Advanced Pedestrian Dataset Augmentation for Autonomous Driving

Antonín Vobecký[1,2], Michal Uřičář[3], David Hurych [2], Radoslav Škoviera [1]
[1]CIIRC - Czech Technical University in Prague; [2]Valeo.ai; [3]Valeo R&D Prague
vobecant@fel.cvut.cz, michal.uricar@valeo.com, david.hurych@valeo.com, radoslav.skoviera@cvut.cz

Abstract

Having the ability of generating people images in arbitrary, yet admissible, pose is a crucial prerequisite for Autonomous Driving applications. Firstly, because the existing datasets are quite limited in the human pose variation and appearance. Secondly, because the strict safety requirements call for the ability of validation on rare situations. Generating realistically looking people images is very challenging problem due to various transformations of individual body parts [2, 6] self occlusions etc. We propose a novel approach for person image generation. Our approach allows generating people images in a required pose, indicated by specific pose keypoints and deals with occlusions. We build on top of the recent prevailing success of Generative Adversarial Networks [10]. Our contributions comprise of the networks architecture, as well as the novel loss terms specifically designed to generate visually appealing pedestrians fitting the surrounding environment well.

1. Introduction

Good quality data are the basis of every system based on statistical machine learning and should be a representative sample of the world. Usually, we do not get such good quality data due to limited resources for capturing and annotation, or inability to identify all the use cases in advance.

The need for diverse high-quality datasets is essential in the automotive industry due to the high expectations to handle complex situations under all conditions (weather, daytime, glare, etc.) due to a high safety requirement and reliability of automated driving assistance systems.

In this paper, we present a method for controlled dataset augmentation by synthesizing person images and merging with the background by a specially trained Generative Adversarial Network (GAN), see Figure 1. We aim at images in the wild. There are no assumptions on the background, visibility of body parts to be generated, the number of people present in images (at training time nor inference) or lighting conditions. Our generator is able to produce luminance consistent people images fitting the background that do not break the overall contrast and brightness.

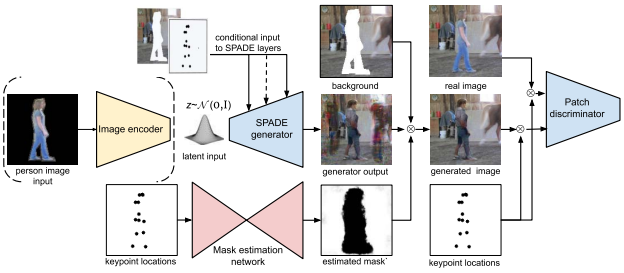

Figure 1: Person image generating framework.

Our contribution is as follows: 1) Novel GAN topology and loss term, 2) From a background crop and pose keypoints we generate a person image and realistically merge it into the background, and: a) generate a person at the place of original one, which is very handy nowadays because of the recent changes due to GDPR rules or b) generate realistic people painting into an arbitrary image with complex occlusions (even non-compact polygons). This is beneficial, e.g., for generating partially occluded pedestrians, that are typically problematic in a pedestrian detection task.

2. Problem Formulation and Related Work

We build on top of GANs, as proposed by [9] and followed by many authors [1, 18, 30, 7, 22, 33, 5, 26, 28, 21]. In our setup, beside a latent vector, we provide as input also the background image and specific pose key points.

In [14], they propose to grow both the generator and discriminator progressively, starting from a very coarse resolution up to super-resolution patches. The core idea is that the networks are trained to capture finer and finer details via progressively increasing complexity of the network architecture. We make use of this method in our approach.

A novel normalization layer called SPADE working with a given semantic map input is proposed in [25]. This layer applies spatially varying affine transformation. We adopt this principle and use it in the generator, see section 3. The most relevant works in the field of person image synthesis are [6, 2, 20, 19, 8, 24, 3, 15]. We highlight the main ideas

and key differences to our work in the following section.

