Structure-transformed Texture-enhanced Network for Person Image Synthesis

Munan Xu[1], Yuanqi Chen[1], Shan Liu[2], Thomas H. Li[3], Ge Li[1]
[1]School of Electronic and Computer Engineering, Peking University Shenzhen Graduate School
[2]Tencent America [3]Advanced Institute of Information Technology, Peking University

Abstract

Pose-guided virtual try-on task aims to modify the fashion item based on pose transfer task. These two tasks that belong to person image synthesis have strong correlations and similarities. However, existing methods treat them as two individual tasks and do not explore correlations between them. Moreover, these two tasks are challenging due to large misalignment and occlusions, thus most of these methods are prone to generate unclear human body structure and blurry fine-grained textures. In this paper, we devise a structure-transformed texture-enhanced network to generate high-quality person images and construct the relationships between two tasks. It consists of two modules: structure-transformed renderer and texture-enhanced stylizer. The structure-transformed renderer is introduced to transform the source person structure to the target one, while the texture-enhanced stylizer is served to enhance detailed textures and controllably inject the fashion style founded on the structural transformation. With the two modules, our model can generate photorealistic person images in diverse poses and even with various fashion styles. Extensive experiments demonstrate that our approach achieves state-of-the-art results on two tasks.

1. Introduction

Person image synthesis has drawn a great deal of attention, due to various applications in the movie industry, e-commerce, person re-identification, etc. There are two critical tasks in person image synthesis: *pose transfer* [17, 22, 15, 40, 20, 23] and *pose-guided virtual try-on* [39, 5, 11, 27]. As shown in Figure 1, pose transfer task aims to transfer person images from one pose to other poses, and pose-guided virtual try-on task is to modify the clothing item based on pose transference. These two tasks have strong correlations and similarities, yet existing methods do not explore their correlations. Especially most methods of pose-guided virtual try-on [39, 11, 27] only implicitly model the pose transformation via learning the concatenation of the target pose or human parsing map and the source image. This implicit transformation is likely to result in blurry and implausible issues when dealing with large misalignment among different poses, which affects the quality of generated results.

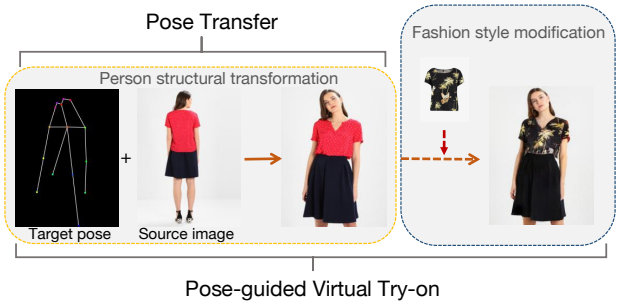

Figure 1: The correlations between two tasks. *Pose transfer* corresponds to a person's structural transformation between the source and the target images, while *pose-guided virtual try-on* aims to modify the fashion styles based on this transformation.

The key to obtaining high-quality results for these two tasks is to generate sharp human body structure and fine-grained textures (*e.g.*, patterns of clothes and hairs). For the structure generation, the target human body structure can be obtained via a transformation from the source one, such as from the back to the front structure in the Figure 1. Transformation simulation is the primary concern of pose transfer task. Previous methods exploit flow-based warping, obtaining this transformation with promising results [8, 15, 20]. However, due to the challenge of accurate flow computation and precise warping operation, it may cause artifacts around the structure [28, 24]. In addition to structure generation, most existing methods adopt perceptual-related constraints (*e.g.,* perceptual loss [14] and adversarial loss [7]) to guide the texture generation and ameliorate the visual quality of the results. However, these losses tend to optimize high-level perception, which does not only focus on textures but also includes other components such as style information. Therefore, the texture generation cannot be supervised effectively in the optimization process. For example, fine-grained region reconstruction can be further improved [30].

To address the above issues, in this paper, we propose a structure-transformed texture-enhanced network for person image synthesis. The proposed model comprises two key components: structure-transformed renderer and texture-enhanced stylizer. First, **structure-transformed renderer** aims at deforming the structure, explicitly solving the concern of pose alignments for pose transfer and pose-guided virtual try-on tasks. Here, we present a cross-modality deformable convolution to deal with this transformation, which avoids artifacts caused by the flow-based warping and fuses multi-modality information to better capture the structural motions. Then, **texture-enhanced stylizer** is designed to further enhance texture details due to the failure of retaining fine-grained regions. Specifically, the pose-guided high-frequency attention is introduced to enhance the high-frequency components with spatial contextual information. We establish a multi-modality long-range dependency to deal well with the invisible regions in the source image. Besides, this component enables users to manipulate fashionable garments controllably, modifying fashion styles for pose-guided virtual try-on task. Benefit from the two novel components, our model combines pose transfer task with pose-guided virtual try-on task, which can model the transformation between the source and the target poses explicitly and generate high-quality person images in diverse poses and even with various fashion styles.

