Learning Semantic Person Image Generation by Region-Adaptive Normalization

Zhengyao Lv[1], Xiaoming Li[1], Xin Li[2], Fu Li[2], Tianwei Lin[2], Dongliang He[2], Wangmeng Zuo[1,3](✉)
[1]School of Computer Science and Technology, Harbin Institute of Technology, China
[2]Department of Computer Vision Technology (VIS), Baidu Inc.
[3]Pazhou Lab, Guangzhou, China
cszy98@gmail.com {csxmli, wmzuo}@hit.edu.cn

Abstract

Human pose transfer has received great attention due to its wide applications, yet is still a challenging task that is not well solved. Recent works have achieved great success to transfer the person image from the source to the target pose. However, most of them cannot well capture the semantic appearance, resulting in inconsistent and less realistic textures on the reconstructed results. To address this issue, we propose a new two-stage framework to handle the pose and appearance translation. In the first stage, we predict the target semantic parsing maps to eliminate the difficulties of pose transfer and further benefit the latter translation of per-region appearance style. In the second one, with the predicted target semantic maps, we suggest a new person image generation method by incorporating the region-adaptive normalization, in which it takes the per-region styles to guide the target appearance generation. Extensive experiments show that our proposed SPGNet can generate more semantic, consistent, and photo-realistic results and perform favorably against the state of the art methods in terms of quantitative and qualitative evaluation. The source code and model are available at https://github.com/cszy98/SPGNet.git.

1. Introduction

Given an image of a specific person under a certain perspective or pose, the human pose transfer task aims to generate images of the person with the same appearance and meanwhile under the given target pose, which more importantly, are expected to be as photo-realistic as possible. The task has a wide range of applications, such as generating videos with pose sequences [2, 8, 12, 19, 40], data augmentation for person re-identification task [28, 47] and multi-view display for virtual try-on [14, 41, 42].

Basically, human pose transfer is a non-rigid deformation of the 3D human. Only using one 2D source image and target pose to generate another view of the human body is still a challenging task due to the following difficulties: (i) spatial rearrangement of the appearance features, (ii) the inference of the self-occlusion region, (iii) the photo-realistic results. All these make the task valuable and remain an active topic in the community of computer vision.

On the one hand, along with the rapid development of deep learning, cGAN [24] based human pose transfer methods [22, 27, 36] have achieved significant progress. However, this task is an unaligned image to image translation due to the inconsistent poses between the source and target images. Directly taking the concatenation of the source appearance and target pose into a general encoder-decoder framework cannot fully exploit the correspondence between the source and target appearance, thus resulting in less realistic results. On the other hand, to facilitate the spatial rearrangement of the source appearance, deformation [18, 19, 29, 31] and disentanglement [5, 23] based methods have been suggested. Albeit the appearance alignment problem can be addressed, the generated distortion and blurry appearance caused by the warped and disentangled features are inevitable, giving rise to visually unpleasing results. The recent progressive generation [35, 47] also achieved plausible performance, but the final results usually lose semantic details. Though improvements have been obtained, the semantic generation remains uninvestigated reasonably in the human pose transfer task.

In this paper, we present a two-stage framework to address the above issues. In the first stage, we predict the target semantic parsing maps. Compared with these directly generating the target image, the prediction of target parsing maps is much easier, because we do not need to consider the texture translation and this will make the network focus on only one single task, *i.e.*, pose transfer without considering the effect of appearance that may bring the burden to the learning of the network. With the given pose and source parsing map, we suggest a SPATN model to generate the target semantic maps in a progressive manner. We observe that the target semantic maps can bring the following benefits: (i) it can not only provide the pose information, but

also the specification of each body region, making our model more robust when dealing with complex poses, (ii) with the per-region styles, it can help to transfer them to the target region separately, which is more effective in generating the semantic and photo-realistic results. In the second stage, we propose a SPGNet by incorporating the semantic region adaptive normalization to generate the target image. To be specific, it takes the per-region styles that are extracted from the source appearance and then broadcasted to the target regions to assist the semantic generation on each body part, resulting in the final photo-realistic results. This manner can well address the inconsistent poses and self-occlusion problems. We note that there are also some works that attempt to utilize the predicted target parsing maps to improve the generation of the final results [4, 23, 33]. However, Dong *et al*. [4] and Song *et al*. [33] both directly take the predicted target parsing map and source image as input. This is more like a cGAN, which can bring limited improvements to the final semantic reconstruction. Men *et al*. [23] also suggest the extraction of per-region style codes, but these codes are utilized in the reconstruction process by the AdaIN layer [11], in which the learned affine parameters from the source appearance are uniform across spatial coordinates. We argue that compared with the global normalization, the region adaptive normalization is more flexible and suitable for this task by specifying the style to each target region.

Experiments are conducted to evaluate the effectiveness of our proposed SPGNet on two challenging datasets, *i.e*., DeepFashion [21] and Market-1501 [45]. Results show that our SPGNet performs favorably against the state of the arts, and yields more consistent and photo-realistic results. The main contributions of this work include:

- We predict the target semantic maps as a supplement to the pose representation, and then further utilize it to the per-region style translation by region adaptive normalization, thereby eliminating the difficulties of the pose transfer and improving the reconstruction quality.

- In comparison with these state of the arts methods, experiments demonstrate the superiority of our SPGNet in generating favorable and photo-realistic person image.

