Artificial Dummies for Urban Dataset Augmentation

Antonín Vobecký[1], **David Hurych**[2], **Michal Uřičář**, **Patrick Pérez**[2], **Josef Sivic**[1]
[1]Czech Institute of Informatics, Robotics and Cybernetics at the Czech Technical University in Prague
[2]valeo.ai
antonin.vobecky@cvut.cz, david.hurych@valeo.com, uricar.michal@gmail.com, patrick.perez@valeo.com,
josef.sivic@cvut.cz

Abstract

Existing datasets for training pedestrian detectors in images suffer from limited appearance and pose variation. The most challenging scenarios are rarely included because they are too difficult to capture due to safety reasons, or they are very unlikely to happen. The strict safety requirements in assisted and autonomous driving applications call for an extra high detection accuracy also in these rare situations. Having the ability to generate people images in arbitrary poses, with arbitrary appearances and embedded in different background scenes with varying illumination and weather conditions, is a crucial component for the development and testing of such applications. The contributions of this paper are three-fold. First, we describe an augmentation method for the controlled synthesis of urban scenes containing people, thus producing rare or never-seen situations. This is achieved with a data generator (called DummyNet) with disentangled control of the pose, the appearance, and the target background scene. Second, the proposed generator relies on novel network architecture and associated loss that takes into account the segmentation of the foreground person and its composition into the background scene. Finally, we demonstrate that the data generated by our DummyNet improve the performance of several existing person detectors across various datasets as well as in challenging situations, such as night-time conditions, where only a limited amount of training data is available. In the setup with only day-time data available, we improve the night-time detector by 17% log-average miss rate over the detector trained with the day-time data only.

1 Introduction

A high-quality dataset is a crucial element for every system using statistical machine learning and should represent the target scenarios. Usually, however, we do not have access to such high-quality data due to limited resources for capture and annotation or the inability to identify all the aspects of the target use-case in advance. A key challenge is to cover all the corner-case scenarios that might arise to train and deploy a reliable model.

Dataset diversity is crucial for the automotive industry, where handling highly complex situations in a wide variety of conditions (weather, time of day, visibility, etc.) is necessary given the strict safety requirements. Our goal is to address these requirements and enable a controlled augmentation of datasets for pedestrian detection in urban settings for automotive applications. While we focus here only on *training* dataset augmentation as a first, crucial step, the approach also aims at generating data for the system *validation*. Validation on such augmented data could complement track tests with dummy dolls of unrealistic appearance, pose, and motion, which are the current industry standard (Fig. 1). By analogy, we named our system DummyNet after this iconic doll that replaces real humans.

The key challenge in automotive dataset augmentation is to enable sufficient control over the generated distributions via input parameters that describe important corner cases and missing situations. In this work, we take a step in that direction and develop a method for controlled augmentation of person datasets, where people with adjustable pose and appearance are synthesized into real urban backgrounds in various conditions. This is achieved by a new Generative Adversarial Network (GAN) architecture, coined DummyNet, which takes as input the desired person pose, specified by skeleton keypoints, the desired appearance, specified by an input image, and a target background scene. See the image synthesis diagram in Fig. 2. The output of DummyNet is the given background scene containing the pedestrian with target pose and appearance composited into one image.

We demonstrate that augmenting training data in this way improves person detection performance, especially in low-data regimes where the number of real training examples is small or when the training and target testing distributions differ (e.g., day/night). The basic assumption of matching training and testing distributions is typically not satisfied when developing real detection systems (e.g., different country, time of day, weather, etc.). Our method allows for adapting the model to known or estimated target (real world) distribution via controlled augmentation of the training data.

Contributions. Our contributions are three-fold: (1) we develop an approach (DummyNet) for controlled data augmentation for person detection in automotive scenarios that enables independent control of the person's *pose*, *appearance* and the *background* scene; (2) the approach relies on a novel architecture and associated *appearance loss* that take into account the segmentation of the foreground pedestrian and its composition into the background scene. (3) we

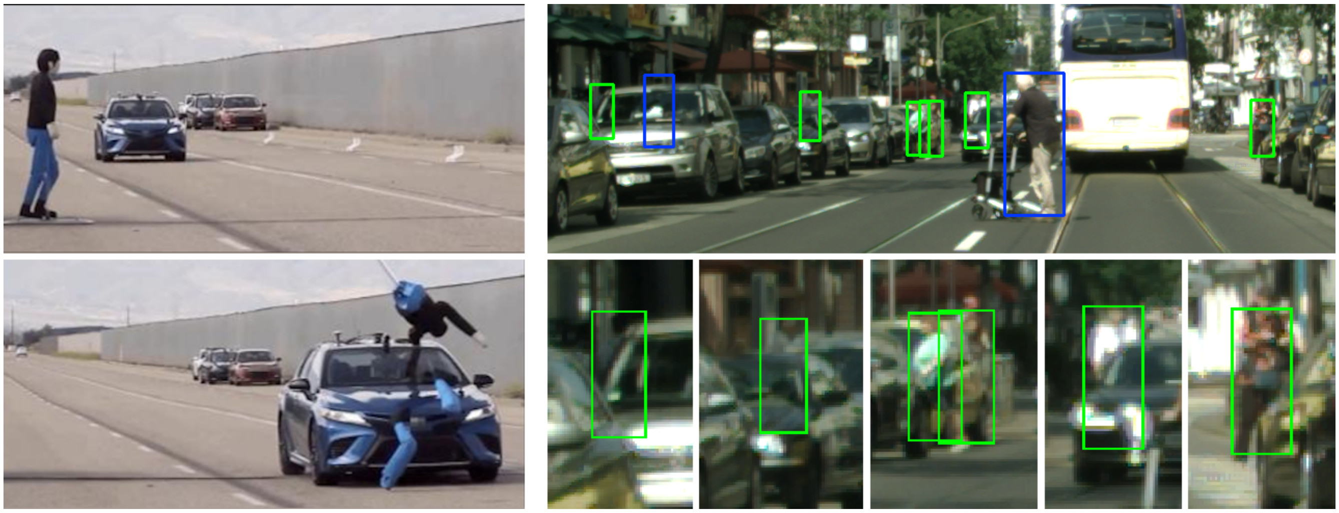

Figure 1: *Left*: Testing emergency breaking with a standard pedestrian dummy with a fixed pose and unified clothing. *Right:* Illustration of the improved performance of state-of-the-art person detector ("Center-and-Scale-Prediction", CSP, (Liu et al. 2019b)) trained on a dataset set augmented with samples produced by our DummyNet generator. Top: full testing scene. Bottom: close-ups. Detections (all correct) produced by both vanilla and DummyNet augmented CSP person detector are in blue. Additional detections (all correct) produced only by the DummyNet augmented CSP detector are shown in green. Note that the additional detections often correspond to hard occluded examples and are missed by the state-of-the-art CSP person detector.

