Generating Person Images with Appearance-aware Pose Stylizer

Siyu Huang[1], **Haoyi Xiong**[1], **Zhi-Qi Cheng**[2], **Qingzhong Wang**[3], **Xingran Zhou**[4], **Bihan Wen**[5], **Jun Huan**[6] and **Dejing Dou**[1]

[1]Baidu Research [2]Carnegie Mellon University [3]City University of Hong Kong
[4]Zhejiang University [5]Nanyang Technological University [6]Styling AI

{huangsiyu, xionghaoyi, doudejing}@baidu.com, zhiqic@cs.cmu.edu, qingzwang2-c@my.cityu.edu.hk, xingranzh@zju.edu.cn, bihan.wen@ntu.edu.sg, lukehuan@shenshangtech.com

Abstract

Generation of high-quality person images is challenging, due to the sophisticated entanglements among image factors, e.g., appearance, pose, foreground, background, local details, global structures, etc. In this paper, we present a novel end-to-end framework to generate realistic person images based on given person poses and appearances. The core of our framework is a novel generator called Appearance-aware Pose Stylizer (APS) which generates human images by coupling the target pose with the conditioned person appearance progressively. The framework is highly flexible and controllable by effectively decoupling various complex person image factors in the encoding phase, followed by re-coupling them in the decoding phase. In addition, we present a new normalization method named adaptive patch normalization, which enables region-specific normalization and shows a good performance when adopted in person image generation model. Experiments on two benchmark datasets show that our method is capable of generating visually appealing and realistic-looking results using arbitrary image and pose inputs.

1 Introduction

Generating realistic-looking human images is of great value in many tasks such as surveillance data augmentation [Zheng *et al.*, 2017] and video forecasting [Walker *et al.*, 2017; Wang *et al.*, 2018b]. In this work[1], we focus on the pose-guided person image generation [Ma *et al.*, 2017] which aims to transfer person images from one pose to other poses. The generated person is expected to accord with the conditioned pose structure as well as preserving the appearance details of the source person. Fig. 1 provides some pose transfer results generated by our proposed framework as examples. The pose-guided person image generation is very challenging due to the following aspects: (1) The distributions of clothes, body appearance, backgrounds, and poses vary largely between human images; (2) One person of different poses may have very different visual features; (3) The generative model usually needs to infer the appearance details of body parts which are unobserved in input images.

Figure 1: Examples of pose-guided person generation. The conditioned source images are shown at left. The target postures and the person images generated by our method are shown at right. Our method shows realistic and appealing results.

To address the above challenges, it is essential to decouple the complex entanglements, such as the interplay between appearance and pose [Esser *et al.*, 2018], in person generation procedure. Towards that goal [Ma *et al.*, 2018] introduced a learning-based pipeline to disentangle and encode three factors: image foreground, background, and pose into separated representations and then decode them back to a person image. Although the three factors are successfully decoupled by the encoder, the representations from three factors are simply concatenated into latent codes before they are fed into the decoder, resulting in a lack of controllability and interpretability in the decoding phase. We note that such a mode decoupling is significant for the decoding phase. From the application perspective, the users prefer a more controllable image editing process brought by flexible input modes. From the engineering perspective, the experts would lean to a generator with more interpretability, such that distinct feedback can be collected from disentangled modes to better guide model de-

[1]Code is available at *https://github.com/siyuhuang/PoseStylizer*

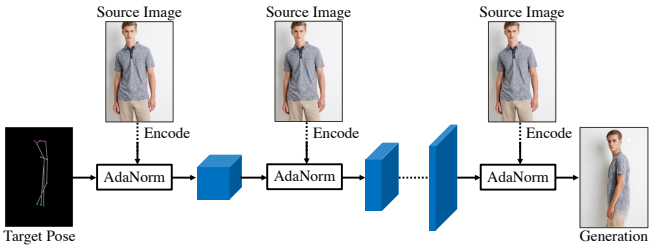

Figure 2: A simplified diagram of our APS model. The target pose is coupled with the encoded source appearance through AdaNorm module, progressively.

sign, tuning, and optimization.

