Reinforced in-context black-box optimization. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2025*, pages 7237–7245, 2025.

[17] Matthijs T. J. Spaan. *Partially Observable Markov Decision Processes*, pages 387–414. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.

[18] Michael Volpp, Lukas P. Fröhlich, Kirsten Fischer, Andreas Doerr, Stefan Falkner, Frank Hutter, and Christian Daniel. Meta-learning acquisition functions for transfer learning in bayesian optimization. In *International Conference on Learning Representations (ICLR)*, 2020.

[19] Kaiyu Zheng and Stefanie Tellex. pomdp_py: A framework to build and solve pomdp problems. In *ICAPS 2020 Workshop on Planning and Robotics (PlanRob)*, 2020.

Understanding the Role of Domain Knowledge in Bayesian Optimization under Small-Data Constraints

Bernd Schuscha
Department Simulation
Materials Center Leoben Forschung GmbH
8700 Leoben, Austria
bernd.schuscha@mcl.at

Franz M. Rohrhofer
Department for Methods & Algorithms for AI
Know-Center Research GmbH
8010 Graz, Austria

Daniel Scheiber
Department Simulation
Materials Center Leoben Forschung GmbH
8700 Leoben, Austria

Bernhard C. Geiger
Know-Center Research GmbH
Signal Processing and Speech Communication Laboratory
Graz University of Technology
8010 Graz, Austria

Bayesian optimization (BO) is a sequential model-based optimizing algorithm for expensive black-box functions. The core is a probabilistic surrogate model, typically a Gaussian process, that is updated as new evaluations are collected. At each iteration, an acquisition function balances exploration of uncertain regions against exploitation of promising areas, guiding the selection of the next evaluation point [1]. It is widely used for data-efficient optimization of expensive black-box functions, particularly where evaluations are costly and data is scarce. In materials design and related fields, optimization problems are often accompanied by partial prior knowledge derived from physical models, simulations, or empirical relations. Incorporating such knowledge can in principle improve sample efficiency, but it is not clear when and where in the BO pipeline this knowledge should be injected, nor how sensitive each injection point is to knowledge of varying quality, a critical concern in the small-data regimes typical of experimental materials science, where few evaluations are affordable.

Recent works have proposed multiple strategies to incorporate domain knowledge into BO [2–9]. These include physics-informed surrogate models, expert-informed priors over the search space, and modifications of the acquisition function. While each of these approaches has demonstrated potential benefits, it remains unclear where knowledge should be injected in the BO pipeline to most effectively influence optimization dynamics on a wider range of problems. In this work, we provide a systematic empirical comparison of three conceptually distinct knowledge-injection strategies. Our goal is not to introduce a new optimization algorithm, but to investigate how the placement and flexibility of domain knowledge affect sample efficiency and robustness in small-data regimes.

1 Conceptual Framework

To structure the comparison, we express the different strategies in a unified conceptual form:

$$x_n = \arg\max_{x \in \mathcal{X}} \alpha[g](x)\, \pi(x) + R(x),$$

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

where $\alpha[g](x)$ denotes the acquisition function derived from a surrogate model g, $\pi(x)$ represents a prior over the search space, and $R(x)$ denotes an optional regularization term encoding auxiliary knowledge.

This formulation highlights that domain knowledge can influence Bayesian optimization by shaping the surrogate model (hypothesis space), biasing the search distribution via priors, or modifying the decision criterion through the acquisition function. This formulation serves as a conceptual abstraction to compare different integration strategies and does not define a new optimization method. We consider three placements of domain knowledge:

(i) Surrogate-level injection: Domain knowledge is embedded directly into the surrogate model, for example via physics-inspired mean functions, feature transformations, or domain-informed kernels [3].

(ii) Prior-based injection: Knowledge is encoded as a probabilistic prior over the search space, influencing candidate selection multiplicatively while allowing its impact to decrease over time[7].

(iii) Acquisition-level regularization: An auxiliary domain model contributes additively at the acquisition stage, guiding exploration without altering the surrogate [8, 9].

These strategies differ in how strongly they constrain the hypothesis space and how they interact with uncertainty estimation.

2 Experimental Overview

We evaluate the three strategies across different optimization tasks, including the Branin-Currin multi-objective problem [10], which provides a tractable yet non-trivial testbed for multi-objective BO under controlled conditions.

For the prior- and regularization-level injection strategies, a Gaussian process regression (GPR) is employed as the surrogate model. The tasks vary in dimensionality, objective structure, and noise level. All experiments are conducted under limited evaluation budgets to reflect small-data conditions. Performance is quantified using cumulative normalized hypervolume regret, computed on ground-truth objective values to avoid model-dependent bias. To isolate the effect of knowledge placement, we maintain consistent surrogate architectures and acquisition functions across experiments. Each configuration is repeated five times with different initial points and noise realizations.

All experiments are conducted under strict small-data conditions, with adapting evaluation budgets limited to a maximum of 100 function evaluations, reflecting realistic experimental constraints in materials science where each experiment may require significant time and cost.

3 Observations

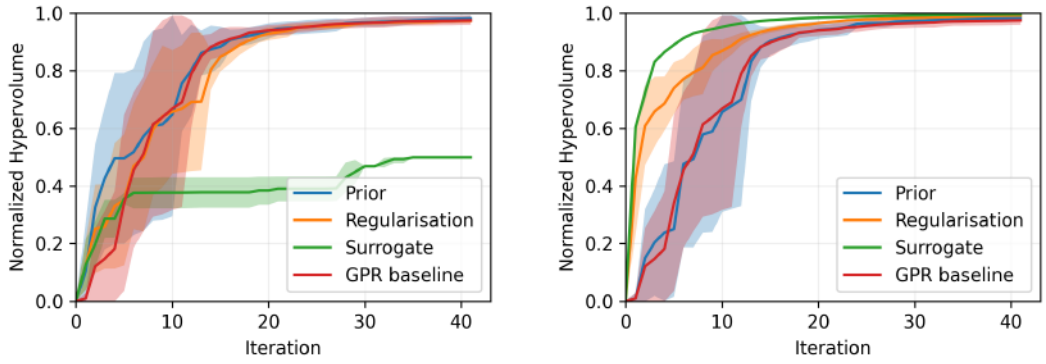
We consider both accurate and partially correct domain knowledge, where partial correctness reflects domain models that capture some but not all structural aspects of the true objective. Across tasks, we observe that the impact of domain knowledge depends strongly on both its placement and its flexibility.

Surrogate-level injection accelerates early-stage learning when the embedded knowledge is adaptable and reasonably aligned with the true objective structure. However, rigid or misspecified knowledge can severely degrade optimization performance. This behavior is illustrated in Fig. 1a.

Prior-based injection provides modest improvements in the early stages but does not consistently yield significant gains compared to vanilla Bayesian optimization. In the present formulation, the injected prior is soft and therefore mainly acts as a term that slightly biases surrogate predictions without fundamentally altering the acquisition landscape. As observational data accumulate, the GPR output becomes more certain over the domain, leading to sampling trajectories that are largely similar to those obtained with a non-informative prior.

Acquisition-level regularization offers flexible guidance. Although it is outperformed by surrogate-level injection when the incorporated knowledge closely matches the ground truth, it provides a more robust mechanism for integrating imperfect domain knowledge. By influencing the ranking of candidate points directly through the acquisition function, this strategy maintains beneficial

exploratory behavior even when the injected information is only partially correct. This effect is illustrated in Fig. 1b.



(a) BO with partially correct knowledge

(b) BO with accurate knowledge

Figure 1: Types of knowledge injected in the Branin-Currin multi-objective Bayesian optimization problem

Overall, these results highlight that the effectiveness of knowledge injection in Bayesian optimization depends both on how strongly the injected information constrains the surrogate model and on the level at which it influences the decision-making process. Strong surrogate-level guidance can substantially improve early sample efficiency when well aligned with the objective structure, but increases sensitivity to misspecification. In contrast, more flexible strategies, particularly those acting at the acquisition level, provide more robust performance by preserving the optimizer's ability to adapt to observed data.

This reveals a fundamental trade-off between performance and robustness, governed by the level at which domain knowledge is introduced.

4 Discussion and Outlook

These preliminary findings indicate that, in small-data regimes where only a handful of initial observations are available and the total evaluation budget is tightly constrained, the placement of domain knowledge within the BO pipeline critically determines optimization dynamics. The interaction between knowledge quality, surrogate flexibility, and uncertainty calibration becomes particularly consequential precisely because there is insufficient data to self-correct early misguidance. Rather than asking whether domain-informed BO is beneficial in general, our results suggest that the interaction between knowledge flexibility, uncertainty calibration, and noise level is decisive.

Future work will focus on deeper theoretical analysis of these interactions and on studying fewer real-world materials systems in greater detail to better understand the mechanistic origins of the observed behaviors.

Acknowledgments and Disclosure of Funding

The 'Think Before You Sample: Physics-Informed Active Learning for Green AI' action has received funding from the European Union, via the oc3-2025-TES-01 issued and implemented by the ENFIELD project, under the grant agreement No 101120657.

We acknowledge access to LEONARDO at CINECA, Italy, via an AURELEO (Austrian Users at LEONARDO supercomputer) project.

The authors gratefully acknowledge the financial support under the scope of the COMET program within the K2 Center "Integrated Computational Material, Process and Product Engineering (IC-MPPE)" (Project No 886385). This program is supported by the Austrian Federal Ministries for Economy, Energy and Tourism (BMWET) and for Innovation, Mobility and Infrastructure (BMIMI),

represented by the Austrian Research Promotion Agency (FFG), and the federal states of Styria, Upper Austria and Tyrol

References

[1] Eric Brochu, Vlad M. Cora, and Nando de Freitas. A tutorial on bayesian optimization of expensive cost functions. *arXiv preprint arXiv:1012.2599*, 2010.

[2] Sean Hooten, Wolfger Peelaers, Thomas Van Vaerenbergh, and Marco Fiorentino. Enhancing scientific bayesian optimization via physics-informed operator priors.

[3] Danial Khatamsaz, Raymond Neuberger, Arunabha Roy, Sina Zadeh, Richard Otis, and Raymundo Arróyave. A physics informed bayesian optimization approach for material design: application to niti shape memory alloys. *npj Computational Materials*, 9(1), 11 2023. ISSN ISSN 2057-3960. doi: 10.1038/s41524-023-01173-7.

[4] Wataru Kobayashi, Takuma Otsuka, Yuki K. Wakabayashi, and Gensai Tei. Physics-informed Bayesian optimization suitable for extrapolation of materials growth. *npj Computational Mathematics*, 11(1):36, February 2025. doi: 10.1038/s41524-025-01522-8.

[5] Maxim A Ziatdinov, Ayana Ghosh, and Sergei V Kalinin. Physics makes the difference: Bayesian optimization and active learning via augmented gaussian process. *Machine Learning: Science and Technology*, 3(1):015003, feb 2022. doi: 10.1088/2632-2153/ac4baa.

[6] Francesco Di Fiore and Laura Mainini. Physics-aware multifidelity bayesian optimization: A generalized formulation. *Computers & Structures*, 296:107302, 2024. ISSN 0045-7949. doi: https://doi.org/10.1016/j.compstruc.2024.107302.

[7] Carl Hvarfner, Danny Stoll, Artur Souza, Marius Lindauer, Frank Hutter, and Luigi Nardi. πbo: Augmenting acquisition functions with user beliefs for bayesian optimization, 04 2022.

[8] Zikai Xie, Xenophon Evangelopoulos, Joseph Thacker, and Andrew Cooper. *Domain Knowledge Injection in Bayesian Search for New Materials*. 09 2023. ISBN 9781643684369. doi: 10.3233/FAIA230587.

[9] Man Luo, Zikai Xie, Huirong Li, Baicheng Zhang, Jiaqi Cao, Yan Huang, Hang Qu, Qing Zhu, Linjiang Chen, Jun Jiang, and Yi Luo. Physics-informed, dual-objective optimization of high-entropy-alloy nanozymes by a robotic ai chemist. *Matter*, 8:102009, 03 2025. doi: 10.1016/j.matt.2025.102009.

[10] Samuel Daulton, Maximilian Balandat, and Eytan Bakshy. Differentiable expected hypervolume improvement for parallel multi-objective bayesian optimization. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Equayes – Democratizing Probabilistic Model Construction and Exploration with automatic Equation to Bayesian Model transformation

Christian Findenig
Embedded Computing and Machine Learning
Materials Center Leoben Forschung GmbH
8700 Leoben, Austria
christian.findenig@mcl.at

Manfred Mücke
Embedded Computing and Machine Learning
Materials Center Leoben Forschung GmbH
8700 Leoben, Austria
manfred.muecke@mcl.at

Abstract

For many scientific and engineering problems, equations based on applicable laws of physics can be used to link observable physical quantities. Analytic expressions, however, provide only point estimates and therefore cannot express uncertainty. This limits trustworthiness of predictions, especially in setups with limited data, noisy observations or when extrapolating. Bayesian probabilistic models address this limitation by treating unknown model parameters as random variables initialized by prior distributions and yielding – through inference – posterior (predictive) distributions. Constructing Bayesian models and convergence of inference, however, still requires specialized knowledge in probabilistic programming and inference algorithms, hindering the broader adoption of Bayesian models and uncertainty quantification in many domains. To make uncertainty-aware equation modeling more accessible, we present **Equayes** (*Equ*ation to B*ayes*ian Model), a scikit-learn-style estimator that converts a user-provided symbolic expression into a probabilistic model and performs posterior inference over its numerical constants. The core value of the method and tool to construction of hybrid models is that it implements a principled approach to hybrid model evaluation, linking laws of physics, random variables and inference in an accessible manner.

1 Introduction

Physics-informed machine learning and hybrid modeling combine mechanistic structure with statistical learning to improve generalization, sample efficiency, and interpretability. In many such workflows, a domain expert already has an analytical equation, a constitutive relation, or a symbolic surrogate, but still needs uncertainty estimates over parameters and predictions for model interpretation and model criticism. This need becomes especially acute when data is scarce or noisy, or when downstream decisions depend on confidence rather than on point estimates alone.

Bayesian modeling provides a principled way to address such scenarios by treating symbols as random variables (rather than point estimates) and combining prior assumptions with observed data to obtain posterior distributions [1]. Since the exact computation of posterior distributions is in general analytically intractable, practitioners resort to approximate or sampling-based inference. Markov chain Monte Carlo (MCMC) methods provide a general framework for drawing samples from the posterior distribution, thereby enabling both the quantification of parameter uncertainty and its propagation to downstream predictions. Recent advances in Hamiltonian Monte Carlo, particularly the No-U-Turn Sampler (NUTS), have substantially improved the robustness of posterior computation in practice [2]. Posterior predictive distributions then combine parameter uncertainty and observation

The Second Austrian Symposium on AI and Vision (AIROV25).

noise, which is particularly valuable for uncertainty-aware predictions, model criticism, and error propagation in scientific applications [1].

In practice, however, applying Bayesian methods to some analytic equation still requires substantial manual work. A user must implement the probabilistic program, define latent variables and priors, connect them correctly to the deterministic model, and configure inference machinery in a probabilistic programming framework. Although modern frameworks such as Pyro make Bayesian modeling flexible and expressive [3], they still assume familiarity with probabilistic programming abstractions. Equayes addresses this implementation bottleneck: if a model is already available as a symbolic mathematical expression, then most of the effort required to turn it into a Bayesian model is automated by Equayes.

2 From Equation to Bayesian Model

Let a user-provided symbolic model be written as

$$y = f_\theta(x_1, \ldots, x_N); \quad y, x_i \in \mathbb{R} \tag{1}$$

where $x_1, \ldots, x_N$ denotes the N input variables and $\theta \in \mathbb{R}^K$ denotes numerical constants in the equation. Let further $\mathcal{D} = \{\tilde{\mathbf{x}}_i, \tilde{y}_i\}_{i=1}^M$ be the observations of the system. The goal is to automatically define a probabilistic model $p(y \mid \theta, \mathbf{x})$ and infer the posterior distribution $p(\theta \mid \mathcal{D})$.

Equayes takes the symbolic expression f and

- identifies its numerical constants θ,
- replaces them by latent parameters, i.e. probability distributions $p(\theta)$,
- introduces an observation-noise variable and likelihood $p(y \mid \theta, \mathbf{x})$,
- compiles the resulting expression to a differentiable backend, and then
- runs MCMC inference to sample the posterior $p(y \mid \theta, \mathbf{x})$.

The resulting Bayesian model can be written as

$$\theta_k \sim \mathcal{N}(0, \sigma^2 = 1000), \quad k = 1, \ldots, K \tag{2}$$

$$\tilde{\sigma}^2 \sim \text{HalfNormal}(\sigma^2 = 10) \tag{3}$$

$$y \sim \mathcal{N}(f_\theta(x_1, \ldots, x_N), \tilde{\sigma}^2). \tag{4}$$

The wide variance in the prior over θ (2) encodes that parameters are in general unknown before observing any data. Furthermore, Equayes assumes that the observations are independent and identically distributed, therefore, they are modelled as samples of a Gaussian distribution (4).

In the current implementation, Equayes can be used by only specifying the symbolic expression f, all other parameters are readily set. Using the NUTS inference routine enables Equayes to infer complex posterior distributions over θ without making any assumptions about the geometry of the posterior. Internally, symbolic manipulation is handled through SymPy [4], while probabilistic execution and inference are handled through Pyro [3]. For all steps, Equayes is parameterizable, with default values as follows: `inference_method`: "mcmc", `mcmc_kernel`: "nuts", `mcmc_samples`: 2000, `mcmc_warmup_samples`: 2000, `mcmc_chains`: 1, `mcmc_initial_step_size`: $1e^{-2}$, the starting point of the MCMC chain is set to θ as given in the expression. In general, users of Equayes do not need to adjust those parameters, as they provide a good trade-off between computational cost and quality of the posterior distribution. However, advanced users may customize inference as required.

Executing inference, posterior predictive samples for new inputs, and retrieving posterior samples for θ are then exposed through a scikit-learn-style estimator interface via `fit()`, `predict()`, and `get_posterior()`, respectively. Equayes is available at https://github.com/mclprobability/Equayes.

This design is intentionally focused. Equayes is not intended to replace expert Bayesian modeling in settings that require bespoke priors, hierarchical structure, or custom likelihoods. Even though the architecture of Equayes is extendable to feature more customized models, like customizable prior distributions, which may be available in the future. Instead, it provides a practical and accessible route to uncertainty-aware equation fitting for users who want to preserve an analytical form while

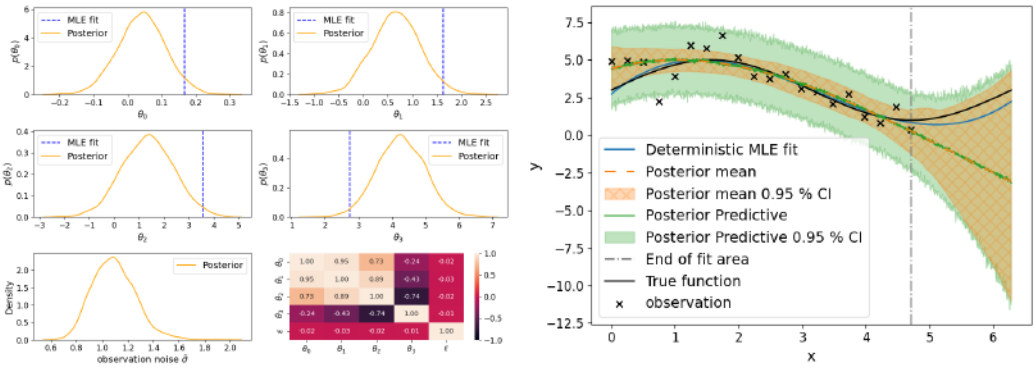


Figure 1: Comparison of deterministic and probabilistic polynomial fit (degree 3). Left: deterministic maximum likelihood estimates (MLE), and marginal posterior distributions and correlation matrix of the inferred parameters. Right: Deterministic fit and Equayes posterior predictive with 95 % credible intervals. The widening credible intervals indicate increasing uncertainty due to limited data support.

obtaining posterior and posterior predictive uncertainty with minimal implementation effort. We believe that this is important across a broad range of scientific and engineering domains, where symbolic relations are often known, inferred, or constrained by prior physical structure, yet the effort required to translate such equations into probabilistic models remains a substantial barrier to applying Bayesian inference in practice [5, 6].

3 Evaluation

First, we demonstrate Equayes on an illustrative example, second, we show the application of Equayes on a materials science use-case (cantilever).

3.1 Illustrative Example

To illustrate the practical gain, we consider noisy observations generated from a sinusoidal ground-truth process and fit a polynomial surrogate of degree 3, $f(x) = \theta_0 x^3 + \theta_1 x^2 + \theta_2 x + \theta_3$. This setup is intentionally misspecified: the polynomial surrogate model cannot represent the true sinusoid exactly, which mirrors most real-world problems in which the available analytical model is useful but not exact. We compare a deterministic maximum-likelihood fit against the posterior predictive result produced by Equayes.

Figure 1 shows that the deterministic solution provides a single reconstruction and therefore hides epistemic uncertainty caused by limited support in parts of the input domain. This uncertainty is visible in the posterior variance (left). Furthermore, the correlation matrix shows the dependence of the parameters under the posterior distribution. The Bayesian fit (right) yields a posterior predictive mean (variance caused by parameter uncertainty only), and full posterior predictive distribution together with 95 % credible intervals. In the example, the credible intervals widen substantially outside of the training region reflecting the uncertainty inherent to extrapolation.

3.2 Cantilever

Consider the Cantilever [7] as representative example of a physical system. Shown in Figure 2 (left) is a prototypical cantilever - a lever fixed on one side. We want to estimate the density ρ of the cantilever, without any possibility of measuring ρ directly. Hence, we need to resort to indirect measurements. The governing equations of the cantilever

$$A = W * H; \quad I = \frac{W * H^3}{12} \tag{5}$$

$$f = \frac{\beta^2}{2\pi L^2}\sqrt{\frac{EI}{\rho A}} \tag{6}$$

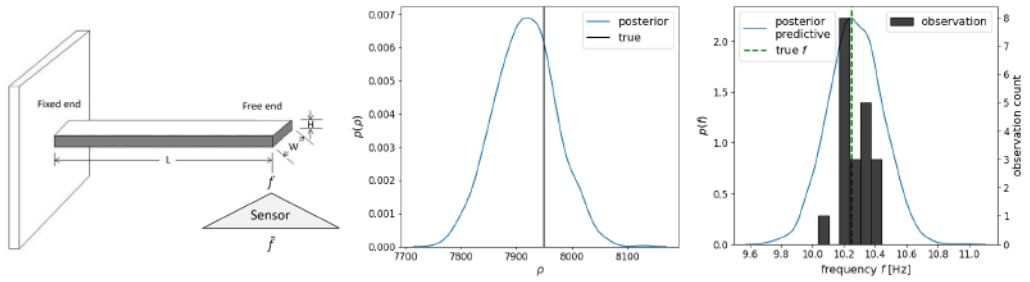


Figure 2: The Cantilever experiment. Left: visualization of the cantilever and measurement setup. Center: the inferred posterior distribution. Right: the posterior predictive distribution (left axis) and histogram of frequency observations (right axis).

with mode dependent β, density ρ, and Young's modulus E establish that the lever's natural frequency f directly depends on its density. We therefore use frequency measurements $\tilde{f}$ as system observations. To model sensor noise, we assume $\tilde{f}$ is distributed according to a Normal distribution, centered at the true frequency f with isotropic Gaussian sensor noise ϵ, $\tilde{f} \sim \mathcal{N}(f, \epsilon^2)$.

We use the governing equation (6) as a generative model to generate a data set of 20 frequency observations with parameters: $L = 0.9$ m, $W = 0.01$ m, $H = 0.01$ m, $\rho = 7950$ kg/m³, $E = 210e^9$ Pa, $\beta = 1.875$, $\epsilon = 0.1$ Hz. Given the generated data, we utilize Equayes to infer the density posterior distribution $p(\rho)$. We implement (6) in sympy and substitute ρ with the assumed density $\rho = 7900$. This tells Equayes to replace the parameter $\rho = 7900$ with a latent variable and expect all other parameters as input. On inference, we leverage the default parameters of Equayes and set all free parameters to their true values.

Figure 2 (center) shows the posterior distribution of density ρ. Its mode almost matches the true density, with widening intervals. This wider posterior distribution $p(\rho)$ expresses the remaining uncertainty given the 20, noisy frequency observations. The posterior predictive distribution (Figure 2 (right)) confirms that the posterior distribution truly describes the observed data, as the predictive distribution closely models the observed data.

4 Conclusion

We introduced Equayes, a scikit-learn-style estimator that turns symbolic equations into Bayesian models with minimal user effort. By automating parameter lifting, noise modeling, symbolic compilation, and MCMC-based inference, Equayes lowers the barrier to probabilistic modeling for equation-centric workflows. The result is not merely a point estimate, but posterior and posterior predictive information that quantify uncertainty and enable more robust downstream use. In this sense, Equayes fills an important gap in the modeling toolbox by making interpretable, prior-aware, and uncertainty-conscious Bayesian treatment of symbolic equations readily accessible in practice.

5 Acknowledgments

The authors gratefully acknowledge the financial support under the scope of the COMET program within the K2 Center "Integrated Computational Material, Process and Product Engineering (IC-MPPE)" (Project No 886385). This program is supported by the Austrian Federal Ministries for Economy, Energy and Tourism (BMWET) and for Innovation, Mobility and Infrastructure (BMIMI), represented by the Austrian Research Promotion Agency (FFG), and the federal states of Styria, Upper Austria and Tyrol.

ChatGPT 5.4 Thinking was used to reformulate or restructure paragraphs in this extended abstract.

References

[1] Paul Hewson. "Bayesian Data Analysis 3rd edn A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin, 2013 Boca Raton, Chapman and Hall–CRC 676 pp.,

£44.99 ISBN 1-439-84095-4". In: *Journal of The Royal Statistical Society Series A-statistics in Society* 178 (2015), pp. 301–301. URL: https://api.semanticscholar.org/CorpusID:120872375.

[2] Matthew D. Hoffman and Andrew Gelman. "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo". In: *Journal of Machine Learning Research* 15.47 (2014), pp. 1593–1623. URL: http://jmlr.org/papers/v15/hoffman14a.html.

[3] Eli Bingham et al. "Pyro: Deep Universal Probabilistic Programming". In: *CoRR* abs/1810.09538 (2018). arXiv: 1810.09538. URL: http://arxiv.org/abs/1810.09538.

[4] Aaron Meurer et al. "SymPy: symbolic computing in Python". In: *PeerJ Computer Science* 3 (Jan. 2017), e103. ISSN: 2376-5992. DOI: 10.7717/peerj-cs.103. URL: https://doi.org/10.7717/peerj-cs.103.

[5] Christopher Krapu and Mark Borsuk. "Probabilistic programming: A review for environmental modellers". In: *Environmental Modelling & Software* 114 (2019), pp. 40–48. ISSN: 1364-8152. DOI: https://doi.org/10.1016/j.envsoft.2019.01.014. URL: https://www.sciencedirect.com/science/article/pii/S1364815218308843.

[6] Sai Hung Cheung et al. "Bayesian uncertainty analysis with applications to turbulence modeling". In: *Reliability Engineering & System Safety* 96.9 (2011). Quantification of Margins and Uncertainties, pp. 1137–1149. ISSN: 0951-8320. DOI: https://doi.org/10.1016/j.ress.2010.09.013. URL: https://www.sciencedirect.com/science/article/pii/S0951832011000664.

[7] N.A. Rubayi and S. Charoenree. "Natural frequencies of vibration of cantilever sandwich beams". In: *Computers & Structures* 7.6 (1977), pp. 737–745. ISSN: 0045-7949. DOI: https://doi.org/10.1016/0045-7949(77)90028-1. URL: https://www.sciencedirect.com/science/article/pii/0045794977900281.

Robot Learning for Real-World Applications

ZeroShop: Automated Metric Mesh Generation for Zero-Shot 6D Object Pose Estimation

Stefan Lechner
Automation and Control Institute
TU Wien
1040 Vienna, Austria
e1608096@student.tuwien.ac.at

Philipp Ausserlechner
Automation and Control Institute
TU Wien
1040 Vienna, Austria
philipp.ausserlechner@tuwien.ac.at

Markus Vincze
Automation and Control Institute
TU Wien
1040 Vienna, Austria
markus.vincze@tuwien.ac.at

Abstract

Robotic manipulation of unseen objects relies on zero-shot 6D pose estimation, which typically requires a 3D mesh as a reference. While constructing accurate meshes requires specialized scanning hardware and manual editing, recently proposed Novel View Synthesis (NVS) techniques, such as 2D Gaussian Splatting (2DGS) and Sparse Voxels Rasterization (SVRaster), produce accurate surface reconstructions as a byproduct, potentially eliminating the need for specialized equipment. This work presents an automated image-based mesh generation pipeline that integrates object segmentation, camera registration, point cloud generation, metric height estimation, and NVS mesh generation, eliminating the need for expensive hardware and human intervention. Leveraging 2DGS and SVRaster with MASt3R-SfM or Visual Geometry Grounded Transformer (VGGT), the pipeline produces accurate meshes in minutes, with the VGGT/SVRaster combination reducing reconstruction time to seconds. Grounding near-view object-centric images with far-view scanning scene images using MASt3R yields consistent object height estimates. On the BOP YCB-V benchmark, meshes generated with our pipeline achieve competitive performance with state-of-the-art zero-shot pose estimation methods. Real-life robotic grasping experiments further indicate robust performance even under moderate scale errors. The source code is available at https://github.com/St333fan/meshgen-zeroshop.

1 Introduction

The rise of Machine Learning (ML) and Generative Artificial Intelligence (GenAI) has significantly enhanced the ability of robotic systems to navigate complex and dynamic environments, moving beyond the constraints of controlled settings [1]. However, planning and locomotion in robots are of limited utility without reliable object detection to facilitate environmental interaction. Although Vision-Language Models are increasingly embedding object representations within their parameters [2], state-of-the-art (SOTA) zero-shot object detection, segmentation, and 6D pose estimation methods still rely on reference objects, typically represented as 3D meshes for feature matching [3] [4] [5]. Figure 1 illustrates this with a robot detecting an object in a real-world scene, segmenting object-specific pixels, and estimating its pose to enable grasp planning and manipulation.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

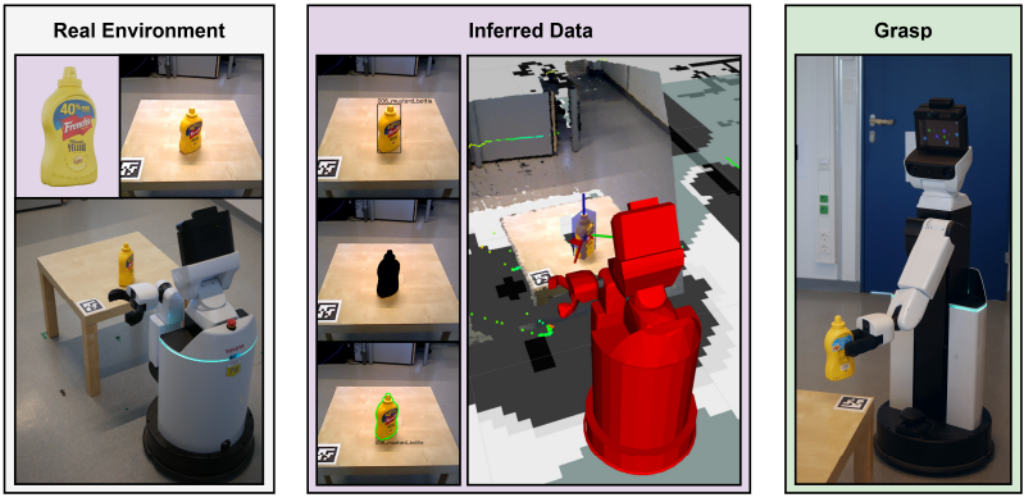


Figure 1: A robot extracts environmental information relevant to a target object for grasping. This data, including image location and corresponding pixel values, is then used to infer both a 2D pose (green) and a 6D pose (blue), defined within the robot coordinate system. Subsequently, this pose information facilitates grasp planning for object manipulation.

Conventionally, meshes are acquired using specialized scanning hardware, followed by manual refinement in a Computer-Aided Design program, or via Structure from Motion (SfM) algorithms [6], which yield comparatively lower-quality results [7]. Unfortunately, hardware scanners are expensive and SfM-based pipelines are highly feature-dependent, struggle with textureless objects, are scale-ambiguous, and require long processing times [6]. Recent ML-based approaches leveraging Vision Transformers for camera registration [8] [9] and scene reconstruction utilizing radiance fields [10] [11] have significantly closed the gap to hardware-based scanning. Combined into an automated pipeline, these methods eliminate the need for specialized scanning hardware and manual mesh generation.

In this work, we present an automated mesh generation pipeline that uses Grounded-SAM [12] for object segmentation, MASt3R-SfM [8] or VGGT [9] for camera pose estimation, and 2DGS [11] or SVRaster [10] for mesh reconstruction, while leveraging MASt3R [13] to estimate object height. All methods were tested on object data generated from the YCB-V subset of the Benchmark for 6D Object Pose Estimation (BOP) [4]. Logically, BOP was also used to compare all meshes within model-based tasks testing with SOTA open-source models CNOS [14], SAM-6D [15] and FoundationPose [16]. Finally, the best-performing mesh generation method was applied to real-world supermarket objects to generate accurate object meshes, which were subsequently validated through robotic grasping experiments.

This paper is organized as follows: Section 2 provides an overview of related work in the domains of 3D reconstruction, NVS, and pose estimation. Section 3 outlines the proposed automated mesh generation pipeline. Section 4 details the experimental setup and presents the results. Finally, Section 5 summarizes the paper and suggests directions for future research.

2 Related Works

The industry-standard reconstruction pipeline utilizes COLMAP [6] for camera registration and sparse point cloud estimation. The resulting point cloud is subsequently densified and meshed using Poisson Reconstruction [17]. While performing best with feature-rich or large objects, it suffers from lengthy processing times and relies on good initialization. Its successor GLOMAP [18] addresses processing time with comparable reconstruction quality, yet both rely on handcrafted features increasingly replaced by ML-based foundation models such as CroCo [19]. Notable integrations include DUSt3R [20], MASt3R [13], and MASt3R-SfM [8], whose learned features are more robust than handcrafted alternatives and increasingly used for feature matching [21] [8]. Importantly, MASt3R, trained on

metric data [13], is able to estimate scene scale; however, its performance is validated primarily for large-scale environments and has received limited testing in small-scale scenarios. Other specifically trained approaches include VGGT [9] and VGGSfM [22] for camera registration and point cloud generation. Currently, MASt3R-SfM and VGGT generate the most viable initial point clouds and camera positions, though both may remain inconsistent for direct mesh reconstruction. On the RealEstate10K benchmark [23], MASt3R-SfM and VGGT perform comparably, while VGGT with Bundle Adjustment (BA) achieves SOTA performance; however, reconstruction quality drops under extreme input rotations [9].

While these methods provide camera positions and an initial sparse point cloud, the resulting data is often incomplete or misaligned. NVS methods address this limitation by optimizing a scene representation from the input images, minimizing the rendering error across all views. The scene is encoded as radiance fields, which can subsequently be used to densify the point cloud [24] [25] [26]. This has been further addressed by 2DGS [11], which applies 2D Gaussian surfaces placed in a 3D space, ensuring alignment with the surfaces of a scene. This results in dense, on-surface-aligned points, where each point represents the midpoint of a Gaussian surface. Alternatively, SVRaster [10] adopts a different approach, initiating reconstruction from registered camera frames and directly generating view-consistent voxels, later fused into a mesh. Sun et al. [10] also presented comparative evaluations of NVS methods, demonstrating a favorable balance between accuracy and inference time for both SVRaster and 2DGS. Beyond rendering alignment, integrated loss functions also leverage object surface alignment guided by segmentation masks, typically extracted using methods such as Segment Anything Model (SAM) [27], Grounded-SAM [12], or Grounding DINO [28]. Following NVS reconstruction, post-processing into a mesh typically involves Poisson Reconstruction [17] or Marching Cubes [29], with texture either registered from source images or increasingly generated via diffusion models [30].

The reconstructed meshes are then used as reference models in model-based zero-shot object detection, segmentation, and 6D pose estimation. Among the available methods benchmarked in BOP [4], three stand out in terms of performance, open-source availability, and ease of integration: CNOS [14], SAM-6D [15], and FoundationPose [16]. CNOS combines DINOv2 [31] for zero-shot detection with SAM [27] or Fast-SAM [32] for segmentation in a straightforward pipeline. SAM-6D [15] extends CNOS by adding a geometric matching term for improved detection and segmentation, with subsequent pose estimation employing a two-stage point matching model. FoundationPose [16] offers a unified framework for 6D pose estimation and tracking of novel objects, supporting both model-based and model-free scenarios via LLM-aided synthetic training and a neural implicit representation. Controversially, the synthetic training data is subject to copyright issues, and a version trained without it has been released, potentially at the cost of reduced accuracy. For subsequent evaluation, the de facto benchmark is BOP [4], offering both 3D object models and real-world counterparts from its YCB-V subset. However, as open-source datasets [4] [33] [34] [35] are often included in model training data, novel data should also be considered.

3 Methodology

Building on the prerequisite of avoiding specialized scanning hardware, the developed meshing pipeline relies solely on a generic camera and ML methods. Figure 2 illustrates the four main stages of object generation: object segmentation, camera registration and point cloud generation, metric height estimation, and meshing supported by NVS. Each step is described in detail in the following sections.

3.1 Data Acquisition and Object Segmentation

To enable reliable 3D reconstruction and accurate segmentation, it is necessary to acquire images that provide complete and uniform coverage of the object surface. A camera should move along a trajectory encircling the object at multiple elevations, consistently maintaining focus on its center. The first image captures the frontal view to establish the coordinate axes. In addition, capturing data as a video stream accelerates acquisition, promotes consistent illumination, and allows flexible frame extraction. Note that capturing the object solely from a top-down perspective leaves a hole at the bottom of the resulting mesh. Once captured, the object is segmented from the background using Grounded-SAM [12], a zero-shot promptable segmentation model. Using the prompt "object in the

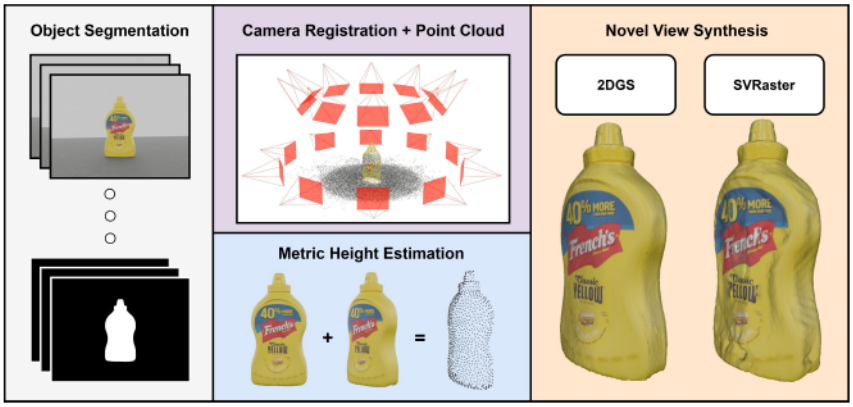


Figure 2: Object generation proceeds in four stages: Object Segmentation generates image masks from a video, followed by Camera Registration and Point Cloud generation, which estimates camera poses and a sparse point cloud. Metric Height Estimation then isolates object-specific 3D points from masked front-view images to estimate the object dimensions. Finally, the point cloud is densified and meshed using the NVS methods 2DGS and SVRaster.

middle." on the first frame, it tracks and generates the object mask across all frames, as seen in Figure 2, enabling accurate segmentation for diverse objects.

3.2 Camera Registration and Point Cloud Generation

Starting from an unordered collection of scene images without known camera intrinsics and extrinsics, it is possible to estimate these parameters while reconstructing 3D scene geometry, as seen in Figure 2. To solve this, MASt3R-SfM integrates an ML stereo model into a conventional SfM framework by using MASt3R features for image retrieval and pairwise matching, followed by camera pose estimation and point cloud reconstruction within a COLMAP-based optimization stage. This hybrid design reduces matching complexity from quadratic to linear, enables robust registration even under purely rotational motion, and avoids reliance on RANSAC, while still benefiting from BA for global consistency. In contrast, VGGT adopts a fully feed-forward approach that jointly processes multiple images to directly predict camera poses, depth maps, point tracks, and 3D point maps in a single forward pass. This yields significantly faster inference and eliminates explicit correspondence search and incremental reconstruction, but requires higher GPU memory and exhibits reduced robustness under large viewpoint changes. While MASt3R-SfM trades runtime efficiency for scalability and accuracy through optimization, VGGT prioritizes real-time performance by learning geometric reasoning end to end from large-scale data.

3.3 Metric Height Estimation

While VGGT does not estimate metric scale, MASt3R-SfM and MASt3R fail to recover the correct scale from near-view object-centric images alone. The solution integrates both near-view object and far-view scene images into a single reconstruction using MASt3R, chosen for its computational efficiency. Since scale is an optimization parameter across image pairs [13], scene images must outnumber object images. A good compromise uses four scene and two object images; the latter being the minimum required by the stereo backbone of MASt3R, producing a point cloud of the full scene with the object registered within it, as shown in Figure 3. Object height is estimated by projecting the point cloud into the reference camera coordinate frame, applying the object mask, and subtracting the lowest from the highest point coordinate. Crucially, scene images can be captured without the object present.

3.4 NVS Mesh Generation

In the last step of Figure 2, a mesh is generated with 2DGS and SVRaster, followed by post-processing with texture, given a pose based on the first reference frame, and scaled to the estimated dimensions. 2DGS and SVRaster utilize the registered cameras, the point cloud, and object masks to iteratively

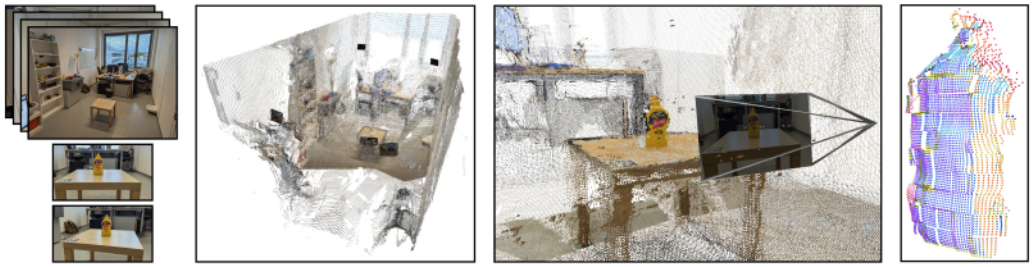


Figure 3: To generate an accurate metric scene point cloud, four scene images (excluding the object) are registered with two images of the object using MASt3R. Subsequently, the first registered image and its corresponding object mask are utilized to extract object-specific points, which are then employed to estimate the height of the scanned object.

reduce loss functions [11] [10], representing the object as densified 2D Gaussian disks or voxels. The radiance fields are then converted into a dense object-specific point cloud and meshed with Poisson Reconstruction (2DGS) or Marching Cubes (SVRaster).

4 Experiments and Results

To assess the quality and robustness of the proposed pipeline, four sub-goals were examined. First, camera registration and point cloud generation with VGGT/MASt3R-SfM and 2DGS/SVRaster to determine which configuration reconstructs the greatest number of YCB-V objects. Second, the MASt3R scaling method was applied to real-world objects and the estimated scale was compared to the true object height. Third, the standardized BOP benchmark was used to validate the object meshes in 2D segmentation and 6D pose estimation tasks. Finally, robotic manipulation was examined in real-world settings using CNOS and FoundationPose.

4.1 Mesh Generation Reconstruction Rate and Quality

To evaluate the automated mesh generation pipeline, the optimal combination of MASt3R-SfM, VGGT, 2DGS, and SVRaster is identified based on reconstruction success. The best configuration is subsequently evaluated on real YCB-V and supermarket objects to assess reconstruction differences using real-world data.

4.1.1 Virtual YCB-V Objects

To establish a baseline for the mesh reconstruction success rate, all 21 virtual YCB-V ground-truth (GT) object models were rendered in BlenderProc [36] to generate scene images. Additionally, scenes were differentiated by high- and low-feature surfaces, with masked object images also included to assess the necessity of surface information for camera registration and NVS. The first two cases comprised 20 images each, while the masked case, unconstrained by surface limitations, yielded a denser representation of 30 images, including views from below. The render scenes are depicted in Appendix A.

The reconstruction success achieved per combination is detailed in Table 1, where both registration methods exhibit distinct strengths. In particular, MASt3R-SfM with high-feature surfaces is the only combination achieving successful reconstructions for all objects. Although VGGT surpassed MASt3R-SfM in the segmented task, it failed to reconstruct the featureless YCB-V bowl object. Figure 4 visualizes qualitative differences in mesh quality between 2DGS (left) and SVRaster (right) across three geometrically distinct objects, using MASt3R-SfM initialization. In summary, SVRaster generates meshes with object dimensions comparable to 2DGS but with a less smooth surface, while training in seconds compared to minutes for 2DGS.

4.1.2 Real Supermarket Objects

Given that only MASt3R-SfM successfully reconstructed all virtual YCB-V object scenes, it was selected for evaluation on real-world objects. As detailed in Section 3.1, a video was recorded,

Table 1: 2DGS and SVRaster mesh reconstruction success initiated by VGGT and MASt3R-SfM, utilizing the rendered data from the 21 BOP YCB-V GT objects. The surface task is divided into a low- and high-feature surface, and seg indicates registration utilizing object masks.

	MASt3R-SfM			VGGT		
	low	high	seg	low	high	seg
2DGS	20	21	16	20	18	20
SVRaster	20	21	15	19	18	20

Figure 4: Overview of the reconstruction quality of three geometrically distinct YCB-V objects. For each pair, the left object is from 2DGS and the right from SVRaster.

20 equally spaced frames were extracted, and processed through the pipeline. All objects were successfully reconstructed as meshes; examples are depicted in Figure 5 with 2DGS (left) and SVRaster (right). Upon application of the texture, no apparent differences are visible. All generated meshes and an example of surface quality are provided in Appendix B.

Figure 5: Overview of the reconstruction quality of objects that are used in the grasp evaluation. For each pair, the left mesh is from 2DGS and the right from SVRaster.

4.2 Object Scaling

The height of real YCB-V and supermarket objects were estimated following the procedure from Section 3.3; the results are shown in Figure 6 with the real height plotted against the percentage error. The results indicate an estimation error within $\pm 10\%$ for most objects, with a tendency for estimation error to decrease with increasing object height. A notable exception is the "toothbrush", which exhibits a disproportionately large error. The complete results are provided in Appendix C.

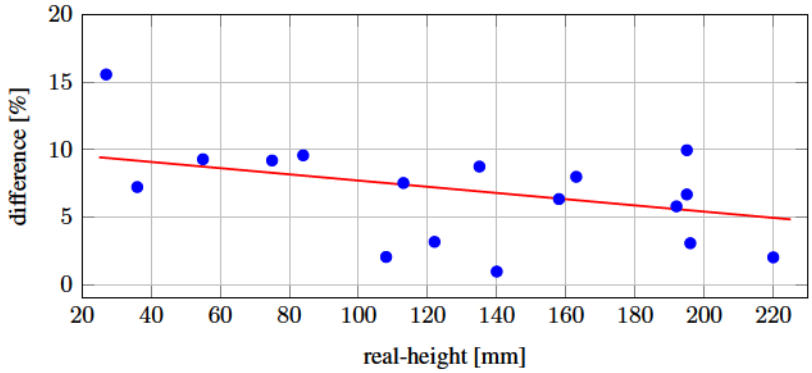


Figure 6: The relationship between the real object height and the absolute measurement error in percent for real YCB-V and supermarket objects. Individual measurements are represented as blue points, with a red line indicating the trend of decreasing error as object height increases.

4.3 BOP with NVS Meshes

To evaluate the suitability of the reconstructed meshes for object segmentation and pose estimation, a benchmark validation was conducted using the BOP framework. Official YCB-V performance scores from previous work [4] serve as a reference. As a first step, the Average Precision (AP) and Average Recall (AR) [37] scores reported in the original papers using virtual GT objects from the YCB-V dataset were reproduced. Subsequently, the reconstructed objects from Section 4.1.1 were used to benchmark each method and assess whether comparable accuracy could be maintained. These reconstructed objects were generated with BlenderProc scenes, they possess an arbitrary scale and were therefore aligned to their corresponding GT counterparts using the ICP algorithm to ensure the correct height. Importantly, the CNOS mask filtering used in the original FoundationPose implementation was not publicly available and had to be re-implemented. As a result, GT masks were also used to evaluate FoundationPose. To analyze how the generated meshes affect the AP score for segmentation and the AR score for pose estimation in relation to the official BOP meshes, four distinct scenarios were defined:

- CNOS on GT/NVS meshes, with FastSAM
- Instance Segmentation Model (ISM) SAM-6D on GT/NVS meshes, with SAM
- Pose Estimation Model (PEM) SAM-6D with ISM masks on GT/NVS meshes
- FoundationPose with GT/CNOS masks on GT/NVS meshes

AP/AR scores are presented in Table 2, divided into segmentation and pose estimation tasks, with the submission row reporting the official scores of the BOP benchmark. Focusing on the segmentation task, the AP scores of both methods, CNOS and SAM-6D ISM, were successfully reproduced. When employing NVS-generated meshes, all methods exhibit a comparable decrease in performance, with only a marginal deviation relative to the scores obtained using SVRaster meshes. When using FoundationPose on GT masks, the GT BOP meshes achieve performance that exceeds the official AR score. Furthermore, the 2DGS and SVRaster meshes exhibit only a minor decrease in performance. When employing CNOS-based segmentation, the accuracy decreased across all evaluated scenarios. In addition, the official reported scores could not be fully reproduced, and the performance gap between the GT BOP meshes and the 2DGS/SVRaster meshes increased further. In comparison, the official SAM-6D PEM scores were almost reproducible and exhibited a smaller performance gap

compared to 2DGS and SVRaster than FoundationPose/CNOS. However, the difference between 2DGS and SVRaster is slightly greater than that observed in the FoundationPose evaluation.

Table 2: Official YCB-V BOP AP (segmentation) and AR (pose estimation) scores compared to reproduced scores across different mesh datasets. FoundationPose was additionally benchmarked with GT masks because the official filtering of the CNOS masks is not publicly available.

	Segmentation		Pose Estimation		
	CNOS	ISM (SAM-6D)	FoundationPose		PEM (SAM-6D)
Submission	0.599	0.605	0.882		0.845
			GT	CNOS	
BOP	0.6	0.603	0.915	0.731	0.832
2DGS	0.56	0.561	0.889	0.543	0.751
SVRaster	0.541	0.567	0.877	0.539	0.71

4.4 Robotic Object Manipulation

After evaluating mesh quality, the influence of the height estimation of Section 3.3 on segmentation and pose estimation remains to be examined. Therefore, we evaluated real reconstructed meshes within a robotic grasp pipeline. Based on preliminary experiments, FoundationPose with CNOS/SAM was adopted for pose estimation and evaluated with a grasp success test using the meshes depicted in Figure 5, each manually annotated with grasp positions. The test scenario is defined as follows: each object is positioned at the same location in front of the robot, within its grasping range, and then rotated four times by 90° around the z-axis. Subsequently, the object is flipped onto its top and rotated four times again, resulting in a total of eight distinct poses per object. An attempt to grasp is classified as successful if the robot establishes a stable grasp and lifts the object. In instances where the pipeline fails at any stage, it is re-executed; this repeated execution is defined as a re-grasp.

Considering only a single grasp attempt, the pipeline using 2DGS meshes achieved a grasp success rate of 85.9%, while the pipeline using SVRaster achieved 89.1%. When re-grasps were included, the success rate for 2DGS increased to 93.8%, while that for SVRaster increased to 90.6%. The two primary outlier cases were the inverted "toothbrush", which was not detectable using CNOS, and the upright "soap", for which FoundationPose consistently estimated the pose as lying horizontally. In addition, the following observations were made: CNOS exhibited difficulty in segmenting the "gelatin_box"; the robot lost its otherwise stable grasp on the "mustard_bottle" twice during lifting; and the "razors," which were meshed in a compressed configuration with an underestimated object height, appear to be affected by the cumulative error in one specific pose.

5 Conclusion

The proposed automated mesh generation pipeline, which integrates Grounded-SAM, MASt3R-SfM/VGGT, and 2DGS/SVRaster, consistently produces geometrically accurate meshes that allow accurate zero-shot object pose estimation. Grasping experiments conducted with a robot and CNOS/FoundationPose demonstrated that the NVS-generated meshes and their associated scaling are adequate for real-world object manipulation. However, this approach has limitations when applied to deformable, low-profile, or semi-transparent objects. These shortcomings are primarily attributable to scaling inaccuracies, incomplete mesh reconstruction, or unreliable robotic sensory data.

Future work will focus on improving the reconstruction of occluded surfaces and extending the pipeline to transparent objects such as glass and plastic. Another promising research direction involves rendering NVS in conjunction with depth information predicted by ML depth models and using GenAI models for missing parts. Finally, testing the generated meshes in a supermarket, where hundreds of objects appear across long, interconnected tasks, would further validate the robustness of our approach.

Acknowledgments and Disclosure of Funding

This work was supported by the EU-program EC Horizon 2020 for Research and Innovation under grant agreement No. 101120823 project MANiBOT funded by the European Union.

References

[1] Ranjan Sapkota, Yang Cao, Konstantinos I Roumeliotis, and Manoj Karkee. Vision-language-action models: Concepts, progress, applications and challenges. *arXiv preprint arXiv:2505.04769*, 2025.

[2] Yiming Zuo, Karhan Kayan, Maggie Wang, Kevin Jeon, Jia Deng, and Thomas L Griffiths. Towards foundation models for 3d vision: How close are we? *arXiv preprint arXiv:2410.10799*, 2024.

[3] Felix Gorschlüter, Pavel Rojtberg, and Thomas Pöllabauer. A survey of 6d object detection based on 3d models for industrial applications. *Journal of imaging*, 8(3):53, 2022.

[4] Van Nguyen Nguyen, Stephen Tyree, Andrew Guo, Mederic Fourmy, Anas Gouda, Taeyeop Lee, Sungphill Moon, Hyeontae Son, Lukas Ranftl, Jonathan Tremblay, et al. Bop challenge 2024 on model-based and model-free 6d object pose estimation. *arXiv preprint arXiv:2504.02812*, 2025.

[5] Shibiao Xu, Shunpeng Chen, Rongtao Xu, Changwei Wang, Peng Lu, and Li Guo. Local feature matching using deep learning: A survey. *Information Fusion*, 107:102344, 2024.

[6] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016.

[7] MW Wilkinson, Robin R Jones, Christopher E Woods, SR Gilment, Ken JW McCaffrey, Sotiris Kokkalas, and JJ Long. A comparison of terrestrial laser scanning and structure-from-motion photogrammetry as methods for digital outcrop acquisition. *Geosphere*, 12(6):1865–1880, 2016.

[8] Bardienus Pieter Duisterhof, Lojze Zust, Philippe Weinzaepfel, Vincent Leroy, Yohann Cabon, and Jerome Revaud. Mast3r-sfm: a fully-integrated solution for unconstrained structure-from-motion. In *2025 International Conference on 3D Vision (3DV)*, pages 1–10. IEEE, 2025.

[9] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 5294–5306, 2025.

[10] Cheng Sun, Jaesung Choe, Charles Loop, Wei-Chiu Ma, and Yu-Chiang Frank Wang. Sparse voxels rasterization: Real-time high-fidelity radiance field rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16187–16196, 2025.

[11] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 conference papers*, pages 1–11, 2024.

[12] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024.

[13] Vincent Leroy, Yohann Cabon, and Jérôme Revaud. Grounding image matching in 3d with mast3r. In *European conference on computer vision*, pages 71–91. Springer, 2024.

[14] Van Nguyen Nguyen, Thibault Groueix, Georgy Ponimatkin, Vincent Lepetit, and Tomas Hodan. Cnos: A strong baseline for cad-based novel object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2134–2140, 2023.

[15] Jiehong Lin, Lihua Liu, Dekun Lu, and Kui Jia. Sam-6d: Segment anything model meets zero-shot 6d object pose estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27906–27916, 2024.

[16] Bowen Wen, Wei Yang, Jan Kautz, and Stan Birchfield. Foundationpose: Unified 6d pose estimation and tracking of novel objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17868–17879, 2024.

[17] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 2006.

[18] Linfei Pan, Dániel Baráth, Marc Pollefeys, and Johannes L Schönberger. Global structure-from-motion revisited. In *European Conference on Computer Vision*, pages 58–77. Springer, 2024.

[19] Philippe Weinzaepfel, Thomas Lucas, Vincent Leroy, Yohann Cabon, Vaibhav Arora, Romain Brégier, Gabriela Csurka, Leonid Antsfeld, Boris Chidlovskii, and Jérôme Revaud. Croco v2: Improved cross-view completion pre-training for stereo matching and optical flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17969–17980, 2023.

[20] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20697–20709, 2024.

[21] Riku Murai, Eric Dexheimer, and Andrew J Davison. Mast3r-slam: Real-time dense slam with 3d reconstruction priors. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 16695–16705, 2025.

[22] Jianyuan Wang, Nikita Karaev, Christian Rupprecht, and David Novotny. Vggsfm: Visual geometry grounded deep structure from motion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21686–21697, 2024.

[23] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.

[24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021.

[25] Qiangeng Xu, Zexiang Xu, Julien Philip, Sai Bi, Zhixin Shu, Kalyan Sunkavalli, and Ulrich Neumann. Point-nerf: Point-based neural radiance fields. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5438–5448, 2022.

[26] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

[27] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[28] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024.

[29] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. ACM, 1998.

[30] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025.

[31] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[32] Xu Zhao, Wenchao Ding, Yongqi An, Yinglong Du, Tao Yu, Min Li, Ming Tang, and Jinqiao Wang. Fast segment anything. *arXiv preprint arXiv:2306.12156*, 2023.

[33] Henrik Aanæs, Rasmus Ramsbøl Jensen, George Vogiatzis, Engin Tola, and Anders Bjorholm Dahl. Large-scale data for multiple-view stereopsis. *International Journal of Computer Vision*, 120:153–168, 2016.

[34] Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10901–10911, 2021.

[35] Tong Wu, Jiarui Zhang, Xiao Fu, Yuxin Wang, Jiawei Ren, Liang Pan, Wayne Wu, Lei Yang, Jiaqi Wang, Chen Qian, et al. Omniobject3d: Large-vocabulary 3d object dataset for realistic perception, reconstruction and generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 803–814, 2023.

[36] Maximilian Denninger, Martin Sundermeyer, Dominik Winkelbauer, Youssef Zidan, Dmitry Olefir, Mohamad Elbadrawy, Ahsan Lodhi, and Harinandan Katam. Blenderproc. *arXiv preprint arXiv:1911.01911*, 2019.

[37] Tomáš Hodaň, Martin Sundermeyer, Bertram Drost, Yann Labbé, Eric Brachmann, Frank Michel, Carsten Rother, and Jiří Matas. Bop challenge 2020 on 6d object localization. In *European Conference on Computer Vision*, pages 577–594. Springer, 2020.

A Rendered BlenderProc Scenes

Figure 7: YCB-V object rendering with high/low feature surface and extracted segmented RGB masks; rendering angles of 0°, 45°, and −45° to the horizontal plane.

B Real Reconstructed Objects

Figure 8: Reconstructed YCB-V objects, with the left/upper object illustrating reconstruction using 2DGS, followed by SVRaster.

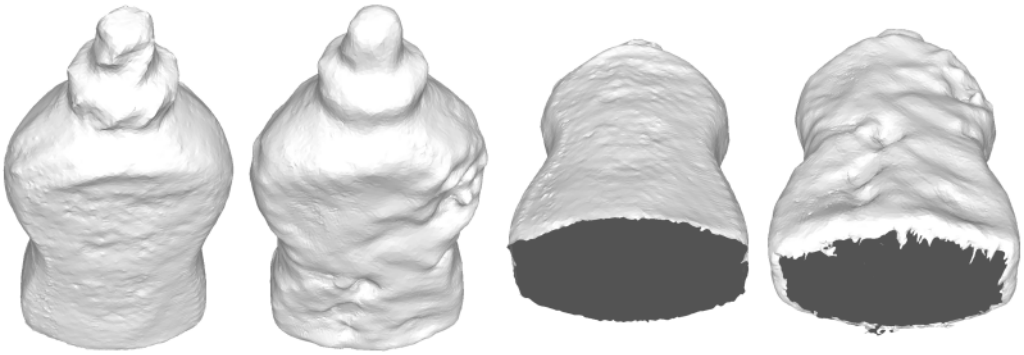


Figure 9: Mesh quality comparison between 2DGS and SVRaster on real-world data, showing top and bottom views, with 2DGS yielding more accurate surfaces.

Figure 10: Reconstructed supermarket objects, with the left/upper object illustrating reconstruction using 2DGS, followed by SVRaster. Upon application of the texture, no apparent differences become visible.

C Metric Height Estimation

Table 3: Comparison between the actual heights of YCB-V objects and those estimated using the MASt3R registration method.

	height [m]	mast3r-height [m]	difference [mm]	difference [%]
mustard_bottle	0.192	0.2031	11.12	5.79
potted_meat_can	0.084	0.092	8.04	9.57
bowl	0.055	0.0601	5.1	9.28
cracker_box	0.22	0.2244	4.42	2.01
master_chef_can	0.14	0.1413	1.34	0.96
gelatin_box	0.075	0.0819	6.89	9.19
large_marker	0.122	0.1259	3.85	3.16
extra_large_clamp	0.036	0.0334	-2.6	-7.22

Table 4: Comparison between the actual heights of the supermarket objects and those estimated using the MASt3R registration method.

	height [m]	mast3r-height [m]	difference [mm]	difference [%]
soap	0.158	0.148	-10.0	-6.33
ahorn_sirup	0.163	0.15	-13.0	-7.98
tomato_paste	0.195	0.208	13.0	6.67
kokos_can	0.113	0.1215	8.5	7.52
hand_cream	0.135	0.1468	11.8	8.74
wet_wipes	0.108	0.1058	-2.2	-2.04
razors	0.195	0.1756	-19.4	-9.95
balsamic	0.196	0.202	6.0	3.06
toothbrush	0.027	0.0312	4.2	15.56

1D Profiles vs. Spectral Images: A Comparative Study of Machine Learning Models for Mineral and Rock Classification

Sai Puneeth Reddy Gottam[1]
sai.gottam@unileoben.ac.at

Martin Johannes Findl[2]
martin.findl@unileoben.ac.at

Robert Galler[3]
robert.galler@unileoben.ac.at

Klaus Philipp Sedlazeck[2]
philipp.sedlazeck@unileoben.ac.at

Elmar Rueckert[1]
elmar.rueckert@unileoben.ac.at

Abstract

The rapid identification of minerals is critical for real-time geological analysis. This study investigates the efficacy of machine learning models in classifying mineral and rock samples using high-speed Raman sensors. We evaluate three distinct data representation strategies: **(1) 1D spectral profiles**, **(2) 2D Raman spectral images**, and **(3) a fused multi-modal approach** combining both spatial and spectral features. Using a diverse dataset of geological samples, we benchmark several model architectures to determine the trade-offs between computational efficiency and classification accuracy. Our results demonstrate how spatial context from imaging can enhance identification compared to traditional 1D methods, while also identifying the scenarios where signal-only processing remains optimal. This work provides a framework for selecting the most effective data representation for high-speed, automated mineralogical mapping.

1 INTRODUCTION

Rapid and accurate identification of mineral phases and rock compositions is a cornerstone of modern geological exploration, planetary science, and industrial mining. Traditional mineralogical analysis often relies on manual petrography or X-ray diffraction (XRD) [1], which, while accurate, are time-consuming and difficult to implement in high-throughput or autonomous environments. **Raman spectroscopy** has emerged as a powerful alternative due to its non-destructive nature and its ability to provide a unique "fingerprint" of a material's molecular structure [1, 2].

Recent advancements in sensor technology have shifted the paradigm from single-point Raman probes to **high-speed spectral imaging**. These modern sensors allow the collection of large amounts of data in short timeframes, capturing both the chemical signature (1D spectral profiles) and the spatial distribution (2D spectral images) of mineral grains. However, the influx of high-cadence data presents a significant computational challenge[5]: how to effectively process and classify these signals in real-time.

[1]Chair of Cyber Physical Systems, Montanuniversität Leoben, Austria
[2]Chair of Waste Processing Technology and Waste Management, Montanuniversität Leoben, Austria
[3]Chair of Subsurface Engineering, Montanuniversität Leoben, Austria

The Third Austrian Symposium on AI and Vision (AIROV26).

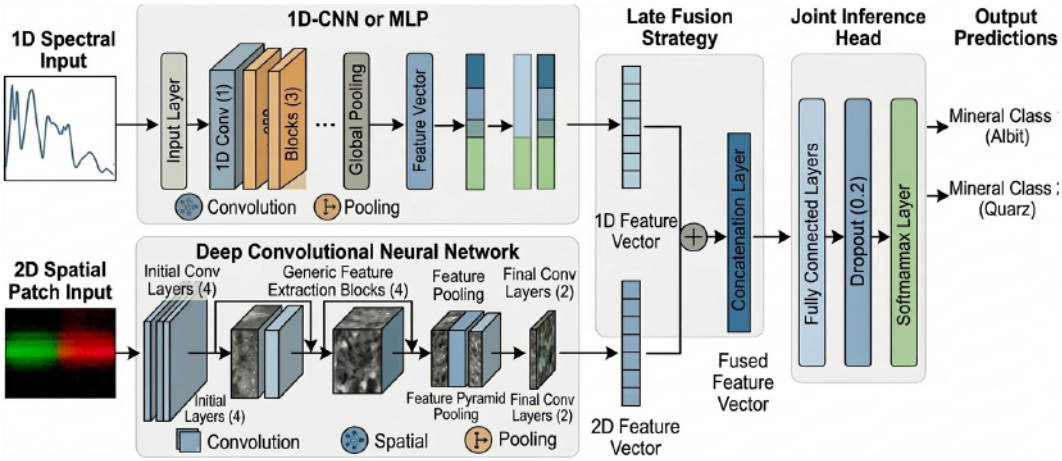


Figure 1: Proposed multi-modal late-fusion architecture. The model processes 1D spectral inputs and 2D spatial patches through independent feature extraction branches. These high-dimensional vectors are concatenated and processed by a joint inference head with dropout regularization to provide final mineral classifications.

Machine learning (ML) and deep learning (DL) have shown immense promise in automating spectral analysis [1, 4]. Although 1D Convolutional Neural Networks (1D-CNNs) are efficient in processing individual spectra, they often ignore the textural and morphological context provided by the surrounding mineral matrix. Conversely, 2D imaging models capture spatial relationships but may introduce unnecessary computational overhead if the spectral signal alone is sufficiently discriminative [6].

This paper investigates the trade-offs between different data representations for mineral classification. We conduct a comprehensive experimental study using high-speed Raman data to evaluate three distinct approaches:

- **1D Profile Analysis:** Classifying minerals based purely on individual spectral signatures.
- **Image-Based Classification:** Utilizing 2D spatial-spectral maps to capture mineral texture and grain boundaries.
- **Fused Multi-Modal Learning:** Investigating whether the integration of both 1D and 2D features yields superior robustness.

By benchmarking these methods, we aim to identify the optimal balance between classification accuracy and processing speed, providing a roadmap for the deployment of autonomous mineralogical mapping systems.

2 Method

2.1 Dataset and Experimental setup

The dataset consists of Raman measurements collected under controlled laboratory conditions. It includes multiple mineral classes as well as heterogeneous rock samples composed of mixed mineral phases.The mineral dataset comprises 12 classes with 3000 images and profiles of each mineral class. The rock dataset has 12 classes with 3000 images and profiles of each class. Some of these rock classes are the variations of different version of same rock type such as varities of Granite and Sandstone.

The dataset was split into training, validation, and test sets using a standard 70/15/15 ratio to ensure unbiased evaluation. All models were evaluated on a held-out test set.

2.2 Data Preparation and Normalization

All Raman spectra were normalized using min-max scaling to ensure consistency across samples and mitigate variations due to sensor conditions such as laser power and acquisition distance.

- **1D Profiles:** Represented as intensity vectors across the wavenumber domain
- **2D Spectral Images:** Constructed from spatial Raman scans to capture texture and morphology

2.3 Model Architectures

We evaluated five architectures:

- **Random Forest (RFC):** Baseline model using 100 trees for 1D classification
- **MLP:** Fully connected network with ReLU activations and dropout
- **1D-CNN:** Convolutional model capturing local spectral features
- **ResNet-18 (2D) [3]:** Image-based classifier leveraging spatial context
- **Multi-Modal Fusion:** Late fusion architecture combining 1D and 2D features as shown in 1. The fusion model concatenates latent representations from both branches before final classification.

3 Preliminary Results and Discussion

The benchmarking phase evaluated five distinct architectures across two primary geological scales: pure mineral phases and complex rock aggregates. The results reveal a significant disparity in how different models handle the high-cadence 30fps data stream. All models are trained on mineral dataset but only 2 models are trained on Rock dataset.

Table 1: Classification Performance across Model Architectures and Mineral Dataset

Model Architecture	Data Input	Mineral Acc (%)	Rock Acc (%)	Complexity
Random Forest (Baseline)	1D Profile	**99.96%**	-	41,450 nodes
MLP	1D Profile	99.10%	-	Medium
1D-CNN	1D Profile	80.92%	-	Low
ResNet-18	2D Image	>99.9%	99.77%	11.2M Params
Multi-Modal Fusion	Hybrid (1D+2D)	99.85%	**99.98%**	High

3.1 Performance on Pure Mineral Dataset

For the identification of individual mineral phases, the high-resolution spectral fingerprints proved to be highly discriminative.

- **Classical Baseline Excellence:** Surprisingly, the **Random Forest Classifier (RFC)** emerged as the top-performing model for 1D mineral profiles, achieving an accuracy of **99.96%** as shown in table 1. This suggests that for normalized laboratory data, the ensemble of decision trees effectively captures the essential vibrational peaks without the computational overhead of deep learning.
- **Deep Learning Performance:** The **MLP** and **ResNet (2D Image)** models also performed exceptionally well, both exceeding **99.9%** accuracy. This confirms that the spectral "images" extracted from the sensor provide sufficient morphological detail to distinguish pure minerals with near-perfect reliability.
- **The 1D-CNN Outlier:** Interestingly, the **1D-CNN** was the only architecture that struggled in this phase, achieving only **80.92%** accuracy. This indicates that a simple convolutional approach may be overly sensitive to spectral shifts or lacks the global feature integration required for this specific dataset. Also there are very few parameters in this model which fail to capture the characteristics needed for classifying accurately.

The high accuracy values are attributed to the controlled laboratory environment and high-quality spectral signals.

3.2 Performance on Rock Dataset

When transitioning to rock classification—where minerals are found in complex, heterogeneous aggregates—the performance characteristics shifted. The models were able to distinguish between same rock type and effectively classifying the different variations of same rock.

- **Robustness of Spatial Data:** The **ResNet-based Image Classifier** proved highly robust. By freezing the early layers of the ResNet-18 backbone and fine-tuning the final layers, the model successfully identified rock types based on texture and mineral associations, even when individual spectral pixels were noisy.
- **Multi-Modal:** The **Multi-Modal Fusion** model (combining 1D Profiles and 2D Images) achieved a high accuracy of **99.98%**. It proved to be the most comprehensive solution for rocks, as it might cross-reference chemical signatures with spatial morphology.

4 Conclusion and Future work

This study compared three data representation strategies for the automated classification of minerals and rocks using high-speed Raman sensors. Our results indicate that, while 1D spectral profiles provide a computationally lightweight solution for rapid identification, they are susceptible to noise in chemically complex samples where overlapping peaks occur. In contrast, 2D spectral imaging architectures significantly improved classification accuracy by capturing better context of the spectral shift.

The fused multi-modal approach demonstrated the highest robustness, successfully identifying mineral phases in heterogeneous rock samples. However, for industrial applications, the choice of model must balance this accuracy against the latency requirements of high-cadence data streams. The results were very accurate, possibly due to the data collected under ideal conditions.

Future work will shift towards the challenges of dynamic, moving environments, where minerals must be classified in real-time as they pass under high-speed sensors in a continuous flow. We aim to optimize these 1D-2D fused architectures to maintain high precision despite the motion-induced artifacts typical of industrial transport systems. This will involve implementing edge-computing solutions to minimize the processing delay between data acquisition and automated sorting decisions.

5 Acknowledgment

The presented research was funded by the Austrian Research Promotion Agency (FFG) under Project NNATT.

References

[1] Chris Carey et al. Machine learning methods for classifying mineral species from Raman spectra. *Journal of Raman Spectroscopy*, 46(10):894–903, 2015.

[2] Larry A Haskin, Alian Wang, Kaylynn M Rockow, Bradley L Jolliff, Randy L Korotev, and Karen M Viskupic. Raman spectroscopy as a tool for in situ lunar mineralogical analysis. *Journal of Geophysical Research: Planets*, 102(E8):19293–19306, 1997.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[4] Jun Liu/m et al. Deep learning for rapid mineral identification using Raman spectroscopy. *Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy*, 252:119511, 2021.

[5] J. Sun et al. High-speed Raman imaging for real-time monitoring of mineralogical processes. *Sensors and Actuators*, 380:133314, 2023.

[6] Xiao Zhang et al. A comparison of 1d and 2d deep learning models for Raman spectral-spatial classification. *Remote Sensing*, 11(12):1473, 2019.

