Transfer Learning for Driver Pose Estimation from Synthetic Data

Daniel Sagmeister[1], Dominik Schörkhuber[1], Matej Nezveda[2],
Fabian Stiedl[2], Maria Schimkowitsch[2], Margrit Gelautz[1]
{daniel.sagmeister, dominik.schoerkhuber, margrit.gelautz}@tuwien.ac.at,
{mne, fst, msc}@emotion3d.ai

***Abstract*—The training of computer vision models for human pose estimation requires large amounts of data. Since labelling image data with pose keypoints is very time consuming and costly, we aim to alleviate this requirement by using synthetic data during pre-training and thus relax the need for large amounts of real data samples during fine-tuning. To this end, we investigate the impact of synthetic data on the performance of a 2D keypoint detection model in the context of driver body pose estimation. We present our approach for synthetic data generation to automatically provide large amounts of in-cabin views as training data. The utilization of the generated synthetic data is evaluated in different learning schemes. We achieve a notable performance gain of +30.5% by pre-training with our in-cabin synthetic data when only 1% of real training data from the DriPE dataset is available. The proposed approach also outperforms pre-training with PeopleSansPeople by +8.3% when the reduced DriPE dataset is used for fine-tuning.**

***Index Terms*—transfer learning, driver pose estimation, synthetic data**

I. Introduction

Human pose estimation has seen great progress in the last decade, partly owed to the large labeled datasets that have been published [1], [2]. These datasets require high effort to create, mainly because labeling the data is very time and cost intensive [3]. Moreover, the inclusion of real-life data is often limited for security, ethical reasons, or privacy regulations [4], [5]. A remedy is synthetic-generated data on which machine learning algorithms can be trained and validated. The impact of transfer learning via synthetic data on human pose estimation has been studied in work such as [6]–[11], essentially using domain-generalized synthetic data, which was generated with the idea of enabling a broad range of applications for transfer learning in human centered tasks.

In this paper, we address the impact of the domain in which the synthetic data is generated on transfer learning. We compare the effect of domain-generalized data versus domain-specific data on the performance of human pose estimation models. In domain-specific environments, there is often only a limited amount of real-world data available to train and validate models, which suggests improving models using transfer learning with synthetic data as a viable alternative. We investigate the impact of domain-specific synthetic data for human pose estimation in the special context of driver monitoring.

Despite the great progress in the field of human pose estimation in general and in applications such as monitoring pedestrians in traffic scenes, the related task of driver pose estimation in car interiors has been less addressed in the literature so far. A major reason for this limited coverage is the lack of suitable datasets for training and validating these models. This shortage of data and studies motivates our work on human pose estimation in vehicle interiors with a focus on the impact of synthetic-generated data for real-world driver pose estimation. More precisely, we investigate how driver pose estimation with various-sized real-world datasets can be improved by pre-training with synthetic datasets generated in different domains. We train and validate the popular HRNet approach for pose estimation [12] on the synthetic dataset PeopleSansPeople [7] and fine-tune on the real-world dataset DriPE [13]. Furthermore we introduce a methodology for synthetic data generation for keypoint detection in a simulated vehicle interior and the so created dataset SimulatedCabin. The main contributions of our work are:

- We present our approach for synthetic data generation called SimulatedCabin, which is used to render human drivers from multiple views in a virtual car in large quantity and great variety for the (pre-)training of human pose estimation models.
- We train and evaluate HRNet for human pose estimation by applying different learning schemes to utilize synthetic data efficiently and to reduce the synthetic-to-real domain gap. By utilizing our new synthetic data for pre-training, we achieve a significant gain in accuracy of +30.5% on DriPE when only 1% of the (real) training data is available.
- We conduct a performance comparison between several models trained on the general-purpose synthetic dataset PeopleSansPeople and our specialized synthetic dataset SimulatedCabin. We find that pre-training on SimulatedCabin outperforms the models pre-trained on PeopleSansPeople by +8.3% when real-world data for fine-tuning is limited.

The remainder of the paper is organized as follows. Firstly, we discuss related work in transfer learning and knowledge transfer from other domains including methods for driver pose

This work was supported by the project SyntheticCabin (Project Number 884336), which is funded through the Austrian Research Promotion Agency (FFG) on behalf of the Austrian Ministry of Climate Action (BMK) via its Mobility of the Future funding program.

estimation and available datasets in II. We then elaborate our approach for synthetic data generation based on a simulation environment to generate data close to our target domain in III. In IV, we continue to explain our method for training a human pose estimation model with the aid of synthetic data and afterwards show our experiments and results in V. To reproduce the evaluations in this paper, we make our human pose estimation models publicly available[1].

