16th International Conference on Machine Vision Applications (MVA)
National Olympics Memorial Youth Center, Tokyo, Japan, May 27-31, 2019.

Spectral Normalization and Relativistic Adversarial Training for Conditional Pose Generation with Self-Attention

Yusuke Horiuchi
Waseda University
Tokyo, Japan
y.horiuchi@suou.waseda.jp

Edgar Simo-Serra
Waseda University
Tokyo, Japan
ess@waseda.jp

Satoshi Iizuka
University of Tsukuba
Tsukuba, Ibaraki
iizuka@cs.tsukuba.ac.jp

Hiroshi Ishikawa
Waseda University
Tokyo, Japan
hfs@waseda.jp

Abstract

We address the problem of conditional image generation of synthesizing a new image of an individual given a reference image and target pose. We base our approach on generative adversarial networks and leverage deformable skip connections to deal with pixel-to-pixel misalignments, self-attention to leverage complementary features in separate portions of the image, e.g., arms or legs, and spectral normalization to improve the quality of the synthesized images. We train the synthesis model with a nearest-neighbour loss in combination with a relativistic average hinge adversarial loss. We evaluate on the Market-1501 dataset and show how our proposed approach can surpass existing approaches in conditional image synthesis performance.

1 Introduction

Conditional pose generation refers to synthesizing a new image of a person, given a reference image of the person and a target pose. It has diverse applications such as generating animations from single images and visualizing fashion outfits. Due to having to generate large occluded parts of the image, it poses a very challenging problem. Not only is it necessary to synthesize new limbs, but it is also important to generate novel views of the garments worn by the person.

In this paper, we tackle this problem by building on the Deformable GAN [12] model, which is an approach that uses a fully convolutional network with deformable skip connections. We use this architecture as a base, and propose incorporating self-attention [15], which can leverage the similarity between non-local parts of the image, unlike standard convolution operations. Furthermore, we show the advantage of using the nearest-neighbor loss in combination with a relativistic average hinge adversarial loss [5], jointly with spectral normalization [9], instead of training with a nearest-neighbor and adversarial loss. Our approach is able to generate new images of persons with given poses by extracting the pose information of the input image using off-the-shelf 2D pose estimators. We base our approach on adversarial training, which has shown good results in a variety of conditional image generation tasks, such as image inpainting [3].

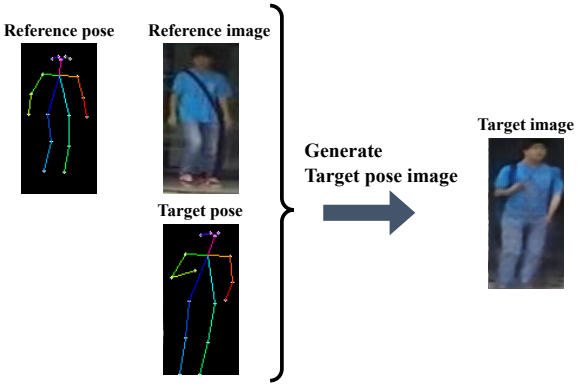

Figure 1. The conditional pose generation task. Given a reference image of a person, the pose of the person, and a target pose, we generate an image of the person in the target pose. In this work, the pose of the person in the input image is automatically extracted from the input image.

We evaluate our approach on the Market-1501 dataset, and compare against existing approaches. Results show that our model compares favorably with existing approaches, evaluated with image generation metrics such as the Inception score.

