Next Steps for Image Synthesis using Semantic Segmentation

Saeed Saadatnejad
Siyuan Li
Taylor Mordan
Alexandre Alahi

Visual Intelligence for Transportation (VITA), EPFL

May 2020

STRC

20th Swiss Transport Research Conference
Monte Verità / Ascona, May 13 – 15, 2020

Visual Intelligence for Transportation (VITA), EPFL

Next Steps for Image Synthesis using Semantic Segmentation

Saeed Saadatnejad, Siyuan Li, Taylor Mordan, 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 2020

Abstract

Image synthesis in the desired semantic can be used in many tasks of self-driving cars giving us the possibility to enhance existing challenging datasets by realistic-looking images which we do not have enough. Our goal is to improve the image quality generated by the conditional Generative Adversarial Network (cGAN). We focus on the class of problems where images are generated given semantic inputs, such as scene segmentation masks or human body poses. To do that, we change the architecture of the discriminator to better guide the generator. The improvements we present are generic and simple enough that any architecture of cGAN can benefit from. Our experiments show the benefits of our framework on different tasks and datasets. In this paper, the preliminary achievements of our study on the discriminator structure are described.

Keywords

Image synthesis, Generative Adversarial Networks, Semantic Segmentation, Self-driving cars

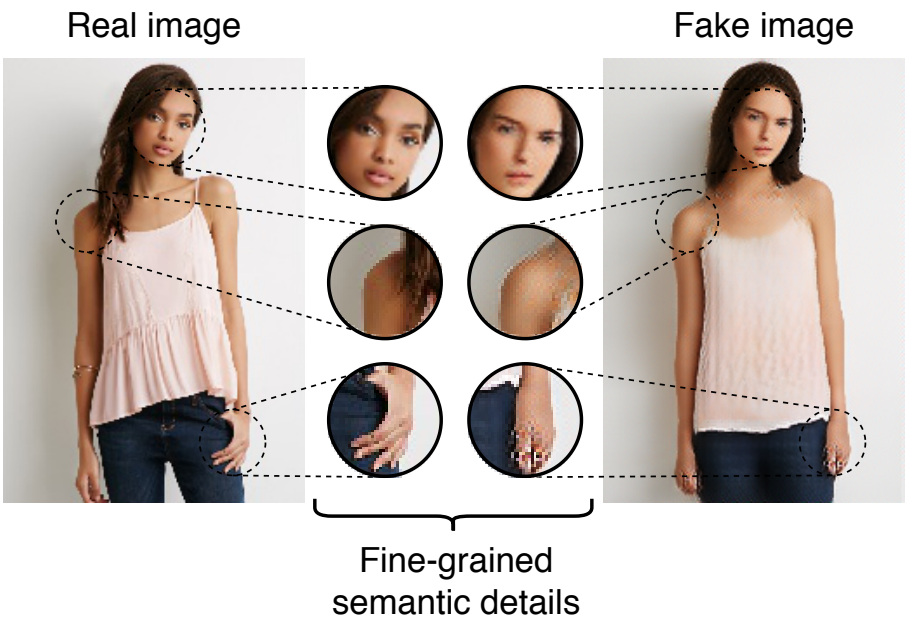

Figure 1: Although a fake image can look realistic from a global view, looking at fine-grained details that carry the semantic content of the scene helps distinguishing between real and fake images.

1. Introduction

To guarantee the safety of self-driving cars, all rare cases should be anticipated in the simulation phase. For instance, synthesizing an image of a kid running in front of the car in the foggy weather is the task of a good simulator. In addition to developing simulators (Wang *et al.,* 2018b), a powerful image synthesis method can be used in augmenting datasets with non-trivial cases as demonstrated in video synthesis (Vondrick *et al.,* 2016) and person re-identification (Liu *et al.,* 2018*,* Zheng *et al.,* 2019).

However, realistic image synthesis is a notoriously difficult task due to a high dimensional output space and an ill-posed objective. In this work, we tackle the semantically-driven image synthesis task: given a semantic mask, *e.g*, human body poses or scene segmentation masks, we aim to generate a realistic image with the same semantics.