II. Related Work

After motivating the need for transfer learning, we start by examining related literature using synthetic and/or real data. We then review relevant work on driver pose estimation and discuss available datasets in this field.

A. Transfer Learning

Transfer learning in computer vision or machine learning aims to improve the performance by transferring knowledge between different but related domains [14]. With this approach, the dependence on large datasets for the target domain can be reduced by adding data of a related domain. To further combat the problem of data scarcity, a recently emerging approach is to transfer knowledge from synthetic-generated data to real data domains. Transfer learning with synthetic data was applied previously for person detection and tracking [6], [15] and human pose estimation [7], [8].

In [8], the authors generate a dataset with purely synthetic humans and a real dataset augmented with synthetic humans in a general context. While both data generation approaches performed equally well, synthetic data was shown to improve detection results. The goal in [15] is to utilize synthetic data for human pose estimation to reduce the impact of occlusion by larger variety in data. The authors of [7] also present a data generation approach for synthetic humans, but different to our approach they put a focus on variety rather than realistic simulation. In contrast to [7], [8], [15], we concentrate in our work on applying transfer learning with synthetic data to a specific application context.

In works like [9], transfer learning is used in human pose estimation for segmentation tasks, and the authors of [10] used synthetic humans for action recognition of daily activities. In [7], the authors investigate the impact of transfer learning with synthetic data for human pose estimation when only a limited amount of real data is available for training. They pre-train on their developed synthetic data PeopleSansPeople and fine-tune on the real data set COCO [16]. The focus of the work is on general applications of human pose estimation, independent of specific tasks or environments. We include their data generation approach for comparison and evaluation in our work.

Some recent works have published use cases of transfer learning for driver pose estimation in car interiors. The authors of [17] attempt to transfer knowledge from a human pose estimator trained on real data to synthetic data to investigate the influence of position and type of cameras to capture image data for driver pose estimation in (virtual) car interiors. While in that work a knowledge transfer from real to synthetic data is intended, we analyze the transfer from synthetic to real data.

In our work, we investigate how driver pose estimation with various-sized real-world datasets can be improved by transfer learning with synthetic-generated datasets. We are not aware of any study that examines the influence of the synthetic domain for transfer learning in the context of driver pose estimation in the car interior. We aim to fill this gap by contrasting the influence of a domain-specific dataset over a general-purpose human pose estimation dataset.

B. Driver Pose Estimation

The task of driver pose estimation or human pose estimation for car interiors has started to raise interest in recent years. The pose of the driver is relevant in various contexts, like head pose estimation or gaze recognition [18]–[20], with applications such as calibration [21] or drowsiness detection [22], [23], and action recognition of the driver [24], [25].

Our work specifically looks at driver pose estimation in car interiors. This can be performed based on 2D images [13], [25], but point clouds or depth images can also be utilized [26], [27]. The authors of [28] and [26] present fast and compact algorithms for driver pose estimation that are especially suited for embedded systems, addressing constraints imposed by the car such as the limited computational and storage capabilities. In [29], human pose estimation is performed on synthetic persons (adults, children and babies) in the backseat of a virtual car interior; contrarily, our focus is on pose estimation of the driver. The authors of [13] investigate driver pose estimation in a real car interior, without the use of synthetic data. In our work we also deal with 2D keypoint detection of the driver. As opposed to [13], however, we use synthetic images of the driver taken from 3 different perspectives to train the models.

C. Datasets for Driver Pose Estimation

While a wide range of general-purpose data sets exists for human pose estimation (e.g., [16]), data sets for domain-specific tasks are often limited. This is especially true for driver pose estimation. One reason for the scarcity of ground truth data in this domain is the effort to record human poses in car interiors that need to be specially set up for data acquisition. In addition to the labor intensive manual annotation of human keypoints, regulations on data privacy protection also need to be taken into account.

Since research on driver pose estimation is currently gaining attention, several relevant datasets have been published only recently. Some datasets like [30] were not recorded in a real car, but in a re-created environment. The dataset presented in [29] is a synthetic dataset which shows people in the backseat of cars. A notable feature of this dataset is the inclusion of children and infants, along with their human pose keypoints. DriPE [13] is a dataset for estimating human posture in car interiors under real driving conditions. Images were acquired

[1] https://mega.nz/folder/ob0HCJrB\#ye7Pg3btLqiL3icc8gdpyg

using an RGB camera positioned above the passenger door, oriented towards the driver. The dataset encompasses 10.000 images featuring 19 drivers under diverse conditions, with driver poses annotated utilizing the 17 COCO 2D keypoints.

