# 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).

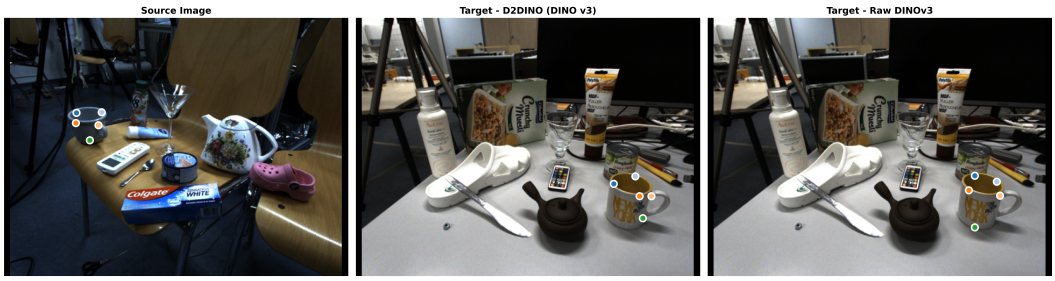



**Figure 1:** Qualitative comparison of point matching for unseen *cup* instances from the HouseCat6D dataset between D$^2$DINO trained with a DINOv3 encoder and 'raw' DINOv3 features. In this example, D$^2$DINO is trained on objects of the *cup* category and then used to match a small set of query pixels (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

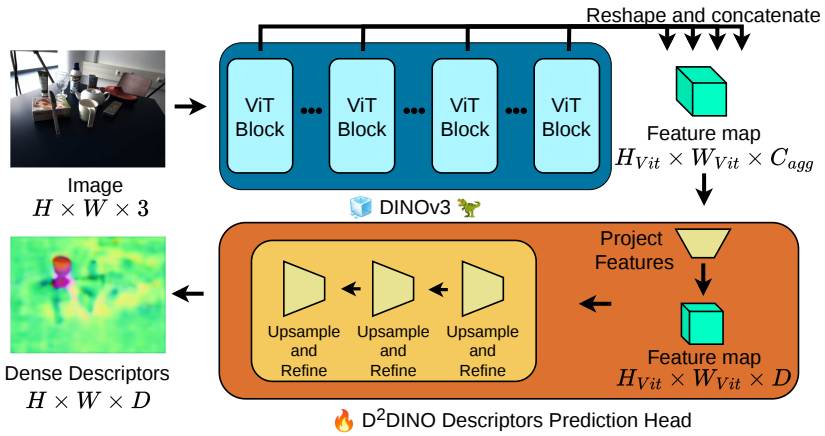



**Figure 2:** Overview of the D$^2$DINO descriptor prediction head. Given an input RGB image of size $H \times W \times 3$, we first extract 4 sets of intermediate token embeddings from a frozen DINOv3 transformer encoder, obtaining a sequence of features that are reshaped and concatenated into a 2D 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 D$^2$DINO 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 $\mathbf{x} \in \mathbb{R}^3$ and search in the target image for pixels for which the NOCS coordinate is equal (up to a small tolerance) to $\mathbf{x}$. 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 $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, D$^2$DINO 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

$$\{\mathbf{F}^{(\ell)} \in \mathbb{R}^{H_{\text{Vit}} \times W_{\text{Vit}} \times C} \mid \ell = 1, \ldots, L\}, \tag{1}$$

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 $\mathbf{F}^{(\ell)}$ are concatenated along the channel dimension:

$$\mathbf{F}_{\text{agg}} = \text{Concat}\big(\{\mathbf{F}^{(\ell)}\}_{\ell=1}^{L}\big) \in \mathbb{R}^{H_{\text{Vit}} \times W_{\text{Vit}} \times C_{\text{agg}}}, \quad C_{\text{agg}} = L \cdot C. \tag{2}$$

We apply batch normalization and a $1 \times 1$ convolution to project this aggregated tensor into a descriptor space of dimension $D$:

$$\mathbf{D}_{\text{low}} = \text{Conv}_{1\times1}\big(\text{BN}(\mathbf{F}_{\text{agg}})\big) \in \mathbb{R}^{H_{\text{Vit}} \times W_{\text{Vit}} \times D}. \tag{3}$$

This yields a low-resolution dense descriptor map aligned with the ViT patch grid.