2 Related Work

Recently, Generative Adversarial Networks (GANs) [2] have become one of the most popular deep network-based generative models, which have been applied to various image generation tasks, such as image super-resolution [7] and image inpainting [3]. In particular, Isola et al. [4] proposed a framework of conditional GANs for image translation, which converts a given image into another image. Although it shows several plausible image translations such as maps

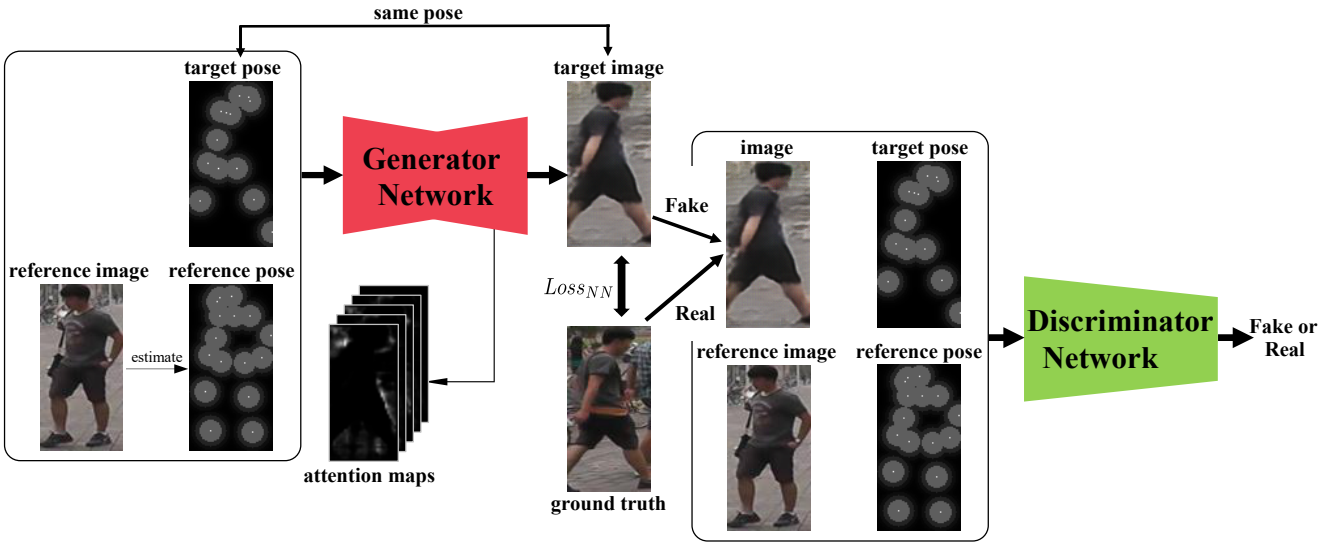

Figure 2. Task overview. The generator network is trained using nearest-neighbor loss and adversarial loss with correct images.

to aerial photos, it has difficulty in addressing large deformations between the input and output images, because the network is trained to spatially share the low-level information between the paired images.

For image translation with large spatial deformations, Ma et al. [8] proposed the pose-guided person generation network which allows synthesizing person images in arbitrary pose. This approach contains two stages: the first stage focuses on pose integration and generation based on the U-Net [10]-like convolutional network; and the second stage refines the initial result via adversarial training. Siarohin et al. [12] further improved the pose-based human image generation by incorporating deformable skip connections that can move local information according to the structural deformations. They also introduced a nearest-neighbor loss, instead of the more common losses such as L_1 and L_2, in order to match the details of the model outputs and the target images. Unlike the previous methods, it can be trained end-to-end, while achieving qualitatively better results.

3 Approach

Our approach builds on Deformable GAN (DGAN) [12] and improves it by incorporating spectral normalization [9] and self-attention [15], as well as changing the loss function from the standard Generative Adversarial Network (GAN) to the Relativistic Hinge GAN [5].

3.1 Deformable GAN

The Deformable GAN extends the Generative Conditional Networks by incorporating deformable skip connections in the generator network and employs a nearest-neighbor loss instead of the commonly used L_1 and L_2 losses. Our model is based on the Deformable GAN model and unless we note otherwise, we use the same architecture.

For a given input person image and a target pose, DGAN first extracts the pose of the person in the form of 2D skeleton with a human pose estimator model [1]. The data is then processed with a fully convolutional network encoder to obtain a feature representation of the input image, the extracted pose, and the target pose. Afterwards, using the pose information for each specific body part, an affine transformation is computed and applied to "move" the feature-map content corresponding to that body part.

