Pacific Graphics 2020
E. Eisemann, A. Jacobson, and F.-L Zhang
(Guest Editors)

*Volume 39 (2020), Number 7*

# Human Pose Transfer by Adaptive Hierarchical Deformation

Jinsong Zhang†, Xingzi Liu† and Kun Li ‡

Tianjin University, Tianjin 300350, China.


**Figure 1:** *Our method can generate person images in different target poses (left) and transfer upper clothing textures to a person image (right).*

**Abstract**

*Human pose transfer, as a misaligned image generation task, is very challenging. Existing methods cannot effectively utilize the input information, which often fail to preserve the style and shape of hair and clothes. In this paper, we propose an adaptive human pose transfer network with two hierarchical deformation levels. The first level generates human semantic parsing aligned with the target pose, and the second level generates the final textured person image in the target pose with the semantic guidance. To avoid the drawback of vanilla convolution that treats all the pixels as valid information, we use gated convolution in both two levels to dynamically select the important features and adaptively deform the image layer by layer. Our model has very few parameters and is fast to converge. Experimental results demonstrate that our model achieves better performance with more consistent hair, face and clothes with fewer parameters than state-of-the-art methods. Furthermore, our method can be applied to clothing texture transfer. The code is available for research purposes at https://github.com/Zhangjinso/PINet_PG.*

**CCS Concepts**

• ***Computing methodologies*** → ***Image processing;***

## 1. Introduction

Human pose transfer aims to synthesize a person image from a source pose to a target pose while preserving the appearance details, which has potential applications in video generation [WMGH17], person re-identification [QFX*18, ZYY*19] and image-based animation [LYL*17]. It is a very challenging and ill-posed problem due to misaligned transformation, occlusion, and high variance in poses.

Existing methods can be categorized into direct deformation and flow/transformation-based methods. Direct deformation method [MJS*17] concatenated the source image with its condition and target pose as the inputs of the generator to synthesize the target image by adversarial learning. However, without considering spatial correspondences, the results of this method are a little blurry. Zhu *et al*. [ZHS*19] proposed a progressive pose transfer network which utilized pose features to guide the image feature transfer, but the results often lost details (*e.g.* hair style) due to using vanilla con-

† Contribute equally to this work.
‡ Corresponding author: lik@tju.edu.cn

volution in the encoder and decoder. More methods estimated flow or transformation matrix to guide the image generation. Siarohin *et al*. [SSLS18] computed an affine transformation matrix based on the condition pose and the target pose to transform the image features. Dong *et al*. [DLG*18] designed a soft-gated warping-block to learn feature-level mapping with the guidance of segmentation map, but the results rely on the accuracy of estimated transformation matrix. Moreover, the matrix uses warped image features and is implemented only once, which is difficult to generate reasonable results for unknown regions. Han *et al*. [HHHS19] proposed a three-stage framework by adding a rendering network at the final stage to avoid the artifacts induced by wrong flows. However, these flow/transformation-based methods are difficult to deal with large transformation between source image and target image.

All the above methods use vanilla convolution that treats all pixels as valid information, which cannot select significant regions to deform and usually have a large number of parameters. Vanilla convolution benefits aligned generation task by extracting alignment information, but for unaligned generation task, *e.g*. , human pose transfer, the ability to select key areas to deform is more important. Moreover, human pose transfer, as a challenging image deformation problem, is difficult to successfully generate the target image with only one warping, especially for large deformation. It is more reasonable to adaptively select important information and gradually deform the image from coarse to fine.

In this paper, we propose a hierarchical deformation framework for learning-based human pose transfer. Unlike flow/transformation-based methods that generate the image in target pose with only one deformation, we use an encoder-decoder architecture with gated convolutions to dynamically select important features and adaptively deform the image layer by layer. To simplify the challenging misaligned problem, we design a hierarchical deformation framework to transfer the human pose from coarse to fine, which includes a parsing generator and an image generator. The parsing generator generates human semantic parsing aligned with the target pose, and the image generator generates the final textured person image in the target pose with the semantic guidance to retain the clothing style and texture. We conduct ablation study to verify our hypothesis. Comparative results with state-of-the-art methods demonstrate that our method achieves better human pose transfer results with fewer parameters. Our method can also be applied to image editing tasks, *e.g*. clothing texture transfer. Figure 1 shows some examples generated by our method.

Our main contributions are summarized as follows:

- We propose a human pose transfer network with two hierarchical deformation levels. The first level generates human semantic parsing aligned with the target pose, and the second level generates the final textured person image in the target pose by fine deformation. This provides a coarse-to-fine deformation framework and alleviates the difficulty of direct deformation from source to target.
- We introduce gated convolution to avoid the drawback of vanilla convolution that treats all the pixels as valid information. This copes well with the unaligned image generation task by learning a dynamic feature selection mechanism and adaptively deforming the image layer by layer.
- Our model has very few parameters and is fast to converge.
- Our model can be applied to image editing tasks, *e.g*. clothing texture transfer.

The rest of this paper is organized as follows. Section 2 presents a brief review of related work. Section 3 describes the proposed network. Experimental results are presented in Section 4, and the paper is concluded in Section 5.

## 2. Related Work

### 2.1. Person Image Generation

Lassner *et al.* [LPMG17] combined variational auto-encoder [KW13] and Generative Adversarial Network (GAN) to generate random person images with different appearance for full body. Zhu *et al.* [ZSW*18] proposed a novel pipeline for synthesizing human bodies from monocular image. Balakrishnan *et al.* [BZD*18] decomposed human image generation tasks into multiple foregrounds with different body parts and backgrounds. Si *et al.* [SWWT18] proposed a pose converter network, in which the foreground converter network and the background converter network use a multi-stage confrontation loss to generate more realistic images. Several previous research work [HWW*18, LCT18, WZL*18] focused on virtual try-on applications and made great progress in transferring clothes for a given character image, but the pose and shape of the character was unchanged. Unlike these methods which generate person images with unchanged pose, we propose a new person image generation method with various poses and viewpoints, which can also change the texture of clothing.

### 2.2. Human Pose Transfer

Human pose transfer can synthesize a new image with changed poses from a single person image. Several methods used 2D keypoints as pose representation to directly synthesize the target image. Ma *et al.* [MJS*17] proposed the first human pose transfer framework, which generated a coarse result and then refined it. However, this model is computationally inefficient and complicated to train. Ma *et al.* [MSG*18] improved their previous work by using a decomposition strategy. Esser *et al.* [ESO18] adopted variational autoencoder to sample appearance, and used U-Net [RFB15] to keep shape information, for interactive modeling. Li *et al.* [LHL19] introduced 3D flow graph with conditional poses, target poses and the visibility map to guide the transformation of image features and pixels. Zhu *et al.* [ZHS*19] proposed to deform image features in the latent space progressively. However, these method used 2D keypoints as pose representation, which is difficult to extract semantic correspondence directly between the source and target images. Some methods used human parsing maps as semantic guidance and estimated a transformation matrix or flow to synthesize the target image. Dong *et al.* [DLG*18] first generated human parsing map and predicted a transformation matrix with the help of human parsing maps aligned with the source pose and the target pose, and then deformed image features using the transformation matrix. Song *et al.* [SZLM19] proposed an unsupervised method to synthesize human parsing map and the target image by designing a cycle loss.