Similar to [29], we use a simulated car interior to generate our SimulatedCabin dataset. However, we concentrate on the pose of the driver, which is not included in [29] due to its focus on the backseat. In contrast to the dataset DriPE [13], we simulate data from different camera perspectives and can also change settings for the environment like backgrounds, texture of cloths, and lighting conditions. With our approach for synthetic data generation for car interiors and drivers, we fill the gap of synthetic datasets for driver pose estimation.

III. Synthetic Data Generation

The SimulatedCabin generator is created based on a Unity version 2020.3.17f1 simulation environment. It features three different camera viewpoints with 50.000 RGB images with ground truth information in JSON format for each camera and puts a special focus on in-cabin monitoring. In this section, we provide information on the data generation approach including 3D assets, camera characteristics, scene background, lighting and ground truth. Statistical distributions of the randomized parameters are outlined in Table I.

3D Assets We use 16 different textured 3D human characters from RenderPeople [31] and 3 different textured 3D vehicles from Hum3D [32] (see Figure 1). All characters are varied in age, gender and ethnicity. To increase the variation in texture, a similar approach to [33] is applied to the characters. In particular, we keep the original normal maps, which preserves the structures of the original textures such as wrinkles in jeans, while the color textures are randomly varied to increase the diversity of the base characters appearance. Figure 2 shows a selection of these texture variations used, as well as an example result when applied to the characters. To simulate typical movements in the vehicle interior, such as grabbing the steering wheel, procedural animations based on an inverse kinematic workflow are applied. This involves first defining the final positions of the external limbs such as the hands and, based on this, we compute the angles and positions of the remaining body parts such as the shoulders and elbows. Figure 3 shows examples of animations using this procedural animation workflow.

Camera Characteristics We use three different camera positions, namely the A-pillar at the driver's side, the A-pillar at the co-driver's side, and the rear-mirror. The same scene is rendered from all three camera positions with a resolution of 960x540 pixels, in RGB modality and without any distortion. We further vary the camera position and orientation and add bloom, Gaussian blur, and chromatic aberration to the images during every render pass as a post-processing step.

Scene Background and Lighting We choose a random background from 10 different samples selected from Poly-Haven [34] and rotate them randomly along the horizon's axis. Those samples feature different lighting situations, including indoor and outdoor environments. For illumination, we use a directional light and alter the rotation as well as the lux intensity.

Ground Truth For every image we provide a 2D bounding box surrounding the person and the 2D key point annotations in both COCO and MPII dataset format. Moreover, we provide 3D information for every key point in the camera coordinate system.

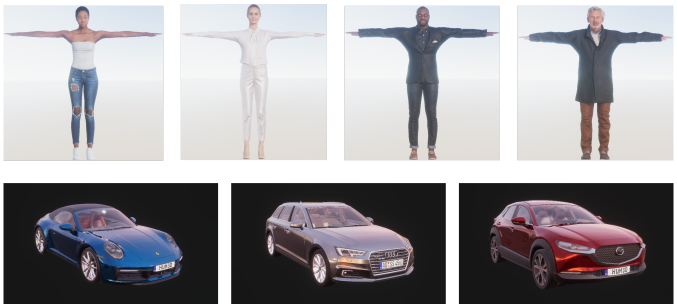

Fig. 1: Illustration of 4 out of the 16 3D human characters (first row) and the 3 different 3D vehicles (second row).

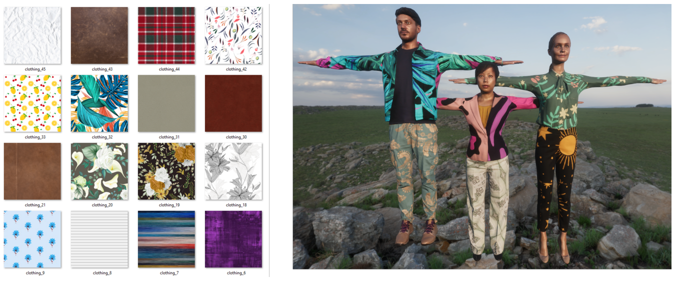

Fig. 2: Illustration of texture synthesis. Left: Various texture samples. Right: Result of texture synthesis applied on three of the characters.