The proposed approach exhibits superior performance over existing methods on the DeepFashion [16] and the FashionTryOn dataset [39]. Furthermore, we perform ablation experiments to validate the contribution of key components in our approach. The main contributions of our paper are summarized as:

- We propose a structure-transformed texture-enhanced network to handle pose-guided virtual try-on task with pose transfer task jointly. Experimental results show the superiority of our model on generating highly photorealistic and fashion-diversified results for person image synthesis.
- We design the structure-transformed renderer based on cross-modality deformable convolutions to process the person's structural transformation, which fuses multi-modality information to capture the structural motions.
- The texture-enhanced stylizer is proposed to enhance detailed textures and enable users to manipulate the fashion style, avoiding blurry textures and generating various fashion styles.

2. Related Works

2.1. Person Image Synthesis

Pose Transfer. Ma *et al*. [17] first define this task and devise a two-stage GAN to produce a coarse result and then refine it. Siarohin *et al*. [22] further improve the results via aligning local features with structure information. They introduce deformable skip connections to decompose the overall deformation by a set of local affine transformations. Zhu *et al*. [40] improve the transformation strategy by proposing a pose-attentional transfer block to transform the condition image to the target pose progressively. However, appearance information may be lost in multiple transfer processes. Han *et al*. [8] first exploit the optical flow to warp the clothing item to the target structure at the pixel level, and then generate the complete person image. Ren *et al*. [20] further introduce an unsupervised flow estimator at the feature level. However, the high-quality estimation of optical flow is hard to be obtained, and the flow-based warping is prone to produce artifacts. Therefore, our model uses deformable convolutions instead of the flow-based warping to transform source features.

Pose-guided Virtual Try-on. The single pose virtual try-on has been studied a lot [9, 26, 36, 35]. Since online customers have the desire to obtain multi-views of themselves wearing the desired clothing, Zheng *et al*. [39] present pose-guided virtual try-on task. They devise a pose-guided virtual try-on model that captures the deformation of the desired clothing and then produces the person image with the deformed clothing and target poses. Dong *et al*. [5] devise a coarse-to-fine model with the human parsing map. Wang *et al*. [27] design a Tree-Block to capture details of the image based on the multi-stage network. All the above approaches simply learn the concatenation of target structural information (*e.g.*, pose, parsing map) and the clothing as the structural transformation. Different from these methods, our model contains the structure-transformed renderer that processes the structural transformation specifically.

2.2. Deformable Convolution

Dai *et al*. [4] first present the deformable convolution, which generates kernel offsets from input features to learn information away from its regular local receptive field. Deformable convolutions have been widely used in several detection and recognition tasks, such as object detection [1] and action recognition [32, 18]. Recently, it is also used in other vision tasks. Yuan *et al*. [37] employ it with optical flow for dynamic scene deblurring. Wang *et al*. [28] and Tian *et al*. [24] adopt it to align the original and reference frames for video super-resolution. Inspired by these methods, we first apply it to address the task of person image synthesis.

2.3. Attention Mechanism

Attention mechanism has been applied to address many tasks, such as object detection [12] and semantic segmentation [33, 3]. Self attention [25] is a subset of the attention mechanism in the neural language process. Wang *et al*.

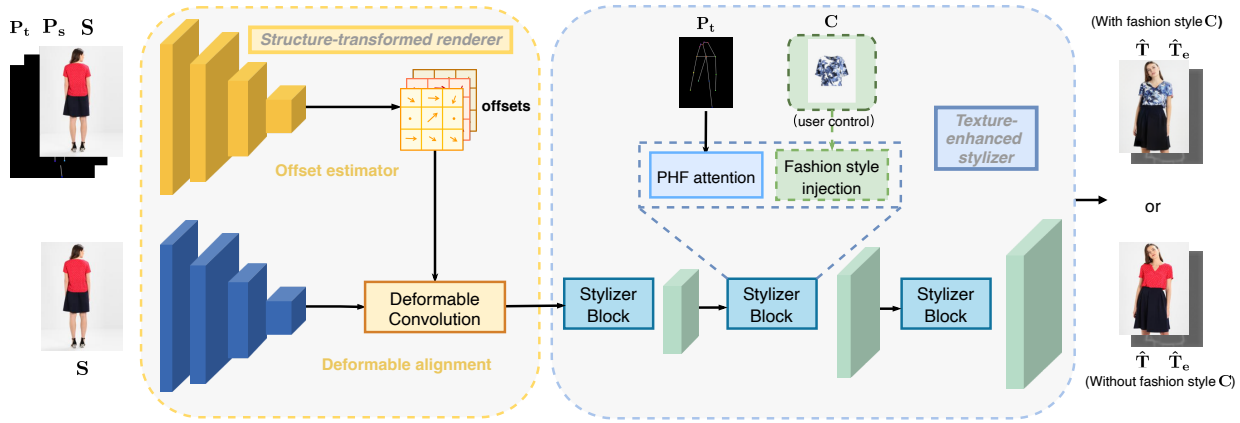

Figure 2: Overview of the proposed model. The offset estimator module generates the cross-modality offsets while the deformable alignment module applies these offsets to the deformable convolution and transforms the source image features. Then, the texture-enhanced stylizer, which consists of the PHF attention and the controllable fashion style injection, enhances textures and modifies the fashion style flexibly. Eventually, we obtain the synthesized person image and a corresponding edge map. With the fashion style injection, the final person image is synthesized with the desired clothing, while keeping the original clothing without this injection.