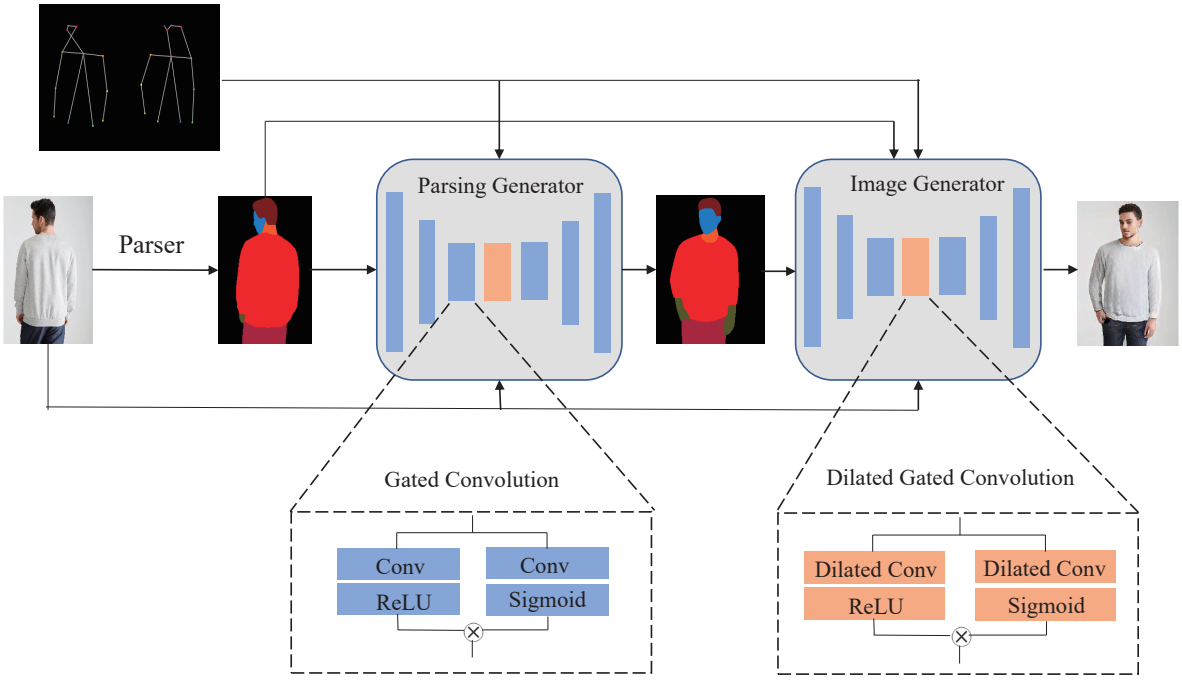


**Figure 2:** *Overview of our framework. Given an input person image together with the source and target poses, our method generates the person image in the target pose by hierarchical deformation with a parsing generator and an image generator. We use gated convolution to dynamically select important features and adaptively deform the image layer by layer.*

Han *et al.* [HHHS19] used synthesized parsing map to estimate a cloth flow mapping, warped image features to generate the clothing image without body, and concatenated the source image to synthesize the final result. Dong *et al.* [DLS*19] proposed a similar method with coarse-to-fine stage to enhance image details. Hsieh *et al.* [HCC*19] decomposed this task into three stages: pose-guided parsing translation, segmentation region coloring, and salient region refinement. Unlike these methods that use vanilla convolution, in this paper, we adopt gated convolution to learn a dynamic feature selection mechanism and adaptively deform the image layer by layer. Instead of estimating a transformation matrix or flow to deform features, we propose to extract and deform image features at the same time, which can preserve more important information and predict unknown regions.

### 2.3. Feature-wise Gating

Feature-wise gating is widely used in speech [ODZ*16], language [DFAG17], and vision [HSS18, VdOKE*16, WWZ*17]. WaveNets [ODZ*16] achieved promising results by applying a special feature gating $O_{x,y} = \tanh(w_1 x) \odot softmax(w_2 x)$ to model audio signals. Gated pixelCNN [VdOKE*16] also used this feature gating to model vision signals. Yu *et al.* [YLY*19] introduced this formulation as gated convolution into the image inpainting task, which significantly improves the inpainting quality of free-form masks and inputs. Inspired by [YLY*19], we adopt gated convolution to deal with the unaligned image generation task.

## 3. Our Approach

As shown in Figure 2, we propose a semantic-guided human pose transfer network with two hierarchical deformation levels. In the first level, parsing generator $G_p$ generates a human parsing map aligned with desired pose. In the second level, image generator $G_i$ synthesizes the final textured person image in the target pose with the semantic guidance. Details of our network architecture can be found in the supplementary video.

We apply Human Pose Estimator (HPE) [CSWS17] to extract 18 heatmaps as pose representation. An off-the-shelf human parser [GLL*18] is used to produce human segmentation maps with 20 labels. Because the predicted segmentation maps by human parser have some ambiguous parts (*e.g.*, left leg and right leg), we re-organize the map into 12 categories: background, hair, upper clothes, dress, pants, neck, skirt, face, hands, legs, shoes and hat.

In this section, we first explain the motivation to use gated convolution in both parsing generator and image generator, and then describe the details of our model. Finally, we show an application of texture transfer with our model.

### 3.1. Gated Convolution

Vanilla convolutions are widely used in convolution neural networks, which achieve great progress in object detection, image segmentation, and image-to-image translation. All the existing methods for human pose transfer used vanilla convolutions to comprise

their models. The vanilla convolution is formulated as

$$O_{y,x} = \sum_{i=-k_h}^{k_h} \sum_{i=-k_w}^{k_w} W_{k_h+i,k_w+j} \cdot I_{y+i,x+j}, \tag{1}$$

where $k_h = \frac{k_{sh}-1}{2}$ and $k_w = \frac{k_{sw}-1}{2}$ in which $k_{sh}$ and $k_{sw}$ are the kernel sizes (*e.g.*, $3 \times 3$). $W$ represents convolutional filters.

