GENERATIVE GUIDING BLOCK: SYNTHESIZING REALISTIC LOOKING VARIANTS CAPABLE OF EVEN LARGE CHANGE DEMANDS

*Minho Park, Hak Gu Kim, and Yong Man Ro**

Image and Video Systems Lab, School of Electrical Engineering, KAIST, South Korea

ABSTRACT

Realistic image synthesis is to generate an image that is perceptually indistinguishable from an actual image. Generating realistic looking images with large variations (e.g., large spatial deformations and large pose change), however, is very challenging. Handing large variations as well as preserving appearance needs to be taken into account in the realistic looking image generation. In this paper, we propose a novel realistic looking image synthesis method, especially in large change demands. To do that, we devise generative guiding blocks. The proposed generative guiding block includes realistic appearance preserving discriminator and naturalistic variation transforming discriminator. By taking the proposed generative guiding blocks into generative model, the latent features at the layer of generative model are enhanced to synthesize both realistic looking- and target variation- image. With qualitative and quantitative evaluation in experiments, we demonstrated the effectiveness of the proposed generative guiding blocks, compared to the state-of-the-arts.

***Index Terms*—** Deep learning, adversarial learning, variation image synthesis, and feature enhancement

1. INTRODUCTION

Generating realistic-looking images draws great attention and considered as an important task in generative models for image synthesis. Recently, deep learning-based generative models have achieved remarkable success in various synthesis tasks such as face, human, and scene generation. In data acquisition, it is time consuming and costly to collect or capture the images with desired variations (e.g., pose, illumination, facial expression, and viewpoint). Generative models that can automatically synthesize images with the desired variations are needed in practice.

For generating realistic-looking images of objects, it is required to understand both their appearance and variants. The object has inherent appearance properties characterized by color and texture such as hair color and fashion style. On the other hand, there are variants including the shape and geometrical layout of the object. One of the most challenging points in the image generation is to preserve the appearance properties of input image (e.g., color, texture, the identity of person) while performing spatial deformation according to variants (e.g., pose variation and illumination variation).

For this task, so far, various methods have been proposed based on Variational Auto-Encoders (VAEs) [1], Generative Adversarial Networks (GANs) [2] and Autoregressive models (ARMs) (e.g., PixelRNN [3]) [4–12]. Recently, a wide range of methods including conditional GANs [13] or conditional VAEs [14] have been proposed for synthesizing the images whose appearances depend on a given conditioning variable (e.g., label). However, most of them could not deal with the large variations (e.g., large spatial deformation [15]) between the input and the target image while preserving the appearance of a given input. Due to the high dimensionality of images and the complex configuration of image contents, it is difficult for a complete end-to-end framework to generate both the correct target variation and the detailed appearance simultaneously [16–19].

In this paper, we focus on realistic appearance and naturalistic variation in target image generation. The generative features are enhanced with appearance preservation and variant transformation. Our objective is to propose new generation method that addresses two problems, which are realistic appearance and naturalistic large-variation. To cope with the problems, we propose a novel generative guiding blocks (GGBs). Each generative guiding block consists of realistic appearance preserving discriminator (RAPD) and naturalistic variation transforming discriminator (NVTD). In the proposed RAPD, to preserve the object appearance of input image (e.g., identity of person), the overall image distribution is considered by determining whether the appearance is preserved in the target image or not. Simultaneously, in the proposed NVTD, to generate the target image with large variation, the change information of deformation is considered by focusing on the variation between the input and the generated target image. We hierarchically integrate the proposed GGBs with the decoding module of the generator to enhance generative feature in multiple resolution levels. The proposed generative model with GGBs enables to synthesize the realistic-looking image robustly even with large variations while maintaining naturalistic variants. Experimental results showed the effectiveness of the proposed GGBs.

* Corresponding author (ymro@ee.kaist.ac.kr). This work was partly supported by IITP grant (No. 2017-0-00780).