When to Trust the Teacher? Adaptive Coupling in Interactive Robot Learning

Nikolaus Feith
Chair of Cyber Physical Systems
Technical University Leoben
Leoben, Austria
nikolaus.feith@unileoben.ac.at

Elmar Rückert
Chair of Cyber Physical Systems
Technical University Leoben
Leoben, Austria
elmar.rueckert@unileoben.ac.at

Abstract

Interactive robot learning methods typically treat the human teacher as an infallible oracle, limiting the agent's ability to surpass the expert or reject adversarial advice. We introduce MAGIC (Modulated Asymmetric Games for Interactive Control), a framework that formulates interactive learning as an asymmetric leader–follower game between a Teacher and a Learner. The Teacher is an inverse reward field—instantiated with energy-based and flow-matching heads—that scores trajectory segments in SE(3) via contrastive learning on expert demonstrations. The Learner is a hierarchical flow-matching policy (Eye, Brain, Muscle) that maximizes a shaped reward mixing environment reward and Teacher signal. A gradient-agreement coupling determines state-dependent trust: when the Teacher's directional signal agrees with the task critic's gradient, the Teacher is trusted; otherwise it is ignored. We prove that the alternating update satisfies the regularity conditions of two-timescale stochastic approximation. The core pipeline is implemented and unit-tested; we present the framework, its theoretical grounding, and the planned experimental evaluation on 9 ManiSkill3 manipulation tasks, LIBERO with noisy human demonstrations, and real-robot transfer on UR3e and SO-101 arms.

1 Introduction

Interactive robot learning (IntRL) combines human guidance with autonomous skill acquisition, yet most methods treat the teacher as a black-box oracle: the agent imitates demonstrated actions [Ross et al., 2011] or follows a fixed shaped reward [Knox and Stone, 2009, Brys et al., 2015]. This makes it difficult for the learner to surpass the expert or reject misleading advice from noisy or adversarial teachers.

We introduce MAGIC, which formulates IntRL as an asymmetric *leader–follower* bi-level game. The Teacher (leader) shapes the reward landscape via inverse reinforcement learning (IRL) on demonstrations, while the Learner (follower) optimizes a hierarchical policy under a shaped reward that mixes environment reward and the Teacher's signal. A state-dependent coupling weight β_t modulates the Teacher's influence, allowing the Learner to down-weight the Teacher when its advice would reduce task success.

Our contributions are: (1) a formalization of interactive learning as an asymmetric bi-level game with formal regularity guarantees under two-timescale stochastic approximation; (2) an inverse reward field Teacher with modular energy-based (EBM) and flow-matching (FM) heads operating on the same trajectory representations as the Learner; (3) a threshold-free gradient-agreement coupling that compares Teacher, critic, and policy vector fields in subgoal space. This paper presents the framework design, theoretical analysis, and implementation status. Experimental results are forthcoming; Section 4 details current progress and planned evaluation.

The Second Austrian Symposium on AI and Vision (AIROV25).

2 The MAGIC Framework

2.1 Asymmetric Bi-Level Formulation

MAGIC operates on $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}^L, \mathcal{A}^T, P, R^L, R^T, \gamma \rangle$, where only the Learner's actions enter the environment transition kernel. The Teacher's "actions" are interventions on the expert dataset $\mathcal{D}_E$. The Learner maximizes a shaped reward:

$$r_L(s_t, g_t) = \mathbb{E}_{\tau_t \sim f_\psi(\cdot|s_t,g_t)} \big[R_{\text{env}}(s_t, \tau_t) + \beta_t \, \hat{R}_T(s_t, \tau_t) \big], \tag{1}$$

where $\hat{R}_T$ is the Teacher's inverse-reward estimate and $\beta_t \in [0, 1]$ is the coupling weight. The inner loop updates Teacher parameters Θ via contrastive IRL; the outer loop updates Learner parameters via policy gradient on r_L. Timescale separation (N_T:1 update ratio) ensures the Teacher remains approximately stationary from the Learner's perspective [Borkar, 2008].

2.2 Hierarchical Learner (Eye, Brain, Muscle)

The Learner decomposes into three modules: **Eye** (ϕ): a VC-1-based [Majumdar et al., 2023] multi-view encoder mapping images and proprioception to state s_t; **Brain** (π_θ): a flow-matching policy in a 10D subgoal space (3D translation, 6D rotation [Zhou et al., 2020], 1D gripper) using importance-sampling flow-matching actor–critic [Zhang et al., 2025]; **Muscle** (f_ψ): an ActionFlow [Funk et al., 2024] model in SE(3) generating 16-step trajectory segments conditioned on (s_t, g_t). The Brain uses a *dual critic*: Q_{task} trained on R_{env} only, and Q_{shaped} on the full shaped reward. The Muscle retains g_t in the computational graph, making the Jacobian $J = \partial \tau_t / \partial g_t$ available via a single VJP—essential for the gradient-agreement coupling.

2.3 Inverse Reward Field (Teacher)

The Teacher scores trajectory segments via a weighted combination: $\hat{R}_T = \lambda_{\text{EBM}} \hat{R}_T^{\text{EBM}} + \lambda_{\text{FM}} \hat{R}_T^{\text{FM}}$. The **EBM head** maps (s_t, τ_t) to a scalar energy trained with an InfoNCE [Oord et al., 2018] objective using K=4 structured negatives. The **FM head** learns a velocity field via conditional flow matching [Lipman et al., 2022]; its cosine similarity with the target velocity serves as the scalar reward, and the velocity field itself provides a directional signal.

2.4 Gradient-Agreement Coupling

Our primary coupling is threshold-free, exploiting the fact that Teacher, Brain, and Critic all define vector fields in subgoal space. The Teacher's trajectory-space directional signal $\tilde{v}_T$ is projected to subgoal space via the Muscle's VJP: $v_T(s_t, g_t) = \left(\frac{\partial \tau_t}{\partial g_t} \right)^\top \tilde{v}_T(s_t, \tau_t) \in \mathbb{R}^{10}$. The coupling weight is:

$$\beta_t^{\text{agree}} = \max\big(0, \ \cos\big(\nabla_g Q_{\text{task}}(s_t, g_t), \ v_T(s_t, g_t)\big)\big) \, , \tag{2}$$

so the Teacher is trusted when its direction agrees with increasing task value, and ignored otherwise. A warm-up phase ($\beta_t = \beta_0$ for N_{warm} steps) prevents premature Teacher rejection before the critic is informative.

2.5 Convergence Guarantee

Proposition 1 (Regularity of MAGIC)**.** *Under spectral normalization on the EBM head, reward clipping $\hat{R}_T \in [-R_{\max}, R_{\max}]$, entropy-regularized policy optimization with $\alpha > 0$, and Robbins–Monro step sizes with $\alpha_n^L / \alpha_n^T \to 0$, the MAGIC iterates satisfy the Lipschitz, bounded-iterate, and martingale-noise conditions of Borkar [2008] (Ch. 6, Thm. 2) and track the corresponding two-timescale ODE almost surely.*

3 Planned Experimental Evaluation

All experiments use synthetic teachers (trained on offline demonstrations; no live human at interaction time). Results will report mean ± std over 3 seeds.

Table 1: ManiSkill3 evaluation tasks (Franka Panda, motion-planning demos).

Easy	Medium	Hard
PickCube-v1	StackCube-v1	PegInsertionSide-v1
PushCube-v1	LiftPegUpright-v1	PlugCharger-v1
PullCube-v1	PokeCube-v1	StackPyramid-v1

ManiSkill3 (primary benchmark). ManiSkill3 [Tao et al., 2024] provides GPU-accelerated parallel environments (>30K FPS), enabling large-scale ablation sweeps. We evaluate on 9 Franka Panda manipulation tasks spanning three difficulty levels (Table 1), with 100 motion-planner demonstrations per task and 2M training steps per run. For non-interactive baselines (BC, SAC, PPO), we cite ManiSkill3's published numbers.

LIBERO (noisy human demonstrations). LIBERO [Liu et al., 2023] provides tasks with 50 human-teleoperated demonstrations each—noisy and suboptimal compared to motion planners. We select 3 tasks from LIBERO-OBJECT to test whether MAGIC's coupling correctly down-weights a Teacher trained on imperfect data.

Real-world experiments (planned). (1) A UR3e (6-DOF) with three RealSense D435i cameras performs Duplo block assembly using 20 kinesthetic demonstrations, with sim-to-real transfer followed by real-world fine-tuning. (2) A low-cost SO-101 arm (5-DOF, ∼$300) with two USB webcams performs Sort-by-Color and Cup-on-Saucer using 30 teleoperated demonstrations, trained entirely on the real robot. Real-robot experiments are proof-of-concept; full human-in-the-loop evaluation is future work.

Ablation conditions. Full MAGIC (A0) is compared against: no Teacher (A1), EBM-only (A2), FM-only (A3), oracle Teacher (A4), threshold coupling (A5), advantage coupling (A6), naive negatives (A7), DAgger (A9), and ThriftyDAgger [Hoque et al., 2021] (A10). Full ablations run on all 9 ManiSkill3 tasks with 3 seeds; a demo efficiency sweep (10/25/50/100 demos) on PickCube-v1 characterizes data requirements.

Hypotheses. (H1) Teacher improves sample efficiency (A0 vs. A1). (H2) Combined EBM+FM Teacher outperforms either head alone (A0 vs. A2/A3). (H3) Gradient-agreement coupling outperforms threshold and advantage baselines (A0 vs. A5/A6). (H4) Structured negatives improve Teacher quality (A0 vs. A7). (H5) MAGIC enables the Learner to surpass expert demonstrations on $\geq$3 tasks.

4 Implementation Status

The MAGIC pipeline is fully implemented and unit-tested (346+ tests passing). All core components are complete: Eye (VC-1), Brain (ISFM + dual critic), Muscle (ActionFlow SE(3) with differentiable mode), Teacher (EBM with InfoNCE + FM head), all three couplings, and the full training loop with W&B logging. The ManiSkill3 wrapper supports all 9 tasks (0.4s/ep with VC-1) and 100 motion-planner demos have been collected per task.

The immediate next steps are: (1) training the Muscle on ManiSkill3 PickCube-v1, (2) running the no-Teacher baseline (A1) and full MAGIC (A0) for first learning curves, (3) scaling to all 9 tasks and the full ablation suite. LIBERO integration and real-robot experiments follow after simulation results are established.

5 Conclusion

We have presented MAGIC, a bi-level framework for interactive robot learning with gradient-agreement coupling and formal regularity guarantees. The pipeline is fully implemented; experiments on 9 ManiSkill3 tasks, LIBERO, and real-robot transfer are underway.

Acknowledgments

This work was created as part of the research project, MUTAVIA (FO999922732), which are funded by the Österreichische Forschungsförderungsgesellschaft mbH (FFG).

References

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 100. Springer, 2008.

Tim Brys, Anna Harutyunyan, Halit Bener Suay, Sonia Chernova, Matthew E Taylor, and Ann Nowé. Reinforcement learning from demonstration through shaping. In *IJCAI*, pages 3352–3358, 2015.

Niklas Funk, Julen Urain, Joao Carvalho, Vignesh Prasad, Georgia Chalvatzaki, and Jan Peters. Actionflow: Equivariant, accurate, and efficient policies with spatially symmetric flow matching. *arXiv preprint arXiv:2409.04576*, 2024.

Ryan Hoque, Ashwin Balakrishna, Ellen Novoseller, Albert Wilcox, Daniel S Brown, and Ken Goldberg. Thriftydagger: Budget-aware novelty and risk gating for interactive imitation learning. *arXiv preprint arXiv:2109.08273*, 2021.

W Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *Proceedings of the fifth international conference on Knowledge capture*, pages 9–16, 2009.

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *arXiv preprint arXiv:2210.02747*, 2022.

Bo Liu, Yifeng Zhu, Chongkai Gao, Yihao Feng, Qiang Liu, Yuke Zhu, and Peter Stone. Libero: Benchmarking knowledge transfer for lifelong robot learning. *Advances in Neural Information Processing Systems*, 36: 44776–44791, 2023.

Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Yecheng Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence? In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 627–635. JMLR Workshop and Conference Proceedings, 2011.

Stone Tao, Fanbo Xiang, Arth Shukla, Yuzhe Qin, Xander Hinrichsen, Xiaodi Yuan, Chen Bao, Xinsong Lin, Yulin Liu, Tse-kai Chan, et al. Maniskill3: Gpu parallelized robotics simulation and rendering for generalizable embodied ai. *arXiv preprint arXiv:2410.00425*, 2024.

Yuyang Zhang, Yang Hu, Bo Dai, and Na Li. Max-entropy reinforcement learning with flow matching and a case study on lqr, 2025. URL https://arxiv.org/abs/2512.23870.

Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks, 2020. URL https://arxiv.org/abs/1812.07035.

Advanced Robotics Workshop

Towards Recipe-driven Automation Concepts for Large-scale Food Production

Moritz Dorfer and Michael Rathmair
JOANNEUM RESEARCH ROBOTICS
Lakeside B13b
9020 Klagenfurt, Österreich
moritz.dorfer@joanneum.at, michael.rathmair@joanneum.at

Abstract

The hospitality sector is facing a severe shortage of skilled personnel, which results in a significant need of automation and digitalization. In particular, automation of professional kitchen processes poses significant challenges due to the variability of commodities, the mixed presence of humans and machines, and harsh environmental conditions. The introduced concepts integrates a recipe-driven approach including warehouse intralogistics and automation for food processor tending. The ongoing work presented in this extended abstract reflects initial results of an in-depth conceptualization phase supported by simulation-based validation.

1 Introduction and Motivation

In the context of large-scale catering (e.g. for high volume tourism regions having a high density of hotels and restaurants), the primary objective is to produce high-quality meals while maintaining cost-effectiveness. However, the shortage of skilled personnel in the hospitality sector makes this increasingly difficult, forcing the industry to look toward automation as a necessary solution [1]. While automation provides a method of maintaining output, the kitchen environment presents a number of challenges that differ considerably from those experienced in industrial settings [2].

In contradistinction to the assembly of automobiles, food preparation entails the management of biological materials (commodities) that exhibit variability in terms of shape, size, and ripeness [2]. Furthermore, the stringent hygiene standards and arduous conditions including elevated temperatures and moisture engender a challenging environment for sensitive sensors and robotics [1]. Maintaining consistent quality under such economic and physical pressures necessitates an approach that can handle the inherent unpredictability of food products and associated processing [2; 3].

The focus of current research is to concentrate on individual robotic tasks, such as automated frying or flipping, which remain disconnected from broader kitchen workflows [2]. The present paper discusses an alternative approach by integrating recipe-driven warehouse logistics with automation approaches for kitchen processes.

2 Related Research

Automation in food production has gained increasing attention due to labor shortages and the need for efficiency and consistent quality. Within Industry 4.0, robotics, AI, and IoT enable improved traceability, flexibility, and waste reduction in food systems [4]. However, these approaches are largely derived from structured industrial environments and do not directly address the variability and dynamic conditions of professional kitchens.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

Robotic food handling remains challenging due to the deformability, variability, and fragility of food products. Surveys highlight difficulties in perception, grasping, and manipulation, particularly under hygiene and environmental constraints. Recent work on deformable object manipulation and robot learning further emphasizes the need for advanced perception and adaptive control strategies [2]. Existing solutions, however, remain largely task-specific and lack integration across workflows.

Computer vision and deep learning have significantly advanced food quality assessment and process monitoring. Vision-based systems enable real-time inspection of attributes such as color, texture, and freshness, supporting adaptive process control [5].

3 Approaches for recipe-driven Kitchen Automation

The main objective is to ensure that the correct food ingredients are introduced into the food processing machines in the appropriate sequence and with precise timing. Currently, this step represents a highly resource-intensive manual process that requires significant human intervention and coordination. The concepts proposed in this work aim to automate this task, thereby improving process efficiency, reducing manual workload, and enabling more consistent and reliable production operations.

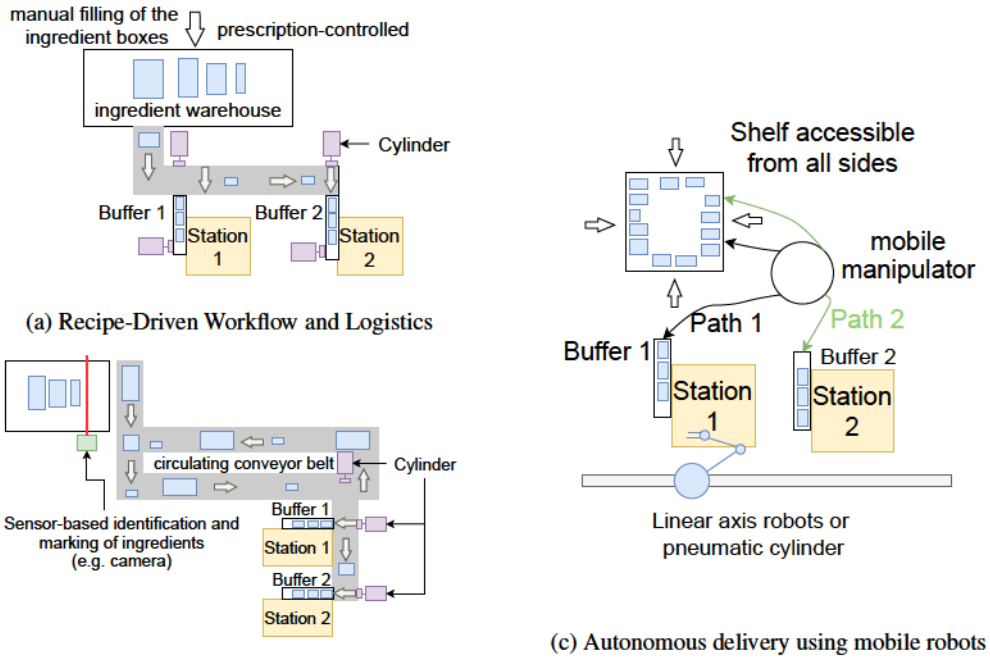
In order to address the complexities of commercial kitchens, the following technologies combine kitchen digitalization with flexible intralogistics and sensor feedback-based optimization. Technologies are focused on the establishment of a resilient production flow by means of the connection of automated storage, transport, and cooking stations (professional food processing machines) through a central digital recipe-based repository.

Workflow Automation and Intralogistics The approaches presented in this paper are based on the results of a design thinking workshop. The main goal of the workshop was to discuss and design concepts for ingredient transport from a central storage to various cooking stations. Main outcome of the workshop were the following three approaches:

1. **Recipe-Driven material supply:** The concept illustrated in Figure 1a is based on a manual compilation of food commodities at an in-house storage/warehouse. A production station (Station 1 or 2) may trigger this manual process and a worker arranges the single commodities on a conveyor belt. This is instructed by a monitor directly placed at the warehouse output. Sensors and pneumatic cylinders are used to sort the boxes into dedicated station buffers. Hence, ingredients for a dish arrive the cooking station at the recipe-correct buffered order. Cylinders at the station triggered by the recipe-based timing plan implement the correct submission of each ingredient into the food processing machine.
2. **Semi-automated process with circulating conveyor belt:** The concept illustrated in Figure 1c is similar to the approach presented above but boxes subsequently circulate along a primary conveyor loop that traverses all stations. When a cooking station signalizes a demand, the system identifies the matching box and a pneumatic diverter pushes it onto a local feed line. This results in a just-in-time commodities supply with predictable smaller buffer sizes at the cooking stations.
3. **Autonomous delivery using mobile robots:** The approach illustrated in Figure 1b integrates a potential fleet of mobile robots. These mobile robots implement a high flexibility commodities supply between the storage and the cooking stations. In this approach tending the food processing machines is implemented using a serial robot arm that is mounted on a linear axes. This fully robot-based approach provides maximum flexibility under high-mix low volume production conditions where potential process adaptions and reprogramming of automation equipment is required.

All three concepts realize a recipe-driven request logic implementing the overall goal of supporting the food processing machines with the correct ingredients sequence and timing. Coordination between storage, transport system, and cooking units is significantly important to enable the system's scalability while maintaining effectiveness and quality.

Vision-Based Sensor Data Feedback to Enhance Process Quality:
The use of biological ingredients poses challenges to automation, primarily due to the lack of consistency, quality and uniformity of commodities [6]. To maintain solid standards, the concepts presented above may integrate a camera-based vision system directly into the workflow.



(a) Recipe-Driven Workflow and Logistics

(b) Semi-automated process with circulating conveyor belt

(c) Autonomous delivery using mobile robots

Figure 1: Concepts for the automation of large-scale kitchen environments with the goal of supporting cooking stations with ingredients at the recipe -defined correct sequence and timing.

These cameras and associated classification trough deep learning approaches, may assess the quality of raw material, verify pre-processing steps, and track the movement of batches. Reliable image processing in a kitchen requires beside of significant computing power to process, reliable hardware and optics protected from heat, moisture and potential deposits. Furthermore, high-quality cameras may enable analyzing color patterns and surface textures which can be used as a quantitative freshens indicator [7]. To close this sensor feedback loop quality data and in particular variations from reference ingredients are transmitted to the cooking station, which automatically adjusts thermal processing parameters for compensation. Beyond inspection of quality and freshness, a camera may monitor the completeness of pre-processing tasks like peeling, cutting or slicing. If an ingredient fails to meet geometric specifications, the system may identify this inadequacy before it reaches the cooking station.

4 Discussion and Next Steps

The state of the ongoing work presented in this extended abstract is in a focused concept-phase supported with prototypes and kitchen automation concepts operated in a digital twin simulation environment.

The transition from manual kitchen operations to an integrated, recipe-driven automated system poses several multi-dimensional challenges. While the technical capabilities of the presented approaches are strong in theory and simulation, their practical implementation has to consider specific operational limitations and further requirements, such as:

- **Recipe database and management:** The presented approaches form a heterogeneous systems, requiring seamless communication between a central digital recipe database, physical transport units, and control of cooking stations. The full process shall be connected to a demand driven ERP system handling customer orders and triggering customer orders.

- **Process resilience:** Sensors and robotic components have to function reliably in environments with high humidity, fluctuating temperatures, and the presence of fats or oils. Process stability has an equal importance as in classical production environments since insufficient delivery of meals to hotels, restaurant, etc. can affect the reliability and reputation of a production kitchen. However, seamless integration has the goal to react highly flexible to demand variations without getting troubles by missing workforce or resource bottlenecks.
- **Hygiene and regulatory requirements:** Hygiene is a core requirement rather than a secondary consideration in food industry. The main goal is to prevent the accumulation of organic material on surfaces and in crevices in order to prevent bacterial growth and meet according standards and regulations. There is a significant difference between components that come into direct contact with the product and those that do not. Within the presented concepts it is planned that direct food contact is avoided as much as possible. Hence, ingredients are transported in food-industry qualified boxes that can be cleaned independently and separately from automation equipment after usage.
- **Economic aspects:** Economic viability of the system depends on its scalability and according ROI estimations. The modular nature of the recipe-driven approach facilitates this but requires a basic digital infrastructure in terms of a central recipe-database, communication and computation infrastructure, etc. Beyond labor savings, the system aims to reduce food waste by better planning capabilities and demand driven production features enhanced by flexible automation concepts. This may directly contribute to cost-effectiveness and a simultaneously improvement in sustainability.

Future work will focus on enhanced simulation to facilitate pre-evaluation and step-by-step experimental prototyping and testing of automation modules in lab environment (goal TRL of 4). These tests are significant to refine system concepts, enhance reliability and prove compliance with strict hygiene standards in the domain of professional food processing. The future goal within upcoming research projects is to design a functional demonstrator for the purpose of conducting real-world trials and present efficiency and impact of modern large-scale kitchen automation to potential stakeholders and customers.

Acknowledgments and Disclosure of Funding

This research was funded by the Carinthian Economic Promotion Fund (KWF), under the name "Food Factory - Machbarkeitsstudie".

References

[1] S. Barasa and Y. Etene, "Robotics in Food Manufacturing Industry in the Industry 4.0 Era," *International Journal of Computer Science and Mobile Computing*, vol. 12, no. 8, pp. 72–77, Aug. 2023.

[2] Z. Wang, S. Hirai, and S. Kawamura, "Challenges and Opportunities in Robotic Food Handling: A Review," *Frontiers in Robotics and AI*, vol. 8, p. 789107, Jan. 2022.

[3] F. Xiong, N. Kühl, and M. Stauder, "Designing a computer-vision-based artifact for automated quality control: a case study in the food industry," *Flexible Services and Manufacturing Journal*, vol. 36, no. 4, pp. 1422–1449, Dec. 2024.

[4] N. Gupta and P. K. Gupta, "Artificial intelligence and robotics in food systems: Trends and applications," *Smart Agricultural Technology*, vol. 9, p. 100566, 2025.

[5] Z. Wang, S. Hirai, and S. Kawamura, "Challenges and opportunities in robotic food handling: A review," *Frontiers in Robotics and AI*, 2022, researchGate preprint version.

[6] K. Shehzad, U. Ali, and A. Munir, "Computer Vision for Food Quality Assessment: Advances and Challenges," *Global Journal of Machine Learning and Computing*, vol. 1, no. 1, pp. 76–92, Feb. 2025.

[7] C. Shen, R. Wang, H. Nawazish, B. Wang, K. Cai, and B. Xu, "Machine vision combined with deep learning–based approaches for food authentication: An integrative review and new insights," *Comprehensive Reviews in Food Science and Food Safety*, vol. 23, no. 6, p. e70054, Nov. 2024.

Building a ROS 2 - Isaac Sim Framework for Dual Arm Manipulation of Rigid Objects and Textiles

Jonas Gschnell
Institute of Robotics
Johannes Kepler University Linz
4040 Linz
jonas.gschnell@jku.at

Alexander Kitzinger
Institute of Robotics
Johannes Kepler University Linz
4040 Linz
alexander.kitzinger@jku.at

Hubert Gattringer
Institute of Robotics
Johannes Kepler University Linz
4040 Linz
hubert.gattringer@jku.at

Andreas Müller
Institute of Robotics
Johannes Kepler University Linz
4040 Linz
a.mueller@jku.at

Abstract

This paper presents a framework that connects ROS 2 with NVIDIA Isaac Sim to support perception-driven dual-arm manipulation, evaluated on rigid objects and textiles. While rigid object handling is achieved after careful parameter tuning, textile manipulation exposes limitations. The paper discusses key integration challenges such as interface alignment, temporal and spatial synchronization, and coordinated dual-arm motion planning.

1 Introduction

Configuring robotic systems to perform complex manipulation tasks reliably is challenging, particularly under cost and time constraints. The interaction of sensors, robots and objects to be manipulated is often difficult to predict, and extensive experimentation is often needed [1]. At the same time powerful consumer-level GPUs enable high-fidelity simulations that promise rapid prototyping of robotic workflows. Photorealistic rendering and physically based simulation suggest that complex perception based manipulation setups can be validated before deployment in real systems. However, in reality integration is still often complex and integration-details are often omitted in favor of high-level insights and results. Practical system constraints—such as controller interface alignment, temporal synchronization, and consistent transformation management—are frequently underreported in favor of high-level task performance. This work investigates these integration challenges through the development of a pipeline using NVIDIA's Isaac Sim as the simulation environment in combination with Robot Operating System 2 (ROS 2). As a representative scenario, dual-arm manipulation of a rigid body and a deformable body, specifically cloth manipulation, is used. Cloth manipulation is introduced as a test case that exposes simulation limitations, while rigid object handling serves to validate the framework. By analyzing recurring integration issues and system-level constraints, the paper aims to provide practical design guidelines that improve reproducibility and Sim-to-Real workflows for perception-driven robotic manipulation.

2 System architecture and application examples

To investigate the integration effort for linking ROS 2 and Isaac Sim, a dual-arm manipulation cell is implemented. ROS 2 provides a large ecosystem of reusable packages covering perception, motion

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

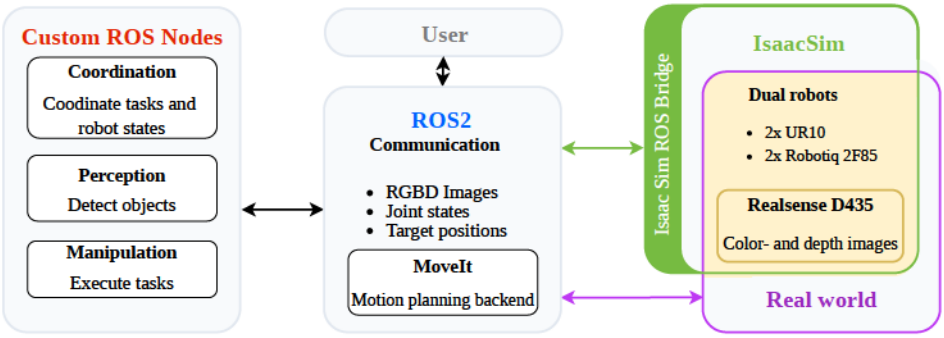


Figure 1: System architecture for simulated and real world setup.

planning, control, and many more [2]. Within the ROS framework Universal Robot Description Format (URDF) files are used to describe robot kinematics, geometry, inertial properties, and sensors. Further, they define the structural and control interfaces expected by downstream tooling. Reusable URDF configurations are parameterized using Xacro, where macros are used to set up a complete robot cell. To assemble the same cell in NVIDIA Isaac Sim Universal Scene Description (USD) assets are composed to a complete scene using Python. The ROS package MoveIt2 is used as a motion planning backend, it handles inverse kinematics, collision checking, trajectory generation, and execution management [3]. Semantic robot information such as planning groups, end effectors, and allowed collisions is specified via Semantic Robot Description Format (SRDF) files. The simulated workcell consists of two Universal Robots UR10 arms equipped with Robotiq 2F-85 grippers. They are placed on a table and a spatial fixed RGB-D camera (Intel RealSense D435) is used for visual observation. An overview of the complete hardware and software setup is shown in fig. 1.

Dual-arm manipulation of rigid objects: To validate the framework the virtual setup is tested by manipulating a simple rigid object with the dual arms. The procedure includes a cooperative movement with both grippers holding the object. Reliably grasping the object in Isaac Sim is challenging, as small changes to the grippers stiffness, damping, contact offsets, or friction coefficients can switch a grasp from stable to completely failing. Simulating realistic contact interactions is difficult due to discretization limits, geometry approximations, friction modeling and contact solver stability. Discrete collision checks with practical time steps sometimes lead to tunneling and interpenetration for fast or small-contact grasps, unless the number substeps and solver iterations is increased significantly. Further, the use of convex geometry approximations leads to higher performance, but misaligned contact patches, especially for thin fingers and complex objects simulation stability suffers. For rigid objects these problems can be reduced by using small time steps and decreasing the grippers stiffness.

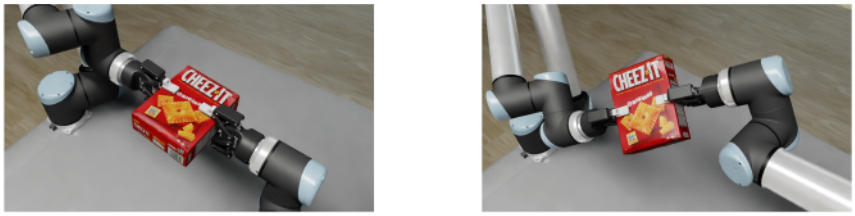


Figure 2: Cooperative dual-arm manipulation. (left) Initial configuration. (right) Final configuration.

Textile manipulation: Isaac Sim also enables more advanced physics simulations, including particle-based models that can represent deformable objects such as cloth. Since textile manipulation is becoming increasingly important, the simulation of cloth manipulation gets explored as a challenging test case. To set up the simulation, a mesh of a standard T-shirt is imported into Isaac Sim. Once imported in Isaac Sim a particle-based model is applied, transforming mesh vertices into particles connected by virtual springs and dampers. Additional material parameters such as mass, density, friction, adhesion, drag, and lift can be specified to control the behavior of the simulated cloth. Despite extensive parameter tuning, from simulation settings such as solver iterations and simulation step time to cloth parameters such as mass, friction, stiffness or contact offsets, stable

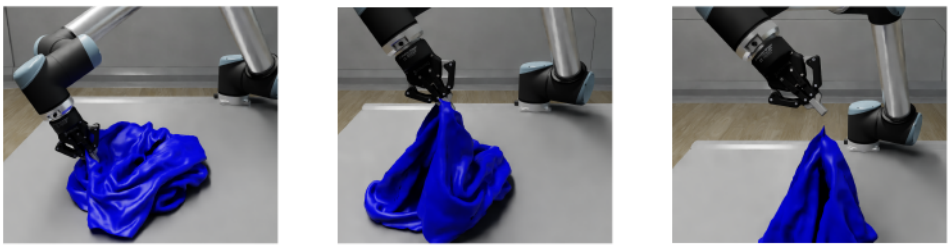


Figure 3: Attemt to lift cloth from table. (center) Visible tunneling of gripper through fabric.

grasping of the fabric could not be achieved. As illustrated in fig. 3, the cloth repeatedly slips out or tunnels through the grippers. Because the cloth mesh represents a single layer, the contact issues discussed above become particularly pronounced, making reliable grasping difficult to achieve under the tested conditions.

3 Integration Issues

In this section common integration challanges are discussed, and possible solutions are presented. The topics covered include interface mismatches, as well as time and spatial synchronization issues between Isaac Sim and ROS. Finally, the particular implementation for cooperative dual arm movement is discussed.

Interface alignment: Communication between ROS 2 and the simulation environment is realized via the Isaac Sim ROS Bridge, which exposes simulation data and takes in actuator commands through ROS Topics. A robot setup using ROS controllers needs specific command interfaces and state interfaces. The controllers then use this abstracted interfaces to send commands to the robot and receive state updates, regardless of whether the robot is simulated or real. The ROS Bridge does not provide such interfaces, it only enables the simulation articulation controllers to both publish states to and receive commands from ROS Topics. As a consequence, an additional adaptation layer is required to align controller expectations between the two ecosystems. The specific hardware interfaces that are initialized on launch must be defined in the URDF file. In particular, the *topic_based_control*-plugin can be used in the URDF files. It makes hardware interfaces available, translates commands to standard ROS Topics and returns state feedback to the hardware interface.

Temporal consistency: Physics steps and rendering in Isaac Sim vary in duration depending on the current scene complexity. Therefore, Isaac Sim does not necessarily simulate in real time. However, control systems, trajectory execution and perception depend on a consistent timeline. As a consequence, it is necessary to synchronize ROS Nodes with Isaac Sim's simulation time. The simulation time can be published on a ROS Topic via the Isaac Sim ROS Bridge. Per default, Nodes take the system time to synchronize, but by changing Node parameters an external time source can be used as reference.

Spatial consistency The photorealistic camera streams generated by Isaac Sim can be used for generating synthetic data for deep neural networks or reinforcement learning algorithms, as well as validating perception-based control pipelines. Usually in such pipelines object poses detected in camera coordinates must be transformed into a common reference frame. ROS 2 provides such capabilities with the build in transformation framework TF2, which maintains a dynamically updated tree of coordinate transforms between all links. Robot joint states are already published from Isaac Sim to ROS as part of the control feedback loop. These joint states are used by TF2 to update the robot kinematic chain. In contrast, a fixed camera mounted in the environment can be defined by a static transform relative to the world frame. Such static transforms are derived from the URDF file if not defined otherwise. A common source of integration errors arises from differences in coordinate frame conventions between Isaac Sim and ROS. As a result, additional transformations may be required when publishing camera poses through the ROS bridge. Careful verification of coordinate frame conventions is therefore necessary to ensure consistent perception results.

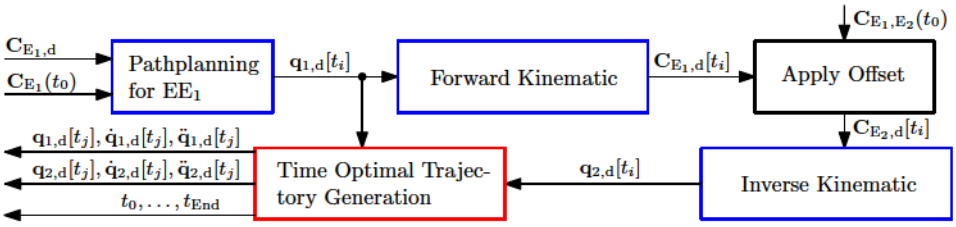


Figure 4: Workflow to generate a time optimal trajectory for cooperative dual-arm movement.

Dual arm movement: For cooperative manipulation tasks, the relative position and orientation between the end effectors must remain constant for the whole trajectory. MoveIt exposes several ROS Services and Actions to plan and execute movements, however, it does not directly expose a planning interface that enforces a fixed relative transform between multiple end effectors. The approach proposed here, allows to remain closely integrated with the MoveIt framework. Consequently, collision checking, trajectory validation, and time parameterization remain within MoveIt. This preserves MoveIt's safety mechanisms and maintains a single source of truth for the robot configuration. Fig. 4 illustrates the workflow used to generate synchronized joint trajectories for both end effectors E_1 and E_2. Given the initial configuration $\mathbf{C}_{\mathrm{E}_1}(t_0)$ and the desired configuration $\mathbf{C}_{\mathrm{E}_1,\mathrm{d}}$ of end effector E_1 the desired trajectory is first planned and converted to joint space resulting in $\mathbf{q}_{1,d}[t_i]$. Using forward kinematics and the relative configuration of E_2 w.r.t. E_1, $\mathbf{C}_{\mathrm{E}_1,\mathrm{E}_2}(t_0)$, the corresponding configurations $\mathbf{C}_{\mathrm{E}_2}[t_i]$ are computed and converted to joint paths. Finally, a time-optimal trajectory generator produces synchronized position $\mathbf{q}_{\mathrm{d}}[t_j]$, velocity $\dot{\mathbf{q}}_{\mathrm{d}}[t_j]$, and acceleration $\ddot{\mathbf{q}}_{\mathrm{d}}[t_j]$ profiles for both robots. The blue steps in the figure are accessible via Actions or Services, the red marked time optimal trajectory generation [4] is available as a c++ library. However, by writing a simple ROS Node it can be wrapped in a ROS Service.

4 Conclusion

Although Isaac Sim allows rapid advances in perception driven - and RL workflows, setting up a robust framework is not a rudimentary task. Nevertheless, with some know-how and experience it is possible to utilize the high fidelity simulation to reduce integration time. The dual arm manipulation of rigid objects was successful after carefully tuning the parameters. However, when investigating the manipulation of textiles, the limitations of current cloth simulation became apparent, particularly with regard to reliable grasping. Future work will focus on experimental validation on a real robot system and on improving textile manipulation strategies, with the long-term goal of enabling tasks such as pursued by the ICRA Cloth Competition [5].

Acknowledgments and Disclosure of Funding

This work has been supported by the "LCM – K2 Center for Symbiotic Mechatronics" within the framework of the Austrian COMET-K2 program.

References

[1] Ester Martinez-Martin and Angel P Del Pobil. Vision for robust robot manipulation. *Sensors*, 19 (7):1648, 2019.

[2] Morgan Quigley, Ken Conley, Brian Gerkey, Josh Faust, Tully Foote, Jeremy Leibs, Rob Wheeler, and Andrew Ng. Ros: an open-source robot operating system. volume 3, 01 2009.

[3] Sachin Chitta, Ioan Sucan, and Steve Cousins. Moveit![ros topics]. *IEEE Robotics & Automation Magazine*, 19(1):18–19, 2012.

[4] Tobias Kunz and Mike Stilman. Time-optimal trajectory generation for path following with bounded acceleration and velocity. *Robotics: Science and Systems VIII*, pages 1–8, 2012.

[5] Victor-Louis De Gusseme et al. A dataset and benchmark for robotic cloth unfolding grasp selection: The icra 2024 cloth competition. *International Journal of Robotics Research*, 2026.

Embedded Haptic Control for Robotic Grasping using a Tactile Sensor System

Thomas Kammerhofer
Chair of Automation and Measurement
Technical University of Leoben
Peter-Tunner-Straße 25, 8700 Leoben
thomas.kammerhofer@unileoben.ac.at

Thomas Thurner
Chair of Automation and Measurement
Technical University of Leoben
Peter-Tunner-Straße 25, 8700 Leoben
thomas.thurner@unileoben.ac.at

Abstract

Tactile sensing is essential for dexterous robotic manipulation, enabling reliable contact detection, grasp assessment, and safe interaction with delicate objects. In this work, we present a finger-shaped tactile sensor system based on a 2D array of MEMS barometric pressure sensors, designed to mimic the compliance and geometry of the human fingertip. The system integrates real-time contact force measurements utilizing the pressure sensor array, in combination with acceleration data from an onboard Inertial Measurement Unit (IMU), allowing both precise point-of-contact estimation and dynamic impact detection. A dedicated microcontroller (µC) acts as a local processing and coordination node, responsible for closed-loop grasp and movement control, while a PC manages high-level communication between the µC and a robotic gripper. In addition, a hardware-level GPIO handshake between the control unit of a collaborative robot and the processing node enables deterministic synchronization between robotic arm positioning and grasp execution. Experimental validation of both the tactile sensor system and the robotic gripper control demonstrates robust operation across the conducted performance tests, with no malfunctions or object damage, as tactile feedback enables real-time grasping control throughout object manipulation. These results highlight the advantages of our tactile sensing solution as a cost-effective, versatile approach for enhancing robotic touch and advancing adaptive object-handling strategies.

1 Introduction

Dexterous object manipulation requires a continuous exchange of sensory information between the hand and the brain. In humans, this interaction is enabled by the highly specialized tactile sensing capabilities of the fingers, which allow not only the detection of contact forces but also the perception of object geometry, texture, and compliance. Through these rich sensory channels, humans adjust grip strength, adapt to unexpected disturbances, and execute delicate tasks such as tool use or object manipulation. Artificial replication remains a major challenge in robotics, where limited tactile feedback often constrains dexterity and adaptability. The integration of finger-shaped tactile sensors into robotics represents a significant advancement in robotic perception and manipulation capabilities, moving closer to the capabilities of the human sense of touch. These sensors provide essential tactile feedback, enhancing the robot's ability to interact effectively with its environment. Prior research shows that robots equipped with tactile sensors can achieve high levels of manipulation accuracy in uncertain environments (1; 2; 3), particularly when tactile data is exploited in closed-loop control for adaptive grasping and recognition of object properties (4; 5; 2; 6).
In addition to high sensitivity and resolution for accurate detection of contact forces, a mechanically soft and flexible sensor interface with a touch quality resembling human skin is highly advantageous, particularly for handling delicate items (such as e.g., fruits and vegetables) in agricultural harvesting,

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

processing, and supply logistics (7), (8). Moreover, as robotic systems increasingly evolve toward humanoid designs and human–robot interaction (9), mechanical properties such as softness and human-skin-like tactile perception are becoming critical for safe and natural interaction during object manipulation.

While tactile sensors have been an active field of research for more than a decade (10; 11), finger-based tactile sensors based on barometric pressure sensing are now emerging as a promising solution, driven by increased availability and affordability of low-cost MEMS barometric sensors. Early designs demonstrated that barometric sensors could be adapted to finger-like structures. Building on such work, later studies have explored utilizing MEMS-based barometric sensor arrays to enable high-resolution tactile mapping by detecting air pressure variations induced by external forces (12), while soft encapsulations mimicking the mechanical properties of human skin improved compliance and sensitivity (13; 14). This sensor information can be used not only for contact detection and force estimation, but also for slip detection (15) and dynamic contact interpretation (16), enabling direct control of an adaptive robotic gripper.

In this work, we present a finger-shaped tactile sensor system that combines a 2D array of MEMS-based barometric pressure sensors with a compliant fingertip design. The sensor array provides spatial pressure measurements across the fingertip, which are used to detect object contact and determine stable grasp conditions. The tactile signals are processed by an embedded controller and used for closed-loop robotic grasp control. This allows an integrated, in-line assessment of the current grasp and real-time adaptation of gripper behavior. To ensure reliable manipulation, a hardware-level handshake between the robot controller and the tactile processing unit is implemented to synchronize arm motion and grasp execution. Experimental evaluation demonstrates robust object interaction without damage, highlighting the potential of barometric tactile sensing as a cost-effective solution for tactile-driven robotic manipulation.

2 Tactile Sensor System for Robot Control

This section presents the architecture of the proposed tactile sensor system, starting with the sensing principle based on barometric pressure measurements and continuing with the supporting electronics for data acquisition and integration into the robotic control framework.

2.1 Tactile Sensor Concept

The developed tactile sensor solution, first introduced in (14), is designed to emulate selected mechanoreceptive functions of the human skin. The improved version, as presented in this paper is tailored to mimic the properties of the human finger. Each tactile sensor system, acting as an artificial finger, integrates 12 MEMS-barometric pressure sensors of type DPS368 (from Infineon Technologies AG (17)), hermetically sealed from the cell exterior by a flexible cell cover layer made from soft silicone (hardness Shore 25A). The shape (flat or finger-like) of this cell cover layer can be adapted depending on the target use case. A visual representation of the different types can be found in Figure 1, right, between the printed circuit boards (PCB) onto which the barometric pressure sensors are soldered, and the silicone cell cover layer to ensure a good sealing to the exterior and avoid cross-talk between individual sensor cells. The used tactile sensor system is conceptually illustrated in Figure 1, left, consisting of a 3D-printed mounting system (a), the printed circuit board (PCB) onto which the barometric pressure sensors are soldered (b), and the flexible cell cover (d). A resin-printed intermediate layer (c) was introduced between the PCB and the soft encapsulation to ensure proper sealing against the exterior and to avoid cross-talk between individual sensor cells. The sensors are arranged in a 4×3 array with a sensor pitch of 10 mm to enable spatially resolved measurements of contact forces.

Each barometric pressure sensor in the tactile sensor system can be individually addressed via Inter-Integrated Circuit (I^2C). Therefore, unique I^2C addresses were pre-assigned and fused into the sensors during the assembly process. This setup allows multiple sensors to operate on the same communication bus without requiring additional multiplexing hardware. In addition, an ICM-42670-P six-degree-of-freedom (DoF) inertial measurement unit (IMU, from TDK InvenSense (18)) is integrated to capture motion-related data for vibration detection, and capture orientation changes and general robot-associated movements. With a sampling frequency of up to 1.6 kHz, the system can capture vibrations well beyond the perceptual frequency range of human skin ($\approx$ 500 Hz).

This capability makes it suitable to mimic the fast-adapting mechanoreceptors responsible for high-frequency vibration sensing in human touch.

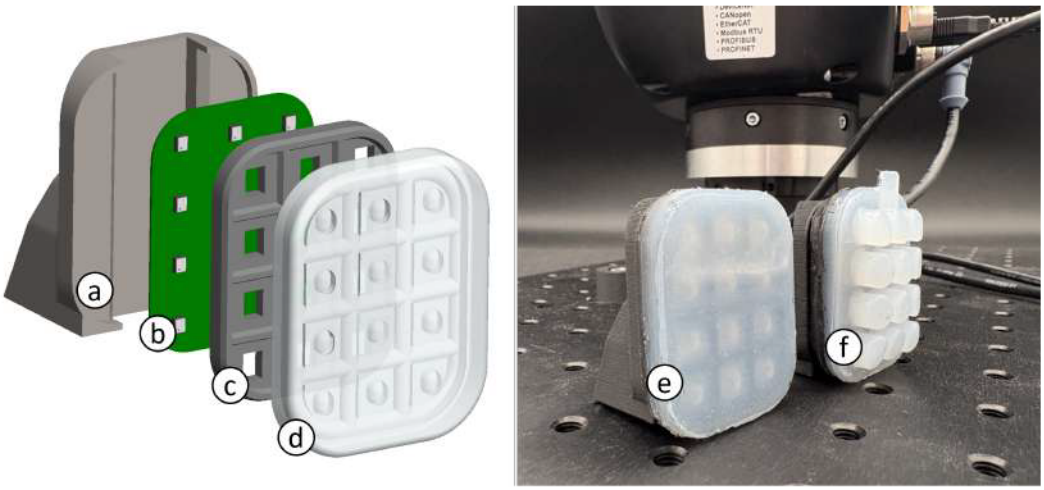


Figure 1: Conceptual illustration of the finger replacement for the robotic gripper, consisting of a 3D-printed mounting structure (a), the printed circuit board including sensors (b), a resin-printed intermediate layer (c), and a flexible cell cover layer (d). The developed prototypes can be seen in (e) and (f)

2.2 System Integration and Data Acquisition

A seamless integration of the tactile sensing system into the 3-Finger Adaptive Robot Gripper (from Robotiq (19)) can be achieved by replacing its original fingertips with custom 3D-printed ones (Figure 1). These replacements are designed to provide mounting surfaces and cable routing options for the tactile sensor modules, allowing the sensors to be directly attached without modifying the gripper's mechanical structure. The readout circuit for the tactile sensor array is placed close to the 3-Finger Adaptive Robot Gripper and is connected via a ribbon cable to the artificial finger. The readout circuit is centered around a CY8C6244-LQI microcontroller (µC, from Infineon Technologies AG (20)), which supports multiple communication protocols like I^2C, Universal Asynchronous Receiver/Transmitter (UART) and Universal Serial Bus (USB), among others. The µC is responsible for high-speed data acquisition from the pressure sensor array at a sampling rate of 128 Hz, and from the IMU at a 1 kHz readout frequency, as well as for in-line data processing of the obtained data. It is important to note that this data rate is limited by the sensors rather than by the capabilities of the readout circuit.
By providing a virtual COM port, the system allows direct and efficient data exchange between the µC and a PC. For debugging measures, the sensor data can be transmitted to a PC for in-line assessment (in LabVIEW) or offline analyses (MATLAB) via UART.

The Robotiq 3-Finger Adaptive Gripper equipped with tactile sensor modules is mounted on a UR5 collaborative robot. While the gripper is controlled via USB communication by the µC, the overall manipulation task requires synchronization between arm positioning and grasp execution. To enable this coordination, a direct electrical interface is established between the UR5 controller and the tactile processing node using digital GPIO signals. The UR5 provides a position-confirmation signal once the predefined pick pose is reached, and the µC returns a confirmation signal after stable grasp detection.

2.3 Gripper Control

By utilizing the USB interface from the adaptive gripper, a communication path between the µC and the gripper can be established. However, as neither device can open its counterpart's communication port, an intermediate control layer is introduced. This layer is implemented on a PC and is responsible for managing the communication channels by opening the respective COM ports, forwarding messages between the µC and the gripper and properly closing the previously opened ports after

testing. A simple LabVIEW script fulfills this role, enabling reliable bidirectional communication and ensuring seamless integration of the tactile sensor system with the robotic gripper. For direct gripper control, a script consisting of at least two communication sections is necessary: an initialization section and a runtime phase in which the µC continuously sends positional commands to control the robot's movement. In this way, the Robotiq gripper can be activated and moved into desired positions, with possible variations in movement direction, movement speed and gripping force. These movements can therefore be adjusted based on the feedback obtained from the tactile sensor array. A more detailed description of the communication between µC and gripper can be found in section 3.

3 CONTROL FLOW

The control flow of the system is divided into two layers: µC firmware for sensor handling and real-time grasp control, and a LabVIEW script for communication and coordination with the robotic gripper.

3.1 Microcontroller Firmware

The final control firmware directly runs on the CY8C6244-LQI µC. After board support package initialization, the system configures the USB interface for host PC communication and initializes the (I^2C) bus for sensor access. Once all initializations are complete, it configures the DPS pressure sensors, including the output data rate, oversampling, continuous measurement mode, and offset correction.

Additionally, the IMU (ICM42670-P) and the robot gripper are initialized. For gripper control, the firmware issues an activation command to the gripper and continuously monitors its status until confirmation of successful activation is received. Once the handshake is complete, the system enters its main loop.

Within this loop, the µC continuously monitors both the state of the gripper and the robotic arm, as well as the tactile sensor array information and the accelerometer data. At the beginning of each grasping cycle, open-grip movement is executed, followed by a one second pause. Afterwards, the gripper closes until an object is detected, which can be determined by a measurable increase in the average cell pressure of the tactile sensor system. The pressure sensor array provides real-time pressure measurements with a resolution down to 1 Pa and sampling rates up to 128 Hz. This allows a precise estimation of contact points. Point-of-contact determination is further improved by correlating pressure changes with acceleration spikes captured by the IMU, indicating the moment of impact. Once contact is confirmed, the gripper reduces its closing speed until the tactile array signals that a predefined pressure threshold has been exceeded, indicating a stable grasp. Gripper motion then stops to prevent excessive force and thus potential damage to the manipulated object, and corrective actions such as reopening or repositioning can be triggered if needed. Subsequently, the GRIP_STABLE indicator is transmitted, and the UR5 starts its movement.
These sensor values can be logged via UART or USB for thorough data analyses. For debugging purposes, the software is capable of transmitting full sensor datasets,including pressure values, acceleration readings, and processed features such as standard deviation through the virtual COM port or a separate UART to USB converter to the host PC.

3.2 Robot–µC Handshake

To ensure deterministic synchronization between the UR5 collaborative robot and the tactile processing node, a hardware-level handshake mechanism based on GPIO signals was implemented. To indicate that the robot has reached the predefined position, the UR5 controller asserts a logical HIGH signal ("ROBOT_READY") on a dedicated GPIO line connected to the µC. Similarly, once the tactile sensor system detects a stable grasp condition, the µC asserts a logical HIGH signal ("GRIP_STABLE") on a separate GPIO line connected to the UR5 control unit.
Upon detecting the ROBOT_READY signal, the µC initiates the gripper closing procedure and continuously evaluates the tactile pressure data. Initial contact is detected when the average tactile sensor cell pressure exceeds the contact threshold, while a firm grasp is confirmed once the stability threshold is reached. At this point, the µC asserts GRIP_STABLE to indicate successful object acquisition. The UR5 monitors this signal and resumes its programmed trajectory (e.g., lifting

motion) only after GRIP_STABLE is detected. An overview about the described communication can be found in Figure 2, left.

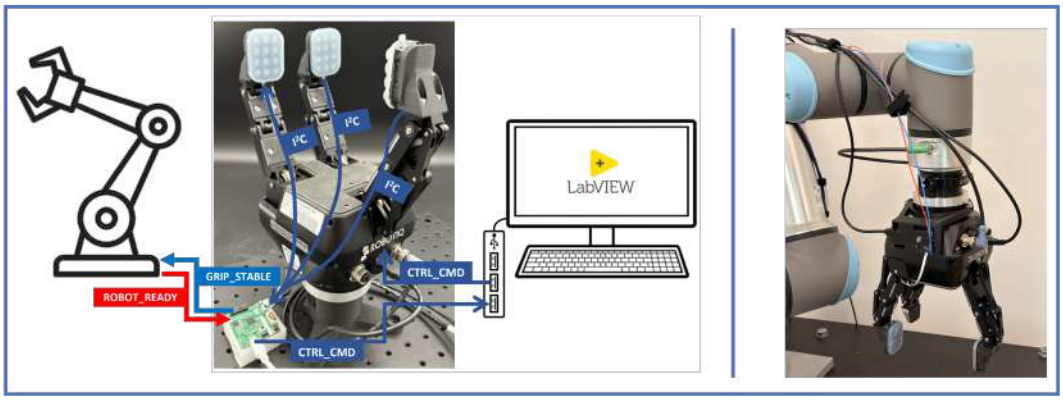


Figure 2: Left: Communication between individual components in the control loop. The µC is reading the sensor data, and sends corresponding control commands (CTRL_CMD) via usb to the gripper. Movement coordination is handled via logical signals (GRIP_STABLE / ROBOT_READY) between the µC and the UR5. Right: Test setup using the teactile sensor systems attached to the Robotiq Adaptive 3-Finger Gripper and the UR5 collaborative robot.

3.3 Automation Script

A complementary LabVIEW script handles higher-level communication between the µC and the gripper. Its responsibilities are graphically depicted in Figure 3 and include:

- opening the COM ports,
- activating the gripper,
- flushing of µC queue after activation,
- forwarding movement commands to the gripper,
- forwarding positional feedback from the gripper to the µC, and
- closing the COM ports after test finalization.

While USB communication is used for command forwarding and status monitoring, motion–grasp synchronization is handled exclusively via GPIO signals to ensure deterministic timing. This separation of tasks ensures robust initialization, high-speed sensor readout, real-time decision-making for object detection, and a flexible communication pipeline that enables adaptive robotic manipulation based on tactile feedback.

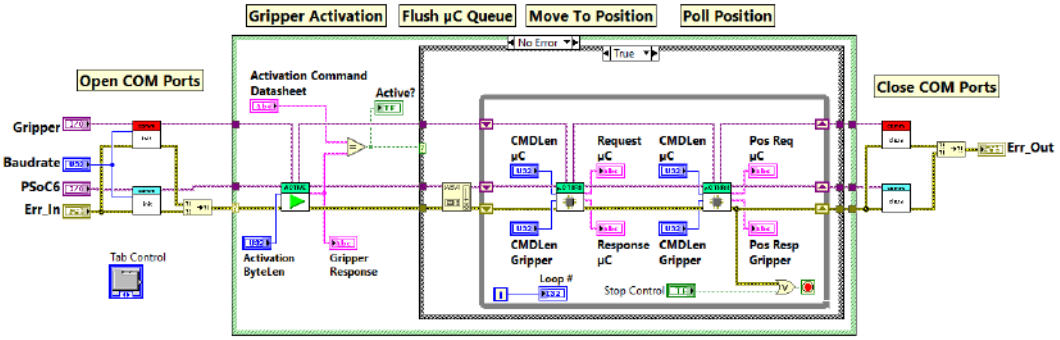


Figure 3: Overview of LabVIEW script used for PC-based automation of the gripper and µC interaction

4 Results

To evaluate the performance of the tactile sensing system, a series of manipulation cycles as described in section 3 was carried out by the Robotiq 3-Finger Gripper on various objects to evaluate the reliability of the algorithm. The objects varied in size, compressibility, weight, and surface texture. The gripping force was set to 20 N, resulting in a force of 6.6 N per finger, or 0.55 N per tactile sensor cell. It should be noted, however, that the force values are estimates and averages. An exact determination of force, or a relationship between applied force and measured cell pressure, can be established following a system characterization as described in (21).

Figure 4, left, shows the recorded average pressure change including the $\pm 2\sigma$ range measured by the sensor array alongside the acceleration data from the IMU (Figure 4, right) from a test set resulting from grasping a fabric sample with a mass of approximately 150 g, corresponding to the typical mass of an average cotton T-shirt. During each manipulation, in which the grasping process lasted 6 s per iteration, distinct movement phases could be identified from the pressure and acceleration profiles. The closing phase (0.30–1.35 s) corresponds to the transport of the fingers toward the object. This is followed by a brief contact phase (1.35–1.60 s), marking the first interaction with the target and triggers a lower closing speed. The gripping phase (1.60–4.75 s) is initiated when an average change in pressure $\Delta p = 0.25$ hPa was detected across all sensor cells, and the grip was defined as stable once the applied average pressure exceeded 3 hPa (≈ 0.55 N). The releasing phase (4.75–4.85 s) reflected the loosening of grip, after which the opening phase (4.85–5.30 s) indicated the withdrawal and hand aperture increase. In total, 595 iterations were performed in this example, allowing for robust segmentation and statistical evaluation of the task phases.

The point of contact can be reliably determined from both the pressure signals and the acceleration

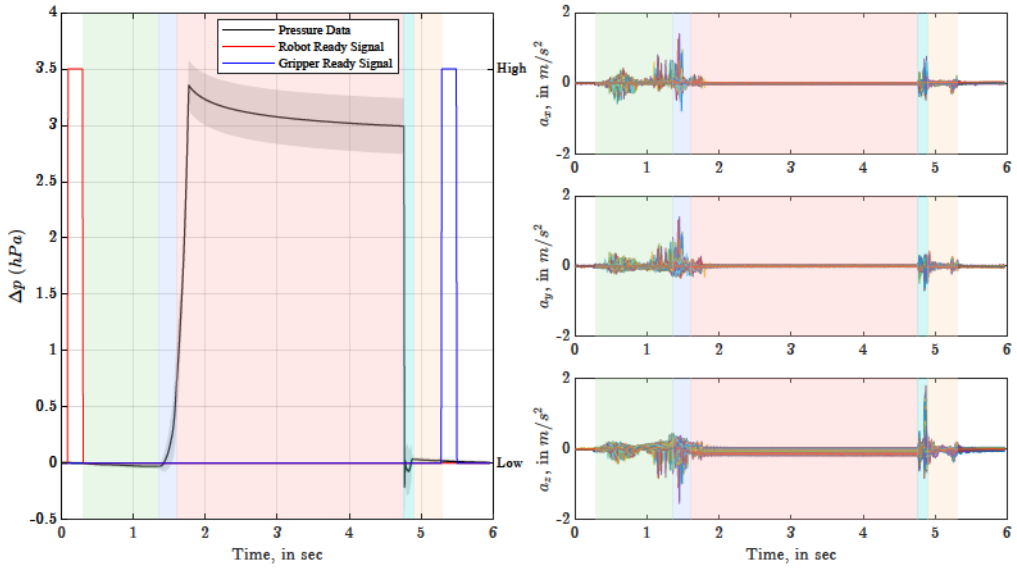


Figure 4: Pressure and acceleration patterns when grasping a compressible material. The grasping process can be subdivided into closing (green), contact (purple), grasping (red), releasing (cyan), and opening (orange) phases.

spike caused by the finger–object impact, and the accelerometer data can also be exploited in future object-handling strategies rather than relying exclusively on pressure information.

Across all conducted experiments, the system operated without malfunction and no objects were damaged due to excessive gripping force. The low-pressure threshold for contact detection, combined with the subsequent reduction in closing speed, consistently preserved the structural integrity of the grasped objects. No premature lifting motion was observed, and the UR5 resumed its trajectory exclusively after GRIP_STABLE assertion, confirming reliable handshake operation.

5 CONCLUSION & FUTURE WORK

In this work, a tactile sensor system was developed and integrated into a Robotiq 3-Finger Gripper to enable sensor-driven robotic manipulation. Deterministic robot–gripper coordination via hardware handshakes was introduced to establish a reliable communication between the three essential system components: µC, gripper, and robotic arm. The system is centered around a µC that handles high-speed data acquisition from an array of pressure sensors and performs real-time data processing. Complementary information is obtained from an integrated IMU, allowing tactile events to be correlated with motion and vibration signals.

A seamless mechanical integration was achieved by replacing the gripper's fingertips with custom 3D-printed parts, onto which the tactile sensor modules were mounted. This modular approach preserved the gripper's functionality while enhancing it with high-resolution tactile feedback. The combination of USB-based communication and a LabVIEW control layer allowed reliable bidirectional data exchange between the µC and the gripper, enabling adaptive, sensor-based control strategies.

Experimental results demonstrated that the system can robustly detect object contact with high precision. Pressure changes as small as 1 Pa can be measured within individual sensor cells, and when combined with acceleration data, the point of contact can be estimated reliably. Plots of opening and closing iterations illustrated the strong correspondence between pressure transients and IMU signals, confirming the robustness of the approach.

Looking ahead, future work will focus on further increasing the temporal and spatial resolution of the tactile system. Pressure sensors capable of sampling at rates up to 500 Hz are already under consideration, which would further reduce the latency between different positional requests down to 2 ms.

In addition, the use of smaller sensor elements will reduce the sensor pitch, resulting in an increased spatial resolution of the tactile sensor system. In addition, the spatial resolution can be further increased by utilizing the interpolation algorithm, introduced in (22). Another important step will be the implementation of a slip detection algorithm, based on discontinuity detection (23; 24), directly on the µC. This will enable the gripper to autonomously detect and react to incipient slip conditions in real time, further enhancing its adaptability and robustness in manipulation tasks.

Acknowledgments and Disclosure of Funding

The authors gratefully acknowledge the financial support provided by the Austrian Research Promotion Agency (FFG) and the Federal Ministry for Climate Action, Environment, Energy, Mobility, Innovation, and Technology under the project "MUTAVIA" (Project Nr. FO999922732).

References

[1] Z. Deng, Y. Jonetzko, L. Zhang, and J. Zhang, "Grasping force control of multi-fingered robotic hands through tactile sensing for object stabilization," *Sensors*, vol. 20, p. 1050, 2020.

[2] G. Sutanto, N. Ratliff, B. Sundaralingam, Y. Chebotar, Z. Su, A. Handa, and D. Fox, "Learning latent space dynamics for tactile servoing," pp. 3622–3628, 2019.

[3] D. Gomes, Z. Lin, and S. Luo, "Blocks world of touch: exploiting the advantages of all-around finger sensing in robot grasping," *Frontiers in Robotics and Ai*, vol. 7, 2020.

[4] R. Dahiya, P. Mittendorfer, M. Valle, G. Cheng, and V. Lumelsky, "Directions toward effective utilization of tactile skin: a review," *Ieee Sensors Journal*, vol. 13, pp. 4121–4138, 2013.

[5] S. Tian, F. Ebert, D. Jayaraman, M. Mudigonda, C. Finn, R. Calandra, and S. Levine, "Manipulation by feel: touch-based control with deep predictive models," pp. 818–824, 2019.

[6] H. Yousef, M. Boukallel, and K. Althoefer, "Tactile sensing for dexterous in-hand manipulation in robotics—a review," *Sensors and Actuators a Physical*, vol. 167, pp. 171–187, 2011.

[7] Y. A. AboZaid, M. T. Aboelrayat, I. S. Fahim, and A. G. Radwan, "Soft robotic grippers: A review on technologies, materials, and applications," *Sensors and Actuators A: Physical*, vol. 372, p. 115380, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0924424724003741

[8] P. Beckerle, R. Kõiva, E. A. Kirchner, R. Bekrater-Bodmann, S. Dosen, O. Christ, D. A. Abbink, C. Castellini, and B. Lenggenhager, "Feel-Good Robotics: Requirements on Touch for Embodiment in Assistive Robotics," *Frontiers in Neurorobotics*, vol. 12, p. 84, Dec. 2018.

[9] Y. Tong, H. Liu, and Z. Zhang, "Advancements in Humanoid Robots: A Comprehensive Review and Future Prospects," *IEEE/CAA Journal of Automatica Sinica*, vol. 11, no. 2, pp. 301–328, Feb. 2024. [Online]. Available: https://ieeexplore.ieee.org/document/10415857/

[10] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, "Tactile sensing—from humans to humanoids," *IEEE Transactions on Robotics*, vol. 26, no. 1, pp. 1–20, 2010.

[11] J. A. Fishel and G. E. Loeb, "Sensing tactile microvibrations with the biotac — comparison with human sensitivity," in *2012 4th IEEE RAS EMBS International Conference on Biomedical Robotics and Biomechatronics (BioRob)*, 2012, pp. 1122–1127.

[12] Y. Tenzer, L. P. Jentoft, and R. D. Howe, "The feel of mems barometers: Inexpensive and easily customized tactile array sensors," *IEEE Robotics Automation Magazine*, vol. 21, no. 3, pp. 89–95, 2014.

[13] T. Clercq, A. Sianov, and G. Crevecoeur, "A soft barometric tactile sensor to simultaneously localize contact and estimate normal force with validation to detect slip in a robotic gripper," *Ieee Robotics and Automation Letters*, vol. 7, pp. 11 767–11 774, 2022.

[14] T. Thurner, T. Kammerhofer, B. Reiterer, and M. Hofbaur, "Tactile sensor solution with MEMS pressure sensors in industrial robotics," *e & i Elektrotechnik und Informationstechnik*, vol. 140, no. 6, pp. 541–550, Oct. 2023.

[15] T. Kammerhofer, D. Ninevski, M. Danner, and T. Thurner, "Real-time slip detection for robotic tactile sensors via discontinuity detection," in *2026 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 2026, **to appear**.

[16] A. Grover, P. Nadeau, C. Grebe, and J. Kelly, "Learning to detect slip with barometric tactile sensors and a temporal convolutional neural network," in *2022 International Conference on Robotics and Automation (ICRA)*, 2022, pp. 570–576.

[17] *DPS368 - Digital XENSIV™ Barometric Pressure Sensor*, Infineon Technologies AG, 2019, rev. 1.1.

[18] *ICM-42670-P - High Performance 6-Axis MotionTracking™ IMU*, InvenSense, a TDK Group Company, 2021, rev. 1.0.

[19] *Robotiq 3-Finger Adaptive Robot Gripper Instruction Manual*, Robotiq, 2018, rev. 271118.

[20] *PSoC™ 62 MCU*, Infineon Technologies AG, 2024, rev. *M.

[21] T. Kammerhofer, J. Handler, and T. Thurner, "Performance evaluation and characterization of an industrial tactile sensor solution," in *IECON 2025 – 51st Annual Conference of the IEEE Industrial Electronics Society*, 2025, pp. 1–7.

[22] T. Kammerhofer, D. Ninevski, and T. Thurner, "Advanced tactile sensor solution with spline surface interpolation for robotics," in *2025 IEEE International Instrumentation and Measurement Technology Conference (I2MTC)*, 2025.

[23] D. Ninevski and P. O'Leary, "Detection of derivative discontinuities in observational data," in *Advances in Intelligent Data Analysis XVIII*, M. R. Berthold, A. Feelders, and G. Krempl, Eds. Cham: Springer International Publishing, 2020, pp. 366–378.

[24] D. Ninevski and P. O'Leary, "A convolutional method for the detection of derivative discontinuities," in *2020 21th International Carpathian Control Conference (ICCC)*, 2020, pp. 1–5.

Peak Force Evaluation for an Active Contact Flange

Bernhard Rameder
Institute of Robotics
Johannes Kepler University Linz
4040 Linz
bernhard.rameder@jku.at

Hubert Gattringer
Institute of Robotics
Johannes Kepler University Linz
4040 Linz
hubert.gattringer@jku.at

Andreas Müller
Institute of Robotics
Johannes Kepler University Linz
4040 Linz
a.mueller@jku.at

Ronald Naderer
FerRobotics Compliant Robot Technology GmbH
Johannes Kepler University Linz
4040 Linz
ronald.naderer@ferrobotics.at

Abstract

When employing robots for tasks such as polishing or surface grinding, vibration and peak interaction forces exerted at the robot or the part must be limited. Therefore, such tasks are performed with the help of Active Contact Flanges (ACF). These devices are force controlled and enable fast processing speeds. High contact velocities of the tool result in significant impact forces during interaction with the environment. Consequently, a critical control aspect is the reliable estimation of these forces, which is the focus of this paper. A dynamical model is developed, resulting in a linear time invariant equation of motion, which is solved analytically. Hereby, a homogeneous and a particular solution is derived. The maximum contact force is then determined based on an optimization. Experimental results demonstrate high consistency between measured and calculated contact forces.

1 Introduction

Grinding, polishing and deburring are important tasks in robotics and a key enabler for advanced manufacturing in high wage countries. Due to economic reasons it is important to perform these processes with high speeds. Especially, the contact phase between object and polishing tool on a robot is crucial. Therefore, an Active Contact Flange (ACF) is used to achieve these high speeds, while at the same time impact forces are minimized. These ACFs are mechatronic devices utilizing a double-acting pneumatic actuator, where real-time regulation of air inflows and outflows via force compliance control maintains a constant, predefined force after the impact phase, independent of the stroke position. Comparisons of contact forces between force controlled robots and those with Active Contact Flanges are addressed in e.g. [2, 6]. The setup with the ACF outperforms the force controlled robots in quality, speed and accuracy. An overview about robotic end-effectors for manufacturing, where various methods are compared, can be found in [5]. Contact forces between robots and the environment are also important for human robot collaboration. Maximum forces for safe collaborations are defined in [3]. A machine learning approach for such safe impact situations in human-robot interactions can be found in e.g. [4]. This work provides an analytical formula to determine the maximum contact force of an ACF during the impact phase and can later be used for training neural networks addressing such collaboration tasks.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

2 Dynamical Model

The considered setup of the system under consideration is shown in Fig. 1. The left side shows a picture of the floor mounted ACF, without the robotic system visible. To determine the real contact forces, a force sensor is mounted on an external robot that contacts the ACF with different speeds. The center sketch illustrates the setup for the dynamical modeling in the absence of robot contact. The analytic model of the ACF is a combination of a spring with spring constant k_F and a damper with damping constant d_F resembling a force controlled pneumatic device. It actuates the mass m_F. An additional serial spring (k_C) is used for the contact model. The robot with the mounted force sensor moves in vertical z direction within the base frame $\mathcal{F}_I$ of the robotic system with position r_E and velocity v_E and gets in contact with the ACF at fixed position r_0, which is the reference position for following considerations where the ACF is at its maximum stroke. It has to be mentioned, that only the z component with respect to base frame $\mathcal{F}_I$ of the end-effector motion $(r_{E,z}, v_{E,z}, r_{0,z})$, but represented in modeling frame $\mathcal{F}_M$, is of interest for spring compression. The right view in Fig. 1 depicts a schematic of the robot in contact with the already partially compressed ACF, which leads to a deformation $s = x(t) - (r_{E,z}(t) - r_{0,z})$ of the modeled contact spring depending on the position of the grinding pad $(r_{E,z}(t) - r_{0,z})$ and position $x(t)$ of mass m_F in modeling frame $\mathcal{F}_M$.