Generating data given specific high-level semantics is commonly done with conditional Generative Adversarial Networks (cGANs) (Isola *et al.,* 2017*,* Siarohin *et al.,* 2018). However, state-of-the-art approaches use the structured semantic description only as an input to the generator network. The discriminator of the cGAN is in charge of classifying the whole image as real or synthetic. Yet, it gives the same weight to all regions and do not learn a specialized network for a specific semantic class. For instance, what makes a hand real might be different than what

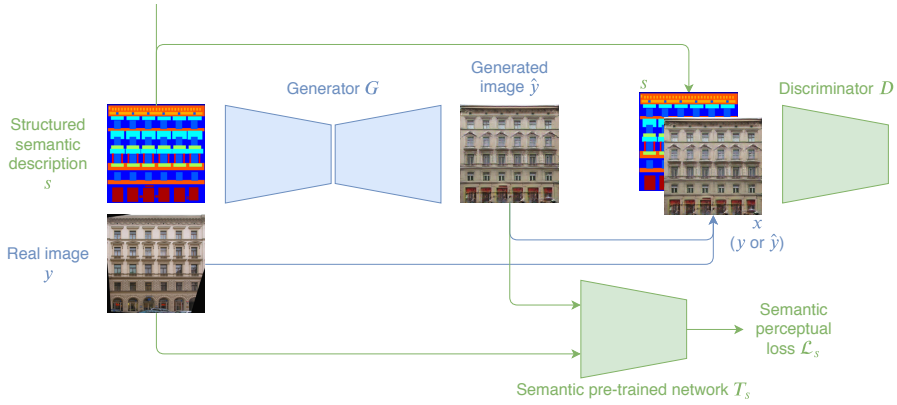

Figure 2: Conditional GAN training with semantic guiding. In addition to the discriminator, a semantic pre-trained network is also leveraged to compute a semantic perceptual loss, matching the given constraint in a suitable space rather than in a pixel one. Visual and semantic information is in blue and green.

makes a left shoulder. Moreover, as illustrated in Figure 1, locations associated with semantic features are arguably some of the most important ones when assessing the realism of an image. For instance, the faces, shoulders, or hands are usually more significant than the middle of the person.

In this paper, we address the aforementioned issues by introducing a simple yet effective modification to the architecture of the cGAN discriminator, further described in Figure 2. The proposed approach better guides the image generator network during learning, yielding more realistic details, hence globally improving the quality of rendered images.

The improvements we present in this paper are generic and simple enough that any architecture of cGAN could benefit from regardless of the specific setup or context considered. Interestingly, as only the discriminator is modified, it should be independent to the particular generator architecture used, and should therefore also be complementary to any approach based on generator enhancement, *e.g*, Siarohin *et al.* (2018), Karras *et al.* (2018b), Park *et al.* (2019), Karras *et al.* (2018a). Since the discriminator is used during training only, it is noticeable that all the changes we apply do not bring any run-time overhead, both in forward time and memory footprint.

2. Related Works

Image generation. Most recent deep learning methods for image synthesis use Generative Adversarial Networks (GANs) (Goodfellow *et al.,* 2014) and Variational AutoEncoders (VAEs)

(Kingma and Welling, 2014). GANs use two separate networks, a generator and a discriminator, that jointly optimize exclusive objectives. In this process, the generator learns to generate more realistic images and the discriminator learns to distinguish between real and fake images more accurately. VAEs are another type of generative models that rely on probabilistic graphical models. Although they have been shown to disentangle features, the generated images are usually not as realistic as those from GANs (Esser *et al.,* 2018). In this paper, we mainly consider GANs.

Several methods have modified the design of the generator of GANs to get better results. Using mapping networks, adaptive instance normalization (AdaIN) (Karras *et al.,* 2018b) and spatially adaptive normalization (SPADE) (Park *et al.,* 2019) are among successful ideas in improving its architecture. These kinds of improvements have recently led to stunning results in generation of natural images (Brock *et al.,* 2019) or human faces (Karras *et al.,* 2018a,b). Moreover, it has been shown that these realistic generated images could be used as data augmentation in other tasks to improve accuracy, *e.g*, in person re-identification (Liu *et al.,* 2018, Zheng *et al.,* 2019).