demonstrate that the data generated by our DummyNet improve several existing person detectors with different architectures on standard benchmarks, including both day-time as well as challenging night-time conditions. In the setup with only day-time data available, using our artificial data, we improve the night-time detector by 17% LAMR over the detector trained with the day-time data only. [1]

2 Related Work and Positioning

Inserting humans. Inserting humans into a given scene is an old problem in image editing and visual effects. There are multiple good tools for image/video clipart (Lalonde et al. 2007) and for producing high-quality images (Karsch et al. 2011; Kholgade et al. 2014). However, such methods are meant for interactive manipulation of individual photos/shots, not for an automatic, large-scale generation of data. We do not necessarily need to generate high-quality visual data since what matters is the usefulness of the generated data for improving the person detector performance. There are several lines of work demonstrating improvements in various recognition tasks, even with non-photorealistic training data. The two prime examples are: (i) recent advances in domain randomization in robotics (Tobin et al. 2017; Loing, Marlet, and Aubry 2018) where object detectors are trained using synthetic scenes with random foreground and background textures, and (ii) optical flow estimators trained using the *flying chairs* dataset (Dosovitskiy et al. 2015), which pastes synthetically generated chairs onto a random background images. We follow this line of work and aim primarily at covering the different modes of appearance variation (pose, clothing, background) rather than synthesizing photorealistic outputs.

Generative data augmentation. We build on Generative Adversarial Networks (GANs) (Goodfellow 2016; Arjovsky, Chintala, and Bottou 2017; Lin et al. 2018; Salimans et al. 2016; Dumoulin et al. 2016; Nowozin, Cseke, and Tomioka 2016; Xiao, Zhong, and Zheng 2018; Donahue, Krähenbühl, and Darrell 2016; Radford, Metz, and Chintala 2015; Reed et al. 2016; Metz et al. 2016), which have shown a great progress recently in generating visual data (Mirza and Osindero 2014; Wang et al. 2018; Isola et al. 2017; Zhu et al. 2017; Liu, Breuel, and Kautz 2017; Huang et al. 2018; Park et al. 2019; Karras et al. 2018; Karras, Laine, and Aila 2019; Shaham, Dekel, and Michaeli 2019). GANs have also been useful for various forms of data augmentation, including (i) adding synthetic samples to expand a training set of real examples (Dwibedi, Misra, and Hebert 2017); (ii) improving the visual quality of synthetic samples generated via computer graphics (Huang and Ramanan 2017); or (iii) generating variants of original real samples (Wang et al. 2018; Choi et al. 2018; Pumarola et al. 2018). Recent progress in conditional image generation has enabled the generation of plausible images, and videos of never seen before humans (full body and faces) (Wu et al. 2019b; Karras, Laine, and Aila 2019), but these methods generate full images and cannot insert people into given backgrounds. Others have considered re-animating (re-targeting, puppeteering) real people in real scenes (Chan et al. 2019; Thies et al. 2016; Ma et al. 2017; Dong et al. 2018; Balakrishnan et al. 2018). These methods have demonstrated impressive visual results in image/video editing set-ups but are not geared towards large-scale generation of training data. We build on this work and develop an approach for disentangled and controlled augmentation of training data in automotive settings, where in contrast to previous work, our approach produces a full scene by explicitly estimating the foreground mask and compositing the

[1] Code is available at https://github.com/vobecant/DummyNet

pedestrian(s) into the background scene.

Synthesizing people. Other related methods, aiming to synthesize images of people, typically use conditional GANs. One line of work aims at changing the pose of a given person or changing the viewpoint. This is achieved either by using a two-stage approach consisting of pose integration and image refinement (Ma et al. 2017), by performing the warping of a given person image to a different pose with a specially designed GAN architecture (Dong et al. 2018), or by requiring segmentation of the conditional image into several foreground layers (Balakrishnan et al. 2018). Some works consider different factors of variation in person generation (Ma et al. 2018), but do not insert people into given full scenes in automotive settings as we do in our work. Others aim at changing only the pose (Balakrishnan et al. 2018; Liu et al. 2019a; Siarohin et al. 2018) or viewpoint (Si et al. 2018) of a given person in a given image keeping the appearance (e.g., clothing) of the person and the background scene fixed. In contrast, our model is designed to control independently the person's pose, appearance and the background scene.

The recent work (Wu et al. 2019a) uses as an input a silhouette (mask) sampled from an existing database of poses, which limits the diversity of the generated outputs. Our approach uses a generative model for poses (represented as keypoints), and we estimate the silhouette automatically. This allows for more complete control and higher diversity of output poses, including rare ones.

Others have also considered detailed 3D modeling of people and the input scene (Zanfir et al. 2020), or modeling individual component attributes (e.g., mixing clothing items such as shorts and a tank top from different input photographs (Men et al. 2020). While the outputs of these models are certainly impressive, their focus (and their evaluation) is on generating visually pleasing fashion photographs in, typically, indoor scenes with a limited variation of imaging conditions. In contrast, our model does not require 3D models of people or individual clothing attributes as input; we focus on demonstrating improved person detection performance in automotive settings and consider the appearance, pose, and the background scene in a single data generator model, which allows us to handle scenes with widely changing imaging conditions (e.g., day, night, snow).

The problem of enlarging the training dataset has been recently explored in (Liu et al. 2019a) where the generative model and a detector are optimized jointly, and in (Wu et al. 2019b) where a class-conditional GAN is used for synthesizing pedestrian instances. In contrast to these works, we train the generator and classifier separately and focus on having full control of generated person images as well as integration into complete scenes.