[29] extend it to a non-local attention module for computer vision, encoding the long-range dependencies. There are several methods based on the non-local module in various tasks [13, 34]. In contrast to the general non-local module, our texture-enhanced stylizer contains a pose-guided high-frequency attention module that fuses multi-modality information and enhances high-frequency visual appearances.

3. Our Approach

Given a source image $\mathbf{S}$, the source pose $\mathbf{P_s}$ and the target pose $\mathbf{P_t}$, our model aims to transfer the source image $\mathbf{S}$ to the target pose $\mathbf{P_t}$ for pose transfer task. The pose representation includes 18 human keypoints extracted by [2]. If injecting the fashion style $\mathbf{C}$, our model can synthesize a new person image $\hat{\mathbf{T}}$ with the target pose and the injected fashion item, achieving pose-guided virtual try-on task.

Our network architecture is shown in Figure 2, the **structure-transformed renderer** is based on the cross-modality deformable convolution to deal with the structural transformation. After the structure-transformed renderer, the transformed features are fed into the **texture-enhanced stylizer**. Our model comprises three stylizer blocks, each of which contains a pose-guided high-frequency (PHF) attention module and a fashion style injection module. The PHF attention module is designed to avoid blurry fine-grained textures and preserve the contextual relationship simultaneously. Then, users can control injecting the fashion style. It indicates that the green line in Figure 2 is optional. In the following, we will give a detailed description of each part of our model.

3.1. Structure-transformed Renderer

Structure-transformed renderer consists of two submodules: the offset estimator module and the deformable alignment module. The offset estimator is responsible for learning the structural motions between the source pose and the target pose. Here, we define these motions as the coordinate offsets Θ and learn them from the source image $\mathbf{S}$, the source pose $\mathbf{P_s}$, and the target pose $\mathbf{P_t}$. Let f_ϕ represent the network with learning parameters ϕ. The estimating process can be defined as:

$$\Theta = f_\phi\left(\mathbf{S}, \mathbf{P_s}, \mathbf{P_t}\right) \tag{1}$$

Let p_n mean each location in the sampling grid $\mathcal{R}$ of the conventional convolution. Θ can be expressed as: $\{\triangle p_n | \quad n = 1, 2, \ldots, |\mathcal{R}|\}$. It fuses multi-modality information (pose and image), which can effectively capture the deformation between the source and the target poses.

With the cross-modality offsets Θ, the deformable alignment module adopts deformable convolutions to learn the transformed features $\mathbf{F_{S_a}}$ from the source image $\mathbf{S}$, aligning the source image with the target pose at the feature level. Specifically, we obtain the source image features $\mathbf{F_S}$ from $\mathbf{S}$ via an encoder architecture and then we utilize the deformable convolution to learn $\mathbf{F_{S_a}}$ from $\mathbf{F_S}$. Let f_{dc} refer to the deformable convolution. For each position p_0 on the transformed features $\mathbf{F_{S_a}}$, we have:

$$\begin{aligned}\mathbf{F_{S_a}}\left(p_0\right) &= f_{dc}\left(\mathbf{F_S}, \Theta\right) \\ &= \sum_{p_n \in \mathcal{R}} \mathbf{w}\left(p_n\right) \cdot \mathbf{F_S}\left(p_0 + p_n + \Theta\right)\end{aligned} \tag{2}$$

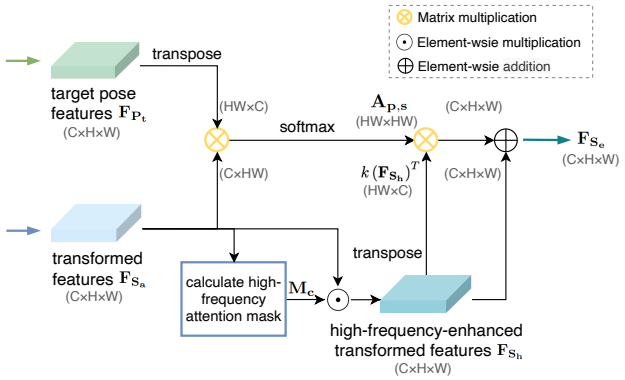

Figure 3: The framework of the proposed PHF attention module, which takes the target pose features $\mathbf{F}_{\mathbf{P_t}}$ and the transformed features $\mathbf{F}_{\mathbf{S_a}}$ as inputs and learns high-frequency-aware contextual dependencies.

where $\mathbf{w}$ means the convolution kernel weight and $p_n + \Theta$ refers to the augmented receptive field of convolutional operation. Distinguished from general deformable convolutions that produce kernel offsets from the original features to expand the receptive field, we utilize Θ to learn the structural deformation away from the regular local kernel neighborhood.

3.2. Texture-enhanced Stylizer

Pose-guided High-frequency (PHF) Attention. The PHF attention is designed to encode high-frequency-aware contextual dependencies under the guidance of the target pose, where high-frequency components correspond to textural information. Here, we explore the spatial contextual information to reduce suffering from occlusions in the source image and generate realistic textures. The framework of this module is shown in Figure 3. To simplify the notation, we use the first PHF attention block as an example to describe this module in this section.