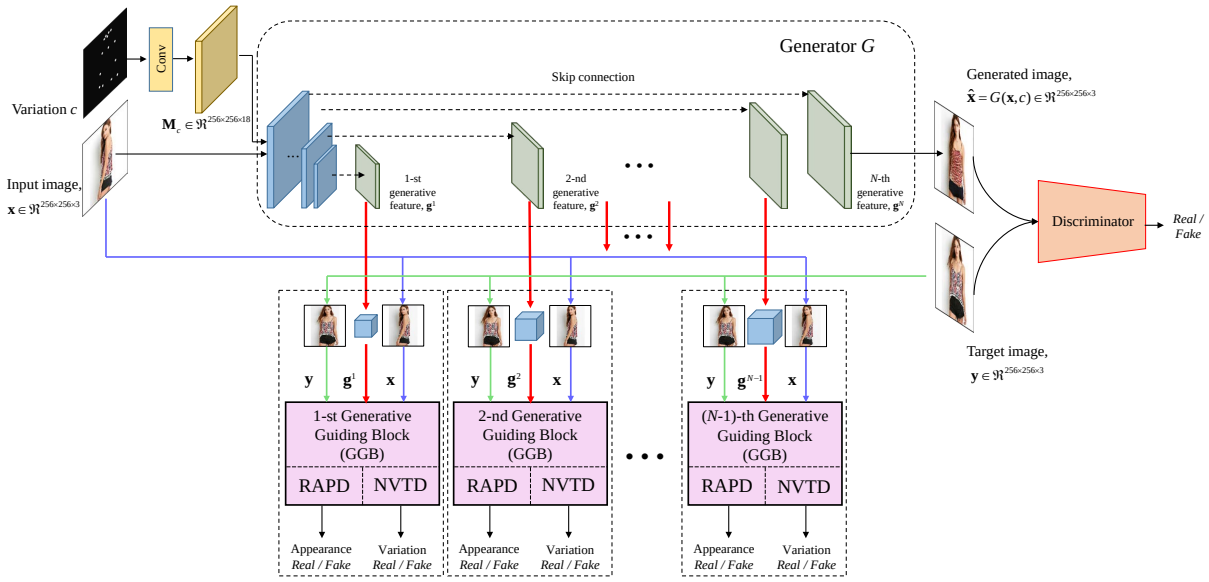

Fig. 1. Overall Architecture of the proposed generative model with generative guiding blocks (GGBs). Note GBBs is hierarchically integrated in decoding modules of generator in multiple resolution levels.

The rest of this paper is organized as follows. In section 2, we describe the proposed generative model with GGBs. In section 3, the experimental results are presented. Finally, conclusion is drawn in section 4.

2. PROPOSED METHOD

Fig. 1 shows the proposed generative model with generative guiding blocks (GGBs). The generator synthesizes the fake image having the appearance of the input image and the target variants. The discriminator determines whether the fake image is real or not. As shown in Fig. 1, the generative guiding blocks (GGBs) are attached to multi-level generative features of multiple layers in the decoder of generator. The GGBs determine whether the generated multi-resolution images have realistic appearance (operated by RAPD in GGB) and naturalistic variation (operated by NVTD in GGB). Variant transformation is performed hierarchically in a multi-resolution manner so that the proposed generator can process large variant demand. In the following subsections, we describe in detail about the generator, discriminator and GGBs.

2.1. Generative model with discriminator

Let $\mathbf{x} \in \mathbb{R}^{256\times256\times3}$ denote the input image and $\mathbf{y} \in \mathbb{R}^{256\times256\times3}$ denote the ground-truth target image. c denotes the target variation and $\hat{\mathbf{x}} \in \mathbb{R}^{256\times256\times3}$ (i.e. $G(\mathbf{x}, c)$) denotes the generated image. Let $\mathbf{g}^n$ denote n-th generative feature. Let G denote the generator, D denote the discriminator and $\mathbf{M}_c \in \mathbb{R}^{256\times256\times3}$ denote the label map which is encoded from c. By encoding c, abundant condition information of the desired variation is provided to the G. In this paper, a U-Net-like structure is employed as G [20, 21]. The encoder and decoder of G consist of 7 convolution layers and deconvolution layers, respectively (i.e. N=7) with 4×4 kernel and stride of 2. D consists of 5 convolution layers with 4×4 kernel and stride of 2.