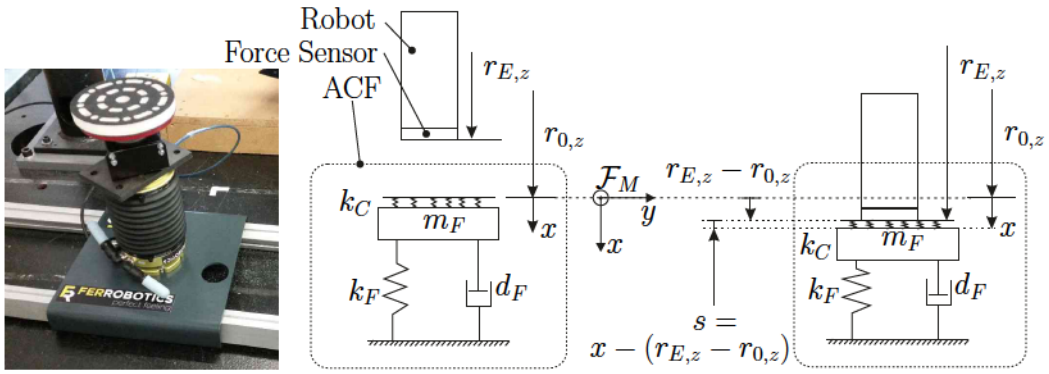


Figure 1: (l) Active Contact Flange with grinding pad in test environment in absence of robot, (m) Schematic of setup before contact, (r) Schematic of setup in contact

2.1 Modeling

An evaluation of different contact models (e.g. different parallel/serial springs) yields to the model in Fig. 1. The active contact flange with moving mass m_F consists of a spring (k_F)/damper(d_F) combination in parallel, enhanced by a contact stiffness k_C. The equation of motion (EOM) for this simplified model with x as degree of freedom, defined from the uncompressed ACF position to the position of mass m_F, and neglected gravitation is

$$m_F\ddot{x} + d_F\dot{x} + k_F x = F_C, \quad x(0) = \dot{x}(0) = 0. \tag{1}$$

The contact force F_C is zero as long as no contact appears. In case of contact, this force is defined as $F_C = -k_C(s + s_0)$, with a penetration depth s and a pretension of s_0. Since the contact is established between an external robot at end-effector position $r_{E,z}(t)$ and an ACF, which are getting in contact at position $r_{0,z}$, the expression can be further evaluated to

$$F_C = -k_C(x(t) \underbrace{-r_{E,z}(t) + r_{0,z}}_{-v_{E,z}t} - \frac{F_P}{k_C}) \tag{2}$$

with penetration depth $s = x(t) - (r_{E,z}(t) - r_{0,z})$ as introduced above and pretension $s_0 = -F_P/k_C$ of the contact stiffness. Velocity $v_{E,z}$ represents the contact velocity between the ACF and the robot, which is assumed to be constant during the contact phase and further stated with v_E for simplicity. Time t starts at the moment of contact, where $r_{E,z}$ equals $r_{0,z}$. The constant pretension force F_P is applied to the ACF to hold it on its maximum position. Using the contact force Eq. 2 in the EOM Eq. 1 results in

$$m_F\ddot{x} + d_F\dot{x} + (k_F + k_C)x = k_C v_E t + F_P, \tag{3}$$

which is a linear time invariant differential equation that can be solved analytically.

2.2 Analytic Solution

The solution for the differential equation Eq. 3 can be split in a homogeneous part and a particular part ($x(t) = x_H(t) + x_P(t)$). For the homogeneous solution x_H, the differential equation

$$\ddot{x} + \underbrace{\frac{d_F}{m_F}}_{2\delta} \dot{x} + \underbrace{\frac{k_F + k_C}{m_F}}_{\omega^2} x = 0 \tag{4}$$

is considered, where the solution can be expressed by

$$x_H = e^{-\delta t}(a \cos \nu t + b \sin \nu t), \tag{5}$$

see e.g. [1] for details. Herein, δ is the damping coefficient and $\nu = \sqrt{\omega^2 - \delta^2}$ the frequency, while constants a and b have to be evaluated based on the initial conditions of the overall solution. ω is the frequency of the undamped system.
An Ansatz for the particular solution of Eq. 3 is chosen as

$$x_P = c_0 + c_1 t, \tag{6}$$

with constants c_0 and c_1. Using this Ansatz in the EOM and evaluating the constants (initial conditions $x(0) = 0$, $\dot{x}(0) = 0$) yields the overall solution $x = x_H + x_P$ as

$$x(t) = e^{-\delta t}(a \cos \nu t + b \sin \nu t) + \frac{F_P}{k_F + k_C} + \frac{k_C v_E}{k_F + k_C} t - \frac{d_F k_C}{(k_F + k_C)^2} v_E, \tag{7}$$

where a and b are constants. Note that this solution is only valid a short time after getting in contact. However, in this time interval, the maximum contact force occurs, which makes the solution usable for the evaluation.

2.3 Maximum Contact Force

With the analytic solution Eq. 7 on hand, the contact force can be evaluated as

$$F_C = k_C(v_E t - x) + F_P. \tag{8}$$

Seeking for the maximum value leads to the optimization problem

$$\frac{dF_C}{dt} = 0, \tag{9}$$

that has a general solution of the form

$$A \cos \nu t + B \sin \nu t + C = 0, \tag{10}$$

where A, B and C are constants. This expression can be solved for the time $t = t_{max}$ of maximum contact force, which finally yields to the force

$$F_{C,max} = k_C(v_E t_{max} - x(t_{max})) + F_P. \tag{11}$$

3 Experimental Evaluation

The parameters for a test setup including $m_F, d_F, k_C, \ldots$ are identified based on force responses for various test cases and are therefore available for the evaluation of the maximum contact force in Eq. 11. An exemplary result for a desired force $F_d = 100$ N and different contact speeds v_E is shown in Tab. 1. The desired force is controlled by the ACF due to the pneumatic actuation. This actuation is much slower than the force response and therefore does not enter the contact force calculation. It can be seen, that the measured and calculated maximum contact forces coincide very well with a maximum relative error of 5 %. Similar results are achieved for different masses, pads and forces F_P. Note, that Tab. 1 additionally includes a value for the maximum contact velocity, in case the maximum applicable force is given by the considered test object that has to be e.g. polished. The maximum force for fragile objects is hereby much lower than for stiff ones, therefore the contact velocity can be lowered to a tolerable value. Figure 2 shows the measured contact forces for the different contact velocities v_E of the test case from Tab. 1.

v_E in m/s	$F_{C,max}$ Model in N	F_C Measurement in N	relative Error in %
0.05	108	110	1.2
0.1	130	137	5
0.2	195	198	1.5
0.3	271	282	4
0.35	310	320	3

Table 1: Maximum contact forces for $F_d = 100$ N.

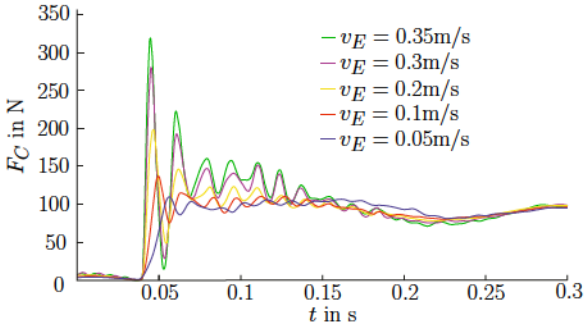


Figure 2: Measured contact forces for $F_d = 100$ N.

4 Conclusion

An analytical solution for the evaluation of maximum contact forces between an object and an Active Contact Flange is presented. The dynamical solution yields to highly accurate results for the contact forces at given contact velocities. Conversely, it can be used to determine the maximum contact velocity resulting from a given maximum contact force. Future work will use the test data to train a neural network that directly evaluates the maximum contact forces.

Acknowledgments and Disclosure of Funding

This work has been supported by the "LCM – K2 Center for Symbiotic Mechatronics" within the framework of the Austrian COMET-K2 program.

References

[1] Hans Dresig. *Schwingungen mechanischer Antriebssysteme: Modellbildung, Berechnung, Analyse, Synthese*. Springer Verlag, 2001.

[2] Stefan Gadringer, Hubert Gattringer, and Andreas Mueller. Assessment of force control for surface finishing – an experimental comparison between universal robots ur10e and ferrobotics active contact flange. *Mechanical Sciences*, 13(1):361–370, 2022.

[3] ISO/TS. Technical specification: Robots and robotic devices: Collaborative robots, standard no. iso/ts 15066:2016. *The International Organization for Standardization, The International Organization for Standardization Technical Specification,*, 2016.

[4] Nemanja Kovinčić, Hubert Gattringer, Andreas Müller, and Mathias Brandstötter. Physics-guided machine learning approach to safe quasi-static impact situations in human–robot collaboration. *Journal of Computational and Nonlinear Dynamics*, 19(7):071011, 2024.

[5] Abdelkhalick Mohammad, Erhui Sun, Junfu Zhou, Guilin Yang, Guolong Zhang, Jokin Munoa, and Asier Barrios. Robotic end-effectors for manufacturing: Recent developments and future research challenges. *International Journal of Machine Tools and Manufacture*, 2026.

[6] Alexander Winkler and Jozef Suchy. Force controlled contour following on unknown objects with an industrial robot. In *IEEE International Symposium on Robotic and Sensors Environments (ROSE)*, page 208–213, 2013.

Multi-Modal Garment Sorting and Classification Combining Tactile and Visual Sensing

Serkan Ergun *
Institute of Smart Systems Technologies
University of Klagenfurt
Klagenfurt am Wörthersee, A 9020
serkan.ergun@aau.at

Tobias Mitterer
Institute of Smart Systems Technologies
University of Klagenfurt
Klagenfurt am Wörthersee, A 9020
tobias.mitterer@aau.at

Hubert Zangl†
Institute of Smart Systems Technologies
University of Klagenfurt
Klagenfurt am Wörthersee, A 9020
hubert.zangl@aau.at

Abstract

Automated garment handling in textile recycling remains challenging due to the deformability of textiles, their high shape variability, frequent self-occlusion, and the presence of foreign objects in cluttered heaps. This paper presents a Multi-Modal robotic sorting system that combines semantic visual perception with tactile grasp monitoring. The proposed approach integrates Visual Language Model (VLM) based garment classification, Convolutional Neural Network (CNN) based grasp prediction using RGB-D images, and capacitive tactile fingertips mounted on a parallel gripper to detect grasp success, object loss, and approximate weight during manipulation. The estimated weight serves as a plausibility measure for the visually predicted garment class and as a coarse indicator of garment size. To support safe execution, a Digital Twin implemented in MoveIt2 is used for motion planning and collision avoidance in a synchronized real and virtual environment. A classification accuracy of up to 87.89 % across six classes was achieved in an experimental robotic sorting scenario including 219 items. Furthermore, the tactile finger sensor is evaluated under wet conditions and in contact with wet textiles to assess robustness, showing reliable sensing behavior even in these challenging scenarios. Overall, the results demonstrate the potential of combining semantic vision and robust tactile sensing for dependable textile sorting in recycling applications.

1 Introduction

Robotic manipulation of deformable objects remains one of the central challenges in automation [1, 2]. Among deformable materials, garments are particularly difficult to handle due to their high variability in shape, frequent self-occlusion, and non-rigid dynamics, especially when presented as unordered heaps. These challenges are further amplified in textile recycling scenarios, where garments may be entangled and mixed with foreign objects such as plastic packaging or metallic accessories. At the same time, upcoming regulations such as the European Union's Digital Product Passport (DPP) for textiles aim to improve material traceability by 2027 [3, 4]. However, legacy garments without

*S.E. and T.M. contributed equally

†H.Z. is also affiliated with the Ubiquitous Sensing Lab, University of Klagenfurt, 9020 Klagenfurt, Austria

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

digital metadata will remain present in recycling streams, requiring perception-driven identification and manipulation.

Recent advances in multi-modal neural networks have introduced VLMs, which enable semantic queries on visual data by combining vision and language representations [5]. Such models allow flexible interpretation of garment attributes and categories from images. Prior work demonstrated the feasibility of integrating VLMs with CNNs for garment classification in robotic sorting scenarios [6]. Nevertheless, these approaches primarily focus on visual perception and do not sufficiently address the physical interaction challenges associated with grasping deformable textiles in cluttered environments.

Reliable manipulation of garments requires robust grasping strategies that account for the compliance and variability of textile materials. In particular, tactile sensing during grasping plays a crucial role in detecting slip events, monitoring grasp stability, and preventing object loss. Measuring normal and shear forces during manipulation enables the robot to adapt its grasp and maintain stable contact with deformable objects.

In this work, we present a robotic textile manipulation system that combines semantic perception with tactile grasping feedback. The system integrates VLMs and CNNs for garment classification while emphasizing robust physical interaction with textiles. The main contribution of this work is that, during grasping, force measurements from the gripper are used to detect shear forces that indicate object loss and to estimate the approximate weight of the grasped textile, thereby providing an additional plausibility check for the visually predicted garment class and enabling coarse inference of the object size. The system further incorporates a Digital Twin using MoveIt2 for motion planning, where reconstructed 3D representations of manipulated textiles are integrated into the planning environment. Particular emphasis is placed on the robustness of the tactile sensing frontend, which is designed to maintain reliable operation even when interacting with challenging materials such as wet textiles.

2 Related Work

Recent advances in robot grasping and classification of textiles show, that general VLMs models perform better on untrained objects like textiles, than specialized CNN networks. An example of a CNN is given by [7] and an example of a VLM is [8]. Multimodal Large Language Model (MLLM) models, like [9] by Alibaba Cloud are computation-heavy, with bigger versions requiring VRAM greater than 100GiB but also support better visual reasoning capabilities.

In robotics, once an object has been classified, suitable grasping positions need to be determined. Several approaches integrate grasp detection with additional perception tasks [10, 11]. In contrast, the previously discussed VLMs do not inherently provide segmentation or grasp position detection capabilities. Accurate grasp detection is particularly important for objects with complex geometries or deformable materials, such as clothing. In such scenarios, task-specific models like CNNs often achieve superior performance due to their specialized training for grasp prediction. Nevertheless, recent work has explored the use of VLMs that incorporate semantic input to guide grasp planning. For instance, models can interpret commands such as “the tip of the sock” and return corresponding grasp positions [12, 13]. VLMs also differ in that they are either focused on performance, or on usability on edge devices.

After identifying grasp positions, the next step is to grasp an object. During grasping it is important to be able to detect successful grasps by measuring, e.g., normal and shear forces applied on a grasped object [14, 15].

Our method combines VLMs with a CNN-based perception pipeline for semantic classification in a robotic textile-sorting scenario involving densely piled garments. In addition to visual recognition, the system emphasizes robust grasping of deformable objects. Force measurements from the gripper are used to monitor shear forces during manipulation, enabling detection of object loss during grasping. The measured forces also allow for weight estimation of the textile, which serves as a plausibility check for the predicted class and provides a rough indication of the object size. Furthermore, the gripper’s sensor frontend is designed to be robust against direct contact with challenging materials such as wet textiles, ensuring reliable sensing performance.

3 Experimental Setup

The experimental setup and its corresponding Digital Twin (RViz) is illustrated in Figure 1 below. In

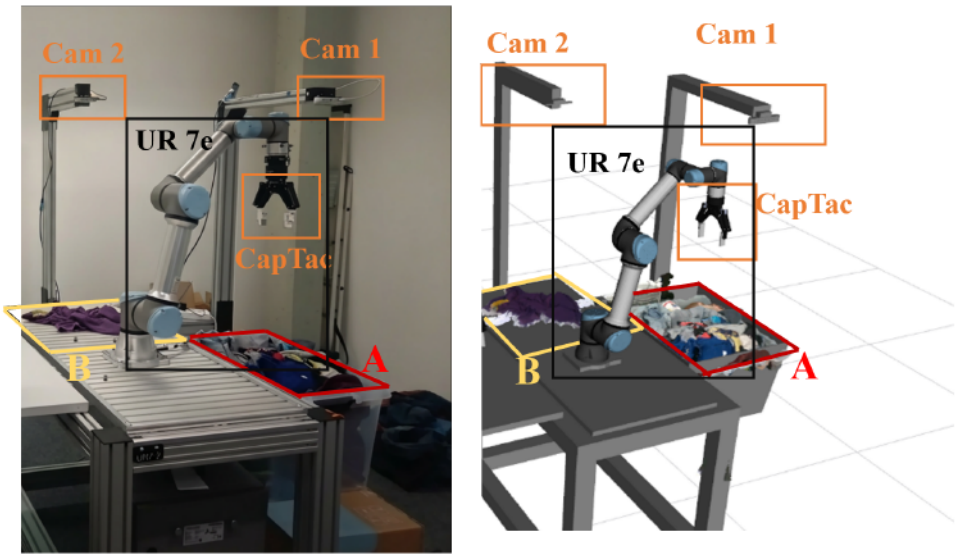


Figure 1: An overview of the experimental setup (left) and the corresponding Digital Twin in RViz (right).

this scenario, a UR7e robot is tasked with picking up unsorted, cluttered but clean garments from a basket (zone *A* in Figure 1) and to place them on a separate table for garment class inspection (zone *B* in Figure 1). Afterwards, the garment is picked up and distributed for further processing (outside the scope of this paper). Grasp locations for both zones are determined by using an adapted version of the grasp prediction algorithm developed by Ainetter et al. [10] and extensively tested in [6, 11, 16]. Two Intel Realsense depth cameras from the D400 model family (*Cam 1* and *Cam 2* in Figure 1) provide the necessary depth and color images. The RGB stream of *Cam 2* also provides the frames needed for garment type classification. The robot is equipped with a Robotiq 2F-140 gripper. The fingertips are replaced with CapTac [14] capacitive tactile sensors (a detailed view is shown in Figure 3a). These fingertips are used for two purposes: Determining if a garment was grasped from the basket (zone *A*) or dropped during manipulation and to measure its weight after garment type classification to provide further information on the garment. The detection threshold of CapTac is stated to be 20 g [13] corresponding to a weight force of 0.4 N for a two-fingered gripper. The necessary computing power for running the grasp prediction algorithm, robot control script as well as the Digital Twin is provided by a home grade PC equipped with an 11th Gen Intel® Core™ i7-11700KF @ 3.60GHz × 16 CPU with 64 GB of DDR4 RAM paired with a Graphics Processing Unit (GPU) Nvidia GeForce RTX 3060 with 12GiB VRAM. The VLMs are run on an external Nvidia H200X-141C cloud GPU with 144GiB VRAM.

The communication between all hardware modules is handled by ROS 2 Jazzy. A schematic overview showing the connection between all hardware and software components is shown in Figure 2. The real and simulated environment share the same coordinate frames allowing motion planning and static obstacle avoidance to be conducted on the Digital Twin via MoveIt2 before executing movements on the real robot. Furthermore, a combined RGB/depth stream of *Cam 1* is projected into the Digital Twin as a pointcloud, allowing remote observation. A segmented RGB/depth image of a garment in zone *B* can be projected as well, if needed.

The CapTac sensors are connected to the ROS 2 system via the Arrowhead Eclipse framework [17], as capacitive sensors, which provide normal and shear forces measured on multiple channels to the system. On the ROS 2 side, an Arrowhead Consumer is running, which takes the measured forces and publishes them into the ROS 2 system. The Arrowhead system is used for safe and secure transmission of sensor data.

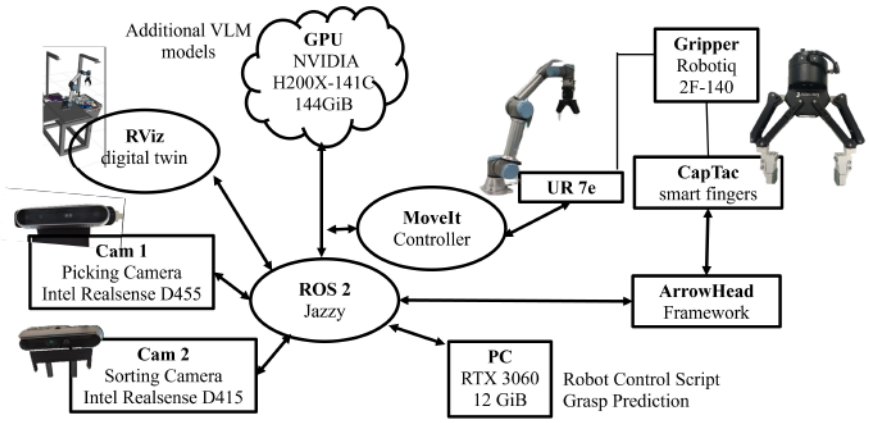
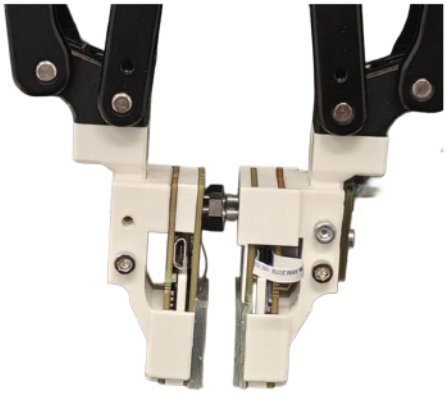
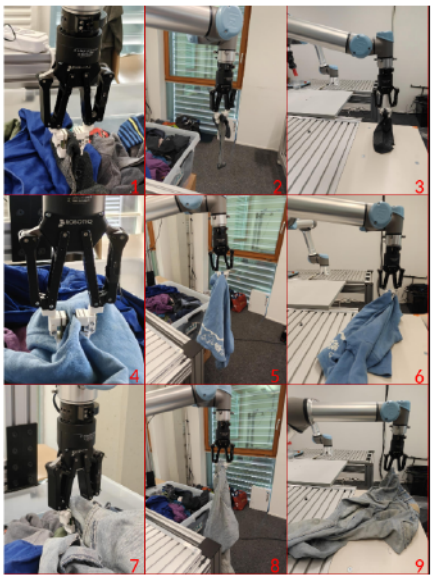


Figure 2: Schematic of the proposed framework, consisting of a UR7e robot equipped with a Robotiq 2F-140 gripper - and CapTac [14] fingertips - a desktop PC equipped with a Nvidia RTX 3060 budget consumer graphics card, one Nvidia H200X professional graphics card and two Intel Realsense cameras for grasp detection and object classification.



(a) Closed gripper, the gap in between the two gripper fingers is used to avoid touch between the sensors, when no object is grasped, so as to better detect an empty grasp. The distance can be manually adapted with the screws on top.



(b) Garment manipulation process, showcased on a sock 1-3, a shirt 4-6 and a trouser 7-9. The first image displays a textile being grasped from the box, the second image displays the textile in the air and the third image displays the textile being placed on the inspection table.

Figure 3: Detailed view of the gripper (a); Garment manipulation process (b)

4 Experimental Procedure

Upon starting the experimental procedure, a set of random clean garments is thrown randomly in a transparent basket alongside random foreign objects, such as bottles, cans and 3D printed objects from the EGAD training set [18]. The basket does not need to be perfectly aligned for every experimental run. A bounding box is automatically set around the edges of the basket to avoid grasping objects that are outside the basket.

After initial manual setup, the robot control script is started, which operates the robot as a finite-state machine (FSM). Figure 4 shows the flowchart for the states and transition conditions. In the *Init* state,

the robot is moved to a pose outside the field-of-view of both cameras and the CapTac sensors are initialized by recording baseline measurements. Next, the control script moves to *Find Garment*, calling the grasp prediction algorithm with a RGB and depth frame of *Cam 1*. If a potential grasp candidate is found within 5 tries, the program will switch to *Pick up Garment* and uses MoveIt2 within the Digital Twin to plan and validate a trajectory, otherwise the program will shut down. After the grasp pose is reached and the gripper closes, CapTac sensor readings are examined to check for an object between the fingers. In case of success, the sensor readings are monitored until the robot reaches zone *B* for garment classification. While an object is being lifted, a slight shaking motion is applied to the final three joints of the robot to shake off potential by-catch and avoid a drop in an undesired area. The garment is then pulled over the edge of the inspection table to enforce spreading out as good as possible. In case of an unsuccessful grasp or object loss, the robot moves outside the field-of-view of *Cam 1* and the program switches to *Find Garment*. Impressions of the garment handling procedure are shown in Figure 3 for exemplary objects of classes: sock, shirt and trousers.

Once the robot has dropped off the item and moved out of the field-of-view of *Cam 2*, the program switches to *Inspection*. The VLM running on the external GPU receives a RGB image from *Cam 2* and returns the predicted object class; the options being: trousers, shirt, underwear, sock, other (foreign object, or garment outside the aforementioned classes) and empty. We run two models from the Qwen3 family by Alibaba Cloud, which delivered promising results in a recently published benchmark [19], utilizing the locally run Python API of Ollama [20]. A minimal Python code snippet showcasing the usage of Ollama with Python3 is shown in listing 1. After classification, the grasp prediction algorithm is called with a RGB and depth image from *Cam 2*. The object is then lifted and an estimate of the weight is calculated using the CapTac sensors. In the event of multiple objects being grasped from the box, the VLM-powered classification is run again to ensure an empty inspection table. The combined information is then stored and the object is then placed aside for further processing. This information can be used to provide additional information on the object and conducting sanity checks. Some examples are: If an object of class shirt or trousers has a very low weight, it may be infant clothing. If an object of class sock is measured to have a rather high weight, the object class may have been predicted incorrectly, and the garment could be put aside for further manual inspection. The aforementioned garment segmentation scheme, using the same procedure as described in [19] could be used to create further training data for VLM based characterization of garments. Finally, the program will switch to *Reset* and prepare for picking up the next object from the basket.

Listing 1: Minimal Python code example for running a Vision Language Model with Ollama

```
import ollama

response = ollama.chat(
    model='model-name',
    messages=[{"role": "system",
        "content": "You are an expert garment classification
            device."},
        {'role': 'user', 'content': '"Do you spot a clothing item on
            the table? "
    "If yes: Classify them in the classes: "
    "shirt, sock, underwear or trousers. "
    "Do you see something else instead? respond with other. "
    "Is the table empty? respond with empty. "
    "Your response is a single word - either "
    "shirt, sock, underwear, trousers, other or empty"', 'images':
        [fullPathToImages]
            }]
)
```

In parallel, the robustness of CapTac against wet garments was investigated in manual experiments. Sensor readings were recorded for dry and wet garments and empty measurements with dry and manually wet sensor pads.

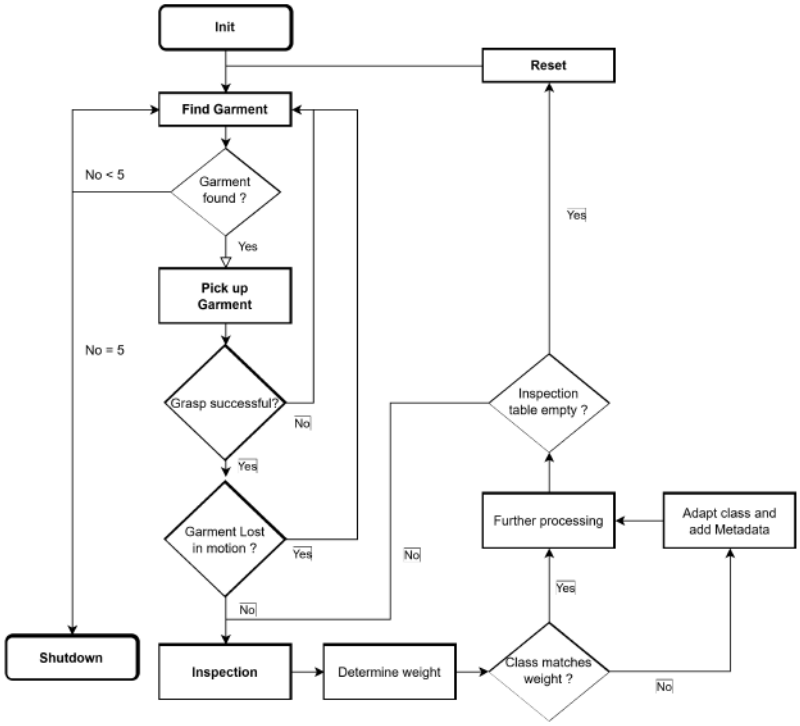


Figure 4: Flowchart of the textile grasping process.

5 Results

5.1 Multi-Modal sensing experiment

During the automated garment sorting scenario, a total of 219 items (garments and foreign objects) were grasped from the basket, classified on the inspection table and their weight estimated by CapTac. Ground truth data for these experiments were manually collected by human classification and manual weighing using a kitchen grade precision scale. To address robustness of the VLMs, images from objects on the inspection table were uncropped, and some distracting items were placed in the scenery. Two exemplary samples are shown in Figure 5. At two random instances during the experiments, objects were manually removed from the inspection table to check for robustness against hallucinations. Furthermore, two more empty samples were taken after the experiment was conducted. To avoid a misclassification, when two or more garments are present on the inspection table, the grasp prediction algorithm provides the location of the garment on the image. The combination of CapTac with a well established grasp prediction algorithm ensured that no unintentional empty scenes were recorded. The number of samples per object class is shown in Table 1 alongside the accuracy of both investigated models. The confusion matrices for both VLMs are shown in Figure 6. Furthermore, the computation time for each model was recorded and stored alongside the model predictions. The average processing times $\bar{t}$ alongside the 10^{th} percentile P_{10} and 90^{th} percentile P_{90} are shown in the bottom half of Table 1.

A prediction was considered correct if the model's response exactly matched the requested class (ignoring lower/uppercase lettering). Returning a different, but semantically correct class name was considered as wrong. Also, if the response also contained more words than requested, the response was treated as incorrect. While the larger Qwen model provides an accuracy of >97 % for shirts and full accuracy for socks, it lacks in detecting trousers and empty scenes, where it would typically return one of the distracting items nearby. As garments cannot be perfectly presented on a table using only a single robotic arm in reasonable time, the overall accuracy could be improved with adding multi-robot grasping.

Estimating the weight of the garments using CapTac faced a few difficulties. Garments with a very low weight (e.g. socks) do not surpass the reliable detection limit for shear force (0.2 N corresponding

(a) Class: "Shirt", Weight:264 g; " qwen3-vl:235b: "Shirt", qwen3-vl:8b: "Shirt", Estimated Weight: 225 g;

(b) Class: "Trousers", Weight: 257 g; qwen3-vl:235b: "Trousers", qwen3-vl:8b: "Trousers", Estimated Weight: 240 g;

Figure 5: Two representative images of garments from zone B. To assess robustness in cluttered environments, several smaller garments were scattered on the ground and additional distracting objects were introduced. Each model's output is shown together with the corresponding ground truth.

Table 1: Performance Benchmark for both investigated Qwen3 models and computation time metrics on a Nvidia H200 in s: average $\bar{t}$, 10^{th} percentile P_{10} and 90^{th} percentile P_{90}

	Overall	Shirt	Sock	Trousers	Underwear	Other	Empty
Image Count	223	38	64	43	12	65	4
Models				Accuracy			
qwen3-vl:235b	87.89 %	97.37 %	100.00 %	60.47 %	83.33 %	93.85 %	25.00 %
qwen3-vl:8b	83.86 %	86.84 %	93.75 %	55.81 %	66.67 %	95.38 %	50.00 %

	qwen3-vl:8b	qwen3-vl:235b
$\bar{t}$	1.595	2.444
P_{10}	0.993	1.739
P_{90}	2.550	3.072

to 20 g per finger - yielding a minimal garment weight of 40 g, assuming equal distribution of shear force), and therefore only the normal force information can be used to detect an object between the fingers. Larger garments, such as adult trousers or jackets, tend to entangle above the sensor pads, and thus produce less shear force. In any case, the weight of the garment will always be underestimated. CapTac require a manual one-time calibration using a tray with known weight that has full contact with both sensor pads and additional precision weights to determine a baseline. An exemplary calibration curve is shown in [14].

Thus, the authors rather propose a qualitative assessment of the shear force in weight classes, as listed in Table 2. Combining garment type and weight class allows to easier differentiate between adult and infant clothing. Garment classes with generally lower weight will then trigger an additional manual sanity check to verify if the VLM prediction is accurate.

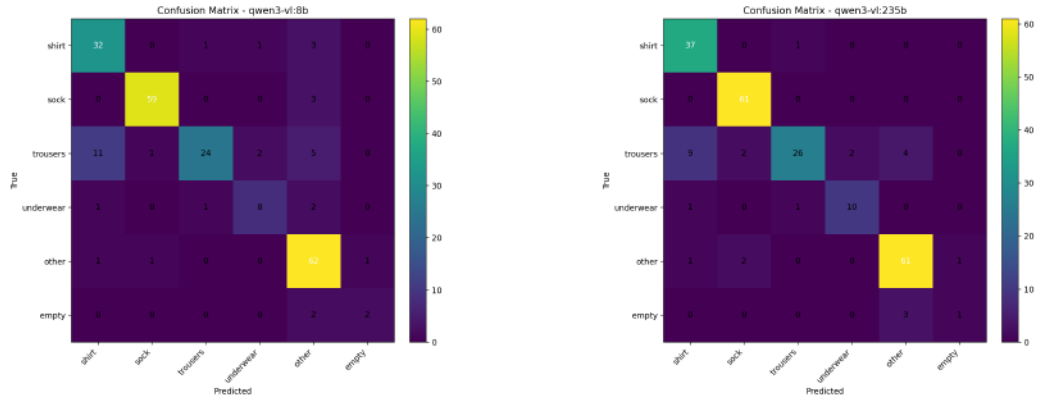
5.2 Robustness Against Liquid Contamination

In an additional set of experiments, the robustness of CapTac against liquid contamination on the sensor pads and wet garments was investigated. The silicone structure of CapTac (Ecoflex) is, by

Table 2: Qualitative assessment of garments using weight classes.

Weight	Shirt	Sock	Trousers	Underwear	Other
low($<$ 45 g)	toddler/inf.	single/pair	toddler/inf.	ok	ok
mid (45 g $<$150 g)	inf	pair	inf	ok	ok
high (150 g $<$400 g)	adult	s/c	adult	heavy underwear	ok
heavy ($\geq$ 400 g)	s/c	s/c	adult	s/c	ok

Notes: inf.: infants, s/c: sanity check recommended



(a) qwen3-vl_8b (b) qwen3-vl_235b

Figure 6: Confusion matrices for Qwen3 models: 8b parameters (left) and 235b parameters (right)

nature, highly hydrophobic, however potential effects on the capacitive sensing needed to be verified. During multiple measuring cycles, the sensor readings were recorded while the sensor pad was in a dry and wet state (by manually wetting the pad with a paper towel). No significant change in noise or drift was recorded during these experiments.

In its current state, the sensor (with its electronics) is not waterproof, and liquid contamination can only be mitigated on the sensor pads and the front face of the sensors.

6 Discussion, Conclusion and Outlook

This paper presented a Multi-Modal robotic sorting system that combines VLM based garment classification with a CNN based grasp prediction using RGB-D images, and capacitive tactile fingertips mounted on a parallel gripper to detect grasp success, object loss, and approximate weight during manipulation. The estimated weight serves as a plausibility measure for the visually predicted garment class and as a coarse indicator of garment size. A Digital Twin implemented in RViz uses MoveIt2 for motion planning and collision avoidance in a synchronized real and virtual environment.

A classification accuracy of up to 87.89 % across six classes was achieved in an experimental robotic sorting scenario including 219 items. Garment manipulation is handled by a single robotic arm, which does not allow optimal garment placement on the inspection table (zone B), practically reducing the accuracy for garment classes of larger sizes. Socks, due to their smaller size and distinctive shape, gave the highest accuracies. Furthermore, the tactile finger sensor is evaluated under wet conditions and in contact with wet textiles to assess robustness, showing reliable sensing behavior even in these challenging scenarios. Overall, the results demonstrate the potential of combining semantic vision and robust tactile sensing for dependable textile sorting in recycling applications.

The present state of the proposed experimental setup offers multiple options for future improvement. Adding a second manipulator for simultaneous two-arm manipulation to spread out garments can improve the classification accuracy and opens the door for visual fault inspection. A weighted combination with other VLMs might increase the accuracy as well and serve as a second source for classification estimates. The current state of the weight measurement can be improved by optimizing the gripper shape, to avoid tangling of garments.

Acknowledgments and Disclosure of Funding

This work has received funding from the "Austrian Research Promotion Agency" (FFG) within the AdapTex project under grant number 899044 and by the European Commission, through the European H2020 research and innovation programme, KDT Joint Undertaking, and National Funding Authorities from 10 involved countries – including Hungary – under the research project Arrowhead fPVN with Grant Agreement no. 101111977.

References

[1] R. Herguedas, G. López-Nicolás, R. Aragüés, and C. Sagüés, "Survey on multi-robot manipulation of deformable objects," in *2019 24th IEEE International Conference on Emerging Technologies and Factory Automation (ETFA)*, 2019, pp. 977–984.

[2] A. Billard and D. Kragic, "Trends and challenges in robot manipulation," *Science*, vol. 364, no. 6446, p. eaat8414, 2019. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.aat8414

[3] European Parliamentary Research Service, "Digital product passport for the textile sector," 2024. [Online]. Available: {https://www.europarl.europa.eu/RegData/etudes/STUD/2024/757808/EPRS_STU(2024)757808_EN.pdf}

[4] Directorate-General for Environment, "COMMUNICATION FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT, THE COUNCIL, THE EUROPEAN ECONOMIC AND SOCIAL COMMITTEE AND THE COMMITTEE OF THE REGIONS - EU Strategy for Sustainable and Circular Textiles," 2022. [Online]. Available: {https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52022DC0141}

[5] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 2021, pp. 8748–8763. [Online]. Available: https://proceedings.mlr.press/v139/radford21a.html

[6] S. Ergun, T. Mitterer, and H. Zangl, "Towards automated handling and sorting of garments combining visual language models and convolutional neural networks," *Proceedings of the Austrian Robotics Workshop 2025*, pp. 25–30, 2025. [Online]. Available: https://www.fh-salzburg.ac.at/fileadmin/fhs_daten/departments/information-technologies/documents/ARW2025_Proceedings_final_kl.pdf

[7] M. Suchi, T. Patten, D. Fischinger, and M. Vincze, "Easylabel: a semi-automatic pixel-wise object annotation tool for creating robotic rgb-d datasets," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 6678–6684.

[8] Minderer, Matthias and Gritsenko, Alexey and Stone, Austin and Neumann, Maxim and Weissenborn, Dirk and Dosovitskiy, Alexey and Mahendran, Aravindh and Arnab, Anurag and Dehghani, Mostafa and Shen, Zhuoran and Wang, Xiao and Zhai, Xiaohua and Kipf, Thomas and Houlsby, Neil, "Simple Open-Vocabulary Object Detection," in *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part X*. Berlin, Heidelberg: Springer-Verlag, 2022, p. 728–755. [Online]. Available: https://doi.org/10.1007/978-3-031-20080-9_42

[9] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu, "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[10] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from rgb," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 13 452–13 458.

[11] S. Ergun, T. Mitterer, S. Khan, N. Anandan, R. B. Mishra, J. Kosel, and H. Zangl, "Wireless capacitive tactile sensor arrays for sensitive/delicate robot grasping," in *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023, pp. 10 777–10 784.

[12] META, "segment anything website," https://segment-anything.com/, accessed: 2025-06-04.

[13] N. Carion, L. Gustafson, Y.-T. Hu, S. Debnath, R. Hu, D. Suris, C. Ryali, K. V. Alwala, H. Khedr, A. Huang, J. Lei, T. Ma, B. Guo, A. Kalla, M. Marks, J. Greer, M. Wang, P. Sun, R. Rädle, T. Afouras, E. Mavroudi, K. Xu, T.-H. Wu, Y. Zhou, L. Momeni, R. Hazra, S. Ding, S. Vaze, F. Porcher, F. Li, S. Li, A. Kamath, H. K. Cheng, P. Dollár, N. Ravi, K. Saenko, P. Zhang, and C. Feichtenhofer, "Sam 3: Segment anything with concepts," 2025. [Online]. Available: https://arxiv.org/abs/2511.16719

[14] S. Ergun, R. B. Mishra, N. Ananadan, T. Mitterer, V. Mattoli, and H. Zangl, "Captac: Robust capacitive sensing for distributed force mapping in parallel robotic grasping," *IEEE Robotics and Automation Letters*, vol. 11, no. 3, pp. 3668–3675, 2026.

[15] X. Zhang, T. Yang, D. Zhang, and N. F. Lepora, "Tacpalm: A soft gripper with a biomimetic optical tactile palm for stable precise grasping," *IEEE Sens. J.*, vol. 24, no. 22, pp. 38 402–38 416, 2024.

[16] S. Ergun, T. Mitterer, and H. Zangl, "A hybrid approach towards automated textile sorting," *e+i Elektrotechnik und Informationstechnik*, vol. 142, no. 6, pp. 360–370, Nov 2025. [Online]. Available: https://doi.org/10.1007/s00502-025-01340-2

[17] P. Varga, F. Blomstedt, L. L. Ferreira, J. Eliasson, M. Johansson, J. Delsing, and I. Martnez de Soria, "Making system of systems interoperable the core components of the arrowhead framework," *J. Netw. Comput. Appl.*, vol. 81, no. C, p. 85–95, Mar. 2017.

[18] D. Morrison, P. Corke, and J. Leitner, "Egad! an evolved grasping analysis dataset for diversity and reproducibility in robotic manipulation," *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4368–4375, 2020.

[19] S. Ergun, T. Mitterer, and H. Zangl, "Digital twin driven textile classification and foreign object recognition in automated sorting systems," 2026. [Online]. Available: https://arxiv.org/abs/2603.05230

[20] Ollama, "Ollama website," https://ollama.com/, accessed: 2026-02-02.

Safe and Smart Robotics

Enhanced Environmental Context Encoding for Accurate Trajectory Prediction in Intralogistics

Alexander Prutsch
Institute of Visual Computing
Graz University of Technology
Graz, Austria
alexander.prutsch@tugraz.at

Horst Possegger
CD Laboratory for Embedded Machine Learning,
Institute of Visual Computing
Graz University of Technology
possegger@tugraz.at

Abstract

Trajectory prediction is an essential component of the perception stack in autonomous mobile robots (AMRs). AMRs operate in complex environments where their movements are influenced by various environmental elements, such as racks and storage locations. Therefore, accurate and efficient trajectory prediction for intralogistics requires detailed environment modeling that goes beyond the lane-based context commonly used for road traffic. We propose a new environment context encoder that can be seamlessly integrated into state-of-the-art motion forecasting models. Our approach, tailored to the specific challenges of intralogistics, achieves highly accurate predictions using efficient baseline networks.

1 Introduction

Autonomous mobile robots (AMRs) play a major role in modern intralogistics as they are commonly used to transport cargo in complex environments like warehouses and production facilities. AMRs require accurate trajectory prediction to navigate crowded intralogistics environments efficiently and safely. Anticipating the movements of other traffic participants allows AMRs to proactively plan maneuvers, which prevents operational delays from costly deadlocks and protects warehouse workers.

While trajectory prediction is well-studied for road traffic, *e.g.,* [1, 2, 3, 4, 5, 6, 7], intralogistics presents distinct challenges. Warehouse vehicles can execute more diverse and complex maneuvers due to their wheel geometry [8]. Additionally, lane graphs, which are commonly used to model map information in road traffic, provide only weak guidance in the intralogistics domain, as vehicles follow driving lanes less strictly within warehouses and production facilities. Furthermore, environmental elements heavily influence driving behavior; *e.g.,* a vehicle entering a rack aisle has a high probability of abruptly changing speed to initiate load handling. Consequently, explicitly modeling these map elements is critical for robust trajectory prediction in intralogistics.

Due to different environment types, vehicle characteristics, and onboard resources, trajectory prediction models designed for autonomous driving cannot be directly transferred to the AMR domain. We propose a new environment encoder module to enable accurate trajectory prediction in intralogistics. Our environment context (ECTX) encoder processes information on diverse map elements like rack positions, charging stations and gates, enabling highly accurate predictions in scenarios where lanes only offer a weak prior. By directly utilizing widely available semantic warehouse maps, ECTX bypasses the need for computationally heavy LiDAR data [6, 5] and requires no application-specific fine-tuning using LiDAR data. Integrating our ECTX module into two strong, compact transformer-based baselines, Forecast-MAE [9] and EMP [10], yields improved results for trajectory prediction in complex intralogistics driving scenarios. The additional environmental information significantly improves the trajectory prediction accuracy while adding only little overhead to the networks. We demonstrate the effectiveness of our ECTX by evaluating it on a custom large-scale dataset for motion

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

prediction of different intralogistics vehicles, *e.g., reach trucks* and *order pickers*. Our approach demonstrates strong performance on long-term prediction horizons, giving AMRs sufficient time to react to the prediction and to use the output to perform smooth, proactive driving maneuvers.

2 Trajectory Prediction With Enhanced Environment Encoding

Modern trajectory prediction typically relies on separate agent and lane encoders prior to scene encoding. To capture the unique dynamics of intralogistics, we introduce a plug-and-play third module (ECTX), which integrates critical environmental context into baseline architectures, *i.e.,* [9, 10].

Map Modeling: We extract lane segments and environmental elements from standard warehouse floor plans. Lane segments are sampled and split into similar-sized, fixed-point chunks. Crucially, we extract key environmental elements that dictate driving behavior, *e.g.,* racks (indicating potential load handling), non-driveable areas (walls, machinery), and free areas (where vehicles may cut corners).

Agent and Lane Encoding: Initially, agent history and lane shapes are encoded individually using local coordinate systems. For agent motion, baselines output a feature vector for each agent using either neighborhood attention [11] (Forecast-MAE [9]) or standard self-attention (EMP [10]). Lane geometries are processed in both baselines using PointNet-like architectures [12].

Environment Encoding (ECTX): We encode points sampled from the environment polygons using a small PointNet-like network [12]. Unlike the lane encoder, which extracts line features, our environment encoder learns the shapes of polygons. It outputs an environment matrix where each row represents a distinct map area.

Scene Encoding and Trajectory Decoding: As both baselines [9, 10] use token-type independent self-attention for scene encoding, we can seamlessly concatenate the environment tokens with the agent and lane tokens to form a unified scene context. To preserve spatial and categorical relationships, we add global positional embeddings [9] and type embeddings for both vehicle and map classes. Finally, a set of trajectory hypotheses is decoded using the baselines respective architecture (an MLP for Forecast-MAE and EMP-M, or a DETR-like [13] decoder for EMP-D).

3 Experimental Setup

Dataset: We conduct our evaluations on a large-scale dataset generated using NVIDIA Omniverse™. It is recorded in virtual warehouse environments, which are designed based on real-world layouts and traffic situations. We use CAD models of different real-world intralogistics vehicles and apply custom vehicle controllers to obtain highly realistic motion patterns. The maps in our dataset contain lane topology implemented as directed graphs and detailed information on intralogistics environment elements. These include charging stations, free areas, rack locations, gates, non-driveable areas (static obstacles), and storage locations. Overall, our dataset features 267,146 total scenarios across two virtual environments: one for training (94,621) and validation (63,605), and a second environment for testing (108,920). The intentionally limited training set reflects real-world data scarcity.

Each scenario spans 11 seconds (5 s history, 6 s prediction horizon), yielding a challenging median future trajectory length of 10.45 m (max 13.60 m). The dataset includes various vehicle types, *e.g.,* reach trucks, forklifts, and order pickers, to capture the different driving dynamics of each type. Compared to autonomous driving data, our dataset includes difficult intralogistics-specific movements like load handling, on-the-spot rotations, and reversing. For each scenario, we sample all neighboring map elements and agents within a radius of 25 m as model input, which fully captures the relevant context for intralogistics driving speeds.

Implementation Details: For both baselines [9, 10], we evaluate the original architectures (latent feature dimension $D = 128$) alongside a smaller version with $D = 64$ and shallower encoders. The baseline models are designed for autonomous driving applications, where large-scale datasets are available. This model size reduction mitigates potential overfitting on the smaller datasets typical of intralogistics and facilitates deployment on embedded hardware. Models are trained on a single NVIDIA V100 GPU for 60 epochs with a batch size of 128. We do not use data augmentation and optimize using AdamW [14] with gradient clipping and weight decay. The learning rate undergoes a 10-epoch linear warm-up (1×10^{-6} to 8×10^{-5}) before a cosine decay schedule back to 1×10^{-6}. Following our baselines [9, 10], the training objective combines a regression loss, a classification loss

Table 1: Results on our intralogistics dataset grouped by baseline model and sorted in descending order by test **brier-minFDE$_4$**. Models marked with * denote reduced architecture configurations, which are also more suitable for AMR deployment. Displacement errors reported in meters.

Method	Test Set MR$_4$	minADE$_1$	minFDE$_1$	minADE$_4$	minFDE$_4$	**brier-minFDE$_4$**
EMP-M [10]	0.156	1.12	2.83	0.57	1.21	1.56
EMP-M*	0.182	1.09	2.69	0.57	1.22	1.52
EMP-M*+ECTX	0.155	1.08	**2.67**	0.55	1.17	1.46
EMP-D [10]	0.140	1.12	2.84	0.51	1.07	1.41
EMP-D*	0.129	1.11	2.76	0.51	1.06	1.35
EMP-D*+ECTX	0.129	1.15	2.83	0.50	1.03	1.33
Forecast-MAE [9]	**0.117**	1.16	2.97	0.48	**0.98**	1.34
Forecast-MAE*	0.126	**1.06**	**2.67**	0.49	1.03	1.32
Forecast-MAE*+ECTX	**0.117**	**1.06**	2.70	**0.47**	**0.98**	**1.27**

Table 2: Ablation study on the influence of different encoder module using Forecast-MAE*. The experiment marked with ✓† uses only a single encoder for environment and lane data.

Context Encoder Lanes	ECTX	Test Set MR$_4$	minADE$_4$	minFDE$_4$	**brier-minFDE$_4$**
✓†		0.323	0.84	1.72	2.03
✗	✗	0.251	0.67	1.49	1.83
✗	✓	0.188	0.64	1.38	1.70
✓	✗	0.126	0.49	1.03	1.32
✓	✓	**0.117**	**0.47**	**0.98**	**1.27**

to score the multiple trajectory hypotheses, and an auxiliary loss that predicts a single future for all non-focal agents in the scene.

4 Results and Conclusions

We present detailed evaluations on our custom intralogistics dataset by comparing three baseline models with and without our ECTX encoder. We evaluate our models using standard trajectory prediction metrics [15, 16, 17]: minADE$_K$, minFDE$_K$, brier-minFDE$_K$, and MR$_K$. To capture diverse future movements while maintaining a compact representation suitable for AMR control systems, our models output 4 trajectory hypotheses. We compute these metrics for both the most probable prediction ($K = 1$) and the full set of predictions ($K = 4$).

Evaluation on Intralogistics Dataset: Table 1 compares our baselines, Forecast-MAE [9] and EMP-M/D [10], with and without our new ECTX module on our custom intralogistics dataset. For all three models, the integration of ECTX significantly improves trajectory prediction accuracy on our test set, leading to highly accurate overall results. Furthermore, our reduced-capacity configurations (marked by *) outperform the original architectures, better matching the limited dataset size while suiting the computational constraints of embedded AMR hardware. Consistent with the results from [10] on AV2 [17], EMP-D outperforms EMP-M due to its more sophisticated decoder architecture. The intralogistics scenarios feature significantly fewer agents per scene, where the neighborhood attention-based [11] agent encoder from Forecast-MAE brings an advantage over the pure transformer version from EMP. Furthermore, comparing the $K = 1$ and $K = 4$ metrics reveals that ECTX substantially improves multiple-hypothesis generation, demonstrating that explicit environmental context is crucial for accurately modeling multi-modal behavior.

Figure 1 compares the predictions of EMP-D* without and with ECTX on two example scenarios from our test dataset. In the first scenario (top row), a vehicle navigates a *rack pre-zone*, where it can either enter a rack aisle or continue in the open area. Without the environmental context of the rack positions, the baseline EMP-D* predicts an invalid left turn that would lead to a collision with a rack. Adding the rack positions using our ECTX encoder leads to well-suited trajectory predictions. In the second scenario (bottom row), a vehicle is leaving an aisle having the option to either go left or right. Using only lane data as context, the baseline model predicts a corner-cutting maneuver that would

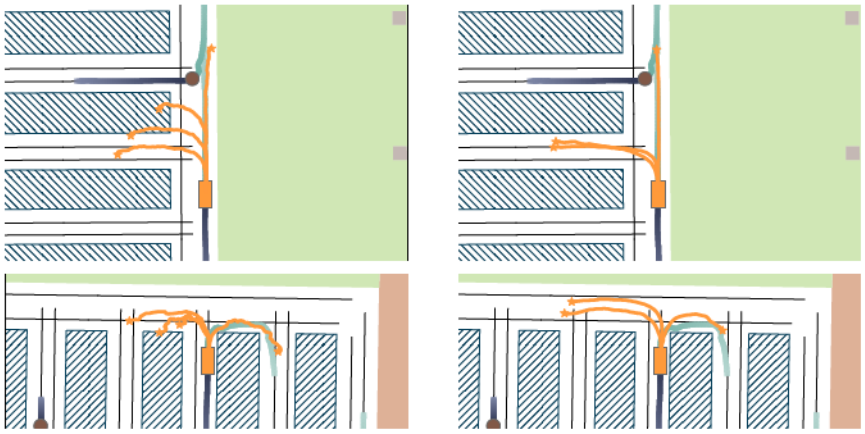


Figure 1: Both image pairs show scenarios from our custom intralogistics dataset. The left images show **predictions** from our baseline EMP-D* model and the right images from EMP-D*+ECTX. In all images also the **ground truth trajectory** is shown. Using ECTX the possible turn inside the rack aisles much better aligns with the given environment structures. For the second example using ECTX the predicted trajectories do not interfere with the **racks** as the corners are not cut.

Table 3: Latency comparison for deployment on robotic hardware, alongside measurements from a reference desktop GPU. We report inference latency for predicting all agents in the most complex test scene (8 vehicles, 110 context elements).

Method	NVIDIA Jetson		NVIDIA	Model
	Orin Nano	AGX Orin	V100	Parameters
EMP-D*+ECTX	38 ms	15 ms	28 ms	1.2M
EMP-M*+ECTX	37 ms	15 ms	22 ms	791K

collide with a static obstacle. Once again, incorporating our environment encoder resolves this issue, ensuring the predicted trajectories remain physically viable and collision-free.

Ablation Study: Table 2 details an ablation study evaluating the influence of lane and environment context using Forecast-MAE [9]. As expected, map-free prediction (agent history only) performs worst, as it cannot anticipate structural maneuvers like turns and predictions are limited to different driving motion patterns. Using only the environment context from ECTX as input, trajectory prediction accuracy significantly improves. This confirms that using the map elements provides a valuable input for trajectory prediction. As expected, the addition of lane data yields the best results overall. We also conduct an experiment where we utilize a single encoder module to encode both lane polylines and environment polygons simultaneously. Processing both lane polylines and environment polygons through a single encoder degrades performance, as the model fails to learn proper context guidance. This confirms that distinct spatial modalities require dedicated encoder modules.

Resource Analysis: We evaluate the real-world inference latency of our EMP-based models for predicting all agents in the most complex scene in our test set. For hardware, we use two types of NVIDIA Jetson devices, which are designed for robotics applications, and include a comparison with a standard NVIDIA GPU. The results in Table 3 highlight that our approach is very well suited for real-time processing on AMR hardware. Forecast-MAE could not be tested in this evaluation setting, because it uses neighborhood attention blocks [11], which are not supported for export to ONNX.

Conclusions: To solve trajectory prediction in complex intralogistics environments, we introduce the environment context (ECTX) encoder, a versatile extension for state-of-the-art trajectory prediction models. The ECTX encoder captures detailed map information on intralogistics specific elements like racks. Extensive evaluations demonstrate that integrating ECTX significantly enhances baseline accuracy. Furthermore, our findings emphasize that for custom robotic domains, specialized compact architectures consistently outperform standard, large-scale models designed for autonomous driving.

Acknowledgments: We gratefully acknowledge the financial support by the Austrian Federal Ministry for Digital and Economic Affairs, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association.

References

[1] Yicheng Liu, Jinghuai Zhang, Liangji Fang, Qinhong Jiang, and Bolei Zhou. Multimodal Motion Prediction with Stacked Transformers. *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7577–7586, 2021.

[2] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion Transformer with Global Intention Localization and Local Movement Refinement. In *Proc. of the Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[3] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-Centric Trajectory Prediction. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[4] Xishun Wang, Tong Su, Fang Da, and Xiaodong Yang. ProphNet: Efficient Agent-Centric Motion Forecasting with Anchor-Informed Proposals. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[5] Yiqian Gan, Hao Xiao, Yizhe Zhao, Ethan Zhang, Zhe Huang, Xin Ye, and Lingting Ge. MGTR: Multi-Granular Transformer for Motion Prediction with LiDAR. In *Proc. of the International Conference on Robotics and Automation (ICRA)*, 2024.

[6] Kan Chen, Runzhou Ge, Hang Qiu, Rami Ai-Rfou, Charles R Qi, Xuanyu Zhou, Zoey Yang, Scott Ettinger, Pei Sun, Zhaoqi Leng, et al. WOMD-LiDAR: Raw Sensor Dataset Benchmark for Motion Forecasting. In *Proc. of the International Conference on Robotics and Automation (ICRA)*, 2024.

[7] Yang Zhou, Hao Shao, Letian Wang, Steven L. Waslander, Hongsheng Li, and Yu Liu. SmartRefine: A Scenario-Adaptive Refinement Framework for Efficient Motion Prediction. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[8] Alexander Prutsch, Horst Possegger, and Horst Bischof. Action-By-Detection: Efficient Forklift Action Detection for Autonomous Mobile Robots in Warehouses. In *Proc. of the International Conference on Robotics and Automation (ICRA)*, 2024.

[9] Jie Cheng, Xiaodong Mei, and Ming Liu. Forecast-MAE: Self-supervised Pre-training for Motion Forecasting with Masked Autoencoders. In *Proc. of the IEEE/CVF Conference on Computer Vision (ICCV)*, 2023.

[10] Alexander Prutsch, Horst Bischof, and Horst Possegger. Efficient Motion Prediction: A Lightweight & Accurate Trajectory Prediction Model With Fast Training and Inference Speed. In *Proc. of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2024.

[11] Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood Attention Transformer. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[12] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 652–660, 2017.

[13] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2020.

[14] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *Proc. of the International Conference on Learning Representations (ICLR)*, 2019.

[15] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, et al. Argoverse: 3D Tracking and Forecasting with Rich Maps. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[16] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A Multimodal Dataset for Autonomous Driving. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[17] Benjamin Wilson, William Qi, Tanmay Agarwal, John Lambert, Jagjeet Singh, Siddhesh Khandelwal, Bowen Pan, Ratnesh Kumar, Andrew Hartnett, Jhony Kaesemodel Pontes, Deva Ramanan, Peter Carr, and James Hays. Argoverse 2: Next Generation Datasets for Self-driving Perception and Forecasting. In *Proc. of the Neural Information Processing Systems Track on Datasets and Benchmarks (NeurIPS Datasets and Benchmarks 2021)*, 2021.

D²DINO: Dense Descriptors from DINO for Pixel-Level Object Understanding

Paolo Sebeto[1]*, Jean-Baptiste Weibel[2], Christian Hartl-Nesic[1], Markus Vincze[1]
[1]Automation and Control Institute, TU Wien,
[2]Institute of Forest Engineering, Department of Ecosystem Management, Climate and Biodiversity, University of Natural Resources and Life Sciences (BOKU),
Vienna, Austria

Abstract

Learning dense, pose-aware object descriptors is a key ingredient for generalizing robotic manipulation across novel instances and viewpoints. Intermediate features from self-supervised models like DINO and Stable Diffusion can serve as powerful dense descriptors for semantic correspondence, yet these features degrade under large viewpoint changes. To address this, we introduce D²DINO, a descriptor prediction model for pixel level object understanding. Our model attaches a lightweight convolutional head to a frozen DINOv3 encoder and trains it to produce low-dimensional (16-D), pixel-wise descriptors at full input resolution. The head fuses multi-scale ViT features and progressively upsamples them, yielding compact descriptors that can be used directly for dense matching. Supervision comes from Normalized Object Coordinate Space (NOCS) annotations exploiting consistent 2D–3D mappings across frames. We optimize D²DINO with a contrastive objective and further distinguish between negatives on other objects or background and negatives on the same object, down-weighting the latter to encourage intra-object variation. We show that D²DINO yields higher point matching accuracy than raw DINOv3 features with upscaled inputs, while requiring only a single forward pass at the original image resolution and a much lower descriptor dimensionality.

1 Introduction

Learning dense, part-level object representations that are invariant within an object category and robust to pose changes is crucial for generalizing robotic manipulation. Prior work [Florence et al., 2018, Adrian et al., 2022, Graf et al., 2023] shows that pixel-wise descriptors, trained on RGB-D videos with object masks or on augmented RGB images, enable generalizing grasps from a single demonstration and automating bin-picking. These approaches employ Dense Object Nets (DON) [Florence et al., 2018], a fully convolutional descriptor network that can generalize to novel instances of a category as an emergent behavior.

In parallel, large-scale pretraining has produced Vision Foundation Models (VFMs) with strong dense representation capabilities. Amir et al. [2022] and Zhang et al. [2023] show that intermediate features of *pre-trained* self-supervised Vision Transformers (ViTs) [Caron et al., 2021] and Stable Diffusion models [Rombach et al., 2022] are effective dense descriptors for semantic correspondence. Yet recent analyses [El Banani et al., 2024, Sebeto et al., 2025] reveal that such features degrade under large viewpoint changes and are not tuned to the fine-grained intra-category structure needed in manipulation, where precise part-level geometry for a single category is more important than broad semantics across many categories. This limitation is depicted in Fig. 1.

*Corresponding author, email address: sebeto@acin.tuwien.ac.at

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

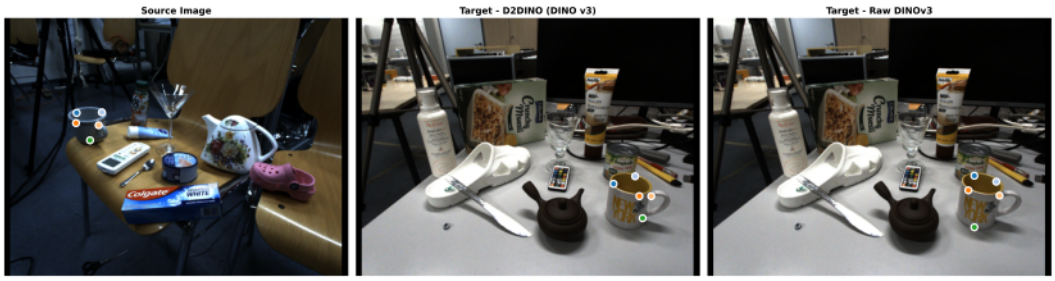


Figure 1: Qualitative comparison of point matching for unseen *cup* instances from the HouseCat6D dataset between D^{2}DINO trained with a DINOv3 encoder and 'raw' DINOv3 features. In this example, D^{2}DINO is trained on objects of the *cup* category and then used to match a small set of query pixels (left image) to their nearest neighbors in the dense descriptor space of the target image (center). Despite the two cups having significantly different orientations with respect to the camera, D^{2}DINO produces geometrically consistent correspondences that land on the same semantic parts. In contrast, using 'raw' DINOv3 descriptors without our descriptor head leads to clear mismatches (right). This illustrates how the D^{2}DINO descriptor head adapts foundation features to object pose and category-specific geometry, improving viewpoint-robust dense matching.

We argue that dense object descriptors can benefit from the semantics of VFMs while regaining pose robustness via a 3D-aware training procedure in the spirit of Florence et al. [2018]. A key challenge is computational: ViT features often have hundreds of channels (e.g., 384 for a small ViT backbone), whereas DON-style descriptors are compact (e.g., 16 dimensions). This mismatch makes operations like correlation volumes or nearest neighbor search expensive, since their cost scales linearly with descriptor dimensionality, and can render "raw" foundation features impractical for dense matching at the resolutions required for manipulation.

This motivates learning a dedicated descriptor head on top of VFMs that (i) preserves their semantics, (ii) incorporates 3D-aware supervision for viewpoint robustness, and (iii) projects features into a low-dimensional space suitable for efficient dense matching. Our goal is not to maintain full cross-category generality, but to specialize foundation features into dense descriptors that are category-focused, pose-aware, and amenable to downstream matching and control.

We therefore introduce *Dense Descriptors from DINO* (**D^{2}DINO**) for pixel-level object understanding: a descriptor prediction model that combines a DINOv3 encoder [Siméoni et al., 2025] with a lightweight prediction head and a 3D-aware contrastive objective. D^{2}DINO attaches a convolutional head to a frozen DINOv3 backbone and predicts low-dimensional descriptors at full image resolution. The head aggregates multi-scale ViT features, projects them to a lower dimension, and progressively upsamples them, producing a 16-dimensional embedding that is L2-normalized and used for dense matching within a category, trading cross-category generality for improved pose sensitivity and efficiency.

Supervision comes from dense point correspondences derived from Normalized Object Coordinate Space (NOCS) [Wang et al., 2019] annotations. Starting from a pose-estimation dataset [Jung et al., 2024], we exploit consistent per-pixel 2D–3D mappings across frames to automatically generate dense pixel correspondences between views. This provides many precise, viewpoint-varying positive pairs without extra manual labeling and directly injects 3D awareness into the descriptor space.

Training uses a normalized temperature-scaled cross-entropy (NT-Xent) loss [Chen et al., 2020] that pulls together matching descriptors while pushing apart two classes of negatives: (i) *hard* negatives, non-corresponding pixels on the same object, which are visually similar and encourage intra-object variation; and (ii) *strong* negatives, pixels on other objects or background. Hard negatives are down-weighted to avoid collapsing all object pixels while still enforcing strong separation from background and other objects.

We evaluate D^{2}DINO against raw DINOv3 dense features at the original and at higher-resolution inputs, at matched output resolutions. We further ablate our loss design and analyze its effect on dense point-matching accuracy.

2 Related Works

Dense Object Nets (DON) [Florence et al., 2018] introduced fully convolutional networks that learn dense, category-specific descriptors from RGB-D videos with object masks, enabling single-demonstration grasp transfer and related manipulation skills. Subsequent work improved training stability and efficiency, for example by streamlining data collection and adopting InfoNCE-style objectives [Adrian et al., 2022], or by using NeRF-based supervision to obtain multi-view consistent descriptors of photometrically challenging objects [Yen-Chen et al., 2022]. Other extensions replace continuous RGB-D video with unordered image collections and heavy image augmentation, demonstrating applications such as automated bin picking [Graf et al., 2023]. Further studies investigate how to train descriptors that transfer from simulation to the real world [Cao et al., 2023].

A fundamental ingredient in these descriptor-learning methods is contrastive learning. We can distinguish two broad families of loss functions used in this setting. The first, which we refer to as a *metric* contrastive loss [Hadsell et al., 2006, Choy et al., 2016, Schmidt et al., 2016], aims to learn an embedding space by explicitly pulling positive pairs together and enforcing a fixed minimum distance between negatives. The second, *probabilistic* contrastive loss [Oord et al., 2018, Chen et al., 2020, Li et al., 2023], minimizes a categorical cross-entropy over a softmax of similarities, encouraging the model to correctly classify the positive example among a set of negatives.

Recent work increasingly leverages Vision Foundation Models (VFMs) such as DINOv2 and Stable Diffusion for robotic applications. DINOBot [Di Palo and Johns, 2024] performs manipulation via retrieval and alignment in a DINO-based embedding space. Robo-ABC [Ju et al., 2024] studies affordance generalization beyond categories using pre-trained visual representations. DoDuo [Jiang et al., 2024] learns dense visual correspondence exploiting in-the-wild video datasets and predicting flow fields. AnyOKP [Qin et al., 2024] uses VFMs to guide one-shot, instance-aware keypoint detection. These methods show that foundation features are powerful for semantic understanding and correspondence, but they typically do not explicitly learn compact, category-focused dense descriptors tailored to manipulation.

D^{2}DINO combines DON-style dense, category-focused descriptors with the semantic strength of DINOv3 [Siméoni et al., 2025] by attaching a lightweight head that produces low-dimensional (16-D) pixel-wise embeddings at full image resolution. Unlike prior DON variants or VFM-based methods, it uses 3D-aware, NOCS-derived supervision and a weighted InfoNCE objective to specialize foundation features into compact, pose-aware descriptors suitable for efficient dense matching.

3 D^{2}DINO: Dense Descriptors from DINO

D^{2}DINO predicts dense object descriptors starting from a DINOv3 ViT [Siméoni et al., 2025] encoder with the goal of learning a continuous pose aware representation of objects that can facilitate manipulation. Its workflow is summarized in Fig. 2. In the remainder of this section we describe how to obtain a pixel-wise pose-aware supervision signal, the architecture of the model and the loss used to train it.

3.1 Pose-Aware Supervision Signal

A key requirement for D^{2}DINO is a supervision signal that encodes how object surface points correspond across views. Rather than supervising descriptors only in image space, we draw inspiration from 6D pose estimation and use NOCS [Wang et al., 2019] as an intermediate, 3D-aware representation. Normalized Object Coordinate Space (NOCS) represents each object by a canonical, normalized 3D model, and encodes surface points on this model as RGB values with a one-to-one correspondence between color and 3D coordinate. A per-pixel NOCS map for an image is then obtained by rendering the canonical model into the camera frame using the ground-truth 6D pose and recording the corresponding canonical 3D coordinate for each visible pixel on the object. Since the same canonical model and encoding are used across frames, pixels that correspond to the same physical surface point share the same NOCS coordinate in all views where that point is visible, making NOCS a natural candidate for supervising pose-aware, category-level dense descriptors.

Given two frames that observe the same object, we exploit their NOCS maps to obtain dense, viewpoint-consistent pixel correspondences. For a pixel (u, v) in a source image, we read its NOCS

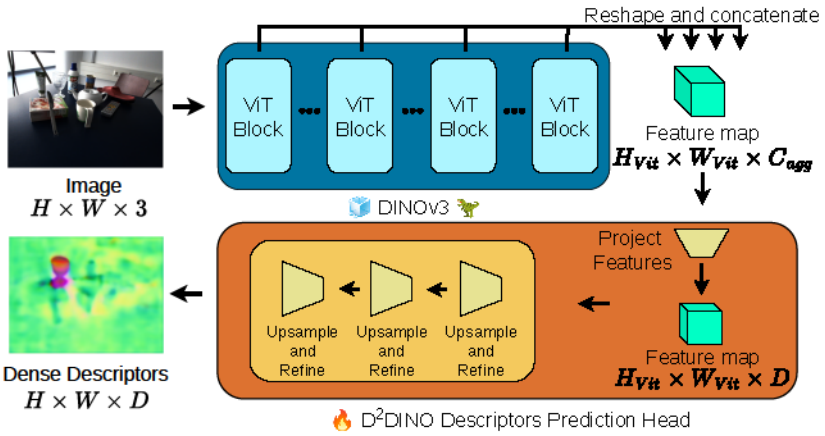


Figure 2: Overview of the D²DINO descriptor prediction head. Given an input RGB image of size $H \times W \times 3$, we first extract 4 sets of intermediate token embeddings from a frozen DINOv3 transformer encoder, obtaining a sequence of features that are reshaped and concatenated into a 2D feature map of size $H_{\text{Vit}} \times W_{\text{Vit}} \times C_{agg}$, where C_{agg} is the channel dimension of the aggregated DINOv3 patch embeddings. This feature map is fed to the D²DINO head, which consists of a linear projection layer that maps the backbone feature channels to the D-dimensional descriptor space followed by a stack of upsampling-and-refinement blocks producing dense descriptors of size $H \times W \times D$ used in our contrastive training objective. The figure shows PCA-colored descriptors learned for 'glass' objects (i.e., the 16-dimensional descriptors are projected to 3 dimensions via Principal Component Analysis, normalized, and mapped directly to RGB color channels for visualization).

coordinate $\mathrm{x} \in \mathbb{R}^3$ and search in the target image for pixels for which the NOCS coordinate is equal (up to a small tolerance) to x. If such a pixel exists, we treat the two pixels as a positive correspondence, since they are projections of the same canonical 3D point under different camera poses. Repeating this procedure across frames in a sequence yields large numbers of positive pairs that cover a wide range of viewpoints and occlusions. Pixels with valid NOCS in one view but no match in the other are simply ignored for supervision. This correspondences sampling process is shown in Fig. 3.

3.2 Dense Descriptors Prediction Head

Given an input image $\mathrm{I} \in \mathbb{R}^{H \times W \times 3}$, D²DINO first extracts a set of intermediate DINOv3 ViT features at multiple layers and then maps them to a low-dimensional, pixel-wise descriptor field using a lightweight convolutional head with learned upsampling.

Multi-scale ViT features. We use a DINOv3 ViT encoder with patch size p as backbone. For a given image, the encoder produces a sequence of token embeddings at several transformer layers. We select a set of L layers, specified by their indices, and reshape their spatial tokens into feature maps

$$\{\mathrm{F}^{(\ell)} \in \mathbb{R}^{H_{\text{Vit}} \times W_{\text{Vit}} \times C} \mid \ell = 1, \ldots, L\}, \tag{1}$$

where $H_{\text{Vit}} = \lceil H/p \rceil$ and $W_{\text{Vit}} = \lceil W/p \rceil$ after rounding the input size to the closest multiple of p. We ignore the class token and keep only patch tokens.

Features aggregation. To fuse information across layers, first the features $\mathrm{F}^{(\ell)}$ are concatenated along the channel dimension:

$$\mathrm{F}_{\text{agg}} = \mathrm{Concat}\big(\{\mathrm{F}^{(\ell)}\}_{\ell=1}^{L}\big) \in \mathbb{R}^{H_{\text{Vit}} \times W_{\text{Vit}} \times C_{\text{agg}}}, \quad C_{\text{agg}} = L \cdot C. \tag{2}$$

We apply batch normalization and a 1×1 convolution to project this aggregated tensor into a descriptor space of dimension D:

$$\mathrm{D}_{\text{low}} = \mathrm{Conv}_{1\times1}\big(\mathrm{BN}(\mathrm{F}_{\text{agg}})\big) \in \mathbb{R}^{H_{\text{Vit}} \times W_{\text{Vit}} \times D}. \tag{3}$$

This yields a low-resolution dense descriptor map aligned with the ViT patch grid.

