Pedestrian Image Generation for Self-driving Cars

Saeed Saadatnejad

Alexandre Alahi

Visual Intelligence for Transportation (VITA), EPFL

May 2019

STRC

19th Swiss Transport Research Conference

Monte Verità / Ascona, May 15 – 17, 2019

Visual Intelligence for Transportation (VITA), EPFL

Pedestrian Image Generation for Self-driving Cars

Saeed Saadatnejad, Alexandre Alahi
Visual Intelligence for Transportation Lab (VITA)
Ecole Polytechnique Federale de Lausanne (EPFL)
Route Cantonale, 1015 Lausanne, Switzerland
phone: +41-21-693 08 94
fax: +41-21-693 26 08
{firstname.lastname}@epfl.ch

May 2019

Abstract

Pedestrian image generation in the desired pose can be used in a wide range of applications e.g., person re-identification and tracking which are among the fundamental challenges in self-driving cars. This is a hard task because it should be invariant to a set of nuisances such as body poses, illuminations, or changes in camera viewpoint. In this work, we want to study the task of synthesizing a latent canonical view of a pedestrian that will potentially be robust to the mentioned factors of nuisances. Our goal is to generate the unique frontalized view of a pedestrian observed in the wild. The generated image should visually be the same regardless of the body pose. We propose a new generative framework that goes beyond the 1 to 1 supervision commonly used. We propose to jointly reason on multiple inputs and outputs thanks to a carefully chosen loss function acting as a regularizer. Our experiments show the benefits of our framework on challenging low-resolution datasets.

Keywords

Image generation, Generative Adversarial Networks, Self-driving cars

Figure 1: The motivation of the proposed method. Given an image in an arbitrary pose, we will be able to synthesize that person in front view

1 Introduction

Thanks to the advancements in generative models, realistic image synthesis is not a dream anymore. Synthesizing images especially in a desired pose or shape can have various applications. For instance, in autonomous cars, one of the main challenges is the re-identification of pedestrians needed for tracking. In pedestrian protection systems, pedestrian detection and tracking is a crucial part which has attracted lots of studies in recent years (Guo *et al.,* 2016). The main challenge in these systems is the nuisances in the images such as pose change, illumination variances, and occlusion. Providing a stable unique representation makes the tracking problem trivial.

The goal of this paper is to generate stable realistic Canonical view images of pedestrians from images in different poses and camera views. The model should be able to synthesize an image of a person in a frontal view pose based on another image of that person in an arbitrary pose and a fixed frontal pose. An example could be found in Figure 1. In the left side, the input images in diverse views can be seen. The picture in the right should be the output of the model for all set of images as the frontal view of a person is unique.

Pioneering work in image generation has successfully tackled some of the challenges in image synthesis. Karras *et al.* (2018) designed a GAN for face image synthesis of humans which led to generating highly realistic images. Wang *et al.* (2018) showed acceptable results in synthesizing realistic city scene images from semantic label maps. They leveraged conditional

generative adversarial network (GAN) and defined a feature matching loss, but their results in pedestrian image synthesis are not satisfactory. Tran *et al.* (2017) announced another GAN to disentangle the features of face pose and appearance. Although it achieved good results in face frontalization, there is not any similar work in learning a representation which is useful for shape (pose) frontalization.

Although previous works have achieved great success in the task of image synthesis, none of them try to generate images in a specific pose. As a result, their synthesized images in a frontalized pose lack stabilization. Besides stabilization, the unsupervised fashion of the task is a challenging problem since the target (canonical view) of a pedestrian is not always provided i.e. the model should be trained in the absence of some paired examples.

In this paper, we propose a model based on GANs which generates stable frontalized images. As a unique image is aimed, the model jointly reasons on multiple images. To do this, we change the loss function of the adversarial training by adding appropriate terms to keep the style of all synthesized images of a person similar to each other. It leads to learning a robust representation per person. Note that our trained model is not limited to canonical view image synthesis and this learned nuisance-invariant representation will help to synthesize images in any desired pose.

The results show that our method has comparative results compared to the state of the art in image synthesis qualitatively and especially it synthesizes stable images for different images out of a person. Also, the learned representation is useful in other tasks such as person re-identification.

Related Works

Most common recent deep learning methods for generating images use Generative Adversarial Networks (GANs) (Goodfellow *et al.,* 2014) and Variational AutoEncoders (VAEs) (Kingma and Welling, 2014). GANs have two adversarial networks (a generator and a discriminator) which each of them tries to beat each other and in this process, the generator learns to generate more realistic images and discriminator becomes capable of detecting between real and fake images. VAEs are another type of generative models that work based on probabilistic graphical models.

