



HAL
open science

Synthetic Driver Image Generation for Human Pose-Related Tasks

Romain Guesdon, Carlos F Crispim-Junior, Laure Tougne

► **To cite this version:**

Romain Guesdon, Carlos F Crispim-Junior, Laure Tougne. Synthetic Driver Image Generation for Human Pose-Related Tasks. International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP), Feb 2023, Lisbonne, Portugal. 10.5220/0011780800003417 . hal-03936401

HAL Id: hal-03936401

<https://hal.science/hal-03936401>

Submitted on 12 Jan 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Synthetic Driver Image Generation for Human Pose-Related Tasks

Romain Guesdon¹, Carlos Crispim-Junior¹, and Laure Tougne Rodet¹

¹*Univ Lyon, Univ Lyon 2, CNRS, INSA Lyon, UCBL, Centrale Lyon
LIRIS UMR5205, F-69676 Bron, France
{romain.guesdon, carlos.crispim-junior, laure.tougne} @liris.cnrs.fr*

Keywords: Dataset, synthetic generation, neural networks, human pose transfer, consumer vehicle

Abstract: The interest in driver monitoring has grown recently, especially in the context of autonomous vehicles. However, the training of deep neural networks for computer vision requires more and more images with significant diversity, which does not match the reality of the field. This lack of data prevents networks to be properly trained for certain complex tasks such as human pose transfer which aims to produce an image of a person in a target pose from another image of the same person. To tackle this problem, we propose a new synthetic dataset for pose-related tasks. By using a straightforward pipeline to increase the variety between the images, we generate 200k images with a hundred human models in different cars, environments, lighting conditions, etc. We measure the quality of the images of our dataset and compare it with other datasets from the literature. We also train a network for human pose transfer in the synthetic domain using our dataset. Results show that our dataset matches the quality of existing datasets and that it can be used to properly train a network on a complex task. We make both the images with the pose annotations and the generation scripts publicly available.

1 INTRODUCTION

The increasing complexity of computer vision tasks over the years has led to a growth in the size of deep learning models. Therefore, more and more data has been required to train the deep neural networks, with more diversity among the images. Large-scale general datasets have been published over the years to answer this problem, such as ImageNet (Deng et al., 2009), COCO (Lin et al., 2015), or DeepFashion (Liu et al., 2016) datasets. However, specific contexts lack sufficiently large datasets, especially because of the high cost of acquisition in comparison with the size of the research field.

Human Pose Transfer (HPT) is an example of a data-demanding task. HPT aims to generate, from a source image of a person, a new image of that same person in a different target pose. Generative Adversarial Networks (GAN) (Goodfellow et al., 2014) achieve good performances on this task (Zhu et al., 2019; Huang et al., 2020; Zhang et al., 2021), mostly in two contexts: fashion and video surveillance images. These two domains correspond to the two main datasets available for this task (Liu et al., 2016; Zheng et al., 2015). However, a substantial number of images, with high diversity in persons, clothes, and environment is required to properly train GAN models.

These requirements are difficult to achieve in specific contexts, for example, images of drivers in consumer vehicles. In this context, data acquisition requires setting up experimentations in a moving car (Guesdon et al., 2021) or at least in a simulator (Martin et al., 2019). These constraints lead to the availability of few images with little variety of subjects.

A commonly used solution to tackle a lack of training data is geometric data augmentation such as random rotation, crop, scaling, etc. (Simard et al., 2003; Krizhevsky et al., 2012). However, these methods may be sufficient for rigid objects but are not fully suitable for articulated ones. An alternative is the use of synthetic data. This process allows the generation of a high number of images with a theoretically infinite diversity and accurate annotations, within a limited time and financial cost. Even if a domain gap exists between synthetic and real images, literature has demonstrated that generated images can be used to assist the training of networks on real-world images for many tasks (Juraev et al., 2022; Wu et al., 2022; Kim et al., 2022). In the driving context, few synthetic public datasets exist (Cruz et al., 2020; Katrolija et al., 2021). Furthermore, these datasets mainly focus on monitoring tasks and emphasize more on actions than on subject diversity.

To address the lack of diversity in driving vehi-



Figure 1: Samples of images from the proposed synthetic dataset.