In [19], they train one generator to get a coarse blurry image that is refined by the second generator. In contrast to our solution, their approach requires a person image as an input. Our approach uses the proposed mask just to blend the generated image to the original one.

Soft-gated Warping-GAN [6] proposes a two-step approach. Similarly to us, they enforce the foreground generator to generate the person silhouette (mask) as a part product of the generator. The disadvantage of their approach is that it deforms the background image behind the person and renders the whole image.

In [2], the authors use segmentation of the condition image into several foreground layers. At the input of their foreground generator, there are already masked out target poses aligned human body parts and target pose keypoints. In our case, the input consists of a sampled latent vector, target pose keypoints, and background image only. Our approach solves a more general problem as it does not impose any limitations on the number of object parts and their segmentation, and does not explicitly model the spatial transformations. Hence it can be used for generating partially occluded people, which can be beneficial in generating challenging samples.

The U-Net architecture [29] was adopted as a part of the overall topology by several authors [20, 19, 8] including us. In [20] for a person image at the input they give the user control of the type of generated change - (a) foreground, (b) background and (c) person pose in any combination.

The quality of results of every machine learning system strongly depends on amounts and quality of training data. We use 7 380 images (128×128) from MS COCO dataset, see Section 4. That is considerably less than in [15] they use (14 411 samples, 256×256) and in [6] (378 352 samples, 256×256) from *fashion datasets*. We did not use the fashion datasets because: (1) pose keypoints are not available, (2) none of them have variable and realistic background which we need for our method for merging and luminance consistency, (3) fashion datasets have a strong bias towards women images (e.g., [15] generates women images only) and towards specic frontal poses with completely unsatisfactory pose variation.

The authors of [16] train the network to insert an object instance into an image in a semantically coherent manner. For this task, they propose an end-to-end trainable network that consists of two generative modules. One module determines *where* the inserted object mask should be (i.e., location and scale) and the other determines *what* the object mask shape should look like. This network can be used in our pipeline to predict the locations of pedestrians.

3. Pedestrian Generating GAN

The overall scheme of our framework is depicted in Figure 1. All 4 networks in our person generation pipeline are trained in an adversarial manner. For the results of person synthesis in the wild, see Figure 3.

Person Style Encoder, yellow in Figure 1, takes person images with masked-out background on the input. It is implemented as a CNN that produces the mean and the variance of the output distribution $\mathcal{N}(\mathbf{0}, \mathbf{1})$ that is enforced by the Kullback-Leibler (KL) divergence in the loss function. It is depicted in detail in supplementary materials.

The Mask Estimation Network, red in Figure 1, is used for predicting which pixels should be taken from the generated images and inserted to the original one. The input to this network consists of keypoint locations that are encoded in 17 channels (one channel per key point), see Figure 2. This network has a U-Net architecture [29], and the idea of progressive growing [14] is applied to the decoder. The output of this network is a one-channel mask $\mathcal{M} \in [0,1]^{H \times W}$, where (H, W) are the height and the width of the resulting image, respectively. Having the values in the range of $[0, 1]$ brings the possibility to control the borders of the synthesized person better, and results in more realistic-looking images than in the case of a simple crop by the original mask. Once the mask $\mathcal{M}$ is computed, it is used for merging the generated image I_G and the input image I_{IN} into the resulting I_{RES}. The resulting image is composed as

$$I_{RES} = \mathcal{M} \odot I_G + (I - \mathcal{M}) \odot I_{IN} \quad (1)$$

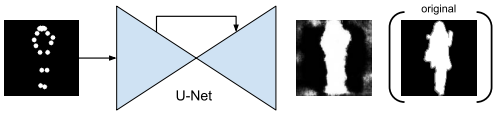

Figure 2: Keypoints are passed through the U-Net to generate the mask used for blending. Original mask shown for comparison only.

The estimated mask does not precisely match the original one since it does not have any information about clothing, carried objects, or occlusions. The mask generator is trained simultaneously with the complete framework. The original mask is not used in training at all, and there is no term in the loss function enforcing its reconstruction.

