Structure-aware Person Image Generation with Pose Decomposition and Semantic Correlation

Jilin Tang[1], Yi Yuan[1]*, Tianjia Shao[2], Yong Liu[3], Mengmeng Wang[3], Kun Zhou[2]
[1] NetEase Fuxi AI Lab
[2] State Key Lab of CAD&CG, Zhejiang University
[3] Institute of Cyber-Systems and Control, Zhejiang University
{tangjilin, yuanyi}@corp.netease.com, {tjshao,mengmengwang}@zju.edu.cn, yongliu@iipc.zju.edu.cn, kunzhou@acm.org

Abstract

In this paper we tackle the problem of pose guided person image generation, which aims to transfer a person image from the source pose to a novel target pose while maintaining the source appearance. Given the inefficiency of standard CNNs in handling large spatial transformation, we propose a structure-aware flow based method for high-quality person image generation. Specifically, instead of learning the complex overall pose changes of human body, we decompose the human body into different semantic parts (e.g., head, torso, and legs) and apply different networks to predict the flow fields for these parts separately. Moreover, we carefully design the network modules to effectively capture the local and global semantic correlations of features within and among the human parts respectively. Extensive experimental results show that our method can generate high-quality results under large pose discrepancy and outperforms state-of-the-art methods in both qualitative and quantitative comparisons.

Introduction

Pose guided person image generation (Ma et al. 2017), which aims to synthesize a realistic-looking person image in a target pose while preserving the source appearance details (as depicted in Figure 1), has aroused extensive attention due to its wide range of practical applications for image editing, image animation, person re-identification (ReID), and so on.

Motivated by the development of Generative Adversarial Networks (GANs) in the image-to-image transformation task (Zhu et al. 2017), many researchers (Ma et al. 2017, 2018; Zhu et al. 2019; Men et al. 2020) attempted to tackle the person image generation problem within the framework of generative models. However, as CNNs are not good at tackling large spatial transformation (Ren et al. 2020), these generation-based models may fail to handle the feature misalignment caused by the spatial deformation between the source and target image, leading to the appearance distortions. To deal with the feature misalignment, recently, appearance flow based methods have been proposed (Ren et al. 2020; Liu et al. 2019; Han et al. 2019) to transform the source features to align them with the target pose, modeling the dense pixel-to-pixel correspondence between the source and target features. Specifically, the appearance flow based methods aim to calculate the 2D coordinate offsets (i.e., appearance flow fields) that indicate which positions in the source features should be sampled to reconstruct the corresponding target features. With such flow mechanism, the existing flow based methods can synthesize target images with visually plausible appearances for most cases. However, it is still challenging to generate satisfying results when there are large pose discrepancies between the source and target images (see Figure 5 for example).

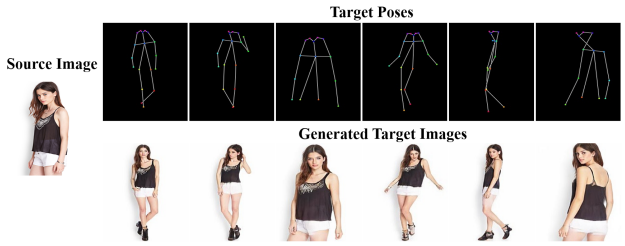

Figure 1: The generated person images in random target poses by our method.

To tackle this challenge, we propose a structure-aware flow based method for high-quality person image generation. The key insight of our work is, incorporating the structure information can provide important priors to guide the network learning, and hence can effectively improve the results. First, we observe that the human body is composed of different parts with different motion complexities w.r.t. pose changes. Hence, instead of using a unified network to predict the overall appearance flow field of human body, we decompose the human body into different semantic parts (e.g., head, torso, and legs) and employ different networks to estimate the flow fields for these parts separately. In this way, we not only reduce the difficulty of learning the complex overall pose changes, but can more precisely capture the pose change of each part with a specific network. Second, for close pixels belonging to each part of human body, the appearance features are often semantically correlated. For example, the adjacent positions inside the arm should have similar appearances after being transformed to a new pose. To this end, compared to the existing methods which generate features at target positions independently with limited

*Corresponding author

receptive fields, we introduce a *hybrid dilated convolution block* which is composed of sequential convolutional layers with different dilation rates (Yu and Koltun 2015; Chen et al. 2017; Li, Zhang, and Chen 2018) to effectively capture the short-range semantic correlations of local neighbors inside human parts by enlarging the receptive field of each position. Third, the semantic correlations also exist for the features of different human parts that are far away from each other, owning to the symmetry of human body. For instance, the features of the left and right sleeves are often required to be consistent. Therefore, we design a lightweight yet effective non-local component named *pyramid non-local block* which combines the multi-scale pyramid pooling (He et al. 2015; Kim et al. 2018) with the standard non-local operation (Wang et al. 2018) to capture the long-range semantic correlations across different human part regions under different scales.