54 cles, we propose a large dataset of synthetic images 77
 55 for pose-related tasks. We develop a pipeline where 78
 56 we diversify the subjects (with 100 driver models), 79
 57 but also the car cockpits, the environment, the light- 80
 58 ing conditions, etc. The images are publicly available, 81
 59 as well as the scripts used for data generation ¹. 82

60 This paper is organized as follows. Section 2 83
 61 presents related work on driver image datasets. In 84
 62 Section 3, we present our proposed process and the 85
 63 synthetic dataset along with the choices made for the 86
 64 generation. We show and evaluate in Section 4 the 87
 65 generated images and an application of our dataset 88
 66 with an HPT architecture. Finally, Section 5 presents 89
 67 our conclusions and future work.

68 2 RELATED WORK

69 Work in the computer-vision field about drivers in 95
 70 consumer vehicles mainly focuses on passenger mon- 96
 71 itoring, mostly for safety-related tasks. Therefore, 97
 72 datasets in real-world conditions or in driving simu- 98
 73 lators have been published for tasks such as driver ac- 99
 74 tivity recognition (Ohn-Bar et al., 2014; Jegham et al., 100
 75 2019; Martin et al., 2019; Borghi et al., 2020), driver 101
 76 pose estimation (Guesdon et al., 2021), driver gaze 102

¹Images and generation scripts are publicly available 104
 on : [https://gitlab.liris.cnrs.fr/aura_autobehave/synthetic_](https://gitlab.liris.cnrs.fr/aura_autobehave/synthetic_drivers) 105
 drivers 106

estimation (Ribeiro and Costa, 2019; Selim et al., 2020), driver awareness monitoring (Abtahi et al., 2014).

Most of these datasets contain RGB images from video clips annotated for the target tasks. However, these datasets usually do not provide pose annotations required for the study of human pose transfer tasks. Drive&Act (Martin et al., 2019) proposes a multi-modal (RGB, NIR, depth) and multi-view dataset in a static driving simulator, with 3D human pose and activity annotations. DriPE dataset (Guesdon et al., 2021) depicts drivers in consumer vehicles in real-world driving conditions, with manually annotated poses. However, these two datasets contain only 15 and 19 subjects, respectively, which is not enough to fully train deep neural networks on a complex task, such as HPT, according to our observations.

Regarding synthetic data for driver monitoring, two datasets have been published. SVIRO (Cruz et al., 2020), a synthetic dataset for scenarios in the passenger cockpit. It depicts people and objects in the car back seat with different placements and provides RGB images along with infrared imitation, depth maps, segmentation masks, and human pose ground-truth keypoints. TICaM (Katrolia et al., 2021) is a dataset with both real and synthetic images for vehicle interior monitoring, with real images recorded in a car cockpit simulator. The dataset provides RGB, depth, and infrared images with action annotations and segmentation ground-truth masks. The two main issues

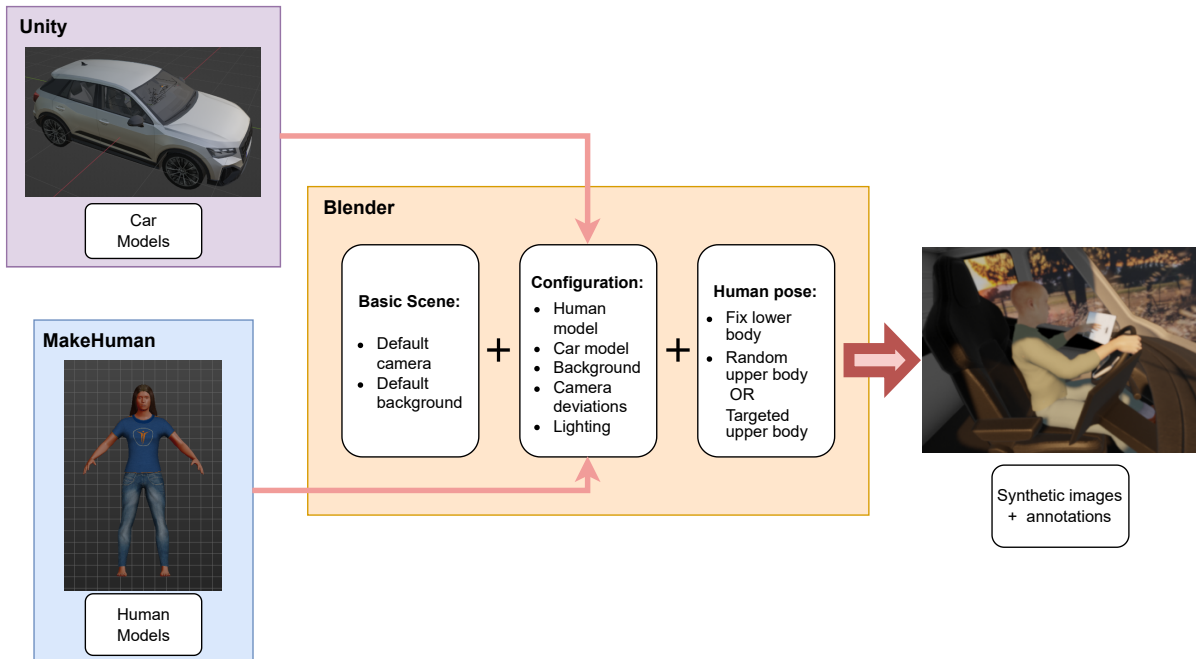


Figure 2: Global process for the generation of the synthetic driver images.