The attention process is divided into two branches. We first calculate an affinity matrix $\mathbf{A}_{\mathbf{p,s}}$ to keep the attention process under the guidance of the target pose. In the notation that follows, we use $\mathbf{F}_{\mathbf{P_t}}$ to denote the target pose features extracted from the target pose via an encoder architecture. $\mathbf{F}_{\mathbf{S_a}}$ equals the transformed features generated by the structure-transformed renderer (in the other two attention modules, it represents the feature generated by the previous stylizer block). In addition, C, H and W are the number of channels, height, and width of features, respectively. $\mathbf{A}_{\mathbf{p,s}} \in \mathbb{R}^{HW \times HW}$ is defined as:

$$\mathbf{A}_{\mathbf{p,s}}(i,j) = \frac{\exp\left(g\left(\mathbf{F}_{\mathbf{P_t}}{}^{i}\right)^T \otimes h\left(\mathbf{F}_{\mathbf{S_a}}{}^{j}\right)\right)}{\sum_j \exp\left(g\left(\mathbf{F}_{\mathbf{P_t}}{}^{i}\right)^T \otimes h\left(\mathbf{F}_{\mathbf{S_a}}{}^{j}\right)\right)} \quad (3)$$

where (i,j) means the coordinate location, $g(\cdot), h(\cdot)$ refer to 1×1 convolution and the reshaping operation, $\otimes$ means matrix multiplication. Both the pose and the appearance information are considered for the affinity matrix.

In the other branch, in order to enhance high-frequency details, we design a high-frequency-aware mask $\mathbf{M_c}$:

$$\mathbf{M_c} = \sigma\left(f_{c_1}\left(\mathbf{F}_{\mathbf{S_a}}\right) - f_{c_2}\left(\mathbf{F}_{\mathbf{S_a}}\right)\right) \in \mathbb{R}^{C\times H\times W} \quad (4)$$

where f_{c_1} and f_{c_2} denote the dilated convolution with 1×1 kernel and *Gaussian blurring* with 5×5 kernel respectively, σ refers to the Sigmoid function. *Gaussian blurring* f_{c_2} can be regarded as filtering high-frequency components of an image. The results of f_{c_2} are subtracted from the outcomes of f_{c_1}, which indicates preserving high-frequency details and reducing the effect of low-frequency ones in all components. After the Sigmoid function, the attention mask $\mathbf{M_c}$ applies more weights to high-frequency components. Then, the high-frequency-enhanced transformed features $\mathbf{F}_{\mathbf{S_h}}$ is defined as:

$$\mathbf{F}_{\mathbf{S_h}} = \mathbf{M_c} \odot \mathbf{F}_{\mathbf{S_a}} \in \mathbb{R}^{C\times H\times W} \quad (5)$$

where $\odot$ means the element-wise multiplication.

Eventually, we obtain the high-frequency-enhanced non-local features $\mathbf{F}_{\mathbf{S_e}}$ as the output of the attention module:

$$\mathbf{F}_{\mathbf{S_e}} = f\left(\left(\mathbf{A}_{\mathbf{p,s}} \otimes k\left(\mathbf{F}_{\mathbf{S_h}}\right)^T\right)^T\right) \oplus \mathbf{F}_{\mathbf{S_h}} \in \mathbb{R}^{C\times H\times W} \quad (6)$$

where $f(\cdot)$ and $k(\cdot)$ refer to 1×1 convolution and the reshaping operation, $\oplus$ represents the element-wise addition. Here, $\mathbf{A}_{\mathbf{p,s}} \otimes k\left(\mathbf{F}_{\mathbf{S_h}}\right)^T$ establishes a high-frequency-aware non-local context under the guidance of pose information, and then $\mathbf{F}_{\mathbf{S_e}}$ is computed in a residual manner to optimize the attention process.

Controllable Style Injection. After the PHF attention, the fashion style (*e.g.*, the clothing image) is injected flexibly to modify fashion textures. First, we employ the TPS transformation as [26, 39] to match the clothing with the target pose geometrically and then learn the fashion style features from the warped clothing with an encoder. Finally, we fuse the fashion style features with the features outputted from the PHF attention module via concatenation and 1×1 convolution. With the controllable style injection, users can control whether modifying the fashion style when generating target person images.

After three stylizer blocks and upsampling layers, the model outputs a result of 4 channels. We obtain the synthesized image $\hat{\mathbf{T}}$ by splitting the 4 channels image into 3 channels RGB synthesized image and 1 channel synthesized edge map $\hat{\mathbf{T}}_{\mathbf{e}}$. The synthesized edge map is a grayscale image and illustrates the structure of the synthesized image. We utilize it to further constrain the structure generation.

