Synthetic Driver Image Generation for Human Pose-Related Tasks
Romain Guesdon, Carlos F Crispim-Junior, Laure Tougne

To cite this version:

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 Guesdon$^1$, Carlos Crispim-Junior$^1$, and Laure Tougne Rodet$^1$

$^1$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; Katrolia 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.

...cles, we propose a large dataset of synthetic images for pose-related tasks. We develop a pipeline where we diversify the subjects (with 100 driver models), but also the car cockpits, the environment, the lighting conditions, etc. The images are publicly available, as well as the scripts used for data generation.

This paper is organized as follows. Section 2 presents related work on driver image datasets. In Section 3, we present our proposed process and the synthetic dataset along with the choices made for the generation. We show and evaluate in Section 4 the generated images and an application of our dataset with an HPT architecture. Finally, Section 5 presents our conclusions and future work.

2 RELATED WORK

Work in the computer-vision field about drivers in consumer vehicles mainly focuses on passenger monitoring, mostly for safety-related tasks. Therefore, datasets in real-world conditions or in driving simulators have been published for tasks such as driver activity recognition [Ohn-Bar et al., 2014; Jegham et al., 2019; Martin et al., 2019; Borghi et al., 2020], driver pose estimation [Guesdon et al., 2021], driver gaze 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...
with these datasets are the front view angle, which
does not allow a clear view of the driver’s full body,
and the subject diversity which is still low for large
models such as GAN (Goodfellow et al., 2014) on
these data without overfitting. We can also mention
Cañas et. al. (Canas et al., 2022) which describe
a global approach to generate synthetic images for
passenger monitoring. However, their work only par-
tially considers the question of random pose genera-
tion, and no script nor images have been made pub-
licly available so far.

In summary, there currently exists no publicly
available dataset suited to study driver pose transfer
with a high variety of driver subjects and a full body
view camera angle.

3 DATASET GENERATION

Because the driver datasets in the literature for hu-
man pose-related tasks lack diversity, deep generative
methods cannot be trained and used to increase the
available data quantity. We propose a process based
on a standard pipeline for 3D scene generation to ren-
der new synthetic images. Using this method, we
build a large dataset depicting one hundred human in-
stances, several car models, variations of luminance,
etc. In this section, we describe the generation pro-
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 (MakeHu-
man, 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 ani-
mating them easily and realistically. We use the de-
fault clothes from MakeHuman along with some pro-
vided by the community. To generate many models,
we use the Mass Produce module which allows set-
ing an interval for each parameter. We also randomly
change the color of the clothes’ textures when gener-
ating 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
in any direction. If we take the forearm for exam-
ple and consider that it is fully open by default, it can
approximately rotate from 0 to 150° around the pitch
and the roll axis and cannot rotate around its yaw axis
(Maik et al., 2010). Secondly, the car cabin is a con-
stricted space, which brings many constraints to avoid
the human and the car models colliding. Therefore, to
address these constraints, we proceed as follows:

1. We define a default pose, which corresponds to
the person sitting straight on the car seat with the
arms close to the upper body.

2. We perform small random rotations on the head,
back, and legs considering the human body con-
straints and the car cabin.

3. We randomly defined a target for each wrist, in
front of the subject and within the arm range.
We also add a constraint to force the targets to
be within a defined box that represents the cabin
space. The boxes are manually defined beforehand
for each car model to best match their shape.

4. We use an inverse kinematic solver integrated
into the 3D modeling software to place the wrists
on the targets. We only move the upper arms
and forearms during this process, which does not
modify the back inclination. This is to avoid un-
natural poses in the car seat. Kinematic angle con-
straints are set on each involved bone to match
real body constraints.

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 3b). 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 3c). 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.

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.
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.

First, we observe in Table 3 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

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Inception Score (IS) ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepFashion</td>
<td>4.247</td>
</tr>
<tr>
<td>Market</td>
<td>4.223</td>
</tr>
<tr>
<td>Drive&amp;Act</td>
<td>1.343</td>
</tr>
<tr>
<td>SVIRO</td>
<td>1.902</td>
</tr>
<tr>
<td>TICam - synthetic</td>
<td>1.276</td>
</tr>
<tr>
<td>TICam - real</td>
<td>1.662</td>
</tr>
<tr>
<td>Ours</td>
<td>2.391</td>
</tr>
</tbody>
</table>

Table 2: Evaluation of the image quality of the full dataset using Inception Score.
the dataset (Salimans et al., 2016). Since the driver datasets present fewer subjects with large and fixed foregrounds, we can expect a lower IS. However, we can see that our synthetic dataset obtains a better score than the other driver datasets. This suggests that its images have an apparent quality similar to those from the other driver datasets while presenting a larger variety.

### 4.2 Human Pose Transfer

As mentioned in Section 1, training a model for a complex task such as human pose transfer, without heavily overfitting the training set, requires many images with a high variety of subjects. Therefore, we train an HPT generative network on our synthetic dataset to evaluate the diversity of its images. We chose from the state of the art the APS architecture (Huang et al., 2020), which presents competitive performances with no need for additional input data such as segmentation maps. We train the network using the scripts provided by the authors in their repository. We adopted the same hyperparameters used for training on the DeepFashion dataset and resize our synthetic images to 192x256 pixels to get closer to the size of the DeepFashion images. The proposed dataset is split into a training set of 180k pictures and a testing set of 20k pictures, and these 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.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>IS ↑</th>
<th>FID ↓</th>
<th>SSIM ↑</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fashion</td>
<td>3.565</td>
<td>16.84</td>
<td>0.669</td>
</tr>
<tr>
<td>Market</td>
<td>3.144</td>
<td>41.49</td>
<td>0.312</td>
</tr>
<tr>
<td>Synthetic</td>
<td>2.456</td>
<td>38.06</td>
<td>0.810</td>
</tr>
</tbody>
</table>

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
the GAN network can easily preserve it almost does not change during the pose transfer.

We can notice that the score measures on the Fashion dataset are better than those on both the Market and our synthetic dataset. This can be explained by the simplicity of the Fashion images context, especially the lack of a complex background, fully visible body parts, etc., in comparison with the real-life images in the two other datasets.

Finally, Figure 3 presents qualitative results of the trained GAN. The generated images show that the network learned to reproduce the pose while preserving most of the visual characteristics of the subject and the global environment. This result indicates that the network can learn and generalize on our dataset.

In the end, the evaluation results combined with the qualitative results suggest that our dataset contains enough diversity to train a network for a complex task without overfitting.

5 CONCLUSION

In this paper, we have presented a dataset of 200k synthetic driver images for human pose-related tasks with a large diversity of human models to answer the lack of available datasets on driver monitoring tasks. Using state-of-the-art metrics, we demonstrated that the quality of our synthetic images is comparable to the one measured in existing datasets, synthetic or real-world. We finally trained a GAN for human pose transfer, a data-demanding task, on our synthetic dataset. The network achieved similar performances to those trained for HPT on real-world datasets for other applications, which demonstrates that the proposed synthetic dataset is diverse enough to train large networks. This dataset is publicly available as well as the script used to generate it.

Future work will investigate the problem of domain adaptation from synthetic to real-world driver images in models for human pose-related tasks. Moreover, the proposed pipeline could be used to extend our dataset with multiple views to approach tasks such as 3D human pose estimation, or with real activities for passenger monitoring.

Acknowledgements

This work was supported by the Pack Ambition Recherche 2019 funding of the French AURA Region in the context of the AutoBehave project.

REFERENCES