107 with these datasets are the front view angle, which 133
 108 does not allow a clear view of the driver’s full body, 134
 109 and the subject diversity which is still low for large 135
 110 models such as GAN (Goodfellow et al., 2014) on 136
 111 these data without overfitting. We can also mention 137
 112 Cañas *et. al.* (Canas et al., 2022) which describe 138
 113 a global approach to generate synthetic images for 139
 114 passenger monitoring. However, their work only par- 140
 115 tially considers the question of random pose genera- 141
 116 tion, and no script nor images have been made pub- 142
 117 licly available so far.

118 In summary, there currently exists no publicly 143
 119 available dataset suited to study driver pose transfer 144
 120 with a high variety of driver subjects and a full body 145
 121 view camera angle. 146

122 3 DATASET GENERATION

123 Because the driver datasets in the literature for hu- 151
 124 man pose-related tasks lack diversity, deep generative 152
 125 methods cannot be trained and used to increase the 153
 126 available data quantity. We propose a process based 154
 127 on a standard pipeline for 3D scene generation to ren- 155
 128 der new synthetic images. Using this method, we 156
 129 build a large dataset depicting one hundred human in- 157
 130 stances, several car models, variations of luminance, 158
 131 etc. In this section, we describe the generation pro- 159
 132 cess and present statistics about the generated images.

3.1 3D Models

To generate synthetic driver images, two objects need to be modeled: cars and humans. Human models are generated using MakeHuman Community (MakeHuman, 2022). This open-source software produces 3D models with many parameters like age, height, muscle mass, ethnicity, face proportions, etc. Models are generated with a rigged skeleton, which allows animating them easily and realistically. We use the default clothes from MakeHuman along with some provided by the community. To generate many models, we use the Mass Produce module which allows setting an interval for each parameter. We also randomly change the color of the clothes’ textures when generating the full scene to increase the diversity. The car models are obtained on the Unity Asset store (Unity, 2022). We select different types of consumer vehicles to represent various car cockpits (*e.g.*, family cars, sports cars, pick-ups), with equipment going from plain dashboards to touchscreens.

3.2 Pose Generation

Human models are animated using the included rigged skeleton (Figure 3-a). Theoretically, each bone can rotate freely around the body joint where its head is attached, which gives it three degrees of liberty. However, several constraints must be considered in our case. First, no real human bone can fully rotate