With an adversarial learning [2], D determines whether the $\hat{\mathbf{x}}$ is a realistic-looking or not, comparing with $\mathbf{y}$. The objective functions of D can be written as

$$\begin{aligned}\mathcal{L}_D = &- \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{y}}}[\log(D(\mathbf{y}))] \\ &- \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}[\log(1 - D(G(\mathbf{x}, c)))].\end{aligned} \tag{1}$$

On the other hand, G tries to fool D by generating the realistic image. To that end, the loss of the generator is composed of two terms, which are the realism loss, ℓ_{real}, and the reconstruction loss, ℓ_{rec}. The realism loss can be written as

$$\ell_{real} = - \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}[\log(D(G(\mathbf{x}, c)))] \tag{2}$$

The reconstruction loss between the ground-truth target image and the generated image at n-th level, ℓ^n_{rec} , in the decoder can be written as

$$\ell^n_{rec} = \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}[\|\mathbf{y}^n - \hat{\mathbf{x}}^n\|_1], \tag{3}$$

where $\hat{\mathbf{x}}^n$ indicates a generated image from $\mathbf{g}^n$ and $\mathbf{y}^n$ indicates an image downsized to the same resolution of $\hat{\mathbf{x}}^n$ from $\mathbf{y}$ (as shown in Fig. 2).

Finally, the total loss function of the proposed generator, G, can be defined as a combination of the realism loss and the reconstruction loss.

$$\mathcal{L}_G = \lambda_{real}\ell_{real} + \ell^N_{rec}, \tag{4}$$

where λ_{real} is a weight parameter to control the balance between ℓ_{real} and ℓ^N_{rec}.

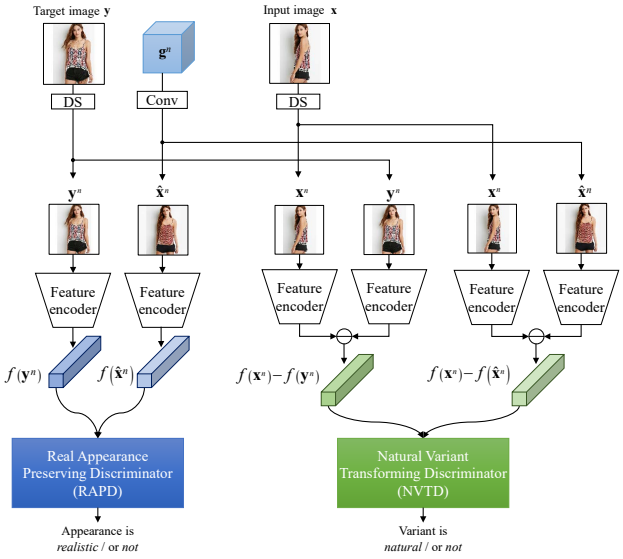

Fig. 2. The architecture of the proposed n-th GGB.

2.2. Generative Guiding Block for realistic appearance and naturalistic variation

Fig. 2 shows the architecture of the proposed n-th GGB, which consists of a realistic appearance preserving discriminator (RAPD), D_{RAPD}, and a naturalistic variation transforming discriminator (NVTD), D_{NVTD}. The GGBs are attached on the multi-level generative features of multiple layers in the decoder as shown in Fig. 1. Let $\mathbf{x}^n$ denote an image downsized to the same resolution of $\hat{\mathbf{x}}^n$ from $\mathbf{x}$. Let $f(\cdot)$ denote the feature encoder. In this paper, D_{RAPD} and D_{NVTD} consist of 3 convolution layers. The feature encoder consists of 2 convolution layers with 4×4 kernel and stride of 2.

First, to deal with feature information of $\mathbf{x}^n$, $\hat{\mathbf{x}}^n$ and $\mathbf{y}^n$, the images are encoded to the latent feature, $f(\mathbf{x}^n)$, $f(\hat{\mathbf{x}}^n)$ and $f(\mathbf{y}^n)$. After that, D_{RAPD} distinguishes whether the encoded features, $f(\hat{\mathbf{x}}^n)$ and $f(\mathbf{y}^n)$, are realistic or not. As shown in Fig. 2, D_{NVTD} distinguishes whether the residual information of encoded features (i.e., $\mathbf{d}^n_{real} = f(\mathbf{x}^n) - f(\mathbf{y}^n)$ and $\mathbf{d}^n_{fake} = f(\mathbf{x}^n) - f(\hat{\mathbf{x}}^n)$) is realistic or not. The reason that the input of D_{NVTD} is residual information is to make D_{NVTD} focus on only the target variation. G tries to fool D_{RAPD}, so that $\hat{\mathbf{x}}^n$ mimics the data distribution of $\mathbf{y}^n$. Through this process, $\mathbf{g}^n$ is enhanced for generating appearance realistic image. Also, G tries to fool D_{NVTD}, so that $\mathbf{d}^n_{fake}$ tries to follow $\mathbf{d}^n_{real}$. $\mathbf{g}^n$ is enhanced for generating the image with naturalistic variation as well.