Motivated by the above observations, we investigate a strategy to decouple and re-couple the entanglement factors including *appearance–pose*, *foreground–background*, and *local detail–global structure* in person images. We propose a novel end-to-end person image generation framework consisting of an appearance encoder and an Appearance-aware Pose Stylizer (APS): The appearance encoder learns the appearance representation of the person image; APS, as shown in Fig. 2, is the image generator which progressively couples the target pose map with the appearance representation, enabling a natural re-coupling of pose and appearance. We additionally adopt an attention mechanism in both encoder and generator to disentangle the foreground and background. In APS, the image is progressively synthesized from small to large scale, thus local details and global structures are fused and preserved in a multi-scale approach..

In summary, the proposed end-to-end framework can effectively decouple the entanglements between appearance, pose, foreground, background, local details, and global structures, and re-couple them in the generator, thus generate high-quality person images in a highly flexible and controllable way. The contributions of this paper are summarized below. (1) We propose a novel person image generation framework to make explicit disentanglement of the complex factors in human images. (2) We propose a new normalization method called adaptive patch normalization. It enables normalization within local regions and is suited to the spatially-dependent generative tasks including person image generation. (3) We have conducted extensive quantitative and qualitative experiments, and ablation studies to validate the effectiveness of the proposed methods.

2 Related Work

Person image generation is very valuable in many real-world applications [Wei *et al.*, 2018; Chan *et al.*, 2019]. Various settings of person image generation have been proposed in the literature. [Lassner *et al.*, 2017] proposed 3D pose representation to generate images of a person with different clothes. [Zhao *et al.*, 2018] generated multi-view cloth images from a single view cloth image. Several work on virtual try-on [Han *et al.*, 2018; Wang *et al.*, 2018a; Zanfir *et al.*, 2018; Dong *et al.*, 2019] including FashionGAN [Zhu *et al.*, 2017] proposed to manipulate the clothes of a given person while maintaining the person identity and pose.

In this work we focus on the pose-guided person image generation [Ma *et al.*, 2017] which aims at generating images of a person with different poses but with the same clothes and identity. Based on generative adversarial networks (GANs) [Goodfellow *et al.*, 2014; Mirza and Osindero, 2014], several efforts [Neverova *et al.*, 2018; Song *et al.*, 2019; Zhou *et al.*, 2019] have been made towards this goal. More related to this work, [Ma *et al.*, 2018] proposed to disentangle image foreground, background, and pose into separated representations by the encoder. [Zhu *et al.*, 2019] proposed to transfer person pose in the encoding phase using an attention-based progressive model [Karras *et al.*, 2018]. Both the above methods disentangles image factors in the encoding phase and combines the disentangled representations before decoding, lacking explicit cross-modal re-coupling in the decoding phase.

Our approach is also related to adaptive normalization-based GANs [Karras *et al.*, 2019; Park *et al.*, 2019] which progressively transforms a constant vector or a random noise (namely, *content*) into an image using a stack of convolutional layers and normalization layers with learned coefficients (namely, *style*). More recently, [Yildirim *et al.*, 2019] applied StyleGAN [Karras *et al.*, 2019] to fashion image generation, where the content input is a constant vector and the style input is a combination of clothes and pose information. Different from existing normalization-based generative models, we set pose and appearance information of a person as content and style inputs, respectively. Our approach enables a more natural re-coupling of cross-modal representations, thus it is more effective and efficient in articulated-object generation problems.

3 Our Approach

The goal of pose-guided person image generation is to generate a person image I_g which is expected to follow a given person pose P_s while keep the appearance details of a given source person image I_s. In this paper, we propose an end-to-end generative framework including an appearance encoder and an Appearance-aware Pose Stylizer (APS) to address this challenging task.