2. Related Work

2.1. Human Pose Transfer

The task of human pose transfer has achieved great development during these years, especially with the unprecedented success of deep learning. Originally, these methods regard this task as a conditional image generation by taking the source appearance image as a condition to guide the generation of the target image, which mainly comes from the conditional generative adversarial networks (cGANs) [24]. Ma *et al*. [22] first present a two-stage model, which directly concatenated the target pose and the source image as input to generate the target image in a coarse-to-fine manner. Pumarola *et al*. [27] suggest an unsupervised manner by taking the concatenation of the generated results and the source pose as input to reconstruct the source image. Similarly, Tang *et al*. [36] propose a cycle-in-cycle way to constrain the learning of pose and appearance translation. However, directly concatenating the target pose and source image usually brings limited improvements due to the inconsistent poses. To solve the above issues that hinder the appearance translation, Essner *et al*. [5] adopt the combination of VAE [17] and U-Net [13] to disentangle the appearance and the pose of person images. Men *et al*. [23] also adopt the disentanglement by extracting the region styles to perform on the whole target pose. Siarohin *et al*. [31] proposed a Deformable GAN, which decomposes the overall deformation by a set of local affine transformations to deal with the misalignments caused by different poses. Subsequently, in order to enhance the spatial rearrangement ability, there are also some works [18, 19, 29] that use flow-based deformation to align the source appearance. Both [18] and [19] adopt additional 3D human models to calculate the flow fields between the source and target image, while Ren *et al*. [29] obtain the global flow fields in an unsupervised manner and further propose a local neural texture render. Recently, Zhu *et al*. [47] propose to progressively transform the source image by a sequence of pose-attentional transfer blocks (PATBs), which is flexible but useful information may be lost during multiple transfers. Based on these observations of this work, Tang *et al*. [35] further propose a XingGAN model, which consists of shape-guided appearance-based generation branch, appearance-guided shape-based generation branch as well as co-attention fusion module to effectively transfer and update person shape and appearance features in a crossing and progressive manner. However, nearly all these methods directly utilize the global appearance features on the target one, and seldom explore the benefits of per-region appearance translation that may bring to the final reconstruction.

On the other hand, most of these aforementioned methods use human keypoints as pose representation due to its cheapness. However, it ignores the body skeleton that is useful to build the human body. There are also some other works that use DensePose [25] or semantic parsing maps [4, 33] as pose representation, which can provide more information about body depth or part segmentation. However, they directly take the semantic maps as a condition that is concatenated to the pose image along the channel dimension. Therefore, they can achieve only limited improvements on the target results especially the semantic appearance of each region due to the misalignment. Here, we also use the target semantic parsing map to assist the image generation but utilize it through the region adaptive normalization on each body region to

separately guide the semantic appearance generation.

2.2. Semantic Image Generation

Semantic image generation aims at synthesizing photo-realistic images from the given semantic layout. Several methods [13, 20, 26, 37, 38, 46] have been suggested for solving this task. SPADE [26] adopts semantic label maps to predict the spatially-varying affine transformation parameters for incorporating the class prior to the target image, which controls the image generation process more precisely and obtains visual fidelity results. Similarly, CC-FPSE [20] proposes to predict the spatially-varying conditional convolutional kernels based on the input semantic layout. LGGAN [37] takes global and local contexts into account and gets promising performance. The image-level global generator learns a global appearance distribution and the class-specific local generator generates different object classes separately. Furthermore, SEAN [46] improves SPADE by introducing per-region style encoding, which is better suited to encode, transfer, and synthesize style in terms of visual quality. Inspired by the benefit brought by SEAN [46], in this work, we handle the person image generation by separate per-region style translation. To be specific, we adopt the region adaptive normalization on each semantic region to retain and transfer the style code between each corresponding semantic part of the source and target images.

3. Method

Given the source person image I_{p_s} in pose p_s and target pose p_t, our goal is to generate a photo-realistic image $\hat{I}_{p_t}$, which has the consistent appearance with I_{p_s} while under the pose p_t. Due to the large changes of views and the unseen regions caused by self-occlusion, directly predicting the desired result tends to be intractable. To address this issue, we propose a two-stage framework to eliminate these difficulties. In the first stage, we generate the semantic parsing maps $\hat{S}_{p_t}$ of the target pose from the source parsing maps S_{p_s} , source pose p_s and target pose p_t in a progressive manner. In the second one, we use the predicted target semantic parsing maps $\hat{S}_{p_t}$ to guide the target image generation by the semantic region-adaptive normalization layers. The overall framework of the method is shown in Fig. 1.

3.1. Target Semantic Parsing Map Generation

Instead of directly predicting the target person image $\hat{I}_{p_t}$, the prediction of the target parsing map tends to be much easier due to the lack of texture translation. However, it still suffers from the large pose changes between the source and the target parsing maps. Inspired by the PATN [47], which proposed a progressive manner to handle the pose image transfer, here we adopt a semantic pose-attentional transfer network (SPATN) to model the semantic translation. The framework is shown in the left part of Fig. 1 (a). It takes the source pose p_s, target pose p_t and source semantic map S_{p_s} as input to predict the target semantic map $\hat{S}_{p_t}$, which can be formulated as:

$$\hat{S}_{p_t} = \mathcal{F}_{SN}(p_s, p_t, S_{p_s}; \Theta_{SN}), \tag{1}$$

where $\mathcal{F}_{SN}$ and Θ_{SN} denote the proposed SPATN model and its learnable parameters, respectively.

We observe that using sparse keypoint to represent the pose can only provide limited body structure information and it ignores the correspondence of each part, which usually fails to deal with some complex poses (*e.g.* crossed arms). In order to better model the pose structure and further benefit the generation of target semantic parsing, we introduce the distance map as a complementary representation of pose structure. Originally, the pose p is constructed by 18-channels heat maps that each one encodes one joint of a human body. With these 18 points, we generate 12 lines $\{L_m|_{m=1}^{12}\}$ to represent the body skeleton. In this work, each skeleton generates one channel distance map. When this skeleton is invisible or occluded, the distance map is set to 0. Thus we can generate a 12-channels distance map $\{M_m|_{m=1}^{12}\}$, which has the same width and height as the source image. The values in (x, y) of each M_m is calculated by the smallest distance between the point (x, y) and the skeleton L_m. The m-*th* distance map can be obtained by:

$$M'_m(x, y) = \min_{(x',y')\in L_m} \{\sqrt{(x - x')^2 + (y - y')^2}\}, \tag{2}$$

where (x', y') denotes the point on the skeleton L_m. Here, we further normalize these values by introducing a negative parameter κ. The final distance map is then defined as:

$$M_m(x, y) = exp(\kappa * M'_m(x, y)), \tag{3}$$

where $m \in \{1, 2, 3...12\}$ represents the m-*th* skeleton. In this way, the closer the point to the skeleton, the larger the value is. Thus it can well model the body structure. In this work, κ is set to -0.1. Finally, the poses p_s and p_t can be represented by 30-channels features, respectively, which consist of 18-channels joint heat maps and the extra 12-channels distance maps. The illustration and analyses of the generated distance map M can be found in our suppl.