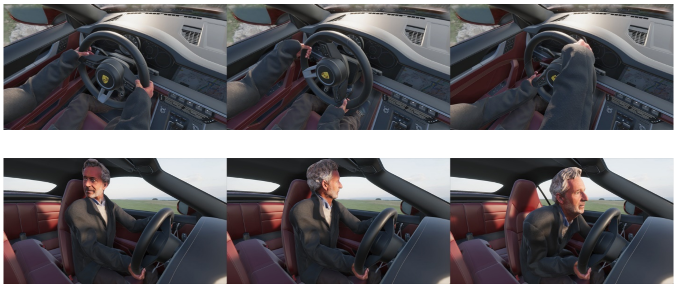

Fig. 3: Illustrations of possible applications of the procedural animation workflow based on inverse kinematics. First row: Grabbing different objects in the vehicle interior such as the steering wheel. Second row: Variation of poses in the vehicle interior.

IV. Driver Pose Estimation

In this section we describe the employed methods to train a human pose estimation model. Our strategy is motivated by the limited availability of data for human pose estimation in

Category	Parameters	Domain
Human textures	Probability to keep default texture Texture	10% $[1, 45] \in \mathbb{N}$ (uniform)
Camera position	Translation (cm) Rotation (degree)	$[0, 1] \in \mathbb{R}$ (uniform) $[0, 5] \in \mathbb{R}$ (uniform)
Bloom	Probability of application	10%
Blur	Probability of application	10%
Chromatic aberration	Probability of application	30%
Background	Background	$[1, 10] \in \mathbb{N}$ (uniform)
Light	Pitch (degree) Yaw (degree) Lux	$[100, 190] \in \mathbb{R}$ (uniform) $[30, 50] \in \mathbb{R}$ (uniform) $[2000, 12000] \in \mathbb{R}$ (uniform)

TABLE I: Domain randomization parameters of SimulatedCabin dataset. To generate a wide variety of data samples, we randomize textures, camera views, lighting and post processing effects.

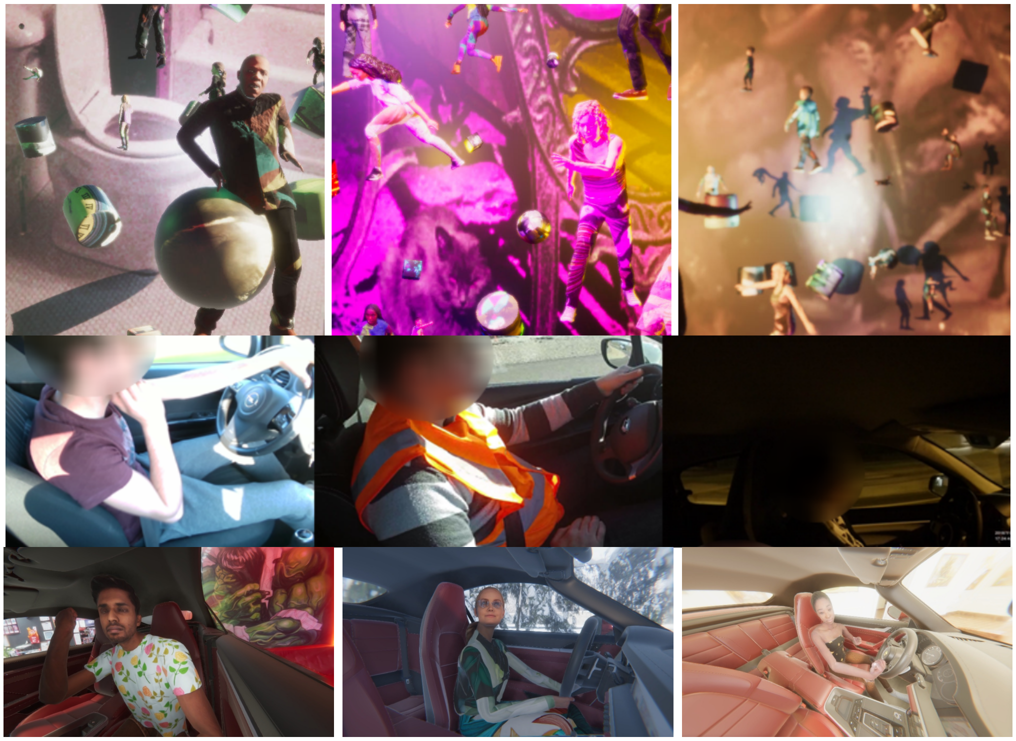

Fig. 4: Sample images from the datasets used to visualize the domain differences between the available synthetic dataset PeopleSansPeople [7] (first row) and the newly generated synthetic dataset SimulatedCabin (last row). Sample images from the real car interior dataset DriPE [13] (middle row) are shown for comparison.