Technically, our network takes as input a source person image and a target pose, and synthesizes a new person image in the target pose while preserving the source appearance. The network architecture is composed of three modules. The part-based flow generation module divides the human joints into different parts, and deploys different models to predict local appearance flow fields and visibility maps of different parts respectively. Then, the local warping module warps the source part features extracted from the source part images, so as to align them with the target pose while capturing the short-range semantic correlations of local neighbors within the parts via the *hybrid dilated convolution block*. Finally, the global fusion module aggregates the warped features of different parts into the global fusion features and further applies the *pyramid non-local block* to learn the long-range semantic correlations among different part regions, and finally outputs a synthesized person image.

The main contributions can be summarized as:

- We propose a structure-aware flow based framework for pose guided person image generation, which can synthesize high-quality person images even with large pose discrepancies between the source and target images.
- We decompose the task of learning the overall appearance flow field into learning different local flow fields for different semantic body parts, which can ease the learning and capture the pose change of each part more precisely.
- We carefully design the modules in our network to capture the local and global semantic correlations of features within and among human parts respectively.

Related Work

Pose guided person image generation can be regarded as a typical image-to-image transformation problem (Isola et al. 2017; Zhu et al. 2017) where the goal is to convert a source person image into a target person image conditioned on two constraints: (1) preserving the person appearance in the source image and (2) deforming the person pose into the target one.

Ma et al. (Ma et al. 2017) proposed a two-stage generative network named PG^2 to synthesize person images in a coarse-to-fine way. Ma et al. (Ma et al. 2018) further improved the performance of PG^2 by disentangling the foreground, background, and pose with a multi-branch network. However, the both methods require a complicated staged training process and have large computation burden. Zhu et al. (Zhu et al. 2019) proposed a progressive transfer network to deform a source image into the target image through a series of intermediate representations to avoid capturing the complex global manifold directly. However, the useful appearance information would degrade inevitably during the sequential feature transfers, which may lead to the blurry results lacking vivid appearance details. Essner et al. (Esser, Sutter, and Ommer 2018) combined the VAE (Kingma and Welling 2013) and U-Net (Ronneberger, Fischer, and Brox 2015) to model the interaction between appearance and shape. However, the common skip connections of U-Net can't deal with the feature misalignments between the source and target pose reliably. To tackle this issue, Siarohin et al. (Siarohin et al. 2018) further proposed the deformable skip connections to transform the local textures according to the local affine transformations of certain sub-parts. However, the degrees of freedom are limited (i.e., 6 for affine), which may produce inaccurate and unnatural transformations when there are large pose changes.

Recently, a few flow-based methods have been proposed to take advantage of the appearance flow (Zhou et al. 2016; Ren et al. 2019) to transform the source image to align it with the target pose. Han et al. (Han et al. 2019) introduced a three-stage framework named ClothFlow to model the appearance flow between source and target clothing regions in a cascaded manner. However, they warp the source image at the pixel level instead of the feature level, which needs an extra refinement network to handle the invisible contents. Li et al. (Li, Huang, and Loy 2019) leveraged the 3D human model to predict the appearance flow, and warped both the encoded features and the raw pixels of source image. However, they require to fit the 3D human model to all images to obtain the annotations of appearance flows before the training, which is too expensive to limit its application. Ren et al. (Ren et al. 2020) designed a global-flow local-attention framework to generate the appearance flow in an unsupervised way and transform the source image at the feature level reasonably. However, this method directly takes the overall source and target pose as input to predict the appearance flow of the whole human body, which may be unable to tackle the large discrepancies between the source and target pose reliably. Besides, this method produces features at each target position independently and doesn't consider the semantic correlations among target features at different locations.

The Proposed Method

Figure 2 illustrates the overall framework of our network. It mainly consists of three modules: the part-based flow generation module, the local warping module, and the global fusion module. In the following sections, we will give a detailed description of each module.