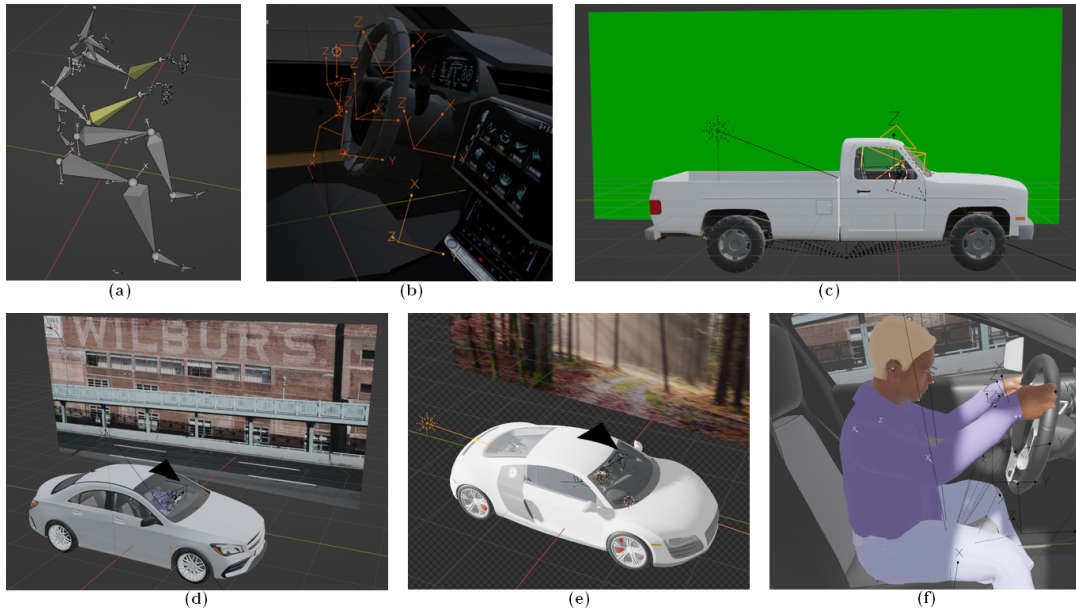


Figure 3: Illustrations of the generation process in Blender with (a) the skeleton rig, (b) the fixed wrist targets (only used for the additional driving images), (c) the default scene perspective, an example of the final scenes without (d) and with (e) the light rendering, and (f) a view from the camera.

160 in any direction. If we take the forearm for exam- 188
 161 ple and consider that it is fully open by default, it can 189
 162 approximately rotate from 0 to 150° around the pitch 190
 163 and the roll axis and cannot rotate around its yaw axis 191
 164 (Maik et al., 2010). Secondly, the car cabin is a con- 192
 165 stricted space, which brings many constraints to avoid 193
 166 the human and the car models colliding. Therefore, to 194
 167 address these constraints, we proceed as follows: 195

- 168 1. We define a default pose, which corresponds to 196
 169 the person sitting straight on the car seat with the 197
 170 arms close to the upper body. 198
- 171 2. We perform small random rotations on the head, 199
 172 back, and legs considering the human body con- 200
 173 straints and the car cabin. 201
- 174 3. We randomly defined a target for each wrist, in 202
 175 front of the subject and within the arm range. 203
 176 We also add a constraint to force the targets to 204
 177 be within a defined box that represents the cabin 205
 178 space. The boxes are manually defined before- 206
 179 hand for each car model to best match their shape. 207
- 180 4. We use an inverse kinematic solver integrated 208
 181 into the 3D modeling software to place the wrists 209
 182 on the targets. We only move the upper arms 210
 183 and forearms during this process, which does not 211
 184 modify the back inclination. This is to avoid un- 212
 185 natural poses in the car seat. Kinematic angle con- 213
 186 straints are set on each involved bone to match 214
 187 real body constraints. 215

This process allows us to easily generate many ran-
 dom plausible poses while taking into consideration
 body and environment constraints.

However, random positioning is very unlikely to
 generate standard driving poses, such as hands on the
 wheel or the gear lever. This is not problematic when
 considering the car as an autonomous vehicle of level
 2 or 3 for example, but can be less realistic for man-
 ual driving tasks (in a vehicle of levels of autonomy
 0 or 1). Therefore, we additionally set in each car
 model fixed wrist targets on the wheel, gear lever, and
 dashboard (Figure 3-b). We use these targets instead
 of random ones to separately generate more realistic
 driving images.

3.3 Generation Process

To set up the full scene and render the images, we use
 Blender 3.2 (Blender, 2022) modeling software. Its
 advantages are that it is free and open-source, accessi-
 ble, and can be fully automated using python scripts.
 The global rendering process is summarized in Fig-
 ure 2.

We first create the default scene by setting up a
 fixed camera, a sunlight source, and a panel for the
 background image (Figure 3-c). We use high-quality
 images of landscapes to simulate the background,
 which allows us to easily leverage a high number
 of different backgrounds from free picture databases.
 The 3D models are then imported into the scene.

Dataset	SVIRO	TICaM	Drive&Act	DriPE	Market	Fashion	Ours
Year	2020	2021	2019	2021	2015	2016	2022
#Frames	25K	126K	9.6M	10k	33k	54k	200k
#Subjects	22 adults	13	15	19	~ 3k	~ 10k	100
#Views	1	1	6	1	-	-	1
Synthetic / Real	Synthetic	Both	Real	Real	Real	Real	Synthetic
Data	Depth, RGB, IR	Depth, RGB, IR	Depth, RGB, IR	RGB	RGB	RGB	RGB
Annotation	Classification labels, 2D box mask, 2D skeleton	2D+3D boxes, 3D segmentation mask, activity	Activity, 2D+3D skeletons	2D boxes, skeleton	2D skeleton	2D skeleton	2D+3D skeletons and boxes

Table 1: Comparison table between different datasets.