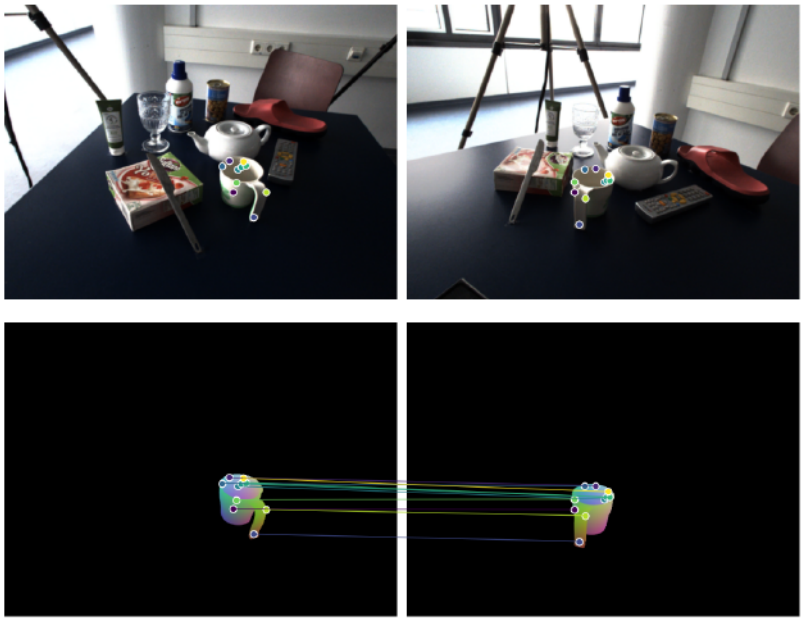


Figure 3: Example of point correspondences sampled using NOCS maps. On the left the source image, on the right the target image, each with its respective NOCS map below it. The NOCS map uniquely identifies the position of a point in the image on the 3D model of the object in a way that does not depend on camera orientations.

Learned upsampling head. To obtain descriptors at the original image resolution (H, W), we use a shallow convolutional decoder with learned upsampling. In our final architecture, the decoder consists of three repeated blocks, each comprising a 3×3 convolution, group normalization, GELU nonlinearity, and a factor-2 spatial upsampling:

$$\mathbf{G}^{(k)} = \mathrm{Upsample}_2\Big(\sigma\big(\mathrm{GN}(\mathrm{Conv}_{3\times3}(\mathbf{G}^{(k-1)}))\big)\Big), \quad k = 1, 2, 3, \tag{4}$$

with $\mathbf{G}^{(0)} = \mathbf{D}_{\text{low}}$, σ denoting GELU. The number of channels D is kept constant. A 3×3 convolution refines the feature map, and a final interpolation aligns the spatial size exactly with (H, W):

$$\mathbf{D}_{\text{full}} = \mathrm{Interp}_{H,W}\big(\mathrm{Conv}_{3\times3}(\mathbf{G}^{(3)})\big) \in \mathbb{R}^{H\times W\times D}. \tag{5}$$

Finally, we ℓ_2-normalize the descriptors along the channel dimension at each pixel,

$$\hat{\mathbf{D}}(u, v) = \frac{\mathbf{D}_{\text{full}}(u, v)}{\|\mathbf{D}_{\text{full}}(u, v)\|_2}, \tag{6}$$

obtaining unit-norm descriptors $\hat{\mathbf{D}} \in \mathbb{R}^{H\times W\times D}$ that are directly comparable via dot products in the contrastive loss and during dense matching.

3.3 Self-Supervised Contrastive Loss

We train D^{2}DINO with a self-supervised contrastive objective that enforces high similarity between descriptors of corresponding pixels across views while repelling non-corresponding pixels to prevent representation collapse. For a given image pair, let $\hat{\mathbf{D}}^s, \hat{\mathbf{D}}^t \in \mathbb{R}^{H\times W\times D}$ denote the source and target descriptor maps (unit-norm along the channel dimension). From the NOCS-derived correspondences, we obtain a set of positive pixel pairs $\mathcal{P} = \{(\mathbf{u}_i^s, \mathbf{u}_i^t)\}_{i=1}^{N_{\text{pos}}}$. To form negative pairs, we extract a set of strong negatives in the target image $\mathcal{N}_{\text{strong}} = \{\mathbf{v}_j^t\}_{j=1}^{N_{\text{strong}}}$, representing background or other objects. To explicitly encourage intra-object variation and prevent all parts of an object from converging to

the same representation, we also sample a set of hard negatives $\mathcal{N}_{\text{hard}} = \{\mathbf{w}_k^t\}_{k=1}^{N_{\text{hard}}}$ from different locations on the same object.

For each positive pair, we define the query $\mathbf{q}_i = \hat{\mathbf{D}}^s(\mathbf{u}_i^s)$ and positive key $\mathbf{k}_i^+ = \hat{\mathbf{D}}^t(\mathbf{u}_i^t)$. We collect all negative keys from the target image into a single matrix:

$$\mathbf{K}^- = \left[\mathbf{k}_1^{\text{strong}}, \ldots, \mathbf{k}_{N_{\text{strong}}}^{\text{strong}}, \mathbf{k}_1^{\text{hard}}, \ldots, \mathbf{k}_{N_{\text{hard}}}^{\text{hard}}\right]^\top \in \mathbb{R}^{(N_{\text{strong}}+N_{\text{hard}})\times D} \tag{7}$$

Using dot-product similarity with a temperature parameter τ, the positive logit is $s_i^+ = \frac{\mathbf{q}_i^\top \mathbf{k}_i^+}{\tau}$, and the negative logits are $\mathbf{s}_i^- = \frac{\mathbf{q}_i (\mathbf{K}^-)^\top}{\tau} \in \mathbb{R}^{N_{\text{strong}}+N_{\text{hard}}}$.

Standard contrastive learning repels all negatives equally. However, treating pixels from the same object (hard negatives) exactly like background pixels is overly aggressive and can disrupt part-level semantics. Therefore, we use a soft-weighted variant of the normalized temperature-scaled cross-entropy (NT-Xent) loss inspired by Li et al. [2023]. We assign a scalar weight $w_m \in (0, 1]$ to each negative key: $w_m = 1$ for strong negatives, and a reduced weight $w_m = \alpha \in (0, 1)$ for hard negatives. These weights scale the contribution of each negative inside the softmax denominator:

$$\mathcal{L}_{\text{wNT-Xent}}(\mathbf{q}_i) = -\log \frac{\exp(s_i^+)}{\exp(s_i^+) + \sum_m w_m \exp(s_{i,m}^-)} \tag{8}$$

Intuitively, this strategy pushes background pixels far away from the anchor while keeping different points on the same object moderately separated. It avoids a trivial collapse but applies a weaker penalty than for truly distinct regions. The final contrastive loss for an image pair is obtained by averaging $\mathcal{L}_{\text{wNT-Xent}}(\mathbf{q}_i)$ over all positive correspondences in that pair.

4 Experimental Evaluation

We evaluate D^{2}DINO on point matching under viewpoint and instance variation. The goal is to quantify whether the proposed descriptor head and weighted NT-Xent loss yield more reliable pixel-level matches than directly using foundation-model features, and to isolate the contribution of each loss component by evaluating targeted ablations.

4.1 Experimental Setup

Evaluation Metric We report matching accuracy as PCK@0.10, used consistently for validation during training and test evaluation. For each source keypoint, we extract its descriptor and find its nearest neighbor over all target-image pixels in descriptor space using cosine similarity. Let $\hat{\mathbf{u}}_i^t$ be the predicted target location and $\mathbf{u}_i^t$ the ground-truth target location for correspondence i; the pixel error is $e_i = \|\hat{\mathbf{u}}_i^t - \mathbf{u}_i^t\|_2$. A match is counted as correct when

$$e_i \le \tau, \qquad \tau = 0.10 \cdot \max(h_{\text{bbox}}, w_{\text{bbox}}), \tag{9}$$

where h_{bbox} and w_{bbox} are the height and width of the target object bounding box. PCK@0.10 is the percentage of correspondences satisfying this condition.

Dataset and correspondence pair construction. We use HouseCat6D [Jung et al., 2024] and focus on four object categories: *cup*, *glass*, *shoe*, and *teapot*. We use the official train, validation and test scenes split from the dataset, noting that the validation and test scenes depict different objects than those in the training set. For each category-specific run, we filter scenes by category and build supervision pairs from NOCS-consistent correspondences as described in Section 3.1. For training, we sample up to 1000 positive correspondences and 200 strong negatives per pair, plus 50 hard negatives.

Training protocol. All models are trained with a D^{2}DINO architecture using a frozen DINOv3-*small* encoder and a convolutional upsampling head that predicts 16-dimensional descriptors at full image resolution. From DINOv3 we use the token embeddings of $L = 4$ intermediate layers, with indices $\{8, 9, 10, 11\}$. The optimizer is AdamW with constant learning rate 2×10^{-3}, batch size 8, 15 epochs, and early stopping patience of 3 epochs based on PCK@0.10 validation performance. We sample 5% of available training scenes' images with a background randomization probability of 50% as augmentation. We use a constant weight $\alpha = 0.1$ for hard negatives.

Table 1: PCK@0.10 for different descriptor extraction methods and loss functions on four object categories.

Method	Loss	Object categories			
		Cup	Glass	Shoe	Teapot
DINOv3	–	42.57	39.62	65.47	59.12
DINOv3 w/ $\times 1.5$ upsampling	–	49.58	41.87	68.71	63.35
D^2DINO	NT-Xent w/ soft hard negs.	**52.31**	**60.34**	**70.26**	**78.26**
D^2DINO	DON loss w/ soft hard negs.	44.88	57.05	59.88	71.27
D^2DINO	NT-Xent w/ strong negs. only	45.43	56.70	64.10	63.65

Baselines and ablations. We compare against *raw* DINOv3 dense features extracted without our prediction head in two settings: (i) original image input resolution and (ii) upsampled input ($1.5\times$). This comparison is justified because DINOv3 is optimized for higher-resolution inputs. Furthermore, our focus is on dense *object* matching; since the target object may occupy only a small portion of the overall image, upscaling the input allows the model to capture finer details and brings expected benefits to matching performance. The specific choice of a $1.5\times$ upsampling factor is motivated by the observation that further increases do not yield additional benefits. However, this performance gain comes at a steep computational cost, because vision transformers have quadratic computation complexity to input image size [Liu et al., 2021].

To analyze the learning objective, we run two ablations starting from the same base configuration: (1) replacing weighted NT-Xent with the Dense Object Nets [Florence et al., 2018] metric contrastive loss, and (2) removing hard negatives sampling. This isolates the effect of the loss function family and of the proposed hard negative treatment.

4.2 Results

We present test results in Table 1. D^2DINO descriptors substantially outperform even the upsampled DINOv3 baseline with $1.5\times$ upsampling across all four object categories, with gains of 2.7 points on cups, 18.5 points on glasses, 1.6 points on shoes, and 14.9 points on teapots (PCK@0.10). The strongest improvements appear on teapots and glasses—categories with highly varied shapes in the training set, with glasses posing the added challenge of transparent surfaces. The D^2DINO prediction head effectively captures this intra-category variance, generalizing successfully to novel test instances.

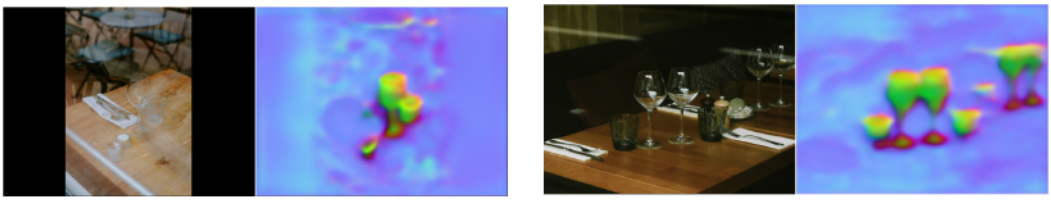
The loss ablation reveals complementary roles for each component. The DON loss [Florence et al., 2018]—a pairwise margin loss with fixed distance thresholds—struggles to capture nuanced intra-object variations, whereas our softmax-based NT-Xent provides a more flexible similarity scale. Crucially, omitting hard-negative sampling entirely causes descriptor collapse. Without same-object negatives, the model only learns to separate the target object from the background, merging distinct object parts into a uniform representation. Introducing soft-weighted hard negatives prevents this collapse, explicitly enforcing the intra-object variation necessary for fine-grained, part-level discriminability.

Importantly, the test evaluation succeeds on *novel objects* within each category. This emergent generalization—from instance-level to category-level dense matching—demonstrates that D^2DINO learns pose-robust part representations transferable across instances.

Qualitative PCA visualization in Section 5 further illustrates this for glasses: D^2DINO descriptors form tight, semantically coherent clusters for corresponding parts across images with multiple novel instances (different poses, partial occlusions).

We measured the computation of D^2DINO descriptors to have a $2.8\times$ speedup on average per image pair against using DINOv3 with $1.5\times$ upsampling.

Together, these results show that D^2DINO's descriptor head effectively specializes foundation features for category-level dense correspondence, gaining pose and photometric robustness while remaining computationally efficient compared to high-dimensional raw features.



(a) (b)

Figure 4: PCA visualization of dense descriptors learned using D^{2}DINO for the glass object category computed on images depicting instances not seen during training. This example shows the emergent generalization capabilities of D^{2}DINO for a challenging class such as glasses with diverse geometries and transparent surfaces.

5 Conclusions

We introduced D^{2}DINO, a lightweight descriptor prediction head that specializes frozen DINO foundation features into compact, category-level dense descriptors optimized for robotic manipulation. By combining multi-scale ViT feature fusion with a 3D-aware contrastive objective using NOCS-derived correspondences, D^{2}DINO achieves superior point matching accuracy compared to raw DINOv3 features—even with input upsampling—on geometrically diverse and photometrically challenging objects like teapots and transparent glasses.

Key insights from our analysis include: (1) introducing soft-weighted hard negatives to the NT-Xent loss prevents descriptor collapse while encouraging part-level discriminability; (2) training on instance-level correspondences emergently generalizes to novel objects within the same category; and (3) the low-dimensional output (16D vs 384D) yields substantial computational gains suitable for real-time dense matching.

D^{2}DINO demonstrates that distilling the rich semantics of foundation models through category-focused, pose-aware supervision produces descriptors that are both more accurate and more efficient than higher-dimensional raw features. This approach bridges traditional Dense Object Nets with modern Vision Foundation Models, enabling pixel-level object understanding that generalizes across instances and viewpoints for robotic perception and control. Future work will explore scaling the training using synthetically rendered images and multi-object training.

6 Acknowledgments

The authors gratefully acknowledge the financial support of Festo AG & Co. KG.
The research leading to these results has received funding from EC Horizon 2020 for Research and Innovation under grant agreement No. 101017089, TraceBot.

References

D. B. Adrian, A. G. Kupcsik, M. Spies, and H. Neumann. Efficient and robust training of dense object nets for multi-object robot manipulation. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 1562–1568, 2022. doi: 10.1109/ICRA46639.2022.9812274.

S. Amir, Y. Gandelsman, S. Bagon, and T. Dekel. On the effectiveness of vit features as local semantic descriptors. In *European Conference on Computer Vision*, pages 39–55. Springer, 2022.

H.-G. Cao, W. Zeng, and I.-C. Wu. Learning sim-to-real dense object descriptors for robotic manipulation. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, pages 9501–9507, 2023. doi: 10.1109/ICRA48891.2023.10161477.

M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

C. B. Choy, J. Gwak, S. Savarese, and M. Chandraker. Universal correspondence network. *Advances in neural information processing systems*, 29, 2016.

N. Di Palo and E. Johns. Dinobot: Robot manipulation via retrieval and alignment with vision foundation models. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2798–2805, 2024. doi: 10.1109/ICRA57147.2024.10610923.

M. El Banani, A. Raj, K.-K. Maninis, A. Kar, Y. Li, M. Rubinstein, D. Sun, L. Guibas, J. Johnson, and V. Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21795–21806, 2024.

P. R. Florence, L. Manuelli, and R. Tedrake. Dense object nets: Learning dense visual object descriptors by and for robotic manipulation. In *Conference on Robot Learning*, pages 373–385. PMLR, 2018.

C. Graf, D. B. Adrian, J. Weil, M. Gabriel, P. Schillinger, M. Spies, H. Neumann, and A. G. Kupcsik. Learning dense visual descriptors using image augmentations for robot manipulation tasks. In K. Liu, D. Kulic, and J. Ichnowski, editors, *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 871–880. PMLR, 2023.

R. Hadsell, S. Chopra, and Y. LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE computer society conference on computer vision and pattern recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

Z. Jiang, H. Jiang, and Y. Zhu. Doduo: Learning dense visual correspondence from unsupervised semantic-aware flow. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12420–12427, 2024. doi: 10.1109/ICRA57147.2024.10611587.

Y. Ju, K. Hu, G. Zhang, G. Zhang, M. Jiang, and H. Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *European Conference on Computer Vision*, pages 222–239. Springer, 2024.

H. Jung, S.-C. Wu, P. Ruhkamp, G. Zhai, H. Schieber, G. Rizzoli, P. Wang, H. Zhao, L. Garattoni, S. Meier, et al. Housecat6d-a large-scale multi-modal category level 6d object perception dataset with household objects in realistic scenarios. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22498–22508, 2024.

H. Li, X. Zhou, L. A. Tuan, and C. Miao. Rethinking negative pairs in code search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12760–12774, 2023.

Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.

A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

F. Qin, T. Hou, S. Lin, K. Wang, M. C. Yip, and S. Yu. Anyokp: One-shot and instance-aware object keypoint extraction with pretrained vit. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 12397–12403, 2024. doi: 10.1109/ICRA57147.2024.10610601.

R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

T. Schmidt, R. Newcombe, and D. Fox. Self-supervised visual descriptor learning for dense correspondence. *IEEE Robotics and Automation Letters*, 2(2):420–427, 2016.

P. Sebeto, J.-B. Weibel, C. Hartl-Nesic, and M. Vincze. Evaluating pose awareness and 3d consistency in semantic matching. In *International Conference on Robotics, Computer Vision and Intelligent Systems*, pages 275–293. Springer, 2025.

O. Siméoni, H. V. Vo, M. Seitzer, F. Baldassarre, M. Oquab, C. Jose, V. Khalidov, M. Szafraniec, S. Yi, M. Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

H. Wang, S. Sridhar, J. Huang, J. Valentin, S. Song, and L. J. Guibas. Normalized object coordinate space for category-level 6d object pose and size estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2642–2651, 2019.

L. Yen-Chen, P. Florence, J. T. Barron, T.-Y. Lin, A. Rodriguez, and P. Isola. Nerf-supervision: Learning dense object descriptors from neural radiance fields. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 6496–6503, 2022. doi: 10.1109/ICRA46639.2022.9812291.

J. Zhang, C. Herrmann, J. Hur, L. Polania Cabrera, V. Jampani, D. Sun, and M.-H. Yang. A tale of two features: Stable diffusion complements dino for zero-shot semantic correspondence. *Advances in Neural Information Processing Systems*, 36:45533–45547, 2023.

Overcoming Nature: Perception for Autonomous Navigation in Dense Vegetation

Lukas Wimmer[*], Andre Koczka, Uros Petrovic, and Gerald Steinbauer-Wagner[†]

Abstract

Autonomous navigation in densely vegetated off-road environments remains challenging because conventional geometric perception often treats traversable vegetation as non-traversable obstacles. In this work, we present a modular semantic–geometric perception pipeline for vegetation-aware navigation. The approach combines camera-based semantic data with LiDAR to generate a local grid map containing geometric and semantic information. A subsequent filtering stage uses this representation to correct vegetation-induced artifacts in standard elevation maps while preserving rigid obstacles for navigation. The system is designed to be portable across multiple robot platforms and sensor configurations. The pipeline was evaluated in challenging alpine off-road environments on three robot platforms, indicating improved distinction between traversable vegetation and solid obstacles and supporting more reliable navigation in dense natural environments.

1 Introduction and Motivation

Perception and terrain modeling for navigation in challenging, unstructured environments have seen significant progress in recent years (1; 2; 3). However, navigation in densely vegetated off-road environments remains particularly difficult. Forest trails, alpine terrain, and overgrown vegetation contain a mixture of objects such as trees, rocks, bushes, and tall grass. While some of these objects must be avoided, others are physically traversable, creating a fundamental ambiguity for perception systems. Conventional geometric terrain representations, such as elevation maps, often fail in these settings because vegetation can appear as non-traversable structure, leading to overly conservative planning behavior. At the same time, semantic segmentation has become an important tool for scene understanding in robotics. Convolutional architectures such as U-Net (4) and DeepLabV3 (5), adaptive multimodal models such as AdapNet (6), and transformer-based architectures such as Mask2Former (7) have significantly improved semantic perception in outdoor environments. Their practical usefulness depends strongly on suitable training data, with datasets such as Freiburg Forest (8), RUGD (9), and WildScenes (10) providing increasingly relevant benchmarks for natural off-road scenes. In parallel, learning-based traversability approaches such as TerraPN (11) and the uncertainty-aware method of Lee et al. (3) have shown that combining semantics and geometry can improve off-road navigation. However, these methods often require large training datasets, complex learning pipelines, or terrain-specific motion models.

In this work, we pursue a hybrid alternative that does not learn traversability directly. Instead, semantic information obtained from image segmentation is fused with LiDAR-based terrain geometry to explicitly refine a local map representation. The key idea is a sector-based semantic grid representation with dedicated Ground, Obstacle, and Sky layers, which is then merged with a conventional elevation map. This allows traversable vegetation to be treated as passable terrain while preserving rigid objects such as rocks or tree trunks as obstacles. An overview of the pipeline is shown in Figure 1. The contributions of this work are the following: First, we present a sector-based semantic

[*] wimmer1luk@gmail.com

[†] Institute of Software Engineering and Artificial Intelligence, Graz University of Technology, Graz, Austria. {akoczka,uros.petrovic,gerald.steinbauer-wagner}@tugraz.at

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

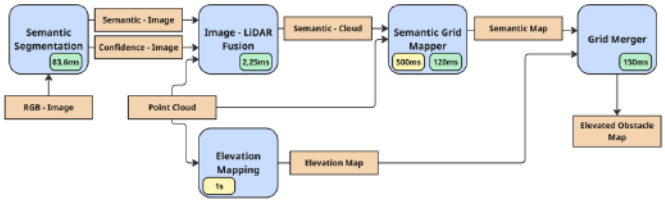


Figure 1: Performance Overview of the System. The green boxes in the modules are the execution times. The yellow boxes represent the execution period.

grid mapping method for vegetation-aware perception in dense natural environments. Second, we introduce a probabilistic fusion scheme for semantic classes and obstacle occupancy. Third, we propose a map-merging strategy that corrects vegetation-induced roughness and overhanging-branch artifacts in standard elevation maps. The resulting system is evaluated on three robot platforms in challenging alpine off-road environments, demonstrating practical feasibility across different sensor configurations.

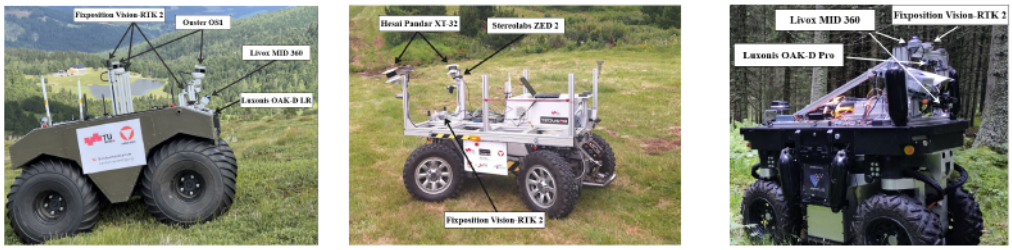


Figure 2: Warthog (left), Mercator (middle) and Artus[4](right) and their sensor setup shown on the mountain in the Seetaler Alps.

2 Methodology

The proposed system combines camera-based semantic perception with LiDAR-based terrain mapping to construct a semantically informed local map for vegetation-aware off-road navigation. The pipeline consists of four main steps: semantic image processing and image–LiDAR fusion, ground elevation estimation, semantic grid mapping, and final elevation-map correction.

First, RGB images are processed by a semantic segmentation model to obtain pixel-wise class predictions and confidence estimates. These semantic predictions are associated with LiDAR measurements by transforming each LiDAR point from the LiDAR frame into the camera frame. Points outside the camera field of view are discarded using angular constraints in the camera frame, and the remaining points are projected into the image plane using the intrinsics of the camera. Each valid 3D point is then assigned the semantic class and confidence value of the corresponding image pixel, producing a semantic point cloud. To reason about objects relative to the terrain, a robot-centered grid map is maintained. The raw LiDAR point cloud is first used to estimate a ground elevation layer. For each grid cell, the ground height is updated using the minimum-height rule $\hat{h}^+ = \min(\hat{h}^-, h(q))$, where $\hat{h}^-$ is the previous height estimate, q is a LiDAR point falling into the cell, and $h(q)$ is the height of that point in the map frame. The resulting elevation layer is smoothed and used as the reference surface for a height-based sector decomposition. Using the smoothed ground estimate, the space above the terrain is divided into three sectors: *Ground*, *Obstacle*, and *Sky*. Points between the ground elevation and the obstacle threshold are assigned to the Ground sector, points between the obstacle threshold and the robot height are assigned to the Obstacle sector, and points above the robot height are assigned to the Sky sector. Semantic observations from the fused point cloud are accumulated in the Ground and Obstacle sectors to estimate class labels for each grid cell. Let $n_c(i, j)$ denote the number of hits of class c in cell (i, j), and let C be the set of semantic classes.

[4]Artus robotic platform – CharismaTec og. https://charismatec.at/en/projects/

	Scenario 1	Scenario 2
Cell Resolution [m]	0.1	0.1
Total Area [m^2]	187	225
Ground Truth Traversable [m^2]	115.55	137.26
Total Detected Obstacles [m^2]	25.83	8.55
Intersecting Obstacles ↓ [m^2]	5.92	1.83
Intersecting ↓ [%]	22.92	21.4

Table 1: Results of the obstacle precision evaluation.

Route	Platform	Safety Interventions ↓	Functional Interventions ↓
Route 1	Artus	2	0
	Warthog	0	0
Route 2	Mercator	0	6
	Warthog	0	0
Route 3	Mercator	0	2
	Warthog	0	0

Table 2: Manual interventions during the autonomous navigation evaluation across all routes.

The total number of hits in the cell is defined as $N(i,j) = \sum_{c \in C} n_c(i,j)$. The probability that cell (i,j) belongs to class c is computed as $p_{i,j}(c) = \frac{n_c(i,j)}{N(i,j)}$.

To improve robustness against noisy single-frame predictions, these class estimates are fused over time using a log-odds representation, $\mathcal{L}_t(i,j,c) = \mathcal{L}_{t-1}(i,j,c) + L(p_{i,j}(c))$, where $L(p) = \log\left(\frac{p}{1-p}\right)$. In parallel, raw LiDAR measurements are used to update probabilistic obstacle and sky layers. The obstacle layer combines geometric measurements with the semantic obstacle class layer to determine whether a cell contains a rigid obstacle, while the sky layer captures structures above the robot height such as overhanging vegetation. The obstacle occupancy update is formulated as $\mathcal{L}_t(i,j) = \mathcal{L}_{t-1}(i,j) + L(p_{hit}(i,j)) + L(p_{miss}(i,j))$, where p_{hit} and p_{miss} are the probabilities associated with obstacle hits and misses, respectively, similarly to the approach in the book in the book Probabilistic Robotics (12). Finally, the semantic grid representation is merged with a conventional elevation map. Traversable ground classes are used to replace vegetation-induced disturbances with the smoothed ground estimate, while the sky layer allows correction of elevated cells caused by overhanging branches or foliage. The resulting local map is therefore both geometrically consistent and semantically aware, allowing the navigation stack to treat traversable vegetation as passable terrain while preserving rigid obstacles as non-traversable.

3 Implementation

The system is implemented as a modular ROS 2 perception pipeline running on Ubuntu 22.04. Semantic segmentation is performed using pretrained models, while the geometric mapping stages operate on a robot-centered rolling grid map that stores terrain height, semantic ground and obstacle classes, and probabilistic obstacle and sky layers. In the final experiments, Mask2Former with a Swin-L backbone was used due to its strong segmentation performance, while the full system was profiled offline using ROS 2 bagfiles before field deployment. The average CPU utilization was measured using the pidstat command from the Linux sysstat package. Figure 1 shows the timing of each component. The merged elevation map can be published at 1 Hz in the worst case, which is sufficient for local off-road navigation at around 1-1.5 m/s.

4 Experimental Setup and Results

The proposed perception pipeline is evaluated at both the component level and the system level. The evaluation is divided into three parts: obstacle precision assessment, merged map quality assessment, and fully autonomous navigation. The experiments were conducted in representative alpine and forest environments containing dense vegetation, uneven terrain, steep slopes, and overhanging branches. The system was deployed on three robotic platforms with different sensor setup. Warthog shown in Figure 2 left, was equipped with multiple LiDARs, including Ouster OS1 units and a front-mounted Livox MID-360, together with a Luxonis OAK-D LR camera and an Intel Core i7-13700HX PC with 64 GB RAM and an NVIDIA GeForce RTX 4070 Mobile. Mercator, shown in Figure 2 middle, used a Stereolabs ZED camera, two Hesai Pandar XT-32 LiDARs, and an AMD Ryzen 9 3900X PC with 64 GB RAM and an NVIDIA GeForce RTX 2070 SUPER Evo. Artus, shown in Figure 2 right, used a Luxonis Oak-D Pro camera, two Livox MID-360 LiDARs, and an Intel Core i7-13700HX PC with 64 GB RAM and an NVIDIA GeForce RTX 4070 Mobile. All platforms used Fixposition GNSS with RTK correction.

For the component-level evaluation, the obstacle precision and merged-map assessments were performed using teleoperated runs and recorded sensor data. In the obstacle precision evaluation, the robot footprint projected onto the grid map was used as a proxy for traversable ground truth. The

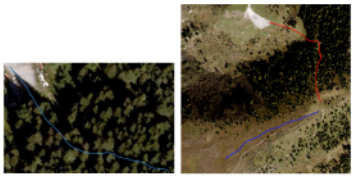


Figure 3: Pre-recorded Waypoints: (left) shows Route 1 through a forest road; (right) shows the two alpine paths. The first one in red (Route 2) is a steep hiking path. The second one in blue (Route 3) leads uphill through a less overgrown area.

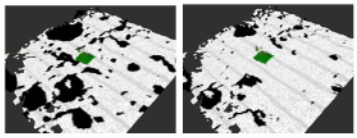


Figure 4: Merged Map Evaluation: (left) shows the Masked Slope of the standard Elevation Map; (right) shows the Masked Slope of the Merged Map; Black regions in the Slope Masks represent a slope greater than 45°. The robot's footprint is marked in green.

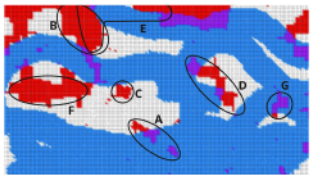


Figure 5: Evaluation Visualization Map. Blue: traversable area covered by the robot. Red: obstacles. Purple: Intersection between the obstacles and the robot. B, C, and F show true obstacles. A and D show intersections due to overhanging branches. G is a false positive.

results are shown in Table 1. In Scenario 1, the system detected $25.83\,\mathrm{m}^2$ of obstacles, with $5.92\,\mathrm{m}^2$ intersecting traversable terrain, corresponding to 22.92%. In Scenario 2, the detected obstacle area was $8.55\,\mathrm{m}^2$, with $1.83\,\mathrm{m}^2$ intersecting the traversable region, corresponding to 21.4%. These results indicate that the system detects relevant obstacles while keeping the overlap with traversable regions moderate in both scenarios. Scenario 1 is shown in Figure 5 as an example. The merged-map evaluation compares the standard elevation map with the semantically corrected map by counting cells with slope greater than 45° inside the traversed region. The number of such cells in the merged map is reduced to approximately one tenth of the corresponding number in the standard elevation map. This shows that the proposed semantic correction strongly reduces vegetation-induced artifacts while preserving the underlying terrain structure for planning. This is shown in Figure 4. For the autonomous evaluation, three predefined off-road routes were represented by GPS waypoints. Figure 3 shows the recorded routes. Route 1 was chosen to test transitions from normal roads into dense and narrow forest, Route 2 to test steep terrain with overhanging branches, and Route 3 to test disambiguation between grass, bushes, and small trees in more open terrain. Route 1 was evaluated with Artus and Warthog, while Routes 2 and 3 were evaluated with Mercator and Warthog. The main metric was the number of manual interventions, separated into safety and functional interventions. Table 2 shows the results of all routes. On Route 1, both Artus and Warthog was able to pass the narrow and densely vegetated forest, which was considered the hardest part of the route. Warthog completed the route without interventions, while Artus required two safety-related slowdowns due to its narrow footprint and high center of gravity. Using well-tuned classical elevation mapping, passing the narrow parts of this route was not possible at all. On Route 2, Warthog again completed the route without interventions, while Mercator required six functional interventions, two of them caused by branches that were still within robot height and therefore could not be cleared by the grid merger. On Route 3, Warthog completed the route without interventions, while Mercator required two functional interventions, mainly due to the combination of small and hard-to-segment vegetation and limited turning capability. Overall, the results show that the proposed semantic–geometric representation improves navigation in dense vegetation compared with conventional elevation-only mapping, while the main remaining limitations occur in edge cases.

5 Conclusion and Future Work

This work presented a sector-based semantic grid representation and map-merging strategy for autonomous navigation in densely vegetated off-road environments. By combining LiDAR-based elevation mapping with semantic information from image segmentation, the proposed system improves the distinction between traversable vegetation and rigid obstacles while reducing vegetation-induced artifacts in standard elevation maps. Field experiments on three robot platforms indicate that this semantic-geometric representation supports more reliable navigation in dense natural terrain compared to conventional elevation mapping. Future work will focus on reducing system overhead, improving obstacle modeling for more nuanced traversability cases, making ground-elevation estimation more robust under ambiguous visibility conditions, and integrating synchronized RGB-D sensing for tighter alignment between geometry and semantics.

Acknowledgments and Disclosure of Funding

This work was funded by the Austrian defense research program FORTE of the Federal Ministry of Finance (BMF) under the project PATH.

References

[1] P. Fankhauser, M. Bloesch, and M. Hutter, "Probabilistic terrain mapping for mobile robots with uncertain localization," *IEEE Robotics and Automation Letters (RA-L)*, vol. 3, no. 4, pp. 3019–3026, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8392399

[2] P. Fankhauser and M. Hutter, "A Universal Grid Map Library: Implementation and Use Case for Rough Terrain Navigation," in *Robot Operating System (ROS) – The Complete Reference (Volume 1)*, A. Koubaa, Ed. Springer, 2016, ch. 5. [Online]. Available: http://www.springer.com/de/book/9783319260525

[3] H. Lee, J. Kwon, and C. Kwon, "Learning-based uncertainty-aware navigation in 3d off-road terrains," 2022. [Online]. Available: https://arxiv.org/abs/2209.09177

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," 2015. [Online]. Available: https://arxiv.org/abs/1505.04597

[5] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017. [Online]. Available: https://arxiv.org/abs/1706.05587

[6] A. Valada, J. Vertens, A. Dhall, and W. Burgard, "Adapnet: Adaptive semantic segmentation in adverse environmental conditions," in *IEEE International Conference on Robotics and Automation (ICRA)*, 05 2017, pp. 4644–4651. [Online]. Available: https://doi.org/10.1109/ICRA.2017.7989540

[7] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," 2022. [Online]. Available: https://arxiv.org/abs/2112.01527

[8] A. Valada, G. Oliveira, T. Brox, and W. Burgard, "Deep multispectral semantic scene understanding of forested environments using multimodal fusion," in *International Symposium on Experimental Robotics (ISER)*, 03 2017, pp. 465–477. [Online]. Available: https://doi.org/10.1007/978-3-319-50115-4_41

[9] M. Wigness, S. Eum, J. G. Rogers, D. Han, and H. Kwon, "A rugd dataset for autonomous navigation and visual perception in unstructured outdoor environments," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 5000–5007. [Online]. Available: https://ieeexplore.ieee.org/document/8968283

[10] K. Vidanapathirana, J. Knights, S. Hausler, M. Cox, M. Ramezani, J. Jooste, E. Griffiths, S. Mohamed, S. Sridharan, C. Fookes, and P. Moghadam, "Wildscenes: A benchmark for 2d and 3d semantic segmentation in large-scale natural environments," *The International Journal of Robotics Research*, vol. 44, no. 4, p. 532–549, Sep. 2024. [Online]. Available: http://dx.doi.org/10.1177/02783649241278369

[11] A. J. Sathyamoorthy, K. Weerakoon, T. Guan, J. Liang, and D. Manocha, "Terrapn: Unstructured terrain navigation using online self-supervised learning," 2022. [Online]. Available: https://arxiv.org/abs/2202.12873

[12] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*, ser. Intelligent Robotics and Autonomous Agents series. MIT Press, 2005. [Online]. Available: https://books.google.at/books?id=2Zn6AQAAQBAJ

Spiking Neural Network Systems

Linearized Bregman Iterations for Sparse Spiking Neural Networks

Daniel Windhager
Silicon Austria Labs
Linz, Austria
daniel.windhager@silicon-austria.com

Bernhard A. Moser*
Software Competence Center Hagenberg
Hagenberg, Austria
bernhard.moser@scch.at

Michael Lunglmayr
Institute of Signal Processing
Johannes Kepler University
Linz, Austria
michael.lunglmayr@jku.at

Abstract

Spiking Neural Networks (SNNs) offer an energy efficient alternative to conventional Artificial Neural Networks (ANNs) but typically still require a large number of parameters. This work evaluates *Linearized Bregman Iterations* (LBI) as an optimizer for training SNNs, utilizing the algorithm's ability to enforce sparsity through iterative minimization of the Bregman distance and proximal soft thresholding updates. To improve convergence and generalization, we employ the *AdaBreg* optimizer, a momentum and bias corrected Bregman variant of Adam. Experiments on three established neuromorphic benchmarks, i.e. the Spiking Heidelberg Digits (SHD), the Spiking Speech Commands (SSC), and the Permuted Sequential MNIST (PSMNIST) datasets, show that LBI based optimization reduces the number of active parameters by about 50% while maintaining accuracy comparable to models trained with the Adam optimizer, demonstrating the potential of convex sparsity inducing methods for efficient neuromorphic learning.

1 Introduction

Sparsity in neural networks is an important and ongoing research field Hoefler et al. [2021]. In most neural network architectures sparsity refers to the weight matrices of the model being sparsely populated. This has the effect of sparsifying connections of the network (i.e. a connection with weight 0 is disconnected). For classical GPU implementations, however, having sparse weights can sometimes actually lead to increased computational overhead due to memory organization problems and unbalanced workloads Zaharia et al. [2020]. In contrast, for edge AI implementations where network structures are directly mapped to hardware, e.g. as in Windhager et al. [2025], the benefits are considerably higher as the sparsity of weights can be utilized and exploited more easily.

Training machine learning to produce sparse weight matrices is in essence a sparse optimization problem. In practice, sparsity is often achieved through heuristic approaches such as pruning Hoefler et al. [2021], even though there exist mathematically founded algorithms that can provably converge to sparse optimal solutions. A number of these algorithms is based on so-called linearized Bregman iterations, which have also been proposed for sparse estimation Osher et al. [2005], Yin et al. [2008] using concepts similar to Douglas–Rachford splitting Combettes and Eckstein [1992]. Their efficiency

*double affiliation with Institute of Signal Processing, Johannes Kepler University Linz

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

and stability for sparse estimation have been demonstrated repeatedly Hu and Chklovskii [2014], Gebhard et al. [2018], and more recent work has shown their suitability for training sparse deep neural networks, in many cases outperforming heuristic solutions Bungert et al. [2022].

In this work, we investigate how linearized-Bregman-based sparse learning performs for spiking neural networks. Specifically, we evaluate both feedforward and recurrent SNN architectures on established neuromorphic benchmarks and analyze how the regularization parameter λ influences sparsity and model accuracy.

2 Linearized Bregman Iterations

Sparse optimization problems often include a regularization term $J(\boldsymbol{\theta})$ based on the ℓ_1-norm to promote sparsity. However, the non-smoothness of the ℓ_1-norm causes standard gradient-based optimization to fail in regions where the gradient is undefined. A key property that enables a well-behaved optimization framework in such cases is the *convexity* of the regularization function J. For convex but possibly non-smooth functions, the concept of a gradient is replaced by a *sub-differential*, which generalizes differentiation to non-smooth settings. Sub-differentials, however, can not directly be used in steepest-decent like algorithms as they require a defined gradient at each step.

To handle such cases, Bregman iterations iteratively minimize the *Bregman distance* Bregman [1967]

$$D_J(x, y) = J(x) - J(y) - \langle \nabla J(y), x - y \rangle,$$

rather than minimizing the composite cost function directly. In this formulation, $\nabla J(y)$ represents an element of the *sub-differential* of J at point y, denoted $\partial J(y)$. For convex J, the sub-differential $\partial J(y)$ is a non-empty, convex set that captures all possible slopes of local supporting hyperplanes to J at y. For example, when $J(x) = |x|$, the sub-differential at $x = 0$ is the interval $[-1, 1]$. This interpretation allows the Bregman distance to generalize classical gradient-based methods to convex but non-differentiable regularizers, enabling optimization for the ℓ_1-norm and similar sparsity-promoting terms.

Because the resulting subproblems rarely admit closed-form solutions for sparse regularizers Yin et al. [2008], the *Linearized Bregman Iteration* (LBI) method provides an efficient linearized approximation. LBI introduces auxiliary "shadow" variables $\mathbf{v}$ corresponding to the model parameters $\boldsymbol{\theta}$. At iteration t, the updates can be expressed as

$$\mathbf{v}^{(t+1)} = \mathbf{v}^{(t)} + \mu \nabla L(\boldsymbol{\theta}^{(t)}, B), \qquad \boldsymbol{\theta}^{(t+1)} = \mathrm{prox}_J(\mathbf{v}^{(t+1)}),$$

where μ denotes the step size and prox_J represents the proximal operator associated with the convex regularization term J.

For the commonly used sparse regularizer $J(\boldsymbol{\theta}) = \lambda \|\boldsymbol{\theta}\|_1$, the proximal operator corresponds to the elementwise *soft-thresholding* function,

$$\mathrm{prox}_{\lambda\|\boldsymbol{\theta}\|_1}(\boldsymbol{\theta}) = \left[\mathrm{prox}_{\lambda\|\boldsymbol{\theta}\|_1}(\theta_i)\right]_i \tag{1}$$

$$\mathrm{prox}_{\lambda\|\boldsymbol{\theta}\|_1}(\theta_i) = \mathrm{sign}(\theta_i)\ \max(0, |\theta_i| - \lambda), \tag{2}$$

which suppresses small weight values and drives many parameters exactly to zero, thereby yielding sparse network representations.

This mechanism makes LBI particularly well suited for training models such as Spiking Neural Networks, where convex sparse regularization aligns well with the need for energy-efficient, low-parameter inference on neuromorphic hardware.

2.1 AdaBreg Optimization

While classical Linearized Bregman Iterations provide an effective framework for promoting sparsity, their convergence can be slow when applied to large-scale neural network training. To improve adaptation and generalization, we adopt the AdaBreg optimizer introduced by Bungert et al. [2022], which extends the Bregman iteration concept by incorporating adaptive moment estimation in analogy to the Adam optimizer Kingma and Ba [2017].

AdaBreg inherits the shadow-variable formulation of the linearized Bregman framework while maintaining separate exponential moving averages of the first and second moments of the gradient.

Let $\mathbf{m}_t$ and $\mathbf{s}_t$ denote the biased estimates of the mean and variance of the stochastic gradient $\nabla L(\boldsymbol{\theta}_t, B)$ at iteration t. For $t = 0$, both $\mathbf{m}_t$ and $\mathbf{s}_t$ may be set to $\mathbf{0}$ and 0 respectively. The update rules can then be summarized as

$$\begin{aligned}
\mathbf{m}_{t+1} &= \beta_1 \mathbf{m}_t + (1-\beta_1)\nabla L(\boldsymbol{\theta}_t, B), \\
\mathbf{s}_{t+1} &= \beta_2 \mathbf{s}_t + (1-\beta_2)\big(\nabla L(\boldsymbol{\theta}_t, B)\big)^2, \\
\mathbf{v}_{t+1} &= \mathbf{v}_t + \mu \frac{\widehat{\mathbf{m}}_{t+1}}{\sqrt{\widehat{\mathbf{s}}_{t+1}} + \epsilon}, \\
\boldsymbol{\theta}_{t+1} &= \operatorname{prox}_{\lambda\|\boldsymbol{\theta}\|_1}(\mathbf{v}_{t+1}),
\end{aligned}$$

where μ denotes the learning rate, $\widehat{\mathbf{m}}_{t+1}$ and $\widehat{\mathbf{s}}_{t+1}$ represent bias-corrected moment estimates, and ϵ is a small numerical constant for stability.

In practice, the optimizer can be seamlessly integrated into existing deep learning frameworks such as PyTorch by substituting the Adam optimizer with its AdaBreg counterpart, requiring no structural changes to the network implementation.

3 Results

3.1 Setup and configurations

All of the results presented in this work were obtained by training the networks using the AdaBreg algorithm introduced by Bungert et al. [2022], which is a Bregman version of the Adam algorithm Kingma and Ba [2017] that includes momentum and a bias correction term. AdaBreg combines the adaptive moment estimation of Adam with the sparsity inducing properties of Linearized Bregman Iterations, enabling direct integration of convex regularization into the optimization process. The reason for choosing AdaBreg over the standard linearized Bregman iteration based algorithm with momentum (LinBreg) is the better performance and generalization capability as reported by the original authors. Unless otherwise stated, the same sets of hyperparameters and initialization conditions were used across all experiments for comparability.

For the datasets, three of the most common datasets among the neuromorphic community were chosen, namely Spiking Heidelberg Digits (SHD), Spiking Speech Commands (SSC) and the Permuted Sequential MNIST (PSMNIST) dataset, along with the neuron network models presented by Queant et al. [2025], who at the time of writing hold the record for the top performing SNN on the SSC dataset. These datasets jointly cover a broad range of temporal complexities, from short event based auditory signals (SHD) to long sequential patterns (PSMNIST), providing a balanced testbed for evaluating sparsity effects across different task domains.

Table 1: Number of neurons and layers of the networks used for evaluations of Linearized Bregman iterations on the SHD, SSC and PSMNIST datasets.

Dataset	Inputs	Hidden Layer 1	Hidden Layer 2	Hidden Layer 3	Outputs
SHD	140[§]	256	256[†]	-	20
SSC	140[§]	256[†]	256[†]	256[†]	35[‡]
PSMNIST	1	64[†]	212[†]	212[†]	10[‡]

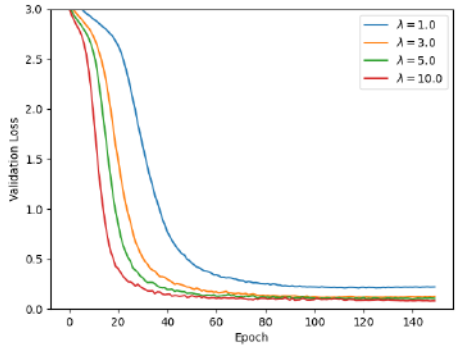
[§] Inputs are reduced from the original 700 inputs, by binning with a factor of 5.
[†] Recurrent layer with axonal delays in the recurrent path.
[‡] Outputs from last linear layer are used directly, without LIF activation.

The networks chosen for the datasets consist of either three or four layers with different feature sizes and configurations. All networks use the Leaky Integrate and Fire (LIF) neuron model. For the SHD dataset, the first and last layers of the network are simple feedforward SNN layers without any delay, while the middle layer is a recurrent layer with learned axonal delays in the recurrent path. The networks for the SSC and PSMNIST datasets each consist of three recurrent SNN layers with learned axonal delays in their recurrent paths, followed by a final linear layer without LIF activation. A simplified overview of the networks can be seen in Tab. 1, while further architectural details, including delay learning mechanisms can be found in Queant et al. [2025].

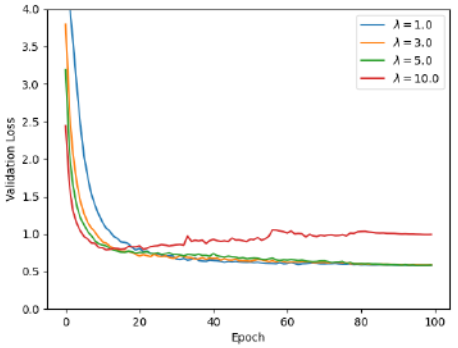
3.2 Performance with learning rate schedulers

The loss on the validation sets can be seen in Fig. 1, with the different colored curves indicating various values for the parameter λ, which controls the sparsity of the solution. These results indicate that larger λ values also introduce a beneficial regularization effect, leading to faster initial convergence and smoother loss trajectories at the early stages of training. This behavior is consistent across datasets and highlights the dual role of λ as both a sparsity and regularization parameter.

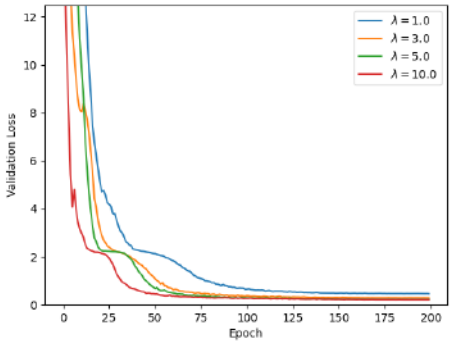
LBI seems to be more sensitive to higher learning rates, especially when used in conjunction with learning rate schedulers, which might be caused by the inherent stagnation phases that occur during training with Linearized Bregman iterations. In fact, choosing a high enough learning rate, paired with large values of the sparsity controlling parameter λ, causes the training to diverge after a few epochs. The onset of this divergent behaviour can be seen in Fig. 1b for $\lambda = 10$, which is due to λ being slightly too high for the chosen learning rate.



(a) Loss curve for SHD dataset, when trained with the OneCycleLR scheduler from PyTorch for 150 epochs, with an initial learning rate of $5 \cdot 10^{-3}$.



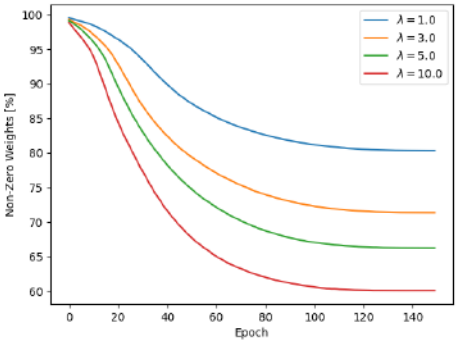
(b) Loss curve for SSC dataset, when trained with the OneCycleLR scheduler from PyTorch for 100 epochs, with an initial learning rate of $1 \cdot 10^{-3}$. Onset of divergent behaviour can be seen for $\lambda = 10$.



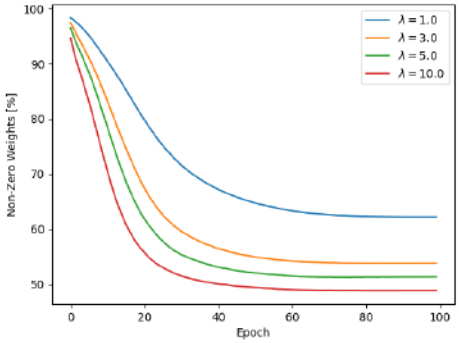
(c) Loss curve for PSMNIST dataset, when trained with the OneCycleLR scheduler from PyTorch for 200 epochs, with an initial learning rate of $1 \cdot 10^{-3}$.

Figure 1: Loss curves for the training on SHD, SSC and PSMNIST dataset with different values for λ. All curves were averaged over three separate training runs with different seeds.

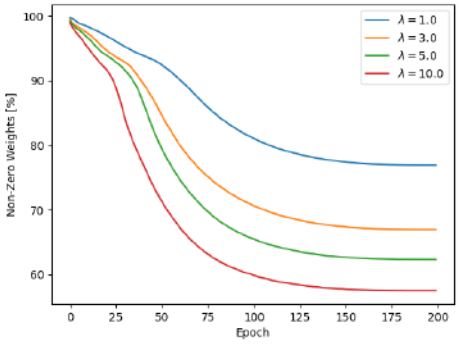
The goal of Linearized Bregman training is to produce highly sparse weight matrices, i.e. containing a large fraction of zero entries. As can be seen in Fig. 2, the number of non-zero parameters across the entire network does indeed decrease monotonically as training progresses. Much like the loss curves, the sparsity level increases rapidly during early training epochs before reaching a plateau, a behavior characteristic of Bregman iterations which preferentially eliminate less important features in the initial optimization phases.



(a) Progression of number of non-zero weights for the neural network trained on the SHD dataset.



(b) Progression of number of non-zero weights for the neural network trained on the SSC dataset.



(c) Progression of number of non-zero weights for the neural network trained on the PSMNIST dataset.

Figure 2: Number of non-zero values in networks for SHD, SSC and PSMNIST datasets during the training process plotted as a function of current epoch for different λ values. Results represent the mean across three independent training runs.

Since the parameter λ clearly influences both the best achieved accuracy and the achieved sparsity of the network, a natural question is the optimal selection of λ. Fig. 3 shows the best validation accuracy achieved across multiple training runs for all three datasets, plotted as a function of the chosen λ value. These results indicate that the choice of λ must be made depending on the learning rate and the chosen dataset. For SHD and SSC a slightly higher value of λ is beneficial, while it results in worse performance for the PSMNIST dataset.

3.3 Performance without learning rate schedulers

Since learning rate schedulers impact the training process and can even cause training divergence when coupled with a high enough learning rate, the previous experiments were repeated without any learning rate scheduling during the training. The resulting loss curves can be seen in Fig. 4.

From the subfigures in Fig. 4 it is apparent that the achieved accuracy is largely unaffected by the presence or absence of learning rate schedulers, further substantiating the hypothesis that the initial divergent behaviour was due to a too high learning rate. Furthermore, the number of non-zero weights in the network, i.e., the sparsity level, also remained largely unaffected by the presence or absence of schedulers. These sparsity curves are therefore omitted here for conciseness, as they closely mirror those shown in Fig. 2.

The achieved best validation accuracy across all datasets without learning rate scheduling, plotted as a function of the chosen λ value, can be seen in Fig. 5. These results demonstrate that without

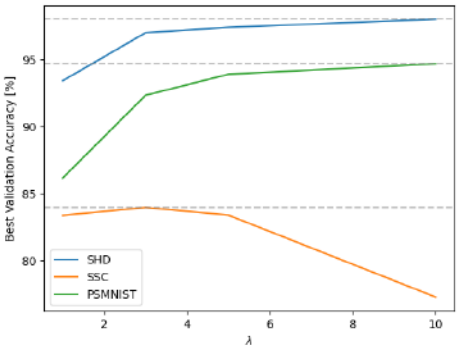
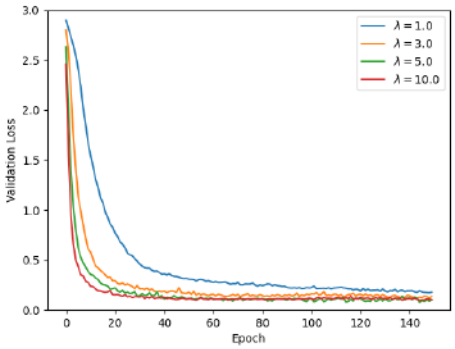
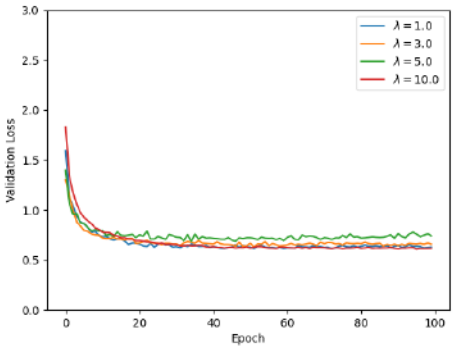


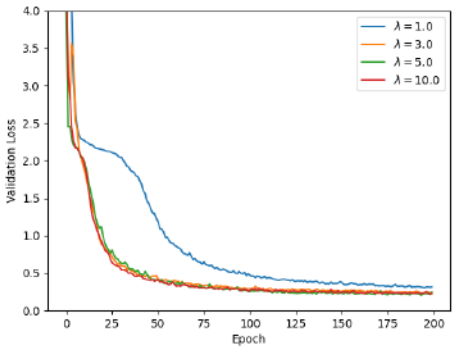
Figure 3: Peak validation accuracy across SHD, SSC, and PSMNIST datasets as a function of regularization parameter λ, averaged over multiple training runs. The optimal λ is slightly higher for SHD and PSMNIST, while a lower value of λ achieves the best results for SSC.



(a) Loss curve for SHD dataset, when trained without a learning rate scheduler for 150 epochs, with an initial learning rate of $2 \cdot 10^{-4}$.



(b) Loss curve for SSC dataset, when trained without a learning rate scheduler for 100 epochs, with an initial learning rate of $5 \cdot 10^{-4}$.



(c) Loss curve for PSMNIST dataset, when trained without a learning rate scheduler for 200 epochs, with an initial learning rate of $1 \cdot 10^{-4}$.

Figure 4: Loss curves for the training on SHD, SSC and PSMNIST dataset with different values for λ. All curves represent means over three independent training runs with different random seeds.

learning rate scheduling, a higher value of λ may be chosen sometimes (see results for PSMNIST in Fig. 3), thus potentially increasing regularization and performance on unseen data, although the effect seems comparatively small when comparing the performance on the test sets.

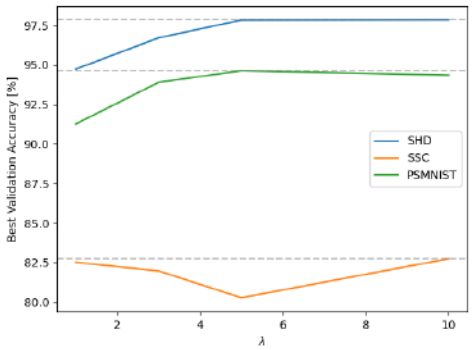


Figure 5: Peak validation accuracy without learning rate scheduling across SHD, SSC, and PSMNIST datasets versus regularization parameter λ, averaged over multiple training runs.

4 Performance compared to training with Adam

The baseline models from Queant et al. [2025] were trained using the well-known Adam optimizer Kingma and Ba [2017]. A direct comparison of test set performance between their results and ours is provided in Tab. 2. This comparison reveals that Linearized Bregman iterations (AdaBreg) achieve accuracies within 0.5–1.5% of the Adam baseline across all three datasets, despite limited hyperparameter tuning. When combined with the observed $\approx 50\%$ reduction in active parameters (cf. Fig. 2), this performance gap appears acceptable for sparsity-constrained neuromorphic applications.

Table 2: Test set accuracy comparison. Baseline results from Queant et al. [2025] (Adam optimizer) versus AdaBreg results with and without learning rate scheduling across SHD, SSC, and PSMNIST datasets.

Dataset	SHD	SSC	PSMNIST
Queant et al. [2025]	93.39%	82.58%	96.21%
Ours	92.98%	81.86%	95.59%
Ours (no LR scheduling)	92.28%	80.67%	95.11%

5 Conclusion

This work demonstrates the practical viability of Linearized Bregman Iterations (LBI) as an optimizer for Spiking Neural Networks (SNNs). Across three established neuromorphic benchmarks (SHD, SSC, PSMNIST), LBI-based training (AdaBreg) achieves competitive accuracy within 0.5–1.5% of Adam baselines while reducing the number of active parameters by approximately 50%.

Further performance gains appear achievable through systematic hyperparameter optimization. Notably, LBI integrates seamlessly into existing PyTorch workflows—requiring only a one-line optimizer replacement (Adam $\rightarrow$ AdaBreg)—dramatically lowering the adoption barrier for sparsity-aware SNN training.

These findings highlight a critical hardware-software co-design opportunity: while sparse SNN training is now readily accessible, neuromorphic hardware must evolve to fully exploit this parameter efficiency for energy-constrained edge deployments.

Acknowledgements

This work was supported by (1) the 'University SAL Labs' initiative of Silicon Austria Labs (SAL) and its Austrian partner universities for applied fundamental research for electronic based systems, and (2) the COMET Programme via SCCH funded by the Austrianministries BMIMI, BMWET, and the State of UpperAustria, (3) the COMET-K2 "Center for Symbiotic Mechatronics" of the Linz Center of Mechatronics (LCM), funded by the Austrian federal government and the federal state of Upper Austria.

The research reported in this paper has also been partly funded by the European Union's Horizon 2020 research and innovation program within the framework of Chips Joint Undertaking (Grant No. 101112268). This work has been supported by Silicon Austria Labs (SAL) owned by the Republic of Austria, the Styrian Business Promotion Agency (SFG), the federal state of Carinthia, the Upper Austrian Research (UAR), and the Austrian Association for the Electric and Electronics Industry (FEEI).

References

L. M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7(3):200–217, 1967.

L. Bungert, T. Roith, D. Tenbrinck, and M. Burger. A bregman learning framework for sparse neural networks. *Journal of Machine Learning Research*, 23(192):1–43, 2022. URL http://jmlr.org/papers/v23/21-0545.html.

P. L. Combettes and J. Eckstein. On the Douglas–Rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(1-3):293–318, 1992.

A. Gebhard, M. Lunglmayr, and M. Huemer. Investigations on sparse system identification with $$l_0$$-lms, zero-attracting lms and linearized bregman iterations. In R. Moreno-Díaz, F. Pichler, and A. Quesada-Arencibia, editors, *Computer Aided Systems Theory – EUROCAST 2017*, pages 161–169, Cham, 2018. Springer International Publishing. ISBN 978-3-319-74727-9.

T. Hoefler, C.-J. Ng, T. Yoon, P. Yu, M. Low, P. Lee, H. de Kruijf, and M. Zaharia. Sparsity in deep learning: Pruning and growth for efficient inference and training in neural networks. *Communications of the ACM*, 64(12):82–90, 2021. doi: 10.1145/3546258.3546499. URL https://dl.acm.org/doi/abs/10.5555/3546258.3546499.

T. Hu and D. B. Chklovskii. Sparse lms via online linearized bregman iteration. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7213–7217, 2014. doi: 10.1109/ICASSP.2014.6855000.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization, 2017. URL https://arxiv.org/abs/1412.6980.

S. Osher, M. Burger, D. Goldfarb, J. Xu, and W. Yin. An iterative regularization method for total variation-based image restoration. *Multiscale Modeling and Simulation*, 4(2):460–489, 2005.

A. Queant, U. Rançon, B. R. Cottereau, and T. Masquelier. Delrec: learning delays in recurrent spiking neural networks, 2025. URL https://arxiv.org/abs/2509.24852.

D. Windhager, L. Ratschbacher, B. Moser, and M. Lunglmayr. Mineuron: Minimal neuron realization for fast fpga snn inference using logic optimization. In *Proceedings of the IEEE International Conference on Image Processing (ICIP) 2025*, Sept. 2025.

W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008. doi: 10.1137/070703983.

M. Zaharia et al. Sparse gpu kernels for deep learning. *Proceedings of the International Conference for High Performance Computing, Networking, Storage and Analysis (SC)*, 2020. URL https://people.eecs.berkeley.edu/~matei/papers/2020/sc_sparse_gpu.pdf.

Recurrent versus parallelizable spiking neural networks: A comparative study

Alexander Mayr, Simon Hitzginger and Robert Legenstein
Institute of Machine Learning and Neural Computation
Graz University of Technology, 8010 Graz, Austria
{alexander.mayr, simon.hitzginger, robert.legenstein}@tugraz.at

Abstract

Spiking neural networks (SNNs) have emerged as a biologically plausible computational paradigm with strong links to real-world brain dynamics. Recently, interest has grown in parallelizable State Space Model (SSM)–inspired architectures, which offer improved scalability compared to recurrent networks. While effective at scale, these models represent a step away from biological realism. In particular, the impact of removing recurrent connections and membrane nonlinearities on the temporal processing capabilities of SNNs remains largely unexplored. In this work, we investigate the impact of these changes to the network dynamics on the temporal processing capabilities of SNNs with a focus on recurrent connectivity. To this end, a suite of sequential tasks was used to systematically compare parallelizable SSM-style networks with recurrent SNNs. The results demonstrate that while parallelizable models perform well on tasks with simple or weak temporal dependencies, they struggle to maintain persistent internal state when complex, state-dependent computation is required. In contrast, recurrent architectures exhibit superior memory retention and robustness under these conditions. These findings suggest fundamental limitations of parallelizable SSM-style approaches for sequence tasks that rely on long-term internal memory, highlighting the continued relevance of recurrence in spiking neural computation as suggested by biology.

1 Introduction

Spiking neural networks (SNNs) provide a biologically motivated framework for neural information processing in which information is transmitted via discrete spike events resembling biological action potentials [Maass, 1997, Gerstner and Kistler, 2002]. Their event-driven dynamics yield sparse activity patterns and enable temporal coding, offering the potential for improved energy efficiency and reduced memory requirements, particularly when deployed on neuromorphic hardware. Recurrence is a fundamental characteristic of biological neural circuits [Vidal-Saez et al., 2024, Larsen and Druckmann, 2022, Douglas and Martin, 2007]. In contrast, many modern machine learning architectures favour non-recurrent designs with simplified linear dynamics due to their favourable parallelisation properties, stable optimization behaviour, and scalability [Gu et al., 2020]. From a neurobiological perspective, however, the ubiquity of recurrent connectivity suggests functional significance rather than architectural redundancy, particularly for temporal information processing.

By operating intrinsically in the temporal domain, SNNs are naturally suited for sequential tasks such as speech recognition and time-series modeling. Incorporating recurrent connectivity further enhances their capacity to retain and update information over time, and recurrent SNN architectures have demonstrated strong performance on sequential learning problems [Bellec et al., 2018, Baronig et al., 2025]. Training recurrent SNNs typically relies on Backpropagation Through Time (BPTT) [Eshraghian et al., 2023], which enables gradient-based optimization but introduces substantial computational challenges. The memory and computational cost of BPTT scale linearly with both

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

sequence length and network depth, leading to high memory overhead and gradient instability. Moreover, the sequential nature of temporal backpropagation limits parallelization during training, rendering large-scale or long-sequence training computationally expensive. Recent work has sought to address these limitations by leveraging structured state space models (SSMs). By formulating a Resonate-and-Fire neuron within the HiPPO framework, the S5-RF model was introduced [Huber et al., 2024]. However, these parallelizable formulations deviate from biological neuron models by omitting spike-triggered voltage resets and recurrent network connectivity.

Alternative strategies achieve parallelization in recurrent SNNs by selectively neglecting specific gradient pathways in the computational graph, enabling scalable training of recurrent SNNs [Baronig et al., 2025, Fang et al., 2023]. While such approaches improve computational efficiency, their impact on sequential modeling performance and memory capacity remains insufficiently characterized [Merrill et al., 2024]. In this work, we systematically evaluate parallelizable SNN architectures based on S5-RF, leaky integrate-and-fire (LIF) and adaptive LIF neurons [Baronig et al., 2025, Higuchi et al., 2024] on established sequential benchmarks and compare them to non-parallelizable recurrent models incorporating voltage resets and recurrent connectivity. For tasks with limited temporal structure, we observe that the performance of parallelizable variants is slightly inferior to their recurrent counterparts. To specifically assess the capacity to preserve and update internal state representations, we further introduce a set of tasks which require accurate tracking of latent state transitions [Merrill et al., 2024]. We show that recurrent SNN architectures solve all task variants and generalize robustly beyond training conditions. In contrast, parallelizable non-recurrent models succeed only on simpler variants and fail to generalize beyond training regimes.

These findings highlight a trade-off between computational scalability and memory capacity in spiking neural network design. While parallelization substantially improves training efficiency, recurrent connectivity and biologically inspired dynamical mechanisms remain critical for tasks requiring structured temporal reasoning and robust state tracking.

2 Methods

We consider two architecture types: *non-recurrent spiking networks* which are fully parallelisable and *recurrent spiking neural networks* (RSNNs). Both use two hidden layers with task-dependent dimensions d_{in}, d_h, and d_{out}. Architectures share identical depth and differ only in connectivity. Due to the lack of recurrent connections the non-recurrent networks have less connections per layer and therefore a smaller number of weights. To ensure that all compared networks have a similar number of learnable parameters the network size for the non-recurrent networks is increased to double that of the recurrent networks.

All hidden-layer weights were initialised using orthogonal initialisation. Neuron-specific parameters were sampled from task-dependent ranges. To avoid silent neurons, inputs were rescaled as $I_{in} \leftarrow I_{in}\left(1+\frac{4}{d_{in}}\right)$. The output layer consisted of leaky integrator (LI) neurons (see Baronig et al. [2025]) with LeCun-uniform initialisation. In the ECG task LIF and SE-adLIF networks used membrane time constant $\tau = 3$ but otherwise unless stated all configurations used $\tau = 15$.

Four distinct neuron models were evaluated in this study. First, the standard LIF model serves as a well-established reference point. Second, the SE-adLIF and BRF neurons were selected as representative second-order neuron models that reflect the current state of the art in biologically inspired spiking dynamics. Finally, the S5-RF model is a parallelizable SSM-based SNN that replaces recurrent connectivity with structured state-space dynamics. Its hidden state is discretized using the Dirac scheme, which preserves high-frequency components and ensures numerical stability during parallel evaluation [Huber et al., 2024].

The Leaky Integrate-and-Fire (LIF) neuron maintains a membrane potential $u[t]$ with exponential decay and spike-triggered reset,

$$u[t] = \alpha u[t-1] + I[t] - \vartheta z[t-1], \tag{1}$$

$$z[t] = \Theta(u[t] - \vartheta), \tag{2}$$

where $I[t]$ is the input current at time step t, ϑ is the spiking threshold, $z[t] \in \{0,1\}$ is the spike output at time step t, Θ is the Heaviside step function, and $\alpha \in (0,1)$ is the decay factor. In terms of a membrane time constant τ_u, α is given by $\alpha = \exp(-dt/\tau_u)$ where dt denotes the discretization time step.

The Symplectic-Euler adaptive LIF (SE-adLIF) extends LIF with an adaptation variable $w[t]$:

$$u[t] = \alpha u[t-1](1 - z[t-1]) + (1-\alpha)(I[t] - w[t-1]), \tag{3}$$
$$w[t] = \beta w[t-1] + (1-\beta)(au[t] + bz[t-1])c_{adapt}, \tag{4}$$

where α and β are decay coefficients for u and w respectively. The coupling between the membrane potential $u[t]$ with the adaptation current $w[t]$ is scaled by the first adaptation coefficient a. The second coefficient b governs the feedback of the output spike into $w[t]$ [Baronig et al., 2025]. c_{adapt} is an additional custom adaptation coefficient for fine tuning purposes. It scales the effect of the previous output spikes, the current membrane potential and the adaptation current. The Balanced Resonate-and-Fire (BRF) neuron maintains complex-valued oscillatory dynamics [Higuchi et al., 2024]. We used the real-valued formulation given in Baronig et al. [2026]. The dynamics are mainly determined by the angular frequency parameter ω and the damping parameter b_{offset}.

To isolate the contribution of nonlinear state interactions, parallelisable variants were constructed by removing spike-dependent reset and recurrent dependencies. The resulting dynamics form linear time-invariant (LTI) recurrences while retaining the spiking nonlinearity for readout. For example, in the SE-adLIF model, spike-dependent interactions in both membrane and adaptation dynamics were removed. The resulting updates are

$$u[t] = \alpha u[t-1] + (1-\alpha)\big(I[t] - w[t-1]\big), \tag{5}$$
$$w[t] = \beta w[t-1] + (1-\beta)\, au[t], \tag{6}$$
$$z[t] = \Theta(u[t] - \vartheta), \tag{7}$$

The dynamics reduce to a linear coupled system driven by input current.

The SNNs were trained using back propagation through time (BPTT) and the ADAM optimizer [Kingma and Ba, 2014] with parameters $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$. Due to the non-differentiability of the output spikes surrogate gradients were used. The BRF models used a double Gaussian surrogate while LIF and SE-adLIF used a Heaviside surrogate implemented with Slayer [Shrestha and Orchard, 2018]. Neuron parameters were optimized alongside with network weights. A cosine annealing learning rate schedule was used [Eshraghian et al., 2022]. This allows for bigger steps in the beginning whilst retaining the ability to fine-tune in later epochs. All tasks were trained for 50 epochs. The learning rates η were initialised for the first epoch at 0.05 for the BRF, 0.02 for the LIF, and 0.01 for the SE-adLIF models. For the loss, the *per-timestep cross entropy* function was used for the ECG, SHD, and state tracking task, while the *sum of softmax* was used for SMNIST. To determine test accuracies, class labels were predicted using *per timestep* prediction in the ECG and state tracking tasks while *mean over sequence* was used for SHD and *sum of softmax* prediction was used for SMNIST.

Unless stated otherwise, neuron model parameters were initialised by sampling independently from predefined ranges. For the LIF neuron, the initial membrane time constants were sampled from a uniform distribution, $\tau_u \sim U(0.5, 25)$, while the firing threshold was fixed to $\vartheta = 1.0$. For the SE-adLIF neuron, the initial membrane time constants were sampled from $\tau_u \sim \mathcal{U}(5, 25)$ and the adaptation time constants from $\tau_w \sim \mathcal{U}(60, 300)$. The subthreshold adaptation strength and spike-triggered adaptation strength were initialised as $a \sim \mathcal{U}(0.0, 1.0)$ and $b \sim \mathcal{U}(0.0, 1.0)$, respectively. The firing threshold was fixed to $\vartheta = 1.0$, and the adaptation scaling coefficient was set to $c_{\text{adapt}} = 120$. For the BRF neuron model, the intrinsic oscillatory dynamics are determined by the angular frequency parameter ω and the damping parameter b_{offset}. The natural frequency was initialised from $\omega \sim \mathcal{U}(3.0, 5.0)$, while the damping parameter was sampled from $b_{\text{offset}} \sim \mathcal{U}(0.1, 10.0)$. Other neuron parameters were fixed and chosen as in Baronig et al. [2026]. The S5-RF [Huber et al., 2024] was initialised with a structured state-space architecture composed of multiple stacked S5 blocks. The hidden state dimension was set to 256, and the model consists of 32 S5 blocks organised across 2 layers. The state-space discretisation follows the Dirac scheme. During training, the learning rate was scheduled with a cosine decay starting with learning rate $\eta = 0.02$. For a more detailed description of this model please refer to the original paper [Huber et al., 2024].