the field of driver monitoring. In particular, we aim to improve human keypoint detection for the DriPE dataset, where training data is not abundant. As a solution to this problem, we resort to synthetically generated data. However, successful utilization of synthetic data for model training is not trivial, because the domain gap typically prevents generalization from the synthetic to the real domain. By generating data close to our domain of application, we aim to reduce this gap preemptively and hence improve performance on the target data distribution. In addition to the SimulatedCabin data, whose generation was described in III, we make use of the PeopleSansPeople [7] dataset for comparison. PeopleSansPeople also provides synthetic data for human pose estimation, but rather focuses on domain randomization to generate data with large variety. In total, it comprises 28 scanned 3D human models and over 1600 different background textures. The authors of PeopleSansPeople provide the simulation environment, which we use to generate a dataset comparable in size to SimulatedCabin for our experiments in Section V.

We design our experiments in order to investigate how knowledge transfer with synthetically generated data can improve the performance of 2D keypoint detection on real data from DriPE. We expect that domain specific datasets, such as our SimulatedCabin dataset, are better suited than general-purpose datasets, such as the PeopleSansPeople dataset, to train detectors for application in the car interior. To prove this hypothesis, we train models on these two synthetic datasets and then fine-tune them on real-world DriPE datasets of various sizes.

For keypoint detection we use the HRNet architecture for human pose estimation [12]. The idea of HRNet is to maintain high-resolution features by connecting high-to-low resolution convolutions in parallel. This architecture features a top-down approach to detect human keypoints. As such, the algorithm detects exactly one pose per prediction. To put the focus onto the person to detect (i.e., driver), the input image is clipped to the region of interest.

Training We rely on the deep high-resolution representation learning for human pose estimation with the HRNet architecture [12] for our benchmark experiments. As input for this approach, we use ground truth bounding boxes. Since we are only interested in the driver, one bounding box is present in each input image. We train our models by initializing the backbone with weights from training with ImageNet [35]. We use an exponential learning rate scheduler for all our models. The learning rate is reduced every epoch by the factor 0.97 with an initial learning rate of γ=1e-3. We perform the learning rate reduction during the whole training process. The remaining training parameters correspond to the standard settings in [12]. We employ an input image size of 192x256px at a batch size of 32. We use an NVIDIA GeForce GTX 1080 GPU to train our models. Parameters for data augmentation are kept the same as in the original HRNet training procedure.

Domain Differences To give an intuition of the differences in domain, we illustrate examples from the used datasets in Figure 4. The first row contains sample images from PeopleSansPeople, showing randomly varied backgrounds, positioning of people, occluding objects and lighting. The second row shows images from DriPE, displaying drivers in real car interiors while driving. In the last row, we present images from our SimulatedCabin data generator. Different to PeopleSansPeople, the domain of SimulatedCabin is visually quite close to DriPE. While we use similar, but not coinciding, camera views, the positioning of the driver and the car interior are mimicked. As a result, our approach based on SimulatedCabin appears better suited to the target domain than the more randomized PeopleSansPeople data.

V. Experiments and Results

In this section, we describe two experiments for transfer learning with synthetic data for driver pose estimation in real car interiors. We go into the procedure of each experiment and discuss the results. We use the metric Average Precision (AP) for 2D-keypoint detection performance [36].

In the first experiment, we investigate the performance of several models trained only on synthetic data in their application to real test data. To assess the impact of domain-specific versus general-purpose synthetic data, we train one model on the common PeopleSansPeople dataset and another model on our SimulatedCabin dataset. For the training of these two models, we first take the complete training dataset with 49k images and then a 4.9k images subset of it. All trained models are evaluated using the DriPE test set. The results can be seen in the first four rows of Table II. The shown AP scores demonstrate that synthetic training data alone is not sufficient to train a reasonable model. However, when comparing the SimulatedCabin and PeopleSansPeople results, we observe a performance difference of 28.3% (29.1% vs. 0.8%) between the AP values of the respective best models. We attribute the performance gain achieved by SimulatedCabin to the domain shift from general-purpose to the in-cabin environment.