There are a lot of works in generating realistic images (Karras *et al.* (2018), Isola *et al.* (2017)). They do not obey any specific shape and have limited performance in pose-based image genera-

tion. A few of them like Liu *et al.* (2018) used these generated set as an augmented dataset for other tasks such as person re-identification.

Among papers that addressed the problem of people image generation based on a specific pose, some used VAEs (Esser *et al.,* 2018) and most of them leveraged GANs (Liqian Ma and Gool, 2017, Ma *et al.,* 2018, Siarohin *et al.,* 2018, Pumarola *et al.,* 2018, Dong *et al.,* 2018). Ma *et al.* (2018) and Liqian Ma and Gool (2017) disentangle background and the person images, then transform the person images to the new pose and try to have a background close to the source image. Pumarola *et al.* (2018) tried to solve the problem by GAN in an unsupervised training leveraging a pose conditioned bidirectional generator. Siarohin *et al.* (2018) introduced deformable skip connections in the generator of their GAN and used the nearest neighbor loss which resulted in more realistic image synthesis. In person image synthesis based on a specific pose, sometimes large geometric transformation is seen. Dong *et al.* (2018) designed a soft-gated Warping GAN to address that. Esser *et al.* (2018) defined a new variational u-net for shape transformation and a VAE for appearance transformation. However, all of those methods do not have enough good qualities and they are not generalizable i.e. their method and their learned representation cannot be used for other tasks like person re-identification.

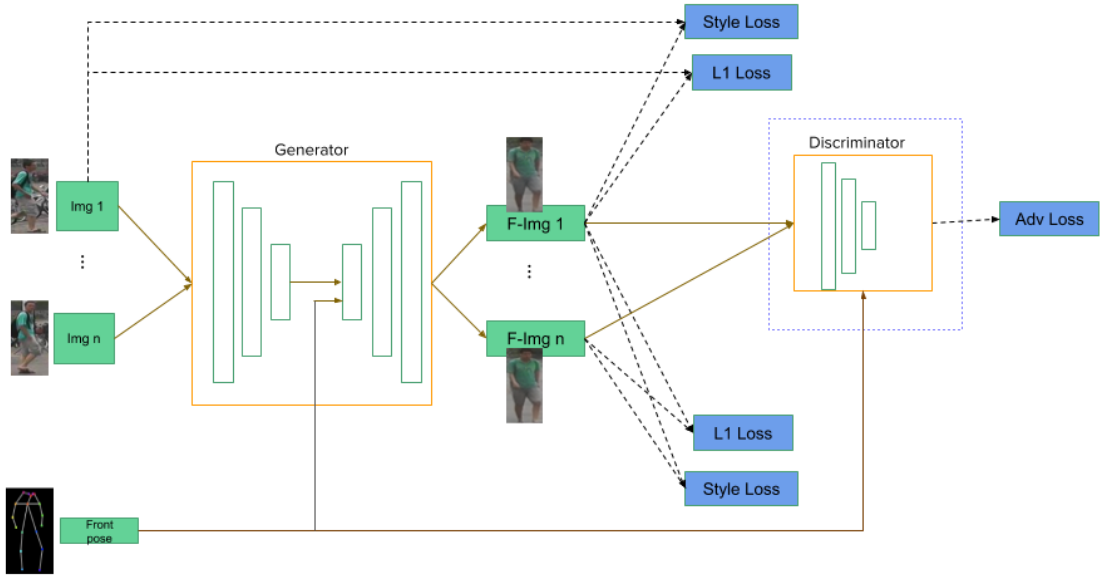
There are also some other papers like (Ge *et al.,* 2018) which designed a method to generate images and used its learned features for person re-identification task. Liu *et al.* (2018) and Qian *et al.* (2018) also addressed the problem of person re-identification but all of them could not generate visually good images. Besides, their results do not show a huge boost in performance.

In this project, we try to learn a latent representation of pedestrians by which we can synthesize that person in any arbitrary pose and especially in front view pose. For instance, consider a fixed camera taking pictures out of a pedestrian walking which results in images from various views. Using the latent representation, the stable pedestrian image in front view can be synthesized even when there is an occlusion. We use the information of a set of images out of a person to stabilize the model's outputs. As this representation is generalizable, it can be used in other tasks.