The formulation of vanilla convolution layers indicates that the output values in all spatial location are calculated with the same filter. It takes all pixels as valid values and extracts local features with a sliding window, which makes sense to object detection, image segmentation and aligned generation tasks. However, for misaligned tasks, *e.g.*, human pose transfer, the features extracted by vanilla convolution do not always have a positive impact on the output. Therefore, it is more important to learn a dynamic feature selection mechanism to deform the image. Inspired by [YLY*19], we introduce gated convolution into human pose transfer, which is formulated as

$$\begin{aligned} O_{x,y} = \phi\Big( \sum_{i=-k_h}^{k_h} \sum_{i=-k_w}^{k_w} u_{k_h+i,k_w+j} \cdot I_{y+i,x+j} \Big) \odot \\ \sigma\Big( \sum_{i=-k_h}^{k_h} \sum_{i=-k_w}^{k_w} v_{k_h+i,k_w+j} \cdot I_{y+i,x+j} \Big), \end{aligned} \tag{2}$$

where $\sigma$ denotes sigmoid function and the output gating values are between 0 and 1. $\phi$ denotes any activation function (*e.g.*, *Tanh* in WaveNet [ODZ*16]). $u$ and $v$ are two different convolutional filters. We use LeakyReLU [MHN13] as $\phi$ in our model.

For the parsing generator, gated convolution can obtain useful information at each spatial location, which is suitable for preserving the semantic parts of the person (*e.g.*, clothing style). For the image generator, gated convolution can preserve important features and deform key areas to generate textured person image. Therefore, we replace all vanilla convolutions with gated convolutions to adaptively extract and deform features.

### 3.2. Parsing Generator

For human pose transfer, the source image and the target image contain the same person with the same clothes and body shape, and hence we design a human parsing generator to build the semantic correspondence between them. Different from previous work, we use gated convolution to compose our parsing generator.

Because the misalignment between the input image and the target image, we design an encoder-decoder architecture for the parsing generator. Table 1 shows the details of our network architecture, where $c_{out}$ is the dimension of the output (12 for parsing generator and 3 for image generator). Given an input person image $I_s$ and a target pose $P_t$, the parsing generator learns to generate the human parsing map $M_g$ conditioned on image $I_s$ and pose $P_t$. Specifically, we first extract source pose $P_s$ and source parsing map $M_s$ from the input person image using a human pose estimator [CSWS17] and a human parser [GLL*18], respectively. Then, we concatenate them with source image $I_s$ as the input of our parsing generator. The processing of our parsing generator $G_p$ can be written as

$$M_g = G_p(I_s, P_s, P_t, M_s). \tag{3}$$

**Table 1:** *Details of our network architecture.*

| Operation | Kernel Size | Stride | Dilation | Output Shape |
|---|---|---|---|---|
| GatedConv | 7x7 | 1 | 1 | (256, 176, 64) |
| GatedConv | 3x3 | 2 | 1 | (128, 88, 64) |
| GatedConv | 3x3 | 1 | 1 | (128, 88, 128) |
| GatedConv | 3x3 | 2 | 1 | (64, 44, 128) |
| GatedConv | 3x3 | 1 | 1 | (64, 44, 256) |
| GatedConv | 3x3 | 1 | 1 | (64, 44, 256) |
| GatedConv | 3x3 | 1 | 1 | (64, 44, 256) |
| DilatedGatedConv | 3x3 | 1 | 2 | (64, 44, 256) |
| DilatedGatedConv | 3x3 | 1 | 4 | (64, 44, 256) |
| Self-attention Module | - | - | - | (64, 44, 256) |
| GatedConv | 3x3 | 1 | 1 | (64, 44, 256) |
| GatedConv | 3x3 | 1 | 1 | (64, 44, 256) |
| Upsample | - | - | 1 | (128, 88, 256) |
| GatedConv | 3x3 | 1 | 1 | (128, 88, 128) |
| GatedConv | 3x3 | 1 | 1 | (128, 88, 128) |
| Upsample | - | - | - | (256, 176, 128) |
| GatedConv | 3x3 | 1 | 1 | (256, 176, 64) |
| GatedConv | 3x3 | 1 | 1 | (256, 176, 32) |
| GatedConv | 7x7 | 1 | 1 | (256, 176, $c_{out}$) |

The reconstruction loss of $G_p$ is defined as $\ell_1$ distance loss between target parsing map $M_t$ and generated parsing map $M_g$:

$$\mathcal{L}_{\ell_1} = ||M_g - M_t||_1. \tag{4}$$

We also apply the categorical cross-entropy loss to encourage the generator to synthesize high-quality parsing maps:

$$\mathcal{L}_{\log} = \mathcal{L}_{\log}(M_t, M_g) = -\frac{1}{N} \sum_{i=0}^{N-1} M_{t_i} \log(S(M_{g_i}), \tag{5}$$

where $N$ is the number of categories of labels ($N = 12$ in our case), and $S$ denotes softmax function. The final loss function of our parsing generator can be formulated as:

$$\mathcal{L}_{parsing} = \mathcal{L}_{\log} + \mathcal{L}_{\ell_1}. \tag{6}$$

### 3.3. Image Generator

With the source human parsing map $M_s$ and the generated human parsing map $M_g$, previous work [DLG*18, HHHS19] used human parsing maps to estimate a flow mapping and warped image features to generate the results using this flow mapping. However, their models may lose clothing style due to imprecise flow mapping and generate artifacts in some unknown regions. Instead of estimating a flow mapping using human parsing maps, we use gated convolution to extract image feature and semantic correspondence, and deform the image feature. The image generator aims to deform the source image $I_s$ based on the parsing maps $M_s$ and $M_g$ by extracting the semantic correspondence. We also feed the target pose $P_t$ into our image generator to encourage the synthesized image to be aligned with the target pose when the generated parsing map is not precise.

We adopt the same encoder-decoder architecture for the image generator. Based on a pair of parsing maps and the source person image, the encoder is responsible for extracting semantic correspondences and encoding the image, while the decoder is used to