3.3. Loss Functions

Structure loss. In order to generate body structure correctly, we introduce the structure loss based on the mean-square error. Here, we extract the edge map from the target image as our ground truth:

$$\mathcal{L}_{structure} = \left\| \hat{\mathbf{T}}_{\mathbf{e}} - \mathbf{T}_{\mathbf{e}} \right\|_2^2 \tag{7}$$

where $\hat{\mathbf{T}}_{\mathbf{e}}$ and $\mathbf{T}_{\mathbf{e}}$ refer to the synthesized edge map and the ground truth edge map, respectively. This loss term is introduced to guide the edge generation and yields an accurate structure partition.

Following previous methods [40, 15, 20], we utilize other loss functions as below:

Adversarial loss. We apply the generative adversarial framework to mimic the distribution of the target images. The discriminator D is used to distinguish the synthesized images generated by the generator G from real images. Therefore, the adversarial loss L_{adv} is defined as:

$$\begin{aligned} \mathcal{L}_{\text{adv}} &= \mathbb{E}[\log D(\mathbf{T})] \\ &+ \mathbb{E}[\log(1 - D(G(\mathbf{S}, \mathbf{P_s}, \mathbf{P_t}, (\mathbf{C}))] \end{aligned} \tag{8}$$

Reconstruction loss. To constrain the synthesized image similar to the target image in the pixel level, we define the L1 loss as:

$$\mathcal{L}_{rec} = \left\| \hat{\mathbf{T}} - \mathbf{T} \right\|_1 \tag{9}$$

Feature similarity loss. We use the cosine similarity and the Euclidean distance to enforce the transformed features $\mathbf{F}_{\mathbf{S_a}}$ be close to the feature maps of the ground truth target image $\mathbf{F}_{\mathbf{T}}$:

$$\mathcal{L}_{\text{fea}} = \lambda_1 \exp\left(-\mu\left(\mathbf{F}_{\mathbf{S_a}}, \mathbf{F}_{\mathbf{T}}\right)\right) + \lambda_2 \rho\left(\mathbf{F}_{\mathbf{S_a}}, \mathbf{F}_{\mathbf{T}}\right) \tag{10}$$

where μ means the cosine similarity and ρ means the Euclidean distance, λ_1 and λ_2 mean their weights in this loss, respectively.

Perceptual and Style loss. Except for pixel-level constraints, we utilize the perceptual loss and the style loss [14] at VGG feature level to ensure perceptually plausible results. The perceptual loss can be defined as:

$$\mathcal{L}_{\text{perc}} = \sum_i \left\| \phi_i\left(\hat{\mathbf{T}}\right) - \phi_i\left(\mathbf{T}\right) \right\|_1 \tag{11}$$

where ϕ_i is the activation map of the i-th layer of a visual perception (pre-trained VGG19) network. Let $\mathcal{G}$ be the Gram matrix. The style loss calculates the statistic error between the activation map as:

$$\mathcal{L}_{\text{style}} = \sum_j \left\| \mathcal{G}_j^{\phi}\left(\hat{\mathbf{T}}\right) - \mathcal{G}_j^{\phi}\left(\mathbf{T}\right) \right\|_1 \tag{12}$$

Model	FID↓	LPIPS↓	SSIM↑
VU-Net [6]	23.708	0.264	0.763
Def-GAN [22]	18.462	0.233	0.760
PATN [40]	20.749	0.253	0.772
GFLA [20]	11.871	0.190	0.770
Ours	**9.888**	**0.182**	**0.774**

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

Model	FID↓	LPIPS↓	SSIM↑
VTOAP [39]	21.205	0.208	0.738
VTDC [27]	9.338	0.154	0.779
Ours	**6.401**	**0.138**	**0.782**

Table 2: Quantitative comparison with state-of-the-art methods on FashionTryOn.

Total loss. In summary, the total loss of our approach can be expressed as:

$$\begin{aligned} \mathcal{L}_{\text{total}} &= \lambda_{\text{structure}}\mathcal{L}_{\text{structure}} + \lambda_{\text{adv}}\mathcal{L}_{adv} + \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} \\ &+ \lambda_{\text{fea}}\mathcal{L}_{\text{fea}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{style}}\mathcal{L}_{\text{style}} \end{aligned} \tag{13}$$

4. Experiments

4.1. Dataset and Evaluation Metrics

Dataset. We conduct experiments on the DeepFashion dataset [16] for pose transfer and the FashionTryOn dataset [39] for pose-guided virtual try-on. The DeepFashion dataset contains 52,712 high-quality model images with a resolution of 256×256. We follow a similar procedure as the prior work [40] to partition training data and testing data for a fair comparison. The training set has 101,966 pairs, and the testing set retains 8,750 pairs. The FashionTryOn dataset [39] consists of 21,209 training pairs and 7,520 testing pairs, with each comprising a clothing item image and two model images in different poses. The image is in the resolution of 256 × 176. Since it is a challenge to collect an ideal dataset with different poses and different clothing, the target image shares the same clothing item with the source person image. Following previous methods [39, 5], we train our model with a masked source image, where the clothing item in the source image is not fed into the network.

Evaluation Metrics. We follow previous methods to use Structure Similarity (SSIM) [31] as our evaluation metrics. We also introduce Learned Perceptual Image Patch Similarity (LPIPS) [38] and Frechet Inception Distance (FID) [10] as our metrics. LPIPS calculates a weighted L_2 distance between the synthesized image and the target ground truth image at the feature level. FID computes the Wasserstein-2 distance between the distributions of synthesized images and target ground truth images.