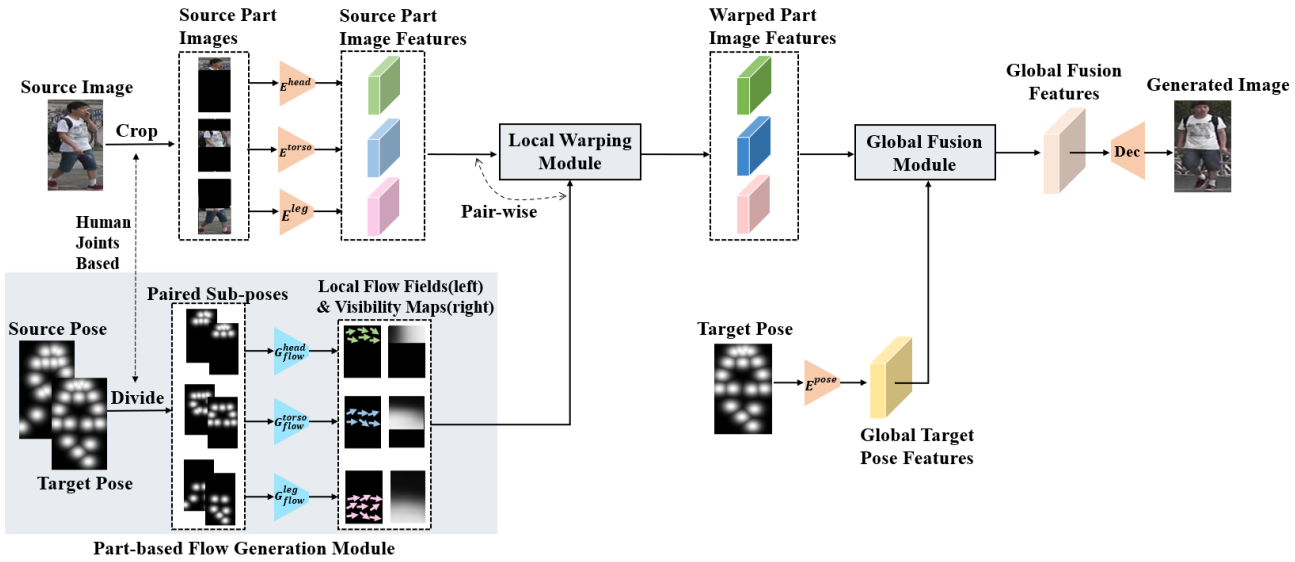

Figure 2: Overview of the proposed method. It mainly consists of three modules: the part-based flow generation module, the local warping module, and the global fusion module.

Part-based Flow Generation Module

We first introduce a few notations. Let $P_s \in \mathbb{R}^{18\times h\times w}$ and $P_t \in \mathbb{R}^{18\times h\times w}$ represent the overall pose of the source image $I_s \in \mathbb{R}^{3\times h\times w}$ and target image $I_t \in \mathbb{R}^{3\times h\times w}$ respectively, where the 18 channels of the pose correspond to the heatmaps that encode the spatial locations of 18 human joints. The joints are extracted with the OpenPose (Cao et al. 2017). As shown in Figure 2, our part-based flow generation module first decomposes the overall pose into different sub-poses via grouping the human joints into different parts based on the inherent connection relationship among them, Then, different sub-models $G^{local}_{flow} = \left\{G^{head}_{flow}, G^{torso}_{flow}, G^{leg}_{flow}\right\}$ are deployed to generate the local appearance flow fields and visibility maps of corresponding human parts respectively. Specifically, let $P^{local}_s = \left\{P^{head}_s, P^{torso}_s, P^{leg}_s\right\}$ and $P^{local}_t = \left\{P^{head}_t, P^{torso}_t, P^{leg}_t\right\}$ denote the decomposed source and target sub-poses, where each sub-pose corresponds to a subset of the 18 heatmaps of human joints. The sub-models G^{local}_{flow} take as input P^{local}_s and P^{local}_t, and output the local appearance flow fields W^{local} and visibility maps V^{local}:

$$W^{local}, V^{local} = G^{local}_{flow}(P^{local}_s, P^{local}_t), \tag{1}$$

where $W^{local} = \left\{W^{head}, W^{torso}, W^{leg}\right\}$ records the 2D coordinate offsets between the source and target features of corresponding parts, and $V^{local} = \left\{V^{head}, V^{torso}, V^{leg}\right\}$ stores confidence values between 0 and 1 representing whether the information of certain target positions exists in the source features.