3.1 Appearance Encoder

We use an attention-based appearance encoder to learn the appearance representation of the source image I_s. The appearance encoder is built upon a stack of L encoder blocks and the architecture is shown in the left part of Fig. 3. The appearance encoder has two network streams, the image stream and pose stream. The two streams take the outputs of two streams in layer $l-1$ as their inputs and output I_s^l and P_s^l, respectively. The input of image stream is source image I_s such that image stream extracts the visual representation of source image. The input of pose stream is source pose P_s such that pose stream learns the structure of source pose to guide the information flow in image stream with attention mechanism [Zhu *et al.*, 2019]. More specifically, the pose input P_s^l from previous layer goes through a convolutional layer and a sigmoid activation layer to obtain foreground attention masks $M_s^l \in (0,1)$ as

$$M_s^l = \sigma\left(\text{conv}_P\left(P_s^{l-1}\right)\right) \quad (1)$$

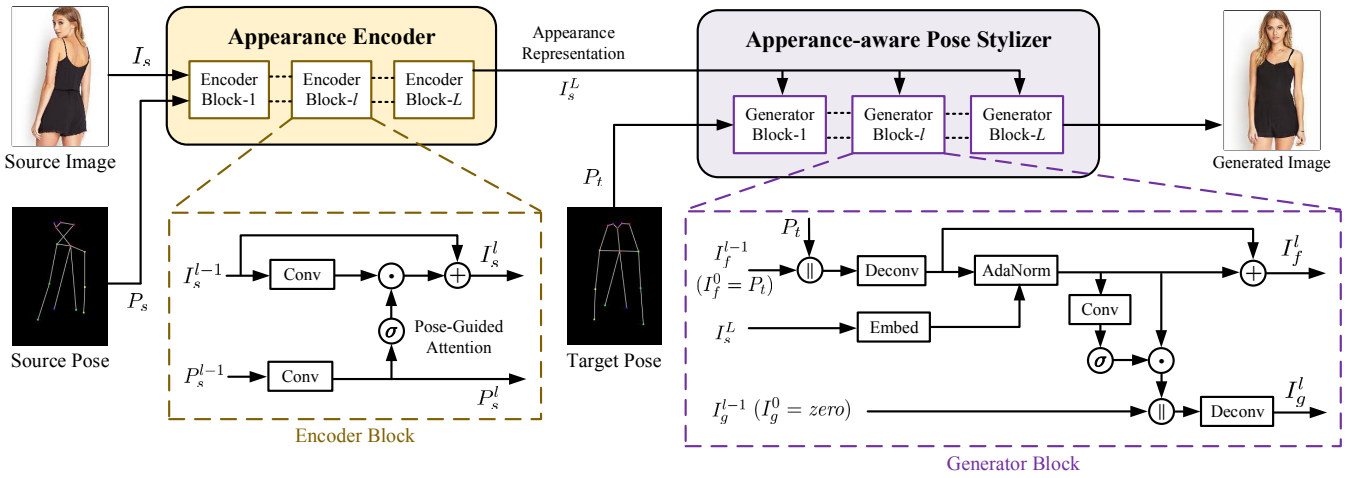

Figure 3: The schematic of the proposed person image generation framework. The scheme consists of an appearance encoder which learns the person appearance representation and an Appearance-aware Pose Stylizer which generates the target image. Notations: σ sigmoid activation, $\cdot$ element-wise multiplication, $+$ element-wise addition, $||$ channel-wise concatenation.

Then, M_s^l masks the image stream using element-wise multiplication as

$$I_s^l = M_s^l \odot \text{conv}_I\left(I_s^{l-1}\right) + I_s^{l-1} \tag{2}$$

The residual connection in Eq. 2 eases the learning of deep generative models in practice. With pose-guided attention, image stream is forced to focus on the foreground of image, i.e., the person.

3.2 Appearance-aware Pose Stylizer

Appearance-aware Pose Stylizer (APS) is a novel generator for realistic person image generation. As shown in Fig. 2, the pose stylizer stylizes the target pose map P_t under guidance of appearance representation, progressively. Similar to appearance encoder, APS consists of L repetitive generator blocks to restore images from small sizes to large sizes. In the following we discuss more details of the generator blocks.