SPADE Generator Topology The SPADE Generator combines the SPADE residual block proposed in [25] with the idea of progressive growing. It is depicted in detail in the appendix. As an input it takes a latent vector $\mathbf{z} \in \mathbb{R}^d$ sampled from $\mathcal{N}(\mathbf{0}, \mathbb{I})$ that is either randomly drawn from the distribution or obtained with the use of the image encoder. This allows us to have better control over the generated person appearance since the generation can be guided by this encoder. This input is reshaped, passed through

Figure 3: Generated people in the wild by our method.

fully-connected layer and further to the convolutional layers. Then it applies n SPADE residual blocks, each followed by a bilinear upsampling. This allows us to perform progressive growing with the resulting image size of $s = 4 \cdot 2^n$. After these n upsampling blocks, the N resulting feature maps are passed to a 3×3 convolutional layer followed by a hyperbolic tangent activation that maps it from N to 3 channels that correspond to RGB color channels.

The topology of the SPADE residual block is similar to the one proposed in [25]. However, we differ in the input, which consists of keypoint locations and the background.

Patch Discriminator Topology We leverage the idea of the patch discriminator that was proposed in [12]. It restricts the discriminator attention to the structure in local image patches only and aims to classify each $M \times M$ image patch. This discriminator is run across the image concatenated with keypoint locations and the responses are averaged on the output.

Training Algorithm and Proposed Loss Term The topology of the whole framework capturing the training procedure is shown in Figure 1. SPADE generator, patch discriminator, and mask estimation networks are trained progressively from the resolution of 8×8 pixels to the final 128×128 pixels (5 upsampling blocks). We use the blending of the new blocks as in [14].

For adversarial training, we use the Improved Wasserstein loss (WGAN-GP) [11] for its stability. In order to enforce the $\mathcal{N}(\mathbf{0}, \mathbf{1})$ distribution generated from person images in the Person Style Encoder, we use the KL divergence.

We propose a novel loss term, which we call **L1 Edge Loss**, that is based on the Local Binary Pattern (LBP) features [23], and that forces the generator to generate samples with a stronger edge structure.

This loss is computed over the masked-out grayscale image of a person. We compute the so-called *soft LBP* features vector, encoding the information about present edges. In the $L1$ Edge loss, we compute the $L1$ distance between this vector and the vector obtained from the generated image. To justify the use of this new loss instead of a standard *L1 loss* we compute the FID scores: 96.15 *L1 loss* and 94.59 for *L1 edge loss*.

The soft LBP feature vector is a non-thresholded version of standard LBPs and is obtained as follows. For every pixel, we obtain a feature vector $f \in [-1, 1]^8$, corresponding to the image gradient in that pixel. This can be efficiently implemented by a convolution with the filter of size $1 \times 8 \times 3 \times 3$ as in [13] that is applied to a grayscale image. The advantage of the $L1$ Edge Loss is that it only keeps the local information about the image gradient and it pays no attention to the color information, unlike the standard identity-preserving losses.

We further use two feature matching losses: the first one compares the features of real and generated samples obtained from the discriminator and the second one from a pre-trained VGG19 network [31].

4. Experiments

COCO dataset For the evaluation of the proposed system, a custom dataset was created from the COCO dataset [17] which contains images of people with pixel-wise annotations and key points – a crucial aspect needed for the pose generation. Only images of standing people of the minimum height of 100 px were kept and scaled to 128×128 px. With larger height, the network can be trained to generate images of higher resolution, but the number of samples in the training set will decrease. Each sample in the new dataset contains: an image with the person masked out; a binary mask extracted from the annotation; a soft mask with values in range $[0, 1]$ estimated from the key points; the key points in the form of a 17-channel tensor (one channel per keypoint). We collected 7380 samples. Examples from the dataset can be found in the appendix.