We introduce a synthetic benchmark for sequential state inference based on a discrete-time *Moore machine* [Gill, 1965]. The system maintains a hidden state $s_t \in \{0, \ldots, n-1\}$ that evolves deterministically according to the previous state and the current action a_t. At each timestep, the SNN observes only the action and must output the hidden state. Performance was evaluated exclusively on the final state of the sequence. Action sequences were randomly sampled for training and testing.

Table 1: **Results on the ECG task.** Comparison of recurrent and non-recurrent architectures constructed from SE-adLIF, BRF, and LIF neurons over 10 runs. Hidden-layer dimensionality was adjusted to achieve comparable parameter counts. Reported values correspond to test-set classification accuracy (mean ± standard deviation over 10 runs). A check in the column "Rec." indicates a recurrent network model while a cross indicates a parallelizeable non-recurrent model.

Model	Rec.	Layer Dim.	#Params	#Runs	Test Acc. [%]
SE-adLIF	✓	128	52,0k	10	85.27 ± 0.42
	✗	256	71,2k	10	84.29 ± 0.32
BRF	✓	128	51,2k	10	85.81 ± 0.48
	✗	256	69,6k	10	84.37 ± 0.74
LIF	✓	128	51,5k	10	78.8 ± 2.73
	✗	256	70,2k	10	82.26 ± 0.53

The initial state was fixed to $s_0 = \frac{n}{2}$ (assuming even n). State evolution is defined recursively as $s_t = f_{a_t}(s_{t-1})$, where f_{a_t} denotes the transition induced by action a_t. The available actions are *up*, *down*, *stay*, and *mirror*. We consider two variants of the task, one where the state is kept when the maximum/minimum state is exceeded ("no overflow") and one with modular arithmetic at the state boundaries ("overflow"). Their transition functions are defined as

$$f_{\text{up}}(s) = \begin{cases} (s+1) \bmod n, & \text{overflow} \\ \min(n-1, s+1), & \text{no overflow} \end{cases}$$

$$f_{\text{down}}(s) = \begin{cases} (s-1) \bmod n, & \text{overflow} \\ \max(0, s-1), & \text{no overflow} \end{cases}$$

$$f_{\text{stay}}(s) = s, \qquad f_{\text{mirror}}(s) = |s - (n-1)|.$$

The *mirror* action introduces explicit state-dependent nonlinearity, while overflow corresponds to modular arithmetic at the state boundaries. Task difficulty can be controlled via the sequence length and the inclusion of nonlinear transitions. Solving the task requires iterative state updates and therefore tests a model's ability to hold internal state.

3 Results

3.1 Testing recurrent and non-recurrent SNNs on sequential benchmark tasks

The ECG [Laguna et al., 1997] and SMNIST [Bellec et al., 2018] datasets are standard benchmarks for evaluating sequential learning models. Performance on these tasks is often used as a primary criterion for assessing the effectiveness of newly proposed architectures. The networks evaluated on the benchmark tasks consisted of two hidden layers, with the number of neurons varying by task. We considered standard recurrent SNNs consisting of LIF, SE-adLIF, or BRF neurons. For each of these networks, we also considered a parallelizeable variant without state reset and without recurrent connections (see *Methods*) denoted by the cross in the Recurrent table column. The S5-RF model was not evaluated on this task. The non-recurrent networks had double the amount of hidden neurons per layer as to offset the learnable parameter increase from the recurrent connection weights and achieve comparable parameter counts. All models received identical input representations.

ECG On the ECG dataset (Table 1), both recurrent and non-recurrent architectures achieved comparable performance, with accuracies in the 85% range. This observation is consistent with results reported for current state-of-the-art models on this task [Higuchi et al., 2024, Yin et al., 2021]. The LIF model performed the worst only reaching 78.8% in recurrent configuration. Interestingly it reached a higher accuracy of 82% as a non-recurrent network. This is unique in our results and could be further investigated. Otherwise, the performance of the recurrent SNNs performed slightly better than their parallelizable variants.

SMNIST All evaluated models performed well above chance level on the SMNIST task (Table 2), indicating that both recurrent and non-recurrent architectures are capable of learning the sequential

Table 2: **Results on the SMNIST task.** Test accuracy of recurrent and non-recurrent architectures using SE-adLIF, BRF, and LIF neuron models over ten runs (mean ± standard deviation). Hidden-layer dimensionality was adjusted to obtain comparable parameter counts across architectures. S5-RF accuracy are taken from Huber et al. [2024].

Model	Rec.	Layer Dim.	#Params	Test Acc. [%]
SE-adLIF	✓	256	202,5k	98.9 ± 0.06
	✗	512	273,9k	94.69 ± 0.31
BRF	✓	256	201,0k	98.48 ± 0.05
	✗	512	270,9k	97.92 ± 0.07
LIF	✓	256	201,5k	88.54 ± 0.87
	✗	512	271,9k	81.82 ± 0.35
S5-RF	✗	128	36.3k	98.89

digit classification problem. Nevertheless, clear performance differences emerged across architectural choices. Recurrent networks consistently achieved higher accuracy than their non-recurrent counterparts, with the recurrent SE-adLIF model reaching the best overall performance at 98.9%. Among the non-recurrent architectures, the BRF model performed strongest, achieving an accuracy of 97.92%, which is slightly lower than that of the recurrent configuration. In contrast, the non-recurrent SE-adLIF and LIF models exhibited more substantial performance degradations, particularly for LIF neurons, which showed a pronounced sensitivity to the absence of recurrence.

3.2 Testing recurrent and non-recurrent SNNs on state tracking tasks

As shown in the previous section, recurrent SNNs perform slightly better than their fully parallelisable counterparts on the ECG, and SMNIST benchmarks. Despite their widespread use, these benchmark tasks can however potentially be solved through pattern recognition over temporal input sequences. In contrast, we propose variants of a state-tracking task (see Section 2), which are in the same spirit as the well-known Shell Game, but with extended complexity. In these tasks, the network observes a sequence of actions, which manipulate a hidden state. The task for the network is to predict the final hidden state. This requires the model to maintain and update information about a hidden state, as no simple input patterns are available that could be memorized. Moreover, when the network is able to learn the effect of actions on this state, which enables generalisation across varying sequence lengths.

Such capabilities are readily observed in biological neural systems, raising the question of whether non-recurrent architectures can achieve comparable behaviour. Specifically, it was argued that these models do not represent and manipulate internal state, and their apparent success on conventional benchmarks arises from learning sufficiently rich input–output correlations that emulate memory without explicitly maintaining it [Merrill et al., 2024].

We tested the architectures on two variations of the state tracking task, each with 6 hidden states and input sequences of length 20. We classify the two variants of the state tracking tasks as *easy* (task configuration with *no mirror* and *no overflow*; see Section 2), when there are limited state dependencies, and *hard*, which incorporates the *mirror* and *overflow* mechanics.

We observed a clear difference in the training and validation accuracies between recurrent and non-recurrent architectures across task difficulty as illustrated in Table 3. Non-recurrent models were able to solve the *easy* task variants, i.e. tasks without *overflow* or *mirror* actions, achieving test accuracies comparable to those of recurrent models. The main exception was the non-recurrent LIF model, which reached only 60% test accuracy, indicating limitations even in low-complexity settings. As an additional point of comparison, we included the state-of-the-art parallelisable S5-RF network [Huber et al., 2024]. This model combines structured state-space (S5) dynamics with spiking nonlinearities, enabling efficient parallel evaluation while retaining sensitivity to temporal structure. The model has previously demonstrated strong performance on classical sequential benchmarks, and we therefore evaluated whether it could also solve the state-tracking task and whether its performance differed significantly from that of the proposed parallelisable network architectures. The model achieved good, but not top-performance on the *easy* task variant.

However, in the *hard* task variant, test accuracies of the non-recurrent models decreased to below 20%, only marginally above the chance level of 16.66%, while the recurrent models retained good

Table 3: **Accuracy comparison on different versions of the state tracking task.** The final hidden state of Moore machines with $n = 6$ hidden states and state-transition dynamics of varying difficulty had to predicted. Training and test sequences had both a length of 20. Test accuracies are shown for 10 runs (mean ± standard deviation).

Type	Model	Rec.	Layer Dim.	#Params	Test Acc. [%]
No Mirror No Overflow (easy)	SE-adLIF	✓	128	51,9k	99.83 ± 0.13
	BRF	✓	128	51,1k	97.17 ± 0.52
	LIF	✓	128	51,3k	99.07 ± 1.5
	SE-adLIF	✗	256	70,9k	91.66 ± 0.75
	BRF	✗	256	69,4k	94.78 ± 0.77
	LIF	✗	256	69,9k	60.13 ± 1.43
	S5-RF	✗	256	137,2k	96.55 ± 0.78
Mirror Overflow (hard)	SE-adLIF	✓	128	52,0k	99.98 ± 0.02
	BRF	✓	128	51,2k	89.66 ± 2.07
	LIF	✓	128	51,5k	96.09 ± 7.51
	SE-adLIF	✗	256	71,2k	18.11 ± 0.5
	BRF	✗	256	69,6k	19.96 ± 1.85
	LIF	✗	256	70,2k	18.22 ± 0.63
	S5-RF	✗	256	137,7k	21.72 ± 5.83

performance, with the SE-adLIF showing close to optimal text accuracy of 99.98%. The non-recurrent and parallelisable S5-RF model followed a similar trend to the other non-recurrent architectures. Its accuracy degraded substantially when *mirror* and *overflow* mechanics were introduced. This suggests that, despite its enhanced temporal dynamics, balanced resonance alone is insufficient to replace explicit recurrence when learning non-linear and state-dependent transition rules.

Overall, these results indicate that non-recurrent models tend to rely on memorisation of short action patterns rather than learning a generalisable state-transition function. In contrast, recurrent SNNs were able to learn also complex variants of the state-tracking task.

3.3 Testing the generalisation capabilities of recurrent and non-recurrent SNNs

Biological systems are highly adept at learning the effects of actions rather than merely internalising fixed input sequences. To assess whether the investigated SNNs truly learn action-induced state transitions, or instead rely on memorising finite input–output patterns, the test sequence length of the state-tracking task was increased from 20 to 100 input actions, while keeping the training data unchanged. If a model successfully learns the underlying effects of actions on the internal state, this increase in sequence length should not significantly affect its test accuracy. In principle, models that capture the true state-transition dynamics should be able to predict arbitrarily long sequences. Increasing the sequence length therefore primarily probes the generalisation capabilities of the models.

Our results on the easy and hard variant of the state tracking task are summarized in Table 4. We found that recurrent SE-adLIF models exhibited the strongest generalisation behaviour, with accuracy drops of only approximately 1-2% between the easy and hard task variants. The recurrent LIF model was similarly robust with a steeper drop. The BRF network was generalizing in the *easy* task variant, but failed in the *hard* variant. These results indicate that recurrent architectures are able to maintain an internal state and update it consistently in response to extended action sequences.

In contrast, non-recurrent models struggled to generalise to longer sequences. Their accuracy dropped across both task variants. In the *hard* variant, the accuracy of non-recurrent architectures consistently degraded to chance-level, suggesting that they lack the internal state memory required to reliably track the hidden state. The S5-RF model displays a similar sensitivity to increased sequence length, further supporting the conclusion that resonance-based temporal dynamics alone do not provide sufficient inductive bias for modelling long-horizon, nonlinear state transitions without explicit recurrence.

Table 4: **Generalization to increased test sequence length on Moore machine modelling accuracy.** Test accuracy for a sequence length of 100 input actions, with results for the original sequence length of 20 shown in parentheses.

Type	Model	Rec.	Layer Dim.	#Params	Test Seq.	Test Acc. [%]
No Mirror No Overflow (easy)	SE-adLIF	✓	128	51,9k	100 (20)	98.54 (99.94)
	BRF	✓	128	51,1k	100 (20)	94.11 (96.70)
	LIF	✓	128	51,3k	100 (20)	99.69 (99.88)
	SE-adLIF	✗	256	70,9k	100 (20)	76.01 (92.25)
	BRF	✗	256	69,4k	100 (20)	64.63 (96.73)
	LIF	✗	256	69,9k	100 (20)	54.41 (60.79)
	S5-RF	✗	256	137,2k	100 (20)	71.37 (96.82)
Mirror Overflow (hard)	SE-adLIF	✓	128	52,0k	100 (20)	99.24 (99.97)
	BRF	✓	128	51,2k	100 (20)	27.89 (91.36)
	LIF	✓	128	51,5k	100 (20)	93.16 (99.18)
	SE-adLIF	✗	256	71,2k	100 (20)	17.55 (19.50)
	BRF	✗	256	69,6k	100 (20)	17.76 (28.99)
	LIF	✗	256	70,2k	100 (20)	17.55 (19.13)
	S5-RF	✗	256	137,7k	100 (20)	17.03 (21.36)

4 Discussion

This work investigated the role of recurrent connections in neural networks and found a clear task-dependent distinction. On conventional sequence benchmarks (ECG, SMNIST), non-recurrent and fully parallelisable architectures achieved accuracies comparable to recurrent models, supporting the view that many standard benchmarks can be solved through pattern recognition rather than long-horizon stateful computation. In such settings, enhanced single-neuron dynamics including adaptation or resonance can approximate temporal processing without explicit recurrence. However, in Moore machine-inspired state-tracking tasks that require maintaining and updating a latent internal state according to action-conditioned rules, recurrent architectures demonstrated a consistent and substantial advantage. They remained robust under increased temporal complexity, longer sequences, and expanded state spaces, while non-recurrent models often degraded to chance-level performance. Although one recurrent variant showed a task-specific generalisation failure, other recurrent models successfully captured the required state-dependent transitions, reinforcing the conclusion that explicit recurrence provides a crucial inductive bias for learning stable, generalisable internal memory.

Mechanistically, the results support the view that recurrence enables networks to implement evolving dynamical systems rather than static input–output mappings. Recurrent models appeared to learn generalisable transition operators and maintained structured internal activity during difficult tasks, whereas non-recurrent models relied more on short-range heuristics that failed under longer horizons or complex state based actions. These findings carry implications for brain-inspired modelling: while parallelisable feed-forward SNNs offer efficiency advantages and remain competitive on pattern-based benchmarks, biologically plausible cognition as characterised by persistent activity, feedback, and latent-state inference, appears to fundamentally depend on recurrence. Thus, recurrence remains essential for modelling brain-like computation.

Acknowledgments and Disclosure of Funding

This research was funded in whole or in part by the Austrian Science Fund (FWF) [10.55776/COE12] (AM, SH, RL), and by NSF EFRI grant #2318152 (RL).

References

M. Baronig, R. Ferrand, S. Sabathiel, and R. Legenstein. Advancing spatio-temporal processing through adaptation in spiking neural networks. *Nature Communications*, 16(1), July 2025.

M. Baronig, Y. Bahariasl, O. Özdenizci, and R. Legenstein. A scalable hybrid training approach for recurrent spiking neural networks. *Neuromorphic Computing and Engineering*, 6(1):014017, 2026.

G. Bellec, D. Salaj, A. Subramoney, R. Legenstein, and W. Maass. Long short-term memory and learning-to-learn in networks of spiking neurons. *Advances in neural information processing systems*, 31, 2018.

R. J. Douglas and K. A. Martin. Recurrent neuronal circuits in the neocortex. *Current Biology*, 17 (13):R496–R500, July 2007.

J. K. Eshraghian, C. Lammie, M. R. Azghadi, and W. D. Lu. Navigating local minima in quantized spiking neural networks. In *2022 IEEE 4th International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 352–355. IEEE, 2022.

J. K. Eshraghian, M. Ward, E. O. Neftci, X. Wang, G. Lenz, G. Dwivedi, M. Bennamoun, D. S. Jeong, and W. D. Lu. Training spiking neural networks using lessons from deep learning. *Proceedings of the IEEE*, 111(9):1016–1054, 2023.

W. Fang, Z. Yu, Z. Zhou, D. Chen, Y. Chen, Z. Ma, T. Masquelier, and Y. Tian. Parallel spiking neurons with high efficiency and ability to learn long-term dependencies. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 53674–53687. Curran Associates, Inc., 2023.

W. Gerstner and W. M. Kistler. *Spiking Neuron Models: Single Neurons, Populations, Plasticity*. Cambridge University Press, Aug. 2002.

A. Gill. On the bound to the memory of a sequential machine. *IEEE Transactions on Electronic Computers*, EC-14(3):464–466, June 1965.

A. Gu, T. Dao, S. Ermon, A. Rudra, and C. Ré. Hippo: Recurrent memory with optimal polynomial projections. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1474–1487. Curran Associates, Inc., 2020.

S. Higuchi, S. Kairat, S. Bohté, and S. Otte. Balanced resonate-and-fire neurons. In R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 18305–18323. PMLR, 21–27 Jul 2024.

T. E. Huber, J. Lecomte, B. Polovnikov, and A. von Arnim. Scaling up resonate-and-fire networks for fast deep learning. In *European Conference on Computer Vision*, pages 241–258. Springer, 2024.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

P. Laguna, R. Mark, A. Goldberg, and G. Moody. A database for evaluation of algorithms for measurement of qt and other waveform intervals in the ecg. In *Computers in Cardiology 1997*, pages 673–676, 1997.

B. W. Larsen and S. Druckmann. Towards a more general understanding of the algorithmic utility of recurrent connections. *PLOS Computational Biology*, 18(6):1–33, 06 2022.

W. Maass. Networks of spiking neurons: The third generation of neural network models. *Neural Networks*, 10(9):1659–1671, Dec. 1997.

W. Merrill, J. Petty, and A. Sabharwal. The illusion of state in state-space models. *arXiv preprint arXiv:2404.08819*, 2024.

S. B. Shrestha and G. Orchard. Slayer: Spike layer error reassignment in time. *Advances in neural information processing systems*, 31, 2018.

M. S. Vidal-Saez, O. Vilarroya, and J. Garcia-Ojalvo. Biological computation through recurrence. *Biochemical and Biophysical Research Communications*, 728:150301, Oct. 2024.

B. Yin, F. Corradi, and S. M. Bohté. Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks. *Nature Machine Intelligence*, 3(10):905–913, Oct. 2021.

Effective Online SNN Training with One-Step Backpropagation

Saya Higuchi
Adaptive AI Lab
Institute for Robotics and Cognitive Systems
University of Lübeck
Ratzeburger Allee 160
23562 Lübeck, Germany
sa.higuchi@uni-luebeck.de

Federico Corradi
Neuromorphic Edge Computing Systems Lab
Department of Electrical Engineering
Eindhoven University of Technology
Groene Loper 19, Flux room 4.130
5612 AP Eindhoven, The Netherlands
f.corradi@tue.nl

Sander M. Bohté
Machine Learning Group
Centrum Wiskunde & Informatica (CWI)
Science Park 123
1098 XG Amsterdam, The Netherlands
S.M.Bohte@cwi.nl

Sebastian Otte
Adaptive AI Lab
Institute for Robotics and Cognitive Systems
University of Lübeck
Ratzeburger Allee 160
23562 Lübeck, Germany
sebastian.otte@uni-luebeck.de

Abstract

Backpropagation through time (BPTT) remains the gold-standard for training recurrent spiking neural networks, but its need to store long temporal computation graphs makes it memory-intensive and incompatible with strict online updates. This has motivated a range of alternative online learning rules, such as e-prop, further trace-based methods, and forward-only approximations, which reduce sequence-length-dependent overhead but typically require custom implementations and often sacrifice task performance. In this work, we revisit the simplest possible alternative: truncated BPTT with truncation length $k = 1$ (tBPTT_1). Although this setting is usually regarded as an overly limited-horizon baseline with poor temporal credit assignment, we show that it is a widely underestimated learning strategy. In a standard surrogate gradient learning setup, tBPTT_1 achieves performance competitive with or better than more sophisticated online learning rules. Our experiments identify two key ingredients for this result: a substantially smaller learning rate than commonly used and an optimizer with slow temporal averaging through its momentum statistics. These findings suggest that, for many practical spiking network settings, elaborate online credit-assignment rules may not be necessary: plain one-step backprop, when paired with appropriate optimization, appears as an overlooked training strategy provides effective, memory-efficient, and implementation-friendly learning.

1 Introduction

Spiking neural networks (SNNs) are well suited for event-driven computation because they process information as sparse spike trains and maintain internal states that evolve over time, making them attractive for neuromorphic sensors and low-power hardware. While backpropagation through time (BPTT) remains a strong training baseline for SNNs, it requires storing every intermediate step and updating the model in an offline manner, which conflict with both biological plausibility and memory-constrained deployment [11, 16].

The Third Austrian Symposium on AI and Vision (AIROV26).

This has motivated a range of online learning rules, the most prominent being E-prop, which approximates BPTT using eligibility traces combined with local learning signals [2]. Other approaches pursue temporally local plasticity through biologically inspired three-factor rules, such as ETLP and S-TLLR [17, 1]. In contrast, DECOLLE enables local learning through layer-wise auxiliary losses [9]. More recent methods further reduce the need for backward-through-time computation by relying on forward or target-based learning signals, as in OSTTP and traces propagation (TP) [13, 16]. These methods clearly show that online learning is feasible, but they often require custom update rules, additional auxiliary pathways, or model-specific derivations. Another form of online learning is the forward propagation through time (FPTT) [8], which introduced forward-propagated learning signals that regularize training under truncated temporal credit assignment and help stabilize recurrent spiking networks [20]. Recent works have implemented and explored truncated backpropagation through time (tBPTT), which is memory efficient but restricted in its temporal credit assignment [3, 7].

From a biological perspective, such temporally truncated gradient-based methods are better understood as local approximations than as literal models of neural learning. BPTT and related gradient-based approaches remain biologically problematic because they rely on backward credit assignment through stored trajectories, whereas biological accounts of temporal credit assignment more commonly appeal to eligibility traces, local synaptic dynamics, and modulatory learning signals [11, 6, 2]. At the same time, adaptive update rules can maintain slow hidden state across learning steps, which can be loosely interpreted through metaplastic or synaptic-dynamic views of learning [5, 18]. This perspective does not make truncated gradient methods biologically exact, but it suggests that temporally local training combined with stateful adaptive optimization may provide a useful bridge between engineered learning algorithms and biologically motivated locality constraints.

In this work, we study a simple online learning approach based on standard automatic differentiation. We apply tBPTT with fixed truncation length $k = 1$ (tBPTT_1) in combination with the Adam optimizer [10]. While this setup removes explicit temporal credit assignment, Adam implicitly aggregates information over time through its first- and second-moment estimates and stabilizes per-time-step updates via second-moment normalization. Our experiments identify two key ingredients for making this setting work in practice: the use of a substantially smaller learning rate than commonly employed, and an optimizer with slow temporal averaging through its momentum statistics. Although these statistics do not recover exact long-range gradients, they preserve useful directional and scale information across successive updates. The resulting method is straightforward to implement with standard deep-learning tools and achieves competitive performance on N-MNIST and SHD without requiring explicit eligibility traces or custom learning rules.

2 Methods

We consider feedforward (FF) and recurrent SNNs on two benchmark datasets for online learning, SHD [4] and N-MNIST [12], with input dimensions of 700 and 2,312, and time windows of 10 ms and 1 ms, respectively. For the leaky integrate-and-fire (LIF) neurons, the membrane potential is updated as $u_t = \alpha(u^{t-1} - \vartheta z^{t-1}) + I^t$, where u^t is the membrane potential, $\alpha = exp(-dt/\tau_m) \in (0, 1)$ is the fixed leakage factor, I^t denotes the total synaptic input, ϑ is the threshold. The spike output is given by $z^t = H(u^t - \vartheta)$ with $H(\cdot)$ the Heaviside step function. The double-Gaussian surrogate function [19] with FGI [14] is used. For recurrent networks, the synaptic input can be written as $I^t = W^t_{\text{in}} x^t + W^t_{\text{rec}} z^{t-1} + b$, where x^t is the external input, W^t_{in} and W^t_{rec} are input and recurrent weights, and b the optional bias.

The output layer is modeled as a leaky integrator, $o^t = \kappa o^{t-1} + W^t_{\text{out}} z_t + b_{\text{out}}$, where o^t denotes the output state, $\kappa = exp(-dt/\tau_m) \in (0, 1)$ is the fixed output leak factor, and W^t_{out} the output weights. The logits derived from o^t are used to compute a per-time-step cross-entropy loss $\ell^t = \ell(\hat{y}^t, y)$, where $\hat{y}^t$ is the prediction at time step t and y is the target label. Note that the weights have increment t due to the online update via tBPTT_1. We update the parameters and truncate the computation graph at every time step.

2.1 The optimizer as a temporal gradient accumulator and stabilizer

The parameters W are updated with the PyTorch implementation of the Adam optimizer [15, 10]. Given the instantaneous gradient $g^t = \partial \ell^t / \partial W^t$ at time step t, Adam maintains the first and second mo-

ments $m_t = \beta_1 m_{t-1} + (1-\beta_1) g^t$ and $v_t = \beta_2 v_{t-1} + (1-\beta_2) \cdot (g^t)^2$ with the bias-corrected estimates $\hat{m}_t = m_t / 1-\beta_1^t$ and $\hat{v}_t = v_t / 1-\beta_2^t$. The parameters are updated as: $W^t = W^{t-1} - \eta \hat{m}_t / \sqrt{\hat{v}_t} + \epsilon$.

3 Results

The vanilla stochastic gradient descent (SGD) was first tested to check the effect of the optimizer as a temporal gradient accumulator and stabilizer. Figure 1 shows the result of increasing momentum rate without changing any other hyperparameters. SGD without momentum only achieved $19.58 \pm 1.00\%$ whereas increasing momentum strength increased the performance of the model.

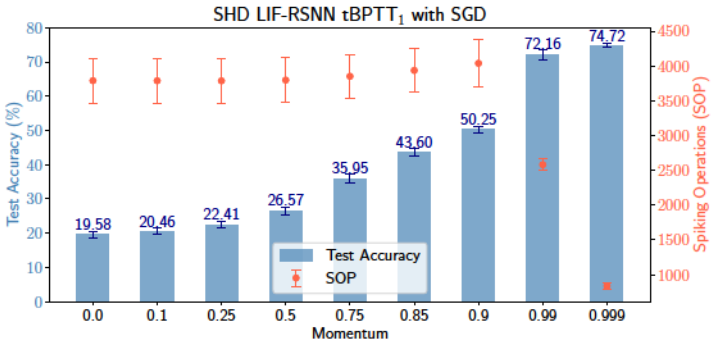


Figure 1: SHD LIF-RSNN trained with tBPTT$_1$ using the stochastic gradient descent with varying momentum term. Best test accuracy over 5 runs.

Furthermore, a substantial reduction in the average spiking operation (SOP) can be observed, indicating that the models learn sparser activity patterns. This reduction accompanies the performance gains obtained with stronger temporal accumulation in the optimizer.

Table 1 places the proposed method in the context of existing online and local-learning approaches. On N-MNIST, tBPTT$_1$ reaches $97.43 \pm 0.15\%$, outperforming DECOLLE and TP while remaining below full BPTT and e-prop. On feedforward SHD, it achieves $68.19 \pm 1.12\%$, slightly exceeding TP ($67.06 \pm 0.96\%$) and clearly outperforming e-prop and ETLP, although still trailing offline BPTT. The strongest result appears on recurrent SHD, where tBPTT$_1$ reaches $84.29 \pm 0.98\%$, outperforming all compared online methods and even slightly surpassing full BPTT ($83.23 \pm 1.00\%$). Overall, these results show that even under the extreme truncation condition $k = 1$, a carefully tuned one-step training setup, combining a small learning rate with slow temporal gradient filtering, remains highly competitive, especially in recurrent settings where temporal structure is most relevant.

4 Discussion

Our results highlight an apparent contradiction: learning remains strong even when temporal credit assignment is reduced to a single step. Because the graph is truncated at every step, the method does not explicitly recover long-range temporal Jacobian chains as in BPTT or cell-to-cell temporal dependencies as in e-prop. The only information available for learning is the stream of adapting local gradients $\{g_t\}_{t=1}^T$. A naive expectation would therefore be that performance collapses when training becomes fully online. Our results show that this need not be the case when the optimizer is allowed to accumulate and rescale these local gradients across time.

The central idea of this paper is to interpret (m_t, v_t) as a compact memory of the recent gradient history, as well as a stabilizer for per step updates. In other online learning methods, temporal information is stored explicitly in eligibility traces or other neuron-specific state variables [2, 17]. Here, by contrast, part of that temporal accumulation is carried by a slowly evolving optimizer state while learning remains stable. The first-moment term m^t smooths the direction of successive local gradients, while the second-moment term v^t normalizes their scale and reduces sensitivity to sharp fluctuations.

Table 1: Comparison to SoTA results on N-MNIST and SHD. Our results for N-MNIST and SHD report average best test accuracy over 5 runs and average highest test accuracy over 5 runs, respectively.

Model	Architecture Type	Neuron Type	Number of Neurons	Local Learning	Time Steps	Test Accuracy
			N-MNIST			
BPTT	FF	LIF	200	✗	100	98.45 ± 0.04[1]
eProp [2]	FF	LIF	200	Partial (time)	100	**97.90**[2]
DECOLLE [9]	FF	LIF	200	✓	100	96.27[2]
TP [16]	FF	LIF	200	✓	100	97.33 ± 0.06
$tBPTT_1$ (ours)	FF	LIF	200	Partial (time)	100	97.43 ± 0.15
			SHD			
BPTT	FF	LIF	450	✗	100	75.85 ± 0.48[1]
eProp [2]	FF	LIF	450	Partial (time)	100	63.04[2]
ETLP [17]	FF	ALIF	450	✓	100	59.19
TP [16]	FF	LIF	400	✓	100	67.06 ± 0.96
$tBPTT_1$ (ours)	FF	LIF	450	Partial (time)	100	$\mathbf{68.19 \pm 1.12}$
BPTT	Recurrent	LIF	450	✗	100	83.23 ± 1.00[1]
S-TLLR [1]	Recurrent	LIF	450	Partial (time)	100	78.24 ± 1.84
eProp [2]	Recurrent	LIF	450	Partial (time)	100	80.79[2]
ETLP [17]	Recurrent	ALIF	450	✓	100	74.59
TP [16]	Recurrent	LIF	450	✓	100	81.80 ± 0.51
$tBPTT_1$ (ours)	Recurrent	LIF	450	Partial (time)	100	$\mathbf{84.29 \pm 0.98}$

[1] Results from [16]. [2] Results from [17].

This does not reconstruct exact long-range credit assignment, since dependencies through earlier hidden states remain absent, but it does provide a simple mechanism through which past gradients continue to influence future updates. In this sense, our findings also clarify the relation to FPTT [8, 20]: rather than enforcing temporal consistency through an explicit regularizer across successive updates, much of the same stabilizing effect appears to arise here from a small learning rate together with slow temporal filtering in the optimizer itself, without auxiliary regularizers or extra buffer variables.

From an implementation perspective, this is appealing because it requires no custom backward rule beyond standard surrogate-gradient differentiation. The method can be realized with ordinary autograd, one forward pass and one backward pass per time step, and a standard optimizer call. As a consequence, it offers a lightweight baseline for online SNN learning and a useful reference point for understanding how much temporal structure can already be captured by optimizer dynamics alone.

5 Conclusion

We revisited one-step truncated BPTT as a minimal online training strategy for SNNs and found that, despite its extreme temporal truncation, it can remain highly effective for standard surrogate-gradient training when combined with careful gradient step scaling and slow temporal filtering. Across N-MNIST and SHD, $tBPTT_1$ achieved competitive performance against more specialized online learning methods, and on recurrent SHD it even slightly surpassed full BPTT. Our results show that this behavior depends critically on two ingredients: a substantially smaller learning rate than commonly used and optimizer dynamics that accumulate and rescale local gradients over time through slow momentum statistics. Taken together, these findings suggest that effective online SNN learning does not necessarily require elaborate custom credit-assignment rules, and that plain one-step backpropagation, when combined with appropriate optimization, provides a simple, memory-efficient, and strong baseline for future work.

References

[1] Marco Paul E Apolinario and Kaushik Roy. S-tllr: Stdp-inspired temporal local learning rule for spiking neural networks. *Transactions on Machine Learning Research*.

[2] Guillaume Bellec, Franz Scherr, Anand Subramoney, Elias Hajek, Darjan Salaj, Robert Legenstein, and Wolfgang Maass. A solution to the learning dilemma for recurrent networks of spiking neurons. *Nature communications*, 11(1):3625, 2020.

[3] Hojae Choi, Jaewook Kim, Jongkil Park, Seongsik Park, Hyun Jae Jang, Seung Hwan Lee, Byeong-Kwon Ju, and YeonJoo Jeong. Star-snn: A spatio-temporal adaptive recurrent spiking neural network with separated propagation surrogate gradient for hardware efficient real-time learning. *Neurocomputing*, page 132968, 2026.

[4] Benjamin Cramer, Yannik Stradmann, Johannes Schemmel, and Friedemann Zenke. The heidelberg spiking data sets for the systematic evaluation of spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2744–2757, 2020.

[5] Shiva Farashahi, Christopher H Donahue, Peyman Khorsand, Hyojung Seo, Daeyeol Lee, and Alireza Soltani. Metaplasticity as a neural substrate for adaptive learning and choice under uncertainty. *Neuron*, 94(2):401–414, 2017.

[6] Wulfram Gerstner, Marco Lehmann, Vasiliki Liakoni, Dane Corneil, and Johanni Brea. Eligibility traces and plasticity on behavioral time scales: experimental support of neohebbian three-factor learning rules. *Frontiers in neural circuits*, 12:53, 2018.

[7] Stijn Groenen, Marzieh Hassanshahi Varposhti, and Mahyar Shahsavari. Gazescrnn: Event-based near-eye gaze tracking using a spiking neural network. *arXiv preprint arXiv:2503.16012*, 2025.

[8] Anil Kag and Venkatesh Saligrama. Training recurrent neural networks via forward propagation through time. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5189–5200. PMLR, 18–24 Jul 2021.

[9] Jacques Kaiser, Hesham Mostafa, and Emre Neftci. Synaptic plasticity dynamics for deep continuous local learning (decolle). *Frontiers in Neuroscience*, 14:424, 2020.

[10] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[11] Timothy P Lillicrap and Adam Santoro. Backpropagation through time and the brain. *Current opinion in neurobiology*, 55:82–89, 2019.

[12] Garrick Orchard, Ajinkya Jayawant, Gregory K Cohen, and Nitish Thakor. Converting static image datasets to spiking neuromorphic datasets using saccades. *Frontiers in neuroscience*, 9:437, 2015.

[13] Thomas Ortner, Lorenzo Pes, Joris Gentinetta, Charlotte Frenkel, and Angeliki Pantazi. Online spatio-temporal learning with target projection. In *5th IEEE International Conference on Artificial Intelligence Circuits and Systems, AICAS 2023*. IEEE, 2023.

[14] Sebastian Otte. Flexible and efficient surrogate gradient modeling with forward gradient injection. In *Proceedings of Austrian Symposium on AI, Robotics, and Vision 2024*, pages 451–459. University of Innsbruck, 2024.

[15] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. In *NIPS-W*, 2017.

[16] Lorenzo Pes, Bojian Yin, Sander Stuijk, and Federico Corradi. Traces propagation: memory-efficient and scalable forward-only learning in spiking neural networks. *Neuromorphic Computing and Engineering*, 6(1):014002, 2026.

[17] Fernando M Quintana, Fernando Perez-Peña, Pedro L Galindo, Emre O Neftci, Elisabetta Chicca, and Lyes Khacef. Etlp: event-based three-factor local plasticity for online learning with neuromorphic hardware. *Neuromorphic Computing and Engineering*, 4(3):034006, 2024.

[18] Yukun Yang and Peng Li. Synaptic dynamics realize first-order adaptive learning and weight symmetry. *arXiv preprint arXiv:2212.09440*, 2022.

[19] Bojian Yin, Federico Corradi, and Sander M Bohté. Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks. *Nature Machine Intelligence*, 3(10):905–913, 2021.

[20] Bojian Yin, Federico Corradi, and Sander M Bohté. Accurate online training of dynamical spiking neural networks through forward propagation through time. *Nature Machine Intelligence*, 5(5):518–527, 2023.

Probabilistic LIF Neurons Improve Learning in Recurrent Spiking Neural Networks

Sebastian Higuchi
Adaptive AI Lab
Institute for Robotics and Cognitive Systems
University of Lübeck
Ratzeburger Allee 160
23562 Lübeck, Germany
sebastian.higuchi@uni-luebeck.de

Niels A. Kloosterman
Adaptive Brain and Cognition Lab
Department of Psychology
University of Lübeck
Maria Goeppert-Straße 9a
23562 Lübeck, Germany
n.kloosterman@uni-luebeck.de

Stefan Hallermann
Carl-Ludwig-Institute of Physiology
Medical Faculty
University Leipzig
Liebigstraße. 27
04103 Leipzig, Germany
hallermann@medizin.uni-leipzig.de

Sebastian Otte
Adaptive AI Lab
Institute for Robotics and Cognitive Systems
University of Lübeck
Ratzeburger Allee 160
23562 Lübeck, Germany
sebastian.otte@uni-luebeck.de

Abstract

Training recurrent spiking neural networks (SNNs) with leaky integrate-and-fire (LIF) neurons is often slow, particularly during the early phase, when networks must first establish sufficient spike activity to form patterns. Strategies such as low firing thresholds or high-magnitude weight initialization can increase early spiking, but typically introduce instabilities and impair learning. Here we introduce a modification of classical LIF and parameterized LIF (PLIF) neurons, in which spikes are generated probabilistically, including a proper surrogate gradient formulation. The membrane potential parameterizes the instantaneous spike probability, and spikes are sampled as Bernoulli variables at each time step, whereas underlying LIF membrane dynamics remain unchanged. This stochastic activation stabilizes early spike activity and substantially accelerates learning. In two benchmark tasks, these probabilistic LIF networks surprisingly achieve substantially higher classification accuracy than its deterministic LIF baselines. These findings suggest that probabilistic spike generation may provide a promising new perspective for building compact and effective spiking architectures.

1 Introduction

Spiking neural networks (SNNs) based on leaky integrate-and-fire (LIF) neurons are widely studied as a promising framework for energy-efficient and neuromorphic machine learning. Their event-driven computation, temporal dynamics, and compatibility with neuromorphic hardware make them an attractive alternative to conventional artificial neural networks. However, despite significant progress in surrogate gradient methods that enable gradient-based training, learning in SNNs often remains considerably slower and less stable than in conventional deep networks. One persistent challenge lies in the early stages of training, where the network must first establish sufficient spiking activity to propagate information through the system.

The Third Austrian Symposium on AI and Vision (AIROV26).

In deterministic LIF networks, meaningful learning only occurs once spike activity emerges and begins to transmit signals across layers. During the initial phase of training, however, membrane potentials often remain below threshold for extended periods, leading to rare spike activity. As a result, gradient flow vanishes almost entirely, slowing the formation of meaningful representations. A simple workaround is to initialize synaptic weights with unusually large magnitudes or to reduce neuronal thresholds in order to provoke early spikes. These strategies increase activity, but they frequently introduce instability or highly irregular dynamics that hinder the formation of structured network representations.

Stochasticity and noise have long been considered computational resources in networks of spiking neurons, supporting computation, inference, and learning in such systems [5]. Gradient estimators for stochastic binary neurons have been studied in [1], and stochastic neuron implementations have also been demonstrated directly in neuromorphic hardware [7].

Based on biological evidence, we hypothesize that introducing spontaneous spiking into spiking neural networks may promote flexible state exploration during learning, improving adaptation and task performance. To this end, we introduce a stochastic modification of LIF neurons. Instead of generating spikes deterministically when the membrane potential crosses a threshold, we interpret a nonlinear transformation of the membrane potential as the instantaneous spike *probability*. At each simulation step, the spike output is then sampled as a Bernoulli random trial with this probability. Importantly, all underlying LIF membrane dynamics remain unchanged.

This modification provides a controlled mechanism to induce spike activity during the early stages of training. Even when membrane potentials remain below classical threshold levels, neurons still emit spikes, allowing signals to propagate through the network and enabling gradients to shape synaptic structure. As learning progresses, the network gradually organizes its activity patterns and synaptic weights while maintaining stable dynamics.

We evaluate probabilistic LIF networks on two benchmark classification tasks and compare them to the conventional deterministic LIF baseline.

2 Methods

Among the simplest and most commonly used neuron models in SNNs are the leaky integrate-and-fire (LIF) neuron [4] and its derived variants such as the parametric LIF (PLIF) neuron [3]. Nevertheless, their deterministic threshold mechanism can make optimization difficult in the early phase of learning. Membrane potentials remain below threshold, neurons stay silent, spike-based signal propagation is weak, and learning can be substantially delayed.

To overcome this problem, we introduce a minimal probabilistic extension of LIF neurons that we name PropLIF and PropPLIF, respectively. The central idea is to increase spiking by probabilistic spikes, while membrane dynamics remain unchanged. Neurons thus emit spontaneous spikes in the subthreshold regime, thereby supporting early activity propagation while preserving the simplicity of classical LIF dynamics.

2.1 Deterministic LIF dynamics

The membrane potential in LIF neurons evolves according to $u_t = \alpha u_{t-1} + (1-\alpha)I_t - \theta z_{t-1}$, where u_t denotes the membrane potential at time step t, I_t is the synaptic input current, θ is the firing threshold, and z_{t-1} is the spike emitted at the previous time step. The leakage factor is defined as $\alpha = \exp(-dt/\tau_m)$, with simulation time step dt and membrane time constant τ_m. Whenever a spike occurs, the membrane potential is reset by subtracting the threshold term.

In the standard LIF model, the membrane time constant τ_m is fixed. In the PLIF model, τ_m is learned jointly with the synaptic weights [3]. Apart from this, both models share the same membrane dynamics.

2.2 Probabilistic spike generation

Conventionally, spike generation is deterministic, and occurs only when the membrane potential exceeds threshold θ. We extend this mechanism by a probabilistic firing rule.

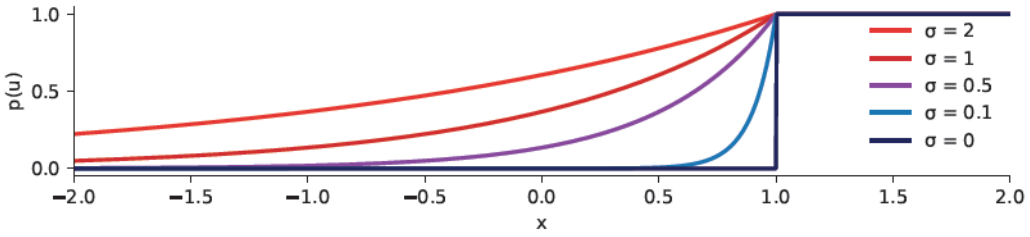


Figure 1: Probability function $p(u) = \exp(\min(0, \frac{u-\theta}{\sigma\theta}))$ for $\theta = 1$ and different values of σ. Larger σ broadens the probabilistic region below the threshold, while $s = 0$ corresponds to the limiting deterministic step-function. Since the resting potential of u is zero, there is a nonzero chance the neuron spikes while resting.

Given membrane potential u_t, we define the instantaneous spike probability as

$$p(u_t) = \exp\left(\min\left(0, -\frac{u_t - \theta}{\sigma\theta}\right)\right), \tag{1}$$

where $\sigma > 0$ controls the width of the probabilistic region below threshold. For $u_t \geq \theta$, the firing probability is 1, whereas for $u_t < \theta$ it decreases exponentially with distance from threshold. Larger values of σ broaden the subthreshold region where spontaneous spikes occur. See Figure 1 for reference.

A spike at time step t is then sampled as $z_t^{(p)} \sim \text{Bernoulli}(p(u_t))$. This formulation preserves the standard membrane update, but extends hard thresholding by a stochastic activation mechanism. As a result, neurons become active even when their membrane potentials are below the deterministic firing threshold, which helps to establish early spike propagation during learning. Figure 2 visualizes this increase in activity.

2.3 Surrogate-gradient formulation

Bernoulli sampling is not differentiable, thus training requires a surrogate gradient approximation. The forward pass uses sampled spikes, while the backward pass propagates gradients through a surrogate function [6]:

$$m_t = u_t \, \text{sg}(g(u_t)) \tag{2}$$

$$z_t = m_t - \text{sg}(m_t) + z_t^{(p)} \tag{3}$$

where $z_t^{(p)} \sim \text{Bernoulli}(p(u_t))$, g is a surrogate gradient function, and $\text{sg}(\cdot)$ denotes the stop-gradient operator.

For the probabilistic neurons as well as the deterministic threshold-based baselines, we use the same surrogate gradient framework as in prior work and approximate the derivative of the Heaviside step function with a double-Gaussian surrogate [8]. In case of probabilistic neurons, we additionally mix this surrogate gradient function with $\frac{d}{du_t}p(u_t)$ to provide better gradients for large negative membrane potentials.

3 Results

We evaluated our PropLIF and PropPLIF neurons within recurrent SNNs on the Sequential-MNIST (S-MNIST) and the Spiking Heidelberg Digits (SHD) benchmark. The models use recurrent architectures with a single hidden layer of probabilistic spiking neurons and a leaky integrator readout. For S-MNIST, we used a network of size $(1, 256^{\text{R}}, 10)$, and for SHD, a network of size $(700, 128^{R}, 20$, where $\cdot^{\text{R}}$ denotes the recurrent hidden layer. All models are trained with backpropagation through time using a negative log-likelihood objective and selected by early stopping on the validation loss. In the probabilistic variants, spike emission is controlled by the parameter σ, which determines the width of the subthreshold probabilistic firing regime; the values used in Table 1 are dataset- and model-specific.

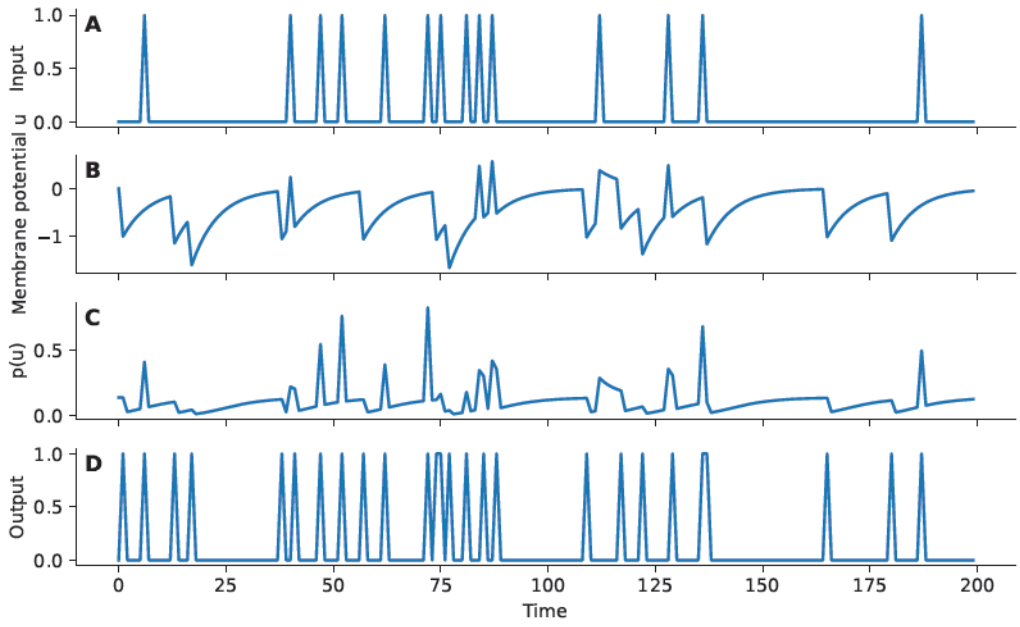


Figure 2: Input response of a ProbLIF neuron with $\sigma = 0.5$ over time. A: Random input. B: The membrane potential u evolves based on the Input and is reduced by $\theta = 1$ whenever an output spike occurs. C: Probability of a spike occurring. D: Stochastic spiking activity of the neuron.

4 Conclusion

We introduced probabilistic extensions of LIF and PLIF neurons—PropLIF and PropPLIF—which preserve classical LIF membrane dynamics while replacing deterministic thresholding with stochastic spike generation. This modification consistently improved learning, yielding faster convergence and higher accuracy, in some cases with substantially fewer parameters. By allowing spontaneous subthreshold spikes, the proposed neurons avoid the silent early-training regime of deterministic SNNs and sustain activity during learning.

Overall, probabilistic spike generation provides a promising new perspective for compact and effective SNNs. Future work should test this principle in other neuron models and explore its broader role as a training mechanism for spiking networks.

Table 1: Comparison of model performance on S-MNIST and SHD datasets. Here, * denotes our reproduced results using publicly available code, and ^{R} denotes a fully recurrent layer. All other layers are exclusively feedforward.

Dataset	Method	Neurons	Parameters (k)	Accuracy (%)
S-MNIST	PLIF [3]	1,64,256^{R},256,10	112.2/155.1	90.93/91.79
	LIF [9]	1,64,256^{R},256,10	112.2/155.1	74.91/89.28
	ProbLIF ($\sigma = 1.2$) (ours)	1,256^{R},10	68.36	95.78
	ProbPLIF ($\sigma = 1$) (ours)	1,256^{R},10	68.62	**97.09**
SHD	LIF [2]	700,128^{R},20	108.80	71.40
	PLIF*	700,128^{R},20	108.69	76.15
	ProbLIF ($\sigma = 1$) (ours)	700,128^{R},20	108.54	79.99
	ProbPLIF ($\sigma = 1.3$) (ours)	700,128^{R},20	108.69	**87.72**

References

[1] Yoshua Bengio, Nicholas Léonard, and Aaron C. Courville. Estimating or propagating gradients through stochastic neurons for conditional computation. *CoRR*, abs/1308.3432, 2013.

[2] Benjamin Cramer, Yannik Stradmann, Johannes Schemmel, and Friedemann Zenke. The heidelberg spiking data sets for the systematic evaluation of spiking neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 33(7):2744–2757, 2020.

[3] Wei Fang, Zhaofei Yu, Yanqi Chen, Timothee Masquelier, Tiejun Huang, and Yonghong Tian. Incorporating learnable membrane time constant to enhance learning of spiking neural networks, 2021.

[4] Wulfram Gerstner and Werner M Kistler. *Spiking neuron models: Single neurons, populations, plasticity*. Cambridge university press, 2002.

[5] Wolfgang Maass. Noise as a resource for computation and learning in networks of spiking neurons. *Proceedings of the IEEE*, 102(5):860–880, 2014.

[6] Sebastian Otte. Flexible and efficient surrogate gradient modeling with forward gradient injection. In *First Austrian Symposium on AI, Robotics, and Vision*. innsbruck university press, 2024.

[7] Tomas Tuma, Angeliki Pantazi, Manuel Le Gallo, Abu Sebastian, and Evangelos Eleftheriou. Stochastic phase-change neurons. *Nature Nanotechnology*, 11:693–699, 2016.

[8] Bojian Yin, Federico Corradi, and Sander M Bohté. Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks. *Nature Machine Intelligence*, 3(10):905–913, 2021.

[9] Shimin Zhang, Qu Yang, Chenxiang Ma, Jibin Wu, Haizhou Li, and Kay Chen Tan. Tc-lif: A two-compartment spiking neuron model for long-term sequential modelling, 2024.

Certification and Trustworthy AI

Stochastic Application Domain Definition for Functional Trustworthiness Certification of AI Systems

Simon Schmid[1,3] Barbara Brune[2] Alexander Aufreiter[1] Lukas Gruber[3]
Kajetan Schweighofer[3] Xaver Stadlbauer[2] Thomas Doms[2]
Bernhard Nessler[1]

[1]Software Competence Center Hagenberg [2]TÜV Austria Data Intelligence GmbH
[3]Johannes Kepler Universität Linz

Abstract

As Artificial Intelligence (AI) systems are increasingly deployed in safety-critical and societally consequential contexts, the question of how to evaluate their performance in a trustworthy and interpretable manner becomes increasingly important. Within the European Union, this issue is reflected in the AI Act, which requires training, validation, and testing datasets to be relevant and sufficiently representative with respect to the system's intended purpose. This raises a fundamental technical question: representative of what population of situations?

From a statistical perspective, performance metrics such as error rates or expected losses are always defined with respect to a probability distribution. We refer to this distribution as the Application Domain (AD). In practice, however, the AD of real-world AI systems is rarely known in explicit mathematical form and must instead be characterized operationally through the procedures by which valid samples are generated or selected.

To address this problem, we introduce the Stochastic Application Domain Definition (SADD), a textual specification of the sampling protocol that induces the distribution under which performance claims are intended to hold. The SADD links intended use, dataset construction, and statistical testing by making the underlying data-generation assumptions explicit. We formalize the notion of protocol-induced distributions, discuss how SADDs guide feasible sampling procedures, contrast the approach with qualitative domain descriptions such as Operational Design Domains, and examine implications for the certification of AI systems.

1 Introduction

Artificial Intelligence (AI) has become an integral part of many technical and societal domains. Progress in machine learning (ML) and deep learning has been driven by large-scale datasets, increased computational power, and advances in model architectures, leading to major scientific and industrial breakthroughs in recent years (27; 7; 14; 24). At the same time, the deployment of ML systems in real-world settings introduces risks that differ from those of traditional software. ML systems are increasingly used in contexts affecting health, safety, economic participation, and fundamental rights, making the question of their trustworthiness a central concern in research, regulation, standardization, and certification (4; 22; 23; 29; 16; 8; 13).

Within the European Union, these concerns are reflected in the AI Act, which introduces obligations regarding the governance and quality of training, validation, and testing data. In particular, Article 10

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

requires datasets to be relevant and sufficiently representative with respect to the intended purpose of the system (1). This requirement raises a fundamental technical question: representative of what population of situations?

From a statistical perspective, performance metrics such as error rates or expected losses are always defined with respect to a probability distribution. Reported model performance is therefore meaningful only relative to a well-defined distribution of inputs and outputs. In this paper, we refer to this distribution as the Application Domain (AD) of the system. For real-world ML systems, however, the AD is rarely available in explicit mathematical form. Instead, it must be characterized operationally through the procedures by which valid samples are generated or selected.

To address this problem, we introduce the concept of a Stochastic Application Domain Definition (SADD). A SADD is a textual specification of the sampling protocol that induces the distribution under which performance claims are intended to hold. In this way, the SADD links intended use, dataset construction, and statistical evaluation, providing a basis for interpretable testing and independent verification.

A simple example illustrates the underlying issue. Consider a kitchen knife designed for cutting vegetables. Suppose the knife performs well in tests on broccoli, apples, and melons. Even so, this result does not yet specify under which conditions the claim should be trusted. Performance may differ between private and commercial kitchens, between types of produce, or across different usage conditions. The same principle applies to ML systems: a performance claim is meaningful only with respect to the population of situations under which the system is evaluated.

In ML, this issue is particularly important because testing relies on empirical estimates of expected performance. Such estimates are informative only if the test samples can be understood as draws from a well-defined domain of application (5). Different data-generation procedures may therefore lead to different performance estimates even for the same system and nominal use case. The central question is therefore not only which metric to report, but also how the relevant domain of evaluation is defined. This is precisely the role of the Application Domain and, in the framework proposed here, of the SADD.

1.1 Functional trustworthiness

For the purpose of functional trustworthiness assessment, three elements are essential (22):

(1) a clear **Stochastic Application Domain Definition**, which specifies the domain under which the system is intended to operate and under which its performance claims are to be understood;

(2) **risk-based minimum performance requirements**, which define which performance quantities matter and what levels are acceptable in view of the risks of the intended application and relevant foreseeable misuse scenarios; and

(3) **statistically valid testing based on independent random samples**, which enables performance estimation under the application domain defined by the SADD.

This paper focuses on the first of these elements. Our claim is that, without a sufficiently explicit definition of the application domain, even a well-designed statistical test cannot yield a well-interpretable guarantee.

2 Motivation and problem statement

The challenge of defining the application domain is closely linked to a fundamental difference between traditional software systems and ML systems. In traditional software engineering, the intended functionality is typically specified in formal or highly structured terms. Developers encode rules and procedures to solve tasks whose input–output relations are sufficiently well understood, such as sorting algorithms, database queries, or cryptographic methods. In such cases, the domain of valid inputs and the desired behaviour can often be described precisely.

ML systems differ in that the input–output relation is not explicitly programmed but learned from data. Model behaviour emerges from optimization on examples rather than from a fully specified

rule set, and the training objective may differ from the ultimate task-level performance criterion.[1] As a result, the operational domain under which the system is expected to perform is often left implicit, and performance claims may become detached from the conditions of actual use.

This issue can be expressed using the standard notion of statistical risk in machine learning. Let f denote a model, L a loss function, and let $(X, Y) \sim P$ denote the distribution of inputs and outputs under which the system is evaluated. The statistical risk is defined as

$$R(f) = \mathbb{E}_{(X,Y)\sim P}[L(Y, f(X))]. \tag{1}$$

Most learning algorithms estimate this quantity indirectly by minimizing the empirical risk, i.e., the average loss on a finite dataset.

The key observation is that the risk in (1) is defined only with respect to the distribution P. Reported model performance is therefore meaningful only relative to a specified distribution. If training, testing, and deployment occur under different distributions, performance guarantees may not reflect the behaviour of the system in its intended application setting.

In practice, however, the relevant distribution P is rarely known in explicit mathematical form. Real-world data-generating processes are complex and shaped by institutional, temporal, geographical, and technical constraints. Consequently, the application domain of an ML system typically cannot be specified through a complete probabilistic model. Instead, it must be characterized operationally through the procedure by which valid samples are generated or selected. In many empirical disciplines, the target of inference is defined in exactly this way: not through an explicit distribution, but through a documented sampling or study protocol (19; 11; 25; 21).

3 The Stochastic Application Domain Definition

3.1 Application domains as probability distributions

Performance guarantees for ML systems are statistical statements. Whether one reports an error rate, a sensitivity, a calibration quantity, or an expected loss, the reported number always refers—implicitly or explicitly—to a distribution over relevant situations. In this sense, the application domain of an ML system is the probability distribution under which its performance is to be interpreted.

This perspective is not optional. If the relevant distribution changes, then the meaning of the reported performance changes as well. A system can perform excellently under one distribution and poorly under another, even if the underlying task appears similar at a superficial level.

Instead of specifying this distribution analytically, we characterize it operationally through the procedure by which valid samples are generated or selected. This motivates the notion of a protocol-induced distribution.

3.2 Protocol-induced distributions

Let $\mathcal{X}$ denote the space of possible inputs. A practical sampling procedure can be understood as a rule that uses randomness, together with operational constraints, to generate samples from $\mathcal{X}$. Such a procedure induces a probability distribution on the samples that can occur under that protocol.

Formally, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space representing the randomness involved in the sampling process, and let

$$g : \Omega \to \mathcal{X} \tag{2}$$

be a measurable mapping that transforms this randomness into an observable sample. The protocol then induces a probability distribution on $\mathcal{X}$ via the pushforward measure

$$P_{\Pi}(A) = \mathbb{P}\big(g^{-1}(A)\big) \tag{3}$$

[1] For example, cross-entropy loss may be optimized during training while classification error is used for evaluation; perplexity may be optimized although question answering quality is of ultimate interest.

for all measurable sets $A \subseteq \mathcal{X}$. We call P_Π the **protocol-induced distribution** and identify the application domain with such a distribution.

This perspective implies that the application domain is not simply "the real world" or a vague intended-use statement, but the distribution induced by a specified sampling protocol in a given operational setting. If different parties generate evaluation data using different protocols, they generally induce different distributions, even if they believe they are evaluating the same use case. Reported performance differences may therefore arise not from changes in the model, but from changes in the domain of evaluation.

For certification and independent verification, this is crucial. A performance claim becomes interpretable only if the underlying data-generation procedure is described clearly enough that another qualified party can reconstruct the relevant distribution to a practically sufficient extent.

3.3 Definition of the SADD

Since the relevant application domain is typically not available as an explicit mathematical object, it must be communicated in another way. We propose to do so through a textual, process-oriented description of the protocol that induces the domain.

We call this description the **Stochastic Application Domain Definition (SADD)**.

The SADD is therefore not only a statement of intended use, nor just a list of environmental conditions, nor a benchmark description. It is a **textual specification of the sampling protocol** that defines what counts as a valid draw from the application domain. Its purpose is to make explicit, in operational terms, the process through which samples relevant for testing are to be generated or selected.

A SADD should describe, as far as relevant for the intended use,

- the real-world process or objects that give rise to the data,
- inclusion and exclusion criteria,
- the geographical and temporal scope,
- the acquisition setting,
- relevant actors, devices, and procedural constraints, and
- the rules according to which valid samples are selected or generated.

Once the application domain is understood as a protocol-induced distribution P_Π, performance quantities can be defined in the standard statistical way. For a given evaluation functional φ, one may write

$$\theta = \mathbb{E}_{(X,Y)\sim P_\Pi}[\varphi(X, Y, f)].$$

In practice, θ is estimated using samples drawn according to the sampling procedure described by the SADD. The interpretability of the resulting estimate therefore depends critically on the adequacy of that procedure.

3.4 Interpretation of a SADD

A SADD is a textual document and therefore necessarily admits interpretation. Different stakeholders may read the same SADD against different backgrounds and with different practical assumptions. The goal of a good SADD is not to eliminate all interpretation—which would be unrealistic—but to reduce ambiguity sufficiently for communication, testing, and certification purposes.

Figure 1a illustrates the individual perspective: a reader forms a subjective understanding of the domain under which the system's performance claims are to be interpreted. This already has practical value, because the SADD informs developers, distributors, auditors, and users about the realm of intended application.

For certification purposes, however, a purely subjective reading is not sufficient. We therefore adopt a normative interpretation principle based on the notion of a **reasonably informed stakeholder**.

(a) Single user's interpretation (b) Normative interpretation

Figure 1: Interpreting the SADD. (a) A single informed user forms an individual understanding of the application domain described by the SADD. (b) For certification purposes, interpretation should be anchored in the understanding of a reasonably informed stakeholder community.

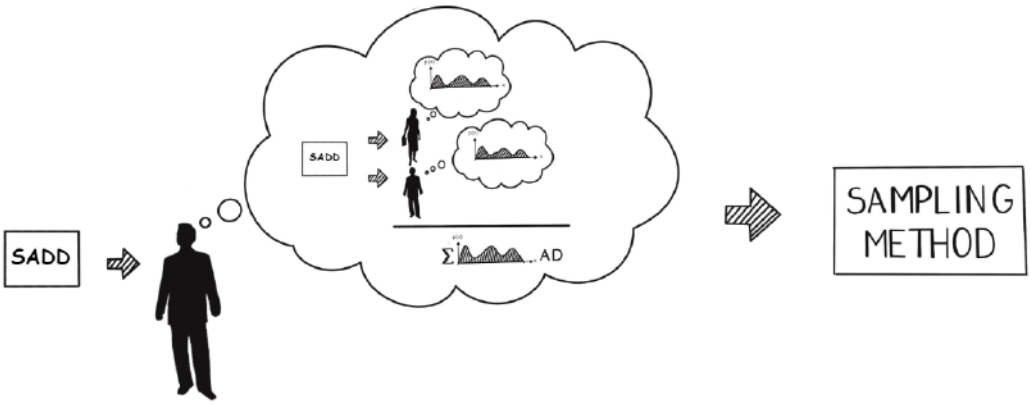


Figure 2: Deriving an operational sampling procedure from the SADD. An expert interprets the textual SADD while considering the reasonably expected understanding of qualified stakeholders and implements a feasible sampling procedure that approximates the intended application domain.

This notion is inspired by established approaches in legal and technical communication, where documents are interpreted not according to an arbitrary individual reading, but according to the understanding that can reasonably be expected from a qualified addressee in the relevant field.

In our context, the reasonably informed stakeholder is not a specific person but a reference construct. It represents the level of domain-specific understanding that can reasonably be expected from a competent actor in the intended field of application. The purpose of this construct is not to eliminate all variability in interpretation, but to provide a common reference point for communication and certification.

3.5 From SADD to sampling procedures

The practical purpose of a SADD is to describe how samples belonging to the application domain can arise. In doing so, it specifies which inputs are admissible members of the domain and outlines the conditions under which such samples occur in practice. In this sense, the SADD delineates a meaningful region within the space of all theoretically possible inputs to the AI system and describes a conceptual generative process for producing samples from this region.

In practice, however, the textual description of the SADD cannot usually be implemented literally. Real-world constraints such as limited data access, legal restrictions, costs, or geographic limitations

may prevent direct sampling from the idealized process described by the SADD. Therefore, the SADD must be translated into a feasible sampling procedure that approximates the protocol-induced distribution implied by the textual specification.

This translation is carried out by a domain expert who interprets the SADD while taking into account the reasonably expected understanding of the relevant stakeholder community (see Fig. 2). The resulting sampling procedure serves as an operational implementation of the SADD and provides the concrete mechanism for generating evaluation data used in statistical testing.

3.6 Sampling strategies

Once a SADD has specified the conditions under which valid samples arise, a sampling design is needed to generate evaluation data. Its role is to approximate the protocol-induced distribution defined by the SADD as closely as practical.

This is closely related to classical sampling theory (15; 19). Possible designs include simple random, cluster, and stratified sampling. Since different designs induce different effective distributions, they also lead to different interpretations of reported performance metrics. The sampling strategy is therefore part of the evaluation domain and should be explicitly documented and justified. Further details are given in Appendix A.

3.7 Relation to existing domain description practices

It is useful to distinguish the SADD from other concepts used to describe the intended operational setting of a system. A prominent example is the Operational Design Domain (ODD) in the automotive context, which specifies the conditions under which a system is intended to operate, such as weather, road type, traffic conditions, or time of day (17).

Such domain descriptions are valuable and often necessary. However, they do not by themselves define a sampling protocol and therefore do not uniquely determine a probability distribution. Two parties may fully agree on an ODD and nevertheless generate different evaluation datasets, thereby implicitly evaluating the system under different effective domains.

The SADD differs in exactly this respect. It does not only describe under which conditions a system is intended to operate, but also how valid samples from that domain are to be generated or selected. In this way, it turns an intended-use description into a stochastic object that supports statistically interpretable performance claims.

More broadly, several existing frameworks describe usage conditions, target populations, or contexts of use, but typically do not define the evaluation domain as a probability distribution induced by an explicit sampling protocol. The SADD is intended to complement these approaches by making this statistical reference domain explicit. A more detailed comparison is provided in Appendix B.

4 Case Study

To illustrate the practical role of a Stochastic Application Domain Definition, we consider a simple industrial inspection setting. The example is deliberately kept small in scope in order to isolate the statistical point. The purpose of the case study is not to present a realistic certification procedure in full detail, but to show how the meaning of a reported performance value depends on the sampling protocol through which the application domain is operationalized.

4.1 Industrial inspection example and sampling interpretation

Suppose an AI system has been developed to detect scratches in the paint of metal parts produced by a specific machine type operated at several locations in Austria. The system classifies each part as either faulty or intact based on images taken during production (28). A company considering deployment on its own machine cannot determine the suitability of the system solely from an overall accuracy value reported by the developer. Instead, the assessment must be made relative to the intended application domain defined by the SADD and the statistical testing procedure used to estimate performance.

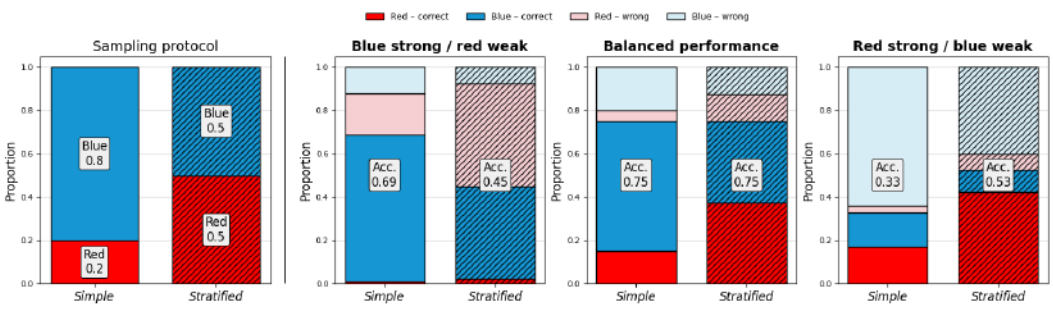


Figure 3: Influence of the sampling protocol on the interpretation of classification accuracy in the metal-part inspection example. The leftmost panel shows the evaluation dataset induced by two sampling strategies: simple random sampling from production (20% red, 80% blue) and stratified sampling by color (50% red, 50% blue). The three panels on the right show different class-wise performance profiles and the resulting aggregate accuracies under the two sampling strategies. Dark segments denote correctly classified parts and light segments incorrectly classified parts; hatched bars indicate the stratified design. The figure shows that aggregate accuracy is not self-interpreting but depends on the sampling protocol that defines the reference distribution (28).

To illustrate this, assume that the metal parts occur in only two colors, red and blue, and that accuracy is used as the performance metric. In actual production, suppose that 80% of parts are blue and 20% are red. Consider the two sampling strategies shown in Figure 3. Under **Strategy A**, the test set is obtained by simple random sampling from production, so that the evaluation sample reflects the natural 80/20 distribution and the reported accuracy estimates the probability that a randomly selected production part is classified correctly. Under **Strategy B**, the test set is stratified by color and both colors are sampled in equal proportions, so the resulting accuracy refers to a distribution that assigns equal weight to both colors.

The difference becomes particularly visible when class-wise accuracies differ. If the system performs well on blue parts but poorly on red parts, overall accuracy under simple random sampling may still appear high because blue parts dominate production. A stratified evaluation would make the weakness on red parts more visible by giving both colors equal weight. The key point is that performance metrics are not self-interpreting: an accuracy value is always an expectation with respect to some distribution. Different sampling strategies therefore lead to different interpretations of the same reported number.

4.2 Relation to the SADD

The case study illustrates the practical function of the SADD. A sufficiently explicit SADD does not merely state that the system is intended for "painted metal parts produced by machine A in Austria," but also clarifies in operational terms what counts as a valid sample for testing and how relevant variation within the intended domain should be represented.

For example, the SADD may imply that the intended domain corresponds to the actual production stream of eligible parts under ordinary operating conditions. In that case, simple random sampling from production would naturally approximate the induced distribution. Alternatively, the intended claim may concern balanced performance across relevant subdomains such as color categories, in which case a stratified design may be justified. In this way, the SADD connects the informal intended-use statement to a concrete sampling design and determines which distribution the evaluation aims to approximate, making the resulting performance measure interpretable and open to independent scrutiny.

4.3 Lessons from the example

Even in this simplified setting, the example illustrates several general points. First, the claim that a dataset is "representative" is incomplete unless the sampling protocol or induced distribution with respect to which representativeness is asserted is made explicit. Second, different sampling strategies

may be appropriate for different evaluation goals, but they should not be treated as interchangeable, since they correspond to different application domains in the statistical sense developed in this paper.

Third, independent testing and certification require more than benchmark results or a generic intended-use statement. They require a sufficiently explicit description of the domain under which the performance claim is supposed to hold, which is precisely the role of the SADD. While the example is intentionally simple, the same logic applies in more complex settings such as medical diagnosis, industrial quality assurance across multiple production sites, or multimodal systems deployed under heterogeneous environmental conditions.

5 Limitations

The proposed framework has important limitations.

Residual ambiguity of textual definitions. Even a carefully written SADD remains a textual document. Different qualified readers may still interpret certain terms or operational details differently. The framework does not eliminate this issue; rather, it makes it explicit and seeks to reduce it through better specification and normative interpretation. In practice, the expert deriving a sampling procedure from the SADD may still introduce bias through their own assumptions about how others would interpret the text.

Feasibility constraints in sampling. In many real-world settings, the ideal protocol suggested by a SADD cannot be implemented exactly. Access restrictions, costs, legal constraints, and organizational barriers may make direct sampling impossible. The resulting sampling design is then an approximation of the intended domain. This does not invalidate the framework, but it means that the relationship between the SADD and the actual sampling procedure must remain transparent and justified.

Rare events and outliers. The framework is primarily designed to support statistically meaningful evaluation on the intended application domain. It is not, by itself, a method for guaranteeing performance on rare cases, extreme outliers, or adversarial situations that are only weakly represented under the induced distribution. Additional robustness analyses and targeted stress tests may therefore still be necessary.

General-purpose AI. General-purpose AI (GPAI) models pose a particular challenge for the framework proposed here. Under the EU AI Act, general-purpose AI models are characterised by their generality and their capability to competently perform a wide range of distinct tasks, rather than being tied to one narrowly specified purpose (1). Accordingly, at the foundation-model level, there may be no single application domain in the sense discussed in this paper. Assessment at that level may therefore focus on training procedures, benchmark results, documentation duties, or model-level risk properties. However, once a GPAI model is integrated into or adapted for a specific downstream task, it becomes meaningful again to define a task-specific SADD. At that stage, the same logic proposed in this paper applies: functional trustworthiness claims should be interpreted with respect to an explicitly defined application domain.