Generator Block

In a generator block, there are two network streams including foreground image stream and synthesized image stream, as shown in the right part of Fig. 1. Foreground image stream I_f generates foreground person images based on pose maps P_t and appearance representations I_s^L. Synthesized image stream I_g synthesizes complete images including both foregrounds and backgrounds. In the foreground stream, we adopt adaptive normalization mechanism (AdaNorm) to stylize pose maps based on appearance representations, such that our generator is named as Appearance-aware Pose Stylizer. AdaNorm(x, y) accepts content feature x and style feature y

$$x = \text{deconv}_f\left(\text{concat}\left(I_f^{l-1}, P_t\right)\right) \tag{3}$$

$$y = \text{embed}\left(I_s^L\right) \tag{4}$$

deconv is a deconvolutional layer, concat is the channel-wise concatenation operation, embed is a convolutional layer with a kernel size of 1×1. $I_f^0 = P_t$ such that content x is derived from pose map P_t. Style y is derived from appearance representation I_s^L. Note that the setting of content and style in this work is distinctly different from existing normalization-based generative models such as StyleGAN [Karras *et al.*, 2019] and SPADE [Park *et al.*, 2019], in which the content x is generally set as a constant vector or a random vector. In our setting, AdaNorm naturally disentangles pose and appearance of a person as its content and style inputs, leading to a reasonable feature fusion as well as an effective person generation pipeline.

After computing $z = \text{AdaNorm}(x, y)$, the foreground stream outputs $I_f^l = z + x$. The foreground stream fuses into the synthesized stream with attention mechanism,

$$z_{\text{Att}} = \sigma\left(\text{conv}_f\left(z\right)\right) \odot z \tag{5}$$

The synthesized stream outputs I_g^l as

$$I_g^l = \text{deconv}_g\left(\text{concat}\left(I_g^{l-1}, z_{\text{Att}}\right)\right) \tag{6}$$

Specifically, I_g^0 is feature maps with *zero* values. I_f^L and I_g^L output by the last generator block are concatenated and decoded as the generated image with an 1×1 convolutional layer.

Adaptive Patch Normalization

Here we discuss the AdaNorm module used in our method. In existing literature, StyleGAN and SPADE successfully applied adaptive instance normalization (AdaIN) [Huang and Belongie, 2017] to progressive generative models. AdaIN is formulated as

$$\text{AdaIN}(x_c, y) = y_c^w\left(\frac{x_c - \beta(x_c)}{\gamma(x_c)}\right) + y_c^b \tag{7}$$

where x and y are content and style features respectively. c denotes the channel number, β and γ denote mean and standard deviation. The weight term y_c^w and bias term y_c^b are embedded from input style feature y. In Eq. 7, content feature

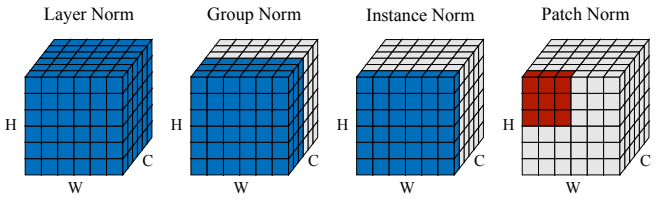

Figure 4: Illustrations of normalization methods. Compared with existing methods, patch normalization enables region-specific normalization parameters.

x is first normalized by instance normalization (IN) and then scaled by parameters conditioned on style feature y.

In this paper we develop a new normalization method, i.e., adaptive patch normalization (AdaPN)

$$\text{AdaPN}(x_{c,i,j}, y) = y^w_{c,i,j} \left(\frac{x_{c,i,j} - \beta(x_{c,i,j})}{\gamma(x_{c,i,j})} \right) + y^b_{c,i,j} \quad (8)$$

where i, j denote the spatial position of a patch on feature map. The content feature is first normalized by IN and then scaled with region-specific parameters $y^w_{c,i,j}$ and $y^b_{c,i,j}$. [2] Fig. 4 illustrates the difference between existing normalization methods. The proposal of patch normalization is motivated by the observation that human body parts within a specific spatial region are relatively fixed among well-cropped person images. For instance, the head usually appears in the top-center area of an image while the legs usually appear in the bottom area of an image. Therefore, it is natural to normalize spatial regions with different factors. Compared with AdaIN, AdaPN induces the generator to learn appearance details of different body parts more effectively. In Section 4.4, we take insights into AdaPN by conducting more empirical studies.