Cityscapes dataset We collected cropped images of size 128×128 px containing pedestrians from the training and validation split of the Cityscapes dataset [4]. We used the annotations from [34]. The resulting training set contains $19,237$ images of pedestrians. Since the [34] annotations do not contain annotations for the test set, we split the validation set in half. This results in the validation set with $1,926$ samples and in the test set of size $1,925$. Each set contains the same amount of the negative samples that don't contain pedestrians that were cropped randomly from the image such that they don't contain the bounding box of an annotated pedestrian. This dataset was not used for training.

Generated samples We experimented with extending the dataset of real pedestrians with the synthesized people images. First, for each scene in the dataset, we obtained 5 locations that would likely contain a pedestrian by [16]. Then, in these positions, we generated a person in a random pose with our proposed network. To compare the results, we also used [32]. Our method uses only the information about the position from initialization by [16]. The pose keypoints are selected randomly from test annotations. In contrast, [32]

(a) Result by our method.

(b) Result by [32]

Figure 4: Comparison of the resulting scene with inserted pedestrians generated by our method (4a) and by [32] (4b).

Figure 5: Generated samples with occlusions. No post-processing was used.

uses the person silhouette as well and was trained on the Cityscapes dataset while our network was trained on COCO dataset. For comparison, see Figure 4.

Evaluation of the Generated Samples Quality It is complicated to evaluate the *quality* of generated samples as it is hard to quantify and is application dependent. Since our primary goal is to improve the performance of detection algorithms, we did test the performance of the proposed method on such an algorithm first. To evaluate the visual appearance and realism of the generated images, we performed a user study in the form of a survey that can be found in the supplementary materials. We also experimented with augmenting the training set with generated samples. In Figure 5, you may observe generated samples with occlusions.

Person detector score for samples quality evaluation We used the pretrained YOLOv3 [27] as the human detector. The mean IoU was **0.765 for real images** and **0.657 for generated images**. The detector performed better on the original images. However, we claim that the difference is sufficiently small to state that the approach is promising.

Augmenting training set with person generator As our primary goal is to provide means to augment existing datasets, we tested the performance of a CNN classifier trained on a pure dataset vs. a dataset augmented with our method. The classifier has 6729 parameters (realistic size for slow DSP processor in a car) and is trained *from scratch*.

We test the influence of the augmentation based on the size of both the baseline dataset (with real images) and the synthesized dataset (generated images). We use $10, 20, 35, 50, 75$ and 100 percent of the samples from the cityscapes dataset. We compare the training with samples generated by our system and with samples generated by [32] as described in 4. Note that our method is trained on the samples from the different dataset in contrast to [32] which is trained on the samples from the complete baseline dataset. Thus, using [32] may better capture the distribution of the samples in the test dataset.

The test set includes original Cityscapes images only. To suppress the stochasticity of the results, we run each experiment 5 times. The results show that with the small number of real samples, the generated samples bring a considerable further boost in performance – in the case of having just 10% of the baseline dataset, our method improved the average test set accuracy by 4.9% from 77.0% to 81.9% (vs to 81.1% with [32]). For 20% of the baseline dataset, the test set accuracy improved by 3.9% from 78.9% to 82.8% with our method (82.5% with [32]). From these results, one can see that extending the dataset with artificial data in the case of a small number of samples leads to a further boost in the accuracy. Note that our method, that is trained on a different than the baseline dataset increases the performance more than [32] which is trained on the complete baseline dataset. For complete results, please see the supplementary.

5. Conclusion

We have proposed a novel approach for person image generation based on GANs that consists of novel network architecture and novel loss terms specifically designed to generate visually appealing person fitting in the surrounding environment and handles occlusions which are not handled by competitive methods. The generation process involves a specification of a person's pose via its keypoints. The target application of the proposed algorithm is in the automotive industry, where it serves to fill the gap of insufficient training and validation examples for a pedestrian detector, via means of the advanced data augmentation. The experimental evaluation shows that even though the visual quality of the generated person instances still has some limitations, it is competitive with state of the art and provides a dataset augmentation that improves the human detector accuracy.

