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`

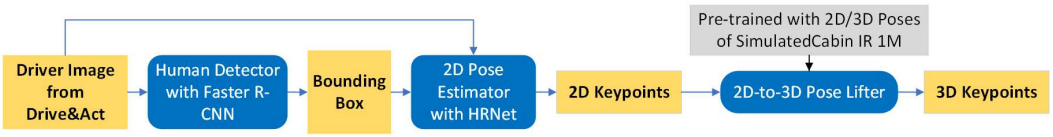


Figure 1: The setup of the 3D Driver Pose Estimation pipeline. The blue rounded boxes denote estimation models, while the yellow boxes represent data.

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.

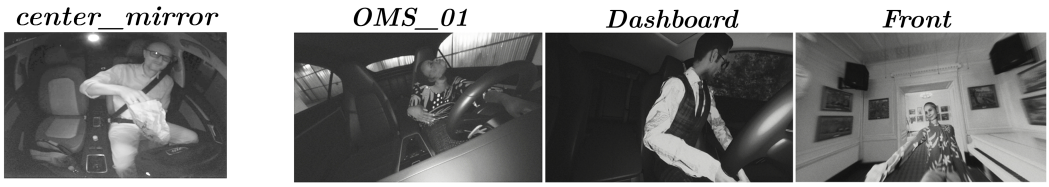
SimulatedCabin IR 1M (Figure 2b) is a simulated data set that contains one million driver poses, extending Sagmeister et al. [24]. It provides simulated NIR images of 3D human models seated in a vehicle under diverse movements, interiors, lighting conditions, and 10 camera viewpoints. Poses are randomly 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.

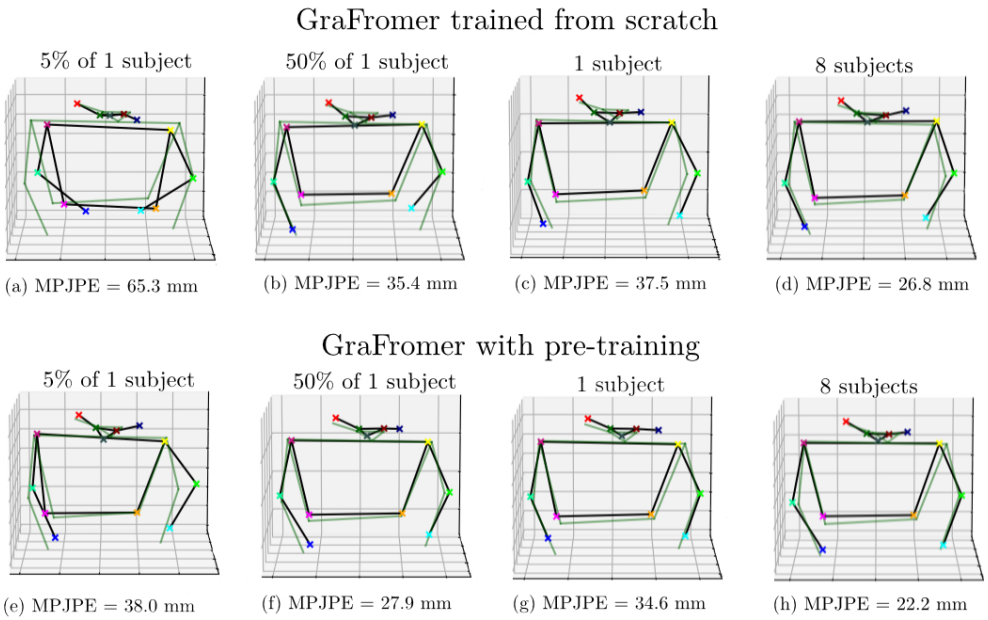


Figure 3: Example 3D pose estimations from GraFormer trained with different amounts of Drive&Act training data. Green indicates ground truth; black indicates the estimate. The input frame is from Subject 12.

	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.