2 Method

Figure 2 shows the block diagram of our algorithm. The pose is detected using the open-pose (Cao *et al.,* 2017) pre-trained model. There is also another encoder which extracts image features. These features (image features + poses) are fed to the decoder of the generator. The

Figure 2: The proposed method.



discriminator takes these generated images and real images and classifies them as fake or real. They are trained based on the adversarial training loss with help of training pairs (source and target). It is shown in Eq. (1), where x is the source image, P is the target pose and z is the noise which is added as usual GANs. This loss helps to generate realistic images in the desired pose.

$$\mathcal{L}_{cGAN}(G, D) = \mathbb{E}_{(x)\in\mathcal{X}}[\log D(x, P)] + \mathbb{E}_{x\in\mathcal{X}, z\in\mathcal{Z}}[\log(1 - D(G(z, x, P), P))], \tag{1}$$

In order to generate a latent nuisance invariant view we need to add other supervisions. Hence, we define L1 loss and style loss (Johnson *et al.,* 2016). The L1 loss is described in Eq. (2), where x and $\hat{x}$ are target and generated images, respectively.

$$\mathcal{L}_{1_1}(\hat{x}, x) = \|x - \hat{x}\|_1, \tag{2}$$

The style loss calculates the style similarity of two images by calculating the gram matrix and is defined in Eq. (3).

$$\mathcal{L}_{s1}(\hat{x}, x) = \|\psi(\hat{x})\psi^T(\hat{x}) - \psi(x)\psi^T(x)\|_1, \tag{3}$$

where ψ is the feature output of a feature extractor in the form of $C * HW$ (C, H, and w are equal

to number of channels, height and width of the extracted feature).

In the proposed method, generating images in an unseen pose would become challenging since there is no target image in the canonical view image synthesis. So instead of a set of pairs of target and generated images, we should define a supervision to leverage a set of generated images. Thus those two losses are also applied on the generated images themselves as the unpaired training part. They are responsible for calculating the difference among the generated images only. It is shown in Figure 2 as well. The formulas are described in Eq. (4) and Eq. (5).

$$\mathcal{L}_{1_2}(x) = \|x_1 - x_2\|_1 + \|x_2 - x_3\|_1 + ... + \|x_n - x_1\|_1, \tag{4}$$

$$\begin{aligned}\mathcal{L}_{s2}(\hat{x}) = \|\psi(\hat{x_1})\psi^T(\hat{x_1}) - \psi(\hat{x_2})\psi^T(\hat{x_2})\|_1 + \|\psi(\hat{x_2})\psi^T(\hat{x_2}) - \psi(\hat{x_3})\psi^T(\hat{x_3})\|_1 + ... \\ +\|\psi(\hat{x_n})\psi^T(\hat{x_n}) - \psi(\hat{x_1})\psi^T(\hat{x_1})\|_1,\end{aligned} \tag{5}$$

where $\hat{x}$ is the batch of n generated images including $\hat{x_1}$ ... $\hat{x_n}$.

The objective function is the combination of those losses and is shown in Eq. (6).

$$G^* = \arg\min_G \max_D \mathcal{L}_{cGAN}(G, D) + \lambda_1 \mathcal{L}_{1_1}(G) + \lambda_2 \mathcal{L}_{1_2}(G) + \lambda_3 \mathcal{L}_{s1}(G) + \lambda_4 \mathcal{L}_{s2}(G), \tag{6}$$

We train the discriminator and generator with equal ratios. It is worth mentioning that the discriminator weights are updated only with paired examples in order not to be biased but the back-propagation losses remain in unpaired examples to train the generator.

3 Experiments

3.1 Dataset and Metrics

Market-1501 dataset (Zheng *et al.,* 2015) is used here which contains 32,668 images of 1,501 people with image size 128 * 64 captured from 6 different cameras. This is a challenging dataset because of its low-resolution images and their pose diversity. 751 ids are used for training and the remaining for testing as the standard split (Zheng *et al.,* 2015) suggests.

3.2 Implementation Details

In style loss formula, the output of the first convolution layer of a resnet-50 (He *et al.,* 2015) pretrained on Market-1501 dataset is used as the feature extractor. We train our network for 100 epochs, with Adam optimizer (learning rate: 10e-4 and linear decreasing rate after each 20 epochs) and batch size of 128.

3.3 Image synthesis

The qualitative results of the proposed method are shown in Fig. 3. To have a comparison with the state-of-the-art methods, the ones with available online codes were selected. (Siarohin *et al.,* 2018, Esser *et al.,* 2018, Ge *et al.,* 2018). The figure shows our results are comparable to the state-of-the-art.

One of the advantages of the proposed method is the capability of synthesizing stable images for a person in the desired pose. Different images of two distinct pedestrians are fed to the network and the results are shown in Fig. 4. It generates more stable and distinct images per person compared to the state of the art. It shows that our method learned a meaningful representation that can be used in other tasks.