3.3 Training and Optimization

The loss function that we use to train our person image generation model is comprised of the adversarial loss $\mathcal{L}_{\text{GAN}}$, the reconstruction loss L_1, and the perceptual loss $\mathcal{L}_{\text{per}}$ as follow:

$$\mathcal{L} = \arg\min_G \max_D \alpha\mathcal{L}_{\text{GAN}} + \lambda_1\mathcal{L}_{L1} + \lambda_2\mathcal{L}_{\text{per}} \quad (9)$$

In Eq. 9, L_1 loss $\mathcal{L}_{L1} = \|I_g - I_t\|_1$ where I_g is the generated image and I_t is the ground-truth target image. The perceptual loss [Johnson *et al.*, 2016] computes L_1 distance over feature maps. The adversarial loss $\mathcal{L}_{\text{GAN}}$ consists of the appearance-consistency term and the pose-consistency term, formulated as

$$\begin{aligned}\mathcal{L}_{\text{GAN}} = \mathbb{E}[& \underbrace{\log D_a(I_s, I_t) + \log(1 - D_a(I_s, I_g))}_{\text{appearance-consistency term}} \\ & + \underbrace{\log D_p(P_t, I_t) + \log(1 - D_p(P_t, I_g))}_{\text{pose-consistency term}}]\end{aligned} \quad (10)$$

where D_a and D_p are discriminators. $(I_s, I_t) \sim \mathbb{I}_{\text{real}}, I_g \sim \mathbb{I}_{\text{fake}}, P_t \sim \mathbb{P}$. $\mathbb{I}_{\text{real}}$, $\mathbb{I}_{\text{fake}}$, $\mathbb{P}$ are the distribution of real images, fake images, and person poses, respectively.

[2] In implementation of AdaPN, in each generator block we embed I^L_s to latent features and then tile the features into y^w and y^b. The sizes of y^w and y^b are the same to the size of corresponding x.

Network architectures. For both encoder and generator, we adopt a total block number $L = 4$ on Market-1501 dataset and $L = 5$ on DeepFashion dataset, respectively. The first layer of encoder and the last layer of generator has 64 channels. The number of channels in every block is doubled/halved in encoder/generator until a maximum of 512. The size of feature map in every block is halved/doubled in encoder/generator using stride-2 convolutions/deconvolutions. The appearance representation I^L_s has a shape of $512 \times \frac{H}{2^L} \times \frac{W}{2^L}$, where H and W is height and width of the input image. Every AdaNorm module is built up with AdaPN-Conv-AdaPN in practice for establishing a deeper model.The standard/adaptive normalization layer and ReLU are applied after every convolutional or deconvolutional layer.

Training details. We implement our model on the PyTorch framework [Paszke *et al.*, 2017]. The model is trained with an Adam optimizer [Kingma and Ba, 2014] for 800 epochs. The initial learning rate is 0.0002 and it linearly decays to 0 from 400 epochs to 800 epochs. Following [Zhu *et al.*, 2019], the loss weights (α,λ_1,λ_2) are set as (5, 10, 10) on Market-1501 and (5, 1, 1) on DeepFashion. In training Market-1501, we additionally apply Dropout [Hinton *et al.*, 2012] with a rate of 0.5 after every generator block in case of overfitting.

4 Experiments

4.1 Experimental Setups

Datasets. We conduct experiments on two benchmark person image generation datasets including Market-1501 [Zheng *et al.*, 2015] and DeepFashion (*In-shop Clothes Retrieval Benchmark*) [Liu *et al.*, 2016]. Market-1501 is a challenging person re-identification dataset which contains 32,668 images of 1,501 person identities. The images in Market-1501 are low-resolution (128×64 pixels) and the person pose, viewpoint, illumination, and background vary largely. DeepFashion contains 52,712 in-shop clothes images (256×256 pixels). We adopt OpenPose [Cao *et al.*, 2017] as our pose keypoints detector. By following the settings in [Zhu *et al.*, 2019], for Market-1501, we collect 263,632 training pairs and 12,000 testing pairs. For DeepFashion, we collect 101,966 pairs for training and 8,570 pairs for testing. Each pair is composed of two images of the same identity but different poses. The person identities in training sets do not overlap with those in testing sets.