Local Warping Module

The generated local appearance flow fields W^{local} and visibility maps V^{local} provide important guidance on understanding the spatial deformation of each part region between the source and target image, specifying which positions in the source features could be sampled to generate the corresponding target features. Therefore, our local warping module exploits this information to model the dense pixel-to-pixel correspondence between the source and target features. As shown in Figure 2, we first crop different part images from the source image, and encode them into the corresponding source part image features $F^{local}_s = \left\{F^{head}_s, F^{torso}_s, F^{leg}_s\right\}$. Then, under the guidance of generated local appearance flow fields W^{local}, our local warping module warps F^{local}_s to obtain the warped source features $F^{local}_{s,w} = \left\{F^{head}_{s,w}, F^{torso}_{s,w}, F^{leg}_{s,w}\right\}$ aligned with the target pose. Specifically, for each target position $p = (x, y)$ in the features $F^{local}_{s,w}$, a sampling position is allocated according to the coordinate offsets $\triangle p = (\triangle x, \triangle y)$ recorded in the flow fields W^{local}. The features at target position are fetched from the corresponding sampling position in the source features by the bilinear interpolation. Further details are available in our supplementary material. The procedure can be written as:

$$F^{local}_{s,w} = G_{warp}(F^{local}_s, W^{local}). \tag{2}$$

Considering not all appearance information of the target image can be found in the source image due to different visibilities of the source and target pose, we further take advantage of the generated local visibility maps V^{local} to select the reasonable features between $F^{local}_{s,w}$ and the local target pose features $F^{local}_{pose} = \left\{F^{head}_{pose}, F^{torso}_{pose}, F^{leg}_{pose}\right\}$ which are encoded from the target sub-poses. The feature selection using visibility maps is defined as:

$$F^{local}_{s,w,v} = V^{local} \cdot F^{local}_{s,w} + (1 - V^{local}) \cdot F^{local}_{pose}, \tag{3}$$

where $F^{local}_{s,w,v} = \left\{F^{head}_{s,w,v}, F^{torso}_{s,w,v}, F^{leg}_{s,w,v}\right\}$ denotes the selected features for different parts.

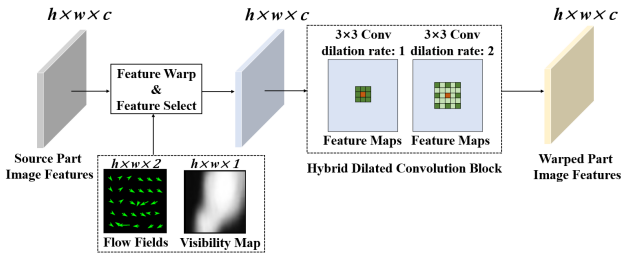

Figure 3: The local warping module. It warps the source features encoded from the corresponding part images to align them with the target pose while capturing the short-range semantic correlations of local neighbors within the parts.

At last, in order to perceive local semantic correlations inside human parts, as shown in Figure 3, we further introduce a *hybrid dilated convolution block* which is composed of sequential convolutional layers with different dilation rates (e.g., $\{1, 2\}$ in our implementation) to capture the short-range semantic correlations of local neighbors within parts by enlarging the receptive field of each position. Specifically, a dilated convolution with rate r can be defined as:

$$y(m,n) = \sum_i \sum_j x(m + r \times i, n + r \times j) w(i,j), \quad (4)$$

where $y(m,n)$ is the output of dilated convolution from input $x(m,n)$, and $w(i,j)$ is the filter weight. Let G_{hdcb} represent the *hybrid dilated convolution block*. The final warped local image features of different human parts $F_{warp}^{local} = \left\{F_{warp}^{head}, F_{warp}^{torso}, F_{warp}^{leg}\right\}$ can be obtained by:

$$F_{warp}^{local} = G_{hdcb}(F_{s,w,v}^{local}). \quad (5)$$

Global Fusion Module

Let F_{pose}^{global} denote the global target pose features encoded from the overall target pose P_t, which can provide additional context as to where different parts should be located in the target image. Concatenating the warped image features of different parts F_{warp}^{local} and the global target pose features F_{pose}^{global} together as input, the global fusion module first aggregates these local part features into the preliminary global fusion features F_{fusion}:

$$F_{fusion} = G_{fusion}\left(F_{warp}^{local}, F_{pose}^{global}\right). \quad (6)$$

Due to the symmetry of human body, there can also exist important semantic correlations for the features of different human parts with long distances. We therefore design a lightweight yet effective non-local component named *pyramid non-local block* which incorporates the multi-scale pyramid pooling with the standard non-local operation to capture such long-range semantic correlations across different human part regions under different scales. Specifically, as shown in Figure 4, given the preliminary global fusion features F_{fusion}, we first use the multi-scale pyramid pooling to adaptively divide them into different part regions and select the most significant global representation