The discriminators in GGB, D_{RAPD} and D_{NVTD}, are trained by adversarial learning with G. Therefore, we adopt generative adversarial loss. First, the objective function of D_{RAPD} is defined as

$$\begin{aligned}\mathcal{L}^n_{D_{RAPD}} = &- \mathbb{E}_{\mathbf{y}\sim p_{\mathbf{y}}}[\log(D^n_{RAPD}(f(\mathbf{y}^n)))] \\ &- \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}[\log(1 - D^n_{RAPD}(f(\hat{\mathbf{x}}^n)))],\end{aligned} \tag{5}$$

where D^n_{RAPD} indicates D_{RAPD} in n-th GGB. Similarly, the objective function of D_{NVTD} is defined as

$$\begin{aligned}\mathcal{L}^n_{D_{NVTD}} = &- \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}},\mathbf{y}\sim p_{\mathbf{y}}}[\log(D^n_{NVTD}(\mathbf{d}^n_{real}))] \\ &- \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}[\log(1 - D^n_{NVTD}(\mathbf{d}^n_{fake}))],\end{aligned} \tag{6}$$

where D^n_{NVTD} indicates D_{NVTD} in n-th GGB.

D^n_{RAPD} and D^n_{NVTD} are trained to minimize $\mathcal{L}^n_{D_{RAPD}}$ and $\mathcal{L}^n_{D_{NVTD}}$, respectively. Contrary, G with GGBs is trained to minimize ℓ^n_{RAPD} and ℓ^n_{NVTD} for learning to fool D^n_{RAPD} and D^n_{NVTD}. These objective functions can be written as

$$\ell^n_{RAPD} = - \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}[\log(D^n_{RAPD}(f(\hat{\mathbf{x}}^n)))], \tag{7}$$

$$\ell^n_{NVTD} = - \mathbb{E}_{\mathbf{x}\sim p_{\mathbf{x}}}[\log(D^n_{NVTD}(\mathbf{d}^n_{fake}))]. \tag{8}$$

In particular, to preserve the appearance information, we adopt the L1 norm as our reconstruction loss, Eq. 3. Finally, the objective function of G with our GGBs is defined as

$$\mathcal{L}_{GGB} = \sum_{n=1}^{N-1} \lambda^n_{RAPD}\ell^n_{RAPD} + \lambda^n_{NVTD}\ell^n_{NVTD} + \ell_{rec}, \tag{9}$$

where Σ is used for weighted sum of multi-level GGB losses.

2.3. Training strategy

Every iteration, $\mathbf{x}$ and c are given to G. Then, G generates $\hat{\mathbf{x}}$. In the D, $\mathcal{L}_D$ is calculated with $\hat{\mathbf{x}}$ and $\mathbf{y}$ (see Eq.1). In the n-th GGB, $\mathcal{L}^n_{D_{RAPD}}$ and $\mathcal{L}^n_{D_{NVTD}}$ are calculated with $\mathbf{x}^n$, $\mathbf{y}^n$ and $\hat{\mathbf{x}}^n$ (see Eq.5 and 6). After that, the weights of D are updated to minimize $\mathcal{L}_D$. Also, the weights of n-th GGB are updated to minimize $\mathcal{L}^n_{D_{RAPD}}$ and $\mathcal{L}^n_{D_{NVTD}}$ (n=1,2,...,N-1). The weights of G except for $\mathbf{g}^N$ are firstly updated to minimize $\mathcal{L}_{GGB}$ (see Eq. 9). Finally, the weights of G are updated to minimize $\mathcal{L}_G$ (see Eq. 4). Until the weights are optimized, this process is repeated.

Table 1. Quantitative comparison with the state-of-the-art methods on DeepFashion dataset.