Evaluation metrics. In this work we use Structural Similarity (*SSIM*) [Wang *et al.*, 2004] to measure the structure similarity between images, i.e., the appearance-consistency. We use the Inception Score (*IS*) [Salimans *et al.*, 2016] to measure the image quality. Following PG2 [Ma *et al.*, 2017] we adopt their masked versions *mask-SSIM* and *IS* to evaluate the image foreground only via masking out the background, since no background information is provided for person generation models. In addition, we use Percentage of Correct Keypoints (*PCKh*) [Andriluka *et al.*, 2014] to measure the pose joints alignment, i.e., the pose-consistency.

4.2 Results

Qualitative comparison. Fig. 5 shows a qualitative comparison of state-of-the-art person image generation methods on

Figure 5: Qualitative comparison of existing pose-guided person generation methods on DeepFashion. Please zoom in for details.

Model	Market-1501					DeepFashion		
	SSIM	IS	mask-SSIM	mask-IS	PCKh	SSIM	IS	PCKh
Real Data	1.000	3.890	1.000	3.706	1.00	1.000	4.053	1.00
PG2 [Ma *et al.*, 2017]	0.261	**3.495**	0.782	3.367	0.73	**0.773**	3.163	0.89
Disentangled [Ma *et al.*, 2018]	0.099	**3.483**	0.614	3.491	-	0.614	3.228	-
VUNet [Esser *et al.*, 2018]	0.266	2.965	0.793	3.549	0.92	0.763	**3.440**	0.93
Deform [Siarohin *et al.*, 2018]	0.291	3.230	0.807	3.502	**0.94**	0.760	**3.362**	0.94
PATN [Zhu *et al.*, 2019]	**0.311**	3.323	**0.811**	**3.773**	**0.94**	**0.773**	3.209	**0.96**
Ours	**0.312**	3.132	**0.808**	**3.729**	**0.94**	**0.775**	3.295	**0.96**

Table 1: Quantitative comparison of our proposed method to the state-of-the-art methods on Market-1501 and DeepFashion. The **best** and the **second-best** performances are highlighted (Higher is better for all reported metrics).

DeepFashion. The source images, pose maps, target images, and the generations are shown from the left to the right, respectively. We compare our method with Variational U-Net (VUNet) [Esser *et al.*, 2018], Deformable GANs (Deform) [Siarohin *et al.*, 2018], and Pose-Attentional Transfer Network (PATN) [Zhu *et al.*, 2019]. The example images shown in Fig. 5 vary in poses, scales, and colors and types of clothes. Our method shows appealing results in the following aspects: (1) *Realistic generations*: The generated images show natural facial details, body postures, clothing collocations, and skin appearances; (2) *Clothing-consistency*: Clothes colors and styles are consistent with those in source images; (3) *Identity-consistency*: Person identity details including facial appearances, body figures, and skin colors are well maintained in generated images; (4) *Pose-consistency*: Poses of generated persons well follow the conditioned poses.

Fig. 6 shows qualitative results on Market-1501. Although the images of this dataset are low-resolution and their visual details are somewhat blurry (as shown in Source and Target columns), we observe that our method shows a good performance in comparison with the other methods considering the sharpness of generated images.

Quantitative comparison. Table 1 quantitatively evaluate the person generation methods under a series of metrics. Our method shows a competitive performance compared with the existing methods. It achieves the best PCKh on both datasets, indicating that the generations have a good consistency with the conditioned poses. We attribute it to the natural disentanglement of pose and appearance in our APS model. On Market-1501, our model performs well on SSIM, mask-SSIM, and mask-IS. On DeepFashion, our model performs well on SSIM. However, it is relatively worse on IS. We con-

Figure 6: Qualitative comparison of existing pose-guided person generation methods on Market-1501.