For training, instead of the commonly used L_1 and L_2 losses, a nearest neighbor loss is used. For an input image $\boldsymbol{x}$ and a target image $\boldsymbol{x}^*$, it is defined as

$$\mathcal{L}_{\mathrm{NN}}(\boldsymbol{x}, \boldsymbol{x}^*) = \sum_{\boldsymbol{p} \in \boldsymbol{x}^*} \min_{\boldsymbol{q} \in N(\boldsymbol{p})} \|C_{\boldsymbol{x}^*}(\boldsymbol{p}) - C_{\boldsymbol{x}}(\boldsymbol{q})\|_1 \,, \quad (1)$$

where $N(\boldsymbol{p})$ is a neighborhood of $\boldsymbol{p}$, and pre-trained VGG19 [13] network with respect to the spatial position $\boldsymbol{p}$.

In addition to the nearest-neighbor loss L_{NN}, a conditional adversarial loss is used and defined as

$$\begin{aligned}\mathcal{L}_{\mathrm{cGAN}} = \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{x}^*) \in X}[\log D(\boldsymbol{x}, H, \boldsymbol{x}^*, H^*)] \\ + \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{x}^*) \in X, \boldsymbol{z} \in Z}[\log(1 - D(\boldsymbol{x}, H, \hat{\boldsymbol{x}}, \hat{H}) \,, \quad (2)\end{aligned}$$

where $\mathbb{E}$ denotes the expectation value, X is the set of training pair images, D is the discriminator model that outputs a probability, $\hat{\boldsymbol{x}} = G(\boldsymbol{z}, \boldsymbol{x}, H, H^*)$, G is the generator model, Z is a random distribution, and H, H^*, and $\hat{H}$ are the 2D skeletons of $\boldsymbol{x}$, $\boldsymbol{x}^*$, and $\hat{\boldsymbol{x}}$, respectively.

The objective function for optimizing the Deformable GAN model thus becomes:

$$\arg \min_G \max_D \mathcal{L}_{\mathrm{cGAN}}(G, D) + \lambda \mathcal{L}_{\mathrm{NN}}(G) \,, \quad (3)$$

where $\mathcal{L}_{\mathrm{NN}}(G) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{x}^*)\in X,\boldsymbol{z}\in Z} L_{\mathrm{NN}}(\hat{\boldsymbol{x}}, \boldsymbol{x}^*)$.

3.2 Spectral Normalization

Spectral normalization [9] normalizes the spectral norm of the weight matrix W of a layer so that it satisfies the Lipschitz constraint $\sigma(W) = 1$:

$$\bar{W}_{\mathrm{SN}}(W) = \frac{W}{\sigma(W)} \tag{4}$$

Instead of naïvely applying singular value decomposition to compute $\sigma(W)$, the power iteration method is used to estimate $\sigma(W)$ with a small computational footprint. The spectral normalization is used in both the discriminator and the decoder part of the generator, and improves results by decreasing the degradation of the error signal during back-propagation.

3.3 Self-Attention

Convolutional layers process information in only a local neighborhood, making fully convolutional networks computationally inefficient for modeling long-range dependencies in images, e.g., human arms or legs. In a self-attention layer [15], the image features from the previous layer $\boldsymbol{\phi} \in \mathbb{R}^{C\times N}$ with N channels are first transformed into two features spaces $g(\boldsymbol{\phi}) = W_g\boldsymbol{\phi}$ and $h(\boldsymbol{\phi}) = W_h\boldsymbol{\phi}$, where W_g and W_h are learnable matrices. Let spatial locations or regions indexed by u, v. Then the attention map β is defined by:

$$\beta_{v,u} = \frac{\exp(s_{u,v})}{\sum_{u=1}^{N}\exp(s_{uv})}, \quad s_{uv} = g(\boldsymbol{\phi}_u)^\intercal h(\boldsymbol{\phi}_v). \tag{5}$$