4 Conclusion and Future Works

In this paper, a generative model was designed that not only generates realistic images but also learns a nuisance-invariant representation for other tasks like person re-identification. This is due to the loss function we defined.

At this phase, the quality of generated images are not very good and it needs some hyper-parameter tuning. As it is shown in the images, the background style is transferred which is not desirable and adds some kind of noise to the generated images. We should mask the body and apply those losses in order to avoid penalizing the transferring of background style. Exploring other supervisions will be the next step, also. Besides, leveraging other datasets can increase the generalizability of the model.

At the moment we do need the visualization to make sure the generator learns to synthesize

Figure 3: Qualitative results on the Market-1501 dataset. Columns 1 represents the source image. The last four columns show the output of different methods and ours

Figure 4: Frontal image synthesis. The network generated images based on the first row and a frontal pose.

images in the desired pose. However, when the desired performance is achieved, the learned representation can be applied directly to other tasks. Thus, we won't need to transfer the representation back to the high dimension which injects the inserted noise.

Acknowledgement

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 754354.

5 References

Cao, Z., T. Simon, S.-E. Wei and Y. Sheikh (2017) Realtime multi-person 2d pose estimation using part affinity fields, paper presented at the *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dong, H., X. Liang, K. Gong, H. Lai, J. Zhu and J. Yin (2018) Soft-gated warping-gan for pose-guided person image synthesis, paper presented at the *Advances in Neural Information Processing Systems*, 474–484.

Esser, P., E. Sutter and B. Ommer (2018) A variational u-net for conditional appearance and shape generation, paper presented at the *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Ge, Y., Z. Li, H. Zhao, G. Yin, S. Yi, X. Wang and h. Li (2018) Fd-gan: Pose-guided feature distilling gan for robust person re-identification, paper presented at the *Advances in Neural Information Processing Systems*, 1222–1233.

Goodfellow, I., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio (2014) Generative adversarial nets, paper presented at the *Advances in neural information processing systems*, 2672–2680.

Guo, L., L. Li, Y. Zhao and Z. Zhao (2016) Pedestrian tracking based on camshift with kalman prediction for autonomous vehicles, *International Journal of Advanced Robotic Systems*, **13** (3) 120.

He, K., X. Zhang, S. Ren and J. Sun (2015) Deep residual learning for image recognition, *arXiv preprint arXiv:1512.03385*.

Isola, P., J.-Y. Zhu, T. Zhou and A. A. Efros (2017) Image-to-image translation with conditional adversarial networks, paper presented at the *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Johnson, J., A. Alahi and L. Fei-Fei (2016) Perceptual losses for real-time style transfer and super-resolution, paper presented at the *European Conference on Computer Vision*.

Karras, T., S. Laine and T. Aila (2018) A style-based generator architecture for generative adversarial networks, *CoRR*, **abs/1812.04948**.

Kingma, D. P. and M. Welling (2014) Auto-encoding variational bayes, paper presented at the *Proceedings of the International Conference on Learning Representations (ICLR)*.

Liqian Ma, Q. S. B. S. T. T., Xu Jia and L. V. Gool (2017) Pose guided person image generation, paper presented at the *Advances in neural information processing systems*.

Liu, J., B. Ni, Y. Yan, P. Zhou, S. Cheng and J. Hu (2018) Pose transferrable person re-identification, paper presented at the *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Ma, L., Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele and M. Fritz (2018) Disentangled person image generation, paper presented at the *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Pumarola, A., A. Agudo, A. Sanfeliu and F. Moreno-Noguer (2018) Unsupervised Person Image Synthesis in Arbitrary Poses, paper presented at the *Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR)*.

Qian, X., Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang and X. Xue (2018) Pose-normalized image generation for person re-identification, paper presented at the *The European Conference on Computer Vision (ECCV)*, September 2018.

Siarohin, A., E. Sangineto, S. LathuiliÃ¨re and N. Sebe (2018) Deformable gans for pose-based human image generation, paper presented at the *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

Tran, L., X. Yin and X. Liu (2017) Disentangled representation learning gan for pose-invariant face recognition, paper presented at the *In Proceeding of IEEE Computer Vision and Pattern Recognition*, Honolulu, HI, July 2017.

Wang, T.-C., M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz and B. Catanzaro (2018) High-resolution image synthesis and semantic manipulation with conditional gans, paper presented at the *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Zheng, L., L. Shen, L. Tian, S. Wang, J. Wang and Q. Tian (2015) Scalable person re-identification: A benchmark, paper presented at the *Computer Vision, IEEE International Conference on*.