Figure 4: Examples of qualitative comparison results with VUNet [6], Def-GAN [22], PATN [40], GFLA [20] for pose transfer and VTOAP [39], VTDC [27] for pose-guided virtual try-on.

4.2. Implementation Details

We use PyTorch to implement our model. The encoder architecture used in our model is similar to that in U-Net [21]. We train our model from scratch. The ADAM optimizer is used to back-propagate gradients, where we set β_1 and β_2 to 0.9 and 0.999. The initial learning rate of the generator is 10^{-4} and that of the discriminator is 10^{-6}. The weights for the loss terms are set to λ_{edge} = 50, λ_{fea} = 4, λ_{adv} = 2, λ_{rec} = 3, λ_{perc} = 1, and λ_{style} = 400. The λ_1 and λ_2 in feature similarity loss are set to 1 and 0.001, respectively. On the DeepFashion dataset, we train our model without the fashion style injection while the fashion style is injected on the FashionTryOn dataset. Due to the limitations of the dataset, the injected clothing is the same as the one in the source image at the training and the testing phases. For the qualitative evaluation, we randomly shuffle the test set and test different clothing items with diverse source images on the FashionTryOn. We train our model for 100 epochs with a batch size of 16.

4.3. Comparisons

We compare our model with several state-of-the-art approaches including VUNet [6], Def-GAN [22], PATN [40], and GFLA [20] for pose transfer task. For pose-guided virtual try-on, we compare our model with VTOAP [39] and VTDC [27]. We evaluate our proposed method with both qualitative and quantitative comparisons. The quantitative results of pose transfer and pose-guided virtual try-on are listed in Table 1 and 2, respectively. Figure 4 gives the typical qualitative examples of two tasks.

Quantitative results. The proposed method outperforms these competing methods on three metrics. For pose transfer task, VUNet does not deal with the structure transformation between the source appearance information and the target pose, where it shows relatively weak results on two metrics. Def-GAN and PATN introduce various strategies to combine the source appearance and the target pose. It can be seen that both of them perform better than VUNet. Furthermore, GFLA designs the flow-based operation to warp source image features locally, which can further cope with complex structure transformations. Compared with GFLA, our model has its unique advantages in dealing with occlusions and avoiding artifacts that come from flow-based warping. In addition to these, all these methods ignore the significance of texture enhancement. This is the reason that our model achieves superior results over other methods.

For pose-guided virtual try-on task, note that both VTOAP and VTDC do not expressly handle the multi-pose transformation. The target structure information, such as the pose or the parsing map, is simply concatenated with the clothing to generate the synthesized image. Both textures and structure are prone to be weakened in the generation process. Therefore, our method achieves the best performance among these state-of-the-art pose-guided virtual try-on methods. Furthermore, as depicted in Figure 5, the human parsing maps used as input guidance in VTDC are inclined to result in some inaccurate geometric division.

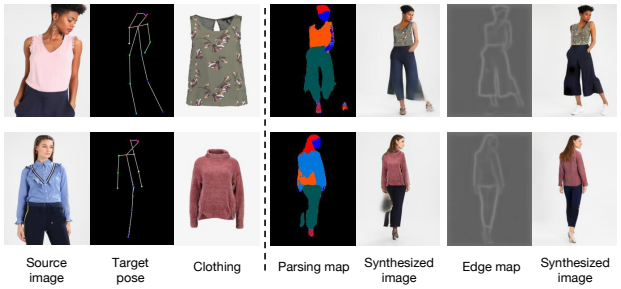

Figure 5: Examples of the synthesized person image with human parsing maps in VTDC (left) and ours with edge maps (right).

Figure 6: Examples of the qualitative results of ablation study on the DeepFashion dataset.

Compared with the human parsing map, the edge map is more precise in the structure division. Therefore, our model utilizes the edge map as explicit supervision and constrains the deformation of the pose, and thus boosts the performance.

Qualitative results. Qualitative comparisons are visualized in Figure 4. The left part is on DeepFashion, which is the result of pose transfer. Compared with other state-of-the-art methods, our model performs the capacity to model textures and structure for person image synthesis where VUNet, Def-GAN, and PATN obviously fail to generate realistic textures of the face, hairs, and clothing. For example, VUNet losses textures of polka dots on clothing in the second row (from top to bottom). GFLA is more accurate in generating textures. However, due to the limitations of flow-based warping, GFLA suffers from the artifacts around the person's structure. Thanks to the cross-modality deformable convolution, our model demonstrates its strength in preventing artifacts, which brings us a sharper body structure than GFLA. Meanwhile, the PHF attention helps generate clear detailed textures, for example, the white lace generated by our model in the third row.