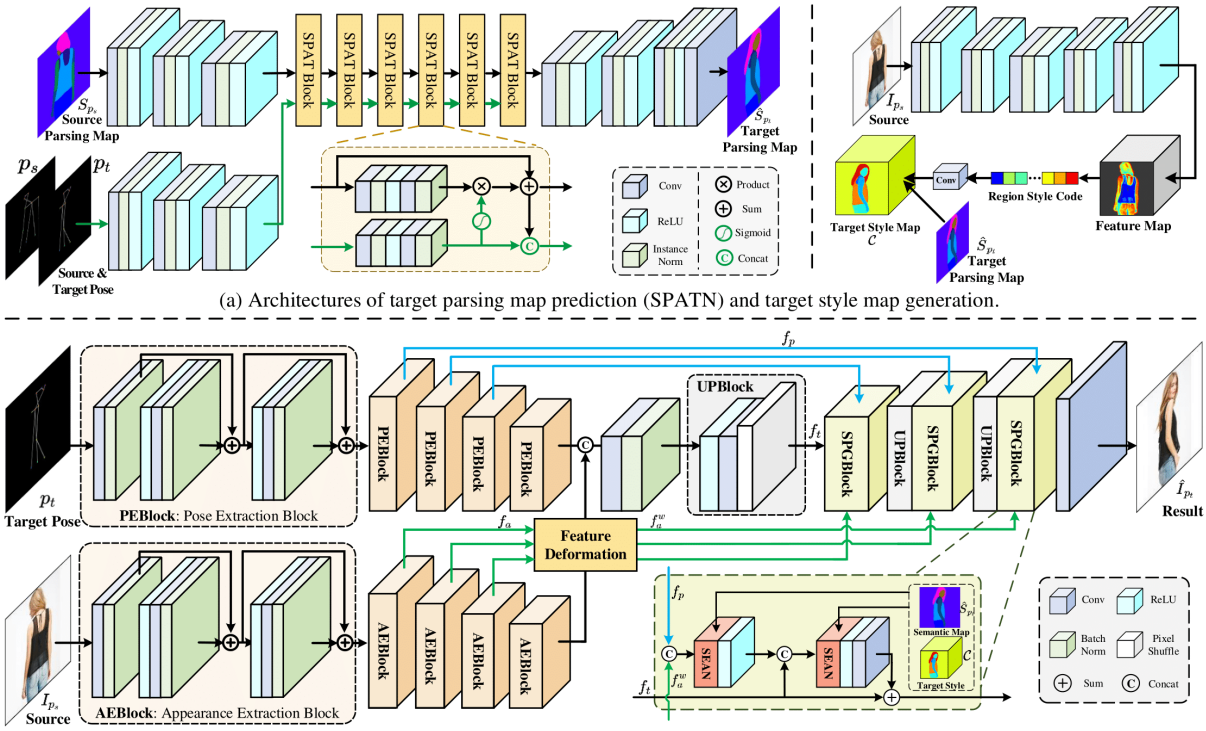
3.2. Target Image Generation

Fig. 1 (b) illustrates the framework of our proposed semantic person generation network (SPGNet). It takes the target pose p_t, source image I_{p_s}, as well as target semantic parsing map $\hat{S}_{p_t}$ (predicted through SPATN in the first stage) as input to generate a photo-realistic target image $\hat{I}_{p_t}$, which has the same appearance with I_{p_s} and the same pose with p_t. The whole model can be formulated as:

$$\hat{I}_{p_t} = \mathcal{F}_{SPG}(p_t, I_{p_s}, \hat{S}_{p_t}; \Theta_{SPG}), \tag{4}$$

where Θ_{SPG} denotes the SPGNet model parameters.

Following [18, 31, 44], we adopt a dual-path UNet [30] to encode the appearance and pose features separately. The pose features are extracted by the suggested PEBlocks, while



(a) Architectures of target parsing map prediction (SPATN) and target style map generation.

(b) Architecture of semantic person image generation.

Figure 1. Overview of our proposed method. It mainly contains two stages. In the first one, SPATN is adopted to generate the target parsing map $\hat{S}_{p_t}$. In the second one, feature deformation is utilized to warp the source appearance feature to the target pose, and SPGBlock is progressively introduced to incorporate the semantic region adaptive normalization on the generation of the target result $\hat{I}_{p_t}$.

the appearance features are obtained through the AEBlocks (see Fig. 1 (b)). The PEBlock and AEBlock have the same architectures, except the channel numbers in the first block (30 for PEBlock and 3 for AEBlock). Each block is constructed by two residual blocks [9]. Inspired by [18], here we adopt a feature deformation operation that is performed on the appearance features f_a to solve the inconsistent poses. With the warped appearance feature f_a^w and target pose feature f_p, we further propose a semantic person generation block (SPGBlock) by the region-adaptive normalization layers. The final result is generated by several stacks of UpBlock and SPGBlock in multi-scale feature spaces.