Cut-and-paste data augmentation methods. Cut-and-paste techniques can be also used to insert humans or objects into the given background image either randomly (Dwibedi, Misra, and Hebert 2017), ensuring a global visual consistency (Georgakis et al. 2017; Lee et al. 2018), or using 3D human models (Varol et al. 2017; Chen et al. 2016; Pishchulin et al. 2017; Zhou et al. 2010). The automotive setting related to our work has been considered in (Huang and Ramanan 2017), emphasizing the importance of detecting corner cases such as people in rare poses, children playing on the street, skateboarders, etc. To this end, the authors have collected an annotated dataset of pedestrians in dangerous scenarios obtained from a computer game engine. Using straight-up computer graphics generated samples is an approach parallel to ours. While it may offer higher image quality, it struggles to capture the diversity of the real world (e.g., texture, clothing, backgrounds) (Hattori et al. 2018; Marín et al. 2010) or require to control a lot of precise scene parameters (geometry, environment maps, lighting) and some user intervention (Alhaija et al. 2018).

3 Proposed Architecture and Loss

Our objective is the complete control of the target person's *pose*, *appearance*, and *background*. For example, suppose it is difficult to collect large amounts of training data of people in the night or snowy conditions, but "background" imagery (with no or few people only) is available for these conditions. Our approach allows synthesizing new scenes containing people with various poses and appearances embedded in the night and snowy backgrounds.

To achieve that, we have developed *DummyNet*, a new generative adversarial network architecture that takes the person's *pose*, *appearance* and *target background* as input and generates the output scene containing the composited person. Our approach has the following three main innovations. First, the control is achieved by explicitly conditioning the generator on these three different inputs. Second, the model automatically estimates a mask that separates the person in the foreground from the background. The mask allows focusing compute on the foreground or background pixels as needed by the different components of the model and is also used in calculating the loss function. Third, the architecture is trained with a new loss ensuring better appearance control of the inserted people. The overview of the entire approach is shown in Fig. 2.

The rest of this section describes the individual components of the model along with the three main areas of innovation. Additional details of the architecture, including the detailed diagrams of training and inference, as well as losses, are in the supplementary material (Vobecký et al. 2020).

3.1 Controlling Pose, Appearance and Background

As illustrated in Fig. 2, at inference time, the model takes three inputs (green boxes) that influence the *conditional generator*: (i) person's appearance (clothing, hair, etc.) is specified by an image sampled from a dataset of training images containing people and encoded by an *image encoder*. (ii) background scene (typically an urban landscape) is sampled from a dataset of background scenes; finally, (iii) a person's keypoints are produced by a *keypoint generator*. The main components of the architecture are described next.

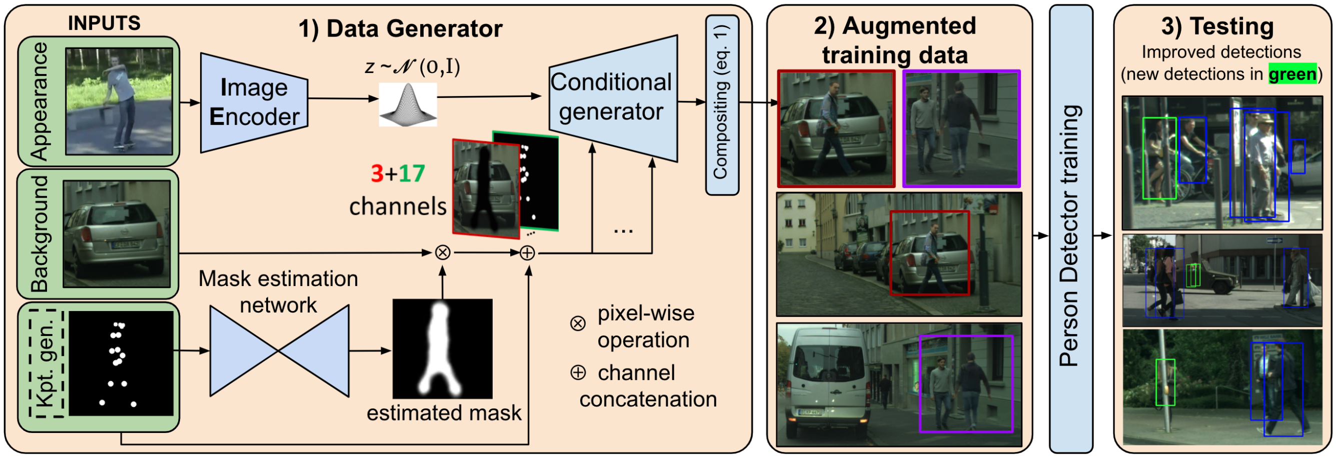

Figure 2: Augmenting training data with our DummyNet approach. The inputs are the desired pose (keypoints), desired pedestrian appearance (specified by an image), and the target background scene. The output is a scene with the composited pedestrian. The DummyNet data generator (1) augments the training data (2), which leads to improved person detector performance (3). The data generator box (1) displays the inference-time setup. For additional details, please see supp. mat. (Vobecký et al. 2020).

Conditional Generator. The generator takes as input the latent person appearance code, the background scene (with masked-out foreground pixels), and target keypoints in the form of 17 individual-channel heatmaps, one for each keypoint. The keypoint representation allows for fine control over the person's pose, including the ability to distinguish between frontal/back as well as left/right views, which would not be possible when using a person mask only. Conditioned on these inputs, the generator outputs an image of a person placed into the given background image in the pose given by the person keypoints and with the appearance specified by the provided appearance code (latent vector). This latent vector is passed through a fully-connected layer and further to the convolutional layers of the generator. The background image is concatenated with the keypoint heatmaps and used as an additional input to the generator. The generator architecture is based on residual blocks containing SPatially-Adaptive (DE)normalization layers (Park et al. 2019) and uses progressive growing. In each such block, the keypoints and the background are injected into the network through the SPADE layer. The corresponding patch discriminator based on (Isola et al. 2017) is described in the supplementary (Vobecký et al. 2020).

Person Appearance Encoder. We pre-train a variational autoencoder (VAE) separately from DummyNet. During DummyNet training, we use the encoder part of this VAE as a person appearance encoder. Its architecture comprises convolutional layers that take as input a 64×64 px person image with masked-out background and produce a latent vector encoding the appearance.