For pose-guided virtual try-on task, typical visual comparisons are shown in the right part of Figure 4. Our method and VTDC generate vivid images while the face and body of VTOAP are blurry. Although VTDC adds the human parsing map as input to enhance body structure, the results still have some regions that have structure confusions, such as the arm in the second row. Compared with VTDC, the proposed approach explicitly models the structural transformation, tackling the displacements of body regions between different poses. Thus, the structure of ours is sharper. Meanwhile, our model performs better on generating fine-grained textures than others (*e.g.*, the face in the third row and the pants in the last row). Furthermore, we visualize our results without editing the fashion style. The results are generated with the same clothing as the source image, which can be considered as the result of pose transfer task.

	PHFA	SL	CDC	FID ↓	LPIPS↓	SSIM ↑
ST-SL	×	✓	✓	12.771	0.191	0.766
STTE	✓	×	✓	11.296	0.184	0.771
ENTE-SL	✓	✓	×	15.503	0.211	0.762
FLTE-SL	✓	✓	×	11.054	0.188	0.767
Ours	✓	✓	✓	**9.888**	**0.182**	**0.774**
ST-SL	×	✓	✓	8.240	0.156	0.757
STTE	✓	×	✓	7.407	0.152	0.762
ENTE-SL	✓	✓	×	14.004	0.182	0.744
FLTE-SL	✓	✓	×	7.146	0.158	0.760
Ours	✓	✓	✓	**6.401**	**0.138**	**0.782**

Table 3: Evaluation results of the ablation study. The top half is on the DeepFashion dataset and the bottom half is on the FashionTryOn dataset. PHFA represents the PHF attention module. SL means the structure loss. CDC represents the cross-modality deformable convolution.

4.4. Ablation study

In this subsection, we investigate how each component contributes to the proposed method. Several variants are provided to verify the effectiveness of the structure loss, PHF attention, and the structure-transformed renderer.

ST-SL. In this model, We eliminate the PHF attention module of the texture-enhanced stylizer from the original full model.

STTE. This model is our proposed model without the structure loss. This variant has the full structure-transformed renderer and the PHF attention module as the proposed model.

ENTE-SL. We replace the structure-transformed renderer with a typical encoder architecture. In this variant, the input of the encoder is set to the concatenation of the source image, the source pose, and the target pose.

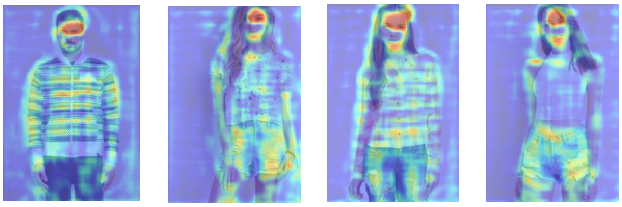

Figure 7: Visualization results of the high frequency attention mask $\mathbf{M_c}$ in the PHF attention module.

FLTE-SL. In this model, we replace the cross-modality deformable convolution of the structure-transformed renderer with the operation of flow-based warping.

Full model (Ours). We employ the full structure-transformed renderer and the attention module in this model and train it with all loss functions adopted in this paper.

Effectiveness of PHF attention. From Table 3, we can observe that our full model outperforms ST-SL. It benefits from the PHF attention module. Meanwhile, as depicted in qualitative results, ablating the PHF attention module incurs messy fine-grained details. For example, the result of ST-SL in the second row shows irregular textures on the clothing region due to the lack of texture enhancement, thus demonstrates the ability of the PHF attention module to strengthen texture details. Furthermore, the attention masks in PHF attention are visualized in Figure 7. It can be seen that it indeed assigns higher weights to complex textures in clothing and face. The regions with smooth textures (*e.g.*, the upper clothing in the far left result) obtain less attention.

Effectiveness of the structure loss. We employ the structure loss to guide the structure generation. According to the quantitative comparison (Table 3), the full model outperforms STTE in all metrics. Moreover, as illustrated in Figure 6, the structure of the full model demonstrates nature-looking better than STTE. Both quantitative and qualitative results confirm the effectiveness of this proposed loss.

Effectiveness of the structure-transformed renderer. Evaluation results are illustrated in Table 3. The full model has a better performance than ENTE-SL and FLTE-SL, which means that the structure-transformed renderer can effectively solve the problem of the transformation between the source and the target poses. The cross-modality deformable convolution also leads to a stable performance gain. In Figure 6, we observe that the full model generates more reasonable results than ENTE-SL. Furthermore, since the flow-based warping tends to have wrong sampling regions, it incurs unreal textures such as the leg near the feet in the last row (from top to bottom) of FLTE-SL.

4.5. Applications

The proposed method can ameliorate virtual fitting room applications, where customers can try the fashion style of the images of models. Several examples of applications are shown in Figure 8. Another person image replaces the input of the garment in the proposed model, and the fashion style is extracted from this person image. Concretely, we utilize the Style-Encoder proposed by [19] to extract the fashion textures from the style image in Figure 8 and then inject it back into the source image by the texture-enhanced stylizer. In this way, users are able to select various fashion styles from other person images to generate a new person image, giving more options to the users.