Feature Deformation. Several works have been proposed to address the appearance warping problem [2, 18, 29, 31], which usually learn the global or local spatial transformation. Among these methods, Intr-Flow [18] proposed to learn the dense and intrinsic $3D$ appearance flow by fitting a $3D$ body model, which is more suitable to the pose transfer task. With the visibility map and the accurate flow prediction, this model can well handle the deformation of the appearance feature. Denote by the feature deformation module $\mathcal{F}_W$, the input appearance feature f_a. Following [18], the warped appearance feature f_a^w can be formulated as:

$$f_a^w = \mathcal{F}_W(f_a, \Phi_{2D}, V; \Theta_w), \tag{5}$$

where Φ_{2D} is obtained from the projection of the predicted $3D$ flow and V is the visibility map. Both of them are generated by the pre-trained Intr-Flow model [18]. Θ_w is the learnable parameters of the deformation module $\mathcal{F}_W$. We adopt the same setting as [18] by incorporating a gating layer to manually exclusive the visible and invisible regions based on the visibility map V. More details can be referred to our supplementary material. In this work, we utilize the flow regression module of Intr-Flow [18].

Semantic Person Image Generation. Basically, except for the pose transfer, keeping the appearance texture is also important for the person image generation task. However, recent works usually cannot well capture the appearance and then fail to transfer to their target semantic regions. We observe that even though the source and target appearance have somewhat differences due to the large pose changes, their corresponding semantic regions should share the same appearance. For instance, the color or texture of the target clothes which can be regarded as a kind of style should be consistent with the source one. Inspired by the SEAN [46] which proposed a semantic region adaptive normalization to individually control the style of each semantic part, we utilize it in our SPGNet to facilitate the appearance translation. To obtain the per-region styles, here we firstly adopt a general encoder-decoder network by taking the source image I_{p_s} as input (as shown in the right part of Fig. 1 (a)). For each semantic region, we then generate the per-region

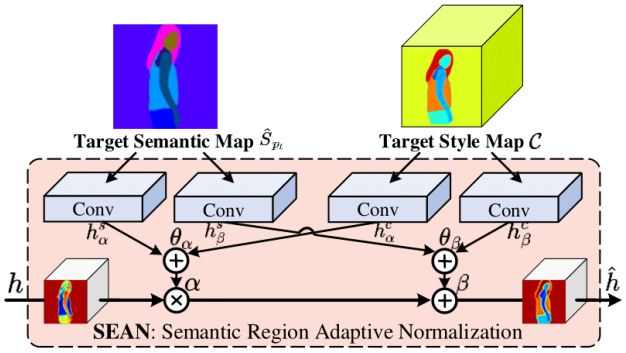


Figure 2. The details of the SEAN for semantic person generation.

style code through a region-wise average pooling layer on the feature map. Each style code is expected to contain the appearance features for each body region. After obtaining the region style codes, we adopt several convolution layers to further encode each one by excluding the effect of the other irrelevant regions. With the predicted target semantic map $\hat{S}_{p_t}$ which is generated by SPATN in the first stage, we further broadcast the style code to their target region to generate the target style map $\mathcal{C}$(see Fig. 1 (a)).

With the pose feature f_p, the warped appearance feature f_a^w, the decoder feature f_t, the predicted target semantic map $\hat{S}_{p_t}$ and the target style map $\mathcal{C}$, we suggest a semantic person generation block (SPGBlock) by incorporating the semantic region normalization to generate the target person feature. As is shown in Fig. 1 (b), each block contains two SEAN operations. The first one takes the f_p and f_a^w as input to generate the preliminary target feature. Then the second one takes the former output and the decoder feature f_t to generate the residual features of the f_t. The output of the SPGBlock is expected to fuse the appearance to the pose features on each semantic region, respectively. The details of SEAN are shown in Fig. 2.

Denote by the input feature as h, the output of SEAN in the position (n, c, y, x) is given by :

$$\hat{h}_{n,c,y,x} = \alpha_{c,y,x}\frac{h_{n,c,y,x} - \mu_c}{\sigma_c} + \beta_{c,y,x}, \tag{6}$$

where μ_c and σ_c are the mean and standard deviation of h in the channel c. α and β are the weighted sum of these features from the convolution output of the target semantic map $\hat{S}_{p_t}$ and the target style map $\mathcal{C}$, which are defined as:

$$\begin{aligned}\alpha &= \theta_\alpha \cdot h^s_\alpha + (1-\theta_\alpha)\cdot h^c_\alpha \\ \beta &= \theta_\beta \cdot h^s_\beta + (1-\theta_\beta)\cdot h^c_\beta\end{aligned} \tag{7}$$

where θ_α and θ_β are the learnable parameters. It should be noted that the target style map $\mathcal{C}$ is shared in all the SPGBlock and the parameters in the style map generation are optimized by the gradient from the final objective $\mathcal{L}_{full}$.

Finally, after several stacks of the UpBlock and the SPGBlock on different feature scales, we can generate the target results $\hat{I}_{p_t}$ in a coarse-to-fine manner.

3.3. Learning Objective

To train the proposed SPATN network, which is suggested in the first stage to predict the target semantic map $\hat{S}_{p_t}$, we adopt a cross-entropy loss to constrain the predicted $\hat{S}_{p_t}$ close to its ground-truth label S_{p_t}, which is defined as:

$$\mathcal{L}_{ce} = -\sum_{i,j}\sum_{c} S_{p_t}(i,j,c) log(\hat{S}_{p_t}(i,j,c)), \tag{8}$$

where i, j and c denote the positions of each semantic maps.

For the second stage, our proposed SPGNet is trained to generate a photo-realistic person image with the target pose p_t. The predicted $\hat{I}_{p_t}$ is expected to approximate the ground-truth image I_{p_t} in both pixel and perceptual levels, and more importantly, should follow the real person image distribution. Thus, the learning objective to train the SPGNet contains two parts, *i.e.*, reconstruction loss and photo-realistic loss.