Keypoint Generator. The keypoint generator was created from the OpenPose (Cao et al. 2018) training set via viewpoint and pose clustering followed by principal component analysis within each pose cluster. This set of simple models fully captures the pose distribution and allows us to sample from it and use the result as input to the generator and the mask estimation network.

3.2 Mask Estimation and fg/bg Compositing

An essential component of our approach is the *mask estimator (ME)* that predicts the foreground/background (fg/bg) segmentation mask from the input set of keypoints. This is a U-Net-type network (Ronneberger, Fischer, and Brox 2015) that ingests keypoint locations encoded in 17 channels (COCO format, one channel per keypoint), and outputs a mask predicting which pixels belong to the person. This output is a one-channel map $\mathcal{M} \in [0,1]^{H \times W}$, where (H, W) are the height and the width of the output image, respectively. Output mask values are in the range of $[0,1]$ to composite the generated person image smoothly into the target real background image. We pre-train this network separately from DummyNet using person mask and keypoint annotations in MS COCO dataset (Lin et al. 2014). Using the pre-trained mask estimator network ME, we obtain the output mask $\mathcal{M}$ from the given input skeleton keypoints as $\mathcal{M} = \mathrm{ME}\left(\mathtt{kpts}\right)$. The estimated fg/bg mask is used at different places of the architecture as described next.

First, it is used to produce the foreground person image as input to the encoder producing the appearance vector and the background scene with masked out pixels prepared for foreground synthesis that are used as inputs to the conditional generator. Details are in the learning and inference diagrams in the supplementary material (Vobecký et al. 2020).

Second, the fg/bg mask is used to composite the final output. Given a background image $\mathtt{I}_{\text{bg}}$ and the output of the conditional generator $\mathtt{I}_{\text{gen}}$, the final augmented training image $\mathtt{I}_{\text{aug}}$ is obtained by compositing the generated foreground image with the background scene as

$$\mathtt{I}_{\text{aug}} = \mathcal{M} \odot \mathtt{I}_{\text{gen}} + (1 - \mathcal{M}) \odot \mathtt{I}_{\text{bg}}, \qquad (1)$$

where $\odot$ denotes the element-wise multiplication and $\mathcal{M}$ the estimated mask. In contrast to other works, we do not need to use manually specified masks during the inference time as we estimate them automatically with the Mask Estimator. This allows us to process more data and hence a more diverse set of poses. If the manual mask is available, it can be used as well. Finally, the estimated person mask is also used to compute fg/bg appearance losses as described next.

3.3 Masked Losses and Training

At training time, the model takes as input an image separated into the foreground person image and the background scene using the estimated fg/bg mask. The model is then trained to produce a realistic-looking output that locally fits the background scene and uses the appearance conditioning from the input foreground image. The network is trained with a weighted combination of four losses, described next.

For the discriminator, we use the Improved Wasserstein loss (WGAN-GP) (Gulrajani et al. 2017) as we found it more stable than other adversarial losses. The loss measures how easy it is to discriminate the output image composite given by Equation 1 from the real input training image.

Next, we introduce a new loss term that we found important for inserting pedestrians into various backgrounds. This person appearance consistency loss $\mathcal{L}_{\text{app}}$ encourages the generator to change the appearance of the inserted pedestrian with the change of the input person appearance latent vector. This is achieved by enforcing the latent appearance vector of the output image to match the latent appearance vector provided at the input of the generation process, measured by L_1 distance. This is implemented as

$$\mathcal{L}_{\text{app}} = \left\| \text{ENC}\left(\mathcal{M} \odot \mathtt{I}_{\text{in}}\right) - \text{ENC}\left(\mathcal{M} \odot \mathtt{I}_{\text{gen}}\right) \right\|_1 , \quad (2)$$

where ENC is the Image Encoder, $\mathcal{M}$ is the estimated person mask, $\mathtt{I}_{\text{in}}$ is the input image, and $\mathtt{I}_{\text{gen}}$ is the generated output. Please note how the estimated foreground mask $\mathcal{M}$ allows focusing the person appearance consistency loss on the person foreground pixels. An experiment showing the effect of appearance preconditioning can be found in the supplementary material (Vobecký et al. 2020).

We further add two reconstruction losses to our objective function but have them act only on the foreground person pixels via the estimated person mask. The first reconstruction loss, $\mathcal{L}_{\text{Rec-dis}}$, compares features extracted using the discriminator from the real input and generated output samples. The second reconstruction loss, $\mathcal{L}_{\text{Rec-VGG}}$ compares the real input and generated output in the feature space of a VGG19 network (Simonyan and Zisserman 2015) pre-trained for ImageNet classification. In both cases, the losses are modulated by the foreground mask to focus mainly on the foreground person pixels, similar to (2), and use L_1 distance. We found these losses greatly stabilize training. The final loss is a weighted sum of the terms described above:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{WGAN-GP}} + \lambda_2 \mathcal{L}_{\text{Rec-dis}} + \lambda_3 \mathcal{L}_{\text{Rec-VGG}} + \lambda_4 \mathcal{L}_{\text{app}}, \quad (3)$$

where hyperparameters λ_i's are set experimentally to ensure convergence and prevent overfitting. Please see the supplementary material for ablations measuring (among others) the influence of our proposed appearance loss and the importance of having control over the background scene as the input to the generator.

4 Experiments

In this section, we present a series of experiments on controlled dataset augmentation with the aim to improve the accuracy of a person classifier/detector in the context of autonomous driving. The augmentation is controlled as our DummyNet can generate images of people with a specific distribution of poses, appearances, and backgrounds. We consider four experimental set-ups. The first experiment (Sec. 4.1) focuses on augmenting the daytime Cityscapes dataset. We test data augmentation in the low-data regime, i.e., with insufficient real training data for training the person classifier. In the second experiment (Sec. 4.2), we use DummyNet to generate night-time person images and show significant improvements in classifier performance on the NightOwls dataset (Neumann et al. 2018). In the next experiment (Sec. 4.3) we use DummyNet to improve performance of the state-of-the-art person detection network CSP (Liu et al. 2019b) in the standard (full) data regime on the Cityscapes and Caltech datasets. Finally, we demonstrate the benefits of our approach in set-ups (Sec. 4.4) where we have only access for training to day-time annotated images (CityPersons) along with night-time images devoid of people, and we wish to detect pedestrians at night (NightOwls). See Fig. 3 for images generated by DummyNet.