Conditional image generation. Conditional GANs (cGANs) generate images from other images or high-level descriptions of the desired outputs. Applications can be really various, as exemplified by pix2pix (Isola *et al.,* 2017) which applied the image-to-image translation approach to a wide range of Computer Vision problems. More recently, realistic results have been obtained in generation of city scenes using semantic maps (Wang *et al.,* 2018b) and even talking head videos from few examples (Zakharov *et al.,* 2019).

Conditional human image generation. In spite of realistic results in face image generation, human image synthesis is far from looking real, since images need fine details of all body parts for a synthesized image to be considered as real. The problem becomes harder in conditional human image synthesis, where the model has to preserve the identity and texture of the conditioned image. One major issue is large body deformations caused by people's movements or changes in camera viewpoint. Several ideas have been developed. Pumarola *et al.* (2018) added a pose discriminator, Siarohin *et al.* (2018) introduced deformable skip connections in its generator and used a nearest neighbour loss, Ma *et al.* (2018), Ma *et al.* (2017) disentangled foreground people from background to transform them into the new pose while trying to have a background close to the source image. Dong *et al.* (2018) designed a soft-gated Warping GAN to address the problem of large geometric transformations in human image synthesis. Chan *et al.* (2018) and Wang *et al.* (2018a) trained a personalized model for each person, and Wang *et al.* (2019) leveraged a few-shot learning approach needing few images from a new person to refine the network at test time.

Discriminator in image generation. The architecture of the discriminator plays a role in the quality of generated images, through learning of the generator. Patch-wise discriminators (PatchGAN) have outperformed global ones with full-image receptive fields for both neural style transfer (Li and Wand, 2016) and conditional image generation (Isola *et al.,* 2017). Although the discriminator is often discarded after training, some methods leverage the information it learns. Wang *et al.* (2018b) yielded high-quality images by having multiple discriminators at different resolutions, and Chan *et al.* (2018) used two separate networks for synthesizing full-body and face. Liu *et al.* (2019) improved the quality of generated images and prevents mode collapse by leveraging the information stored in the discriminator and reshaping the loss function of GAN during image synthesis.

3. Method

We address conditional Generative Adversarial Network (cGAN) training for general purpose image synthesis, and include structured semantic information to guide learning to focus more on meaningful regions of images. We build on successful cGAN models Isola *et al.* (2017), Siarohin *et al.* (2018) and propose to add fine-grained details to these areas by (i) biasing the discriminator toward semantic features and (ii) relaxing the strict pixel matching constraints with the semantic perceptual loss, which will subsequently influence the learning of the image generator network in the same way.

Our model is composed of a main network G generating an image $\hat{y} = G(s)$ from a structured semantic description $s = (s_1, \ldots, s_K)$ over K feature maps (*e.g*, class masks or heatmaps of keypoints) of the desired output, as depicted in Figure 2. During learning, examples consist of pairs (y, s) of real images y and their corresponding semantic descriptions s. After training, output images $\hat{y} = G(s)$ are expected to be similar to y as they should share the same underlying semantic structures s. However, it is not easy to handcraft a loss function to assess the quality of the outputs $\hat{y}$ of G. For this, a discriminator network D is concurrently trained with it to act as a proxy loss, both networks competing to optimize exclusive loss functions in an adversarial minimax game Goodfellow *et al.* (2014).

As illustrated in Figure 2, the discriminator D takes as input a tuple (x, s) composed of an image x along with its semantic description s. The image x can either be generated by G (in which case $x = \hat{y}$) or be a real image ($x = y$), and D is trained to identify this, through minimization of a classification loss $\mathcal{L}_D$. At the same time, the generator G learns to generate images that both are realistic and match the input constraints.

In order to generate realistic images, the generator G learns to fool the discriminator D, by maximizing its loss $\mathcal{L}_D$. Usually, the training of cGANs does not leverage all the semantic content of the description s: it only uses it as input to the discriminator D to check whether the image matches with it. We first modify the discriminator network D and its associated loss function $\mathcal{L}_D$, which will impact the training of the generator G.

The second objective to be optimized by the generator, *i.e*, having images matching their semantic descriptions, is usually achieved by training the generator network G with a regression loss, *e.g*a L_1 loss (Isola *et al.,* 2017), but this promotes blurry outputs as the model tries to match pixel intensities directly. We claim that this regression is too strict, since the target images y are only examples of possible images to generate. To solve this issue, we replace the regression loss function by a semantic perceptual one $\mathcal{L}_s$, relaxing the requirements by only matching semantic outputs.