Then, we randomly define several configurations, where a configuration is composed of a human model, a car model, a background, small camera deviations, and lighting parameters (Figure 3-d, e. Note that the black triangle in the illustrations represents the up direction of the camera model). We use a Blender add-on that places the sun in a realistic position from GPS coordinates and date time, which we set randomly. We also generate night configurations by selecting night backgrounds and dimming the lights. The night setting is randomly used 20% of the time.

Finally, for each configuration, we generate a pose using the process described in Section 3.2 (Figure 3-f) and render the image. We also save the 2D and 3D coordinates of each body joint, the bounding boxes, and the camera’s intrinsic and extrinsic parameters.

4 RESULTS AND DISCUSSIONS

In this section, we present and discuss methods used to evaluate the relevance of the proposed dataset. We first compare it with other state-of-the-art datasets using metrics from the literature to measure the quality of the images. Then, we use the task of human pose transfer to evaluate whether our synthetic dataset is large and diversified enough for a complex task.

4.1 Dataset Evaluation

We define a total of 1.000 configurations by randomly picking between 7 cars and 100 human models. For each configuration, 200 poses are generated, which results in a dataset of 200k images.

In Table 1, we compare our dataset with several other datasets from the literature. We can see that our dataset possesses more images than both driver synthetic and real-world HPT datasets. The only exception is Drive&Act, which is composed of

video clips instead of single images, which multiplies the total number of frames. However, the proposed dataset presents far more driver models than previous datasets.

Then, we compare the quality of the synthetic images with the ones in other datasets. For this purpose, we use the Inception Score (IS) (Salimans et al., 2016) which is a metric commonly used to evaluate the quality of images generated by GAN (Zhu et al., 2019; Tang et al., 2020; Huang et al., 2020). This metric is based on the predictions from a pre-trained InceptionNet classifier (Szegedy et al., 2016). Since Inception Score is sensitive to image sizes, each dataset is resized to approximately match the same number of pixels. We choose a standard size of 49,152 pixels, which corresponds to a shape of 192 * 256 pixels. The Inception Score is computed on the full datasets using a Pytorch implementation of the original IS algorithm (Pytorch metrics, 2022). Results of the evaluation can be found in Table 2.

Dataset	Inception Score (IS) ↑
DeepFashion	4.247
Market	4.223
DriPE	1.481
Drive&Act	1.343
SVIRO	1.902
TICam - synthetic	1.276
TICam - real	1.662
Ours	2.391

Table 2: Evaluation of the image quality of the full dataset using Inception Score.

First, we observe in Table 2 that the two datasets used for HPT, *i.e.*, DeepFashion and Market, present a score strictly higher than the one measured on driver datasets. This can be explained by the fact that the Inception Score reflects two aspects: the intrinsic quality of each image and the variety among

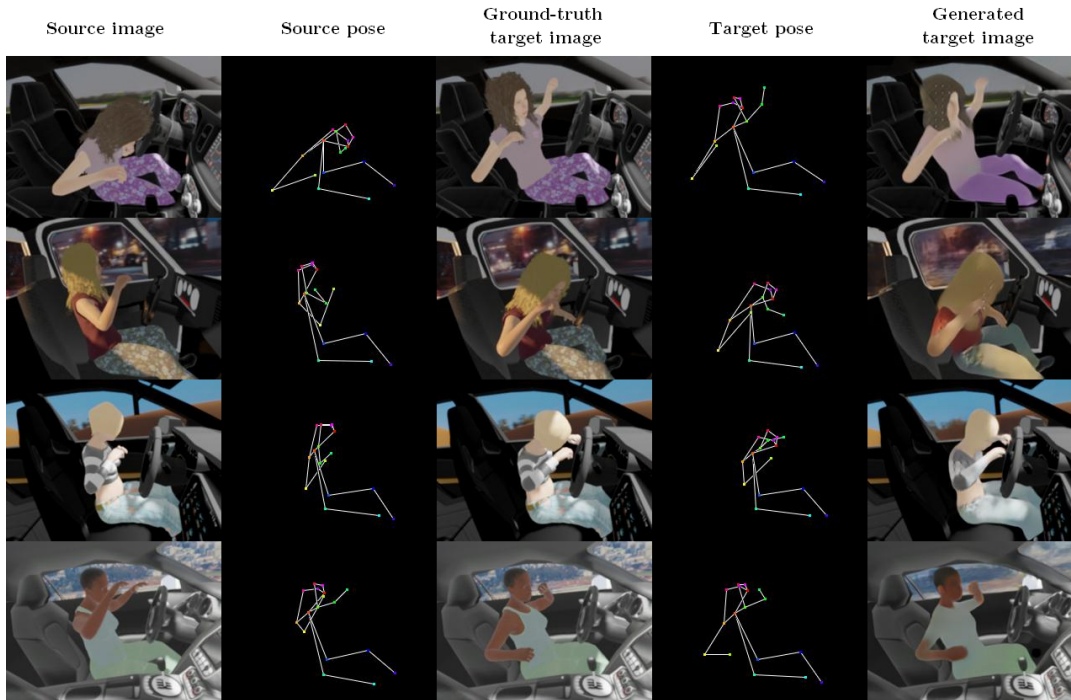


Figure 4: Samples from the test inferences generated by the GAN trained on our synthetic dataset.