Acknowledgement This work was supported by the European Regional Development Fund under the project IMPACT (reg. no. CZ.02.1.01/0.0/0.0/15_003/0000468) and Robotics for Industry 4.0 (reg. no. CZ.02.1.01/0.0/0.0/15_003/0000470).

References

[1] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein GAN. *CoRR*, abs/1701.07875, 2017. 1

[2] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *CVPR*, 2018. 1, 2

[3] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody dance now. *CoRR*, abs/1808.07371. 1

[4] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 3

[5] J. Donahue, P. Krhenbhl, and T. Darrell. Adversarial feature learning. In *arXiv preprint arXiv:1605.09782*, 2016. 1

[6] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *NeurIPS*, 2018. 1, 2

[7] V. Dumoulin, I. Belghazi, B. Poole, A. Lamb, M. Arjovsky, O. Mastropietro, and A. Courville. Adversarially learned inference. In *arXiv preprint arXiv:1606.00704*, 2016. 1

[8] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, pages 8857–8866. IEEE Computer Society, 2018. 1, 2

[9] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. In *arXiv preprint arXiv:1701.00160*, 2016. 1

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27*, pages 2672–2680. Curran Associates, Inc., 2014. 1

[11] Ishaan Gulrajani, Faruk Ahmed, Martín Arjovsky, Vincent Dumoulin, and Aaron C. Courville. Improved Training of Wasserstein GANs. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5769–5779, 2017. 3

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *CVPR*, pages 5967–5976. IEEE Computer Society, 2017. 3

[13] Felix Juefei-Xu, Vishnu Naresh Boddeti, and Marios Savvides. Local binary convolutional neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. 3

[14] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 1, 2, 3

[15] Christoph Lassner, Gerard Pons-Moll, and Peter V. Gehler. A generative model of people in clothing. In *ICCV*, pages 853–862, 2017. 1, 2

[16] Donghoon Lee, Sifei Liu, Jinwei Gu, Ming-Yu Liu, Ming-Hsuan Yang, and Jan Kautz. Context-aware synthesis and placement of object instances. In *Advances in Neural Information Processing Systems*, pages 10393–10403, 2018. 2, 3

[17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 3

[18] Zinan Lin, Ashish Khetan, Giulia Fanti, and Sewoong Oh. PacGAN: The power of two samples in generative adversarial networks. In *arXiv preprint arXiv:1712.04086*, 2018. 1

[19] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *NIPS*, 2017. 1, 2

[20] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. *CoRR*, abs/1712.02621, 2017. 1, 2

[21] Luke Metz, Ben Poole, David Pfau, and Jascha Sohl-Dickstein. Unrolled generative adversarial networks. In *arXiv preprint arXiv:1611.02163*, 2016. 1

[22] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *NIPS*, 2016. 1

[23] Timo Ojala, Matti Pietikäinen, and David Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *ICPR (1)*, pages 582–585. IEEE, 1994. 3

[24] Xi Ouyang, Yu Cheng, Yifan Jiang, Chun-Liang Li, and Pan Zhou. Pedestrian-synthesis-gan: Generating pedestrian data in real scene and beyond. *CoRR*, 2018. 1

[25] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. *CoRR*, abs/1903.07291, 2019. 1, 2, 3

[26] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *arXiv preprint arXiv:1511.06434*, 2015. 1

[27] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. 4

[28] S. Reed, Z. Akata, X. Yan, L. Logeswaran, B. Schiele, and H. Lee. Generative adversarial text to image synthesis. In *arXiv preprint arXiv:1605.05396*, 2016. 1

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 2

[30] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, pages 2234–2242, 2016. 1

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 3

[32] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution Image Synthesis and Semantic Manipulation With Conditional

GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 3, 4

[33] Chang Xiao, Peilin Zhong, and Changxi Zheng. BourGAN: Generative Networks with Metric Embeddings. In *arXiv preprint arXiv:1805.07674*, 2018. 1

[34] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3213–3221, 2017. 3