Then $\beta_{v,u}$ indicates to what extent the layer attends to the u-th location when synthesizing the v-th region. The output of the layer $\boldsymbol{o} = (\boldsymbol{o}_1, \ldots, \boldsymbol{o}_N) \in \mathbb{R}^{C\times N}$ can then be computed as

$$\boldsymbol{o}_v = \gamma \sum_{u=1}^{N} \beta_{v,u} W_o \boldsymbol{\phi}_u + \boldsymbol{\phi}_v \ , \tag{6}$$

where W_o is another learnable matrix and γ is a parameter that is initialized to 0 and gradually increased during training to learn to assign more weight to non-local features. The self attention learns the parameters $W_g \in \mathbb{R}^{\bar{C}\times C}$, $W_h \in \mathbb{R}^{\bar{C}\times C}$, and $W_o \in \mathbb{R}^{C\times C}$ with back-propagation during training.

We add a self-attention after each of the last two layers of the generator's decoder and after each of the last two layers of the discriminator.

3.4 Relativistic Average Hinge Adversarial Loss

Training with conditional adversarial loss is done with a discriminator that is being trained to classify whether the input is real or fake with negative log-likelihood. Relativistic adversarial losses [5] optimize the probability that a given real data is more realistic than a randomly sampled fake data and vice versa. This leads to more stable and robust training in general. Relativistic adversarial training can be done by modifying the output of the discriminator $D(\cdot)$ to be

$$\bar{D}(\boldsymbol{x'}) = \begin{cases} D(\boldsymbol{x'}) - \mathbb{E}_{\hat{\boldsymbol{x}}\in X_{\mathrm{T}}} D(\hat{\boldsymbol{x}}) & \text{if } x' \text{ is real} \\ D(\boldsymbol{x'}) - \mathbb{E}_{\boldsymbol{x}^*\in X_{\mathrm{I}}} D(\boldsymbol{x}^*) & \text{if } x' \text{ is fake} \ , \end{cases} \tag{7}$$

where we denote the set of input images as X_{I} and the set of target images as X_{T}, with $X = (X_{\mathrm{I}}, X_{\mathrm{T}})$.

The discriminator and generator losses can then be written as

$$\begin{aligned} \mathcal{L}_D^{\mathrm{RH}}(G, D) &= \mathbb{E}_{\boldsymbol{x}^*\in X_{\mathrm{T}}}\left[\max(0, 1 - \bar{D}(\boldsymbol{x}^*))\right] \\ &\quad + \mathbb{E}_{\boldsymbol{x}\in X_{\mathrm{I}}}\left[\max(0, 1 + \bar{D}(\hat{\boldsymbol{x}}))\right] \\ \mathcal{L}_G^{\mathrm{RH}}(G, D) &= \mathbb{E}_{\boldsymbol{x}\in X_{\mathrm{I}}}\left[\max(0, 1 - \bar{D}(\hat{\boldsymbol{x}}))\right] \\ &\quad + \mathbb{E}_{\boldsymbol{x}^*\in X_{\mathrm{T}}}\left[\max(0, 1 + \bar{D}(\boldsymbol{x}^*))\right] \ . \end{aligned} \tag{8}$$

We note that this leads to a min-min problem, unlike the standard adversarial loss formulation, which leads to a min-max problem.

3.5 Training

Replacing the DGAN objective function (3), we train our model by using the nearest-neighbor loss $\mathcal{L}_{\mathrm{NN}}$ in conjunction with the relativistic average hinge adversarial loss by minimizing

$$\arg\min_{G,D} \mathcal{L}_{\mathrm{NN}}(G) + \lambda_1 \mathcal{L}_G^{RH}(G, D) + \lambda_2 \mathcal{L}_D^{RH}(G, D) \tag{9}$$

where λ_1 and λ_2 are two hyperparameters.