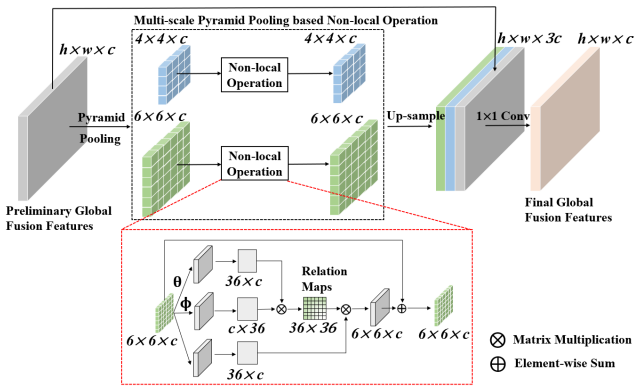

Figure 4: The global fusion module. It aggregates the warped features of different parts into the global fusion features and captures the non-local semantic correlations among different human parts.

for each region, producing hierarchical features with different sizes (e.g., $4 \times 4, 6 \times 6$) in parallel. Next, we apply the standard non-local operations on the pooled features at different scales respectively to obtain the response at a target position by the weighted summation of features from all positions, where the weights are the pairwise relation values recorded in the generated relation maps (which are visualized in our experiments). Specifically, given the input features x, the relation maps R are calculated by $R = softmax(\theta(x)^T \phi(x))$, where $\theta(\cdot)$ and $\phi(\cdot)$ are two feature embeddings implemented as 1×1 convolutions. Let G_{pnb} denote the *pyramid non-local block*. The final global features F_{global} are obtained via:

$$F_{global} = G_{pnb}(F_{fusion}). \quad (7)$$

Finally, the target person image $\hat{I}_t$ is generated from the global features F_{global} using a decoder network Dec which contains a set of deconvolutional layers:

$$\hat{I}_t = Dec(F_{global}). \quad (8)$$

Training

We train our model in two stages. First, without the ground truth of appearance flow fields and visibility maps, we train the part-based flow generation module in an unsupervised manner using the sampling correctness loss (Ren et al. 2019, 2020). Since our part-based flow generation module contains three sub-models corresponding to different parts, we train them together using the overall loss defined as:

$$L_{sam} = L_{sam}^{head} + L_{sam}^{torso} + L_{sam}^{leg}, \quad (9)$$

where L_{sam}^{head},L_{sam}^{torso}, and L_{sam}^{leg} denote the sampling correctness loss for each part respectively. The sampling correctness loss constrains the appearance flow fields to sample positions with similar semantics via measuring the similarity between the warped source features and ground truth target features. Refer to the supplementary material for details.

Then, with the pre-trained part-based flow generation module, we train our whole model in an end-to-end way.

Model	Market-1501							DeepFashion			
	FID↓	LPIPS↓	Mask-LPIPS↓	SSIM↑	Mask-SSIM↑	PSNR↑	Mask-PSNR↑	FID↓	LPIPS↓	SSIM↑	PSNR↑
VU-Net	24.386	0.3211	0.1747	0.242	<u>0.801</u>	13.664	19.102	13.836	0.2637	<u>0.745</u>	16.255
Def-GAN	29.035	0.2994	0.1496	0.276	0.793	<u>14.391</u>	20.425	26.283	<u>0.2330</u>	**0.747**	<u>17.524</u>
PATN	24.917	0.3196	0.1590	<u>0.282</u>	0.799	14.241	<u>20.482</u>	20.399	0.2533	0.671	16.621
DIST	**21.539**	<u>0.2817</u>	<u>0.1482</u>	0.281	0.796	14.337	20.421	**7.629**	0.2341	0.714	17.445
Ours	<u>24.254</u>	**0.2796**	**0.1464**	**0.290**	**0.802**	**14.526**	**20.726**	<u>8.755</u>	**0.1815**	0.726	**18.030**

Table 1: Quantitative comparison with state-of-the-art methods on the Market-1501 and DeepFashion datasets. The first and second best results are bolded and underlined respectively.