Reconstruction Loss. In general, it contains two terms, *i.e.*, L1 loss $\mathcal{L}_{L1}$ and perceptual loss $\mathcal{L}_{perc}$ [15]. The L1 loss is defined on the pixel space to measure the appearance difference between the generated results $\hat{I}_{p_t}$ and the ground-truth I_{p_t}, which is defined as:

$$\mathcal{L}_{L1} = \frac{1}{CHW}||\hat{I}_{p_t} - I_{p_t}||_1, \tag{9}$$

where C, H and W represent the image dimensions.The perceptual loss $\mathcal{L}_{perc}$ is defined on the feature space, which is usually adopted to improve the visual quality of the generated images. In particular, we adopt the pre-trained VGG19 model [32] to extract the features in multi-scale space, which is defined as:

$$\mathcal{L}_{perc} = \sum_{k=1}^{K}\frac{1}{C_k H_k W_k}||\phi_k(\hat{I}_{p_t}) - \phi_k(I_{p_t})||_2^2, \tag{10}$$

where ϕ_k denotes the *k-th* layer of the pre-trained VGG19 model ϕ. In this work, we set $K = 4$.

Photo-realistic Loss. Adversarial loss [7] has been proved to be very effective in generating realistic results. Therefore, we also adopt it in our work to constrain the generated results $\hat{I}_{p_t}$ to be more photo-realistic. In this work, we take the triplet $\{I_{p_t}, I_{p_s}, p_t\}$ and $\{\hat{I}_{p_t}, I_{p_s}, p_t\}$ as the input of discriminator D to constrain the generated result $\hat{I}_{p_t}$ to have the same appearance with I_{p_s} and the same pose with p_t. In this work, we adopt the PatchGAN [13] to judge the realism of image patches. The objective for training the discriminator D and generator G (SPGNet in this work) is defined as:

$$\begin{aligned}\mathcal{L}_{adv}(G,D) =& \mathbb{E}_{I_{p_s},I_{p_t}}[\log(1 - D(G(p_t, I_{p_s}, \hat{S}_{p_t})|I_{p_s}, p_t))] \\ &+ \mathbb{E}_{I_{p_s},I_{p_t}}[\log D(I_{p_t}|I_{p_s}, p_t)].\end{aligned} \tag{11}$$

By taking the reconstruction loss and photo-realistic loss into consideration, the overall learning objective of our framework including SPATN and SPGNet is formulated as:

$$\mathcal{L}_{full} = \lambda_{ce}\mathcal{L}_{ce} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{perc}\mathcal{L}_{perc} + \lambda_{adv}\mathcal{L}_{adv}, \tag{12}$$

where λ_{ce}, λ_{L1}, λ_{perc}, λ_{adv} are the trade-off parameters.

4. Experiments

Extensive experiments are conducted to assess the effectiveness of our SPGNet, and compare it with the recent state of the art methods, including Def-GAN [31], Intr-Flow [18], PATN [47], GFLA [29], ADGAN [23], XingGAN [35], and BiGraphGAN [34]. Quantitative and qualitative results as well as user study are reported for a comprehensive comparison. In addition, we also conduct an ablation study to explore the benefits that our method may bring to the final results. More results can be found in our supplemental materials.

4.1. Dataset and Experimental Details

Dataset. There are two datasets that are usually used in this task, *i.e.*, DeepFashion (In-shop Clothes Retrieval Benchmark) [21] and Market-1501 [45]. As for the first one, it contains a total of 52,712 person images with clean background and resolution of 256×256, which cover various poses and appearances. Following the same settings in [47], we divide this dataset into training and testing subsets with 101,966 and 8570 pairs, respectively. In contrast, Market-1501 [45] is more challenging due to its relatively low resolution (128×64), complex background, diverse light conditions, etc. Here, we follow the data split as [47], in which 263,632 pairs are selected for training and 12,000 ones for testing. These two subsets are not overlapped in terms of either identity and image. For building the body structure, we adopt OpenPose [3] to extract the 18 human joints. Since our SPATN model (the 1-*st* stage) is designed to take the source parsing map S_{p_s} to predict the target one S_{p_t}, it is necessary to obtain the parsing maps S_{p_s} and S_{p_t} for training. For the training data of DeepFashion [21], the off-the-shelf PGN model [6] is utilized to predict the human parsing maps by the given human images, which are regarded as the ground-truth for training the SPATN model. However, due to the adverse effect of low quality and complex background, PGN [6] fails to generate the plausible parsing maps with the given human images from Market-1501, making it unable to predict the target parsing maps, which inevitably result in the failure of the latter per-region translation. To tackle this problem, we train another SPATN* model on DeepFashion by using the source appearance image (rather than the source parsing map) and target pose to predict the target parsing map of Market-1501. This can eliminate the dependency of source parsing maps and achieve superior performance on predicting the target parsing maps for Market-1501.

Implementation Details. We adopt the Adam optimizer [16] with $\beta_1 = 0.5, \beta_2 = 0.999$ in all experiments. The learning rate lr is set to 2×10^{-4} for the two-stage model and 2×10^{-5} for the discriminator, respectively. The lr is decreased by 0.5 when the reconstruction loss on the validation set is no longer decreasing. The trade-off parameters are set as follows: $\lambda_{ce} = 10$, $\lambda_{L1} = 1$, $\lambda_{perc} = 1$, $\lambda_{adv} = 0.01$. The batch size is set to 4 and 32 for training on DeepFashion and Market-1501, respectively. The SPATN and SPGNet are trained simultaneously, in which the SPGNet takes the ground-truth target parsing map S_{p_t} as input. Only in the inference, SPGNet takes the predicted target parsing map $\hat{S}_{p_t}$ from SPATN model as input. More details and analyses about the training manner are given in the suppl. The whole model and experiments are carried out on a PC server with 2 1080Ti. It takes one week to train the two-stage model.