**Figure 3:** *Results of person image synthesis in different poses using our method.*

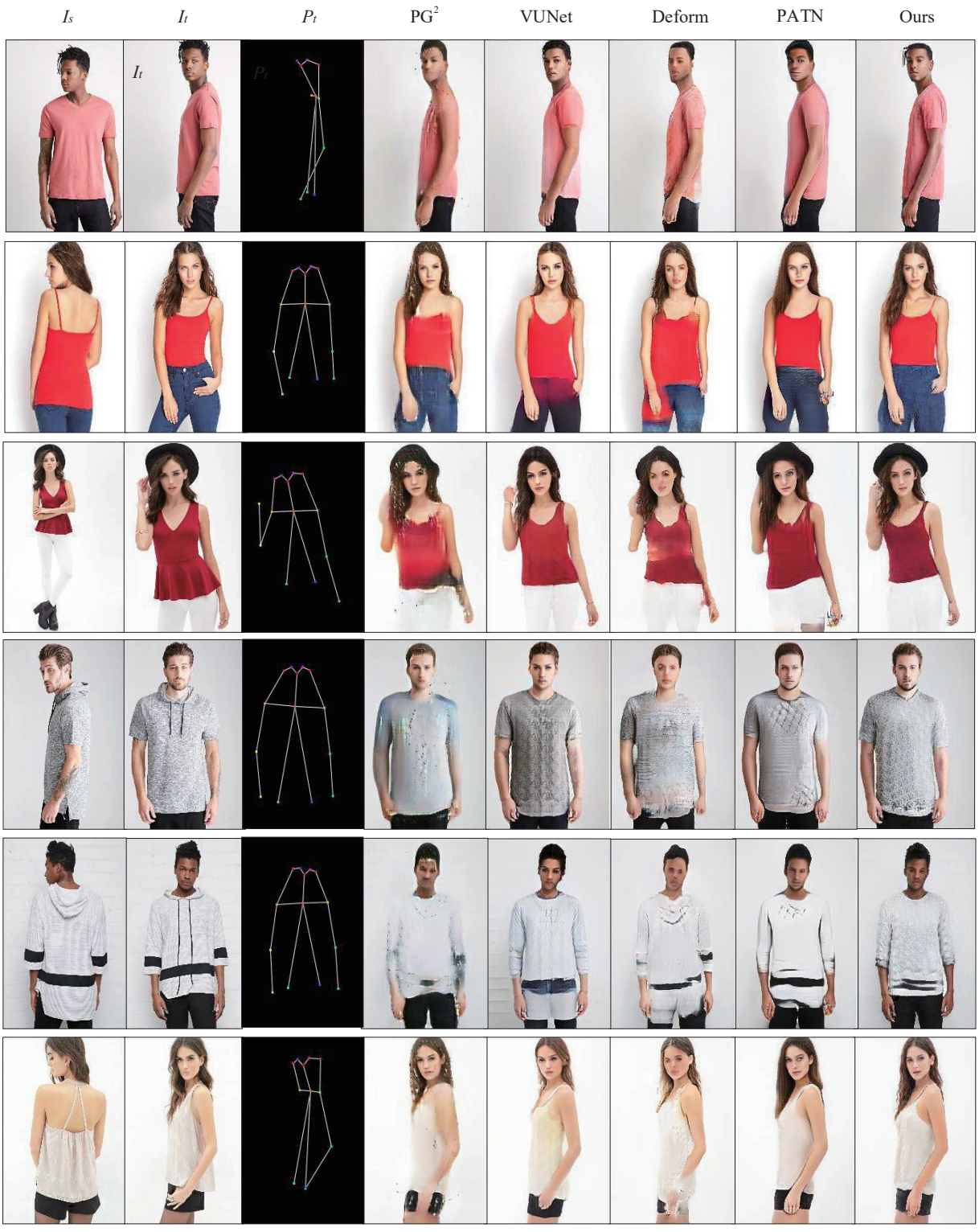


**Figure 4:** *Qualitative results on DeepFashion dataset compared with PG$^2$ [MJS*17], VUnet [ESO18], Deform [SSLS18] and PATN [ZHS*19].*

refine and decode the final feature to deliver the result. The dilated gated convolution [YK16] is applied to expand the receptive filed.

The generative adversarial framework [GPAM*14] is used to generate more realistic images by mimicking the distributions of the ground truth $I_t$. In traditional GAN, the discriminator is used to judge whether the generated image is real or fake to encourage the generator to synthesize realistic images. For conditional image generation task, it is also important to make the generated images meet the requirement of condition image (e.g., the pose of the generated image should be aligned with the target pose in human pose transfer), in addition to generating realistic images. Therefore, we use two discriminators: pose discriminator $D_P$ and appearance discriminator $D_A$, to encourage the generator to synthesize images aligned with the target pose and preserve the texture consistent with the input image. The conditional adversarial loss is defined as:

$$\begin{aligned}\mathcal{L}_{CGAN} = E\{\log[D_A(I_s,I_t)\cdot D_P(I_t,P_t)]\}+\\ E\{\log[(1-D_A(I_s,I_g))\cdot(1-D_P(I_g,P_t))]\}.\end{aligned} \tag{7}$$

We use $\ell_1$ distance loss between the generated image $I_g$ and the ground truth $I_t$, which is defined as:

$$\mathcal{L}_{\ell_1} = ||I_g - I_t||_1. \tag{8}$$

Perceptual loss [JAFF16] has achieved great success in image synthesis [IZZE17, ZPIE17, LHM*19]. We apply a perceptual loss $\mathcal{L}_{percep}$ to compute the distances of high-level features in the pre-trained model between the generated image $I_g$ and the ground truth $I_t$. We formulate the perceptual loss as:

$$\mathcal{L}_{percep} = \sum_{i=1}^{N} \alpha_i ||\phi_i(I_g) - \phi_i(I_t)||_1, \tag{9}$$

where $\phi_i(I_g)$ denotes the feature map of the $i$-th ($i = 0,1,2,3,4$) layer in the pre-trained network $\phi$ for the generated image $I_g$. We use VGG-19 model [SZ14] pre-trained on ImageNet [DDS*09] as $\phi$ and extract feature maps from layers *relu*{1_1, 2_1, 3_1, 4_1, 5_1}.

The final loss function is defined as :

$$\mathcal{L}_{Image} = \lambda_1\mathcal{L}_{\ell_1} + \lambda_2\mathcal{L}_{percep} + \lambda_3\mathcal{L}_{CGAN}, \tag{10}$$

where $\lambda_1, \lambda_2, \lambda_3$ represent the weights of $\mathcal{L}_{\ell_1}$, $\mathcal{L}_{percep}$, $\mathcal{L}_{CGAN}$ that contribute to $\mathcal{L}_{Image}$, respectively.

### 3.4. Implementation Details

For two generators, we first train the parsing generator and the image generator for about 60K iterations, respectively. The input of parsing generator is a triplet ($I_s$ ,$P_s$, $P_t$, $M_s$), and the output is a generated parsing map aligned with target image. For the image generator, we alternatively train the generator and the discriminators. The image generator takes a triplet ($I_s$, $M_s$, $M_t$, $P_t$) as input and delivers the generated image $I_g$. To train the discriminators, the appearance discriminator $D_A$ takes ($I_s$, $I_t$) and ($I_s$, $I_g$) as inputs, and the pose discriminator $D_p$ takes ($P_t$, $I_t$) and ($P_t$, $I_g$) as inputs. Then, we train two generators jointly for around 90K iterations. The initial learning rate is linearly decayed to 0 after 50K iterations.