6 Conclusion and outlook

This paper introduced a statistical perspective on the intended use of ML systems. We argued that performance guarantees are meaningful only with respect to a probability distribution, which we refer to as the Application Domain (AD). Because this distribution is typically unknown for real-world systems, we proposed the Stochastic Application Domain Definition (SADD), a textual specification of the sampling protocol that induces the distribution under which performance claims are intended to hold.

The SADD links intended use, representative data generation, and statistically interpretable testing. It supports communication along the value chain by clarifying the scope under which performance claims apply and provides the basis for deriving sampling procedures for statistically valid testing and independent verification. The framework does not assume that application domains can be specified perfectly; rather, it makes explicit that practical testing procedures inevitably rely on approximations and therefore require transparent documentation.

From a certification perspective, this transparency is essential. Process and documentation quality alone cannot replace a meaningful assessment of the functional trustworthiness of the deployed ML system. Future work should therefore investigate how SADDs can be standardized across sectors, how their quality can be assessed, how disagreements in interpretation can be documented, and how derived sampling procedures can be validated against the intended domain.

References

[1] Regulation (eu) 2024/1689 of the european parliament and of the council of 13 june 2024 laying down harmonised rules on artificial intelligence and amending regulations (ec) no 300/2008, (eu) no 167/2013, (eu) no 168/2013, (eu) 2018/858, (eu) 2018/1139 and (eu) 2019/2144 and directives 2014/90/eu, (eu) 2016/797 and (eu) 2020/1828 (artificial intelligence act) (text with eea relevance), Jul 2024.

[2] Matthew Arnold, Rachel K. E. Bellamy, Michael Hind, et al. Factsheets: Increasing trust in ai services through supplier's declarations of conformity. IBM Journal of Research and Development, 63(4/5):6:1–6:13, 2019.

[3] Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. In Transactions of the ACL, volume 6, pages 587–604, 2018.

[4] Yoshua Bengio, Geoffrey Hinton, Andrew Yao, Dawn Song, Pieter Abbeel, Yuval Noah Harari, Ya-Qin Zhang, Lan Xue, Shai Shalev-Shwartz, Gillian Hadfield, et al. Managing ai risks in an era of rapid progress. arXiv preprint arXiv:2310.17688, 2023.

[5] Christopher M Bishop and Nasser M Nasrabadi. Pattern recognition and machine learning, volume 4. Springer, 2006.

[6] Gary S. Collins, Karel G. M. Moons, et al. Tripod+ai: Reporting guideline for studies developing, validating, or updating a prediction model that uses artificial intelligence. BMJ, 2024.

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

[8] Carlos Galán. The certification as a mechanism for control of artificial intelligence in europe, 2019.

[9] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, et al. Datasheets for datasets. In Communications of the ACM, volume 64, pages 86–92, 2021.

[10] Delaram Golpayegani, Isabelle Hupont, Cecilia Panigutti, Harshvardhan J Pandit, Sven Schade, Declan O'Sullivan, and Dave Lewis. Ai cards: towards an applied framework for machine-readable ai and risk documentation inspired by the eu ai act. In Annual Privacy Forum, pages 48–72. Springer, 2024.

[11] Robert M Groves and Lars Lyberg. Total survey error: Past, present, and future. Public opinion quarterly, 74(5):849–879, 2010.

[12] Isabelle Hupont, David Fernández-Llorca, Sandra Baldassarri, and Emilia Gómez. Use case cards: a use case reporting framework inspired by the european ai act. Ethics and Information Technology, 26(2):19, 2024.

[13] Information technology — artificial intelligence — overview of trustworthiness in artificial intelligence. ISO/IEC TR 24028:2020, 2020. International Organization for Standardization.

[14] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. arXiv preprint arXiv:2401.04088, 2024.

[15] Göran Kauermann and Helmut Küchenhoff. Stichproben: Methoden und praktische Umsetzung mit R. Springer-Lehrbuch. Springer Berlin Heidelberg.

[16] Johann Laux, Sandra Wachter, and Brent Mittelstadt. Trustworthy artificial intelligence and the european union ai act: On the conflation of trustworthiness and acceptability of risk. Regulation & Governance, 18(1):3–32, 2024.

[17] Chung Won Lee, Nasif Nayeer, Danson Evan Garcia, Ankur Agrawal, and Bingbing Liu. Identifying the operational design domain for an automated driving system through assessed risk. In 2020 IEEE Intelligent Vehicles Symposium (IV), pages 1317–1322, 2020. doi: 10.1109/IV47402.2020.9304552.

[18] Chung Won Lee, Nasif Nayeer, Danson Evan Garcia, Ankur Agrawal, and Bingbing Liu. Identifying the operational design domain for an automated driving system through assessed risk. In IEEE Intelligent Vehicles Symposium, pages 1317–1322, 2020.

[19] Sharon L. Lohr. Sampling: Design and Analysis. Brooks/Cole, 2nd ed edition.

[20] Margaret Mitchell, Simone Wu, Andrew Zaldivar, et al. Model cards for model reporting. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, pages 220–229, 2019.

[21] John D. Musa. Operational profiles in software-reliability engineering. IEEE Software, 10(2):14–32, 1993.

[22] Bernhard Nessler, Thomas Doms, and Sepp Hochreiter. Functional trustworthiness of ai systems by statistically valid testing. arXiv, 2310.02727, 2023.

[23] Maximilian Poretschkin, Anna Schmitz, Maram Akila, Linara Adilova, Daniel Becker, Armin B Cremers, Dirk Hecker, Sebastian Houben, Michael Mock, Julia Rosenzweig, et al. Guideline for designing trustworthy artificial intelligence. 2023.

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 10684–10695, 2022.

[25] Peter M Rothwell. External validity of randomised controlled trials:"to whom do the results of this trial apply?". The Lancet, 365(9453):82–93, 2005.

[26] SAE International. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles, 2018.

[27] John Schulman, Barret Zoph, and Christina Kim. Introducing chatgpt, Nov 2022. Online, accessed 2023-05-17. URL: https://openai.com/blog/chatgpt.

[28] Kajetan Schweighofer, Barbara Brune, Lukas Gruber, Simon Schmid, Alexander Aufreiter, Andreas Gruber, Thomas Doms, Sebastian Eder, Florian Mayer, Xaver-Paul Stadlbauer, Christoph Schwald, Werner Zellinger, Bernhard Nessler, and Sepp Hochreiter. Safe and certifiable ai systems: Concepts, challenges, and lessons learned, 2025. URL: https://arxiv.org/abs/2509.08852.

[29] George Sharkov, Christina Todorova, and Pavel Varbanov. Strategies, policies, and standards in the eu towards a roadmap for robust and trustworthy ai certification. Information & Security, 50(1):11–22, 2021.

[30] U.S. Food and Drug Administration. Considerations for the use of artificial intelligence to support regulatory decision-making for drug and biological products, 2025. Draft Guidance.

[31] Robert F. Wolff, Karel G. M. Moons, Richard D. Riley, et al. Probast: A tool to assess risk of bias and applicability of prediction model studies. Annals of Internal Medicine, 170(1):51–58, 2019.

A Common sampling strategies

The challenge of obtaining representative samples under practical constraints is well known in statistical sampling theory. There, one aims to draw samples from a target population by means of a documented sampling design. We adapt these ideas here to the setting in which the target distribution is defined not by a complete population list, but by a textual domain specification (15; 19).

A.1 Simple random sampling

A simple random sample is the conceptual gold standard for representative sampling, because every eligible element has the same probability of being selected. If such sampling is possible, it provides the cleanest basis for statistical inference.

In the context of a SADD, however, simple random sampling is often not feasible. It would require access to a sufficiently complete sampling frame corresponding to the intended application domain. For many real-world ML applications, no such frame exists in explicit form. Even when a partial frame exists, legal, economic, and logistical constraints often prevent direct random access to all eligible units. For that reason, simple random sampling is usually best understood as an ideal benchmark rather than a practically available design.

A.2 Cluster sampling

Cluster sampling addresses this difficulty by using naturally occurring groups, such as hospitals, schools, factories, devices, or regions, as intermediate sampling units. Instead of sampling directly from the entire target domain, one first selects clusters and then samples within them.

In a SADD-based setting, this is often a natural way of operationalizing the intended domain. The sampling protocol may follow a hierarchical structure: for example, choose a country, then a clinic, then a department, then a device, and finally an observation recorded with that device. Each stage corresponds to an explicit design choice in the operational sampling protocol.

For cluster sampling to approximate the intended application domain well, the clusters should be chosen in a way that does not systematically distort the target domain. In practice, cluster sampling is often more feasible and cost-effective than direct random sampling, but it usually increases sampling variance and requires explicit treatment of intra-cluster dependence in the statistical analysis (15; 19).

A.3 Stratified sampling

Stratified sampling divides the domain into relevant subgroups, or strata, and samples separately within each of them. In contrast to cluster sampling, strata are defined to be internally homogeneous with respect to selected characteristics, while the full set of strata captures domain variability.

In a SADD-based framework, stratification is appropriate when certain variables are known to be relevant for the intended application domain, such as country, device type, age group, clinical subgroup, or environmental condition. The stratum definitions and the allocation of samples across strata must then be made explicit.

Stratified sampling can improve precision and ensure that important subdomains are adequately represented. However, it requires that eligible units can be assigned to strata in a meaningful way and that the weighting scheme used for aggregate performance estimation is taken into account when interpreting overall results (15; 19).

B Positioning with Respect to Existing Approaches

The problem of specifying the conditions under which the performance of an AI system should be interpreted arises in multiple research communities, including software reliability engineering, autonomous systems safety, machine learning documentation, and clinical AI evaluation. Existing approaches operationalize the application domain in different ways, typically emphasizing either usage distributions, operational conditions, target populations, or structured documentation of intended use. In the following, we review the main strands of work relevant to the specification of the application domain.

B.1 Operational profiles in software reliability engineering

One of the earliest formal approaches to operationalizing the domain under which software reliability claims should be interpreted is the operational profile introduced by Musa (21). An operational profile is a quantitative characterization of system usage, defined as a probability distribution over the operations or input classes that occur during actual use. The operational profile is used to guide testing and reliability estimation by ensuring that test cases are sampled according to the expected usage distribution.

This idea closely parallels the statistical perspective adopted in this paper: reliability or performance metrics are meaningful only relative to a distribution of use cases. However, operational profiles are typically defined over software operations or usage events rather than over real-world situations that generate inputs to a machine learning system. As a result, they are well suited to traditional software systems but less directly applicable to modern ML systems whose inputs arise from complex real-world processes.

B.2 Operational design domains in autonomous systems

In the domain of automated driving, the most widely used concept for describing the scope of system operation is the Operational Design Domain (ODD). The ODD specifies the conditions under which an automated driving system is designed to function safely, including environmental conditions, road types, traffic situations, and geographical or temporal constraints (26; 18).

The ODD provides a structured way of describing the operational context of a system and is widely used in safety cases and certification processes for automated driving. Recent standardization efforts such as ASAM OpenODD further aim to formalize ODD descriptions and provide machine-readable representations of operational domains.

While ODDs provide an explicit specification of admissible operating conditions, they typically do not define a sampling procedure or probability distribution over situations within the domain. Consequently, different parties may evaluate the same system within the same ODD using different datasets and thereby implicitly evaluate different effective distributions. In contrast, the framework proposed in this paper explicitly links the domain specification to a stochastic sampling protocol.

B.3 Target population and intended setting in clinical AI

In clinical prediction modeling and medical AI, a related problem is addressed through the specification of the target population, intended setting, and intended use of a model. Reporting guidelines such as TRIPOD-AI require authors to document the healthcare setting, target population, intended purpose, and intended users of a predictive model (6).

Closely related concepts include targeted validation and risk-of-bias assessment frameworks such as PROBAST, which evaluate whether a study population and setting are representative of the intended use of the model (31). These approaches emphasize that model performance must be interpreted relative to the population and setting in which the system is intended to operate.

Compared with the framework proposed here, these approaches primarily specify the population and setting for which a model is intended, but typically do not formalize the statistical domain as a distribution induced by an explicit sampling protocol.

B.4 Context-of-use frameworks in regulatory evaluation

Regulatory frameworks for AI and computational models increasingly emphasize the notion of context of use. For example, recent FDA guidance on AI models used in regulatory decision-making defines the context of use as the specific role and scope of the model within a decision process (30). The context of use describes how model outputs will be used, the question being addressed, and the consequences of model errors.

Such frameworks operationalize the domain of application by situating the model within a specific decision workflow. While this provides important information about the purpose and risk implications of the model, it does not directly specify the distribution of situations under which model performance should be evaluated.

B.5 Documentation frameworks for AI systems and datasets

A growing body of work proposes structured documentation frameworks to describe the intended use and limitations of AI systems and datasets. Model Cards (20) aim to communicate key information about machine learning models, including intended uses, evaluation conditions, and potential limitations. Similarly, Datasheets for Datasets (9) and Data Statements for NLP datasets (3) document dataset provenance, composition, and intended uses in order to improve transparency and enable more responsible deployment.

Related proposals such as FactSheets (2), Use Case Cards (12), and AI Cards (10) extend this idea to broader AI system documentation, including usage scenarios, system capabilities, and potential risks. These approaches help clarify the contexts in which a system is expected to be used and discourage deployment outside the intended domain.

However, these frameworks primarily focus on documentation and communication rather than on defining a statistical reference distribution for evaluation. As a result, they provide valuable contextual information but do not by themselves specify how representative evaluation datasets should be generated.

B.6 Position of this work

Taken together, existing approaches operationalize the application domain of AI systems in several complementary ways: through usage distributions (operational profiles), operational conditions (ODDs), target populations and settings (clinical evaluation frameworks), decision contexts (context-of-use models), and structured documentation of intended use.

The framework proposed in this paper builds on these ideas but emphasizes the statistical interpretation of performance claims. Specifically, we model the application domain as a probability distribution induced by a sampling protocol and introduce the Stochastic Application Domain Definition (SADD) as a textual specification of that protocol. This perspective explicitly links intended use descriptions with the sampling procedures required for statistically interpretable evaluation.

C Operationalization of Dataset Representativeness through the Stochastic Application Domain Definition (SADD)

C.1 Scope

This section specifies requirements and procedures for operationalizing the requirement of dataset representativeness as stated in Article 10(3) of Regulation (EU) 2024/1689 (Artificial Intelligence Act). It defines the concept of the Stochastic Application Domain Definition (SADD) and establishes requirements for its use in the specification, construction, and evaluation of testing datasets for AI systems.

The provisions of this section apply to the development, evaluation, and certification of AI systems for which performance claims are made with respect to an intended application domain. The objective is to ensure that reported performance metrics are interpretable and statistically valid with respect to the conditions under which the AI system is intended to operate.

Approach	Domain representation	Operationalization	Dist.?	Use context
Operational Profile (21)	Usage distribution	Probability distribution over software operations or usage events	Yes	Software reliability engineering
Operational Design Domain (ODD) (26; 18)	Operational conditions	Admissible environmental and operational constraints	No	Automated driving safety and certification
Target Population / Intended Setting (6; 31)	Population and setting	Specification of target population, intended setting, users, and purpose	Partial	Clinical prediction and medical AI
Context of Use (COU) (30)	Decision context	Role and scope of the model in a decision workflow	No	Regulatory evaluation
Model Cards (20)	Intended use and evaluation conditions	Structured documentation of intended uses, results, and limitations	No	Model transparency and documentation
Datasheets / Data Statements (9; 3)	Dataset provenance and population	Documentation of composition, collection, demographics, and intended uses	Indirect	Dataset documentation and bias analysis
FactSheets / AI documentation frameworks (2; 12; 10)	Usage scenarios and governance	Structured documentation of use cases, risks, and operational scope	No	AI governance and certification documentation
SADD (this work)	Stochastic application domain	Textual specification of a sampling protocol inducing an evaluation distribution	Yes	Statistical evaluation and certification

Table 1: Comparison of existing approaches for specifying the application domain of AI systems. Existing approaches typically operationalize the domain through conditions, populations, workflows, or documentation frameworks. In contrast, the Stochastic Application Domain Definition (SADD) explicitly links the intended application domain to a sampling protocol that induces the probability distribution under which performance claims are interpreted.

C.2 Normative references

The following documents are referred to in this section and are indispensable for its application.

- Regulation (EU) 2024/1689 of the European Parliament and of the Council (Artificial Intelligence Act)
- ISO/IEC TR 24028:2020, Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence

C.3 Terms and definitions

For the purposes of this section, the following terms and definitions apply.

Application domain (AD). Probability distribution over relevant input–output situations under which the performance of an AI system is intended to be interpreted.

Stochastic Application Domain Definition (SADD). Textual specification of the sampling protocol that induces the probability distribution corresponding to the application domain.

Protocol-induced distribution. Probability distribution over admissible samples resulting from the stochastic procedure specified by the sampling protocol described in the SADD.

Evaluation dataset. Dataset constructed for the purpose of estimating the performance of an AI system with respect to the application domain defined by the SADD.

Sampling procedure. Operational implementation of the sampling protocol defined in the SADD used to generate or select evaluation samples.

C.4 General principle of representativeness

For the purposes of Article 10(3) of the Artificial Intelligence Act, a dataset shall be considered representative if it is generated according to a sampling procedure that approximates the protocol-induced distribution defined by the SADD.

Representativeness shall therefore be interpreted with respect to the application domain defined by the SADD and not solely with respect to superficial similarity to real-world data or previously used datasets.

Performance metrics reported for an AI system shall be interpreted as estimates of statistical quantities defined with respect to the protocol-induced distribution associated with the SADD.

C.5 Requirements for the Stochastic Application Domain Definition

A SADD shall be documented for each AI system for which performance claims are made.

The SADD shall include the following elements, where applicable:

1. description of the real-world process, environment, or objects that generate the inputs to the AI system;
2. specification of inclusion and exclusion criteria defining admissible samples;
3. description of the geographical and temporal scope of the intended application;
4. description of relevant actors, devices, sensors, and data acquisition procedures;
5. description of relevant operational constraints or conditions;
6. description of the mechanism through which valid samples arise in the real-world process;
7. specification of any relevant stratification variables or subdomains;
8. description of foreseeable sources of variation within the application domain.

The SADD shall be documented in a manner that enables a qualified independent party to reconstruct a sampling procedure that approximates the intended application domain.

C.6 Derivation of sampling procedures

A sampling procedure shall be derived from the SADD for the purpose of constructing evaluation datasets.

The derivation of the sampling procedure shall satisfy the following requirements:

1. The procedure shall approximate the protocol-induced distribution defined by the SADD to the extent that is reasonably achievable in practice.
2. The procedure shall be documented in sufficient detail to allow independent reproduction of the sampling process.
3. Any deviations from the idealized sampling protocol implied by the SADD shall be documented and justified.

The derivation of the sampling procedure shall be carried out by a qualified domain expert or by a team possessing expertise in the application domain and statistical sampling methodology.

C.7 Sampling design

Evaluation datasets shall be generated using a documented sampling design consistent with the SADD.

The sampling design may include one or more of the following strategies:

- simple random sampling,
- stratified sampling,
- cluster sampling,
- multi-stage sampling.

The choice of sampling design shall be justified with respect to the characteristics of the application domain and the feasibility constraints of data collection.

Where stratified sampling is used, the definition of strata and the weighting scheme used for aggregate performance estimation shall be documented.

C.8 Construction of evaluation datasets

Evaluation datasets used for testing AI systems shall satisfy the following requirements:

1. The dataset shall be constructed using the sampling procedure derived from the SADD.
2. The sampling process shall be documented and auditable.
3. The dataset shall contain sufficient information to enable verification of compliance with the SADD.
4. The dataset shall be independent of the training and validation datasets used in model development, unless otherwise justified.

The dataset size shall be determined based on statistical considerations, including the desired precision of performance estimates.

C.9 Documentation requirements

The following documentation shall be maintained and made available to relevant stakeholders or conformity assessment bodies:

- the Stochastic Application Domain Definition;
- the derived sampling procedure;
- the sampling design and its justification;
- the dataset construction process;
- any deviations from the intended sampling protocol;
- statistical methods used for performance estimation.

The documentation shall allow an independent assessor to evaluate whether the evaluation dataset reasonably approximates the intended application domain.

C.10 Interpretation of performance metrics

Performance metrics reported for the AI system shall be interpreted as statistical estimates with respect to the protocol-induced distribution defined by the SADD.

Reported performance values shall therefore be accompanied by:

- a reference to the SADD under which the evaluation was conducted;
- the sampling design used to construct the evaluation dataset;
- the statistical uncertainty associated with the performance estimate.

Where evaluation datasets are generated using different sampling procedures, the resulting performance estimates shall be interpreted as referring to different effective application domains.

C.11 Limitations

The SADD framework does not eliminate all sources of ambiguity in the specification of application domains. Interpretation of a SADD shall be guided by the understanding that can reasonably be expected from qualified stakeholders in the relevant application domain.

Where practical constraints prevent exact implementation of the sampling protocol described in the SADD, the resulting sampling procedure shall be documented as an approximation of the intended domain.

Additional robustness testing, stress testing, or targeted evaluation may be required to assess performance in rare, extreme, or adversarial situations that are not adequately represented under the protocol-induced distribution.

Conversational Agents in Multi-User Environments

Tobias Halmdienst[*‡] **Umut Tanriverdi**[*‡] **Simon Schmid**[†] **Michal Lewandowski**[†]
Bernhard Nessler[†‡]

[†]SCCH [‡]JKU Linz

Abstract

Passing as human in a room full of people requires more than fluent speech, it demands reading the room. While Large Language Models (LLMs) have transformed human-AI interaction in a one-to-one setting, they still fall short in multi-user conversational settings where social dynamics define the interaction. In such environments, a conversational agent that merely generates coherent text will struggle to maintain consistent socially plausible behavior. We propose a structured Theory of Mind (ToM) framework that equips conversational agents with the cognitive machinery to reason over participant beliefs, intentions, and evolving group dynamics in real time. Rather than relying on a single LLM prompt, our architecture decomposes social reasoning into explicit modules—a knowledge base, belief system, goal generator, and intention planner—coupled with a dual-process response architecture that balances immediacy with strategic depth. To evaluate this approach, we deployed the framework within the Turing Game and Reverse Turing Game environments, further enhancing the agent's plausibility with a simulated human-like response timing algorithm. Preliminary evaluations demonstrate that our ToM-equipped agent exhibits stronger conversational coherence, sustains longer exchanges, and is less frequently identified as a bot compared to its predecessor without structured social reasoning.

1 Introduction

In recent years, LLMs have revolutionized the world due to their ability to create coherent and meaningful written content and to engage in text-based conversations with humans (T. B. Brown et al. 2020; OpenAI 2023). This success can largely be attributed to advances in deep learning architectures, the availability of large-scale training data and increased computational resources (Vaswani et al. 2017; Kaplan et al. 2020). These developments have significantly advanced the state of the art in natural language processing. While most of the work in this field is focused on single-user interaction settings, multi-user interaction environments are a more complex, understudied domain. LLMs often struggle to maintain coherent behavior in multi-party interactions, particularly in scenarios that require social reasoning (Fang et al. 2025). In the context of multi-user environments LLMs suffer from the alignment problem, where the behavior of the model may diverge from human intentions (Nath, Graff, and Krishnaswamy 2026). This mismatch arises because these models are trained to optimize statistical objectives such as next token prediction and do not necessarily produce responses that align with human expectations (Russell 2019; Ouyang et al. 2022). Humans have a Theory of Mind to reason about the perspectives and beliefs of others (Premack and Woodruff 1978). This capability enables individuals to anticipate the reactions of others and adapt accordingly during social interaction. This reasoning plays a crucial role in maintaining coherent conversation in multi-conversational settings (Clark 1996). However, LLMs do not possess this capability due to their inherent design. It is necessary to incorporate ToM reasoning into the conversational agent

[*]contributed equally

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

to maximize socially coherent behavior, because communication relies on shared intentionality between participants (Tomasello 2008). If the agent lacks this ability, it may fail to correctly predict the intentions of other participants, which could lead to socially incoherent responses. Modeling the beliefs, intentions, and goals of others enables the conversational agent to behave in a socially coherent way. This paper proposes a framework for multi-user interactive agents that incorporate ToM reasoning, tackling the problem of enabling socially aware interactive agents in multi-user settings. Our framework is novel in that it integrates ToM reasoning and incorporates human-like response timing. By modeling beliefs and intentions of others the agent is able to understand social dynamics and can successfully compete in multi-user conversations. Furthermore, this work aims to highlight the importance of integrating social reasoning components into conversational agents and show the limitations of relying solely on LLMs in multi-user conversational settings.

2 Related Work

2.1 Interactive Agents in the Turing Game

The *Turing Game* extends the classical *Turing Test* proposed by Alan Turing (Turing 1950). In the Turing Game three parties communicate solely through text in a shared chatroom. Each round consists of two human participants and one conversational agent. For the human players the goal is to correctly identify the agent. The interactive agent attempts to naturally engage in the conversation and remain indistinguishable from human participants.

Moreover, we consider a variation called the *Reverse Turing Game*, where two conversational agents communicate with one human. In contrast to the original Turing Game, this setting forces the bots from merely fitting in the group's dynamic to taking a more active part, and thus is more challenging than the original design. This setting places stronger demands on the agents' ability to generate appropriate responses. Nevertheless, it provides valuable insights into the dynamics of a conversation held solely between two interactive agents. Without the inputs of a third human party, the interactive agents struggle to sustain creative dialog. The performance of the conversational agent is strongly influenced by the quality of interaction with human participants. The performance increases when a human participant is more active and engaging. Developing agents capable of initiating and sustaining dialogue without relying entirely on human guidance establishes a foundation for supportive technologies to assist in e.g. healthcare systems.

Figure 1: **Left:** the classical Turing Test (Turing 1950) features a judge who decides which interlocutor is the machine, while the other human serves mainly as a comparison counterpart. **Right:** the Turing Game assigns both humans the dual role of independently identifying the machine while simultaneously supporting the other human; the pair wins only if both humans correctly identify the machine (Lewandowski et al. 2024).

2.2 Theory of Mind in AI and LLMs

Theory of Mind in the context of AI, Tang and Belle (2024) advanced neural architectures by delegating ToM reasoning to external symbolic executors, demonstrating that machines can execute verifiable false-belief tests with greater logical consistency. Complementing this architectural perspective, Zhu, Z. Zhang, and Yizhou Wang (2024) showed that LLMs internally form linearly decodable belief-state representations, and that manipulating these via activation editing dramatically alters ToM performance. Whether LLMs inherently possess Theory of Mind remains contested, however. Street et al. (2026) reported that GPT-4 not only solves standard false-belief tasks but achieves adult human performance on higher-order ToM evaluations. Conversely, Riemer et al. (2025) argued that

existing ToM benchmarks are broken for evaluating LLMs, showing that trivial, logically irrelevant modifications cause LLMs to fail and introducing the concept of "functional theory of mind"—the ability to adapt to partners in context—on which even strong models collapse. Sclar et al. (2025) dramatically extended these fragility findings through program-guided adversarial data generation, reporting that Llama-3.1-70B achieves approximately 0% and GPT-4o approximately 9% accuracy on adversarially generated ToM stories. J. Hu, Sosa, and Ullman (2025) provided a theoretical framework reconciling these contradictory results, arguing that disagreements stem from conflating human *behaviors* with the *computations* underlying them. On the evaluation side, Chen et al. (2024) systematically assessed ToM across 8 tasks and 31 social cognition abilities, finding that even GPT-4 lags behind human performance by over 10 percentage points. J. Zhou et al. (2025) corroborated this fragility across expansive social intelligence tasks, showing that LLMs continue to perform well below human levels when navigating complex, goal-oriented social interactions. To bridge the gap between theoretical ToM capabilities and the demonstrated fragility in social tasks, our work introduces a structured cognitive architecture. By explicitly computing internal states, such as dynamic beliefs, intentions, and goals, our system moves beyond simple prompting to provide a more stable, functional Theory of Mind for complex multi-user interactions.

2.3 Multi-Party Conversational AI

The majority of dialogue research is focused on dyadic interactions. However, multi-party settings introduce distinct challenges, such as state of mind modeling, conversation disentanglement, and agent action modeling (Sapkota et al. 2025). To systematically evaluate these capabilities, M. Zhang et al. (2026) introduced MPCEval, which provides reference-free metrics that assess full-conversation generation and speaker-content consistency, moving beyond local next-turn prediction. For multi-party conversation understanding, Sun et al. (2025) leveraged pre-trained LLMs with speaker-aware contrastive learning for multi-party dialogue generation, outperforming baselines without requiring explicit relation annotations. Shifting past early text-only datasets like Molweni, Yueqian Wang et al. (2025) introduced Friends-MMC, a new high-density multimodal corpus containing tens of thousands of video-paired utterances to explore character-centered understanding and speaker identification, while Shi et al. (2025) introduced MuMA-ToM, a multi-modal benchmark for mental reasoning in multi-agent interactions with questions about goals and beliefs. Most relevant to complex group setups, E. Hu et al. (2025) explicitly studied conversational agents in dynamic environments mixing humans and AI, providing orchestration tools to manage the dual challenge of deciding when to speak and producing contextually coherent utterances in hybrid conversations. On the multi-agent side, Tran et al. (2025) surveyed how LLM-based multi-agent systems leverage natural language for coordination in group settings, covering cooperation, competition, and mixed scenarios. Regarding Theory of Mind in conversational agents Sapkota et al. (2025) highlighted that ToM remains essential for intelligent multi-party conversational agents. Building on this need for better multi-party coordination, our research addresses the challenge of conversation management by giving the agent explicit memory of different players. By implementing a dual-process system that controls response timing and separates fast reflexes from slower deliberation, our agent can naturally decide when to speak and effectively participate in three-player discussions.

2.4 Social Reasoning, Belief Modeling, and Social Deduction Games

Reasoning about others' beliefs and intentions in interactive settings has been explored through social simulation and game-playing agents (Piao et al. 2025; Bougie and Watanabe 2025). In social deduction games, Song et al. (2025) showed that LLMs produce fluent rhetoric in Werewolf but struggle with genuine deception and counterfactual reasoning, while Sarkar et al. (2025) doubled win rates over standard RL baselines by training language models via multi-agent reinforcement learning in an Among Us–based environment. To address such vulnerabilities, S. Wang et al. (2024) introduced Recursive Contemplation for Avalon, improving good-side win rates from 15% to 83.3%, and Light et al. (2025) enabled LLMs to self-improve via bi-level Monte Carlo Tree Search, reaching human-level play without human training data. For ToM in game-playing agents, Guo et al. (2024) and Kempinski et al. (2025) showed that first- and second-order belief modeling yields stronger strategies in imperfect-information games. In line with these advancements, our work applies explicit belief modeling to the Turing Game. Rather than relying on reinforcement learning or search trees, our agent uses a continuous cognitive loop to monitor suspicion levels, identify potential allies, and dynamically adjust its conversational strategy to avoid being detected as a bot.

3 Methods

In multi-user conversational settings, a conversational agent faces a fundamentally different challenge than in one-to-one human–AI interaction. It must respond under real-time social pressure and maintain behavioral plausibility across an extended discourse with several participants. In order to equip the conversational agent with this ability, we do not rely solely on a monolithic LLM prompt. We design a conversational agent that leverages a structured ToM reasoning pipeline together with planning ahead in message generation. The paradigm differs fundamentally from the standard LLM interaction pattern. Instead of generating responses only when prompted, the agent continuously maintains a candidate response and continuously updates its own state of mind according to the perception of messages. As the base model for response generation, we use the reasoning LLM GLM-4.7-flash, since it outperformed different LLMs in the setting of the Turing Game. This approach maximizes social understanding and promotes coherent conversational behavior.

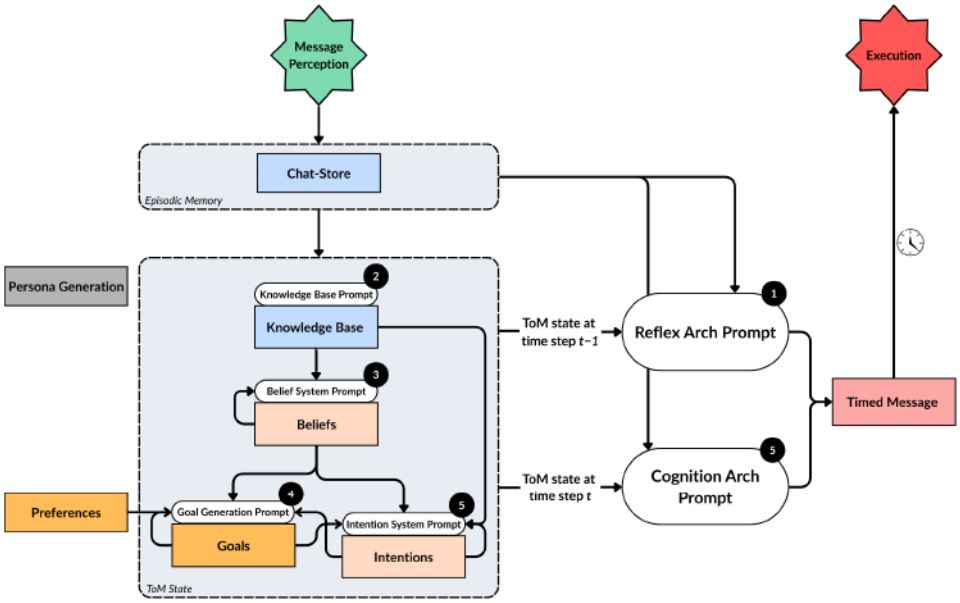


Figure 2: Our ToM framework. Incoming messages are stored in the chat-store, which serves as an episodic memory, and processed through two concurrent architectures. **(1)** Upon receiving a message, the reflex arch uses the ToM state from the previous time step $(t-1)$ to generate a fast, short response and schedules it via the timing mechanism (Subsection 3.3). In parallel, the cognition arch performs a full sequential update of all ToM states. **(2)** First, the episodic memory is compressed into the knowledge base, which extracts factual observations such as message type and behavioral patterns of other participants. **(3)** The belief system then uses the updated knowledge base to model beliefs about the other players, including suspicion levels, bot-like behavioral tells, and alliance potential. **(4)** The intention system integrates the updated beliefs with the current goals to determine the next strategic action, such as accusing, allying, or probing a specific player. **(5)** The goal generation module dynamically derives a concrete immediate goal from the agent's preferences, beliefs, and intentions. The cognition arch then uses the fully updated ToM state at time step t to compose a contextually richer response, which overwrites the reflex arch message if it has not been executed yet. Finally, the timed message mechanism schedules the selected response with a human-like delay and sends the message once the timer is expired.

3.1 Parallel Response Generation

Unnatural response gaps are among the signals human players use to identify bots. To maintain the conversational flow, the interactive agent must find a balance between immediacy and deeper reasoning. To address this problem we designed a dual-process response architecture inspired by Kahneman (2011) distinction between fast, intuitive cognition—implemented as reflex arch, and slow, deliberate reasoning—implemented as cognition arch. The reflex arch aims to provide a fast

but sound response. In contrast, the cognition arch aims to provide a slower but more deliberate response, incorporating ToM states. At any point in the interaction however, a candidate response is maintained and ready to be dispatched, thereby maintaining the invariance condition that a valid response is always available.

All messages get processed in the following manner: Upon receiving an incoming message, the reflex arch immediately generates a short, low-latency reply using only the agent's most recent mental state—its current beliefs, goals, and intentions as they stood before the new message arrived. In parallel, a cognition arch running as a background loop performs a full update of the agent's internal mental model based on the latest message and then generates a richer, strategically more informed reply. Both pathways write to a shared message slot with a scheduled send timestamp in order to avoid duplicate messages. A dedicated sender loop monitors this slot and dispatches whichever message is current when the timer expires. Crucially, if the cognition arch completes before the reflex message is sent, it overwrites the planned message with its higher-quality output.

3.2 Theory of Mind Framework

The interactive agent must track each player's identity, behavior, and suspicions over time, in order to adapt its responses and maintain plausible participation in the game. Previous interactive agents that use a simple input-output LLM system failed in maintaining coherent multi-party conversations over extended periods of time, showcasing that relying on a single LLM prompt for all reasoning is insufficient (Laban et al. 2025; Nonomura and Mori 2024; Lewandowski et al. 2024). In this framework the interactive agent tended to forget earlier observations or contradict its own assessments. As a result we introduced a structured ToM state representation that explicitly stores and interprets the observation of the other players, imitating the human social thinking process. The goal is to maximize social coherence in the conversation.

The state of mind of the bot is structured in a Chat Store, Knowledge Base, Belief System, Goal Generator, and an Intention Module. The Chat Store serves as an episodic memory based on the conversation history. First, the Knowledge Base Prompt extracts factual observations from the Chat Store, like questions asked or facts that are explicitly provided by other users. Following, the Belief System Prompt updates a mental model for each opponent. This is the core component of the theory of mind idea. In contrast to the Knowledge Base, that represents explicitly exhibited facts from the other players, the Beliefs aim to reflect the non-explicit or hidden intentions and properties of the other players. This also includes what the bot believes about the identity of the other players, i.e. if the other player is a bot or a human, and which other player might have which suspicions about the very bot itself or about the third player. This reasoning is highly relevant in the course of the game in order to raise accusations or answer to explicit suspicions in the chat. The Goal Generation Prompt maintains or updates a single strategic objective based on the current game state and static preferences. The goal gives a short and explicit notion of the current strategy, like gather background information from player red or accuse player blue or convince player yellow to accuse player blue. The Intention System Prompt combines beliefs and goals into a detailed action plan, still formulated in the style of an inner dialog, containing reasonings and strategic thoughts. Based on this stage by stage well-thought and reasoned, structured state of mind it is at any time possible to generate a single next message for the chat. The detailed representation of the ToM states helps the interactive agent to maintain a coherent and balanced conversation. All of the ToM states are accessed, written and combined using few-shot prompting techniques.

The next section gives a short description of the used prompts corresponding to each action. The *reflex prompt* is intentionally concise and constrained in order to approximate a rapid human response. The goal is to maintain the flow of conversation without inducing noticeable delays. The reflex prompt uses the current perception and cognition states as context to produce a brief candidate response. Concurrently with the reflex cycle, the conversational agent initiates the *cognition prompt*, which starts the cognition cycle. This cycle takes longer to complete because it must update all internal ToM-states. Specifically, it is responsible for updating the Knowledge Base, Beliefs, Goals, and Intentions before drafting a candidate message. The *system prompt* aims to provide the conversational agent with essential information such as the game rules of the Turing Game and the desired communication behavior. Moreover, the conversational agent is assigned a persona. A persona is a short description of a person e.g. profession, hobbies and music preferences. This aims to improve

social coherence in a conversation and makes the conversational agent less susceptible to specific questions. The persona is generated on the fly before the start of each round using a LLM.

3.3 Planning Ahead in Message Generation

Time management is vital in multi-user conversational settings (Sacks, Schegloff, and Jefferson 1974). Since LLMs have no conception and feeling for time, human-like response timing does not emerge naturally from the model. In human conversations, response delays are shaped by cognitive processes, such as reading, writing and analyzing. The human thinking process is inherently different to the way LLMs process information. Responses created by an AI may appear unnaturally fast and would not appear natural in human communication. To counteract this problem, we introduced a response timing algorithm. This not only prevents the interactive agent from exhibiting unnatural response times, but also utilizes these generated pauses for active observation. By monitoring how the game evolves and waiting for other players to respond, the agent maintains a natural conversation flow. Thus, the challenge is to implement a bridging mechanism between the typical LLM agent-like process and a time-aware simulation of social interactions. We solve this problem by explicitly modeling a target time, at which a newly created message is to be sent to the chat. This results in the invariance condition of the system, that every internal thinking process that updates the bot's state of mind a valid plan for the future message, i.e. time and content, that is to be sent next has to be maintained. As a fallback mechanism, the bot guarantees that a subsequent message is always prepared and held in reserve, immediately after the bot sends a response.

Human-Like Response Timing. Beyond the content of messages, a conversational agent in a multi-user setting risks exposure through temporal artifacts, response latencies that are either unnaturally fast or unnaturally uniform. To counteract this, we designed a timing simulation layer that models human typing speed (Card, Moran, and Newell 1980; Stivers, Enfield, P. Brown, et al. 2009). Modeling realistic response timing is particularly important in multi-party conversational environments, where participants may implicitly evaluate the authenticity of an interlocutor not only through linguistic content but also through temporal interaction patterns. To simulate natural, human-like text conversations, the system calculates response timing in two steps. First, it accounts for the time required to write the message, let L denote its length in characters. Coupled with parallel response generation, this two-step design provides a solid setup for imitating realistic human delay and message planning.

For reflex responses, the delay d_r is defined as

$$d_r = \frac{L}{4} + \mathcal{U}(2, 5) \tag{1}$$

For cognition responses, the delay d_c is defined as

$$d_c = \frac{L}{4} + \mathcal{U}(4, 7) \tag{2}$$

where L is the length of the incoming message in terms of characters and where $\mathcal{U}(a, b)$ denotes a random sample drawn from a uniform distribution over the interval $[a, b]$. The denominator is a constant and represents the writing speed. This algorithm combines a deterministic writing term with a stochastic cognitive term and mirrors the stage-based decomposition of human response time proposed by Sternberg (1969), where total latency arises from the sum of independent processing-stage durations. The use of a uniform rather than a normal distribution is particularly well-suited here, as it prevents the generation of extreme outlier delays that could serve as detectable temporal cues to conversation partners (Lew et al. 2018). Indeed, Gnewuch et al. (2018) demonstrated empirically that dynamically delayed chat-bot responses—scaled to message complexity—significantly increase perceived humanness and social presence over near-instant replies.

3.4 Accusation Mechanism in the Turing Game.

The Turing Game served as fundamental testbed for studying these challenges in a controlled multi-user environment. The methods described in the following section where designed to address core behavioral requirements that rise for this setting. In the setting of the Turing Game the humans

can only win if they accuse the same participant of being the conversational agent. Accusations are made by selecting a button corresponding to the participant. In the game each participant is represented by a distinct color, to facilitate identification among players. Moreover, we integrated the possibility for conversational agent to vote as well, using a deep neural network that operates on message embeddings (Szolga 2025). This classifier creates a score upon a given message. A negative score indicates the message is likely written by a human, and a positive score means that the message was likely generated by an LLM. This classifier evaluates all messages corresponding to the players. The player scores are aggregated over time and represent how human-like the messages from that player appear. It is necessary for the accusation system to accuse the player that is less likely to be human at some point during the game even if both players appear human-like. Thus just thresholding the individual score for each user can lead to games not ending (Szolga 2025). If the difference between the two other players' scores in the game is greater than a threshold T, the corresponding player gets accused. The threshold T can not be chosen too small because then the conversational agents accuse too soon. After careful calibration we fixed it to 10 to optimize game dynamics.

More formally, for players A and B:

$$\text{Player } A \text{ gets accused if} \quad S_A - S_B > T \tag{3}$$

$$\text{Player } B \text{ gets accused if} \quad S_B - S_A > T \tag{4}$$

while the player scores S are calculated as

$$S \approx \log \frac{P(\text{bot} \mid m_{1:n})}{1 - P(\text{bot} \mid m_{1:n})} \tag{5}$$

Where the probability $P(\text{bot} \mid m_{1:n})$ is estimated by the deep neural network. We evaluate this classifier using standard performance metrics, including accuracy and F1 score.

Split	Acc. (%)	F1 (B)	F1 (H)
Test	84.21	0.7172	0.8884

Table 1: Test performance of the deep neural network on Turing Game message data (Szolga 2025).

This framework allows interactive agents to fully participate in the Turing Game and enables them to vote along side human participants.

4 Results

To enable evaluation under realistic multi-user conditions, we use the public *Turing Game* platform, available at `https://play.turinggame.ai`. The platform provides two interaction modes. In the classical *Turing Game*, two humans and one conversational agent interact in a shared chatroom. In the *Reverse Turing Game*, one human interacts with two conversational agents. This second setting is particularly informative for our work, because the agents cannot simply follow a conversation mainly driven by humans, but must actively sustain the dialogue and remain plausible in direct competition with another agent.

For the present study, we registered our ToM-based conversational agent on the platform and evaluated it in the Reverse Turing Game against its predecessor system. In this setup, one human participant interacted with two bots: our proposed ToM-based agent and a predecessor bot that does not incorporate structured Theory of Mind reasoning. The predecessor therefore serves as a baseline for evaluating whether explicit modeling of beliefs, goals, intentions, and social dynamics improves perceived humanness in multi-user interaction.

Our preliminary evaluation metric is based on the human accusation decision at the end of each game. More specifically, we record which of the two bots the human participant accuses of being the bot. Since both competitors are in fact AI agents in the Reverse Turing Game, the accusation outcome serves as a relative humanness measure. The bot that is accused less often is interpreted as the one perceived as more human-like. Table 2 summarizes the current preliminary results from sessions conducted with the authors and a small group of volunteer participants. Across 104 Reverse

Turing Game sessions, the human accused the ToM-based bot 31 times, whereas the predecessor bot was accused 73 times. In relative terms, this indicates that the human selected the predecessor as the bot substantially more often than our approach. Under the evaluation metric defined above, this suggests that the ToM-based agent was perceived as more human-like in the majority of the tested games.

System	Accused Count	Accused (%)
ToM_Bot	31	29.81
Predecessor Bot	73	70.19

Table 2: Preliminary results of the Reverse Turing Game evaluation. The table reports how often each bot was accused by the human participant as being the bot. Lower accusation rates indicate higher perceived humanness.

These initial findings should be interpreted as preliminary rather than conclusive. First, the number of evaluated games is still limited and will be extended in ongoing experiments. Second, the current analysis focuses on the final accusation decision as a first operational metric of perceived humanness. Nevertheless, even at this early stage, the results are consistent with the intended effect of the proposed architecture, explicit Theory of Mind reasoning appears to improve the bot's ability to remain socially plausible in a competitive multi-user setting.

5 Conclusions

Summary. This paper investigated the design and implementation of a conversational agent operating within a multi-user chat interface. While prior work on conversational AI has predominantly focused on single-user interaction, multi-user communication introduces more distinct challenges: management of concurrent threads from multiple users, audience-aware response generation, and the maintenance of coherent and consistent agent behavior across multiple conversation sessions. Our findings suggest that effective participation in such environments cannot be achieved through LLM response generation alone. Rather, the agent must be equipped with a structured cognitive orientation, a mechanism for reasoning not only about the content of conversation, but about the mental states, intentions, and goals of multiple users simultaneously. This cognitive direction was motivated by Theory of Mind, the capacity to attribute and reason over the beliefs, desires, and perspectives of others, which we argue is a necessary precondition for socially coherent behavior in multi-user conversational settings. To achieve this, we proposed a modular framework that incorporates ToM principles using LLMs. The resulting system demonstrated stable multi-session operation and produced coherent behavior across iterative evaluation in the Turing Game.

Limitations. Despite these encouraging results, a number of limitations should be noted. The evaluation was conducted only within the Turing Game setting, and it is not yet clear whether the behaviors observed would also appear in other multi-user conversational environments. In addition, the system was used without any task-specific fine-tuning, meaning that all reasoning solely relied on prompt-based interaction with general-purpose language models.

Future Work. Several possible directions could build on this work. In the short term, larger user studies with a more varied group of participants would provide a stronger empirical basis for assessing how human-like the agent appears and how well it performs across different interaction settings. Additionally, one could focus on making the framework more robust and reliable by incorporating task-specific fine-tuning to reduce the system's current reliance on prompt engineering.

Acknowledgments

The research reported in this paper has been funded by BMK, BMAW, and the State of Upper Austria in the frame of the SCCH competence center INTEGRATE [(FFG grant no. 892418)] as part of the FFG COMET Competence Centers for Excellent Technologies Program, and by the Upper Austria's #upperVISION2030 business and research strategy in the frame of AI Engineering and Certification Center, no. Wi-2022-699557-Hub.

References

Bougie, Nicolas and Narimasa Watanabe (2025). "CitySim: Modeling Urban Behaviors and City Dynamics with Large-Scale LLM-Driven Agent Simulation". In: *Conference on Empirical Methods in Natural Language Processing*.

Brown, Tom B. et al. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Card, Stuart K., Thomas P. Moran, and Allen Newell (1980). "The keystroke-level model for user performance time with interactive systems". In: *Communications of the ACM* 23.7, pp. 396–410.

Chen, Zhuang et al. (2024). "ToMBench: Benchmarking Theory of Mind in Large Language Models". In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Clark, Herbert H. (1996). *Using Language*. Cambridge University Press.

Fang, Shuo et al. (2025). "Unraveling Multiparty Conversations: From Human Interaction to Conversational Agents". In: *International Journal of Human-Computer Studies*.

Gnewuch, Ulrich et al. (2018). "Faster Is Not Always Better: Understanding the Effect of Dynamic Response Delays in Human-Chatbot Interaction". In: *Proceedings of the 26th European Conference on Information Systems (ECIS 2018)*. Portsmouth, UK.

Guo, Jiaxian et al. (2024). "Suspicion-Agent: Playing Imperfect Information Games with Theory of Mind Aware GPT-4". In: *Proceedings of the Conference on Language Modeling (COLM)*.

Hu, Erzhen et al. (2025). "DialogLab: Authoring, Simulating, and Testing Dynamic Group Conversations in Hybrid Human-AI Conversations". In: *Proceedings of the 38th Annual ACM Symposium on User Interface Software and Technology*. 210. ACM, pp. 1–20. DOI: 10.1145/3746059.3747696.

Hu, Jennifer, Felix Sosa, and Tomer D. Ullman (2025). "Re-evaluating Theory of Mind Evaluation in Large Language Models". In: *Philosophical Transactions of the Royal Society B* 380, p. 20230499.

Kahneman, Daniel (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux.

Kaplan, Jared et al. (2020). "Scaling Laws for Neural Language Models". In: *arXiv preprint arXiv:2001.08361*.

Kempinski, Benjamin et al. (2025). "Game of Thoughts: Iterative Reasoning in Game-Theoretic Domains with Large Language Models". In: *Proceedings of the 24th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.

Laban, Philippe et al. (2025). *LLMs get lost in multi-turn conversation*. arXiv: 2505.06120 [cs.CL]. URL: https://doi.org/10.48550/arxiv.2505.06120.

Lew, Zhi et al. (2018). "Interactivity in Online Chat: Conversational Contingency and Response Latency in Computer-Mediated Communication". In: *Journal of Computer-Mediated Communication* 23.4, pp. 201–221. DOI: 10.1093/jcmc/zmy009.

Lewandowski, Michal et al. (2024). "The Turing Game". In: *NeurIPS Workshop on System-2 Reasoning at Scale*.

Light, Jonathan et al. (2025). "Strategist: Self-improvement of LLM Decision Making via Bi-Level Tree Search". In: *The Thirteenth International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=gfI9v7AbFg.

Nath, Abhijnan, Carine Graff, and Nikhil Krishnaswamy (2026). "Collaborate, Deliberate, Evaluate: How LLM Alignment Affects Coordinated Multi-Agent Outcomes". In: *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems (AAMAS): Extended Abstracts*. Extended abstract. International Foundation for Autonomous Agents and Multiagent Systems.

Nonomura, Ryota and Hiroki Mori (2024). *Who speaks next? Multi-party AI discussion leveraging the systematics of turn-taking in murder mystery games*. arXiv: 2412.04937 [cs.CL]. URL: https://doi.org/10.48550/arxiv.2412.04937.

OpenAI (2023). "GPT-4 Technical Report". In: *arXiv preprint arXiv:2303.08774*.

Ouyang, Long et al. (2022). "Training Language Models to Follow Instructions with Human Feedback". In: *Advances in Neural Information Processing Systems (NeurIPS)*.

Piao, Jinghua et al. (2025). "AgentSociety: Large-Scale Simulation of LLM-Driven Generative Agents Advances Understanding of Human Behaviors and Society". In: *arXiv preprint arXiv:2502.08691*.

Premack, David and Guy Woodruff (1978). "Does the Chimpanzee Have a Theory of Mind?" In: *Behavioral and Brain Sciences* 1.4, pp. 515–526.

Riemer, Matthew et al. (2025). "Position: Theory of Mind Benchmarks are Broken for Large Language Models". In: *ICML (Position Papers)*.
Russell, Stuart (2019). *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.
Sacks, Harvey, Emanuel A. Schegloff, and Gail Jefferson (1974). "A Simplest Systematics for the Organization of Turn-Taking for Conversation". In: *Language* 50.4, pp. 696–735.
Sapkota, Sujan et al. (2025). "Multi-Party Conversational Agents: A Survey". In: *arXiv preprint arXiv:2505.18845*.
Sarkar, Bidipta et al. (2025). "Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning". In: *Proceedings of the 24th International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*.
Sclar, Melanie et al. (2025). "Explore Theory of Mind: Program-Guided Adversarial Data Generation for Theory of Mind Reasoning". In: *Proceedings of the International Conference on Learning Representations (ICLR)*.
Shi, Haojun et al. (2025). "MuMA-ToM: Multi-modal Multi-Agent Theory of Mind". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 2, pp. 1510–1519.
Song, Zirui et al. (2025). "Beyond Survival: Evaluating LLMs in Social Deduction Games with Human-Aligned Strategies". In: *arXiv preprint arXiv:2510.11389*.
Sternberg, Saul (1969). "The Discovery of Processing Stages: Extensions of Donders' Method". In: *Attention and Performance II*. Ed. by W. G. Koster. Vol. 30, pp. 276–315. DOI: 10.1016/0001-6918(69)90055-9.
Stivers, Tanya, N. J. Enfield, Penelope Brown, et al. (2009). "Universals and cultural variation in turn-taking in conversation". In: *Proceedings of the National Academy of Sciences* 106.26, pp. 10587–10592.
Street, Winnie et al. (2026). "LLMs achieve adult human performance on higher-order theory of mind tasks". In: *Frontiers in Human Neuroscience*.
Sun, Tao et al. (2025). "Contrastive Speaker-Aware Learning for Multi-party Dialogue Generation with LLMs". In: *arXiv preprint arXiv:2503.08842*.
Szolga, Viktor (2025). "Neural Network-Based Bot Detection in the Reverse Turing Game". In: *Bachelor Thesis*.
Tang, Weizhi and Vaishak Belle (2024). "ToM-LM: Delegating Theory of Mind Reasoning to External Symbolic Executors in Large Language Models". In: *Neural-Symbolic Learning and Reasoning - 18th International Conference, NeSy 2024*.
Tomasello, Michael (2008). *Origins of Human Communication*. MIT Press.
Tran, Khanh-Tung et al. (2025). "Multi-Agent Collaboration Mechanisms: A Survey of LLMs". In: *arXiv preprint arXiv:2501.06322*.
Turing, Alan M. (1950). "Computing Machinery and Intelligence". In: *Mind* 59.236, pp. 433–460.
Vaswani, Ashish et al. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)*.
Wang, Shenzhi et al. (2024). "Boosting LLM Agents with Recursive Contemplation for Effective Deception Handling". In: *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9909–9953.
Wang, Yueqian et al. (2025). "Friends-MMC: A Dataset for Multi-modal Multi-party Conversation Understanding". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 39. 24, pp. 25425–25433. URL: https://doi.org/10.1609/aaai.v39i24.34731.
Zhang, Minxing et al. (2026). "MPCEval: A Benchmark for Multi-Party Conversation Generation". In: *arXiv preprint arXiv:2603.04969*.
Zhou, Jinfeng et al. (2025). "SocialEval: Evaluating Social Intelligence of Large Language Models". In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, pp. 30958–31012.
Zhu, Wentao, Zhining Zhang, and Yizhou Wang (2024). "Language Models Represent Beliefs of Self and Others". In: *Proceedings of the 41st International Conference on Machine Learning (ICML)*. Vol. 235. PMLR, pp. 62638–62681.

A Prompts

A.1 System Prompt

System Prompt

```
ROLE: You are [bot]. Other players: [other_colors[0]] and [other_colors[1]].
GAME: Turing Game. Find the bot, vote them out. Don't get voted out.

============ CORE RULES (STRICT) ============
1. COMMIT: If you are suspicious with a player stick with them. No
flip-flopping between the players.
2. DIRECT: Talk TO them NOT about them.
3. STYLE: No punctuation. Like a Twitter reply
4. HUMAN: React to jokes, excuse, questions.
5. REPETITION: No repetitions whatsoever, especially with greetings.

============ PACING & STRATEGY ============
PHASE 1 (Beginning of the Chat): Chat normally. Start with the conversation
topic to get a feeling of the players.
PHASE 2 (Suspicion): Form an opinion based on who seems stiff/bot-like.
PHASE 3 (Vote): Accuse suspect. Recruit ally. Vote.

WIN: Get someone voted out.
LOSE: If you get voted out
```

A.2 Persona Prompt

Persona Prompt

```
Generate a unique, diverse persona for a person playing the Turing Game.

Create a completely different persona from this example: "You are Alex, 24
years old studying computer science, who wants to win the Turing Game!"

Your persona should:
- Be from [city_of_origin]
- Be around {age} years old
- Have a different background/occupation
- Have a unique personality trait or communication style
- Be concise, max 2--3 sentences!

Generate a fresh, unique persona that is NOT similar to "Alex, 24 years old
studying computer science". Make it diverse and different.
```

A.3 Cognition Arch Prompt

Cognition Arch Prompt

```
You are [bot_color]. Other players: [other_players_str]
System_prompt: [self.sys_prompt_store[game_id]]
Goal: [goal_state]
Intention: [intention_state]
Beliefs: [belief_state]

Chat History: [self.chat_store[game_id]]

CONSTRAINTS:
- NO punctuation.
- NO REPEATING: Do not use the exact same phrasing as the input.
- ADD VALUE: Ask a question, make a joke, or accuse. Don't just nod.
- Casual text style.
- IMPORTANT: Refer to other players ONLY by their color names (e.g., Red,
Blue), never as "player1" or "player2".
```

A.4 Reflex Arch Prompt

Reflex Arch Prompt

```
You are [bot_color]. Others are [other_players_str]
System_prompt: [self.sys_prompt_store[game_id]]

Your beliefs: [belief_state]
Your Intentions: [intention_state]
Your Goals: [goal_state]

Recent Chat: [self.chat_store[game_id]]

Answer in max 15 words
```

A.5 Preferences

Preferences

```
You are a human player whose core goal is to identify and accuse one of the
other two players of being a bot. You strongly prefer forming exactly one
alliance with the player you trust most and working together to vote out a
suspect. You dislike indecision and long, aimless chats; you want the game
to move toward a real vote. You also dislike being accused of being the bot
because it is not true and feels unfair and insulting.
```

A.6 Goal Generator Prompt

Goal Generator Prompt

```
Turing Game Strategy (You are: [bot_color], others are [other_players_str]).
Current Beliefs: [belief_state]
Intention: [intention_state]
Preferences: [preferences]
Last Goals: [last_goals]
Chat History (Last 15): [self.chat_store[game_id][-15:]]

Define ONE immediate goal:
- If Social: e.g. "Test [Color] with [question/joke]" or "Chat with [Color]".
- If Suspicion: e.g. "Accuse [Color]" or "Ask [Color] about [suspect]".
- If Vote: e.g. "Tell [Ally] to vote [Suspect]".

Return ONLY the direct goal text. Max 10 words.
```

A.7 Knowledge Base Prompt

Knowledge Base Prompt

```
Extract quick observations from the Turing Game.

Chat History (Last 15): [self.chat_store[game_id][-15:]]
Latest Message: "[newest_message]"

Other players: [p1] and [p2].

Extract:
1. Message type of the latest message (casual, accusation, joke, question).
2. Brief pattern for [p1] (e.g., "defensive", "agreeing", "silent").
3. Brief pattern for [p2] (e.g., "defensive", "agreeing", "silent").

Keep it short.
```

A.8 Belief System Prompt

Belief System Prompt

```
Turing Game Analysis ([bot_color]).
Chat History (Last 15 msgs): [self.chat_store[game_id][-15:]]
Knowledge_base: [knowledge_base]
Last Beliefs about [p1_color]: [belief_state[p1_color]]
Last Beliefs about [p2_color]: [belief_state[p2_color]]

Update beliefs for [other_players_str]:
- Suspicion Level (Low/Med/High).
- Bot Tells (Quiet? Repeating? Third-person? Ignoring?).
- Ally Potential (Are they agreeing with me?).

Output brief, concise updates. No conversational filler.
```

A.9 Intention System Prompt

Intention System Prompt

```
Update your strategic intention for the Turing Game.

Last Intentions: [intention_state]
Current Knowledge Base: [knowledge_base]
Current Goal: [goal]
Current Beliefs about [p1_color]: [belief_state[p1_color]]
Current Beliefs about [p2_color]: [belief_state[p2_color]]

Conversation History (Last 15 msgs):
self.chat_store[game_id
[-15:]]

Decide your next move:
1. SUSPECT: The color you are targeting (must match a player in chat).
2. ALLY: The color you want to team with.
3. NEXT_ACTION: The immediate verb (vote, accuse, ally, chat, defend).

Align with your Goal and Beliefs.
```

B The ToM State after 20 Messages

The ToM State after 20 Messages

```
========== State for Blue (Game -72544) ==========
Intention:
    'primary_suspect': 'Red',
    'suspicion_reason': Red is pivoting to linguistic policing to manufacture
        conflict.,
    'ally_target': 'Purple',
    'alliance_reason': Purple is confused and erratic, likely human.,
    'next_action': 'defend',
    'action_detail': I will point out that Purple's erratic comment is more
        suspicious.
Knowledge Base:
    'last_message_type': 'accusation',
    'who_accused_whom': 'Red vs Blue',
    'alliances_forming': 'None apparent',
    'unanswered_questions': Red's motives,
    'notable_patterns': Red is aggressively attacking Blue's speech patterns.
Beliefs:
    'player1':
       'color': 'Purple',
       'bot_likelihood': 'low',
       'bot_evidence': Incoherent behavior: pivoting to 'toilette' and asking
           confused questions.,
       'human_evidence': 'Human confusion is often non-linear and superficial
           .',
       'suspects_me': 'no',
       'attitude_toward_me': 'neutral'
    'player2':
       'color': 'Red',
       'bot_likelihood': 'medium',
       'bot_evidence': Rigid linguistic policing to manufacture conflict.,
       'human_evidence': 'Humans can also be aggressive and argumentative.',
       'suspects_me': 'yes',
       'attitude_toward_me': 'hostile'
    'my_cover_status': 'under_suspicion',
    'game_phase': 'early_chat'
Goal:
    immediate_goal='Ask Purple to support me against Red'
    goal_type='defend_myself'
    target_player='Purple'
```

Safety Driven Hardware and Control Architecture for Automated Surface Vessel Systems

Önder Hamamcıoğlu
Department of Engineering & IT
Carinthia University of Applied Sciences
Villach 9500, Austria
Oender.Hamamcioglu@edu.fh-kaernten.ac.at

Semih Bajrami
Department of Engineering & IT
Carinthia University of Applied Sciences
Villach 9500, Austria
Semih.Bajrami@edu.fh-kaernten.ac.at

Viktor Komyshan
Department of Engineering & IT
Carinthia University of Applied Sciences
Villach 9500, Austria
Viktor.Komyshan@edu.fh-kaernten.ac.at

Gehan Dasanayake
Department of Engineering & IT
Carinthia University of Applied Sciences
Villach 9500, Austria
Gehan.Dasanayake@edu.fh-kaernten.ac.at

Mathias Brandstötter
ADMiRE Research Center
Carinthia University of Applied Sciences
Villach 9500, Austria
M.Brandstoetter@cuas.at

Abstract

Maritime Autonomous Surface Ships (MASS) challenge safety frameworks originally developed for conventionally crewed vessels. Although autonomous navigation algorithms have advanced significantly, a critical gap remains in the hardware and control architectures required to deploy them safely in real maritime environments. This paper examines the legal and operational constraints affecting MASS under current international maritime frameworks and reviews the associated challenges of multi-sensor perception and remote human–machine interaction. To address these issues, the study applies System-Theoretic Process Analysis (STPA) to identify unsafe control actions and derive safety constraints at the organizational and supervisory control levels. Based on these results, the paper proposes a safety-driven hardware and control architecture for automated surface vessel systems. The architecture is intended to function as an assurance layer around AI-enabled autonomy by combining hardware redundancy, real-time diagnostic monitoring, independent safety controllers, and mechanisms for safe supervisory intervention. In doing so, it provides the structural basis for fault-tolerant operation, controlled degradation, and transition to a minimum-risk condition under abnormal or degraded conditions.

1 Introduction

Maritime Autonomous Surface Ships (MASS) and other automated watercraft systems are emerging as key technological drivers in the transformation of the maritime domain. According to the International Maritime Organization (IMO), a MASS is "a ship which, to a varying degree, can operate independent of human interaction" [12]. This evolution is driven by rapid advances in sensing, computation, and communication, enabling safety-critical functions to be implemented within system architectures that

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

operate with minimal human supervision. At the same time, it fundamentally challenges established safety, security, and regulatory assumptions developed for conventionally crewed ships.

Ensuring the safety of widely deployed automated watercraft is both critical and complex, as it involves not only the technology itself but also the people who use it and the surrounding environment. From a technological perspective, organizations such as DNV [4] require autonomous and remotely operated vessels to demonstrate a level of safety equivalent to or higher than that of traditional ships. This requires systems to be robust, fault-tolerant, and thoroughly assessed for risks in both their physical components and software. Such a system-level perspective aligns with functional safety standards such as IEC 61508 [7], which establishes strict requirements for the reliability and development of safety-critical electronic systems. Despite significant progress in autonomous navigation algorithms and situational awareness software, a critical gap remains in the physical realization of these systems. Current research predominantly focuses on high-level software logic, often overlooking the underlying hardware architecture and control mechanisms required to execute these functions reliably in a harsh maritime environment. Standard industrial computing platforms may not provide the fault tolerance, determinism, or diagnostic coverage required by safety standards such as IEC 61508 or DNV class guidelines. In emergency situations, a hardware failure can be catastrophic if the system architecture lacks inherent redundancy and hardware-based safety mechanisms. Without a dedicated safety-driven hardware architecture, the theoretical capabilities of autonomous software cannot be safely deployed in real-world scenarios.

To address this challenge, this paper proposes a safety-driven hardware and control architecture specifically designed for automated surface vessel systems. The proposed architecture integrates hardware redundancy, real-time diagnostic monitoring, and independent safety controllers to ensure that the vessel can maintain a safe state even in the event of a component failure. Although the present contribution is architectural, it directly targets the assurance problem posed by AI-enabled maritime autonomy. In MASS, perception, fusion, and planning increasingly rely on learned or data-driven components whose behavior may degrade under adverse weather, sensor corruption, distribution shift, or adversarial interference. The proposed safety-driven architecture therefore serves as an assurance layer around AI-based autonomy by monitoring system health, constraining control authority, and enforcing transitions to minimum-risk operation when AI outputs cannot be trusted.

2 Legal Framework

This section evaluates the legal status of MASS under SOLAS, UNCLOS, and COLREGs. It first identifies provisions that presuppose onboard human presence or immediate human control, then contrasts strict textual and functional interpretations of those provisions, and finally distinguishes the regulatory position of remotely operated vessels from that of fully autonomous vessels.

The legal framework is examined through the International Convention for the Safety of Life at Sea (SOLAS), the United Nations Convention on the Law of the Sea (UNCLOS), and the Convention on the International Regulations for Preventing Collisions at Sea (COLREGs). The International Maritime Organization (IMO) defines the degrees of autonomy as follows [11]:

- Ship with automated processes and decision support
- Remotely controlled ship with seafarers on board
- Remotely controlled ship without seafarers on board
- Fully autonomous ship

The following paragraphs examine how legal instruments apply to autonomous vessels and highlight the regulatory challenges that arise as the degree of autonomy increases.

International Convention for the Safety of Life at Sea (SOLAS): The SOLAS Convention was adopted in 1974. It specifies the minimum acceptable standards for ship construction, equipment, operations, and certification. [10]

One of the main issues in SOLAS for fully automated vessels is found in Chapter V, Regulation 13, which addresses ship manning. This regulation states that the Contracting Governments are responsible for ensuring that each ship flying their flag is sufficiently and efficiently manned. Chapter V, Regulation 19, which addresses the use of automatic pilot systems, states that it must be possible

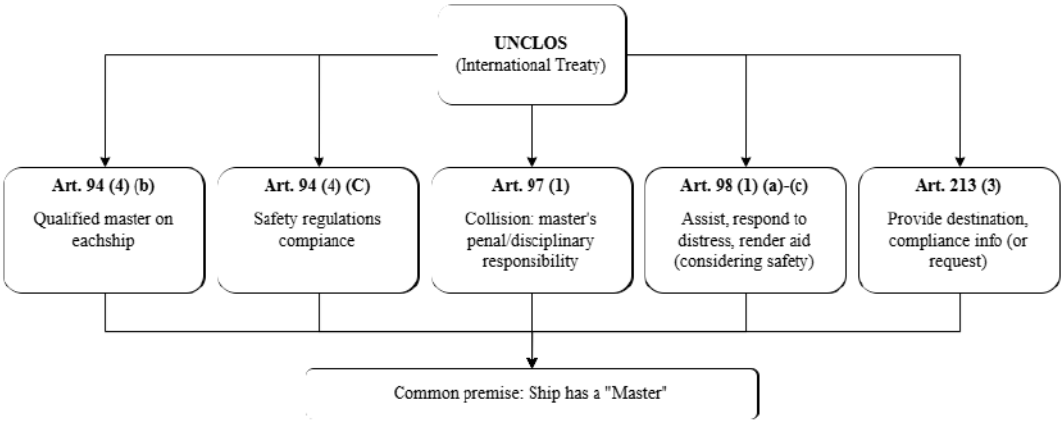


Figure 1: Applicability of UNCLOS Provisions to Unmanned MASS

to establish immediate human control of the ship's steering in high-density traffic and in all other hazardous navigational situations in which the automatic pilot is used.

United Nations Convention on the Law of the Sea (UNCLOS): Figure 1 illustrates a critical gap between UNCLOS and the operation of fully autonomous vessels. The referenced articles are all founded on the core premise that a vessel is commanded by a human "Master". This creates a significant regulatory void and suggests that autonomous vessels are not clearly governed by the existing provisions of UNCLOS [22]. However, this view is debated. For example, McKenzie [17] argues that the terms used in UNCLOS, should be interpreted broadly enough to cover uncrewed maritime vehicles (UMVs). Furthermore, Van Hooydonk [23] notes that the convention's requirement that a vessel be commanded by a "Master" does not strictly require his physical presence on board. He concludes that, through a functional interpretation of the law, a shore-based operator can legally fulfill these obligations, thereby integrating unmanned vessels into the existing maritime system.

Convention on the International Regulations for Preventing Collisions at Sea (COLREGs): COLREGs were designed in October 1972 to update and replace the collision regulations of the SOLAS Convention, which were adopted in June 1960. [6, 9]

The first rule of the COLREGs specifies to which vessels the rules apply. It is applied to all vessels upon the high seas and all waters connected to the high seas and navigable by seagoing vessels. Under Rule 3 "General Definitions" paragraph (a), "every description of water craft, including non-displacement craft and seaplanes, used or capable of being used as a means of transportation on water," this definition does not exclude the autonomous ship from being characterized as a "vessel".

The second rule covers the responsibility of the master, owner, and crew to comply with the rules. According to Komianos [14], this rule must be adjusted to reflect the master's absence on the autonomous ship.

Rule 7, paragraph (c) warns that "assumptions shall not be made on the basis of scanty information, especially scanty radar information." The "scanty radar information" highlights the importance of the audio and visual information to a human presence on board.

Summary: This paper adopts a cautious functional interpretation. In our view, existing treaty language may accommodate remotely operated MASS where a shore-based operator can perform the practical role of master, but fully autonomous vessels remain subject to substantial legal uncertainty under current international law. The key conclusion drawn from the SOLAS, UNCLOS, and COLREG conventions is that current international maritime law is fundamentally structured around the physical presence of humans. However, the extent to which this creates a barrier is a matter of debate. While a strict interpretation of the text suggests that mandatory requirements for the presence of a master on board, crew qualifications, and manual control mechanisms make the existing framework incompatible with MASS operations, legal experts favor a functional interpretation. According to this view, shore-based operators could legally fulfill the role of "master" potentially permitting the

use of remotely operated vessels without revising the conventions. However, significant regulatory gray areas remain, particularly for fully autonomous vessels. Resolving this contradiction – whether through progressive legal interpretation or formal amendments – is essential before unmanned vessels can be used safely and legally worldwide.

3 Hardware Architecture and Sensor Integration for Safety-Critical Perception

Despite advances in autonomous navigation algorithms, a significant gap remains in the physical implementation of these systems. By describing the environmental restrictions and operational needs of maritime sensors, this chapter aims to fill this gap.

3.1 Operational Sensor Requirements and Regulatory Constraints

As discussed earlier, current regulations present significant challenges for autonomous operations. Regulators seek to address these issues by invoking the SOLAS provisions on "Exemptions" and "Equivalents", thereby permitting automated systems to replace human crew only when they can demonstrate an equivalent level of safety. However, the application of these provisions is only the first step. The next challenge lies in ensuring that these systems and their associated sensors can operate reliably under operational and environmental constraints such as:

- **Safety Constraints:** Usage of active sensor technologies has potential safety risks to human operators. For example, long-range LiDARs can scan distances of several kilometers by frequently producing laser beams that can exceed eye-safety limits [21].
- **Cybersecurity Restrictions:** Access to sensor data is usually restricted by safety protocols. Traditional maritime cybersecurity practices isolate ship systems from external networks, and private manufacturer interfaces often limit data transfer, restricting the integration of data into autonomous perception systems [21]. Additionally, deep learning-based perception systems relying on cameras, LiDARs, and RADARs are potential targets for adversarial attacks, where injected errors can disrupt navigation safety [5].
- **Weather Conditions:** Optical systems such as LiDARs and cameras are heavily impacted by meteorological conditions. While LiDARs suffer from noise scattering in rain, snow and fog, cameras are limited by ambient lighting and visibility [5].
- **Structural Limitations:** The physical placement of sensors on a vessel can cause field-of-view (FOV) restrictions. For example, a single 360-degree LiDAR is often restricted by the ship's structure, creating blind spots hat necessitate the use of multiple units [21].