Method	SSIM ↑	FID ↓	PCKh ↑	LPIPS ↓
Def-GAN [31]	0.760	18.475	0.94	0.2330
PATN [47]	0.773	20.739	0.96	0.2533
Intr-Flow [18]	0.778	16.314	0.96	0.2131
GFLA [29]	**0.790**	**10.573**	0.96	0.2341
ADGAN [23]	0.772	14.460	0.96	0.2256
XingGAN [35]	0.778	39.322	0.95	0.2927
BiGraphGAN [34]	0.778	20.951	**0.97**	0.2444
Ours	0.782	12.243	**0.97**	**0.2105**

Table 1. Quantitative comparisons on two DeepFashion test sets.

Evaluation Metrics. Though it remains an open problem to quantitatively evaluate the quality of the appearance and shape consistency of the generated results, here we adopt four metrics, *i.e.*, SSIM, FID, LPIPS, and PCKh, which are common used in the previous works. Among these metrics, SSIM [39] is proposed to evaluate the quality of generated images by computing the global variance and means of the image to assess the structure similarity. Fréchet Inception Distance (FID) [10] is adopted to measure the realism of the generated images by calculating the Wasserstein-2 distance between the distributions of the generated results and its corresponding ground-truth. LPIPS [43] is another common metric to assess the visual quality, which is more consistent with human perception. PCKh [1] is recently suggested to measure the shape consistency, in which the score is computed by measuring the accuracy of the localization of the body joints. In addition, user study is conducted to evaluate the visual quality and the faithfulness (*i.e.*, realistic appearance and consistent shape). Except these quantitative metrics, we report the visual results to compare with these competing methods to show the superiority of our SPGNet.

4.2. Quantitative and Qualitative Comparison

We compare our proposed model with several recent state of the art methods, including Def-GAN [31], Intr-Flow [18], PATN [47], GFLA [29], XingGAN [35], and BiGraphGAN [34]. The quantitative result on DeepFashion is shown in Table 1. We can observe that our proposed method achieves comparable performance on this dataset, which means that results of ours own high visual quality and are realistic. In addition, we also conduct the evaluation on the challenging dataset Market-1501. The comparison is shown in the left part of Table 2. Mask-SSIM (M-SSIM) and Mask-LPIPS (M-LPIPS) are also considered to exclude the effect of irrelevant regions, *i.e.*, background. We can see that our method can also have a superior performance to these competing methods, indicating the great generalization of

Methods	Comparison on Market-1501						User Study					
	SSIM ↑	M-SSIM ↑	FID ↓	PCKh ↑	LPIPS ↓	M-LPIPS↓	R2G↑ (DF)	G2R↑ (DF)	R2G↑ (M)	G2R↑ (M)	Jab↑ (DF)	Jab↑ (M)
Def-GAN [31]	0.290	0.805	25.364	0.94	0.2994	0.1496	12.42	24.61	22.67	50.24	4.87	11.07
PATN [47]	0.311	0.811	22.657	0.94	0.3196	0.1590	19.14	31.78	32.23	63.47	8.27	6.53
Intr-Flow [18]	0.308	0.813	27.163	0.95	0.2888	0.1403	10.01	31.71	36.27	65.33	13.60	17.07
GFLA [29]	0.281	0.796	**19.751**	0.94	0.2817	0.1482	19.53	35.07	35.87	64.93	22.60	16.80
XingGAN [35]	0.313	0.816	22.495	0.93	0.3059	0.1581	21.61	33.75	35.28	65.16	4.73	7.20
BiGraphGAN [34]	**0.325**	**0.818**	28.915	0.94	0.3048	0.1505	**22.39**	34.16	35.76	65.91	13.93	16.20
Ours	0.315	**0.818**	23.331	**0.97**	**0.2779**	**0.1385**	19.47	**36.80**	**37.93**	**66.53**	**32.00**	**25.13**

Table 2. The quantitative comparison with the competing methods on Market-1501 test set and the two groups comparisons of user study. Here, DF (M) represents the evaluation on DeepFashion (Market-1501) test datasets. ↑ (↓) indicates higher (lower) is better.