The full loss function is defined as:

$$L = \lambda_1 L_{sam} + \lambda_2 L_{rec} + \lambda_3 L_{adv} + \lambda_4 L_{per} + \lambda_5 L_{sty}, \quad (10)$$

where L_{rec} denotes the reconstruction loss which is formulated as the L1 distance between the generated target image $\hat{I}_t$ and ground truth target image I_t,

$$L_{rec} = \left\| I_t - \hat{I}_t \right\|_1 . \quad (11)$$

L_{adv} represents the adversarial loss (Goodfellow et al. 2014) which uses the discriminator D to promote the generator G to synthesize the target image with sharp details,

$$L_{adv} = \mathbb{E}\left[log(1 - D(G(I_s, P_s, P_t)))\right] + \mathbb{E}\left[logD(I_t)\right] . \quad (12)$$

L_{per} denotes the perceptual loss (Johnson, Alahi, and Fei-Fei 2016) formulated as the L1 distance between features extracted from special layers of a pre-trained VGG network,

$$L_{per} = \sum_i \left\| \phi_i(I_t) - \phi_i(\hat{I}_t) \right\|_1 , \quad (13)$$

where ϕ_i is the feature maps of the i-th layer of the VGG network pre-trained on ImageNet (Russakovsky et al. 2015).

L_{sty} denotes the style loss (Johnson, Alahi, and Fei-Fei 2016) which uses the Gram matrix of features to calculate the style similarity between the images,

$$L_{sty} = \sum_j \left\| G_j^{\phi}(I_t) - G_j^{\phi}(\hat{I}_t) \right\|_1 , \quad (14)$$

where G_j^{ϕ} is the Gram matrix constructed from features ϕ_j.

Implementation Details. Our model is implemented in the PyTorch framework using one NVIDIA GTX 1080Ti GPU with 11GB memory. We adopt the Adam optimizer ($\beta_1 = 0, \beta_2 = 0.99$) (Kingma and Ba 2014) to train our model and the learning rate is fixed to 0.001 in all experiments. For the Market-1501 dataset (Zheng et al. 2015), we train our model using the images with resolution of 128×64, and the batch size is set to 8. For the DeepFashion dataset (Liu et al. 2016), our model is trained using the images with resolution of 256×256, and the batch size is 6.

Experiment

In this section, we perform extensive experiments to demonstrate the superiority of the proposed method over state-of-the-art methods. Furthermore, we conduct the ablation study to verify the contribution of each component in our model.

Datasets. We conduct our experiments on the ReID dataset Market-1501 (Zheng et al. 2015) and the In-shop Clothes Retrieval Benchmark DeepFashion (Liu et al. 2016). The Market-1501 dataset contains 32,668 low-resolution images (128×64) which vary enormously in the pose, background, and illumination. Meanwhile, the DeepFashion dataset contains 52,712 person images (256×256) with various appearances and poses. For a fair comparison, we split the two datasets following the same setting in (Ren et al. 2020). Consequently, we pick 263,632 training pairs and 12,000 testing pairs for the Market-1501 dataset. For the DeepFashion dataset, we randomly select 101,966 pairs for training and 8,570 pairs for testing.

Metrics. It remains an open problem to evaluate the quality of generated images reasonably. Following the previous works (Siarohin et al. 2018; Zhu et al. 2019; Ren et al. 2020), we use the common metrics such as Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al. 2018), Fréchet Inception Distance (FID) (Heusel et al. 2017), Structural Similarity (SSIM) (Wang et al. 2004), and Peak Signal-to-noise Ratio (PSNR) to assess the quality of generated images quantitatively. Specifically, both LPIPS and FID calculate the perceptual distance between the generated images and ground truth images in the feature space w.r.t. each pair of samples and global distribution, respectively. Meanwhile, SSIM and PSNR indicate the similarity between paired images in raw pixel space. For the Market-1501 dataset, we further calculate the masked results of these metrics to exclude the interference of the backgrounds. Furthermore, considering that these quantitative metrics may not fully reflect the image quality (Ma et al. 2017), we perform a user study to qualitatively evaluate the quality of generated images.

Comparison with State-of-the-art Methods

Quantitative Comparison. As shown in Table 1, we compare our method with four state-of-the-art methods including VU-Net (Esser, Sutter, and Ommer 2018), Def-GAN (Siarohin et al. 2018), PATN (Zhu et al. 2019), and DIST (Ren et al. 2020) on the Market-1501 and DeepFashion datasets. Specifically, we download the pre-trained models of state-of-the-art methods and evaluate their performance on the testing set directly. As we can see, our method outperforms the state-of-the-art methods in most metrics on both datasets, demonstrating the superiority of our model in generating high-quality person images.