The coefficients in the loss function of image generator ($\lambda_1$, $\lambda_2$, $\lambda_3$) are set to (1, 0.5, 5). Spectral Normalization [Yos18] is adopted after every convolution layer in two discriminators to improve the stability of training process. Adam optimizer [KB14] with $\beta_1$=0.5 and $\beta_2$=0.999 is applied to train our model.

### 3.5. Texture Transfer

Because we use human parsing map as an intermediate result to represent semantic correspondences, we can achieve texture transfer by replacing the source body parts with new clothing texture, utilizing human parsing maps. For example, we can replace the texture of upper clothes in the source image $I_s$ with that in the condition image $I_c$. To achieve this, we first take the condition image $I_c$, parsing maps $M_c$ and $M_s$ extracted from $I_c$ and $I_s$ and the source pose $P_s$ as inputs to our image generator to synthesize a new image with the original body shape of the source image and the texture of the condition image. Then, we crop the generated image by the body part mask $M_{s_i}$ ($i \in [1,12]$) from human parsing map $M_s$. Finally, we deliver the result by replacing the region of $M_{s_i}$ in the source image $I_s$ with the generated image we cropped. Formally, the texture transfer process is formulated as

$$I_f = I_s \odot (1 - M_{s_i}) + G_I(I_c, M_c, M_s, P_s) \odot M_{s_i}, \tag{11}$$

where $G_I$ is our image generator. More details can be found in the supplementary video.

## 4. Experimental Results

In this section, we first give some human pose transfer results of our method, and quantitatively and qualitatively compare our method with several state-of-the-art methods. Then, we perform ablation study to verify the effect of different components in our model. Finally, we present an application in texture transfer. More results can be found in the supplementary video.

### 4.1. Results

To verify the effectiveness of our method, we generate person images with the same source person image and several different target poses on the DeepFashion dataset [LLQ*16] that has large variation in pose and appearance. This dataset contains 52712 images with the resolution of 256 × 256. Following PATN [ZHS*19], we adopt the same data configuration, collecting 101966 training pairs and 8570 testing pairs. Note that the person identities in the test set are different from those in the training set. The source and target poses are estimated by a human pose estimator [CSWS17]. Figure 3 shows some results using our method. Our method can generate sharp details by preserving the textures (*e.g.*, hair) and clothing patterns (*e.g.*, dress and hat). Moreover, our results retain facial details. This appealing property attributes to our hierarchical deformation framework with gated convolutions.

### 4.2. Comparison

**Quantitative comparison.** To evaluate the performance of human pose transfer, we use three metrics as our evaluation metrics. Inception Score (IS) is commonly used to measure the quality of image generation [MJS*17]. To coincide with human judgment, Learned

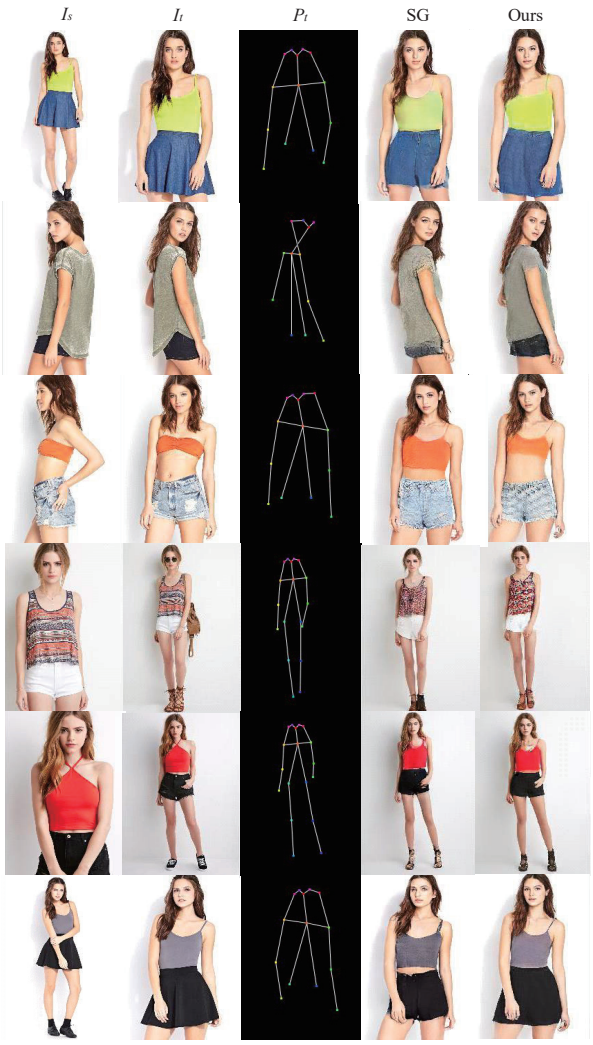


**Figure 5:** *Qualitative results on DeepFashion dataset compared with SG [DLG*18]. Please zoom in for details.*


**Figure 6:** *Qualitative results on DeepFashion dataset compared with CF [HHHS19]. Please zoom in for details.*

Perceptual Image Patch Similarity (LPIPS) [ZIE*18] is used to calculate the reconstruction error between the generated image and the ground truth. Fréchet Inception Distance (FID) [HRU*17] is used to measure the realism of the generated images. Table 2 shows the quantitative results compared with four state-of-the-art methods: $PG^2$ [MJS*17], VUnet [ESO18], Deform [SSLS18] and PATN [ZHS*19]. We run the pre-trained models of these methods and use the same data division as PATN [ZHS*19]. For $PG^2$ [MJS*17], VUnet [ESO18] and Deform [SSLS18], our test images may appear in their training set due to lack of their data division details. As shown in Table 2, our method achieves the best performance in LPIPS and FID, which are more consistent with human judgement. Besides, our model has the fewest parameters.

**Table 2:** *Quantitative comparison with four state-of-the-art methods on DeepFashion dataset.*

| Model | IS ↑ | LPIPS ↓ | FID ↓ | Parameters |
|---|---|---|---|---|
| $PG^2$ [MJS*17] | 3.163 | 0.2901 | 45.288 | 437.09M |
| VUnet [SSLS18] | 3.362 | 0.2637 | 23.667 | 139.36M |
| Deform [ESO18] | **3.440** | 0.2330 | 18.457 | 82.08M |
| PATN [ZHS*19] | 3.209 | 0.2533 | 20.739 | 41.36M |
| Ours | 3.419 | **0.2159** | **12.635** | 19.46M |

**Qualitative comparison.** Figure 4 shows qualitative results compared with $PG^2$ [MJS*17], VUnet [ESO18], Deform [SSLS18] and PATN [ZHS*19]. Our model can generate more sharp images with rich details. For example, the hair in our generated images is more realistic and the style has been preserved. Furthermore, our model keeps shape consistence, *e.g.* , the hat in the third row. The images synthesized by our model also have more consistent texture with source person images, *e.g.* , the color of skin and clothes, which


**Figure 7:** *Qualitative results of ablation study. The fourth and fifth columns indicate one-stage model and two-stage model with vanilla convolution respectively, and the sixth column indicates one-stage model with gated convolution. Last two columns denote the results of our full model with and without adversarial loss, respectively.*

means that our image generator can transfer textures based on human parsing maps.