The complete loss function $\mathcal{L}_G$ to be minimized by the generator network G is therefore

$$\mathcal{L}_G = -\mathcal{L}_D + \lambda \mathcal{L}_s, \tag{1}$$

where λ is a weighting coefficient between the two loss terms.

We introduce a semantic perceptual loss function to relax regression loss constraints, and instead match images in a semantic space. For this, we use a network T_s pre-trained for the task corresponding to the semantic description s, *i.e*, s lies in the output space of T_s. Target and generated images are then compared in the semantic space induced by T_s, *i.e*,

$$\mathcal{L}_s(y, \hat{y}) = |T_s(y) - T_s(\hat{y})|. \tag{2}$$

Note that a L_1 distance is used here, but others are possible. We argue that a matching in a semantic space rather than a pixel one should yield more diversity and less blurring in synthesized images, as the generator is not forced to closely predict target appearances.

Compared to common perceptual losses that use generic activations from ImageNet pre-trained networks (Johnson *et al.,* 2016), using a network specifically trained for a task related to the input semantic description considered makes the learning more specialized to its nature. Indeed, general purpose activations learned from ImageNet might only match global appearance and content, while task-specific ones should lead to a more accurate matching of the semantic features.

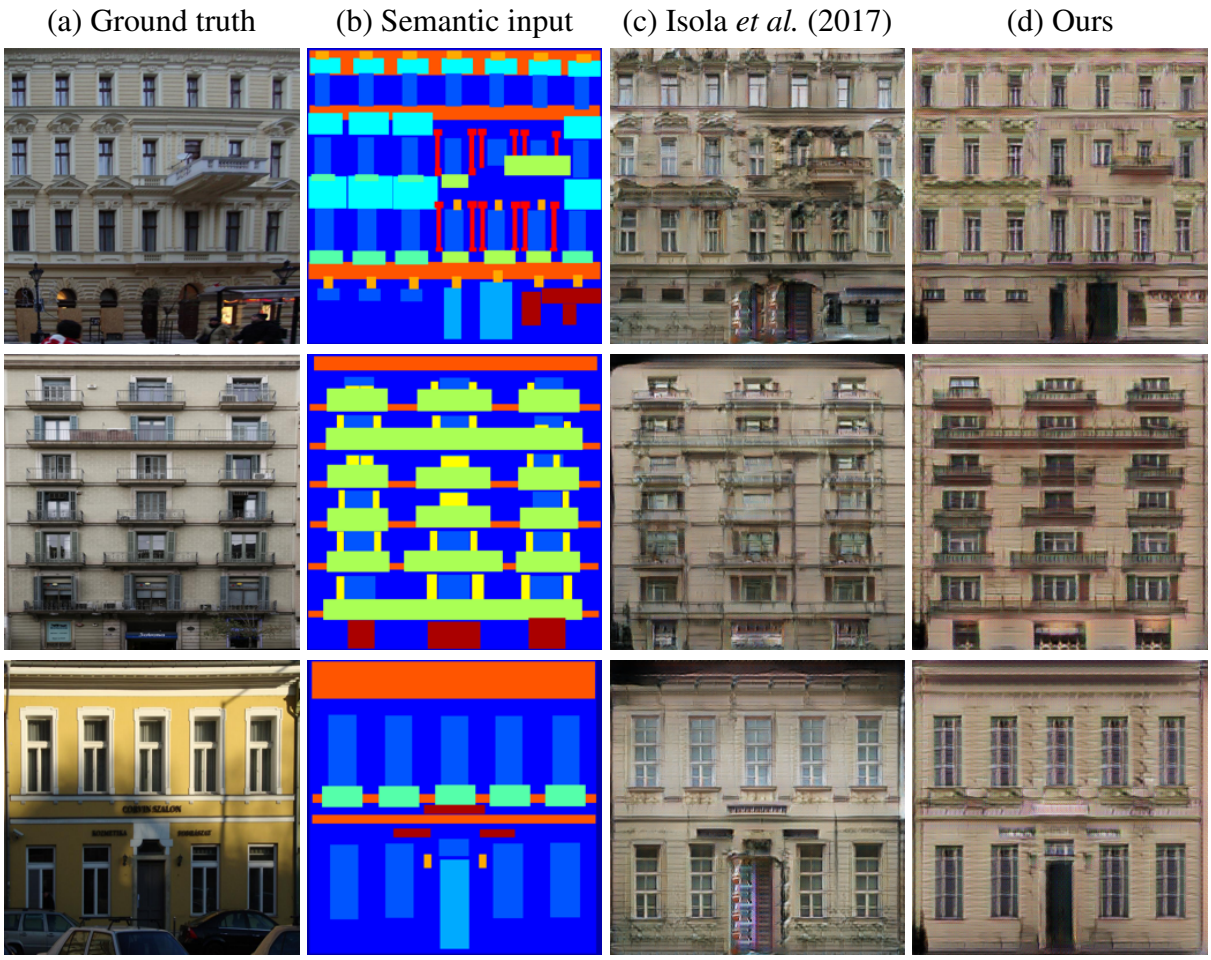

Figure 3: Qualitative results of facade image synthesis on CMP Facades. Column (a) represents the ground truth. Its semantic map is shown in column (b). The results of pix2pix baseline (Isola *et al.*, 2017) using their pre-trained models is shown in column (c) followed by our proposed method in column (d).

4. Experiments

We evaluate our model on two sets of experiments in different domains, showing the benefits of our proposed method:

1. scene synthesis from semantic maps;
2. human image synthesis from keypoints.

Figure 4: Qualitative results of city scenes image synthesis on Cityscapes. Column (a) represents the ground truth. Its semantic map is shown in column (b). The results of pix2pix baseline (Isola *et al.,* 2017) using their pre-trained models is shown in column (c) followed by our proposed method in column (d).

4.1. Scene synthesis from semantic maps