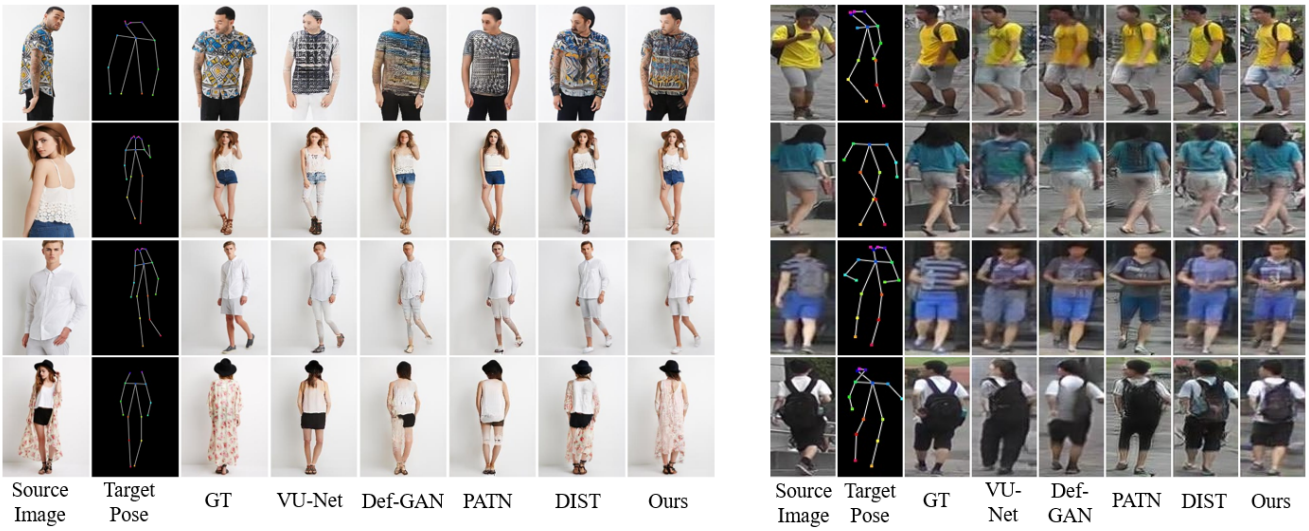

Figure 5: Qualitative comparison with state-of-the-art methods on the DeepFashion(left) and Market-1501(right) datasets.

Qualitative Comparison. Figure 5 shows the qualitative comparison of different methods on the two datasets. All the results of state-of-the-art methods are obtained by directly running their pre-trained models released by authors. As we can see, for the challenging cases with large pose discrepancies (e.g., the first two rows on the left of Figure 5), the existing methods may produce results with heavy artifacts and appearance distortion. In contrast, for the DeepFashion dataset (Liu et al. 2016), our model can generate realistic images in arbitrary target poses, which not only reconstructs the reasonable and consistent global appearances, but preserves the vivid local details such as the textures of clothes and hat. Especially, our model is able to produce more suitable appearance contents for target regions which are invisible in the source image such as the legs and backs of clothes (see the last three rows). For the Market-1501 dataset (Zheng et al. 2015), our model yields natural-looking images with sharp appearance details whereas the artifacts and blurs can be observed in the results of other state-of-the-art methods. More results can be found in the supplementary material.

User Study. We perform a user study to judge the realness and preference of the images generated by different methods. For the realness, we recruit 30 participants to judge whether a given image is real or fake within a second. Following the setting of previous work (Ma et al. 2017; Siarohin et al. 2018; Zhu et al. 2019), for each method, 55 real images and 55 generated images are selected and shuffled randomly. Specifically, the first 10 images are used to warm up and the remaining 100 images are used to evaluate. For the preference, in each group of comparison, a source image, a target pose, and 5 result images generated by different methods are displayed to the participants, and the participants are asked to pick the most reasonable one w.r.t. both the source appearance and target pose. We enlist 30 participants to take part in the evaluation and each participant is asked to finish 30 groups of comparisons for each dataset. As shown in Table 2, our method outperforms the state-of-the-art methods in all subjective measurements on the two datasets, especially for the DeepFashion dataset (Liu et al. 2016) with higher resolution, verifying that the images generated by our model are more realistic and faithful.

Model	Market-1501		DeepFashion	
	G2R↑	Prefer↑	G2R↑	Prefer↑
VU-Net	- -	11.44	- -	1.00
Def-GAN	41.03	10.00	5.23	1.44
PATN	38.03	14.00	10.93	2.22
DIST	47.37	23.11	38.30	28.89
Ours	**50.00**	**41.45**	**43.83**	**66.45**

Table 2: User study(%). "G2R" means the percentage of generated images rated as real w.r.t. all generated images. "Prefer" denotes the user preference for the most realistic result among different methods.