Person Classifier. In experiments 4.1 and 4.2, we consider a CNN-based classifier with $6,729$ parameters, which is a realistic set-up for a digital signal processor in a car (where GPU is not available in a deployed system) and is trained *from scratch*. The classifier consists of 4 convolutional layers with a 3×3 kernel, stride 2, ReLU activations, max-pooling, and one fully connected layer with sigmoid activation. For both experiments, the classifier is trained for $1,000$ epochs and the classifier with the best validation error is kept.

Training Data for DummyNet. For training the generator, we aim for the real-world data with all its artifacts (blur, low/high res, changing lighting, etc.) and with enough samples with a person height of at least 190px. To achieve that, we leveraged the YoutubeBB (Real et al. 2017) dataset that contains a large number of videos with people in a large variety of poses, illuminations, resolutions, etc. We used OpenPose (Cao et al. 2018) to automatically detect people and annotate up to 17 keypoints based on their visibility.

The final dataset contains $769,176$ frames with annotated keypoints. Please note that the keypoints and mask annotations (estimated by our mask estimator) are noisy as they have been obtained automatically, yet, are sufficient for our needs. More details about the dataset, examples of training images with skeleton annotations, and the keypoint generator are in the supplementary (Vobecký et al. 2020).

4.1 Data Augmentation in a Low-Data Regime

In this experiment, we show how adding training person samples generated by DummyNet influences the testing performance of the person classifier. We conduct an experiment where there is only a small number of training samples available and investigate how adding synthesized images of the positive class (person) helps the performance of the resulting classifier. We compare against two baseline methods *Cut, Paste and Learn* (CPL) (Dwibedi, Misra, and Hebert 2017) and pix2pixHD (Wang et al. 2018). Both methods require

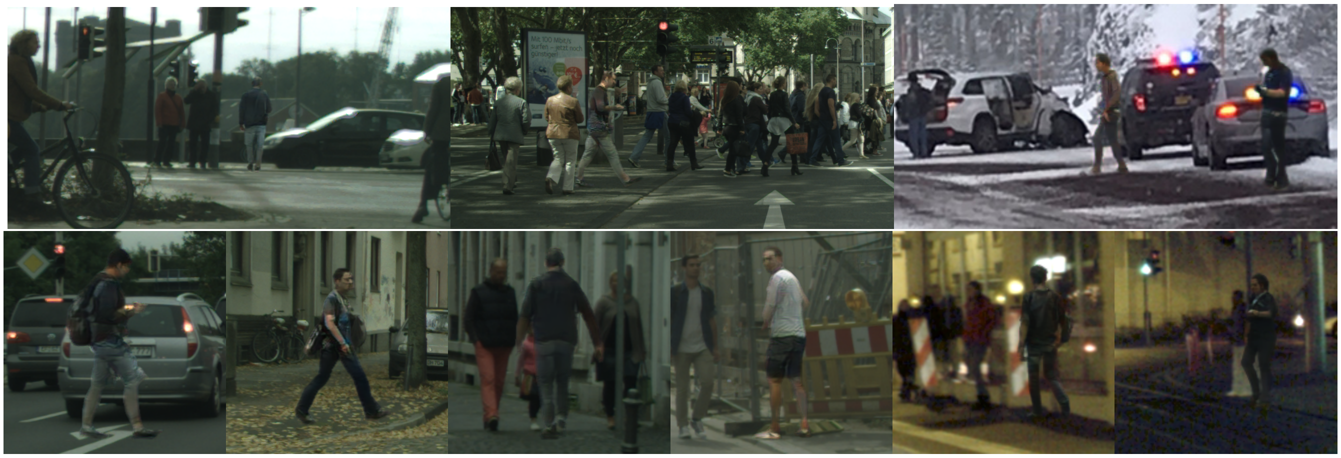

Figure 3: Artificial people in urban scenes produced by our DummyNet approach. Examples are shown at various scales and levels of detail. *Top*: Each full-scene image contains one synthetic person except for the winter scene where there are two. *Bottom*: Close-ups of day and night scenes with one synthetic person in each. Please see the supp. mat. (Vobecký et al. 2020) for qualitative results and ablations evaluating the importance of the different contributions.

generated	pix2pixHD		CPL		DummyNet	
FPR	1%	10%	1%	10%	1%	10%
0	0.980	0.682	0.980	0.682	0.980	0.682
75	0.852	0.424	**0.795**	**0.467**	**0.763**	**0.425**
200	**0.800**	**0.438**	0.809	0.466	0.836	0.490
500	0.865	0.616	0.818	0.567	0.861	0.636
1000	0.922	0.731	0.813	0.514	0.790	0.463

Table 1: Data augmentation in low-data regimes on the Cityscapes dataset. We report classifier test set miss rate (lower is better) at 1% and 10% false positive rate (FPR). 100 real samples were used for training, plus different amounts of generated samples (leftmost column) for augmentation. The best results are marked for 1% FPR in bold.

stronger input information as they need to have a segmentation mask of the inserted object. Besides, pix2pixHD requires to have a complete segmentation of the input scene. Therefore pix2pixHD and CPL have the advantage of additional information that our DummyNet does not require.

In Table 1, we report results in the form of a miss rate at 1% and 10% false positive rate. We compare results to the baseline classifier trained with only 100 real and no synthetic samples (first row) and investigate adding up to 1000 synthetic samples. In this low-data regime, the classifier trained on real data only performs poorly and adding synthesized images clearly helps. However, when too many synthetic examples are added, performance may decrease, suggesting that there is a certain amount of domain gap between real and synthetic data. Compared to the pix2pixHD and CPL, our DummyNet performs the best, bringing the largest performance boost, lowering the baseline miss rate by 21.7% at 1% FPR and by 25.7% at 10% FPR.