Model	SSIM	IS
Disentangled [17]	0.614	3.23
VariGAN [18]	0.620	3.03
PG^2 [16]	0.762	3.09
DPT [19]	0.769	3.17
Ours	**0.799**	**3.26**

Table 2. Effectiveness of using both RAPD/NVTD and multiple GGBs

Model	SSIM	IS
Ours w/o GGBs	0.705	2.81
Ours w/o RAPD	0.709	2.72
Ours w/o NVTD	0.714	2.73
Ours with 1 GGB	0.780	3.14
Ours with 2 GGBs	0.793	3.15
Ours	**0.799**	**3.26**

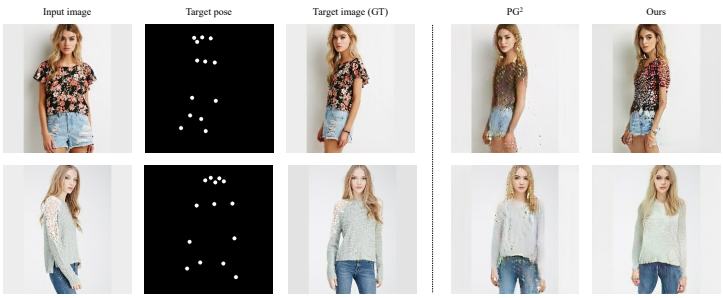

Fig. 3. Qualitative comparison on DeepFashion dataset between the results obtained by our approach and PG^2 [16].

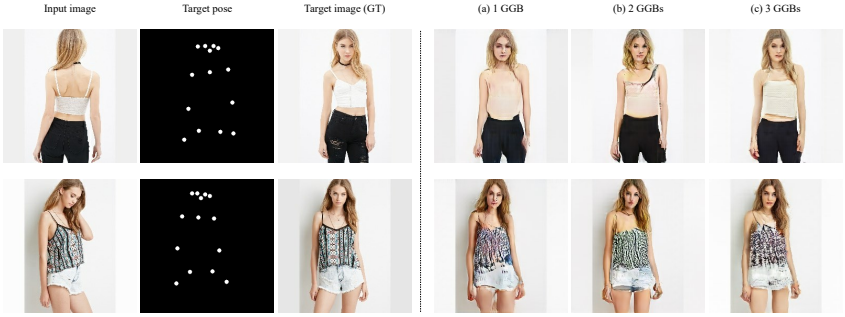

Fig. 4. Generated human pose images obtained by our model on DeepFashion dataset when it is trained with (a) only 6-th GGB, (b) 5-th and 6-th GGBs, (c) 4-th, 5-th and 6-th GGBs.

3. EXPERIMENTS AND RESULTS

3.1. Datasets

For verifying the effectiveness of the proposed generative model with GGBs, we used public datasets: DeepFashion [22]. This dataset consists of 52,712 in-shop clothes images with 256×256 resolution. As similar to [16], for the training set, we have 146,680 pairs. Each pair is composed of two images of the same identity but different poses. For the test set, we randomly selected 12,800 pairs from the test set. To use the human pose landmark of DeepFashion data as the target variation, we applied a state-of-the-art pose estimation [23], as in [16].

3.2. Implementation details

We used Adam optimizer [24] with β_1 = 0.5, β_2 = 0.999, the batch size of 8, and learning rate of 0.0002 to train proposed models. In our experiment, we attached three GGBs on the generative features with 32 × 32, 64 × 64 and 128 × 128 resolutions (i.e. $\mathbf{g}^4$, $\mathbf{g}^5$ and $\mathbf{g}^6$). We empirically set λ_{real} = 0.02 and $\lambda^n_{RAPD} = \lambda^n_{NVTD}$ = 0.01.

3.3. Performance evaluation

Fig. 3 shows comparison between generated images by our model and those by the state-of-the-art model, PG^2 [16]. To obtain the results of PG^2, we used pretrained weight provided by the author of PG^2. As shown in Fig. 3, in the results of PG^2, hair and clothes were blurred a lot. Thus the appearance information was not preserved well. On the other hand, the appearances were preserved well in ours. Fig. 4 shows the effectiveness of refining multi-level features using GGBs. '1 GGB' indicates the generative model with only 6-th GGB. '2 GGBs' indicates the generative model with 5-th and 6-th GGBs. '3 GGBs' indicates the generative model with 4-th, 5-th and 6-th GGBs, same as proposed model. The more GGBs were used in generative model training, the clearer the images and the better the appearance preserved. Table 1 and 2 show the quantitative results of state-of-the-art models [16–19] and the proposed model by measuring Structural Similarity (SSIM) [25] and Inception scores (IS) [7]. As seen in Table 1, the proposed method outperformed the state-of-the-art method. In table 2, 'w/o GGBs' indicates training generative model without any GGB. 'w/o RAPD' and 'w/o NVTD' indicate that there are only NVTD and RAPD in GGB, respectively. As seen in Table 2, the proposed model (i.e. 3 GGBs are used, RAPD and NVTD in GGB) provided the highest performance.