276 the dataset (Salimans et al., 2016). Since the driver 303
 277 datasets present fewer subjects with large and fixed 304
 278 foregrounds, we can expect a lower IS. However, we 305
 279 can see that our synthetic dataset obtains a better score 306
 280 than the other driver datasets. This suggests that its 307
 281 images have an apparent quality similar to those from 308
 282 the other driver datasets while presenting a larger vari- 309
 283 ety. 310

284 4.2 Human Pose Transfer

285 As mentioned in Section 1, training a model for a 312
 286 complex task such as human pose transfer, without 313
 287 heavily overfitting the training set, requires many im- 314
 288 ages with a high variety of subjects. 315

289 Therefore, we train an HPT generative network 316
 290 on our synthetic dataset to evaluate the diversity of 317
 291 its images. We chose from the state of the art the 318
 292 APS architecture (Huang et al., 2020), which presents 319
 293 competitive performances with no need for additional 320
 294 input data such as segmentation maps. We train the 321
 295 network using the scripts provided by the authors in 322
 296 their repository. We adopted the same hyperparam-
 297 eters used for training on the DeepFashion dataset and
 298 resize our synthetic images to 192x256 pixels to get
 299 closer to the size of the DeepFashion images. The
 300 proposed dataset is split into a training set of 180k
 301 pictures and a testing set of 20k pictures, and these
 302 two sets do not share any subject model.

To measure the quality of our results, we evaluate
 the images using several state-of-the-art metrics (Table 3):
 Inception Score (IS), Frechet Inception Distance (FID)
 (Heusel et al., 2017), and Structural Similarity (SSIM)
 (Wang et al., 2004). FID and SSIM are computed using
 the same script as IS (Pytorch metrics, 2022). Unlike
 the evaluation of the datasets in Section 4.1, the metrics
 here are only computed on the images generated by the
 network on the test set.

Dataset	IS \uparrow	FID \downarrow	SSIM \uparrow
Fashion	3.565	16.84	0.669
Market	3.144	41.49	0.312
Synthetic	2.456	38.06	0.810

Table 3: Evaluation of images generated by an APS network trained on different datasets.

First, we observe that the Inception Score of the generated images is close to the one measured on the full synthetic dataset in Table 2. Then, the FID distance between the driver images generated by the GAN and the ground truth images is close to the one observed with the Market dataset. Furthermore, the SSIM score, which measures the structural similarity between two images, is higher on our synthetic dataset than on both Fashion and Market. This can be explained by the fact that more than half the surface of driver images is composed of a fixed background that

323 the GAN network can easily preserve since it almost 369
324 does not change during the pose transfer.

325 We can notice that the score measures on the Fash- 370
326 ion dataset are better than those on both the Market 371
327 and our synthetic dataset. This can be explained by 372
328 the simplicity of the Fashion images context, espe- 373
329 cially the lack of a complex background, fully visible 374
330 body parts, etc., in comparison with the real-life im- 375
331 ages in the two other datasets. 376

332 Finally, Figure 4 presents qualitative results of the 377
333 trained GAN. The generated images show that the 378
334 network learned to reproduce the pose while preserv- 379
335 ing most of the visual characteristics of the subject 380
336 and the global environment. This result indicates that 381
337 the network can learn and generalize on our dataset. 382
338 In the end, the evaluation results combined with the 383
339 qualitative results suggest that our dataset contains 384
340 enough diversity to train a network for a complex task 385
341 without overfitting. 386

342 5 CONCLUSION 387

343 In this paper, we have presented a dataset of 200k 388
344 synthetic driver images for human pose-related tasks 389
345 with a large diversity of human models to answer the 390
346 lack of available datasets on driver monitoring tasks. 391
347 Using state-of-the-art metrics, we demonstrated that 392
348 the quality of our synthetic images is comparable to 393
349 the one measured in existing datasets, synthetic or 394
350 real-world. We finally trained a GAN for human 395
351 pose transfer, a data-demanding task, on our synthetic 396
352 dataset. The network achieved similar performances 397
353 to those trained for HPT on real-world datasets for 398
354 other applications, which demonstrates that the pro- 399
355 posed synthetic dataset is diverse enough to train large 400
356 networks. This dataset is publicly available as well as 401
357 the script used to generate it. 402