4.2 Person Classification at Night Time

Annotated training images with pedestrians at night are hard to obtain. On the other hand, it is relatively easy to get night scenes that contain no people. To this end, we construct an experiment on the NightOwls dataset (Neumann et al. 2018) and vary the number of available real (night-time) training person images. We then add more images of persons at night synthesized by DummyNet. Generated samples were obtained by providing the generator with day-time (but low light, based on thresholding the average brightness) input images of people together with night-time background scenes to get night-time output scenes with people. Please see the supp. mat. (Vobecký et al. 2020) for details.

gen\real	0		100		1000		12000 (full)	
FPR	1%	10%	1%	10%	1%	10%	1%	10%
0			0.88	0.62	0.64	0.34	0.51	0.28
5k	0.76	0.49	0.72	0.44	0.63	**0.33**	0.48	0.24
10k	**0.71**	**0.46**	0.72	**0.42**	**0.58**	**0.33**	**0.47**	0.25
20k	0.76	0.52	**0.72**	0.42	0.62	0.36	0.50	**0.22**

Table 2: Night-time person detection on the NightOwls dataset. We report the mean classifier test set miss rate (lower is better) over 5 runs. Test results are reported at 1% and 10% FPR. The best results for every combination are shown in bold. Samples generated by our method help to train a better classifier, often by a large margin, over the baseline trained only from real images (the first row).

Classification results on *testing NightOwls data* are shown in Table 2 and demonstrate that generated samples help to train a better classifier, which improves over the baseline by a large margin. The reported results are mean miss rates over five runs. In a low-data regime, we can lower the miss rates by nearly 20%. Adding synthetic examples improves performance in all settings, even with an increasing amount of real training data. In particular, note that we can improve performance even when having all the available real training data (column 'full set'). In that case, we improve MR at 10% FPR by more than 6%. Please note that DummyNet was not finetuned on this dataset. Please see additional results and examples of synthesized night-time data in Fig. 3 and the supplementary material (Vobecký et al. 2020).

4.3 Improving State-of-the-Art Person Detector

Citypersons. We conduct an experiment with one of the state-of-the-art person detectors, CSP (Liu et al. 2019b), on the full Citypersons dataset (Zhang, Benenson, and Schiele 2017). We use our DummyNet to extend the training set of this dataset. The augmentation procedure is the same for all the compared methods. It uses the semantic segmentation of the road scenes to place pedestrians at plausible locations (ground, road, sidewalk) and uses existing people in the scene (if available) to choose the position and size of the inserted people. We require that the new person stands on the ground and does not occlude other people already present in the image. The details of the augmentation procedure are given in the supplementary material (Vobecký et al. 2020).

Following (Liu et al. 2019b), we train the detection network for 150 epochs and report log-average miss rate for multiple setups of the detector with the best validation performance, see Table 3. We compare the following data augmentation setups: (a) original results reported in (Liu et al. 2019b); (b) our CSP detector retrained on original Citypersons images to reproduce results of (Liu et al. 2019b); (c) augmentation with the SURREAL (Varol et al. 2017) dataset, (d) CPL augmentation (Dwibedi, Misra, and Hebert 2017); (e) augmentation with recent ADGAN (Men et al. 2020) generative network; (f)-(g) augmentation with the Human3.6M dataset (Ionescu et al. 2014; Catalin Ionescu 2011) using provided segmentation or segmentation with DeepLabv3+ (Chen et al. 2018); and (h) training on DummyNet extended Cityscapes dataset.

We observe a consistent improvement in performance across all setups (see Supplementary for more details) when DummyNet augmented samples are used. Please note that differences between the reported results in (Liu et al. 2019b) (a) and our reproduced results (b) could be attributed to differences in initialization (random seed not provided in (Liu et al. 2019b)); otherwise, we use the same training setup.

Caltech. Following the experiments and the setup in (Liu et al. 2019b), we also train a CSP detector on the Caltech (Dollar et al. 2011) dataset. We initialize the detector weights with the best-performing network on CityPersons. Our data augmentation improves over the results reported in (Liu et al. 2019b) (the reasonable setup), reducing the LAMR from 3.8% to 3.47%, i.e., by almost 0.35%, which is non-negligible given the overall low LAMR. These results demonstrate the benefits of our approach on another challenging dataset.

setup	reasonable	small	partial
(a) CSP reported	11.02%	15.96%	10.43%
(b) CSP reproduced	11.44%	15.88%	10.72%
(c) SURREAL aug.	11.38%	17.39%	10.56%
(d) CPL aug.	11.36%	16.46%	10.84%
(e) ADGAN aug.	10.85%	16.20%	10.55%
(f) H3.6M aug., orig.	11.07%	16.66%	10.58%
(g) H3.6M aug., DLv3	10.59%	16.00%	10.21%
(h) DummyNet aug. (ours)	**10.25%**	**15.44%**	**9.12%**

Table 3: Improving state-of-the-art person detector (Liu et al. 2019b). Log-average miss rate of the detector (lower is better) in multiple testing setups.

setup	reasonable	small	occluded
(a) CityPersons ann.	41.90%	50.55%	65.95%
(b) ADGAN [Men20]	36.60%	48.91%	54.49%
(c) SURREAL [Varol17]	32.94%	44.05%	49.24%
(d) CPL [Dwibedi17]	30.73%	49.61%	54.60%
(e) H3.6M [Ionescu14]	27.83%	43.81%	45.67%
(f) DummyNet (ours)	**24.95%**	**39.73%**	**44.89%**

Table 4: Person detection in night-time scenes. LAMR (lower is better) in multiple testing setups on the NightOwls dataset.

4.4 Person Detection in Night-Time Scenes

Here we address the problem of an insufficient amount of annotated training data again. We conduct an experiment where we have access only to annotated images captured during daytime (CityPersons dataset), and we wish to detect pedestrians at night time (NightOwls dataset). However, we do not have any night-time images annotated with people. This is a similar setup as in Section 4.2, but here we consider a well-known object detection architecture.

As a baseline, we train a Faster-RCNN with ResNet-50 backbone initialized from COCO detection task with the day annotations only (setup (a)). Then, we use the same augmentation and person placement strategy as described in Section 4.3, and retrain the Faster-RCNN detector using the augmented training dataset. The results are summarized in Table 4 and show that our method (f) performs the best, outperforming the baseline (a) by 16%, the-state-of-the-art person generative network (b) by $\sim$ 12%, the compositing approach (c) by $\sim$ 8%, and the nearest competitor (e), which uses much more images depicting people in various poses but with limited appearance variation, by $\sim$ 3% measured by the LAMR (reasonable setup). These results indicate that it is important to have variability and control over (i) the appearance, (ii) the pose, and (iii) the background scene of the generated people for person detection in automotive scenarios. The complete set of the quantitative results is in the supplementary material (Vobecký et al. 2020).