For comparison with models trained only on real data, we train and evaluate additional models on the real DriPE data set using training data sizes of 1%, 10%, 50% and 100% of the training set, amounting to 74, 745, 3727, and 7453 images, respectively. The results are listed in the last four rows of Table II. As expected, we observe a significantly higher AP when both training and evaluation are performed on the real DriPE dataset. While the accuracy increases with growing size of the training data set, we can see that even 1% of the real data training set is enough to outperform the models trained on synthetic data only. This obvious limitation of exclusive training with simulated data leads to the design of our next experiment.

The second experiment deals with transfer learning from synthetic data to the real car interior domain. In this experiment, we use the synthetic data to pre-train the models and then fine-tune them with real data. For this purpose, we pre-train one model on the PeopleSansPeople dataset and one model on the SimulatedCabin dataset. In both cases, we generate two variants with the complete training dataset comprising 49k images and the reduced dataset containing 4.9k images, respectively. These models were already trained in the first experiment. The two models are now fine-tuned on differently sized fractions of the DriPE train set. More precisely, we use the pre-trained models and fine-tune them on subsets of 1%, 10%, 50%, and 100% of the entire DriPE train set. The resulting models are evaluated on the DriPE test set, and the results are shown in Table III. The table also lists the results of models generated without pre-training on synthetic data in the first row of each group. The AP scores presented in Table III reveal that the performance of all models improves with pre-training on synthetic data. The overall best performing model is the one pre-trained on SimulatedCabin and fine-tuned on the whole DriPE dataset. It outperforms the corresponding model without pre-training by 2.0% (97.9% vs. 95.9%). These improvements are more significant when a smaller amount of real-world DriPE data is used for fine-tuning. For the smallest subset of 1% DriPE images, we find an improvement in AP score of 30.5% (84.1% vs. 53.6%).

Regarding the impact of the domain of the simulated dataset, Table III demonstrates that the AP values resulting from using the SimulatedCabin dataset for pre-training in all cases outperform the corresponding PeopleSansPeople results. The

Training data	Data size	Training steps	AP
PSP	4.9 x 10^3	8700	0.0
	49 x 10^3	156300	**0.8**
SC	4.9 x 10^3	8700	1.2
	49 x 10^3	156300	**29.1**
DriPE	74	300	53.6
	745	2400	68.9
	3727	11700	92.3
	7453	23300	**95.9**

TABLE II: Keypoint test metrics for three models trained on different sets of training data with initialized weights from ImgNet and evaluated on the DriPE test set. PSP stands for PeopleSansPeople dataset and SC stands for SimulatedCabin dataset. The highest metrics in each category of training data are in boldface.

Fine-tune data size DriPE	Pre-train data	Fine-tune training steps	AP	Δ
74	None	300	53.6	-
	4.9 x 10^3 PeopleSansPeople	300	64.2	+10.6
	49 x 10^3 PeopleSansPeople	300	75.8	+22.2
	4.9 x 10^3 SimulatedCabin	300	77.3	+23.7
	49 x 10^3 SimulatedCabin	300	**84.1**	+30.5
745	None	2400	68.9	-
	4.9 x 10^3 PeopleSansPeople	2400	69.6	+0.7
	49 x 10^3 PeopleSansPeople	2400	82.1	+13.2
	4.9 x 10^3 SimulatedCabin	2400	80.3	+11.4
	49 x 10^3 SimulatedCabin	2400	**88.2**	+19.3
3727	None	11700	92.3	-
	4.9 x 10^3 PeopleSansPeople	11700	92.5	+0.2
	49 x 10^3 PeopleSansPeople	11700	92.7	+0.4
	4.9 x 10^3 SimulatedCabin	11700	94.8	+2.5
	49 x 10^3 SimulatedCabin	11700	**95.4**	+3.1
7453	None	23300	95.9	-
	4.9 x 10^3 PeopleSansPeople	23300	96.6	+0.7
	49 x 10^3 PeopleSansPeople	23300	96.9	+1.0
	4.9 x 10^3 SimulatedCabin	23300	97.3	+1.4
	49 x 10^3 SimulatedCabin	23300	**97.9**	+2.0

TABLE III: Keypoint test metrics for models pre-trained on PeopleSansPeople (PSP) or SimulatedCabin and fine-tuned on DriPE train sets. For all models, we report the results on the DriPE test set. The highest AP in each category is in boldface. Δ describes the improvement of a model's AP when pre-training is applied.

difference amounts to 8.3% (84.1% vs. 75.8%) for the smallest amount of real data for fine-tuning, and 1% (97.9% vs. 96.9%) when the whole DriPE dataset is used for fine-tuning. Also noteworthy is the better performance of the reduced-size SimulatedCabin dataset (with 4.9k images) compared to the full-size PeopleSansPeople (with 49k images) which we observe for 1%, 50%, and 100% of the DriPE test set. This reinforces the observed benefit of the simulated cabin environment.