358 Future work will investigate the problem of do- 403
359 main adaptation from synthetic to real-world driver 404
360 images in models for human pose-related tasks. 405
361 Moreover, the proposed pipeline could be used to ex- 406
362 tend our dataset with multiple views to approach tasks 407
363 such as 3D human pose estimation, or with real activ- 408
364 ities for passenger monitoring. 409

365 Acknowledgements 410

366 This work was supported by the Pack Ambition 411
367 Recherche 2019 funding of the French AURA Region 412
368 in the context of the AutoBehave project. 413

REFERENCES

- Abtahi, S., Omidyeganeh, M., Shirmohammadi, S., and Hariri, B. (2014). Yawdd: A yawning detection dataset. In *Proceedings of the 5th ACM multimedia systems conference*, pages 24–28.
- Blender (2022). Blender. <https://www.blender.org/>. Accessed: 2022-11-01.
- Borghi, G., Pini, S., Vezzani, R., and Cucchiara, R. (2020). Mercury: a vision-based framework for driver monitoring. In *International Conference on Intelligent Human Systems Integration*, pages 104–110. Springer.
- Canas, P. N., Ortega, J. D., Nieto, M., and Otaegui, O. (2022). Virtual passengers for real car solutions: synthetic datasets. *arXiv preprint arXiv:2205.06556*.
- Cruz, S. D. D., Wasenmuller, O., Beise, H.-P., Stifter, T., and Stricker, D. (2020). Sviro: Synthetic vehicle interior rear seat occupancy dataset and benchmark. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 973–982.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 248–255. Ieee.
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., and Weinberger, K., editors, *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.
- Guesdon, R., Crispim-Junior, C., and Tougne, L. (2021). Dripe: A dataset for human pose estimation in real-world driving settings. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 2865–2874.
- Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30.
- Huang, S., Xiong, H., Cheng, Z.-Q., Wang, Q., Zhou, X., Wen, B., Huan, J., and Dou, D. (2020). Generating person images with appearance-aware pose stylizer. In *IJCAI*.
- Jegham, I., Khalifa, A. B., Alouani, I., and Mahjoub, M. A. (2019). Mdad: A multimodal and multiview in-vehicle driver action dataset. In *International Conference on Computer Analysis of Images and Patterns*, pages 518–529. Springer.
- Juraev, S., Ghimire, A., Alikhanov, J., Kakani, V., and Kim, H. (2022). Exploring human pose estimation and the usage of synthetic data for elderly fall detection in real-world surveillance. *IEEE Access*, 10:94249–94261.
- Katrolia, J. S., El-Sherif, A., Feld, H., Mirbach, B., Rambach, J. R., and Stricker, D. (2021). Ticam: A time-of-flight in-car cabin monitoring dataset. In *32nd British*