We also compare our method with two semantic-guided methods, SG [DLG*18] and CF [HHHS19], in Figure 5 and Figure 6, respectively. Due to lack of their codes, we use the results presented in their papers. As shown in Figure 5, our model can preserve more clothing details and clothing style, *e.g.* , dress in the first/last row and upper clothes in the third row. Besides, our model can generate more reasonable unknown regions with less artifacts, *e.g.* , shoes in the fourth and fifth rows. As shown in Figure 6, our model generates more consistent semantic shape and texture with the source image, *e.g.* , hair in the first row and the third row.

### 4.3. Ablation study

We conduct ablation study to verify the effectiveness of the important parts. We train four ablation models compared with our model:

**One w/ VC.** The one-stage model with vanilla convolution is similar with the first stage in $PG^2$ [MJS*17]. However, we use two discriminators and losses as shown in Section 3.

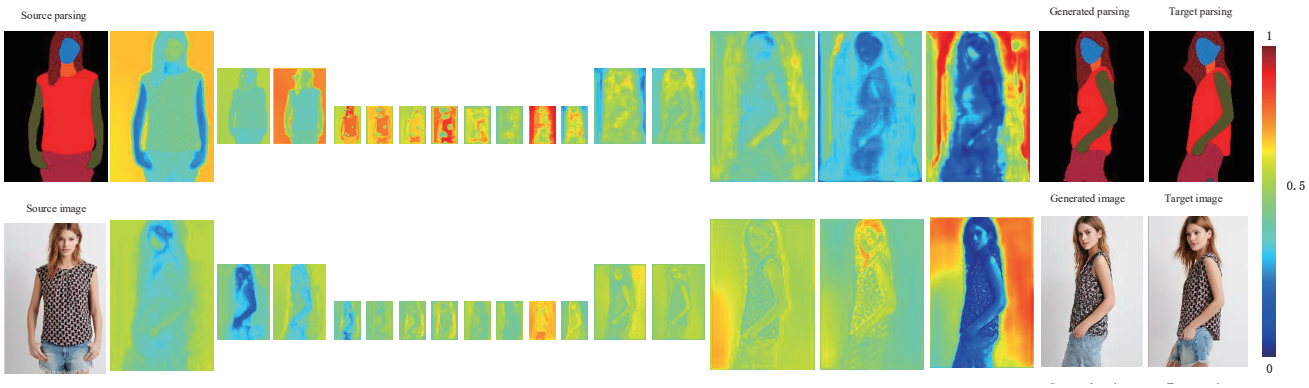


**Figure 8:** *Visualization of our attention mask in each gated convolution layer of the parsing generator and the image generator.*

**Two w/ VC.** The two-stage model with vanilla convolution is also a hierarchical deformation framework with a parsing generator and an image generator, while using vanilla convolution instead of gated convolution. Other settings are the same as our model.

**One w/ GC.** The one-stage model directly synthesizes the results from the input image together with the source pose and the target pose without semantic guidance. We employ gated convolution in the generator and the same discriminators as illustrated in Section 3.

**Full w/o adv.** We retrain our model without adversarial loss to investigate the effect of adversarial loss.

Table 3 and Figure 7 give the quantitative and qualitative results compared with one stage model and vanilla convolution model. As shown in Figure 7, compared with one-stage model based on gated convolution, our image generator generates more realistic images with details. This demonstrates that our image generator can extract richer semantic information and deform image features more reasonable with human parsing map as an intermediate result. The vanilla convolution model also synthesizes reasonable results, but the texture and identity of the results are less consistent with the source images. On the contrast, our model with gated convolution can learn a dynamic selection mechanism and adaptively select important regions to deform. Moreover, compared with $PG^2$, one-stage model with vanilla convolution gets better results, which illustrates the advantages of our two discriminators and reasonable losses. Besides, there is no doubt that adversarial loss encourages the model to synthesize realistic images. The quantitative results given in Table 3 also verify the rationality of our model.

**Table 3:** *Quantitative comparison of ablation study.*

| Model | IS ↑ | LPIPS ↓ | FID ↓ |
|---|---|---|---|
| One w/ VC | 3.129 | 0.271 | 25.908 |
| Two w/ VC | 3.248 | 0.232 | 15.229 |
| One w/ GC | 3.172 | 0.228 | 17.254 |
| Full w/o adv | 2.404 | 0.415 | 81.980 |
| Full w/ adv | **3.419** | **0.216** | **12.635** |

### 4.4. Results analysis

To give an intuitive demonstration on how the gated convolution works in parsing generator and image generator, we visualize the attention masks in all the gated convolution layers in Figure 8. The first column shows the input we need to deform, and the last two columns show the output and the target ground truth. The other columns show the attention masks in each gated convolution layer of the parsing generator and the image generator. The parsing generator aims to deform the source parsing, and the image generator is used to deform the source image with the guidance of parsing maps.

As shown in Figure 8, feature selection, *i.e.* , attention mask, varies from coarse to fine in both image and weight spaces among different layers. Specifically, feature selection focuses on background and large regions in the first layer while pays more attention to foreground and small regions in the small-scale layer. Moreover, the weight variation between different semantic regions becomes smaller as the scale becomes smaller, but weight values vary in the opposite direction. This demonstrates that our multi-layer gated convolutions are able to dynamically select and deform the features from coarse to fine. Besides, the features aligned with the source parsing have large weights in the large-scale layer but have small weights in the small-scale layer. At the same time, the weights of features aligned with the target pose become lager as the scale becomes smaller. This demonstrates that feature selection and feature deformation are achieved together with different importance in different layers.

### 4.5. Texture Transfer

As described in Section 3.5, our model can also achieve texture transfer. Figure 9 shows some visual results. The upper clothes and pants of the input person image can be automatically edited by using the texture of those in the condition images. Our model can edit the images by generating realistic components.