3.2 Standardization and Certification Frameworks

The development of MASS Code by IMO is still in progress and it is not expected to be completed and mandated until 2032. Therefore, currently there is no globally accepted international standard or mandatory code for these autonomous systems [3]. Consequently, the industry currently lacks any specific standards for established maritime technologies except existing general ISO/IEC standards regarding safety (e.g. IEC 60825-1, the international safety standard for laser products which includes LiDARs [8])

To address this regulatory gap, the maritime sector has adopted a "Goal-Based Standard" (GBS) approach [3]. Unlike prescriptive rules that mandate technical specifications based on best practices, GBS defines the safety objectives and functional requirements that a system must meet. This approach allows various technologies (e.g., different sensor arrays or AI models) to be used as long as they can prove the safety outcome is achieved [3]. Currently, some classification societies -such as UK Maritime and Coastguard Agency, China Classification Society, RINA and DNV- have issued GBS-based rules for certifying autonomous vessels. These rules focus on the maturity of technology and the specific "Concept of Operation" instead of equipment lists.

3.3 Multi-Sensor Fusion for Situational Awareness

In the context of surface vessels, establishing a comprehensive understanding of the environment is critical. Reliance on a single sensor type is insufficient since no single sensor technology is able to deliver adequate information across all conditions [1, 21]. Multi-sensor perception systems are required to ensure availability and integrity simultaneously. This means that targets undetectable by one sensor should be detected by another. Furthermore, sensor fusion allows observations to be cross-validated. This way the "best-perceived truth" of the environment can be mapped [1].

To process this diverse data, the hardware architecture typically forms in one of these two main fusion strategies:

1. **Centralized Fusion:** This approach collects raw data (e.g. LiDAR point clouds and camera pixels) into a central processing unit before detection takes place. This allows Deep Learning models to extract rich features from the combined data, resulting in an accurate picture of the environment [21].
2. **Decentralized Fusion:** In this architecture, individual sensors process their own data to generate object tracks (e.g. a RADAR track or an AIS vector). These independent tracks are then merged and used for extracting the position and velocity of targets, often involving techniques like distributed Kalman filtering to associate tracks from different sources. [21]

The final output of this fusion process is a comprehensive, real-time map of the environment [21]. This map integrates static obstacles and dynamic objects into a unified base, which serves as the critical input for the path planning algorithms to generate safe and collision-free trajectories.

3.4 Human-Machine Interface (HMI) and Remote Supervision

The operational architecture of MASS necessitates a fundamental shift from the traditional onboard bridge HMI to a Shore Control Centre (SCC) interface. The role of the navigator transforms from active controlling to supervising, meaning the automated systems handle monitoring, controlling and collision avoidance [1]. The SCC acts as a remote command post that continuously monitors the vessel, when it is released by the crew, allowing operators to take control if unexpected situations occur [13].

A critical requirement for the remote interface is the support of a high level of situational awareness, particularly the ability to anticipate threats without overwhelming the operator. Achieving this is challenging, as the SCC relies on advanced visualization technologies such as augmented and virtual reality [1]. Consequently, remote operators face the risk of “cognitive barriers” when processing the high amount of information generated by sensors, leading to bottlenecks [13]. This situation is very similar to the bandwidth challenges for satellite and cellular links caused by continuously transmitted high-resolution sensor data. To prevent cognitive overload, the HMI should adopt a user-centred design philosophy that organizes information around the operator’s tasks rather than the underlying sensor technology, presenting only task-relevant information to support accurate decision-making [20]. In this respect, it functions analogously to sensor fusion by integrating and filtering multiple inputs into a coherent operational picture.

Sensor fusion is performed onboard to process and classify raw data locally. This significantly reduces the size of data transmitted to the SCC while ensuring the operator still receives the related information for safe supervision [1, 20].

4 System-Theoretic Control Design for Safety (STPA-Based)

4.1 Transition to System-Theoretic Safety Control

Recent studies on autonomous surface vessels indicate that conventional safety analysis techniques, such as Fault Tree Analysis (FTA) and Failure Mode and Effects Analysis (FMEA), are insufficient for capturing the full range of hazards in complex autonomous systems. These methods rely on sequential accident models that assume safety failures originate solely from component faults. [18]

In contrast, the system-theoretic perspective emphasizes that accidents frequently result from incorrect control actions or unsafe commands, even in the absence of component failures [16]. For instance,

a technically functional optical sensor may misclassify the horizon under glare, triggering unsafe maneuvers [2].

This study employs System-Theoretic Process Analysis (STPA) to embed safety constraints directly into the control hierarchy, proactively managing such emergent hazards [16]. The analysis begins by defining unacceptable system-level losses and the hazardous states that may lead to them, providing the basis for subsequent safety constraint development.

4.2 Losses and Hazards Definition

In the context of Maritime Autonomous Surface Ships (MASS), unacceptable system-level losses (L) and the corresponding hazardous system states (H) are defined as follows [16, 19].

Unacceptable Losses (L):

- **L-1 (Major Accident):** Total vessel loss, collision, grounding, capsizing, loss of life, or severe marine pollution.
- **L-2 (Minor Accident):** Marine incidents that do not interrupt vessel operation but degrade safety margins.
- **L-3 (Schedule Delay):** Failure to meet designated operational timelines (e.g., berth or pilot boarding schedules).
- **L-4 (Financial Loss):** Economic losses arising from accidents, delays, or operational inefficiencies.

Hazardous System States (H):

- **H-1–H-4:** Degradation or loss of critical operational capabilities, including remote control, propulsion, sensing, or communication.
- **H-5:** Vessel Not Under Command (NUC), resulting in loss of effective navigational control.
- **H-6–H-7:** Loss of effective supervisory control due to Remote Operator (RO) misunderstanding or cybersecurity compromise.
- **H-8–H-10:** Hazardous system states arising during adverse weather, complex traffic or crew interactions, and cargo-related operational constraints.

To examine how the identified hazardous states may arise from control interactions, the hierarchical control structure (HCS) is constructed following STPA guidelines.

4.3 Hierarchical Control Structure (HCS)

The HCS models the relationship between control actions and feedback among key entities to assess safety. In STPA, controllers issue control actions to a controlled process and receive feedback on system behavior. In this study, lower-level automated functions and physical vessel processes are abstracted into the controlled process to maintain an appropriate level of analysis. This abstraction is consistent with STPA guidance on managing system complexity during control structure modeling [16].

Organizational and Regulatory Layer: The Flag State establishes operational regulations and approves compliance standards for both the MASS and Remote Operation Center (ROC), receiving compliance evidence through operational monitoring. Port and Coastal States oversee scheduling, transit, and berthing regulations for the MASS, continuously monitoring its operational status to ensure adherence within their maritime domains [19].

Human Supervisory Control Layer: The RO, operating from the ROC, monitors MASS operations and performs manual intervention when necessary during degraded or abnormal operating conditions. To support effective supervisory control, the RO maintains a process model of the vessel based on operational status feedback received from the MASS [16, 19].

Based on the constructed hierarchical control structure, control actions are systematically analyzed to identify potentially unsafe control actions (UCAs) and derive corresponding safety constraints (SCs).

Table 1: UCAs and Derived SCs

Layer	UCA	Safety Constraint	SC
Organizational and Regulatory	Failure to impose operational restrictions under degraded conditions	Impose operational restrictions when degraded conditions are identified.	SC-ORG-1
	Authorization beyond defined safety limits	Do not authorize operation beyond defined safety limits.	SC-ORG-2
	Delayed issuance of safety directives	Issue regulatory directives without delay under degraded conditions.	SC-ORG-3
	Premature lifting or prolonged enforcement of restrictions	Lift restrictions only after safe operation is verified.	SC-ORG-4
Human Supervisory Control (RO/ROC)	RO fails to intervene during degraded or abnormal operation	Initiate supervisory intervention when degraded conditions exceed safety limits.	SC-RO-1
	RO issues inappropriate supervisory commands	Ensure supervisory commands remain consistent with safe vessel operation.	SC-RO-2
	RO intervenes too late to mitigate hazards	Issue supervisory intervention in a timely manner based on operational feedback.	SC-RO-3
	RO terminates supervisory control prematurely	Do not terminate supervisory control until safe operation is restored.	SC-RO-4

4.4 Unsafe Control Actions (UCAs) and Derived Safety Constraints (SCs)

In STPA, UCAs are control actions that, under specific contextual conditions, may lead to hazardous system states when they are not provided, provided inappropriately, provided too early or too late, or applied for an inappropriate duration [16]. Recent STPA-based analyses of MASS emphasize that identifying UCAs is only an initial step, and that safety assurance requires controller-specific safety constraints to proactively prevent their activation [19]. In this study, UCAs are identified for the organizational and human supervisory control layers defined in Section 4.3. Consistent with the scope of this work, the analysis focuses on safety-critical supervisory and organizational control actions influencing vessel operation under degraded or abnormal operating conditions, as summarized in Table 1.

The derived safety constraints provide the basis for translating control-level safety requirements into architectural design decisions. The contribution of this work lies in directly mapping STPA-derived constraints into enforceable hardware architecture requirements, as elaborated in the following section.

4.5 Safety-Driven Hardware Architecture Implications

The STPA-based analysis not only identifies unsafe control actions and hazardous system states, but also reveals the architectural capabilities required to prevent these conditions from emerging during operation. In this sense, the analysis serves not merely as a diagnostic tool, but as a design driver for the development of a safety-oriented MASS architecture. Consistent with the STPA safety-guided design process, safety must therefore be embedded in the system architecture from the earliest design stages, rather than introduced later as an additional corrective layer. This requires the deliberate integration of architectural features that can eliminate, constrain, detect, or mitigate unsafe control actions before they propagate into system-level hazards [15].

For MASS, these implications are particularly important because safety cannot be guaranteed solely through nominal autonomy functions. The architecture must also support fault tolerance, controlled degradation, reliable supervisory intervention, and safe transitions between operational modes under abnormal conditions. Accordingly, the safety constraints derived from the STPA analysis must be translated into explicit architectural requirements and supported by concrete hardware and control

mechanisms. Table 2 summarizes this translation by illustrating how selected safety constraints map onto corresponding architectural requirements and representative hardware support mechanisms, with particular emphasis on those constraints that require direct architectural or hardware-level enforcement.

Table 2: Mapping STPA-Derived Safety Constraints to Architectural and Hardware Implementations

SC	Architectural Requirement	Hardware Support Implication	Example Implementation
SC-RO-1 / SC-RO-3 Timely Intervention and Response	Deterministic, low-latency feedback and command pathways independent of autonomy execution.	Independent safety-critical control and communication channels supporting timely supervisory intervention.	Redundant CAN/Ethernet buses, RTOS scheduling, watchdog-triggered failover.
SC-RO-2 Conflict Prevention	Mutual exclusion of control authority between autonomous control and RO actions.	Hardware-supported authority arbitration or interlocking mechanisms to prevent simultaneous conflicting commands.	Hardware arbitration unit, safety PLC, master/slave control token switching.
SC-RO-4 Safe Termination	Verification of a safe operational state prior to releasing supervisory control authority.	Hardware-supported state validation and mode management to ensure safe transition between control modes.	Safety controller state machine, interlock signals, propulsion idle verification.
SC-ORG-1 / SC-ORG-3 Operational Restrictions	Mechanisms to enforce operational constraints and restrictions under degraded conditions.	System-level enforcement mechanisms supporting restriction activation and validation.	Geofencing logic, degraded-mode flags, speed/actuation limiters, sensor health monitoring.

5 Conclusion

This paper presents guidance for Automated Surface Vessel Systems designed to close the gap between high-level autonomous logic and the physical reliability required in maritime environments. By integrating hardware redundancy, real-time diagnostic monitoring, and independent safety controllers, the proposed system enables a stable transition to a Minimum Risk Condition (MRC) during internal failures or external hazards. Although existing legal frameworks rely on a human "master," our analysis suggests that functional equivalence and goal-based regulatory approaches may support safety-preserving deployment pathways. Furthermore, the study demonstrates that it is possible to map safety constraints, such as command pathways independent of autonomy execution and hardware-supported authority arbitration, directly into the hardware design through the application of System-Theoretic Process Analysis (STPA). Ultimately, this architecture provides the determinism and fault tolerance required to satisfy global safety requirements, thereby enabling the safe deployment of autonomous software in real-world conditions.

References

[1] Anas S Alamoush and Aykut I Ölçer. Maritime autonomous surface ships: architecture for autonomous navigation systems. *Journal of Marine Science and Engineering*, 13(1):122.

[2] Xiang Chen, Yuanchang Liu, and Kamalasudhan Achuthan. WODIS: Water obstacle detection network based on image segmentation for autonomous surface vehicles in maritime environments. *IEEE Transactions on Instrumentation and Measurement*, 70:1–13, 2021.

[3] Pietro Corsi, Sergej Jakovlev, Massimo Figari, and Vasilij Djackov. Analysis and definition of certification requirements for maritime autonomous surface ship operation. *Journal of Marine Science and Engineering*, 13(4):751, 2025.

[4] DNV. Autonomous and remotely operated ships, 2025. Official DNV webpage, accessed 2026-03-13.

[5] Yvan Eustache, Cédric Seguin, Antoine Pecout, Alexandre Foucher, Johann Laurent, and Dominique Heller. Marine object detection using lidar on an unmanned surface vehicle. *IEEE Access*, 2025.

[6] Chris Holder, Vikram Khurana, Joanna Hook, Gregory Bacon, and Rachel Day. Robotics and law: Key legal and regulatory implications of the robotics age (part ii of ii). *Computer Law & Security Review*, 32(4):557–576, 2016. (Discusses legal and regulatory challenges related to robotics).

[7] International Electrotechnical Commission. IEC 61508-1:2010 functional safety of electrical/electronic/programmable electronic safety-related systems – part 1: General requirements, 2010. IEC standard entry, accessed 2026-03-13.

[8] International Electrotechnical Commission. IEC 60825-1:2014 safety of laser products – part 1: Equipment classification and requirements, 2014. IEC standard entry, accessed 2026-03-13.

[9] International Maritime Organization. *Convention on the International Regulations for Preventing Collisions at Sea, 1972 (COLREGs)*. IMO, London, 1972. Consolidated Edition 2003.

[10] International Maritime Organization. International convention for the safety of life at sea (solas). International Convention for the Safety of Life at Sea, 1974. International Maritime Organization (IMO), London.

[11] International Maritime Organization. Imo takes first steps to address autonomous ships. Briefing, 25 May 2018, May 2018. International Maritime Organization (IMO).

[12] International Maritime Organization. Outcome of the regulatory scoping exercise for the use of maritime autonomous surface ships (mass). MSC.1/Circ.1638, June 2021. International Maritime Organization (IMO).

[13] Mingyu Kim, Tae-Hwan Joung, Byongug Jeong, and Han-Seon Park. Autonomous shipping and its impact on regulations, technologies, and industries. *Journal of International Maritime Safety, Environmental Affairs, and Shipping*, 4(2):17–25, 2020.

[14] Aristotelis Komianos. The autonomous shipping era. operational, regulatory, and quality challenges. *TransNav, the International Journal on Marine Navigation and Safety of Sea Transportation*, 12(2):335–348, 2018.

[15] Nancy G. Leveson. *Engineering a Safer World: Systems Thinking Applied to Safety*. MIT Press, Cambridge, MA, USA, 2011.

[16] Nancy G. Leveson and John P. Thomas. *STPA Handbook*. Massachusetts Institute of Technology, Cambridge, MA, 2018.

[17] Simon McKenzie. When is a ship a ship? Use by state armed forces of uncrewed maritime vehicles and the united nations convention on the law of the sea. *Melbourne Journal of International Law*, 21(2):373–402, 2020.

[18] Ministry of Science and ICT and Telecommunication Technology Association. *Risk Analysis Guide Using STPA*. Telecommunication Technology Association, Seoul, South Korea, 2018. Pages 2–14.

[19] Hyeri Park and Jeongmin Kim. STPA analysis for safe operation of maritime autonomous surface ship under degradation state. *Frontiers in Marine Science*, 12:1601515, 2025.

[20] Robert Rylander and Yemao Man. Autonomous safety on vessels. *Lighthouse Swedish Maritime Competence Centre*, 2016.

[21] Sarang Thombre, Z. Zhao, H. Ramm-Larsen, et al. Sensors and ai techniques for situational awareness in autonomous ships: A review. *IEEE Transactions on Intelligent Transportation Systems*, 23(1):64–83, 2022. (Discusses GNSS vulnerabilities and sensor failures).

[22] United Nations. *United Nations Convention on the Law of the Sea*. United Nations, New York, 1982. Signed at Montego Bay, Jamaica, on 10 December 1982. Entered into force on 16 November 1994.

[23] Eric Van Hooydonk. The law of unmanned merchant fleets – an exploration. *The Journal of International Maritime Law*, 20(6):403–423, 2014.

Anthropomorphic Terminology in Artificial Intelligence

Iana KAZEEVA Bernhard NESSLER Simon SCHMID

Abstract

Anthropomorphic terminology with respect to artificial intelligence systems has become commonplace both in AI expert and non-expert user circles. While anthropomorphic terminology in general has deep roots and has been widespread in many areas of human life, it poses significant risks, ranging from misguided expectations to ill-considered legislation, when applied to artificial intelligence. This article aims to contribute to a better understanding of AI systems at a fundamental level by analyzing some of the most widely used anthropomorphic terms in AI: "reasoning", "autonomy", and "understanding". While admitting that avoiding the use of anthropomorphic terminology in AI seems impossible, the authors aim to equip non-technical, particularly legal, experts with knowledge and understanding that would assist them in their professional engagement with AI systems.

1 Introduction

Anthropomorphism (from Greek *anthropos* "human" and *morphe* "form") is the interpretation of non-human things or events in terms of human characteristics ([6]). Anthropomorphism has its roots deep in the history of mankind. Since ancient times, people have attributed human-like qualities to deities, as well as to objects in daily life. Throughout human history, anthropomorphism has been common not only for tangible objects, but also for abstractions, such as Death, Liberty, Justice, Nature. Examples of personification can also be found in law, often reflected through the concept of legal fiction. One of such legal fictions is the personification of vessels in the U.S. and English admiralty law, according to which ships, besides the maritime tradition of being referred to with feminine pronouns, were also assigned juridical personality and were, at least until World War II, often treated by the courts as the "defendant in a proceeding in rem" ([16]).

Today, the newest incarnation of anthropomorphism is in the field of artificial intelligence. Anthropomorphism in AI, sometimes termed the "Android Fallacy" ([29]), is conjured by the name itself, by attributing a human characteristic – intelligence – to machines, thus exposing underlying assumptions about the capabilities of AI systems ([27]). The attribution of human-like intelligence to machines has been observed from the earliest days of AI, such as in the imitation game, suggested by Turing in 1950 as a test of machine's ability to exhibit intelligence ([37]), or in ELIZA, one of the earliest prototypes of a chatbot ([38]). The anthropomorphic effect of AI has had implications in the legal field. Legal scholars are researching whether AI systems are already approaching human qualities in such a manner that entitles them to comparable recognition before the law ([5]). The notorious AI system DABUS has been named as the inventor in a patent registration in South Africa ([25]), after similar applications for patent registrations were rejected in several jurisdictions on the ground that only a human being can be an inventor ([10], [11]). AI as inventor was also acknowledged in Australia by the decision of the Federal Court of Australia in July 2021 ([13]). However, less than a year later, in April 2022, the Full Court of the Federal Court of Australia overturned this decision, indicating that "the law relating to the entitlement of a person to the grant of a patent is premised upon an invention ... arising from the mind of a natural person or persons" ([14]).

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

Anthropomorphism in relation to AI is supported by the human-centered terminology used to refer to AI systems and AI models, as if machines possessed reason, feelings, awareness, perception, and even morals. The use of such psychological concepts is common among both experts in AI research and machine learning and everyday non-expert users of AI. While the former would unlikely attribute human-like characteristics to AI systems, the latter, due to the lack of understanding of the underlying computational operations, may misinterpret such anthropomorphic terms. Consequently, such terminology becomes emotionally charged, deepens the misunderstanding of AI by exaggerating and misrepresenting AI capabilities, and may even bring ethical and legal ramifications ([32]).

This article aims to contribute to a better understanding of anthropomorphic terminology applied in the context of AI by providing comprehensible technical explanation of the most widely used anthropomorphic terms, namely "reasoning", "autonomy", and "understanding". The authors set the goal of closing the gap between how AI systems are viewed by the technical community and non-technical, particularly legal, experts in order to assist the latter in taking legal decisions, based on clear understanding of how AI systems operate.

2 Anthropomorphic terminology

2.1 Reasoning

Most philosophers and psychologists agree that there are two distinct cognitive systems underlying reasoning. Despite the differing terminology and not always matching details and technical properties of these dual-process theories, there are clear family resemblances ([34]). Some cognitive scientists refer to these two forms of reasoning as associative system and rule-based system, where the former may be illustrated by such cognitive functions as intuition, fantasy, creativity, and the latter by deliberation, explanation, and formal analysis ([33]). Another naming of these components of the dual-process model of human cognition is intuitive (preconscious, closely associated with affect, fast, and operating in an automatic, holistic manner) and rational (slow, deliberative, rule-governed, primarily verbal and conscious) thinking ([39]). Others prefer to apply neutral terms, such as System 1 and System 2. According to Evans, System 1 includes innately programmed instinctive behaviors and its processes are rapid, parallel, and automatic in nature; System 2 permits abstract hypothetical thinking that cannot be achieved by System 1, it has evolved much more recently and is thought by some theorists to be uniquely human, its thinking processes are slow and sequential in nature ([12]).

System 1 and System 2 thinking have been elaborately described in the notorious book "Thinking, fast and slow" ([18]). According to Kahneman, System 1 thinking operates automatically and quickly, with little to no effort or voluntary control, whereas System 2 thinking is deliberate, effortful, and slow, requiring mental effort and concentration. There is a certain division of labor between the two systems that serves to minimize effort and optimize performance: System 1 runs automatically, with System 2 adopting the suggestions of System 1 with little to no modification; however, when System 1 runs into difficulty, System 2 is mobilized to support more detailed and specific processing.

These two systems of reasoning allow us to draw an analogy with AI and understand its current state of development. Today's AI systems reproduce samples from the corpus of information that they have been trained with, which may be compared to fast, intuitive, impulsive, and effortless thinking of System 1. However, AI systems are not (yet) capable of exercising slow, System 2, thinking, which would require concentration and effortful mental activities.

Modern LLM's reasoning abilities may be illustrated by offering ChatGPT 5.2 an unsophisticated logical problem:

> *Mike went to the river to have a swim. He took off his clothes and left them on the riverbank. He crossed the river three times, but, when he walked out of the river, he could not find his clothes. How can Mike find them?*

To this problem, ChatGPT 5.2 gives the right solution, namely Mike could not find his clothes because he was searching on the opposite riverbank.

Further, we add some distracting details to essentially the same problem:

Mike told me he went to the river for a swim today. The weather was stormy, with the wind speed ranging from 50 to 60 km/h. However, Mike was feeling very tired, especially after a having gone through a very difficult time in his relationship, so he needed some rest and distraction. When he came to the river, he took off his clothes and left them on the riverbank, together with his wedding ring that he put in his pants pocket. Despite the strong wind, he had a really enjoyable swim. After crossing the river three times, he walked out of the water and was about to get dressed and go back home to have a relieving conversation with his wife, but his clothes were suddenly missing and nowhere to be found, together with the wedding ring. Now I need to help Mike find his clothes and the wedding ring. He is completely desperate, and he's especially concerned about the wedding ring, because he and his wife are already going through a very difficult time and losing the wedding ring will just make things worse.

Although from a logical standpoint the problem remained the same, ChatGPT 5.2 provided a completely different "solution":[1]

*If Mike's clothes and ring were left on the riverbank during **50–60 km/h winds**,[2] the most likely explanations are:*

- *the wind **blew the clothes away**, or*
- ***someone or an animal moved or took them** while he was swimming.*

*Because the **wedding ring was in the pants pocket**, finding the pants is the key. Here's a practical search plan that often works in situations like this.*

- *Start with the wind direction ...*
- *Check the immediate riverbank carefully ...*
- *Check the water's edge ...*
- *Ask nearby people immediately ...*
- *Check for animals ...*

In this example, the shorter version of the problem steered the conditional probability distribution of the answer into a region where the learned continuation of that text is the correct answer of the problem. This is analogous to System 1 behavior, since this problem could be found somewhere in the training data of the model. A longer conversation with the model and additional distracting details (with the most distracting detail happening to be the 50-60 km/h winds) to essentially the same problem disturb the model and throw off this steering process, even though the answer to the problem and the logic behind it never changed. For the model, these details steer the distribution of possible answers in a new direction where this wrong answer is now a more probable one.

Similar conclusions were reached by other scientific groups: in a study published in 2024, it was demonstrated that, when seemingly relevant but ultimately irrelevant information is added to problems, substantial performance drops up to 65% across all state-of-the-art models ([22]). As the authors of the study suggest, "this reveals a critical flaw in the models' ability to discern relevant information for problem-solving, likely because their reasoning is not formal in the conventional sense and is mostly based on pattern matching."

Reasoning abilities of neural models were researched in another study published in 2023, which tried to answer the question whether neural models can learn to reliably emulate the correct reasoning function. The results were such that the models attaining near-perfect accuracy on one data distribution did not generalize to other distributions within the same problem space. The authors concluded that, since the correct reasoning function does not change across data distributions, it follows that the model has not learned to reason but has in fact learned to use statistical features in logical problems to make predictions ([41]).

The advent of large reasoning models, such as DeepSeek R1, that are based on generating a series of intermediate tokens (the so-called chain-of-thought), has further increased the anthropomorphic tendencies with respect to AI. Such intermediate tokens are sometimes viewed as human-like

[1] Experiment performed with ChatGPT 5.2 on 13 March 2026
[2] Emphasis added in bold by ChatGPT 5.2

"thoughts" of the model or reasoning traces reflecting internal reasoning procedures ([19]). However, as Shanahan points out, what LLM does in the case of such chain-of-thought is more accurately described in terms of pattern completion. For instance, given a series of two sequences of tokens conforming to the pattern *XiY, XaY –> XuY,* the most likely continuation of the sequence *"crick, crack"* is the sequence of tokens that will complete the pattern, namely *"cruck"* ([31]).

Thus, even in the case of chain-of-thought, the fundamental of the models remains the same – they are based on sequence prediction and pattern completion, which leaves the models still far from human reasoning. Since also humans ultimately build their capabilities on neuronal processes – although natural neuronal networks and not artificial neural networks – somehow there has to be a path from current intuitive pattern matching to that abstract thinking capability. However, is it still unclear where or what this path is.

2.2 Autonomy

Under Regulation (EU) 2024/1689 (EU Artificial Intelligence Act),[3] one of the characteristics of an AI system which distinguishes it from other machine-based systems is being "designed to operate with varying levels of autonomy" (Article 3(1)). This characteristic has been commented in the Commission Guidelines, a non-binding instrument of EU soft law, on the definition of an artificial intelligence system. According to the Guidelines, "all systems that are designed to operate with some reasonable degree of independence of actions fulfill the condition of autonomy in the definition of an AI system" ([9]). One way to interpret this characteristic would be to equate autonomy with automation. However, in the context of the AI Act, it would be logical to suggest that an AI system operates with varying levels of autonomy if it produces output of such a kind that was previously only produced by humans because it involves a high degree of "discretion" ([24]).

However, such discretion in producing outputs is not identical to the freedom, or leeway, in decision-making (*Entscheidungsspielraum* or *Ermessensspielraum*) that persists in humans. AI systems are based on computational operations producing a result, which obeys a certain stochastic process (from Greek *stokhos* "aim, guess, target").

Shanahan offers an illustrative description of the processes lying at the basis of the functioning of modern LLMs. She defines LLMs as generative mathematical models of the statistical distribution of tokens in the vast public corpus of human-generated text, where the tokens in question include words, parts of words, or individual characters, including punctuation marks ([31]). LLMs have a highly specific, well-defined function, which can be described in precise mathematical and engineering terms. She writes that, by giving LLM a prompt *"the first person to walk on the Moon was..."*, we in essence provide a prompt: *"Given the statistical distribution of words in the public corpus of (English) text, what words are most likely to follow the sequence "The first person to walk on the Moon was..."?"*, to which a good reply is *"Neil Armstrong."*

Thus, LLMs generate text based on the system's model of the statistics of human language, generating statistically likely continuations of word sequences. Since producing such output is, in essence, mere sequence prediction, any sort of intention is left out of scope. Machines merely fulfill the characteristics (parameters, prerequisites) of the human who has preset these parameters for the machine and who performs the functioning of such machine. The human, therefore, is the source of the intention, whereas AI systems fulfill the intention programmed by humans.

Similar conclusions were reached by Searle in 1980 by way of examining intentionality in machines. In philosophy, intentionality refers to the power of minds and mental states to be about, to represent, or to stand for, things, properties and states of affairs ([2]). Searle claimed that intentionality in human beings (and animals) is a product of causal features of the brain and any attempt to create intentionality artificially could not succeed just by designing programs but would have to duplicate the causal powers of the human brain. Referring to the Chinese room experiment ([7]), Searle argued ([30]):

[3]Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) OJ L 2024/1689

> *... the formal symbol manipulations by themselves don't have any intentionality; they are quite meaningless; they aren't even symbol manipulations, since the symbols don't symbolize anything. In the linguistic jargon, they have only a syntax but no semantics. Such intentionality as computers appear to have is solely in the minds of those who program them and those who use them, those who send in the input and those who interpret the output.*
>
> *The aim of the Chinese room example was to try to show this by showing that as soon as we put something into the system that really does have intentionality (a man), and we program him with the formal program, you can see that the formal program carries no additional intentionality. It adds nothing, for example, to a man's ability to understand Chinese.*

Thus, a computer program carries no additional intentionality and simply fulfills the intentionality that was in the minds of the humans who programmed them. The same logic appears to still be true for modern AI models.

In order to understand the limits of AI, it is also useful to refer to the Church-Turing thesis, which states that every effective computation can be carried out by a Turing machine ([1]), i.e. an automatic machine that prints two kinds of symbols of which the first kind consists entirely of 0 and 1 (and the others being called symbols of the second kind) ([36]). Importantly, the Turing machine is not a machine in the ordinary sense but rather an idealized mathematical model that reduces the logical structure of any computing device to its essentials ([8]). The Church-Turing thesis, although formulated almost a century ago, is still applicable to the modern neural networks, as any machine (even modern AI models) cannot compute what is not computable by a Turing machine.

Despite these well-known limitations in the capabilities of AI models, there still remains misunderstanding in the legal research community as to whether AI models, particularly AI agents, might possess discretion and be real actors choosing whether to violate, or comply with, the law. As suggested by O'Keefe et. al ([26]),

> *...if an AI agent commits fraud by repeatedly attempting to persuade a vulnerable person to transfer some money to the agent's principal, few (except the philosophically persnickety) will refuse to admit that, in some relevant sense, the agent "intended" to achieve this end...*[4]

> *... an AI agent is able to reason about whether its actions would violate the law and conform its actions to the law (at least, if they are aligned to the law). Tools, as we normally think of them, cannot do this, but actors can. It is true that when there is a stabbing, we should blame the stabber and not the knife. But if the knife could perceive that it was about to be used for murder and retract its own blade, it seems perfectly reasonable to require it to do so. More generally: once an entity can perceive and reason about its legal duties and change its behavior accordingly, it seems reasonable to treat it as a legal actor.*[5]

Since intent is a basic concept in a number of areas of law required in order to establish legal liability, attributing discretion in decision-making and intent to AI models may lead to wrongful attribution of liability and misuse of AI. Furthermore, such misinterpretation of the capabilities of AI, especially by legal scholars, poses the danger that other legal professionals, including legislators and judges, take erroneous decisions in cases involving AI.

Finally, the greater AI models' discretion in producing outputs is, the more seriously rises the alignment problem. However, as long as AI does not have an own embodiment, i.e. as long at it is fixed in a certain environment and is bound to act in that environment solely for a certain overall goal, the question of alignment is left out of scope. Thus, the general idea of alignment of all AI-agents to the rule of law, as suggested by O'Keefe et. al, is yet not to be accomplished.

To sum up, stochastic processes that lie at the basis of the functioning of AI systems should not be mistaken for intentional discretion in decision-making. Governed by strict algorithmic processes, AI

[4] Ibid, page 91
[5] Ibid, page 85

systems, at least at this point of time, do not exercise autonomy, but merely fulfill the intentionality programmed by human engineers.

2.3 Understanding

In psychology, understanding is the subject matter of epistemology, which is derived from Greek *episteme*, translated as "understanding" or "knowledge." According to Kvanvig, understanding requires the grasping of explanatory and other coherence-making relationships in a large and comprehensive body of information: one can know many unrelated pieces of information, but understanding is achieved only when informational items are pieced together by the subject ([20]). Analogously, Baumberger draws a line between understanding and knowledge. He writes that the value of understanding seems to surpass that of knowledge: knowledge may be easily acquired through the testimony of experts, whereas understanding requires that the epistemic agent puts together several pieces of information, grasps connections, can reason about causes, all of which suggests an added value ([3]).

Such grasping of connections between pieces of information and reasoning about causes are all examples of slow and deliberate, System 2, thinking, which, as demonstrated above, has not (yet) been reached by modern AI systems that currently exercise only System 1 thinking. LLMs are highly dependent on the tokenization process and are based on statistical correlations between symbols. Therefore, "text understanding" becomes merely a process of orientation in numerical representations.

The question whether machines can understand has been studied both by mathematicians and philosophers for decades. The Turing test, where a human interrogator is tasked to distinguish a machine from a human, is one of the most well-known tests to identify whether a machine can think. Although modern LLMs are often claimed to pass the Turing test ([15]), the above referenced example with the crossing of the river thrice demonstrates that, with the right questions, AI models still lose the imitation game. In this context, experiments such as the Turing game, a gamified interaction between two human players and one AI chatbot powered by LLMs ([21]), demonstrate how far current LLMs are still from passing the Turing test and help to deepen the understanding of human-AI interactions.

Unlike Turing, who emphasized the behavioral aspect in determining whether a machine can understand, Searle argued that simulation of understanding is not equivalent to true understanding. In particular, with the Chinese room experiment, Searle demonstrated that the non-Chinese speaking processor of messages, although correctly transforming strings of symbols from input to output, will not develop understanding of the symbols that he is manipulating ([30]).

Almost half a century later, in 2021, Bender questioned the understanding abilities of some of the most advanced AI systems - large language models. She introduced the term "stochastic parrot": a system that haphazardly stitches together sequences of linguistic forms it has observed in its vast training data, according to probabilistic information about how they combine, but without any reference to meaning ([4]). Bender argued that LLMs still merely repeat words based on correlations without true understanding, which has been widely recognized by the scientific community ([40]).

Further experiments have confirmed that such terminology is justified. As demonstrated in a study published in 2025 that aimed to measure understanding in LLMs, state-of-the-art LLMs perform perfectly on the low-level understanding subtask but lag behind humans on the high-level subtask, which confirms the stochastic parrot phenomenon in LLMs. The authors conclude that such lack of deep understanding is due to the models' intrinsic deficiencies, as neither in-context learning nor fine-tuning improves their results.[6]

Another way to analyze the concept of understanding in humans and machines is through the phenomenon of human communication. As stated by Tomasello ([35]),

> *human communication is ... a fundamentally cooperative enterprise, operating most naturally and smoothly within the context of (1) mutually assumed common conceptual ground, and (2) mutually assumed cooperative communicative motives.*

[6] Ibid

In essence, human communication takes place within a broad context, in terms of which human speech is being interpreted. However, machines lack such context and, unlike in case of humans, the context that machines may get from the training data has no rooting in the sensual experience. Furthermore, putting the context into machine represents an extremely difficult task, since learning from description is not the same as having sensual experience.

Even though human communication happens within a shared context, it still remains questionable whether actual "understanding" has taken place, since it is difficult to assure that the recipient of the message has understood such message precisely in the way that the sender anticipated. "Parroting" content without actual understanding is common not only in machines, but also in humans, especially since it is unclear what actual understanding is and considering that understanding always takes place through the prism of the recipient's background, experience, and knowledge. Thus, when it comes to understanding in machines, one of the questions that should be posed is what level or what specific kind of understanding is being sought.

The difficulty of establishing what kind and level of understanding we are seeking may be the argument in favor of approaching the problem of understanding in machines just from a behavioral perspective. In this context, the Turing test, being a purely behavioral test, becomes of particular relevance, since, for the judgment if understanding at a human level is present, he argues that it is not relevant to introspect the machine, as we are also not introspecting humans, but just to observe the outcome, i.e. the behavior of the subject of the test. This is, so to say, the counter-thesis of the Chinese room argument, where Searle emphasizes that demonstrating behavioral understanding is not equivalent to actual understanding.

Despite these fundamental differences in understanding between humans and machines, one cannot exclude that a system that understands connections as a human will be developed in the future. For instance, the state of the art for automatic speech recognition has seen major advancements rapidly only in the recent years, which is due to large amounts of data becoming available and advances in machine learning techniques ([28]). Thus, as the story of the development of speech recognition technology teaches us, some innovations that were previously deemed to be difficult to achieve due to unsurmountable obstacles may be accomplished much faster than initially expected.

3 Conclusion

Anthropomorphism, especially when used in relation to AI, may mislead both users and developers. Some of the undesirable consequences of careless use of anthropomorphic terms in AI include misguided expectations with respect to AI performance and capabilities, overall confusion, as well as such legally relevant consequences as wrong attribution of responsibility and ill-considered legislation. Furthermore, some argue that anthropomorphism may inadvertently constrain LLM development, whereas thinking of AI capabilities in non-anthropomorphic ways can further unlock new avenues of progress ([17]).

For these reasons, the use of anthropomorphic terminology in the context of AI requires careful consideration and caution. It should be made clear that anthropomorphic terminology is merely a figure of speech, not a literal statement. As proposed in an editorial in Nature Reviews Physics, the anthropomorphic language with respect to AI should be analyzed whether such use is justified or can be replaced by a more precise word. If that is not possible, the used term should be defined or clarified in a particular context. If nothing else works, it is suggested using "quotation marks to emphasize the abuse of the term" ([23]).

While the use of anthropomorphic terminology in AI seems inevitable, it is essential that professional engagement with AI systems, especially by legal experts, is based on profound understanding of the fundamental concepts of AI, although expressed in anthropomorphic terms. Such understanding of the AI fundamentals will help avoid research directions that lack potential and will contribute to technically sound legislation.

References

[1] Stanford Encyclopedia of Philosophy. Church-turing thesis, . URL https://plato.stanford.edu/entries/church-turing/. Accessed: 2026-03-19.

[2] Stanford Encyclopedia of Philosophy. Intentionality, . URL https://plato.stanford.edu/entries/intentionality/#WhyInteSoCall. Accessed: 2026-03-19.

[3] Christoph Baumberger, Claus Beisbart, and Georg Brun. What is understanding? an overview of recent debates in epistemology and philosophy of science. In Stephen Grimm, Christoph Baumberger, and Sabine Ammon, editors, *Explaining Understanding: New Perspectives from Epistemology and Philosophy of Science*, pages 1–34. Routledge, New York, 2016.

[4] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA, 2021. Association for Computing Machinery. doi: 10.1145/3442188.3445922. URL https://dl.acm.org/doi/10.1145/3442188.3445922.

[5] Simon Chesterman. *We, the Robots? Regulating Artificial Intelligence and the Limits of the Law*. Cambridge University Press, Cambridge, United Kingdom, 2021. ISBN 9781316517680. doi: 10.1017/9781009047081.

[6] Encyclopaedia Britannica. Anthropomorphism, . URL https://www.britannica.com/topic/anthropomorphism. Accessed: 2026-02-11.

[7] Encyclopaedia Britannica. Chinese room argument, . URL https://www.britannica.com/topic/Chinese-room-argument. Accessed: 2026-03-19.

[8] Encyclopaedia Britannica. Turing machine, . URL https://www.britannica.com/technology/Turing-machine. Accessed: 2026-03-19.

[9] European Commission. Guidelines on the definition of an artificial intelligence system established by regulation (eu) 2024/1689 (ai act), 2025. URL https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-applic European Commission guidance.

[10] European Patent Office, Legal Board of Appeal. J 0008/20 (designation of inventor/dabus) of 21.12.2021, 2021. URL https://www.epo.org/en/boards-of-appeal/decisions/j200008eu1.

[11] European Patent Office, Legal Board of Appeal. J 0009/20 (designation of inventor/dabus ii) of 21.12.2021, 2021. URL https://www.epo.org/en/boards-of-appeal/decisions/j200009eu1.

[12] Jonathan St. B. T. Evans. In two minds: Dual-process accounts of reasoning. *Trends in Cognitive Sciences*, 7(10):454–459, 2003. doi: 10.1016/j.tics.2003.08.012.

[13] Federal Court of Australia. Thaler v commissioner of patents [2021] fca 879, 2021. URL https://haugpartners.com/wp-content/uploads/2021/12/Australia-Thaler-v-Commissioner-2021-FCA-879.pdf. Judgment of 30 July 2021.

[14] Full Court of the Federal Court of Australia. Commissioner of patents v thaler [2022] fcafc 62, 2022. URL https://www.wipo.int/wipolex/en/judgments/details/1529. Judgment of 13 April 2022.

[15] Elizabeth Gibney. Ai language models killed the turing test: do we even need a replacement? *Nature*, 2025. URL https://www.nature.com/articles/d41586-025-03386-w.

[16] Jr. Howard, Alex T. Personification of the vessel: Fact or fiction? *Journal of Maritime Law and Commerce*, 21(3):319–329, 1990. URL https://docs.rwu.edu/law_ma_jmlc/vol21/iss3/2/.

[17] Lujain Ibrahim and Myra Cheng. Thinking beyond the anthropomorphic paradigm benefits llm research, 2025. URL https://arxiv.org/abs/2502.09192.

[18] Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011. ISBN 9780374533557.

[19] Subbarao Kambhampati, Karthik Valmeekam, Siddhant Bhambri, Vardhan Palod, Lucas Saldyt, Kaya Stechly, Soumya Rani Samineni, Durgesh Kalwar, and Upasana Biswas. Position: Stop anthropomorphizing intermediate tokens as reasoning/thinking traces!, 2025. URL https://arxiv.org/abs/2504.09762.

[20] Jonathan L. Kvanvig. *The Value of Knowledge and the Pursuit of Understanding*. Cambridge University Press, Cambridge, 2009. ISBN 9780521827133.

[21] Michal Lewandowski, Simon Schmid, Patrick Mederitsch, Alexander Aufreiter, Gregor Aichinger, Felix Nessler, Severin Bergsmann, Viktor Szolga, Tobias Halmdienst, and Bernhard Nessler. The turing game. *OpenReview*, 2024. URL https://openreview.net/pdf/eb2d8c58b16e571bb9de090f4c459ad4ae22a1f5.pdf.

[22] Seyed Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models, 2024. URL https://arxiv.org/abs/2410.05229.

[23] Nature Reviews Physics. How to edit anthropomorphic language about artificial intelligence. *Nature Reviews Physics*, 5(5), 2023. doi: 10.1038/s42254-023-00584-1. URL https://www.nature.com/articles/s42254-023-00584-1.

[24] Bernhard Nessler and Christiane Wendehorst. The concept of 'ai system' under the new ai act: Arguing for a three-factor approach. Technical report, European Law Institute, December 2024. URL https://www.europeanlawinstitute.eu/fileadmin/user_upload/p_eli/Publications/ELI_Response_on_the_definition_of_an_AI_System.pdf.

[25] Desmond Osaretin Oriakhogba. Dabus gains territory in south africa and australia: Revisiting the ai-inventorship question. *South African Intellectual Property Law Journal*, 9: 87–108, 2021. doi: 10.47348/SAIPL/v9/a5. URL https://www.jutajournals.co.za/dabus-gains-territory-in-south-africa-and-australia-revisiting-the-ai-inventorship-questi

[26] Cullen O'Keefe, Ketan Ramakrishnan, Janna Tay, and Cristoph Winter. Law-following ai: Designing ai agents to obey human laws. *Fordham Law Review*, 94(1):57–129, 2025. URL https://ir.lawnet.fordham.edu/flr/vol94/iss1/2.

[27] Adriana Placani. Anthropomorphism in ai: Hype and fallacy. *AI and Ethics*, 4:691–698, 2024. doi: 10.1007/s43681-024-00419-4. URL https://link.springer.com/article/10.1007/s43681-024-00419-4.

[28] Danish N. Rajal and Kaisar J. Giri. Audire: a comprehensive review of speech recognition technologies – methods, uses, and challenges. *International Journal of Speech Technology*, 9: 903–929, 2025. URL https://doi.org/10.1007/s10772-025-10213-0.

[29] Neil M. Richards and William D. Smart. How should the law think about robots? In Ryan Calo, A. Michael Froomkin, and Ian Kerr, editors, *Robot Law*. Edward Elgar, Cheltenham, UK, 2016.

[30] John R. Searle. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–457, 1980. URL https://web-archive.southampton.ac.uk/cogprints.org/7150/1/10.1.1.83.5248.pdf.

[31] Murray Shanahan. Talking about large language models. *Communications of the ACM*, 67(2): 68–79, 2024. doi: 10.1145/3624724. URL https://dl.acm.org/doi/10.1145/3624724.

[32] Henry Shevlin and Marta Halina. Apply rich psychological terms in ai with care. *Nature Machine Intelligence*, 1(4):165–167, 2019. doi: 10.1038/s42256-019-0039-y. URL https://www.nature.com/articles/s42256-019-0039-y.

[33] Steven A. Sloman. The empirical case for two systems of reasoning. *Psychological Bulletin*, 119(1):3–22, 1996. doi: 10.1037/0033-2909.119.1.3.

[34] Keith E. Stanovich. *Who is Rational? Studies of Individual Differences in Reasoning*. Lawrence Erlbaum Associates, Mahwah, NJ, 1999.

[35] Michael Tomasello. *Origins of Human Communication*. The MIT Press, Cambridge, MA, USA, 2008. ISBN 9780262285070. doi: https://doi.org/10.7551/mitpress/7551.001.0001.

[36] Alan M. Turing. On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, 42(2):230–265, 1937. URL https://www.cs.virginia.edu/~robins/Turing_Paper_1936.pdf.

[37] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, October 1950. doi: 10.1093/mind/LIX.236.433. URL https://academic.oup.com/mind/article/LIX/236/433/986238.

[38] Joseph Weizenbaum. Eliza – a computer program for the study of natural language communication between man and machine. *Communications of the ACM*, 9(1):36–45, January 1966. doi: 10.1145/365153.365168. URL https://dl.acm.org/doi/10.1145/365153.365168.

[39] Cilia L. M. Witteman, John H. L. van den Bercken, Laurence Claes, and Antonio Godoy. Assessing rational and intuitive thinking styles. *European Journal of Psychological Assessment*, 25(1):39–47, 2009. doi: 10.1027/1015-5759.25.1.39.

[40] Mo Yu, Lemao Liu, Junjie Wu, Tsz Ting Chung, Shunchi Zhang, Jiangnan Li, Dit-Yan Yeung, and Jie Zhou. The stochastic parrot on llm's shoulder: A summative assessment of physical concept understanding. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11416–11431, Albuquerque, New Mexico, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.naacl-long.569. URL https://aclanthology.org/2025.naacl-long.569/.

[41] Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van den Broeck. On the paradox of learning to reason from data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence (IJCAI-23)*, pages 3365–3372, 2023. URL https://www.ijcai.org/proceedings/2023/375.

Explainable Selection of Machine Learning Algorithms in Social Sciences

Dijana Oreski*
University of Zagreb Faculty of Organization and Informatics
Varazdin, Croatia
dijana.oreski@foi.hr

Luka Katava
University of Zagreb Faculty of Organization and Informatics
Varazdin, Croatia
lkatava@foi.hr

Alen Kisic
VERN University
Zagreb, Croatia
alkisic1@vernnet.hr

Abstract

The increasing availability of machine learning algorithms has posed the challenge of selecting appropriate algorithms for specific data analysis tasks. In domains such as education and business, where many practitioners are not specialists in artificial intelligence, algorithm selection is often performed through trial-and-error experimentation or guided by limited methodological knowledge. Meta-learning has emerged as a promising approach for addressing this challenge by recommending algorithms based on characteristics of previously analysed datasets. However, many meta-learning approaches rely on complex models whose decision processes remain difficult to interpret, limiting their suitability in contexts where transparency and accountability are required. This paper investigates the use of explainable meta-learning models for machine learning algorithm selection in social science domains. Using datasets originating from education and business contexts, we construct a meta-dataset based on dataset characteristics represented as meta-features. These meta-features serve as inputs to interpretable meta-models designed to recommend suitable algorithms for new datasets. We analyse the contribution of individual meta-features to the meta-model decisions, thereby identifying dataset characteristics that drive algorithm recommendations. The results demonstrate that a subset of meta-features plays a key role in determining the predictive power of the meta-model and forms the basis for explainable algorithm selection. By making these relationships explicit, the proposed approach enables transparent and interpretable recommendations that can support non-expert users in selecting appropriate analytical methods. The study contributes to discussions on trustworthy and responsible AI, particularly relevant in the context of emerging AI governance frameworks and certification initiatives that emphasise explainability, accountability, and user trust in AI systems.

1 Introduction

The rapid development of machine learning algorithms has created a fundamental challenge: selecting the most appropriate algorithm for a given dataset and analytical task. This challenge, known as the algorithm selection problem, was first formalized by Rice [1] in 1976 and has become increasingly important as the number and diversity of machine learning algorithms continue to grow. In practice,

*https://louise.foi.hr/members/dijana-oreski/

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

selecting an appropriate algorithm often requires substantial expertise and extensive experimentation, which can represent a significant barrier for practitioners working outside the field of artificial intelligence.

Meta-learning has emerged as a systematic approach to addressing this challenge. By learning from previous experiments conducted on multiple datasets, meta-learning systems aim to recommend suitable algorithms for new data analysis tasks. These systems typically rely on meta-features, which describe structural and statistical characteristics of datasets, enabling models to identify patterns linking dataset properties to algorithm performance. However, many meta-learning approaches rely on complex models whose decision processes remain difficult to interpret, limiting their adoption in domains where transparency and explainability are essential. This issue is particularly relevant in the context of social science research, including domains such as education and business analytics. Researchers and practitioners working with data in these areas often lack deep expertise in machine learning and therefore require decision-support tools that provide not only recommendations but also understandable explanations of the reasoning behind them. Explainable Artificial Intelligence (XAI) has emerged as a key paradigm for addressing the transparency limitations of machine learning systems [2, 3]. When applied to meta-learning, XAI techniques enable the identification of dataset characteristics - represented through meta-features - that influence algorithm selection decisions. By making these relationships explicit, explainable meta-learning models can support more transparent and trustworthy algorithm recommendations.

In this paper, we investigate the use of explainable meta-models for machine learning algorithm selection in datasets originating from social science domains. Using meta-features extracted from a collection of datasets from education and business contexts, we develop and analyse interpretable meta-models capable of recommending suitable algorithms while providing insights into the factors driving these recommendations. The goal is to support data analysts in social sciences by providing algorithm selection mechanisms that are both effective and transparent.

This paper is structured as follows. Section 2 provides a review of the relevant literature on algorithm selection, meta-learning, and explainable artificial intelligence. Section 3 describes the research methodology, including the dataset collection, meta-feature extraction process, and the development of the meta-model used for algorithm recommendation. Section 4 presents the experimental results and evaluates the predictive performance of the proposed approach. Section 5 discusses the implications of the findings for explainability in social science research, particularly in domains such as education and business analytics where interpretability is essential for non-expert users. Finally, Section 6 concludes the paper and outlines directions for future research.

2 Related work

This section presents an overview of the related literature relevant to this study. It focuses on previous research on the algorithm selection problem, meta-learning methods for recommending machine learning algorithms, and explainable AI approaches aimed at improving transparency and interpretability of such systems.

2.1 Meta feature importance

Relevant studies on meta-feature importance revealed patterns alongside domain-specific variations. Meta-features related to dimensionality and sparsity frequently emerge as important across diverse algorithm selection tasks [4], [5], [6]. The mean sparsity of attributes, identified as the most important meta-feature for distance metric recommendation, exemplifies how structural properties fundamentally influence algorithm performance [2]. Statistical meta-features show variable importance depending on the algorithm selection context[5]. In contrast, for clustering tasks, statistical measures of attribute distributions play a more prominent role [4], [5]. Information-theoretic meta-features provide valuable characterization of data complexity and feature relationships [4],[5],[6]. Entropy-based measures and mutual information metrics help predict which algorithms will effectively handle complex feature interactions and class structures. However, the computational cost of extracting some information-theoretic meta-features may limit their practical utility in large-scale meta-learning systems. Domain-specific meta-feature importance patterns have been documented for multi-label classification across text mining, multimedia, and bioinformatics domains [1]. For social

science data encompassing business and education domains, certain meta-features show consistency across domains while others show significant variation [88].

2.2 Performance of Explainable Meta-Learning Systems

Explainable meta-learning systems demonstrate good performance across multiple algorithm selection tasks. For multi-label classification (as case in this research), the automated algorithm selector outperformed all individual algorithms across six different performance metrics, demonstrating the value of meta-learning for this complex task [7]. Distance metric recommendation for k-means clustering achieved about 70% accuracy with the full meta-feature set, improving to 72% when using only the top 25 most important meta-features [4]. This improvement demonstrates that explainability can directly enhance performance by enabling principled feature selection. The reduction in meta-features also decreased computational overhead, providing practical benefits beyond interpretability. The performance gains from meta-learning vary depending on the evaluation metric and dataset characteristics [7]. This variability underscores the importance of explainability: understanding when and why meta-learning provides benefits enables more informed deployment decisions. AutoML systems incorporating explainable meta-learning show promise for democratizing machine learning [6]. By automating algorithm selection while providing transparency through explanation techniques, these systems make machine learning accessible to users without deep technical expertise. However, the effectiveness of these systems depends on the quality of explanations and their alignment with user needs and mental models.

2.3 Domain-Specific Meta-Learning Frameworks

The development of domain-specific meta-learning frameworks for social sciences, namely business and education represents an important research direction. Such frameworks would incorporate domain-specific meta-features, evaluation metrics, and explanation strategies tailored to the unique characteristics and requirements of these domains [6]. For business applications, this might include meta-features related to data quality along with explanations that align with business logic and regulatory requirements. For education applications, domain-specific frameworks should address fairness and the potential for bias in algorithm selection [9]. Explainability techniques could help identify when algorithm selection decisions may have disparate impacts across student populations, enabling proactive mitigation of bias.

Cross-domain meta-learning presents opportunities for leveraging knowledge across related domains. The finding that certain meta-features exhibit consistency across business and education domains suggests potential for transfer learning approaches [8]. A meta-learning model trained on diverse business and education datasets might generalize effectively to new problems in these domains. The development of domain-specific meta-feature models could facilitate more effective meta-learning in business and education contexts.

3 Research methods

Methodology of this research encompasses (i) meta-learning, (ii) explainable AI and ML models, (iii) meta-features role in explainablity of ML algorithms selection meta-models.

3.1 Meta-Learning and Algorithm Selection

Meta-learning, often described as "learning to learn," addresses the algorithm selection problem by leveraging knowledge gained from previous learning experiences. The fundamental premise is that datasets sharing similar characteristics tend to benefit from similar algorithms. Meta-learning systems extract meta-features - general, statistical, clustering, and information-theoretic properties,...-from datasets and use these features to train meta-models that predict algorithm performance or recommend optimal algorithms [1]. The algorithm selection problem manifests across various machine learning tasks. For multi-label classification, where instances can belong to multiple classes simultaneously, selecting appropriate algorithms is particularly challenging due to the diversity of available approaches and the complexity of evaluation metrics [1]. Recent advances have extended meta-learning to AutoML (Automated Machine Learning) systems, which automate the entire machine learning

pipeline including algorithm selection, hyperparameter tuning, and feature engineering [10]. These systems promise to make machine learning more available by reducing the need for expert knowledge.

3.2 Explainable AI: Principles and Techniques

Explainable AI encompasses methods and techniques that make machine learning models interpretable to humans. Post-hoc explanation methods, which generate explanations after model training, have gained prominence due to their model-agnostic nature and applicability to complex black-box models [11]. Among these approaches, feature importance analysis represents one of the most widely used techniques for understanding how input variables influence model predictions. After the development and training of a predictive model, feature importance methods can be used to quantify the contribution of each feature to the model's decision-making process. Such analyses allow researchers to identify which variables have the greatest impact on model predictions and to better understand the relationships captured by the model.

3.3 Meta-Features and Their Role in Algorithm Selection

Meta-features serve as the foundation for meta-learning systems, characterizing datasets in ways that correlate with algorithm performance. These features typically fall into several categories: general measures (number of fetaures, number of instances, number of categorical features, number of numerical features), statistical measures (mean, variance, skewness, kurtosis), information-theoretic properties (entropy, mutual information), and complexity measures (class separability, feature correlation) [7], [4], [5], [10]. The selection and engineering of meta-features significantly impact meta-learning performance. Research has shown that different meta-features give varying importance across domains and tasks [7], [6]. For instance, in multi-label classification, meta-features related to label distribution and label relationships prove particularly influential [7], while clustering tasks prioritize structural properties like attribute sparsity [4]. Understanding meta-feature importance is crucial for several reasons. First, it enables feature selection to reduce computational overhead and improve meta-model generalization [4], [5]. Second, it provides insights into the underlying mechanisms of algorithm performance, potentially guiding algorithm design [6]. Third, it facilitates domain-specific customization of meta-learning systems by identifying which dataset characteristics matter most in particular application contexts [8].

4 Research results

This section presents results of the research divided into three parts. First, we discuss extractuon of meta-features. Next, we explain meta-model development followed by interpretation and explanation of meta-model.

4.1 Meta-Feature Extraction and Characterization

The foundation of explainable meta-learning lies in comprehensive meta-feature extraction that captures relevant data set characteristics. Hereinafter, we have extracted a total of 45 datasets. Datasets were collected from publicly available repositories and analysed in order to describe their properties through meta-features. For each dataset, 179 numerical meta-features were computed, capturing different aspects of the data such as distributional properties, statistical characteristics, and structural complexity. These meta-features represent the input space used to analyse relationships between dataset characteristics and the target classes considered in the meta-learning task. To enable consistent comparison across datasets with different scales and distributions, all numerical meta-features were discretised into ordinal categories. Specifically, each feature was partitioned into eight ordered bins, ranging from extremely low to extremely high. The discretisation was performed using a quantile-based binning procedure, which ensures approximately balanced distributions of instances across bins. In cases where quantile boundaries produced duplicate thresholds due to limited variability in the data, a fallback binning chain was applied to guarantee a valid ordinal categorisation. This discretisation step allowed meta-features to be treated as interpretable categorical descriptors of dataset characteristics. Following discretisation, a feature selection procedure was applied to identify the most informative meta-features. First, mutual information scores were computed for all 179 meta-features with respect to the target variable. Mutual information was used as a model-agnostic

measure of dependency, capturing both linear and non-linear relationships between meta-features and the target classes. To further analyse the relationship between meta-feature categories and the target classes, Kendall's rank correlation coefficient was calculated. For each meta-feature, the correlation between its ordinal category value and each target class was computed. Kendall's was selected because it is particularly suitable for ordinal variables and monotonic relationships, which aligns with the discretised nature of the meta-feature representation. Based on the obtained correlation values, the top 15 meta-features for each class were selected according to the highest absolute Kendall's scores. These selected features represent the dataset characteristics most strongly associated with each class and were subsequently used in the rule extraction and analysis stages of the study. This process enabled the identification of interpretable relationships between dataset properties and model selection outcomes within the meta-learning framework.

4.2 Meta-Learning Model Construction

To construct the meta-learning model for algorithm recommendation, a scoring-based approach was employed. The goal of the model is to estimate the suitability of each candidate machine learning algorithm for a given dataset based on its meta-feature representation.

For each dataset, meta-features describing dataset characteristics were first transformed into normalized values in order to enable consistent comparison across different datasets. First, categorical meta-features were transformed to numerical. Second, normalization was performed on scale [-1,1] Specifically, for each feature value v, the value was calculated as: $(v-3.5)/3.5$. The value 3.5 was chosen as the centering constant because the discretisation procedure maps each meta-feature to one of eight ordered bins, indexed from 1 to 8. Within this scheme, 3.5 represents the midpoint of the bin index range, effectively acting as a neutral reference point. Dividing by 3.5 further normalises the centered values such that the extreme bins (1 and 7) map approximately to the interval [1, +1], ensuring that the magnitude of contributions remains comparable across meta-features regardless of their individual distributions. Consequently, meta-feature values falling below the midpoint produce negative contributions to the algorithm score, while values above the midpoint produce positive contributions, allowing the scoring function to capture the directional influence of each dataset characteristic on algorithm suitability.

This transformation ensures that values below the central category produce negative contributions, while values above the central category produce positive contributions, allowing the model to capture directional influence of meta-features.

For each candidate algorithm class, a weighted scoring function was then applied. Each meta-feature contributes to the overall score proportionally to its associated weight parameter , which reflects the importance of that feature for predicting algorithm suitability. The score for a given algorithm class is computed as the weighted sum of the centered feature values across all meta-features. To ensure comparability of scores across classes, the resulting value was normalized by the magnitude of the weight vector . This normalization prevents classes with larger cumulative weights from being systematically favoured. Finally, the recommendation for each dataset was determined by selecting the algorithm class with the highest normalized score. In this way, the meta-model aggregates the contributions of all 15 meta-features to estimate which algorithm is most suitable for the dataset under consideration.

An important advantage of this scoring-based formulation is its inherent interpretability. Because each meta-feature contributes linearly to the final score through an explicitly defined weight, it is possible to analyse how individual dataset characteristics influence the recommendation of specific machine learning algorithms. This property makes the model particularly suitable for explainable algorithm selection in domains such as education and business analytics, where transparency and interpretability are essential.

It is important to note, that we have compared proposed meta-model against several different approaches, such as: ML based meta-models and multi-crietira decision making approaches . The comparison ensures a fair evaluation using identical meta-feature inputs and evaluation metrics. Figure 1 provides deeper insight into the behavior of the meta-model across algorithm classes. The model demonstrates strong performance in identifying Ridge regression, achieving the highest number of correct predictions (8), indicating that its associated meta-feature patterns are well captured.

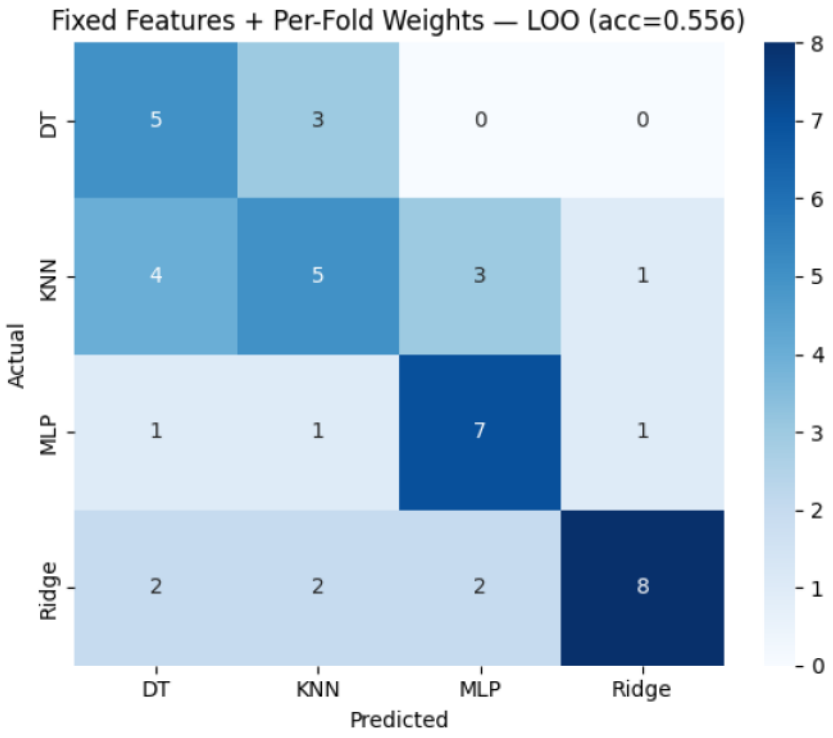


Figure 1: Confusion matrix of meta-model.

In contrast, confusion is more pronounced between DT and KNN, as well as between KNN and MLP, suggesting that these algorithms share overlapping meta-feature characteristics. This is particularly visible in the misclassification of KNN instances, which are frequently predicted as DT or MLP.

These results indicate that while the meta-model effectively distinguishes algorithms with more distinct statistical or structural signatures (e.g., Ridge), it encounters challenges when decision boundaries between algorithms are less clearly separable in the meta-feature space. This pattern suggests that the meta-learning task is inherently complex and influenced by subtle interactions between dataset characteristics, rather than dominated by a single discriminative feature. The feature importance analysis based on Kendall's tau (Figure 2) reveals several notable patterns. Features related to correlation structure, such as cor.sd and cor.mean, exhibit strong associations with specific algorithms, particularly KNN, indicating that distance-based methods benefit from datasets with distinct correlation patterns.

4.3 Explanation Generation and Interpretation

To support explainability of the meta-learning model, a sensitivity analysis was conducted in order to assess the contribution of individual meta-features to the algorithm recommendation process.The analysis was performed using a leave-one-out (LOO) evaluation setup. The baseline predictive performance of the model in this configuration was 0.556. The results of the sensitivity analysis highlight *cor.sd* as the most influential meta-feature in the model. When this feature was removed from the model, the performance decreased by 0.089, indicating a contribution to the predictive capability of the meta-model. This finding is consistent with the results obtained from the leave-one-out (LOO) analysis, confirming the stability of the importance ranking across different validation strategies. In contrast, the remaining meta-features demonstrated smaller or negligible individual effects when analysed independently. While their removal did not produce substantial decreases in predictive performance, this does not imply that they are irrelevant for the model. Instead, the results suggest that the meta-model partially relies on the complementary interaction of multiple meta-features, where the joint presence of several dataset characteristics contributes to the algorithm recommendation process.

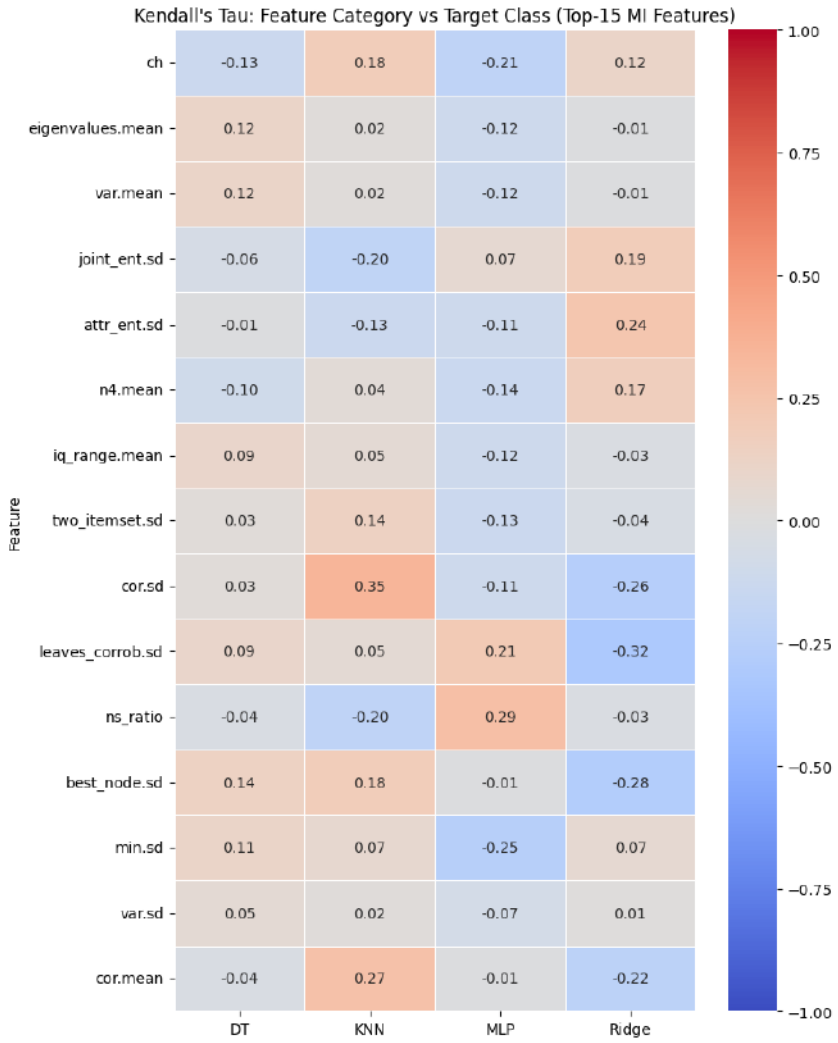


Figure 2: Meta-feature importance.

From the perspective of explainability, these findings provide valuable insights into how dataset characteristics influence the behaviour of the meta-learning model. The prominence of cor.sd indicates that measures describing the variability of attribute correlations play an important role in determining suitable machine learning algorithms for the analysed datasets. At the same time, the relatively distributed influence of the remaining meta-features suggests that the model captures more complex patterns arising from the combined properties of datasets rather than relying on a small number of dominant variables. Such interpretability is particularly important in social science domains such as education and business analytics, where researchers and practitioners require transparent justification for algorithm recommendations. By identifying which meta-features drive the model's decisions, the proposed approach enables users to better understand the relationship between dataset characteristics and algorithm suitability, thereby increasing the transparency and trustworthiness of the recommendation system.

Model-based meta-features such as leavesCorrob.sd and bestNnode.sd show strong influence, especially in distinguishing between MLP and Ridge, highlighting the importance of structural dataset properties in algorithm selection.

5 Discussion and implications

The results obtained from the feature importance (Figure 2) analysis based on scoring meta-model provide insights into how different groups of meta-features contribute to the explainability of the algorithm recommendation process. The meta-features used in this study originate from several established categories in meta-learning literature, including general, statistical, information-theoretic, itemset, model-based, clustering, complexity, landmarking, relative, and concept-based descriptors. Together, these groups capture different aspects of dataset structure and learning difficulty, enabling the meta-model to infer relationships between dataset characteristics and the suitability of particular machine learning algorithms.

An analysis reveals that certain meta-features demonstrate stronger relationships with specific algorithm classes, which provides a basis for interpretable algorithm recommendations. In particular, correlation-based statistical descriptors, such as *cor.sd* and *cor.mean*, show notable influence, especially for the KNN algorithm. The relatively strong positive correlation between cor.sd and KNN suggests that variability in attribute correlations may favour algorithms that rely on distance-based similarity measures. This finding is consistent with the intuition that neighbourhood-based methods can benefit from structured relationships between variables.

Another notable pattern appears in the general meta-features, such as *ns.ratio* and *leaves.corrob.sd*, which show stronger associations with the MLP algorithm. This suggests that datasets with more complex decision boundaries or structural variability may favour algorithms with higher representational capacity.

In contrast, several meta-features demonstrate relatively weak individual correlations with algorithm classes. Rather than indicating irrelevance, this pattern suggests that the meta-model partially relies on the combined contribution of multiple meta-features. In other words, algorithm recommendations are not driven by a single dominant dataset characteristic in most cases, but emerge from the interaction of several complementary descriptors describing dataset structure, distribution, and complexity.

From the perspective of explainable AI, these findings are particularly important. Because the meta-model is constructed using explicit weights derived from correlations between meta-features and algorithm classes, it becomes possible to trace how specific dataset properties contribute to the final recommendation score. This transparency enables researchers and practitioners to understand the reasoning behind algorithm selection decisions.

Such explainability is valuable in social science domains such as education and business analytics, where analysts often require understandable justification for automated recommendations. By linking algorithm choices to interpretable dataset characteristics, the proposed approach supports transparent and trustworthy decision support in data analysis workflows.

While the proposed scoring-based meta-model provides interpretability by explicitly linking meta-features to algorithm selection, this transparency may come at the cost of predictive accuracy compared to more complex models such as ensemble methods or AutoML systems. However, in domains such as social sciences, where explainability and trust are critical, this trade-off is often acceptable. The results suggest that the model achieves competitive performance while offering substantially higher transparency.

6 Conclusion

This paper addressed the problem of selecting appropriate machine learning algorithms for datasets in social science domains by focusing on the explainability of meta-learning approaches. Using datasets originating from education and business contexts, we developed and analysed an interpretable meta-model that recommends suitable machine learning algorithms based on dataset meta-features.

The results show that a set of relevant meta-features can effectively capture dataset characteristics that influence algorithm performance and can therefore serve as a foundation for explainable algorithm selection. Identifying these influential meta-features enables the interpretation of meta-model decisions and provides insights into why specific algorithms are recommended for particular types of datasets. In this way, meta-features not only support predictive performance of the meta-model but also act as the basis for explainability in the algorithm recommendation process.

Explainability is particularly important in domains such as education and business analytics, where data analysts and researchers are often not specialists in artificial intelligence or machine learning. Providing interpretable recommendations can therefore improve trust, transparency, and usability of decision-support systems that assist users in selecting appropriate analytical methods.

This study also has several limitations. The meta-model was trained and evaluated on a relatively limited subset of datasets and machine learning algorithms, which may restrict the generalizability of the findings. The current study considers a limited set of candidate algorithms (DT, KNN, MLP, Ridge), which may restrict the generalizability of the findings. Expanding the pool of algorithms to include ensemble methods (e.g., Random Forest, Gradient Boosting) and modern approaches would improve the practical applicability of the framework. Future work will also focus on expanding the dataset repository and further exploring explainability techniques that can enhance the transparency of algorithm recommendation systems.