5 Conclusion

We have developed an approach for controlled augmentation of person image datasets, where people with adjustable pose and appearance can be synthesized into real urban backgrounds. We have demonstrated that adding such generated data improves person classification and detection performance, especially in low-data regimes and in challenging conditions (like the night-time person detection) when positive training samples are not easy to collect. We have shown that neural networks of various model complexities designed for multiple tasks benefit from artificially generated samples, especially when we have control over the data distributions. These results open up the possibility to control many more parameters (e.g., weather conditions, age, gender, etc.) and make a step towards controllable training and validation process where generated dummies cover challenging corner cases that are hard to collect in real-world situations.

Acknowledgements

This work was supported by Valeo, the Grant Agency of the Czech Technical University in Prague, grant No. SGS18/205/OHK3/3T/37, and the European Regional Development Fund under the project IMPACT (reg. no. CZ.02.1.01/0.0/0.0/15_003/0000468)

References

Alhaija, H. A.; Mustikovela, S. K.; Mescheder, L. M.; Geiger, A.; and Rother, C. 2018. Augmented Reality Meets Computer Vision: Efficient Data Generation for Urban Driving Scenes. *IJCV* .

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein GAN. *CoRR* abs/1701.07875.

Balakrishnan, G.; Zhao, A.; Dalca, A. V.; Durand, F.; and Guttag, J. 2018. Synthesizing Images of Humans in Unseen Poses. In *CVPR*.

Cao, Z.; Hidalgo, G.; Simon, T.; Wei, S.; and Sheikh, Y. 2018. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *CoRR* .

Catalin Ionescu, Fuxin Li, C. S. 2011. Latent Structured Models for Human Pose Estimation. In *ICCV*.

Chan, C.; Ginosar, S.; Zhou, T.; and Efros, A. A. 2019. Everybody dance now. In *ICCV*.

Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; and Adam, H. 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *ECCV*.

Chen, W.; Wang, H.; Li, Y.; Su, H.; Wang, Z.; Tu, C.; Lischinski, D.; Cohen-Or, D.; and Chen, B. 2016. Synthesizing training images for boosting human 3d pose estimation. In *3DV*.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and Choo, J. 2018. StarGAN: Unified Generative Adversarial Networks for Multi-Domain Image-to-Image Translation. In *CVPR*.

Dollar, P.; Wojek, C.; Schiele, B.; and Perona, P. 2011. Pedestrian detection: An evaluation of the state of the art. *TPAMI* .

Donahue, J.; Krähenbühl, P.; and Darrell, T. 2016. Adversarial feature learning. In *arXiv:1605.09782*.

Dong, H.; Liang, X.; Gong, K.; Lai, H.; Zhu, J.; and Yin, J. 2018. Soft-Gated Warping-GAN for Pose-Guided Person Image Synthesis. In *NeurIPS*.

Dosovitskiy, A.; Fischer, P.; Ilg, E.; Häusser, P.; Hazirbas, C.; Golkov, V.; van der Smagt, P.; Cremers, D.; and Brox, T. 2015. FlowNet: Learning Optical Flow with Convolutional Networks. In *ICCV*.

Dumoulin, V.; Belghazi, I.; Poole, B.; Lamb, A.; Arjovsky, M.; Mastropietro, O.; and Courville, A. 2016. Adversarially learned inference. In *arXiv:1606.00704*.

Dwibedi, D.; Misra, I.; and Hebert, M. 2017. Cut, Paste and Learn: Surprisingly Easy Synthesis for Instance Detection. In *ICCV*.

Georgakis, G.; Mousavian, A.; Berg, A. C.; and Kosecka, J. 2017. Synthesizing Training Data for Object Detection in Indoor Scenes. *CoRR* abs/1702.07836.

Goodfellow, I. 2016. Nips 2016 tutorial: Generative adversarial networks. In *arXiv preprint arXiv:1701.00160*.

Gulrajani, I.; Ahmed, F.; Arjovsky, M.; Dumoulin, V.; and Courville, A. C. 2017. Improved Training of Wasserstein GANs. In *NIPS*.

Hattori, H.; Lee, N.; Boddeti, V. N.; Beainy, F.; Kitani, K. M.; and Kanade, T. 2018. Synthesizing a Scene-Specific Pedestrian Detector and Pose Estimator for Static Video Surveillance - Can We Learn Pedestrian Detectors and Pose Estimators Without Real Data? *IJCV* .

Huang, S.; and Ramanan, D. 2017. Expecting the unexpected: Training detectors for unusual pedestrians with adversarial imposters. In *CVPR*.

Huang, X.; Liu, M.-Y.; Belongie, S.; and Kautz, J. 2018. Multimodal unsupervised image-to-image translation. In *ECCV*.

Ionescu, C.; Papava, D.; Olaru, V.; and Sminchisescu, C. 2014. Human3.6M: Large Scale Datasets and Predictive Methods for 3D Human Sensing in Natural Environments. *TPAMI* .

Isola, P.; Zhu, J.; Zhou, T.; and Efros, A. A. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *ICLR*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *CVPR*.

Karsch, K.; Hedau, V.; Forsyth, D.; and Hoiem, D. 2011. Rendering Synthetic Objects into Legacy Photographs. In *SIGGRAPH Asia*.

Kholgade, N.; Simon, T.; Efros, A.; and Sheikh, Y. 2014. 3D Object Manipulation in a Single Photograph Using Stock 3D Models. *ACM Trans. Graph.* .

Lalonde, J.-F.; Hoiem, D.; Efros, A. A.; Rother, C.; Winn, J.; and Criminisi, A. 2007. Photo Clip Art. *SIGGRAPH 2007* .

Lee, D.; Liu, S.; Gu, J.; Liu, M.-Y.; Yang, M.-H.; and Kautz, J. 2018. Context-aware synthesis and placement of object instances. In *NeurIPS*.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *ECCV*.