### 4.6. Failure Cases

Our model can generate images that preserve the semantic shape and texture from the source image. However, as shown in Figure 10,

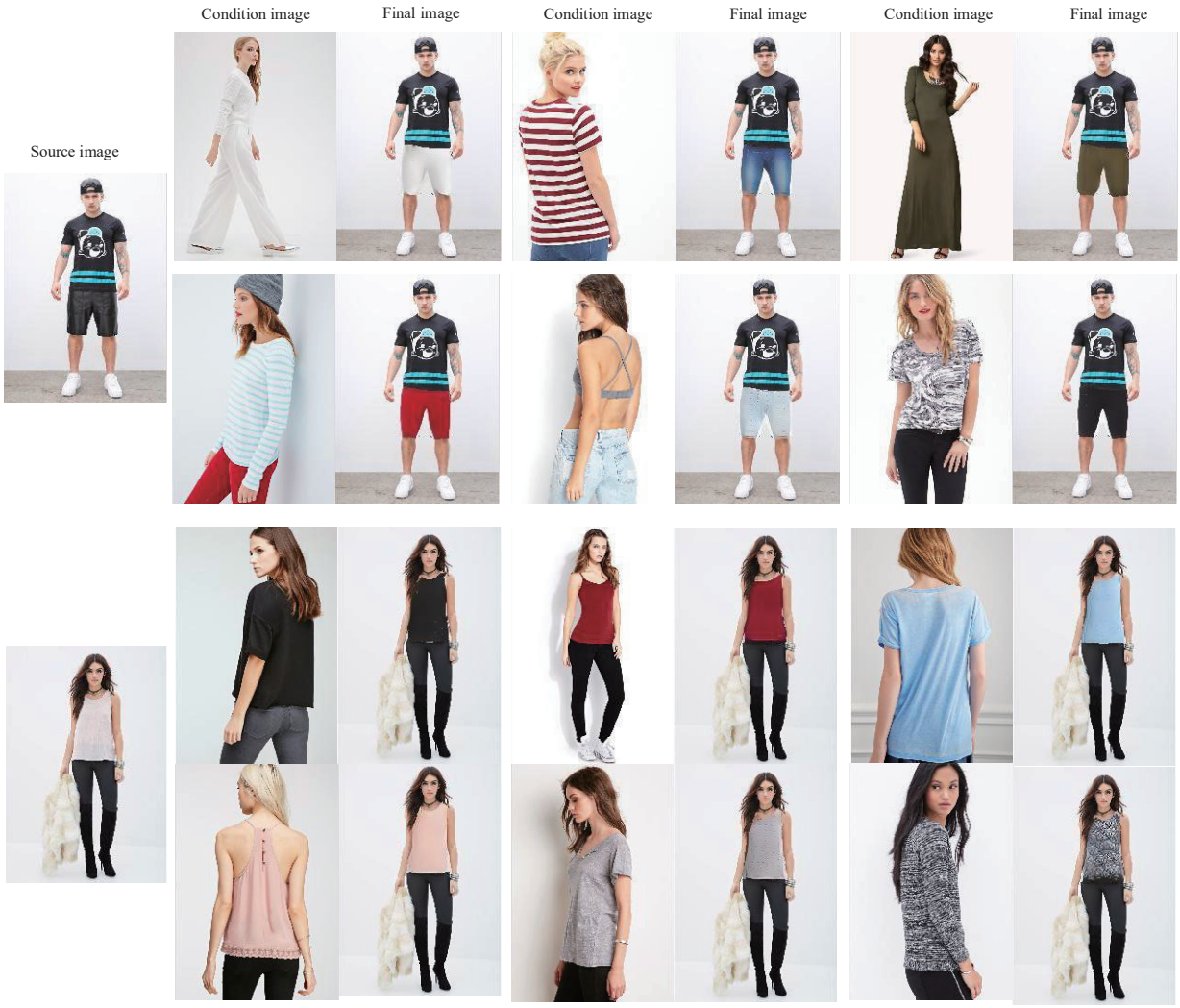


**Figure 9:** *Results of texture transfer using our method. From top to down, the first two rows show the results that transfer the texture of pants in condition images to the source images. The last two rows show the results that transfer the texture of upper clothes in condition images to the source images.*

there are some limitations in our work. For the source images with complex texture (in the first and second rows), it is insufficient to use semantic correspondence extracted from human parsing maps as guidance. Besides, our training set does not have enough images including the specific carton patterns (in the first row), which is also the important reason. Moreover, for some complex poses, the synthesized images may be blurry.

## 5. Conclusion

In this paper, instead of directly synthesizing the target image, we propose an adaptive hierarchical deformation framework for human pose transfer. The first deformation level generates human semantic parsing aligned with the target pose, and the second deformation level generates the final textured person image in the target pose with the semantic guidance. Furthermore, we notice that the vanilla convolution is not suitable for unaligned generation task, and hence we use gated convolution to dynamically select important features and adaptively deform the image layer by layer. Experimental results demonstrate that our method achieves better pose transfer results with fewer parameters. Besides, our model can transfer clothing texture based on component attribute.

## Acknowledgements

This work was supported in part by Tianjin Research Program of Application Foundation and Advanced Technology (18JCYBJC19200).

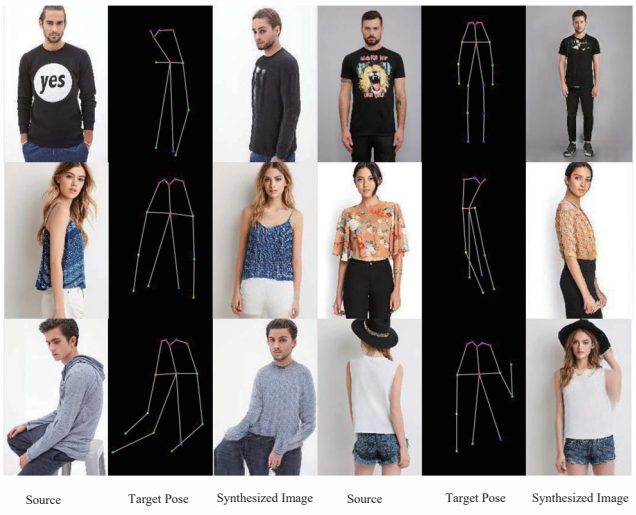


**Figure 10:** *Failure cases of our method.*

## References

[BZD*18] BALAKRISHNAN G., ZHAO A., DALCA A. V., DURAND F., GUTTAG J.: Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8340–8348. 2

[CSWS17] CAO Z., SIMON T., WEI S.-E., SHEIKH Y.: Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 7291–7299. 3, 4, 7

[DDS*09] DENG J., DONG W., SOCHER R., LI L.-J., LI K., FEI-FEI L.: Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2009), IEEE, pp. 248–255. 7

[DFAG17] DAUPHIN Y. N., FAN A., AULI M., GRANGIER D.: Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70* (2017), JMLR. org, pp. 933–941. 3

[DLG*18] DONG H., LIANG X., GONG K., LAI H., ZHU J., YIN J.: Soft-gated warping-gan for pose-guided person image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)* (2018), pp. 472–482. 2, 4, 8, 9