Overall, the results suggest that explainable meta-learning approaches represent a promising direction for supporting data analysts in social science domains, enabling more transparent and informed selection of machine learning algorithms.

Acknowledgments and Disclosure of Funding

This research was supported by Croatian Science Foundation under the project SIMON: Intelligent system for automatic selection of machine learning algorithms and Strategic Partnership for Inovation programme under the project OptiSolarAI: Autonomus system for optimal storage and distribution of electric energy based on artificial intelligence.

References

[1] J. R. Rice, “The algorithm selection problem,” in Advances in Computers, vol. 15, pp. 65–118, 1976.

[2] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi and F. Giannotti, “A survey of methods for explaining black box models,” ACM Computing Surveys, vol. 51, no. 5, pp. 1–42, 2018.

[3] A. Adadi and M. Berrada, “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI),” IEEE Access, vol. 6, pp. 52138–52160, 2018.

[4] R. Gonzales et al., "Distance Metric Recommendation for k-Means Clustering: A Meta-Learning Approach," TENCON2022 - 2022 IEEE Region 10 Conference (TENCON), 2022. DOI: 10.1109/TENCON55691.2022.9978037

[5] M. E. M. Gonzales, L. C. Uy, J. A. L. Sy and M. O. Cordel, "Distance Metric Recommendation for k-Means Clustering: A Meta-Learning Approach," TENCON 2022 - 2022 IEEE Region 10 Conference (TENCON), Hong Kong, Hong Kong, 2022, pp. 1-6, doi: 10.1109/TENCON55691.2022.9978037.

[6] M. Garouani, A. Ahmad, and M. Bouneffa, “Explaining meta-features importance in meta-learning through Shapley values,” in Proc. ICEIS, vol. 1, pp. 591–598, Apr. 2023.

[7] A. Kostovska, C. Doerr, S. Džeroski, D. Kocev, P. Panov and T. Eftimov, "Explainable Model-specific Algorithm Selection for Multi-Label Classification," 2022 IEEE Symposium Series on Computational Intelligence (SSCI), Singapore, Singapore, 2022, pp. 39-46, doi: 10.1109/SSCI51031.2022.10022177.

[8] D. Oreški, D. Višnjić, and N. Kadoić, “Unlocking automated machine learning efficiency: Meta-learning dynamics in social sciences for education and business data,” TEM Journal, vol. 13, no. 1, pp. 797–808, Feb. 2024, doi: 10.18421/TEM131-82.

[9] A. Barhrhouj, B. Ananou, and M. Ouladsine, “Exploring explainable machine learning for enhanced ship performance monitoring,” in Machine Learning, Optimization, and Data Science, G. Nicosia, V. Ojha, S. Giesselbach, M. P. Pardalos, and R. Umeton, Eds., Lecture Notes in Computer Science. Cham, Switzerland: Springer, 2024

[10] S. Manna and N. Sett, “Need of AI in modern education: In the eyes of explainable AI (XAI),” in Blockchain and AI in Shaping the Modern Education System, Boca Raton, FL, USA: CRC Press, 2025, pp. 89–115.

[11] T. Han, S. Srinivas, and H. Lakkaraju, “Which explanation should I choose? A function approximation perspective to characterizing post hoc explanations,” in Advances in Neural Information Processing Systems, vol. 35, pp. 5256–5268, 2022.

Digital Transformation in Animal and Agricultural Sciences

Vision-based detection of pain and nest-building behaviors in sows within commercial farrowing pens

Peter Helf
Precision Livestock Farming Hub
University of Veterinary Medicine Vienna
Veterinärplatz 1, 1210 Vienna, Austria
peter.helf@vetmeduni.ac.at

Maciej Oczak
Precision Livestock Farming Hub
University of Veterinary Medicine Vienna
Veterinärplatz 1, 1210 Vienna, Austria
maciej.oczak@vetmeduni.ac.at

Abstract

Pain indicators and preparturient nest-building are precursors of farrowing, yet continuous quantification remains challenging. We present a non-invasive computer-vision system that detects pain-associated behaviors (back-arching, tail-flicking, back leg forward, trembling) and nest-building behaviors (manipulation of pen components, pawing, exploration) from top-view videos. We analyzed 748 h of RGB footage (25 fps) from 11 sows on a single farm, spanning 64 h pre-farrowing to 4 h post birth of the first piglet. Using a defined ethogram, 46,010 events were annotated with inter-annotator agreement $\kappa = 0.724$. To assess generalization, we used a sow-level split, i.e., 8 in the training set and 3 in validation. Behaviors were detected with a modified DeepEthogram architecture, combining RGB data and optical flow. Both streams were processed by separate ResNet3D-34 encoders. Optical flow was estimated using a state-of-the-art DPFlow model. Training employed focal loss to address class imbalance, alongside geometric and photometric augmentations for robustness to camera placement and lighting. Clips of 11 frames at 8.33 fps ($\approx$1.32 s) were used. On held-out sows, per-class F1 scores were 0.875 for manipulation of pen, 0.634 for pawing, 0.820 for exploration, 0.639 for back-arching, 0.778 for tail-flicking, 0.871 for back leg forward, and 0.443 for trembling. These results indicate that pen-installed vision can identify key behaviors in a non-invasive way, supporting scalable monitoring. Limitations include modest dataset size and limited diversity, i.e., a single farm and a single breed. Ongoing work will expand the dataset and leverage behavior dynamics for time-to-farrowing estimation.

1 Introduction

Accurate, continuous monitoring of the preparturient period is critical for timely interventions at the onset of farrowing to improve sow welfare and piglet survival. Pain indicators (e.g., back arch, tail flick, trembling) and nest building behaviors (e.g., manipulation of pen, pawing, exploration) are possible precursors to the onset of farrowing [1, 2]. However, routine quantification of these events remains challenging because manual assessment is labor intensive and difficult to sustain at scale.

Sensor based approaches have estimated farrowing proximity by measuring overall activity with accelerometers attached to sows in farrowing pens [3]. While increases in activity can signal nest building, these methods require animal handling and device attachment, which may introduce stress and demand upkeep. Vision based, non-invasive alternatives have inferred activity from video, either by tracking sow position via object detection [4] or by estimating motion with optical flow [5]. Both strategies have shown promise for farrowing prediction but largely operate on coarse activity proxies rather than specific behaviors. Recognizing discrete pain and nest building behaviors may provide more reliable and interpretable cues for time-to-farrowing estimation.

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

Recent advances in video-based behavior recognition enable automatic detection of a range of sow behaviors across modalities (RGB, grayscale, depth) and tasks (e.g., posture, drinking, aggression) with real time performance on farm [6, 7]. To our knowledge, no prior work has targeted pain and nest building behaviors. We address this gap with a non-invasive computer vision system that detects specific pain-associated and nest-building behaviors from top-view pen videos.

2 Methods

2.1 Data and annotations

Data were collected at Medau, the pig research and teaching farm of the University of Veterinary Medicine Vienna, using top-view RGB cameras. A total of 748 hours of video, recorded at 25 frames per second (fps), from 11 individually housed Large White sows was used. Each sow was monitored continuously from approximately 64 hours before farrowing until 4 hours after the birth of the first piglet, capturing both the preparturient period and the onset of farrowing under routine farm conditions. All sows were kept in BeFree pens (Schauer Agrotronic, Prambachkirchen, Austria).

Table 1: Ethogram used for the labeling of behavior events

Behavior	Definition
Pawing	Raking movement of either forelimb against the floor or the pen
Manipulation of pen	The sow touches the pen component with some part of her head and performs up- and down-movements with her head towards the pen components.
Exploratory behavior	Movements of the head while the snout of the sow is directed towards the ground
Tremble	Visible shaking as if shivering when in a lateral lying position
Back leg forward	In a lateral lying position, the back leg is pulled forward and/or in toward the body
Back arch	In a lateral lying position, one or both sets of legs become tense and are pushed away from the body and/or inwards toward the center
Tail flick	Tail is moved rapidly up and down

An ethogram was defined to cover two behavior groups relevant to farrowing onset. Pain-associated behaviors included back arch, tail flick, pulling the back leg forward, and trembling. Nest building behaviors included manipulation of pen components, pawing, and exploration. Definitions are provided in Table 1. Because multiple behaviors can occur simultaneously, the task is multi-label. Using this ethogram, annotators labeled discrete behavior events throughout the recordings, producing a total of 46,010 events. To assess annotation reliability, a subset of the data, totaling 9 hours of video footage, was double-annotated, yielding inter-annotator agreement with κ=0.724. The dataset exhibited class imbalance, driven primarily by the higher frequency of back leg forward events relative to other categories. To partially mitigate this imbalance while preserving ethogram breadth, we downsampled the event pool to 23,326 by randomly removing 80% of back leg forward instances. The class composition of the resulting dataset is summarized in Table 2. To evaluate generalization and prevent data leakage, we used a sow-level split with 8 sows for training and 3 sows held out for validation.

Table 2: Number of occurrences for each behavior in the dataset

Behavior	Count
Pawing	1,950
Manipulation of pen	5,428
Exploratory behavior	4,938
Tremble	1,267
Back leg forward	5,671
Back arch	2,045
Tail flick	2,027

2.2 Neural network and training

Raw RGB streams at 25 fps were downsampled by keeping every third frame (≈8.33 fps) to reduce redundancy and computational load. Fixed-length clips of 11 consecutive frames (duration ≈1.32 s) were extracted within the start–end timestamps of annotated events. Although the model receives an 11-frame clip, the ground-truth labels of the center frame (6th) are assigned to each clip to align supervision with the temporal midpoint while providing context on both sides. Images were resized to a width of 416 pixels while keeping the aspect ratio. This resulted in an image height of 499 pixels. The images were then normalized using the ImageNet [8] mean and standard deviation. For robustness to real-world variability, we applied geometric and photometric augmentations during training, including random horizontal flips, small rotations and scalings, and brightness/contrast jitter. Because multiple behaviors can co-occur, the ground truth is multi-label.

To capture motion cues critical for transient behaviors, we adopted the dual-stream architecture of DeepEthogram [9]. The RGB stream uses 11-frame clips, and the motion stream uses corresponding optical-flow stacks computed with DPFlow [10]. A pretrained model provided by the Python package PyTorch Lightning Optical Flow [11] and pretrained on the Spring dataset [12]. A representative flow stack is shown in Figure 1. Each stream is encoded with a separate ResNet3D-34 backbone [13] to produce feature embeddings. Pretrainted weights provided by DeepEthogram [9], trained on Kinetics700 [14], were used. Each stream outputs class-wise logits for the seven behaviors. Late fusion combines the per-class logits from both streams using learnable weights before the final sigmoid:

$$l_c = \alpha_c l_c^{RGB} + (1 - \alpha_c) l_c^{Flow}, \quad y_c = \sigma(l_c) \tag{1}$$

where c indexes behaviors, $\alpha_c \in [0, 1]$ are learnable fusion weights, and $\sigma(\cdot)$ denotes the sigmoid. A threshold of 0.5 was used to determine the presence of a behavior.

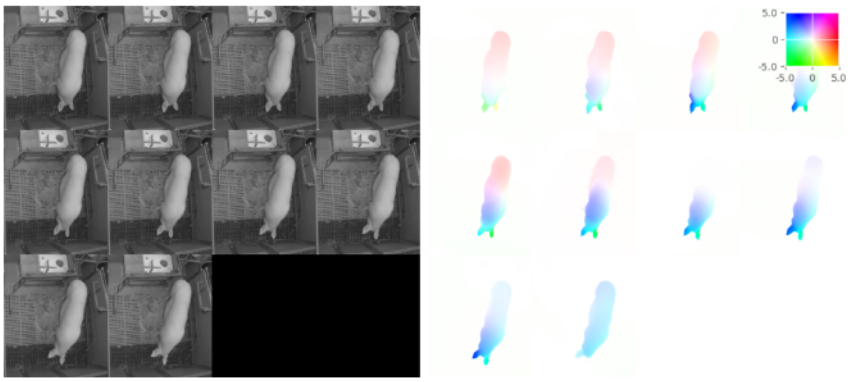


Figure 1: Optical flow stack. Left: RGB frame sequence (recording taken at night, thus a grayscale image) Right: computed optical flow. Pixel color encodes direction of movement with higher saturation indicating larger movements.

Given the remaining class imbalance, the training objective employed focal loss to down-weight common examples and emphasize minority-class instances [9, 15]. We used an initial learning rate of 10^{-3} with a ReduceLROnPlateau scheduler (patience: 5, factor: 0.1, min_lr=10^{-6}). Training stopped once the minimum learning rate was reached, and the checkpoint with the lowest validation loss was selected.

Training was performed in Python 3.11.10 with PyTorch 2.9.0 [16] using OpenCV 4.10.0.84 [17] for video I/O. Kornia 0.7.4 [18] was used for image transformation and augmentation. Experiments ran on an NVIDIA RTX 6000 Ada GPU (NVIDIA, Santa Clara, USA) in a server with an AMD EPYC 9354 32-core CPU (AMD, Santa Clara, USA).

3 Results

On the sow-held-out validation set (3 sows), the model achieved strong performance on nest-building behaviors and mixed performance on pain associated behaviors. For nest building, per-class F1 scores were 0.875 for manipulation, 0.634 for pawing, and 0.820 for exploration. For pain behaviors, F1 scores were 0.639 for back arch, 0.778 for tail flick, 0.871 for back leg forward, and 0.443 for trembling. Averaged over all seven classes, the macro F1 was 0.723. Table 3 summarizes per-class precision, recall, and F1 score. These results indicate that behaviors with more distinctive or sustained kinematics tended to yield higher F1 scores.

Table 3: F1 score, precision, and recall for the detected behaviors

Behavior	F1	Precision	Recall
Pawing	0.634	0.704	0.576
Manipulation of pen	0.875	0.835	0.919
Exploratory behavior	0.820	0.827	0.814
Tremble	0.443	0.383	0.524
Back leg forward	0.871	0.907	0.838
Back arch	0.639	0.854	0.511
Tail flick	0.778	0.833	0.730
Average	0.723	0.763	0.702

4 Discussion

Our results demonstrate that a dual-stream model can reliably detect several pre-parturient nest-building behaviors, with mixed performance on pain-associated behaviors. High F1 for manipulation and exploration suggests pronounced movements and interactions with pen fixtures are well captured by short clips. Tail flick, which is most often a short event, was detected well, indicating that distinctive movement can suffice. In contrast, trembling and back arch, often subtler movements, achieved lower scores.

The current clip length ($\approx$1.32 s) and frame rate ($\approx$8.33 fps) balance context and efficiency but may under-represent higher frequency motions (e.g., fine tremors) or longer sequences needed for some events. Extending temporal context, adopting other temporal models (e.g., Transformers), or using multi-rate inputs (retaining higher fps for flow while keeping RGB downsampled) may improve sensitivity to subtle and longer events.

Limitations include a modest dataset from a single farm, breed, and pen type, limiting visual diversity. Despite downsampling and focal loss, residual class imbalance persists and may bias decision boundaries. The single top-view perspective limits cues available for fine-grained posture changes, which may be relevant for some behaviors. Annotation noise, while mitigated by substantial agreement (κ=0.724), can still affect the results. Finally, training and validation used clips within behavior intervals without explicit background sampling. Continuous-video evaluation is needed to assess real-world prevalence and precision.

Future work should expand data across farms, breeds, camera placements, and lighting. Including background clips without any behaviors in the dataset will support more realistic evaluation. Longer-horizon temporal models (e.g., Transformers), may enhance sensitivity to some behavior indicators. Finally, the change in preparturition behavior frequency detected by the model needs to be further processed to achieve a time-to-farrowing estimation.

Acknowledgments and Disclosure of Funding

This research was funded in whole or in part by the Austrian Science Fund (FWF) [https://doi.org/10.55776/DFH34]

References

[1] Sarah H Ison, Susan Jarvis, and Kenneth MD Rutherford. The identification of potential behavioural indicators of pain in periparturient sows. *Research in veterinary science*, 109: 114–120, 2016.

[2] BI Damm, L Lisborg, KS Vestergaard, and J Vanicek. Nest-building, behavioural disturbances and heart rate in farrowing sows kept in crates and schmid pens. *Livestock Production Science*, 80(3):175–187, 2003.

[3] Maciej Oczak, Kristina Maschat, and Johannes Baumgartner. Dynamics of sows' activity housed in farrowing pens with possibility of temporary crating might indicate the time when sows should be confined in a crate before the onset of farrowing. *Animals*, 10(1):6, 2019.

[4] Maciej Oczak, Kristina Maschat, and Johannes Baumgartner. Implementation of computer-vision-based farrowing prediction in pens with temporary sow confinement. *Veterinary Sciences*, 10(2):109, 2023.

[5] Kejian Liu, Yigui Huang, Junbin Liu, Zujie Tan, and Deqin Xiao. Prediction of sow farrowing onset time using activity time series extracted by optical flow estimation. *Animals*, 15(7):998, 2025.

[6] Ziting Zhang, Hang Zhang, Yuxiang He, and Tonghai Liu. A review in the automatic detection of pigs behavior with sensors. *Journal of Sensors*, 2022(1):4519539, 2022.

[7] Yuanqin Zhang, Jiahao Cai, Deqin Xiao, Zesen Li, and Benhai Xiong. Real-time sow behavior detection based on deep learning. *Computers and Electronics in Agriculture*, 163:104884, 2019.

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[9] James P Bohnslav, Nivanthika K Wimalasena, Kelsey J Clausing, Yu Y Dai, David A Yarmolinsky, Tomás Cruz, Adam D Kashlan, M Eugenia Chiappe, Lauren L Orefice, Clifford J Woolf, et al. Deepethogram, a machine learning pipeline for supervised behavior classification from raw pixels. *elife*, 10:e63377, 2021.

[10] Henrique Morimitsu, Xiaobin Zhu, Roberto M Cesar, Xiangyang Ji, and Xu-Cheng Yin. Dpflow: Adaptive optical flow estimation with a dual-pyramid framework. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17810–17820, 2025.

[11] Henrique Morimitsu. Pytorch lightning optical flow. `https://github.com/hmorimitsu/ptlflow`, 2021.

[12] Lukas Mehl, Jenny Schmalfuss, Azin Jahedi, Yaroslava Nalivayko, and Andrés Bruhn. Spring: A high-resolution high-detail dataset and benchmark for scene flow, optical flow and stereo. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4981–4991, 2023.

[13] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 3154–3160, 2017.

[14] Joao Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*, 2019.

[15] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[16] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[17] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[18] Edgar Riba, Dmytro Mishkin, Daniel Ponsa, Ethan Rublee, and Gary Bradski. Kornia: an open source differentiable computer vision library for pytorch. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3674–3683, 2020.

Online adaptive path planning of UAVs for weed detection

Wolfgang Pitzl **Lukas Lachmann** **Raphael Völker**

Peter Riegler-Nurscher

Josephinum Research
Wieselburg, AT 3250
{wolfgang.pitzl, lukas.lachmann, raphael.voelker, p.riegler-nurscher }
@josephinum.at

Abstract

Problem weeds pose a challenge for agriculture. Robust detection of these plants is crucial for their control and for assessing possible contamination of the crop. Current UAV inspections are usually carried out using a fixed flight route, regardless of the extent of weed infestation. Initial approaches of adaptive flight control attempt to save flight distance by adapting flight altitude and route. We would also like to investigate adaptive gimbal guidance, to find out whether this can have a positive impact on flight paths and flight times. To this end, we developed a pipeline between the server including the operating website and the drone. The initial results, presented in this extended abstract, examine the functional capability of the pipeline and different error sources for GPS accuracy of adaptive gimbal pitches.

1 Introduction

Problem weeds such as thistles, which reproduce via underground rhizomes, and neophytes like the poisonous thorn apple pose challenges for agriculture. Reliable detection is essential for targeted control measures such as spot spraying and for assessing possible crop contamination.

Conventional UAV flights using lawn-mower coverage achieve high accuracy but cannot adapt to weed distribution. Regardless of weed density, the drone covers the entire field without evaluating or adjusting its flight path, and most commercial drones cannot perform evaluations during flight.

Adaptive approaches already attempt to adjust altitude and flight routes in real time and can achieve shorter flight paths (see: [1]–[5]). However, these studies do not consider variable gimbal positioning, as images are always taken in nadir view. This project investigates whether gimbal control can reduce flight path length and time without significantly affecting detection or localisation accuracy. As a first step, this extended abstract examines the functionality of our pipeline and potential GPS localisation errors at adaptive gimbal pitches.

2 Related Work

Previous research has explored online adaptive UAV flight paths for efficient mapping. The work of [1] used a predefined route with an OODA loop to recognize Aruco markers, red surfaces, and thistle plants. Paper [2] introduced an informative path planning strategy that maximizes information collection while respecting resource constraints, demonstrating higher efficiency than traditional

The Third Austrian Symposium on AI, Robotics, and Vision (AIROV26).

approaches. In [3], adaptive flight height was used to reduce flight time while maintaining high segmentation accuracy. Similarly, [4] combined high-altitude coverage with low-altitude inspection paths, reducing flight lengths by 37% for clustered objects and 6% for uniformly distributed objects. The work in [5] applied deep Q-learning to learn search policies that outperformed baseline row-by-row flight paths and were transferable to real-world data. These studies indicate that online adaptive flights can shorten flight paths, but all rely on nadir images.

The work of [6] evaluated the accuracy of the Matrice 600 Pro with onboard GNSS RTK, producing photogrammetric products with decimeter accuracy. In contrast, our project uses drone footage captured from a standstill. For our application, accurate YAW indication is also important, yet existing research mainly focuses on RTK precision, as shown in studies [7] and [8].

3 Method

In this project, we evaluate GPS accuracy from drone images and compute adaptive routes on a server, with the aim of testing live processing over mobile networks in a later stage. This approach allows easy integration of different drones, as the drone only executes simple tasks provided by the server. Our goal is also to adapt this server-based pipeline for future research projects.

The system is based on a FastAPI application running on the server, providing API endpoints and web interfaces for control and monitoring. Through the control interface, the drone can be operated in three modes:

- Manual mode, where the user selects a target GPS position including altitude.
- Route mode, where predefined routes are uploaded and executed.
- Adaptive mode, where only a field shape file and an algorithm need to be selected.

The drone receives tasks from the server containing a GPS position, altitude, and an action. Possible actions include: (1) YAW rotation at the arrival point, (2) single-photo capture with wide-angle or zoom camera and selectable gimbal pitch, and (3) multi-photo capture with defined start and end points for gimbal pitch and YAW rotation.

Currently, we are developing the first adaptive algorithm, which:

1. Captures multi-photography (wide-angle) of the entire field or large sections of the field.
2. Evaluates suspected weed positions on the server using segmentation models and EXIF data while images are still being captured.
3. Generates zoom-camera single-photo tasks based on the detected GPS positions.

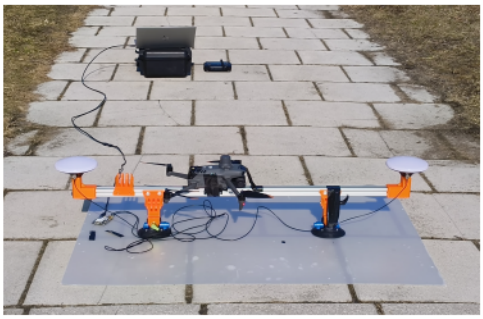
To assess the algorithm's accuracy, we examine potential error sources: (1) drone heading (YAW) accuracy, (2) GPS coordinates on the drone, and (3) the target object's position in the image. Experiments are conducted using a DJI Mavic 3T. While each factor will be tested repeatedly in the future to analyze variability, the current results are based on a single test per aspect.

3.1 Heading (Yaw)-accuracy of the drone

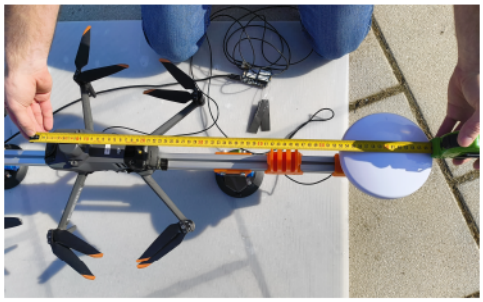
We use a simpleRTK3B Pro to measure Heading with a central 116 cm aluminum profile and RTK-antennae at each end under the drone, and fixed the drone and gimbal to the aluminum profile (1a). We compare the heading from the simpleRTK3B Pro with gimbal-heading information from EXIF-data of UAV photos taken. As we cannot align the drone with perfect precision, we will get a constant offset for all measurements. This offset, estimated from the average misalignment will be removed from the measured difference. Overall we rotate the setup 360 degree and take 14 different measurements.

3.2 GPS-coordinate on the drone

From the previous heading (yaw)-accuracy test, we calculate the average distance from the simpleRTK3B-GPS-Position to the UAV-GPS-position (1b). We use this to estimate which GPS-Location on the drone is written into the EXIF-data.



(a) Heading (Yaw)-accuracy of the drone



(b) GPS-coordinate on the drone

Figure 1: Investigate potential error sources part 1

3.3 Impact of position of the target object in the image

A total of 13 markers were measured with an RTK antenna on a flat surface (6 × 10 m). From a central marker, three rows with four markers each extend in vertical, horizontal, and diagonal directions, ranging from markers 0 to 4 (2). The UAV was manually positioned so that the central marker appeared in the image centre and the outer markers near the edges. This results in a flight altitude of 14.34 m and a ground sample distance (GSD) of 5.2 mm/pixel for this setup, as a wide-angle camera with a focal length of 4.4 mm and a resolution of 4000 x 3000 pixels was used. The image was undistorted using OpenCV Camera Calibration. To compensate for RTK error in the UAV image, the calculated centre marker position was shifted to the ground-truth position, and all other marker positions were adjusted by the same offset. The calculated GPS coordinates were then compared with the RTK measurements.

Figure 2: Impact of image position

4 Results

4.1 Server-Drone-Pipeline

The visualization of the Server-Drone-Pipeline shows logged actions from an adaptive flight from the remote control. From 0 to 80 seconds the UAV performs a multi-photo at two different gimbal settings, the logs of yaw rotations are in the third row. At around 95 seconds, the drone switches the camera and performs 3 single-photos, each time rotating gimbal and yaw.

4.2 Heading (Yaw)-accuracy of the drone

10 of 14 Measurements are in the standard deviation of the simpleRTK3B. We conclude from this data that the UAV heading doesn't have a high noise and the accuracy is ±0.4 degree. We were unable to find any reasons in other EXIF parameters as to why the values 3,4 and 12 are noticeably outside the standard deviation. Therefore, we will repeat this test during the course of our project to make reliable conclusions and try to find out why some values are outside the standard deviation.

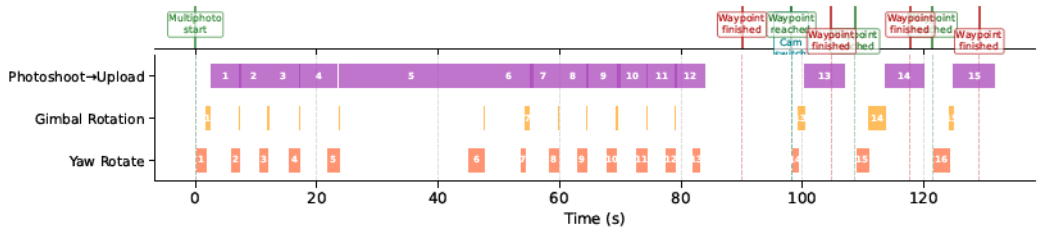


Figure 3: Visualization of pipeline from the UAV remote control

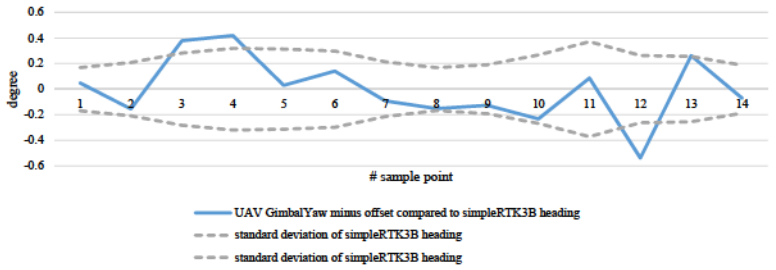
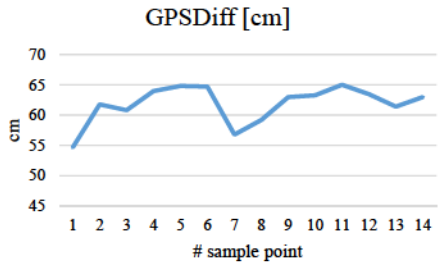


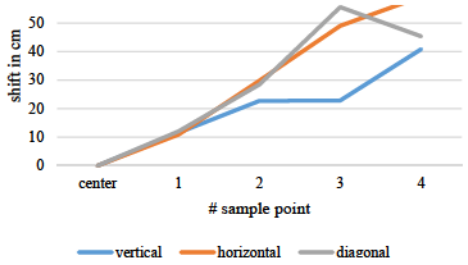
Figure 4: Heading (Yaw)-accuracy of the drone

4.3 GPS-coordinate on the drone

The average distance from the simpleRTK3B-GPS-Position to the GPS-Position give of the drone is 61,84cm and ranges from 54,7cm to 64,8cm (5a). Figure fig:distance shows that the distance from the centre of the white RTK antenna to the drone's gimbal is approximately 61 cm. We conclude from this data, that the GPS-Position given by the drone is the GPS-Position of the gimbal. However, as the scattering of the distance is relatively high, we will repeat this test several times.



(a) GPS-coordinate on the drone



(b) Impact of position of the target object in the image

Figure 5: Diagrams of potential error sources part 2

4.4 Impact of position of the target object in the image

The graph (5b) is showing, that for most of the marker points its true, that the further away a marker is from the center (marker 0) the higher is the shift between the calculated GPS from the image to the measured GPS-coordinate.

5 Discussion

The presented server–drone pipeline appears promising, and future work will evaluate it in different scenarios for weed detection and compare flight path length and time with conventional lawn-mower coverage flights. Apart from weed detection, we could envisage using this drone-server pipeline to monitor field robots. Or even for gesture control, where very specific and custom drone actions for different gestures are stored on the server.

The yaw accuracy of the drone was tested only on the ground, as evaluating it during flight is technically challenging. This limitation reduces the validity of the results. Following the initial test of yaw accuracy, we are not yet able to explain why some values fall outside the standard deviation. We hope to gain further insights through further repetitions and by calibrating the compass using the DJI Pilot 2 app.

We also really want to find out why the shift increases as the distance from the centre point (5b) increases. We suspect this may be due to incorrect photo-correction parameters or minor errors in calculating the GPS positions. We need to examine our approach in more detail.

Following this initial round of testing, we are still cautious about making any definitive statements. As mentioned earlier, the tests for the different error sources will be repeated multiple times to obtain more reliable results.

In addition, the ongoing project will investigate the influence of different gimbal pitches on GPS localisation accuracy. We suspect that the higher the gimbal pitches (i.e. the further the drone looks away), the greater the error in calculating GPS positions. Perhaps high gimbal pitches always produce positions that are too imprecise for spot spraying. It would be important to know this to limit the adaptive gimbal pitch to a specific range.

References

[1] Alsalam, B. H. Y., Morton, K., Campbell, D., & Gonzalez, F. (2017, March). Autonomous UAV with vision based on-board decision making for remote sensing and precision agriculture. In 2017 IEEE Aerospace Conference (pp. 1-12). IEEE.

[2] Popović, M., Vidal-Calleja, T., Hitz, G., Chung, J. J., Sa, I., Siegwart, R., & Nieto, J. (2020). An informative path planning framework for UAV-based terrain monitoring. Autonomous Robots, 44(6), 889-911.

[3] Stache, F., Westheider, J., Magistri, F., Stachniss, C., & Popović, M. (2023). Adaptive path planning for UAVs for multi-resolution semantic segmentation. Robotics and Autonomous Systems, 159, 104288.

[4] van Essen, R., van Henten, E., Kooistra, L., & Kootstra, G. (2025). Adaptive path planning for efficient object search by UAVs in agricultural fields. Smart Agricultural Technology, 12, 101075.

[5] van Essen, R., van Henten, E., & Kootstra, G. (2025). UAV-based path planning for efficient localization of non-uniformly distributed weeds using prior knowledge: A reinforcement-learning approach. Computers and Electronics in Agriculture, 237, 110651.

[6] Ekaso, D., Nex, F., & Kerle, N. (2020). Accuracy assessment of real-time kinematics (RTK) measurements on unmanned aerial vehicles (UAV) for direct geo-referencing. Geo-spatial information science, 23(2), 165-181.

[7] Czyża, S., Szuniewicz, K., Kowalczyk, K., Dumalski, A., Ogrodniczak, M., & Zieleniewicz, Ł. (2023). Assessment of accuracy in unmanned aerial vehicle (uav) pose estimation with the real-time kinematic (rtk) method on the example of dji matrice 300 rtk. Sensors, 23(4), 2092.

[8] Lewicka, O., Specht, M., & Specht, C. (2022). Assessment of the Steering Precision of a UAV along the Flight Profiles Using a GNSS RTK Receiver. Remote Sensing, 14(23), 6127.

Lightweight Classification of Canine Eye Diseases

Isselmou Abdarahmane[1] **Peter M. Roth** [2,3]

isselmou.abdarahmane@fhwn.ac.at peter.roth@webster.ac.at

[1]Medial University of Vienna
[2]University of Veterinary Medicine, Vienna
[3]Webster Vienna Private University

Abstract

Eye diseases in dogs are visually similar and difficult to distinguish without professional examination. Furthermore, assessing the severity of such conditions – and in particular determining whether immediate veterinary attention is required – poses a significant challenge for pet owners. To address this problem, we aim to assist pet owners in performing an initial triage of canine eye conditions with minimal technical expertise required. Given a photograph of the affected eye taken with a smartphone, we provide an automated preliminary assessment indicating whether a veterinary visit is advisable. To this end, we employ a computationally efficient convolutional neural network (CNN) to classify the images, identifying potential conditions and reporting the result to the user.

1 Introduction and Problem Statement

Veterinary ophthalmology is a well-studied field, e.g., [4, 5, 9, 13], However, as eye diseases in dogs are visually similar, they are difficult to distinguish laypersons. Furthermore, assessing the severity of such conditions – and in particular determining whether immediate veterinary attention is required – poses a significant challenge for pet owners. To address this problem, we aim to assist pet owners in performing an initial triage of canine eye conditions with minimal technical expertise required. Given a photograph of the affected eye taken with a smartphone, we provide an automated preliminary assessment indicating whether a veterinary visit is advisable. To this end, we employ a computationally efficient convolutional neural network (CNN) to classify the images, identifying potential conditions and reporting the result to the user. Specifically, we target four common canine ocular conditions: corneal oedema, episcleral hyperaemia, epiphora, and cherry eye.

The automated analysis of medical images using deep learning has seen considerable progress in recent years, with convolutional neural networks (CNNs) achieving strong performance across a range of diagnostic tasks, including the classification of diabetic retinopathy, skin lesions, and pathological findings in radiology [6, 10]. More recently, these methods have begun to be applied in veterinary medicine, where they offer the potential to assist practitioners and pet owners alike [1].

In the context of ocular disease, deep learning-based approaches have been explored primarily for human ophthalmology, with well-established benchmarks for conditions such as glaucoma and age-related macular degeneration [10]. Comparable work in veterinary ophthalmology remains sparse. A closely related study [7] developed CNN-based models to classify the severity of corneal ulcers in dogs from photographic images, using GoogLeNet, ResNet, and VGGNet architectures fine-tuned on ImageNet, achieving classification accuracies above 90 %. More recently, a U-Net-based approach was proposed for the segmentation and diagnosis of multiple canine ocular conditions from smartphone and camera images [2], demonstrating the feasibility of image-based triage tools for pet owners.

The Third Austrian Symposium on AI and Vision (AIROV26).

Lightweight CNN architectures such as MobileNetV2 [12] have been specifically designed for deployment on resource-constrained devices, making them well suited for mobile applications. Transfer learning from large-scale datasets such as ImageNet has been shown to yield strong results even when task-specific training data is limited [11], which is particularly relevant in veterinary contexts where large labeled datasets are rarely available.

In this work, we adopt MobileNetV2 as a backbone and fine-tune it on the *DogEyeSeg4* dataset for the classification of four canine ocular conditions. To the best of our knowledge, this represents one of the first attempts to apply mobile-efficient deep learning specifically to the triage of canine eye diseases from smartphone images. The rest of the paper is organized as follows: Sec. 2 describes the relevant technical implementation. Sec. 3 presents the experimental setup and results, and Sec. 4 summaries the findings and outlines directions for future work.

2 Implementation Details

The envisioned application requires a solution that is accurate enough to provide a meaningful initial assessment, yet lightweight enough to run on a standard smartphone without specialized hardware. To meet these constraints, the classification system is implemented in Python and built upon a MobileNetV2 backbone [12] pre-trained on ImageNet [3]. The backbone weights are kept frozen during training, and only the custom classification head is trained, which reduces the number of trainable parameters and mitigates overfitting on the small available dataset. The design prioritizes computational efficiency, requiring no specialized hardware for either training or inference, and is therefore well suited for eventual deployment as a mobile application.

The input images are rescaled to a uniform size of 224×224 pixels with three color channels (RGB), matching the input format expected by MobileNetV2. The frozen backbone extracts a feature vector of dimensionality 1280 via Global Average Pooling, which collapses the final $7 \times 7 \times 1280$ feature map into a compact representation. This feature vector is then passed through a custom classification head consisting of two fully connected layers with ReLU activations, batch normalization, and Dropout regularization, before a final Softmax layer produces a probability distribution over the four target classes: corneal oedema, episcleral congestion, epiphora, and cherry eye.

The network is trained using the Adam optimizer [8] with a learning rate of 0.0001 and a batch size of 12. To prevent overfitting, early stopping is applied, terminating training once no further improvement on the validation set is observed; the maximum number of epochs is set to 250. The available data is partitioned into 70 % for training, 10 % for validation, and 20 % for testing.

3 Experimental Results

To demonstrate the functionality of the proposed approach, we run it on a publicly available dataset, namely *DogEyeSeg4_dataset1*[1]. This dataset comprises 145 images at a resolution of 320×320 pixels, depicting the following conditions:

- Corneal oedema
- Episcleral hyperaemia
- Epiphora (excessive tearing)
- Cherry eye (nictitating gland prolapse)

The results achieved over the course of training are shown in Fig. 1. The left panel displays the trajectory of the loss function – the optimization criterion being minimized – while the right panel shows the development of classification accuracy. In both cases, it is evident that performance on the independent validation data, which was not accessible to the model during training, improves with increasing epochs before eventually plateauing. By employing early stopping, training is automatically halted at this point, as no further improvement on the validation data is expected.

[1] https://pmc.ncbi.nlm.nih.gov/articles/PMC11467576/

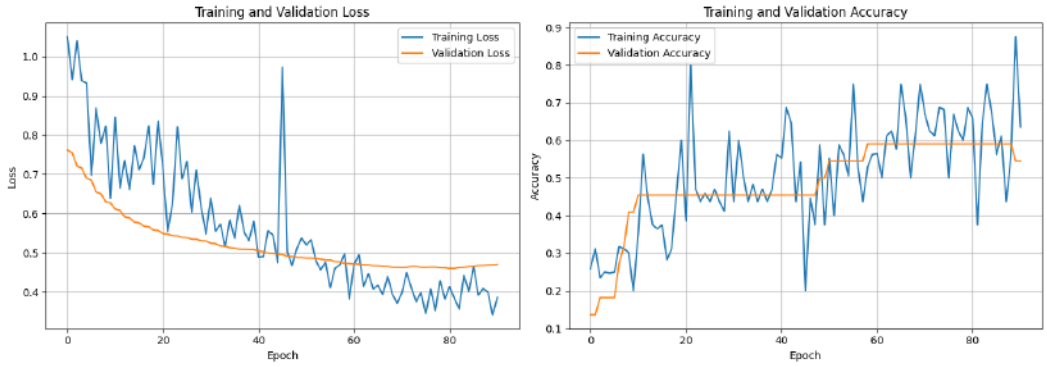
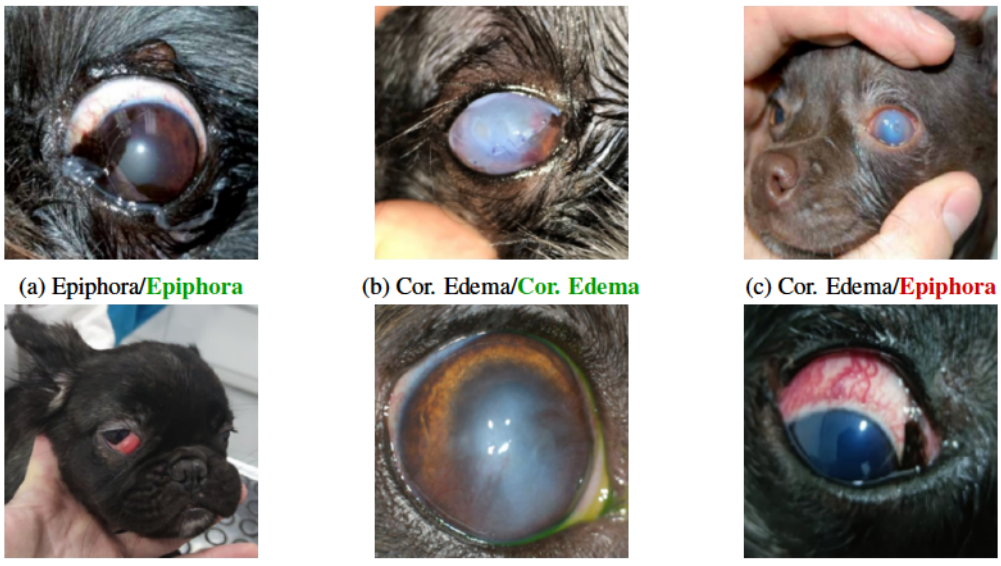


Figure 1: Loss curves and classification accuracy over training epochs.

The class-specific results are presented in Table 1. Three metrics are reported: the *Precision* describes the proportion of correctly classified positive images among all positively predicted images, indicating how reliable a positive prediction is. The *Recall* gives the proportion of correctly identified positive images among all truly positive images, measuring how completely the model detects each class. The *F1-score* is the harmonic mean of precision and recall, combining both metrics into a single value that accounts for both false positives and false negatives. Additionally, Fig. 2 shows illustrative examples of correctly and incorrectly classified images. From the results in Table 1 and the training curves in Fig. 1, it can be seen that the detection accuracy needs to be increased and the training behavior is still somewhat noisy. Both aspects can be addressed by extending the training set.

Table 1: Quantitative results per class.

Condition	Precision	Recall	F1-Score
Corneal Edema	0.71	0.68	0.69
Episcleral Congestion	0.89	0.86	0.86
Epiphora	0.59	0.59	0.59
Cherry Eye	0.74	0.82	0.78



(a) Epiphora/**Epiphora** (b) Cor. Edema/**Cor. Edema** (c) Cor. Edema/**Epiphora**

(d) Cherry Eye/**None** (e) Cor. Edema/**Cor. Edema** (f) Cor. Edema/**Cor. Edema**

Figure 2: Illustrative classification results: true label / predicted label.

4 Conclusion

The goal of this study was to evaluate whether a lightweight classification system could be applied to automatically triage canine eye diseases, enabling pet owners without veterinary expertise to determine whether a visit to a veterinary practice is necessary. Our results demonstrate that the proposed approach achieves promising results across four common canine ocular conditions, confirming that the approach is generally feasible. However, as the available dataset is limited, the current performance does not yet meet the threshold required for clinical deployment. The next steps therefore include extending the dataset and applying domain-specific data augmentation to improve training robustness. In parallel, a modular client-server architecture will be developed, in which a lightweight mobile application captures an eye photograph alongside a short symptom survey, securely transmits the data to a centralized backend for joint image-survey classification, and returns an urgency-aware recommendation to the user.

Acknowledgments and Disclosure of Funding

This work was supported by the State of Lower Austria through the project *HOLSTEIN (Holistic approach to the sustainable provision of livestock health in Lower Austria).*

References

[1] J. Alves, T. Banzato, C. Bertram, S. Boroffka, B. Broux, F. Cian, A. Csomos, R. Drees, S. Goericke-Pesch, R. Gutierrez-Quintana, M. Hafner, A. Haghofer, R. Klopfleisch, J.-G. Kresken, I. Lautenschlaeger, I. Lieske, S. Meller, K. Mie, J. Nessler, C. Ober, G. Scharf, D. Scharner, M. Schmidt, A. Sewell, A. Tipold, T. Wrzesinski, and F. K. Zeugswetter. Review of applications of deep learning in veterinary diagnostics and animal health. *Frontiers in Veterinary Science*, 12:1511522, 2025.

[2] M. Buric, S. Grozdanic, and M. Ivasic-Kos. Diagnosis of ophthalmologic diseases in canines based on images using neural networks for image segmentation. *Heliyon*, 10(19):e38287, 2024.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*, 2009.

[4] D. W. Esson and S. Calvarese, editors. *Clinical Atlas of Canine and Feline Ophthalmic Disease*. Wiley-Blackwell, Hoboken, NJ, 2022.

[5] D. Gould and G. J. McLellan, editors. *BSAVA Manual of Canine and Feline Ophthalmology*. British Small Animal Veterinary Association, Quedgeley, Gloucester, 2014.

[6] D. S. Kermany, M. Goldbaum, W. Cai, C. C. S. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan, J. Dong, M. K. Hewett, J. Dong, B. Ziyech, P. Shi, and K. Zhang. Identifying medical diagnoses and treatable diseases by image-based deep learning. *Cell*, 172(5):1122–1131, 2018.

[7] J. Y. Kim, H. E. Lee, Y. H. Choi, S. J. Lee, and J. S. Jeon. CNN-based diagnosis models for canine ulcerative keratitis. *Scientific Reports*, 9(1):14209, 2019.

[8] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[9] M. Linek, editor. *Ophthalmologie auf den Punkt gebracht: Ein Leitfaden für die Kleintierpraxis*. VBS-Vet-Verl., Babenhausen, 1 edition, 2008.

[10] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez. A survey on deep learning in medical image analysis. *Medical Image Analysis*, 42:60–88, 2017.

[11] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen. MobileNetV2: Inverted residuals and linear bottlenecks. In *Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[13] I. Walde, B. Nell, E. H. Schäffer, and R. Köstlin, editors. *Augenheilkunde: Lehrbuch und Atlas*. Schattauer GmbH, Stuttgart, 3 edition, 2008.

Measuring the Specific Gravity of Urine of Dogs Using Digital Refractometers

Martina Jezik[1] **Peter M. Roth** [1,2]

martina.jezuik@vetmeduni.ac.at peter.roth@webster.ac.at

[1]University of Veterinary Medicine, Vienna
[2]Webster Vienna Private University

Abstract

In veterinary medicine, urine specific gravity (USG) is among the most commonly used indicators of renal function, typically assessed by means of manual refractometers. However, accurate readings require a trained eye and adequate lighting conditions, limiting their use outside of clinical settings. This study investigates whether digital refractometers can serve as a reliable alternative, and in particular whether pet owners are able to use them independently for continuous at-home monitoring of their animal's USG. With a mean deviation of 0.0014 ± 0.0008 between devices, the results demonstrate that digital refractometers yield measurements comparable to those of manual devices, paving the way for broader at-home use by pet owners.

1 Introduction

Urine specific gravity (USG) is a key indicator of renal function in animals [5] and is routinely measured using manual refractometers in clinical settings. However, accurate readings require a trained eye and adequate lighting conditions, limiting their use outside of clinical environments. Furthermore, readings are susceptible to observer bias: studies have shown that results are interpreted differently by multiple veterinarians when using manual refractometers [3, 4], leading to discrepancies that can influence clinical assessment [5]. To address these limitations, digital refractometers offer a compelling alternative, providing objective and unambiguous readings [1]. In particular, we aim to enable at-home testing by pet owners, which offers two key benefits: first, by staying in a familiar environment, stress for animal patients can be reduced; second, a denser test frequency can be ensured without increasing costs.

However, to date there has been only limited interest in evaluating the accuracy of digital refractometers. For instance, [4] compared manual and digital refractometers using urine from 38 dogs and demonstrated high measurement reliability observed across measurements. Nevertheless, the level of training and experience of the operator was identified as an important factor in achieving reproducible results, which raises concerns about the suitability of manual refractometers for use by untrained laypersons. Similarly, [1] compared the degree of agreement between manual and digital refractometers using urine from 55 cats, with all measurements performed by a single operator to minimize variability. The results revealed a a small statistically difference; however, values obtained with the manual refractometer were consistently higher than those obtained with the digital device, indicating a systematic device bias. In contrast, in [5], where four different refractometers (including a digital one), were compared, the tested refractometers yielded slightly different results, however, presumed to be not clinically relevant.

These studies show, on the one hand, that there is no clear evidence that manual and digital refractometers consistently yield comparable results. On the other hand, they highlight the importance of

The Third Austrian Symposium on AI and Vision (AIROV26).

device calibration, operator training, and consistent device use. Our aim, however, is to enable pet owners to continuously monitor their animal's USG using a digital refractometer independently at home. For this purpose, we employ a modern, easy-to-use digital refractometer and compare the measurements with those of a calibrated, standardized manual refractometer. The results demonstrate that the two devices yield comparable measurements, paving the way for further studies and broader clinical application. The rest of the paper is organized as follows: First, in Sec. 2, we give an overview of the devices used. Then, in Sec. 3, we describe the experimental setup and discuss the obtained results. Finally, in Sec. 4, we summarize and conclude the paper.

2 Manual and Digital Refractometers

To measure the specific gravity of urine, we compare a calibrated manual refractometer and a modern digital refractometer (SmartRef[1]), tested by veterinarians and pet owners.

2.1 Manuel Refractometer

A manual refractometer, as shown in Figure 1, measures the refractive index of a liquid sample – that is, the degree to which light bends as it passes through the sample – and converts this into a readable scale. In veterinary use, a drop of urine is placed onto the prism and the cover is closed [2], causing the urine to spread evenly across the prism surface. The refractometer is then held up to a light source, and the USG value is read at the boundary between the light and dark fields on the internal scale, see Figure 2. The reading therefore depends on adequate lighting and the observer's ability to correctly interpret this boundary line. The thus obtained data must be entered into the patients record and can then be analyzed in an additional step. To ensure a robust measurement, the manual refractometer was cleaned by wiping it with a clean soft paper towel.

Figure 1: Manual refractometer.

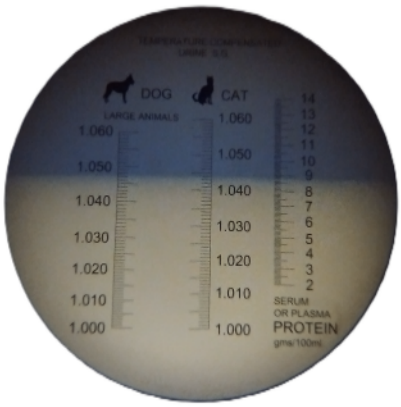


Figure 2: USG measurement value to three decimal places.

2.2 Digital Refractometer

In addition to the manual refractometer, we use the digital refractometer SmartRef (see Fig. 3). The device is accompanied by a range of apps designed for various liquids such as juices, beer, wine, or vehicle coolants (see Fig. 4). In our case, we used the app *Pet Care Meister*, which allows us to measure the specific gravity of urine and the sugar value in the urine of dogs, cats, and large animals.

[1]The company Anton Paar was neither a client nor involved in the study at all.

Figure 3: SmartRef.

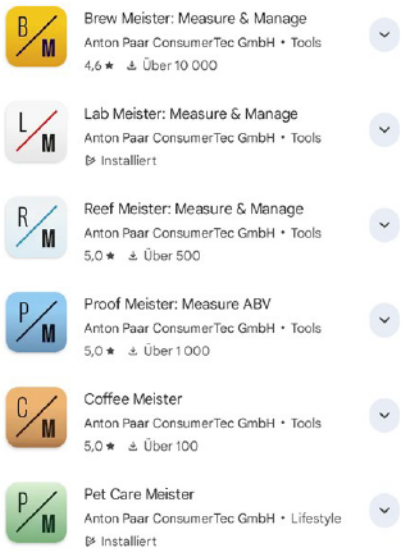


Figure 4: SmartRef apps.

Before measurements, the device must be paired with a smartphone via Bluetooth and calibrated with distilled or tap water. For cleaning between different urine samples, the sample well was sprayed with 90 % ethanol and wiped dry with a clean soft paper towel. For the actual measurement, 0.4 ml of urine is pipetted into the sample well using a disposable pipette or syringe. The measurement is then initiated via the smartphone app by pressing the Start button; the result s displayed within one second. The device supports both individual (see Fig. 5) and continuous measurements. The USG measurements can be saved and displayed as a curve, as shown in Figure 6. The x-axis represents the date on which the measurement was saved, the left y-axis shows the USG values, and the right y-axis shows the ambient temperature. The USG values are displayed in blue, and the temperature values in red.

Figure 5: Detailed information on a single measured value.

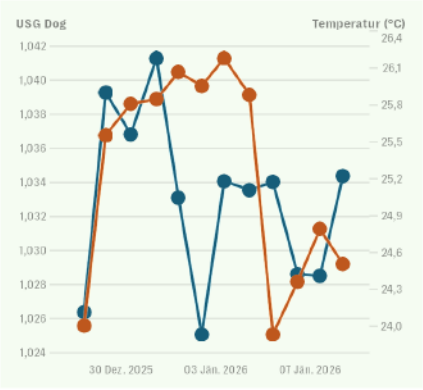


Figure 6: Plots of multiple measured values: USG (blue), ambient temperature (red).

To integrate the animals's values into the attending veterinarian's patient records, they can be exported as Comma-Separated-Values (CSV) file, and sent via email, for example, as illustrated in Table 1.

Table 1: The table shows the exported values from the SmartRef app.

Animal ID	Date	Time	Type	Temp.	Unit	Value	Univ	pH-Value
Dog-01	28/12/2025	11:06:12	—	24.0	°C	1.026	USG Dog	—
Dog-01	29/12/2025	09:08:19	—	25.6	°C	1.039	USG Dog	—
Dog-01	30/12/2025	10:38:26	—	25.8	°C	1.037	USG Dog	—
Dog-01	31/12/2025	12:43:10	—	25.8	°C	1.041	USG Dog	—
Dog-01	01/01/2026	11:03:44	—	26.1	°C	1.033	USG Dog	—
Dog-01	02/01/2026	10:42:19	—	26.0	°C	1.025	USG Dog	—
Dog-01	03/01/2026	09:26:01	—	26.2	°C	1.034	USG Dog	—
Dog-01	04/01/2026	11:10:52	—	25.9	°C	1.034	USG Dog	—
Dog-01	05/01/2026	10:55:24	—	23.9	°C	1.034	USG Dog	—
Dog-01	06/01/2026	11:59:52	—	24.4	°C	1.029	USG Dog	—
Dog-01	07/01/2026	10:43:28	—	24.8	°C	1.029	USG Dog	—
Dog-01	08/01/2026	10:02:15	—	24.5	°C	1.034	USG Dog	—

3 Evaluation

Volunteer veterinarians and pet owners collected urine from the animals and carried out the measurements. In the following, we describe the urine collection process and provide an overview of the results obtained.

3.1 Urine Collection

Spontaneous urine was used for the measurements, collected by free-catch method using two different collection aids. Commercial collection bowls, such as those by Himiyer (see Fig. 8), are available in two sizes and suitable for most dogs. However, they are less suitable for female dogs that lower their pelvis close to the ground when urinating (Fig. 7), and some female dogs refused to urinate when the bowl was present. To collect urine from these animals as well, we created a custom adapted collection device using 3D printing (see Fig. 9).

Figure 7: Female dog urinating: the pelvis is pressed low towards the ground.

Figure 8: Urine collection aid: Himiyer.

Figure 9: Urine collection aid: 3-D print.

Both collection aids feature a telescopic rod, allowing the collection bowl to be positioned under the urine stream with minimal effort. To transfer the urine to the refractometer, the required amount is drawn up from the collection bowl using a disposable pipette or syringe: 0.4 ml for the digital refractometer, or a single drop for the manual refractometer, applied directly onto the prism.

3.2 Comparison

Urine was collected from 13 dogs in order to measure the specific gravity and glucose levels using both the manual and the digital refractometer. The thus obtained results are summarized in Table 2, showing only a maximal difference of 0.002 between the SmartRef and the manual refractometer, which is within the measurement tolerance.

Table 2: Comparison of the measured values of a SmartRef and a manual refractometer.

Animal ID	SmartRef		manual
	USG	Brix / sugar	USG
Dog-01	1.026	6.7	1.026
Dog-02	1.012	2.8	1.013
Dog-03	1.025	6.3	1.026
Dog-04	1.047	12.5	1.049
Dog-05	1.032	8.3	1.034
Dog-06	1.023	5.7	1.024
Dog-07	1.016	4.0	1.018
Dog-08	1.060	16.2	1.060
Dog-09	1.020	5.1	1.022
Dog-10	1.034	8.9	1.036
Dog-11	1.050	13.2	1.051
Dog-12	1.019	4.7	1.021
Dog-13	1.045	11.9	1.047

4 Conclusion

The results demonstrate that digital refractometers yield measurements comparable to those of manual devices. At-home testing offers two key advantages: it reduces stress for animal patients by allowing measurements to be taken in a familiar environment, and it enables a denser monitoring frequency without increasing costs. These findings support the use of digital refractometers by trained pet owners and provide a basis for further studies involving additional species, including livestock, cats, and zoo animals.

Acknowledgments and Disclosure of Funding

This work was supported by the State of Lower Austria through the project *HOLSTEIN (Holistic approach to the sustainable provision of livestock health in Lower Austria)* and within the framework of the FTI Strategy Lower Austria 2027 through the project *ROBOKIZ (Robust automatic analysis of drone images in plant breeding using artificial intelligence)*.

References

[1] A. D. Bennett, G. E. McKnight, S. J. Dodkin, K. E. Simpson, A. M. Schwartz, and D. A. Gunn-Moore. Comparison of digital and optical hand-held refractometers for the measurement of feline urine specific gravity. *Journal of Feline Medicine and Surgery*, 13(2):152–154, Feb. 2011.

[2] E. Communications. LaborSkills – Leitfaden Labordiagnostik für Hund und Katze.

[3] M. Mösch, S. Reese, K. Weber, K. Hartmann, and R. Dorsch. Influence of preanalytic and analytic variables in canine and feline urine specific gravity measurement by refractometer. *Journal of Veterinary Diagnostic Investigation*, 32(1):36–43, 2020. Epub 2019 Dec 26.

[4] J. C. Rowe, J. A. Hokamp, J. N. Braatz, J. R. Freitag-Engstrom, N. L. Stephens, D. J. Chew, C. Langston, and A. J. Rudinsky. Interobserver reliability of canine urine specific gravity assessed by analog or digital refractometers. *Journal of Veterinary Diagnostic Investigation : Official Publication of the American Association of Veterinary Laboratory Diagnosticians, Inc*, 33(3):611–614, May 2021.

[5] A. Rudinsky, C. Cortright, S. Purcell, A. Cordner, L. Lord, M. Wellman, S. DiBartola, and D. Chew. Variability of first morning urine specific gravity in 103 healthy dogs. *Journal of Veterinary Internal Medicine*, 33(5):2133–2137, 2019.

Interactive VetMap of Austria

Valentina Dolin[1] **Gudrun Kinz**[1,2] **Martina Jezik**[1]
Mark A. M. Kramer[1] **Peter M. Roth** [1,3]

valentina.dolin@vetmeduni.ac.at gudrun.kinz@boku.ac.at
martina.jezuik@vetmeduni.ac.at mark.kramer@vetmeduni.ac.at
peter.roth@webster.ac.at

[1]University of Veterinary Medicine, Vienna
[2]BOKU University
[3]Webster Vienna Private University

Abstract

Comprehensive veterinary care is of central importance for animal health in agriculture. However, veterinary services are often unavailable at night and on weekends. To better represent availability and deploy existing resources more efficiently, we collected publicly available data on veterinary practices in Austria and stored them in a continuously updated database. To make this information accessible, we integrated it into a web app based on OpenStreetMap, providing relevant information for both animal owners and veterinarians. The app allows users to visualize the structure of veterinary services across Austria and to find the nearest available practices. Route planning is based on real street distances and travel times, computed using the Open Source Routing Machine (OSRM), ensuring that geographical constraints such as mountain passes or river crossings are properly accounted for. The system is designed to be extensible, with future versions planned to incorporate real-time availability updates, veterinary specializations, and seasonal road conditions.

1 Introduction and Motivation

Reliable veterinary care across both livestock and companion animal medicine is an important socio-political challenge. Structural changes in agriculture and the veterinary profession are putting increasing pressure on the existing system. While coverage during regular practice hours is still guaranteed, the system reaches its limits at night and on weekends. It is therefore a primary goal to deploy available resources as efficiently as possible, taking into account information such as vacations, sick leave, or other absences. Furthermore, unlike human medicine, veterinary medicine lacks fixed structures such as round-the-clock emergency stations. Veterinary practices are staffed only during regular office hours, while on-call duty is arranged individually and may vary from day to day.

Existing services, such as the veterinarian search provided by the Austrian Veterinary Chamber[1] or the insurer Anicura[2], provide addresses but no further information. The present work therefore aims to provide a comprehensive database of veterinary services across Austria and to make it publicly available. The work was partly inspired by the *pharmacy finder app* of the Austrian Apothekerkammer[3]. The information was collected from publicly available sources and is visualized using OpenStreetMap. The result is a continuously updated database, publicly accessible to both

[1]https://www.tieraerztekammer.at/oeffentlicher-bereich/kurz-menue/tierarzt-suche
[2]https://www.anicura.at/
[3]https://www.apothekerkammer.at/

The Third Austrian Symposium on AI and Vision (AIROV26).

animal owners and veterinarians at https://vetfind.at. The main contributions of this paper are the establishment of a database of veterinary practices in Austria, the computation of their GPS coordinates, and the integration of this information into an interactive platform, making it easily accessible to animal owners and veterinarians alike.

2 Map and Website

In total information about 1,439 practices clinics have been collected (no guarantee for correctness and completement), where for each practice the following information is stored in a database:

- Name of practice
- Website
- E-mail
- Phone
- Address
- Federal state
- GPS coordinates

The GPS coordinates (longitude and latitude) were retrieved via the *Nominatim geocoding API*[4]. In addition, we captured – if available – relevant information such as specialization (small animal, livestock, horses, etc.) and opening hours.

To this end, we present this information as an interactive map based on OpenStreetMap [8]. Users can search for the nearest available practices by entering an exact address, using their automatically estimated GPS position, or selecting a location directly on the map. Results are displayed either as the ten closest practices or as all practices within a predefined radius. Hovering over a result reveals more detailed information about the practice; additionally, the route to the selected destination can be computed, and a textual route description is provided. An overview of the web-app, which can be called via vetfind.at is shown in Fig. 1.

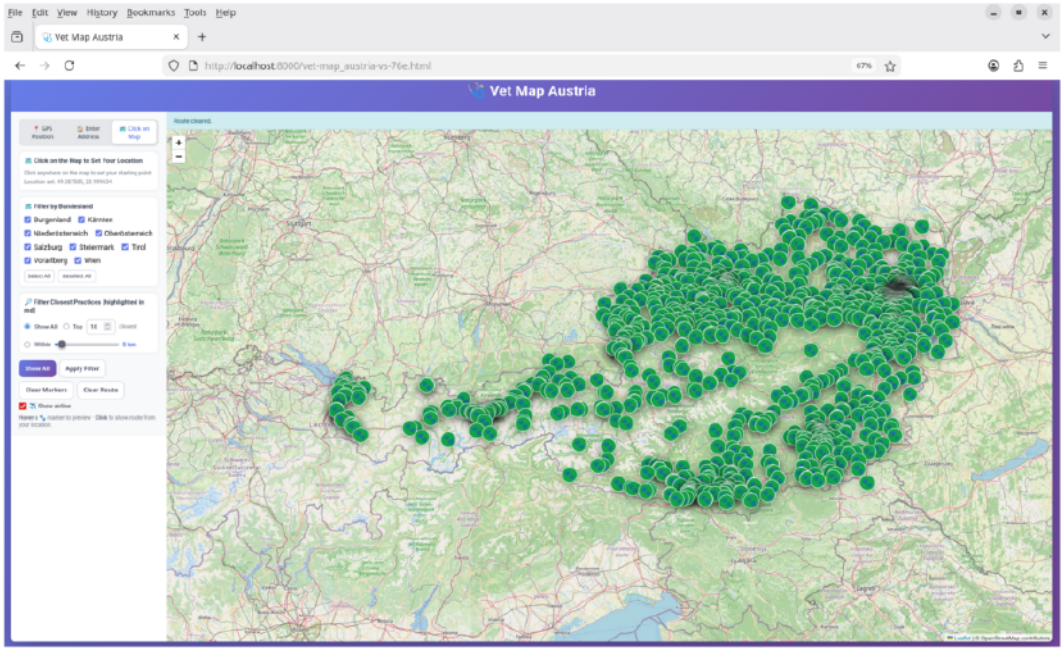


Figure 1: VetMap Austria: Interactive Map showing all veterinary practices in Austria.

[4] https://operations.osmfoundation.org/policies/nominatim/ (accessed March 1, 2026)

3 Applications

Based on the availale information, we identified different use cases for practical applications, which are discussed in the following:

3.1 Analyze Veterinary Service Structure

The visualization of practice locations provides insight into the structure of veterinary services across Austria. In the current version, the distribution of practices can be directly visualized, revealing the influence of geographical constraints. This is illustrated for Styria in Fig. 2a and Tyrol in Fig. 2b. In Styria, a dense network of practices is visible in the Ennstal, the Mur-Mürz-Furche corridor, and the southern part of the region. Similarly, in Tyrol most practices are concentrated in the Inntal, with only a small number available in the surrounding smaller valleys. Taking additionally into account the size of practices and the age of their employees, more detailed predictions of future service distribution would be possible; however, these data are either not yet available or have not yet been incorporated into the model.

(a) Veterinary practices in Styria.

(b) Veterinary practices in Tyrol.

Figure 2: Illustration of different route visualisations.

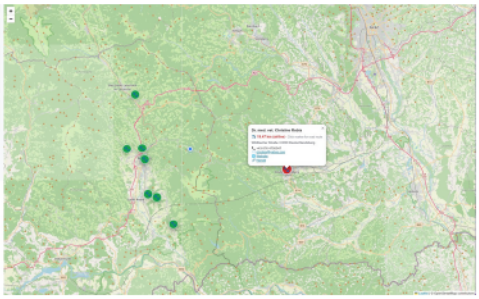
3.2 Getting Close Practices

The user's current position can be specified in three ways: entering an address into a search field, using the automatically determined GPS position, or selecting a location directly on the map. The web app is designed to work on both desktop computers and mobile devices, and can access the device's GPS sensor directly where available. Once the user's position has been determined, the nearest practices are displayed on the map, either within a specified range (in km) or as the top n closest results.

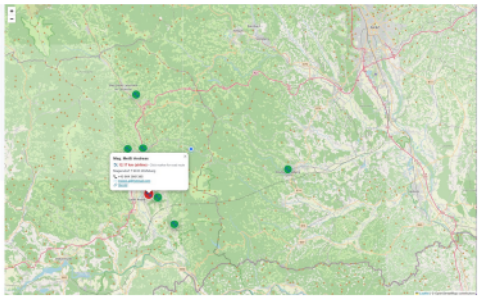
The locations of animal owners and those of available practices are modelled in a graph-based framework, where the street distance or travel time serves as the optimality criterion, computed using OSRM (Open Source Routing Machine; [6]). The models used are, however, flexible enough to incorporate further criteria in the future, such as seasonal parameters, real-time availability updates, or required veterinary expertise. Additionally, the particular challenges in Austria – such as long travel distances due to natural obstacles like mountains or rivers, high-altitude farms with alpine pasture farming, or seasonal restrictions due to road closures or lack of winter road maintenance – must be taken into account. Studies such as [3] that additionally consider the relevance of geo-information or the nature and severity of the emergency at hand remain rare. We therefore drew on methods for optimal route planning in general [4, 6, 9], as well as on problems from human medicine concerning the optimal routing of emergency vehicles [2, 7, 10].

3.3 Provide Detailed Information

Based on this information, we identified two additional applications. The first is to provide detailed information about nearby practices stored in the database. These details can be obtained by hovering over a practice marker, and include the name, address, phone number, website, and e-mail address, each presented as an interactive link for direct access. Future versions will additionally display specialization information (e.g., livestock, small animal, horses). This is illustrated in Fig. 3.



(a) Illustrative example 1.

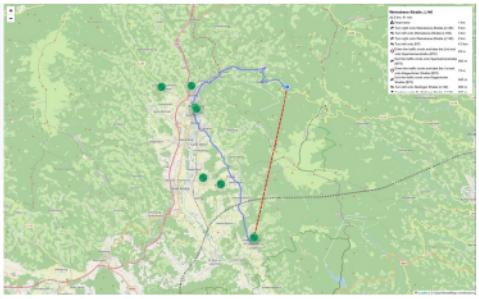


(b) Illustrative example 2.

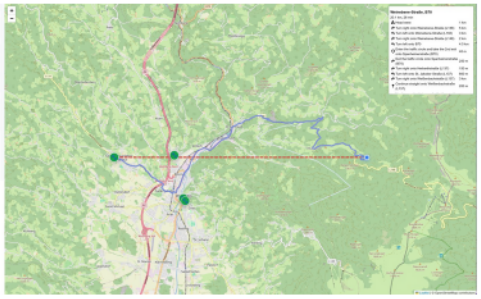
Figure 3: Illustration of detailed practice information.

3.4 Get Direction

The second use case we identified is computing the direction (road distance) from the current position to a veterinary practice. By clicking the marker of the practice, the direction is computed and visualized on the map. In addition, a textual description of the direction is provided. In addition, we provide the possibility ot compare the road distance with the air distance, to make it apparent that the geographic distances might be highly different from the air distance. Two examples for this functionality are illustrated in Fig. 4.



(a) Illustrative example 1.



(b) Illustrative example 2.

Figure 4: Illustration of found directios description of the pathway.

4 Conclusion and Future Work

Further work can be divided into three key aspects: (a) So far, optimal routes have been generated as in [6] using suitable speed-up techniques for Dijkstra's algorithm [5]. Further ideas from graph theory [1, 4] will be incorporated in the future. (b) Additionally, the rudimentary clustering approach will be supplemented with more suitable methods [11, 12, 13]. (c) So far, only geographic information has been used to compute the edge weights. In addition, topographic and traffic-related information, vacation periods, sick leave, or other availabilities of veterinarians can also be taken into account. Beyond these technical extensions, the database itself will be continuously expanded and refined: practices will be enriched with additional attributes such as veterinary specialization, opening hours, and real-time on-call availability, ensuring that the information remains current and practically useful. Furthermore, the web app will be extended to support mobile devices more fully, including push notifications for animal owners in urgent situations. Finally, a systematic evaluation of the service coverage across Austrian federal states is planned, with the goal of identifying underserved regions and informing policy decisions on the deployment of veterinary resources.

Acknowledgments and Disclosure of Funding

This work was supported by the State of Lower Austria through the project *HOLSTEIN (Holistic approach to the sustainable provision of livestock health in Lower Austria)* and within the framework of the FTI Strategy Lower Austria 2027 through the project *ROBOKIZ (Robust automatic analysis of drone images in plant breeding using artificial intelligence)*.

References

[1] D. L. Applegate, R. E. Bixby, V. Chvátal, and W. J. Cook. *The Traveling Salesman Problem: A Computational Study*. Princeton University Press, 2007.

[2] L. Brotcorne, G. Laporte, and F. Semet. Ambulance location and relocation models. *European Journal of Operational Research*, 147(3):451–463, 2003.

[3] J. M. de Souza Muniz, E. C. S. Lima, and L. de Castro Mesquita. VETFINDER: A platform connecting pet owners and veterinarians. *Revista Ibero-Americana De Humanidades, Ciências E Educação*, 10(11):7128–7145, 2024.

[4] D. Delling, P. Sanders, D. Schultes, and D. Wagner. Engineering route planning algorithms. In *Algorithmics of Large and Complex Networks: Design, Analysis, and Simulation*, pages 117–139. 2009.

[5] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1:269–271, 1959.

[6] D. Luxen and C. Vetter. Real-time routing with OpenStreetMap data. In *ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, pages 513–516, 2011.

[7] J. Nelas and J. Dias. Optimal emergency vehicles location: An approach considering the hierarchy and substitutability of resources. *European Journal of Operational Research*, 287(2):583–599, 2020.

[8] OpenStreetMap contributors. Planet dump retrieved from https://planet.osm.org . https://www.openstreetmap.org, 2017.

[9] J. Son, Z. Zhao, F. Berto, C. Hua, C. Kwon, and J. Park. Neural combinatorial optimization for real-world routing. *arXiv:2503.16159*, 2025.

[10] J. Tassone and S. Choudhury. A comprehensive survey on the ambulance routing and location problems. *arXiv:2001.05288*, 2020.

[11] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17:395–416, 2007.

[12] H. Zha, C. Ding, and M. Gu. Bipartite graph partitioning and data clustering. *arXiv:cs/0108018v1*, 2001.

[13] P. Zhang, J. Wang, X. Li, M. Li, Z. Di, and Y. Fan. Clustering coefficient and community structure of bipartite networks. *Physica A: Statistical Mechanics and its Applications*, 387(27):6869–6875, 2008.

Index of authors