4 Experimental Results

We evaluate using the Market-1501 [16] dataset which contains 32,668 images of 1,501 individuals captured by 6 different surveillance cameras. Following [12], the dataset is cleaned up by automatically removing images in which no individual is detected using HPE [1], and training with pairs of images of the same individual in two different poses. This results in 263,631 training pairs, of which we use 2,000 for validation and 10,000 for testing. No person appears in more than one splits. In contrast to [12], which trained all models for a fixed number of iterations, we train for 50,000 iterations and use the model that has the largest Inception Score (IS) [11] on the validation set which we then use to evaluate on the testing images. This reduces the effect of the stochastic nature of training generative adversarial network models during evaluation. We evaluate using three metrics, Inception Score (IS) [11], SSIM [14], and L_1 distance, between the ground truth and the output of the model. We use $\lambda_1 = \lambda_2 = 50$ for our model and train by optimizing Eq. (9) with ADAM [6].

Table 1. Comparison with existing approaches. We compare with the results of Defomable GAN (DGAN) [12], both the result reported in their paper trained for 90 epochs, and the result of retraining their model and using validation to choose the best model. The best results are highlighted in bold.

Model	IS	SSIM	L_1
DGAN [12] (paper)	3.185	**0.290**	-
DGAN [12] (retrained)	3.272	0.274	0.292
Ours	**3.402**	0.279	**0.288**
Real-Data	3.86	1.00	0.00

Table 2. Ablation results. We compare different variants of our proposed approach. SN: Spectral Normalization, RH: Relativistic Hinge loss, SA: Self-Attention. The best results are highlighted in bold.

Model	IS	SSIM	L_1
Ours SN	3.066	0.289	0.289
Ours RH	3.122	0.283	0.289
Ours SA	3.121	0.278	0.295
Ours SN+RH	2.973	**0.296**	**0.288**
Ours SA+SN+RH	**3.402**	0.279	**0.288**

4.1 Comparison with existing approaches

We compare against the approach of Deformable GAN (DGAN) [12] by retraining their model and using the validation set to choose their best model. For reference, we also show their results as reported in their paper. Results are shown in Table 1. Retraining DGAN shows improved performance in Inception Score (IS), although SSIM decreases. Our approach outperforms significantly in IS, and although it beats the SSIM of the retrained model, it underperforms in comparison to the originally reported values. However, we found that removing the self-attention layers improves the SSIM at the cost of IS, outperforming the original values reported in [12].

4.2 Ablation study

We perform an ablation study to analyze the different effects of the components of our model. We show the results in Table 2. We can see that the combination of all the elements gives a large boost in IS, while not using Self-Attention (SA) improves SSIM.

4.3 Qualitative results

We show some qualitative results of our approach in Fig. 3. The output result shows that the overall quality of the image is improved. In particular, the quality of background improved by using Self-Attention.

5 Conclusions

We have presented an approach for the conditional generation of images of individuals given a single reference image and a target pose. Our model incorporates deformable skip connections and self-attention, and is trained with a nearest-neighbour loss in combination with a relativistic average hinge adversarial loss, using spectral normalization. We have evaluated on the Market-1501 dataset and results compare favorable with existing approach.

6 Acknowledgements

This work was partially supported by JST ACT-I (Iizuka, Grant Number: JPMJPR16U3), JST PRESTO (Simo-Serra, Grant Number: JPMJPR1756), and JST CREST (Ishikawa, Iizuka, and Simo-Serra, Grant Number: JPMJCR14D1).

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[2] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Conference on Neural Information Processing Systems*, 2014.

[3] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics*, 36(4):107:1–107:14, July 2017.

[4] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5967–5976, 2017.

[5] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018.

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[7] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew P. Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 105–114, 2017.

[8] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Conference on Neural Information Processing Systems*, pages 405–415, 2017.

[9] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations*, 2018.

[10] O. Ronneberger, P.Fischer, and T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *Med-*

Figure 3. Examples of generated images and attention map. Attention map is sampled from random coordinates. Baseline is output by Deformable GAN [12]

ical Image Computing and Computer-Assisted Intervention (MICCAI), volume 9351, pages 234–241, 2015.

[11] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *Conference on Neural Information Processing Systems*, 2016.

[12] Aliaksandr Siarohin, Enver Sangineto, Stphane Lathuilire, and Nicu Sebe. Deformable gans for pose-based human image generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[13] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[14] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004.

[15] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.

[16] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *International Conference on Computer Vision*, 2015.