VI. Conclusions

In this work, we presented various benchmark results for driver pose estimation using transfer learning from synthetic to real-world data. We investigated the impact of the synthetic environment of the datasets on the 2D-keypoint detection performance. While the models trained purely on synthetic data suffer from poor performance on the real DriPE dataset, we report a remarkable performance gain of up to 30% for the investigated models when combining pre-training on synthetic data with fine-tuning on real data. The improvement becomes particularly visible when only a small amount of real data is available for fine-tuning, which underlines the relevance of our approach for practical applications.

As part of our work, we describe a simulation procedure for generating synthetic images of drivers in a vehicle cockpit. Our evaluation results document the advantage in terms of increased accuracy achieved by shifting the simulation to the application (i.e., car interior) domain. Generally, the simulated data not only helps overcome a lack of real annotated data, but has the additional benefit of avoiding potential violations of data protection and privacy regulations. Regarding ethics and prevention of data bias, variations in age, gender and ethnicity are taken into account in the design of our SimulatedCabin simulator.

We have shown that data generation in a similar synthetic domain improves human pose estimation. While we have verified the applicability of our approach in the domain of driver pose estimation, further research is needed to assess if comparable results can also be obtained from images of other application domains. Another topic for future research would be to explore whether further increasing the complexity and photorealism of the simulation can lead to additional improvements of the estimation results. To overcome a current limitation of our simulation, future work could involve the simulation of authentic lighting conditions, which are inherently complex and challenging to replicate. Future extensions to the simulator could pay specific attention to pose interactions with different objects, such as smartphones, in the car cockpit, and rare scenarios including hazardous driving behavior or accidents.

References

[1] T. L. Munea, Y. Z. Jembre, H. T. Weldegebriel, L. Chen, C. Huang, and C. Yang, "The progress of human pose estimation: A survey and taxonomy of models applied in 2d human pose estimation," *IEEE Access*, vol. 8, pp. 133 330–133 348, 2020.

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

[3] M. Cormier, F. Röpke, T. Golda, and J. Beyerer, "Interactive labeling for human pose estimation in surveillance videos," in *IEEE International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 1649–1658.

[4] P. Voigt and A. Von dem Bussche, *The EU general data protection regulation (GDPR)*. Springer, 2017.

[5] A. Harvey and J. LaPlace, "Exposing.ai," https://exposing.ai, accessed: 2022-10-30.

[6] M. Fabbri, G. Brasó, G. Maugeri, O. Cetintas, R. Gasparini, A. Ošep, S. Calderara, L. Leal-Taixé, and R. Cucchiara, "Motsynth: How can synthetic data help pedestrian detection and tracking?" in *IEEE International Conference on Computer Vision (ICCV)*, 2021, pp. 10 829–10 839.

[7] S. E. Ebadi, Y. Jhang, A. Zook, S. Dhakad, A. Crespi, P. Parisi, S. Borkman, J. Hogins, and S. Ganguly, "Peoplesanspeople: A synthetic data generator for human-centric computer vision," *arXiv*, 2021.

[8] D. T. Hoffmann, D. Tzionas, M. J. Black, and S. Tang, "Learning to train with synthetic humans," in *German Conference on Pattern Recognition (GCPR)*. Springer, 2019, pp. 609–623.

[9] G. Varol, J. Romero, X. Martin, N. Mahmood, M. J. Black, I. Laptev, and C. Schmid, "Learning from synthetic humans," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 4627–4635.

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

[11] C. Doersch and A. Zisserman, "Sim2real transfer learning for 3d human pose estimation: motion to the rescue," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

[12] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 43, no. 10, p. 3349–3364, 2021.

[13] R. Guesdon, C. Crispim-Junior, and L. Tougne, "Dripe: A dataset for human pose estimation in real-world driving settings," in *IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, 2021, pp. 2865–2874.