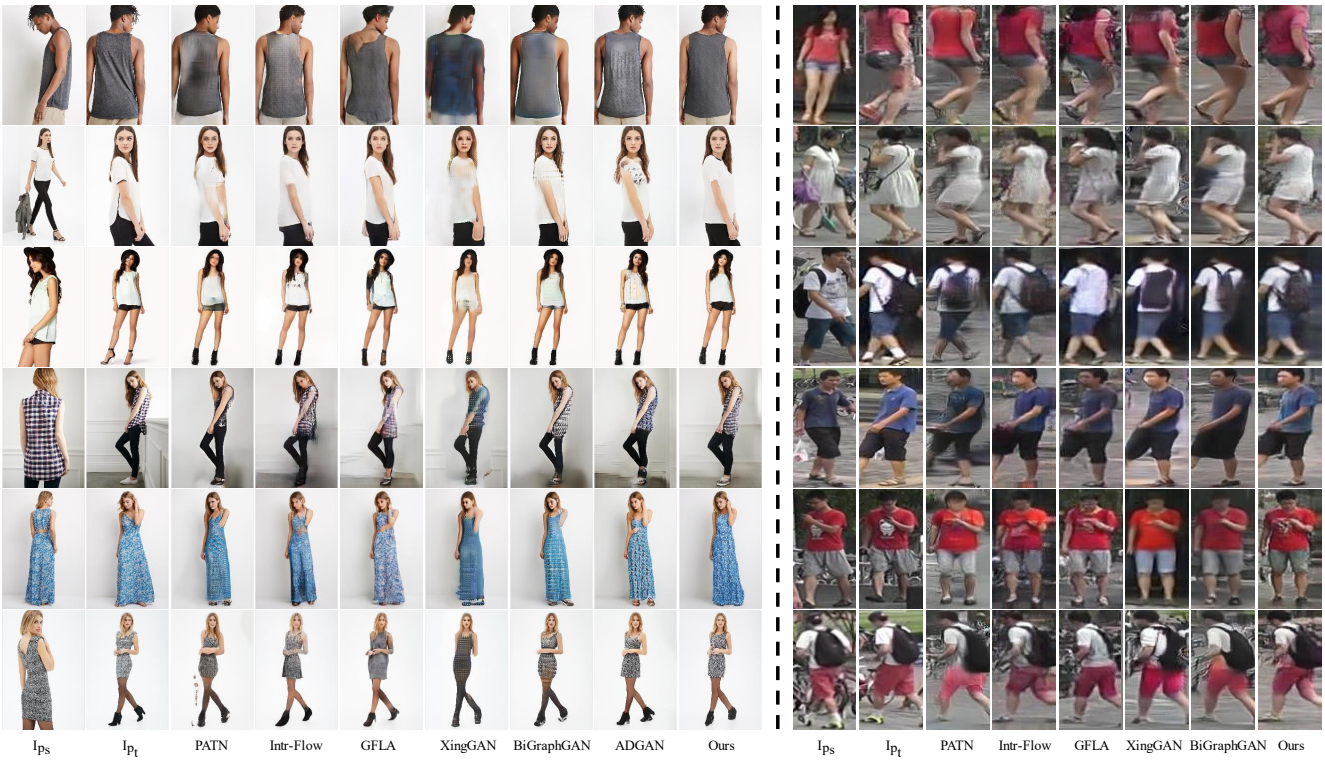


Figure 3. The visual comparison with the competing methods on two types of test sets. Best view it by zooming in the screen.

our method. The superior performance can be attributed to the proposed two-stage model in which the region adaptive normalization is utilized for per-region style translation.

Figure 3 gives the qualitative comparison on both DeepFashion and Market-1501 datasets. We can have the following observations, (i) though the source image I_{p_s} is under an extreme pose (1-*st* and 3-*rd* rows in the left part), results of our method are more realistic, and especially, have the consistent shape with the ground-truth I_{p_t}, indicating the benefits of the SPATN that bring to the generation process. (ii) In terms of complex texture (4∼6-*th* rows in the left part), ours are obviously superior to the competing methods in retaining the appearance textures, which is mainly attributed to the utilization of per-region style translation. (3) Within the complex background, the results of ours on the Market-1501 (right part of Figure 3) still perform favorably in generating the clear and photo-realistic results as well as a consistent appearance with the source image.

4.3. User Study

Although the above evaluation indicators can evaluate the performance of the generated results from different aspects, the person image generation task is likely to be user-oriented. Therefore, we conduct a user study on the two test sets to evaluate the performance from real human perception. To be specific, it mainly contains two groups that are conducted from different aspects. (i) Comparison with ground-truth. Following [22, 31], we randomly select 55 real and 55 generated images from the test set and then shuffle them. The first 10 of the 110 images are used for practice and the remaining 100 images are used to assess the performance. 30 volunteers that cover the bachelor and master students with computer vision background are required to give a choice about whether this one is real or generated within a second. (ii) Comparison with the state of the arts. To achieve this goal, we randomly select 55 image pairs, including source image, target pose, ground-truth and images generated by

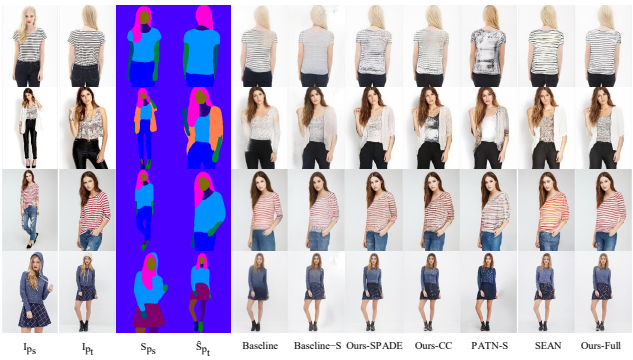


Figure 4. Visual comparison of different SPGNet variants.

these competing methods. Volunteers are required to select the image that is the closest to the ground-truth in terms of the visual quality, and more importantly, the pose and appearance consistency with ground-truth. Results are shown in the right part of Table 2. Here we adopt three metrics, *i.e.*, R2G: the percentage of the real image that is judged as the generated one, G2R: the percentage of the generated image that is judged as real one, Jab: the percentage that the image is judged as the best one. Higher values in these three metrics indicate better performance. We can see that results of ours outperform the state of the arts with a large margin (*e.g.*, 9.40% and 8.06% higher than the 2-*nd* best one in Jab), which further indicates the superiority of our method.

4.4. Ablation Study

In order to explore the effectiveness of our per-region style translation in the human generation task, in this paper, we follow [18] and conduct the ablation study on the DeepFashion datasets with the following variants, (i) Baseline: our baseline model is a dual-path U-Net, which mainly comes from Intr-Flow [18]. There are two encoders, which encode appearance and pose, respectively. The decoder combines the target pose and the warped appearance features to generate the final image. It should be noted that there is not any conditional normalization in this model. (ii) Baseline-S: based on the Baseline model, the pose encoder of Baseline-S takes the additional features (*i.e.*, the target semantic parsing map that is predicted by our SPATN model) as input. (iii) Ours-SPADE: by replacing the per-region adaptive instance normalization (SEAN) with the spatial adaptive normalization (SPADE) [26]. (iv) Ours-CC [20]: by replacing SEAN with CC-FPSE in the decoder module, in which ours-CC takes the target semantic parsing map as input and predicts the conditional convolution kernels to guide image generation. (v) PATN-S: by taking the semantic parsing maps as additional input data to retrain PATN. (vi) SEAN: by using the SEAN generator [46] to directly broadcast the style information to the target semantic layout to generate an image. (vii) Ours-Full. The quantitative results are shown in Table 3. We can see that (1) compared with Baseline and

Variants	SSIM ↑	FID ↓	PCKh ↑	LPIPS ↓
Baseline	**0.784**	18.946	0.96	0.2308
Baseline-S	0.780	16.460	0.96	0.2353
Ours-SPADE	0.779	13.808	0.96	0.2200
Ours-CC	0.778	13.686	0.96	0.2252
PATN-S	0.776	15.488	0.96	0.2557
SEAN	0.778	15.993	0.96	0.2209
Ours-Full	0.783	**12.613**	**0.97**	**0.2133**

Table 3. Quantitative comparison of different SPGNet variants.