4. CONCLUSION

In this paper, we proposed a novel Generative Guiding Block for synthesizing realistic looking images with the large variations while preserving the appearance properties. The proposed GGB consisted of two critic networks which were RAPD for maintaining the appearance characteristic and NVTD for applying the target variants. By hierarchically integrating the proposed GGBs with the generator, the proposed GGBs could enhance the generative features in the decoder from coarse to fine. The experimental results showed that the proposed method outperformed the state-of-the-art methods. Also, the effectiveness of components of GGB (i.e. RAPD and NVTD) and hierarchical multi-level features were shown.

5. REFERENCES

[1] D. P. Kingma and M. Welling, "Auto-encoding variational bayes.," *CoRR*, vol. abs/1312.6114, 2013.

[2] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems 27*, pp. 2672–2680. Curran Associates, Inc., 2014.

[3] A. Van Den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *ICML*, 2016, pp. 1747–1756.

[4] R. A. Yeh*, C. Chen*, T. Y. Lim, A. G. Schwing, M. HasegawaJohnson, and M. N. Do, "Semantic image inpainting with deep generative models," in *CVPR*, 2017, * equal contribution.

[5] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras, "Neural face editing with intrinsic image disentangling," in *CVPR*. IEEE, 2017, pp. –.

[6] D. Pathak, P. Krähenbühl, J. Donahue, T. Darrell, and A. Efros, "Context encoders: Feature learning by inpainting," in *CVPR*, 2016.

[7] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NIPS*, pp. 2234–2242. 2016.

[8] D. Yoo, S. Park Kim, N. Kim, A. S. Paek, and I. Kweon, "Pixel-level domain transfer," in *ECCV*, 10 2016, vol. 9912, pp. 517–532.

[9] Y. Zhou and T. L. Berg, "Learning temporal transformations from time-lapse videos," in *ECCV*, 2016, pp. 262–277.

[10] H. J. Lee, S. T. Kim, H. Lee, and Y. M. Ro, "Lightweight and effective facial landmark detection using adversarial learning with face geometric map generative network," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[11] J. U. Kim, J. Kwon, H. G. Kim, and Y. M. Ro, "Bbc net: Bounding-box critic network for occlusion-robust object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2019.

[12] S. Lee, H. G. Kim, and Y. M. Ro, "Stan: Spatio-temporal adversarial networks for abnormal event detection," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2018, pp. 1323–1327.

[13] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *CoRR*, vol. abs/1411.1784, 2014.

[14] K. Sohn, H. Lee, and X. Yan, "Learning structured output representation using deep conditional generative models," in *NIPS*, pp. 3483–3491. 2015.

[15] A. Siarohin, E. Sangineto, S. Lathuilire, and N. Sebeu, "Deformable gans for pose-based human image generation," in *CVPR*, June 2018.

[16] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *NIPS*, 2017, pp. 405–415.

[17] L. Ma, Q. Sun, S. Georgoulis, L. V. Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *CVPR*, 2018.

[18] B. Zhao, X. Wu, Z. Cheng, H. Liu, Z. Jie, and J. Feng, "Multi-view image generation from a single-view," in *Proceedings of the 26th ACM International Conference on Multimedia*, 2018, pp. 383–391.

[19] N. Neverova, R. Alp Guler, and I. Kokkinos, "Dense pose transfer," in *ECCV*, 2018.

[20] O. Ronneberger, P.Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, vol. 9351 of *LNCS*, pp. 234–241.

[21] M. Park, H. G. Kim, and Y. M. Ro, "Photo-realistic facial emotion synthesis using multi-level critic networks with multi-level generative model," in *MultiMedia Modeling*, Cham, 2019, pp. 3–15, Springer International Publishing.

[22] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations.," in *CVPR*, 2016, pp. 1096–1104.

[23] Z. Cao, T. Simon, S. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017, vol. 00, pp. 1302–1310.

[24] D. Kingma and J. Ba, "Adam: A method for stochastic optimization.," *CoRR*, vol. abs/1412.6980, 2014.

[25] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity.," *IEEE Trans. Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.