[DLS*19] DONG H., LIANG X., SHEN X., WANG B., LAI H., ZHU J., HU Z., YIN J.: Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 9026–9035. 3

[ESO18] ESSER P., SUTTER E., OMMER B.: A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 8857–8866. 2, 6, 8

[GLL*18] GONG K., LIANG X., LI Y., CHEN Y., YANG M., LIN L.: Instance-level human parsing via part grouping network. In *Proceedings of the European Conference on Computer Vision* (2018), pp. 770–785. 3, 4

[GPAM*14] GOODFELLOW I., POUGET-ABADIE J., MIRZA M., XU B., WARDE-FARLEY D., OZAIR S., COURVILLE A., BENGIO Y.: Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)* (2014), pp. 2672–2680. 7

[HCC*19] HSIEH C.-W., CHEN C.-Y., CHOU C.-L., SHUAI H.-H., LIU J., CHENG W.-H.: Fashionon: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proceedings of the 27th ACM International Conference on Multimedia* (2019), pp. 275–283. 3

[HHHS19] HAN X., HUANG W., HU X., SCOTT M.: Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE International Conference on Computer Vision* (2019). 2, 3, 4, 8, 9

[HRU*17] HEUSEL M., RAMSAUER H., UNTERTHINER T., NESSLER B., HOCHREITER S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)* (2017), pp. 6626–6637. 8

[HSS18] HU J., SHEN L., SUN G.: Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7132–7141. 3

[HWW*18] HAN X., WU Z., WU Z., YU R., DAVIS L. S.: Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 7543–7552. 2

[IZZE17] ISOLA P., ZHU J.-Y., ZHOU T., EFROS A. A.: Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 1125–1134. 7

[JAFF16] JOHNSON J., ALAHI A., FEI-FEI L.: Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision* (2016), Springer, pp. 694–711. 7

[KB14] KINGMA D. P., BA J.: Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014). 7

[KW13] KINGMA D. P., WELLING M.: Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013). 2

[LCT18] LAHNER Z., CREMERS D., TUNG T.: Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision* (2018), pp. 667–684. 2

[LHL19] LI Y., HUANG C., LOY C. C.: Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 3693–3702. 2

[LHM*19] LIU M.-Y., HUANG X., MALLYA A., KARRAS T., AILA T., LEHTINEN J., KAUTZ J.: Few-shot unsupervised image-to-image translation. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 10551–10560. 7

[LLQ*16] LIU Z., LUO P., QIU S., WANG X., TANG X.: Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 1096–1104. 7

[LPMG17] LASSNER C., PONS-MOLL G., GEHLER P. V.: A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 853–862. 2

[LYL*17] LI K., YANG J., LIU L., BOULIC R., LAI Y. K., LIU Y., LI Y., MOLLA E.: Spa: Sparse photorealistic animation using a single rgb-d camera. *Transactions on Circuits Systems for Video Technology 27*, 4 (2017), 771–783. 1

[MHN13] MAAS A. L., HANNUN A. Y., NG A. Y.: Rectifier nonlinearities improve neural network acoustic models. In *Proceedings of International Conference on Machine Learning* (2013), vol. 30, p. 3. 4

[MJS*17] MA L., JIA X., SUN Q., SCHIELE B., TUYTELAARS T., VAN GOOL L.: Pose guided person image generation. In *Advances in Neural Information Processing Systems (NeurIPS)* (2017), pp. 406–416. 1, 2, 6, 7, 8, 9

[MSG*18] MA L., SUN Q., GEORGOULIS S., VAN GOOL L., SCHIELE B., FRITZ M.: Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 99–108. 2

[ODZ*16] OORD A. V. D., DIELEMAN S., ZEN H., SIMONYAN K., VINYALS O., GRAVES A., KALCHBRENNER N., SENIOR A., KAVUKCUOGLU K.: Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499* (2016). 3, 4

[QFX*18] QIAN X., FU Y., XIANG T., WANG W., QIU J., WU Y., JIANG Y.-G., XUE X.: Pose-normalized image generation for person re-identification. In *Proceedings of the European Conference on Computer Vision* (2018), pp. 650–667. 1

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2015), Springer, pp. 234–241. 2

[SSLS18] SIAROHIN A., SANGINETO E., LATHUILIÃĹRE S., SEBE N.: Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (June 2018). 2, 6, 8

[SWWT18] SI C., WANG W., WANG L., TAN T.: Multistage adversarial losses for pose-based human image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 118–126. 2

[SZ14] SIMONYAN K., ZISSERMAN A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014). 7

[SZLM19] SONG S., ZHANG W., LIU J., MEI T.: Unsupervised person image generation with semantic parsing transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 2357–2366. 2

[VdOKE*16] VAN DEN OORD A., KALCHBRENNER N., ESPEHOLT L., VINYALS O., GRAVES A., ET AL.: Conditional image generation with pixelcnn decoders. In *Advances in Neural Information Processing Systems (NeurIPS)* (2016), pp. 4790–4798. 3

[WMGH17] WALKER J., MARINO K., GUPTA A., HEBERT M.: The pose knows: Video forecasting by generating pose futures. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 3332–3341. 1

[WWZ*17] WANG H., WANG Y., ZHANG Q., XIANG S., PAN C.: Gated convolutional neural network for semantic segmentation in high-resolution images. *Remote Sensing 9*, 5 (2017), 446. 3

[WZL*18] WANG B., ZHENG H., LIANG X., CHEN Y., LIN L., YANG M.: Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision* (2018), pp. 589–604. 2

[YK16] YU F., KOLTUN V.: Multi-scale context aggregation by dilated convolutions. In *International Conference on Learning Representations* (2016). 7

[YLY*19] YU J., LIN Z., YANG J., SHEN X., LU X., HUANG T. S.: Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 4471–4480. 3, 4

[Yos18] YOSHIDA Y.: Spectral normalization for generative adversarial networks. In *International Conference on Machine Learning Workshop on Implicit Models* (2018). 7

[ZHS*19] ZHU Z., HUANG T., SHI B., YU M., WANG B., BAI X.: Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 2347–2356. 1, 2, 6, 7, 8

[ZIE*18] ZHANG R., ISOLA P., EFROS A. A., SHECHTMAN E., WANG O.: The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (Jun 2018). 8

[ZPIE17] ZHU J.-Y., PARK T., ISOLA P., EFROS A. A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 2223–2232. 7

[ZSW*18] ZHU H., SU H., WANG P., CAO X., YANG R.: View extrapolation of human body from a single image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4450–4459. 2

[ZYY*19] ZHENG Z., YANG X., YU Z., ZHENG L., YANG Y., KAUTZ J.: Joint discriminative and generative learning for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 2138–2147. 1