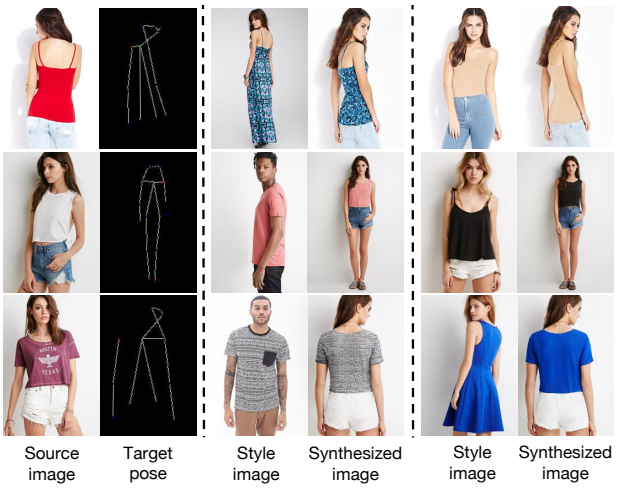

Figure 8: Examples of fashion style transfer.

5. Conclusion

In this paper, a novel approach is presented with the structure-transformed renderer and the texture-enhanced stylizer to synthesize person images, exploring correlations between pose transfer and pose-guided virtual try-on tasks. In contrast to existing methods, we emphasize the structure and texture generation. Specifically, in order to eliminate the blurry artifacts that come from the flow-based operation, we are the first to apply the deformable convolution to capture the structural offsets between the source and the target poses. The structure loss is proposed to constrain the structure generation. Meanwhile, the PHF attention module is designed to enhance textures. Several experiments are conducted for two tasks. The experimental results demonstrate the effectiveness and versatility of the proposed method.

Acknowledgments

This work was supported by Key-Area Research and Development Program of Guangdong Province (No.2019B121204008) and Shenzhen Fundamental Research Program (GXWD20201231165807007-20200806163656003). We thank all reviewers for their valuable comments.

References

[1] Gedas Bertasius, Lorenzo Torresani, and Jianbo Shi. Object detection in video with spatiotemporal sampling networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 331–346, 2018. 2

[2] Zhe Cao, Tomas Simon, Shih En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[3] Liang-Chieh Chen, Yi Yang, Jiang Wang, Wei Xu, and Alan L Yuille. Attention to scale: Scale-aware semantic image segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3640–3649, 2016. 2

[4] Jifeng Dai, Haozhi Qi, Yuwen Xiong, Yi Li, Guodong Zhang, Han Hu, and Yichen Wei. Deformable convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 764–773, 2017. 2

[5] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9026–9035, 2019. 1, 2, 5

[6] 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. 5, 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. 1

[8] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10471–10480, 2019. 1, 2

[9] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 2

[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. 5

[11] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. Fashionon: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 275–283, 2019. 1

[12] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 2

[13] Wentao Jiang, Si Liu, Chen Gao, Jie Cao, Ran He, Jiashi Feng, and Shuicheng Yan. Psgan: Pose and expression robust spatial-aware gan for customizable makeup transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5194–5202, 2020. 3

[14] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 1, 5

[15] 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, 5

[16] 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, 5

[17] 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

[18] Khoi-Nguyen C Mac, Dhiraj Joshi, Raymond A Yeh, Jinjun Xiong, Rogerio S Feris, and Minh N Do. Learning motion in feature space: Locally-consistent deformable convolution networks for fine-grained action detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6282–6291, 2019. 2

[19] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020. 8

[20] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020. 1, 2, 5, 6

[21] 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. 6

[22] 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, 5, 6

[23] Hao Tang, Song Bai, Li Zhang, Philip HS Torr, and Nicu Sebe. Xinggan for person image generation. In *European Conference on Computer Vision*, pages 717–734. Springer, 2020. 1

[24] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3360–3369, 2020. 1, 2

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia

Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 2

[26] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018. 2, 4

[27] Jiahang Wang, Tong Sha, Wei Zhang, Zhoujun Li, and Tao Mei. Down to the last detail: Virtual try-on with fine-grained details. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 466–474, 2020. 1, 2, 5, 6

[28] Xintao Wang, Kelvin CK Chan, Ke Yu, Chao Dong, and Chen Change Loy. Edvr: Video restoration with enhanced deformable convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 1, 2

[29] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018. 3

[30] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018. 1

[31] 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. 5

[32] Junwu Weng, Mengyuan Liu, Xudong Jiang, and Junsong Yuan. Deformable pose traversal convolution for 3d action and gesture recognition. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 136–152, 2018. 2

[33] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *Proceedings of the European conference on computer vision (ECCV)*, pages 3–19, 2018. 2

[34] Kai Xu, Minghai Qin, Fei Sun, Yuhao Wang, Yen-Kuang Chen, and Fengbo Ren. Learning in the frequency domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1740–1749, 2020. 3

[35] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 7850–7859, 2020. 2

[36] 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*, pages 10511–10520, 2019. 2

[37] Yuan Yuan, Wei Su, and Dandan Ma. Efficient dynamic scene deblurring using spatially variant deconvolution network with optical flow guided training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3555–3564, 2020. 2

[38] 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. 5

[39] Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. Virtually trying on new clothing with arbitrary poses. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 266–274, 2019. 1, 2, 4, 5, 6

[40] 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, 5, 6