Baseline-S, PATN and PATN-S, though slight improvements can be obtained, Ours-Full obviously outperforms them with a large margin, especially the FID metric (*e.g.*, Baseline-S is 13.1% higher than Baseline), indicating the benefits that the per-region adaptive normalization brings to this task. (2) Compared with the Baseline model, Ours-SPADE and Ours-CC achieve an obvious improvement by incorporating the spatial adaptive normalization and the conditional convolution, respectively, but they are still inferior to Ours-Full. (3) Directly taking SEAN to handle this task, the poor performance may be caused by the inconsistent poses, indicating the necessity of feature deformation in Ours-Full. (4) With the per-region adaptive normalization, Ours-Full performs superior to others, indicating the effectiveness of our SPGNet in human person generation. The visual comparison is shown in Fig. 4. We can see that compared with other variants, Ours-Full obviously outperforms others in generating photo-realistic and consistent appearance results, which further indicates the benefits of our SPGNet by adopting the per-region style translation in handling this task.

5. Conclusion

In this paper, we propose a two-stage model to deal with the challenging pose transfer task. In the first stage, we generate the target semantic parsing maps to eliminate the difficulties of pose transfer and benefit the latter per-region appearance translation. In the second one, with the predicted semantic map, we adopt the region adaptive normalization to achieve the per-region style translation, which is more effective in the person image generation task. The proposed method decomposes the complex task into two easier problems. Experiments show that both per-region style and semantic map are crucial in generating high-quality body parts and fine-scale textures. Moreover, instead of joint heatmaps, the distance map of the skeleton is adopted for a better generation of target semantic parsing map. Both the quantitative and visual comparison demonstrate the superior performance in generating consistent appearance and photo-realistic results with complex poses and source appearance.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China under grant No. U19A2073.

References

[1] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3686–3693, 2014. 6

[2] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8340–8348, 2018. 1, 4

[3] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7291–7299, 2017. 6

[4] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in Neural Information Processing Systems*, pages 474–484, 2018. 2

[5] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8857–8866, 2018. 1, 2

[6] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision*, pages 770–785, 2018. 6

[7] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 5

[8] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 4

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 6

[11] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 2

[12] Mohamed Ilyes Lakhal, Oswald Lanz, and Andrea Cavallaro. Pose guided human image synthesis by view disentanglement and enhanced weighting loss. In *Proceedings of the European Conference on Computer Vision*, pages 0–0, 2018. 1

[13] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 2, 3, 5

[14] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2287–2292, 2017. 1

[15] Justin Johnson, Alexandre Alahi, and Fei-Fei Li. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 694–711, 2016. 5

[16] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[17] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[18] Yining Li, Chen Huang, and Chen Change Loy. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3693–3702, 2019. 1, 2, 3, 4, 6, 7, 8

[19] Wen Liu, Zhixin Piao, Jie Min, Wenhan Luo, Lin Ma, and Shenghua Gao. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5904–5913, 2019. 1, 2

[20] Xihui Liu, Guojun Yin, Jing Shao, Xiaogang Wang, et al. Learning to predict layout-to-image conditional convolutions for semantic image synthesis. In *Advances in Neural Information Processing Systems*, pages 570–580, 2019. 3, 8

[21] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1096–1104, 2016. 2, 6

[22] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 406–416, 2017. 1, 2, 7

[23] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1, 2, 6

[24] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014. 1, 2

[25] Natalia Neverova, Riza Alp Guler, and Iasonas Kokkinos. Dense pose transfer. In *Proceedings of the European Conference on Computer Vision*, pages 123–138, 2018. 2

[26] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 3, 8

[27] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, June 2018. 1, 2

[28] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European Conference on Computer Vision*, pages 650–667, 2018. 1

[29] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020. 1, 2, 4, 6, 7

[30] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-assisted Intervention*, pages 234–241. Springer, 2015. 3

[31] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3408–3416, 2018. 1, 2, 3, 4, 6, 7

[32] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 5

[33] Sijie Song, Wei Zhang, Jiaying Liu, and Tao Mei. Unsupervised person image generation with semantic parsing transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2357–2366, 2019. 2

[34] Hao Tang, Song Bai, Philip HS Torr, and Nicu Sebe. Bipartite graph reasoning gans for person image generation. In *BMVC*, 2020. 6, 7

[35] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. *arXiv preprint arXiv:2007.09278*, 2020. 1, 2, 6, 7

[36] Hao Tang, Dan Xu, Gaowen Liu, Wei Wang, Nicu Sebe, and Yan Yan. Cycle in cycle generative adversarial networks for keypoint-guided image generation. In *ACM International Conference on Multimedia*, 2019. 1, 2

[37] Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *CVPR*, pages 7870–7879, 2020. 3

[38] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, 2018. 3

[39] 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. 6

[40] Ceyuan Yang, Zhe Wang, Xinge Zhu, Chen Huang, Jianping Shi, and Dahua Lin. Pose guided human video generation. In *Proceedings of the European Conference on Computer Vision*, September 2018. 1

[41] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020. 1

[42] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE International Conference on Computer Vision*, 2019. 1

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6

[44] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. Multi-view image generation from a single-view. In *ACM International Conference on Multimedia*, pages 383–391, 2018. 3

[45] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1116–1124, 2015. 2, 6

[46] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 3, 4, 8

[47] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2347–2356, 2019. 1, 2, 3, 6, 7