Model	Market-1501						DeepFashion		
	SSIM ↑	IS ↑	L1 ↓	mask-SSIM ↑	mask-IS ↑	mask-L1 ↓	SSIM ↑	IS ↑	L1 ↓
Conv enc+Deconv dec	0.205	3.141	0.332	0.750	3.507	0.104	0.758	3.373	0.106
Conv enc+APS dec	0.301	2.985	0.286	0.804	3.661	0.080	0.766	3.335	0.101
PATN enc+Deconv dec	0.218	**3.193**	0.341	0.751	3.447	0.101	0.760	3.334	0.104
StyleGAN	0.251	2.987	0.312	0.777	3.737	0.093	0.766	**3.393**	0.101
APS w/ AdaIN	0.297	3.094	0.293	0.800	**3.755**	0.081	0.763	3.275	0.102
APS w/o attention	0.291	2.879	0.292	0.799	3.653	0.082	0.764	3.305	0.100
APS w/o decoding I_f^L	0.303	2.993	0.285	0.806	3.622	**0.079**	0.768	3.374	0.098
Full model	**0.312**	3.132	**0.281**	**0.808**	3.729	**0.079**	**0.775**	3.295	**0.097**

Table 2: Ablation studies on our person image generation model for evaluating the efficacy of different components. ↑ denotes higher is better and ↓ denotes lower is better.

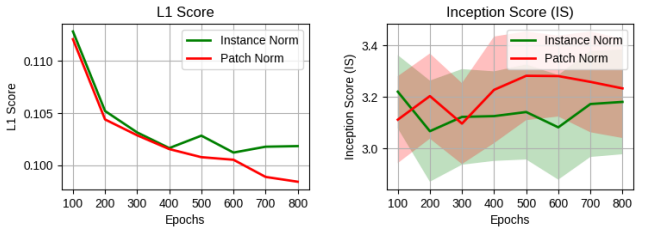

Figure 7: A comparison of AdaIN and AdaPN. We show L1 score (lower is better) and Inception Score (higher is better) vs. training epochs on testing set of DeepFashion. AdaPN is better than AdaIN under both metrics.

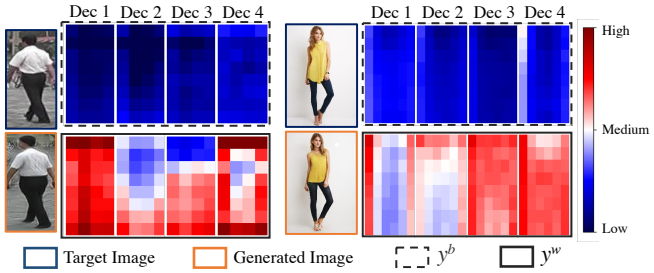

Figure 8: Visualizations of AdaPN statistics, including bias y^b and scaling factor y^w, in individual generator blocks.

jecture it is because the model well restores the visual appearance details, while, a certain level of overfitting may hurt the realness performance.

4.3 Ablation Study

We conduct ablation study on our person generation model to evaluate the efficacy of different components proposed in this paper. The first part of Table 2 evaluates different encoders and decoders. With the same encoders (Conv encoder or PATN encoder), our APS decoder shows large improvements over Deconv decoder. The second part of Table 2 evaluates different components of APS generator. In StyleGAN [Karras *et al.*, 2019], the content input is a constant vector and the style input contains both pose and appearance representation. APS shows significant improvements over StyleGAN, demonstrating that the distinct disentanglement of pose and appearance in generator can benefit the performance of person image generation. APS w/ AdaIN replaces the AdaPN with AdaIN in APS, and the results show that our proposed AdaPN is more suited to spatially-dependent generative tasks. APS w/o attention removes all the attention modules in encoder and decoder, and the results indicate that attention mechanism can slightly help the APS model. Decoding both I_f^L and I_g^L is better than only decoding the synthesized stream I_g^L.

4.4 Study on Adaptive Normalization

Adaptive normalization is the core of our APS model. Fig. 7 compares AdaIN and AdaPN on testing set of DeepFashion, vs., the training epochs. Before 400 epochs, AdaIN and AdaPN shows similar L1 scores. After 400 epochs, L1 scores of AdaIN does not decrease anymore, while, L1 scores of AdaPN is decreasing continuously until the end of training. It reveals that AdaPN has a larger model capacity than AdaIN in learning, thus to better reconstruct the high resolution images. AdaPN also shows better Inception Scores on most of the epochs, demonstrating its superiority in person generation model.