Datasets. For the task of generating scene images from segmentation maps, we use two different datasets. CMP Facades dataset (Tyleček and Šára, 2013) has 606 images of different resolutions of buildings with their 12 classes. We use the same split as Isola *et al.* (2017), composed of 400 training, 100 validation and 106 test examples. Cityscapes (Cordts *et al.,* 2016) is a dataset of road scenes, with 2975 images in training and 500 in validation. Semantic annotations are segmentation maps in 19 classes.

Implementation details. We use the same generator as Isola *et al.* (2017) and a 70×70 PatchGANs as base backbone for our discriminator. Learning is done with Adam optimizer with a learning rate of 0.0002, momentum parameters $\beta_1 = 0.5, \beta_2 = 0.999$ for 200 epochs. The loss

weight is set to $\lambda = 10$. Following Isola *et al.* (2017) implementation, we jitter and crop training images to augment the dataset. Results of pix2pix baseline (Isola *et al.,* 2017) are obtained from publicly available code.[1]

Results. Qualitative results are shown in Figure 3 for CMP Facades and in Figure 4 for Cistyscapes. Having different feature maps each focusing on a specific object could generate more details, *e.g*, the details of the windows and balconies which are less blurry and with more detailed for the facades. By giving equal weight to all classes, our discriminator is able to better synthesize small objects that have few pixels or that are less frequent, such as doors in Facades, and traffic signs and lights in Cityscapes.

4.2. Human image synthesis from keypoints

Datasets. For the task of generating human images in a given pose (described as keypoints) with the same appearance as a source image, we validate our approach on two datasets. Market-1501 dataset (Zheng *et al.,* 2015) contains $32,668$ images of $1,501$ people with the resolution of 128×64 captured from 6 different cameras. This is a challenging dataset because of its low-resolution images and pose diversity. 751 ids are used for training and the remaining for testing as the standard split (Zheng *et al.,* 2015) suggests. This dataset does not have pose information labels. To obtain them, a pre-trained pose detector (Cao *et al.,* 2017) with $K = 18$ keypoints is used. Results of Deformable GAN baseline (Siarohin *et al.,* 2018) are obtained from publicly available code.[2]

Implementation details. The generator and discriminator are similar to these in Siarohin *et al.* (2018). Learning uses the same hyper-parameters as in Section 4.1, but for 100 epochs.

Results. The results on Market-1501 dataset are available in Figure 5. We observe that our discriminator adds more details, especially on faces and hands without penalizing other parts. These details are even more visible in high resolution images.