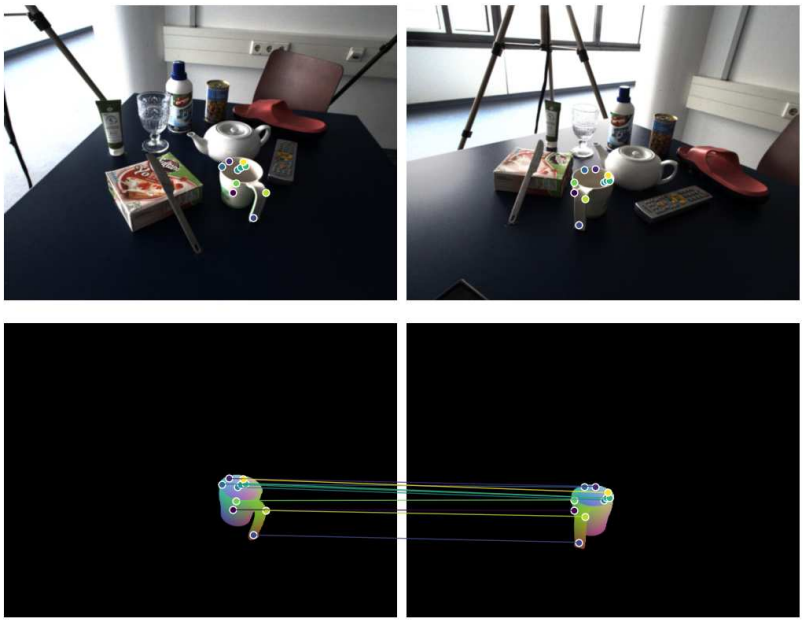


**Figure 3:** Example of point correspondences sampled using NOCS maps. On the left the source image, on the right the target image, each with its respective NOCS map below it. The NOCS map uniquely identifies the position of a point in the image on the 3D model of the object in a way that does not depend 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 $3 \times 3$ 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 $\mathbf{G}^{(0)} = \mathbf{D}_{\text{low}}$, $\sigma$ denoting GELU. The number of channels $D$ is kept constant. A $3 \times 3$ 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 $\ell_2$-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 $\tau$, 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 $h_{\text{bbox}}$ and $w_{\text{bbox}}$ 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 $2 \times 10^{-3}$, 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.

| | | Object categories | | | |
|---|---|---|---|---|---|
| Method | Loss | 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 |
| $D^2$DINO | NT-Xent w/ soft hard negs. | **52.31** | **60.34** | **70.26** | **78.26** |
| $D^2$DINO | DON loss w/ soft hard negs. | 44.88 | 57.05 | 59.88 | 71.27 |
| $D^2$DINO | 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. $D^2$DINO 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 $D^2$DINO prediction head effectively captures this intra-category variance, generalizing successfully to novel test instances.

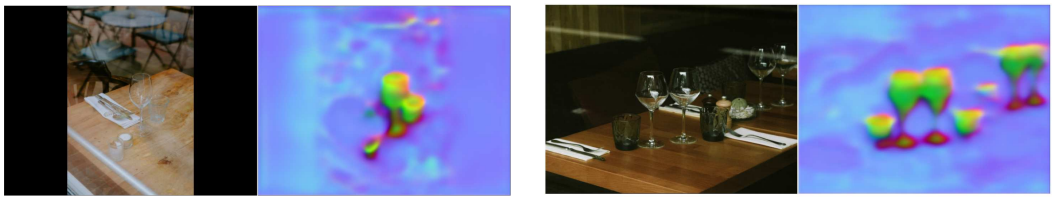
The loss ablation reveals complementary roles for each component. The DON loss [Florence et al., 2018]—a pairwise margin loss with fixed distance thresholds—struggles to capture nuanced intra-object variations, whereas our softmax-based NT-Xent provides a more flexible similarity scale. Crucially, omitting 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 $D^2$DINO learns pose-robust part representations transferable across instances.

Qualitative PCA visualization in Section 5 further illustrates this for glasses: $D^2$DINO descriptors form tight, semantically coherent clusters for corresponding parts across images with multiple novel instances (different poses, partial occlusions).

We measured the computation of $D^2$DINO 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 $D^2$DINO'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.