Lin, Z.; Khetan, A.; Fanti, G.; and Oh, S. 2018. PacGAN: The power of two samples in generative adversarial networks. In *NIPS*.

Liu, L.; Muelly, M.; Deng, J.; Pfister, T.; and Li, L.-J. 2019a. Generative Modeling for Small-Data Object Detection. In *ICCV*.

Liu, M.-Y.; Breuel, T.; and Kautz, J. 2017. Unsupervised Image-to-Image Translation Networks. In *NIPS*.

Liu, W.; Liao, S.; Ren, W.; Hu, W.; and Yu, Y. 2019b. High-Level Semantic Feature Detection: A New Perspective for Pedestrian Detection. In *CVPR*.

Loing, V.; Marlet, R.; and Aubry, M. 2018. Virtual training for a real application: Accurate object-robot relative localization without calibration. *IJCV* .

Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; and Gool, L. V. 2017. Pose Guided Person Image Generation. In *NIPS*.

Ma, L.; Sun, Q.; Georgoulis, S.; Van Gool, L.; Schiele, B.; and Fritz, M. 2018. Disentangled person image generation. In *CVPR*.

Marín, J.; Vázquez, D.; Gerónimo, D.; and López, A. M. 2010. Learning appearance in virtual scenarios for pedestrian detection. In *CVPR*.

Men, Y.; Mao, Y.; Jiang, Y.; Ma, W.; and Lian, Z. 2020. Controllable Person Image Synthesis With Attribute-Decomposed GAN. In *CVPR*.

Metz, L.; Poole, B.; Pfau, D.; and SoChl-Dickstein, J. 2016. Unrolled generative adversarial networks. In *ArXiv:1611.02163*.

Mirza, M.; and Osindero, S. 2014. Conditional Generative Adversarial Nets. *CoRR* abs/1411.1784.

Neumann, L.; Karg, M.; Zhang, S.; Scharfenberger, C.; Piegert, E.; Mistr, S.; Prokofyeva, O.; Thiel, R.; Vedaldi, A.; Zisserman, A.; and Schiele, B. 2018. NightOwls: A Pedestrians at Night Dataset. In *ACCV*.

Nowozin, S.; Cseke, B.; and Tomioka, R. 2016. f-GAN: Training Generative Neural Samplers using Variational Divergence Minimization. In *NIPS*.

Park, T.; Liu, M.-Y.; Wang, T.-C.; and Zhu, J.-Y. 2019. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*.

Pishchulin, L.; Wuhrer, S.; Helten, T.; Theobalt, C.; and Schiele, B. 2017. Building statistical shape spaces for 3d human modeling. *Pattern Recognition* .

Pumarola, A.; Agudo, A.; Martínez, A. M.; Sanfeliu, A.; and Moreno-Noguer, F. 2018. GANimation: Anatomically-Aware Facial Animation from a Single Image. In *ECCV*.

Radford, A.; Metz, L.; and Chintala, S. 2015. Unsupervised representation learning with deep convolutional generative adversarial networks. In *arXiv preprint arXiv:1511.06434*.

Real, E.; Shlens, J.; Mazzocchi, S.; Pan, X.; and Vanhoucke, V. 2017. YouTube-BoundingBoxes: A Large High-Precision Human-Annotated Data Set for Object Detection in Video. In *CVPR*.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative adversarial text to image synthesis. In *arXiv preprint arXiv:1605.05396*.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved Techniques for Training GANs. In *NIPS*.

Shaham, T. R.; Dekel, T.; and Michaeli, T. 2019. SinGAN: Learning a Generative Model from a Single Natural Image. In *ICCV*.

Si, C.; Wang, W.; Wang, L.; and Tan, T. 2018. Multistage Adversarial Losses for Pose-Based Human Image Synthesis. In *CVPR*.

Siarohin, A.; Sangineto, E.; Lathuiliere, S.; and Sebe, N. 2018. Deformable gans for pose-based human image generation. In *CVPR*.

Simonyan, K.; and Zisserman, A. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *ICLR*.

Thies, J.; Zollhofer, M.; Stamminger, M.; Theobalt, C.; and Nießner, M. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *CVPR*.

Tobin, J.; Fong, R.; Ray, A.; Schneider, J.; Zaremba, W.; and Abbeel, P. 2017. Domain randomization for transferring deep neural networks from simulation to the real world. In *IROS*.

Varol, G.; Romero, J.; Martin, X.; Mahmood, N.; Black, M. J.; Laptev, I.; and Schmid, C. 2017. Learning from Synthetic Humans. In *CVPR*.

Vobecký, A.; Hurych, D.; Uřičář, M.; Pérez, P.; and Sivic, J. 2020. Extended verison of the paper with supplementary material. *arXiv* abs/2012.08274. URL http://arxiv.org/abs/2012.08274.

Wang, T.-C.; Liu, M.-Y.; Zhu, J.-Y.; Tao, A.; Kautz, J.; and Catanzaro, B. 2018. High-resolution Image Synthesis and Semantic Manipulation With Conditional GANs. In *CVPR*.

Wu, J.; Peng, Y.; Zheng, C.; Hao, Z.; and Zhang, J. 2019a. PMC-GANs: Generating Multi-Scale High-Quality Pedestrian with Multimodal Cascaded GANs. *CoRR* abs/1912.12799.

Wu, S.; Lin, S.; Wu, W.; Azzam, M.; and Wong, H.-S. 2019b. Semi-Supervised Pedestrian Instance Synthesis and Detection With Mutual Reinforcement. In *ICCV*.

Xiao, C.; Zhong, P.; and Zheng, C. 2018. Bourgan: Generative networks with metric embeddings. In *NIPS*.

Zanfir, M.; Oneata, E.; Popa, A.; Zanfir, A.; and Sminchisescu, C. 2020. Human Synthesis and Scene Compositing. In *AAAI*.

Zhang, S.; Benenson, R.; and Schiele, B. 2017. Citypersons: A diverse dataset for pedestrian detection. In *CVPR*.

Zhou, S.; Fu, H.; Liu, L.; Cohen-Or, D.; and Han, X. 2010. Parametric reshaping of human bodies in images. *ACM Transactions on Graphics (TOG)* 29(4).

Zhu, J.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *ICCV*.