In Fig. 8 we visualize the AdaPN statistics of a fully trained model. The statistics of individual generator blocks, including bias y^b and scaling factor y^w, are shown at right. We notice that the biases of different positions are similar within a layer. Conversely, the scaling factors of different positions vary largely, suggesting that the local-version scaling is valuable in normalization-based person generation, while, the local-version translation is not much necessary. Intuitively, the scaling operation is related with the input variables. Compared with the bias term, the scaling term contributes more to variations within the output. This leads to the necessity of locally sensitive scaling factors.

5 Conclusion

In this paper, we have presented a novel framework for generating realistic person images. The framework decouples the image factors by an attention-based appearance encoder and re-couples the image factors by an APS generator. It is effective, controllable, and flexible since it makes explicit disentanglement of the complex factors in human images. We have also proposed AdaPN which enables local-specific normalization for spatially-dependent generative tasks. Extensive experiments on benchmark datasets have validated the effectiveness of our approach over existing methods.

References

[Andriluka *et al.*, 2014] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *CVPR*, 2014.

[Cao *et al.*, 2017] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017.

[Chan *et al.*, 2019] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A Efros. Everybody dance now. In *ICCV*, 2019.

[Dong *et al.*, 2019] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *ICCV*, 2019.

[Esser *et al.*, 2018] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *CVPR*, 2018.

[Goodfellow *et al.*, 2014] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[Han *et al.*, 2018] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018.

[Hinton *et al.*, 2012] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012.

[Huang and Belongie, 2017] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, 2017.

[Johnson *et al.*, 2016] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016.

[Karras *et al.*, 2018] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *ICLR*, 2018.

[Karras *et al.*, 2019] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

[Kingma and Ba, 2014] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[Lassner *et al.*, 2017] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *ICCV*, 2017.

[Liu *et al.*, 2016] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, 2016.

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

[Ma *et al.*, 2018] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *CVPR*, 2018.

[Mirza and Osindero, 2014] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[Neverova *et al.*, 2018] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *ECCV*, 2018.

[Park *et al.*, 2019] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, 2019.

[Paszke *et al.*, 2017] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in PyTorch. In *NIPS Autodiff Workshop*, 2017.

[Salimans *et al.*, 2016] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NIPS*, 2016.

[Siarohin *et al.*, 2018] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *CVPR*, 2018.

[Song *et al.*, 2019] Sijie Song, Wei Zhang, Jiaying Liu, and Tao Mei. Unsupervised person image generation with semantic parsing transformation. In *CVPR*, 2019.

[Walker *et al.*, 2017] Jacob Walker, Kenneth Marino, Abhinav Gupta, and Martial Hebert. The pose knows: Video forecasting by generating pose futures. In *ICCV*, 2017.

[Wang *et al.*, 2004] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004.

[Wang *et al.*, 2018a] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018.

[Wang *et al.*, 2018b] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Guilin Liu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. Video-to-video synthesis. *arXiv preprint arXiv:1808.06601*, 2018.

[Wei *et al.*, 2018] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *CVPR*, 2018.

[Yildirim *et al.*, 2019] Gokhan Yildirim, Nikolay Jetchev, Roland Vollgraf, and Urs Bergmann. Generating high-resolution fashion model images wearing custom outfits. In *ICCV Workshops*, 2019.

[Zanfir *et al.*, 2018] Mihai Zanfir, Alin-Ionut Popa, Andrei Zanfir, and Cristian Sminchisescu. Human appearance transfer. In *CVPR*, 2018.

[Zhao *et al.*, 2018] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. Multi-view image generation from a single-view. In *ACM MM*, 2018.

[Zheng *et al.*, 2015] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *ICCV*, 2015.

[Zheng *et al.*, 2017] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *ICCV*, 2017.

[Zhou *et al.*, 2019] Xingran Zhou, Siyu Huang, Bin Li, Yingming Li, Jiachen Li, and Zhongfei Zhang. Text guided person image synthesis. In *CVPR*, 2019.

[Zhu *et al.*, 2017] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *ICCV*, 2017.

[Zhu *et al.*, 2019] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *CVPR*, 2019.