[14] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2021.

[15] M. Fabbri, F. Lanzi, S. Calderara, A. Palazzi, R. Vezzani, and R. Cucchiara, "Learning to detect and track visible and occluded body joints in a virtual world," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 430–446.

[16] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European Conference on Computer Vision (ECCV)*, 2014, pp. 740–755.

[17] L. G. T. Ribas, M. P. Cocron, J. L. Da Silva, A. Zimmer, and T. Brandmeier, "In-cabin vehicle synthetic data to test deep learning based human pose estimation models," in *IEEE Intelligent Vehicles Symposium (IV)*, 2021, pp. 610–615.

[18] G. Borghi, M. Venturelli, R. Vezzani, and R. Cucchiara, "Poseidon: Face-from-depth for driver pose estimation," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 5494–5503.

[19] M. Selim, A. Firintepe, A. Pagani, and D. Stricker, "Autopose: Large-scale automotive driver head pose and gaze dataset with deep head orientation baseline," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*, 2020, pp. 599–606.

[20] T. Hu, S. Jha, and C. Busso, "Robust driver head pose estimation in naturalistic conditions from point-cloud data," in *IEEE Intelligent Vehicles Symposium (IV)*, 2020, pp. 1176–1182.

[21] R. Fischer, M. Hödlmoser, and M. Gelautz, "Evaluation of camera pose estimation using human head pose estimation," *SN Computer Science*, vol. 4, no. 3, pp. 301:1–301:15, 2023.

[22] B. Reddy, Y.-H. Kim, S. Yun, C. Seo, and J. Jang, "Real-time driver drowsiness detection for embedded system using model compression of deep neural networks," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2017, pp. 438–445.

[23] R. Jabbar, M. Shinoy, M. Kharbeche, K. Al-Khalifa, M. Krichen, and K. Barkaoui, "Driver drowsiness detection model using convolutional neural networks techniques for android application," in *IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT)*, 2020, pp. 237–242.

[24] J. D. Ortega, N. Kose, P. Cañas, M.-A. Chao, A. Unnervik, M. Nieto, O. Otaegui, and L. Salgado, "Dmd: A large-scale multi-modal driver monitoring dataset for attention and alertness analysis," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 387–405.

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

[26] Z. Yao, Y. Liu, Z. Ji, Q. Sun, P. Lasang, and S. Shen, "3d driver pose estimation based on joint 2d-3d network," in *IEEE International Conference on Image Processing (ICIP)*, 2019, pp. 2546–2550.

[27] M. Martin, M. Voit, and R. Stiefelhagen, "An evaluation of different methods for 3d-driver-body-pose estimation," in *IEEE International Intelligent Transportation Systems Conference (ITSC)*, 2021, pp. 1578–1584.

[28] M. Martin, S. Stuehmer, M. Voit, and R. Stiefelhagen, "Real time driver body pose estimation for novel assistance systems," in *IEEE International Conference on Intelligent Transportation Systems (ITSC)*, 2017, pp. 1–7.

[29] S. D. Da Cruz, O. Wasenmüller, H.-P. Beise, T. Stifter, and D. Stricker, "Sviro: Synthetic vehicle interior rear seat occupancy dataset and benchmark," in *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2020, p. 962–971.

[30] J. S. Katrolia, A. El-Sherif, H. Feld, B. Mirbach, J. Rambach, and D. Stricker, "Ticam: A time-of-flight in-car cabin monitoring dataset," in *British Machine Vision Conference (BMVC)*, 2021, pp. 277–289.

[31] "RenderPeople: Over 4,000 scanned 3D people models," https://renderpeople.com/, accessed: 2022-10-30.

[32] "Hum3D: 3D models for design, AR and visualization," https://hum3d.com/, accessed: 2022-10-30.

[33] S. Borkman, A. Crespi, S. Dhakad, S. Ganguly, J. Hogins, Y.-C. Jhang, M. Kamalzadeh, B. Li, S. Leal, P. Parisi, C. Romero, W. Smith, A. Thaman, S. Warren, and N. Yadav, "Unity perception: Generate synthetic data for computer vision," *arXiv preprint arXiv:2107.04259*, 2021.

[34] "PolyHaven: The public 3D asset library," https://polyhaven.com/, accessed: 2022-10-30.

[35] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 248–255.

[36] "COCO Keypoint Evaluation," https://cocodataset.org/\#keypoints-eval, accessed: 2022-11-09.