[1] `https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix`
[2] `https://github.com/AliaksandrSiarohin/pose-gan`

(a) Source image (b) Target image (c) Siarohin *et al.* (2018) (d) Ours

Figure 5: Qualitative results of human image synthesis on Market-1501 dataset. Column (a) represents the source image. The target image is presented in column (b). The baseline Deformable GAN (Siarohin *et al.,* 2018) is shown in column (c) followed by our proposed method in column (e).

5. Conclusion and Next Steps

In this paper, we have addressed realistic image synthesis by exploring various ways of including semantics into the discriminator network for conditional GAN training. We have introduced a modification in the structure of the discriminator combined with a semantic perceptual loss function in order to match semantic features rather than pixels, which is less strict of a requirement. These improvements only affect the discriminator, so that no additional overhead is incurred at test time. The model has been evaluated on four image synthesis datasets, covering two different contexts and tasks, and has shown better quality in generated images.

In the future, we will study how to guide the generator with a more careful designed discriminator and train the whole network end-to-end.

Acknowledgement

We thank Mohammadhossein Bahari for his helpful comments.

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.

References

Brock, A., J. Donahue and K. Simonyan (2019) Large scale GAN training for high fidelity natural image synthesis, paper presented at the *Proceedings of the International Conference on Learning Representations (ICLR)*.

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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chan, C., S. Ginosar, T. Zhou and A. A. Efros (2018) Everybody dance now, *arXiv preprint arXiv:1808.07371*.

Cordts, M., M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth and B. Schiele (2016) The cityscapes dataset for semantic urban scene understanding, paper presented at the *Proceedings of 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 (NeurIPS)*, 474–484.

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

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 (NeurIPS)*, 2672–2680.

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 *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*, 694–711.

Karras, T., T. Aila, S. Laine and J. Lehtinen (2018a) Progressive growing of GANs for improved quality, stability, and variation, *Proceedings of the International Conference on Learning Representations (ICLR)*.

Karras, T., S. Laine and T. Aila (2018b) A style-based generator architecture for generative adversarial networks, *arXiv preprint arXiv: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)*.

Li, C. and M. Wand (2016) Precomputed real-time texture synthesis with markovian generative adversarial networks, paper presented at the *Proceedings of the IEEE European Conference on Computer Vision (ECCV)*.

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

Liu, Y., P. Kothari and A. Alahi (2019) Collaborative sampling in generative adversarial networks, April 2019.

Ma, L., X. Jia, Q. Sun, B. Schiele, T. Tuytelaars and L. V. Gool (2017) Pose guided person image generation, paper presented at the *Advances in Neural Information Processing Systems (NIPS)*.

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

Park, T., M.-Y. Liu, T.-C. Wang and J.-Y. Zhu (2019) Semantic image synthesis with spatially-adaptive normalization, paper presented at the *Proceedinsg of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

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 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

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

Tyleček, R. and R. Šára (2013) Spatial pattern templates for recognition of objects with regular structure, paper presented at the *Proceedings of the German Conference on Pattern Recognition (GCPR)*.

Vondrick, C., H. Pirsiavash and A. Torralba (2016) Generating videos with scene dynamics, paper presented at the *Advances in Neural Information Processing Systems (NIPS)*.

Wang, T.-C., M.-Y. Liu, A. Tao, G. Liu, J. Kautz and B. Catanzaro (2019) Few-shot video-to-video synthesis, paper presented at the *Advances in Neural Information Processing Systems (NeurIPS)*.

Wang, T.-C., M.-Y. Liu, J.-Y. Zhu, G. Liu, A. Tao, J. Kautz and B. Catanzaro (2018a) Video-to-video synthesis, in *Advances in Neural Information Processing Systems (NeurIPS)*, 1144–1156.

Wang, T.-C., M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz and B. Catanzaro (2018b) 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 (CVPR)*.

Zakharov, E., A. Shysheya, E. Burkov and V. Lempitsky (2019) Few-shot adversarial learning of realistic neural talking head models, May 2019.

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

Zheng, Z., X. Yang, Z. Yu, L. Zheng, Y. Yang and J. Kautz (2019) Joint discriminative and generative learning for person re-identification, paper presented at the *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.