- 427 *Machine Vision Conference 2021, BMVC 2021, On-* 487
428 *line, November 22-25, 2021*, page 277. BMVA Press. 488
- 429 Kim, T. S., Shim, B., Peven, M., Qiu, W., Yuille, A., and 489
430 Hager, G. D. (2022). Learning from synthetic vehi- 490
431 cles. In *Proceedings of the IEEE/CVF Winter Confer-* 491
432 *ence on Applications of Computer Vision*, pages 500– 492
433 508. 493
- 434 Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). 494
435 Imagenet classification with deep convolutional neu- 495
436 ral networks. In Pereira, F., Burges, C. J. C., Bottou, 496
437 L., and Weinberger, K. Q., editors, *Advances in Neu-* 497
438 *ral Information Processing Systems 25*, pages 1097– 498
439 1105. Curran Associates, Inc. 499
- 440 Lin, T.-Y., Maire, M., Belongie, S., Bourdev, L., Girshick, 500
441 R., Hays, J., Perona, P., Ramanan, D., Zitnick, C. L., 501
442 and Dollár, P. (2015). Microsoft coco: Common ob- 502
443 jects in context. 503
- 444 Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X. (2016). 504
445 Deepfashion: Powering robust clothes recognition and 505
446 retrieval with rich annotations. In *Proceedings of the* 506
447 *IEEE/CVF Conference on Computer Vision and Pat-* 507
448 *tern Recognition*, pages 1096–1104. 508
- 449 Maik, V., Paik, D., Lim, J., Park, K., and Paik, J. (2010). 509
450 Hierarchical pose classification based on human phys- 510
451 iology for behaviour analysis. *Computer Vision, IET*, 511
452 4:12 – 24. 512
- 453 MakeHuman (2022). Makehuman community. [http://www.](http://www.makehumancommunity.org/) 513
454 [makehumancommunity.org/](http://www.makehumancommunity.org/). Accessed: 2022-11-01. 514
- 455 Martin, M., Roitberg, A., Haurilet, M., Horne, M., Reiß, 515
456 S., Voit, M., and Stiefelhagen, R. (2019). Drive&act: 516
457 A multi-modal dataset for fine-grained driver behavior 517
458 recognition in autonomous vehicles. In *Proceedings of* 518
459 *the IEEE/CVF International Conference on Computer* 519
460 *Vision*, pages 2801–2810. 520
- 461 Ohn-Bar, E., Martin, S., Tawari, A., and Trivedi, M. M. 521
462 (2014). Head, eye, and hand patterns for driver activ- 522
463 ity recognition. In *2014 22nd international conference* 523
464 *on pattern recognition*, pages 660–665. IEEE. 524
- 465 Pytorch metrics (2022). Pytorch implementation of com- 525
466 mon gan metrics. [https://github.com/w86763777/](https://github.com/w86763777/pytorch-gan-metrics) 526
467 [pytorch-gan-metrics](https://github.com/w86763777/pytorch-gan-metrics). Accessed: 2022-11-01. 527
- 468 Ribeiro, R. F. and Costa, P. D. P. (2019). Driver gaze 528
469 zone dataset with depth data. In *2019 14th IEEE In-* 529
470 *ternational Conference on Automatic Face & Gesture* 530
471 *Recognition (FG 2019)*, pages 1–5. 531
- 472 Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., 532
473 Radford, A., and Chen, X. (2016). Improved tech- 533
474 niques for training gans. *Advances in neural informa-* 534
475 *tion processing systems*, 29. 535
- 476 Selim, M., Firintepe, A., Pagani, A., and Stricker, D. 536
477 (2020). Autopose: Large-scale automotive driver head 537
478 pose and gaze dataset with deep head orientation base- 538
479 line. In *VISIGRAPP (4: VISAPP)*, pages 599–606. 539
- 480 Simard, P. Y., Steinkraus, D., and Platt, J. C. (2003). Best 540
481 practices for convolutional neural networks applied to 541
482 visual document analysis. In *Proceedings of the Sev-* 542
483 *enth International Conference on Document Analysis* 543
484 *and Recognition-Volume 2*, page 958. 544
- 485 Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wo- 545
486 jna, Z. (2016). Rethinking the inception architecture 546
487 for computer vision. In *Proceedings of the IEEE/CVF* 547
488 *Conference on Computer Vision and Pattern Recogni-* 548
489 *tion*, pages 2818–2826. 549
- Tang, H., Bai, S., Zhang, L., Torr, P. H., and Sebe, N. 550
(2020). Xinggan for person image generation. In *Pro-* 551
ceedings of the European conference on computer vi- 552
sion, pages 717–734. 553
- Unity (2022). Unity asset store. [https://assetstore.unity.](https://assetstore.unity.com/) 554
[com/](https://assetstore.unity.com/). Accessed: 2022-11-01. 555
- Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. 556
(2004). Image quality assessment: from error visi- 557
bility to structural similarity. *IEEE transactions on* 558
image processing, 13(4):600–612. 559
- Wu, Y., Yuan, Y., and Wang, Q. (2022). Learning from 560
synthetic data for crowd instance segmentation in the 561
wild. In *2022 IEEE International Conference on Im-* 562
age Processing (ICIP), pages 2391–2395. IEEE. 563
- Zhang, J., Li, K., Lai, Y.-K., and Yang, J. (2021). Pise: 564
Person image synthesis and editing with decoupled 565
gan. In *Proceedings of the IEEE/CVF Conference* 566
on Computer Vision and Pattern Recognition, pages 567
7982–7990. 568
- Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, 569
Q. (2015). Scalable person re-identification: A bench- 570
mark. In *Proceedings of the IEEE/CVF International* 571
Conference on Computer Vision. 572
- Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., and Bai, X. 573
(2019). Progressive pose attention transfer for person 574
image generation. In *Proceedings of the IEEE/CVF* 575
Conference on Computer Vision and Pattern Recogni- 576
tion, pages 2347–2356. 577