Ablation Study

We further perform the ablation study to analyze the contribution of each technical component in our method. We first introduce the variants implemented by alternatively removing a corresponding component from our full model.

w/o the part-based decomposition (w/o Part). This model removes the part-based decomposition in our flow generation module, and directly estimates the whole flow field of human body to warp the global source image features.

w/o the hybrid dilated convolution block (w/o HDCB). This model removes the *hybrid dilated convolution block* in our local warping module, and directly uses the selected part features to conduct the subsequent feature fusion.

w/o the pyramid non-local block (w/o PNB). This model removes the *pyramid non-local block* in our global fusion module, and simply takes the preliminary global fusion features as input to generate the final target images.

Full. This represents our full model.

Table 3 shows the quantitative results of ablation study on the DeepFashion dataset (Liu et al. 2016). We can see that, our full model achieves the best performance on all evaluation metrics except SSIM, and the removal of any components will degrade the performance of the model.

Model	DeepFashion			
	FID↓	LPIPS↓	SSIM↑	PSNR↑
w/o Part	13.736	0.2090	0.716	17.420
w/o PNB	9.302	0.1832	0.728	17.945
w/o HDCB	9.326	0.1829	**0.729**	18.021
Full	**8.755**	**0.1815**	0.726	**18.030**

Table 3: The quantitative results of ablation study on the DeepFashion dataset. The best results are bolded.

Qualitative comparison of different ablation models is demonstrated in Figure 6. We can see that, although the models w/o Part, w/o PNB, and w/o HDCB can generate target images with correct poses, they can't preserve the human appearances in source images very well. Specifically, there exists heavy appearance distortion on the results produced by the model w/o Part, because of the difficulty in directly learning the overall flow fields of human body under large pose discrepancies. The results generated by the model w/o PNB often suffer from the inconsistency in global human appearance since it doesn't explicitly consider the long-range semantic correlations across different human parts. Besides, the images produced by the model w/o HDCB may lose some local appearance details because it can't fully capture the short-range semantic correlations of local neighbors within a certain part. In contrast, our full model can reconstruct the most realistic images which not only possess consistent global appearance, but maintain vivid local details.

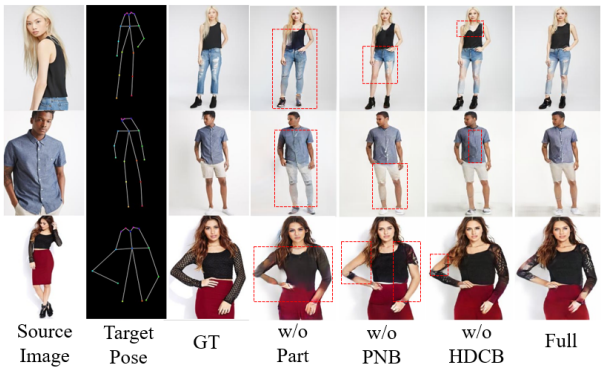

Figure 6: The qualitative comparison of ablation study.

Visualization of The Relation Map

To illustrate the effectiveness of our *pyramid non-local block* in capturing the global semantic correlations among different human parts, in Figure 7 we visualize the generated relation map (e.g., size of 6×6), which represents the relation values of all patches w.r.t a certain target patch. As we can see, for a target patch in a certain image region (e.g., shirt, pants, background), the patches with similar semantics usually have larger relation values w.r.t. this target patch, indicating that our *pyramid non-local block* can capture the non-local semantic correlations among different part regions effectively.

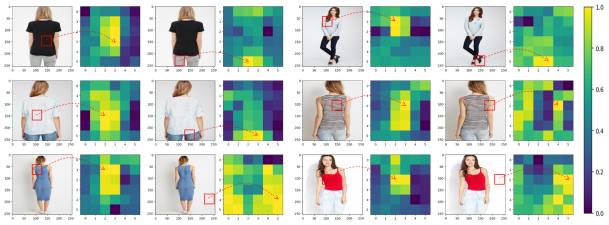

Figure 7: Visualization of the relation map w.r.t. a certain target patch marked by a red rectangle in the image.

Person Image Generation in Random Poses

As shown in Figure 8, given the same source person image and a set of target poses selected from the testing set randomly, our model is able to generate the target images with both vivid appearances and correct poses , demonstrating the versatility of our model sufficiently.

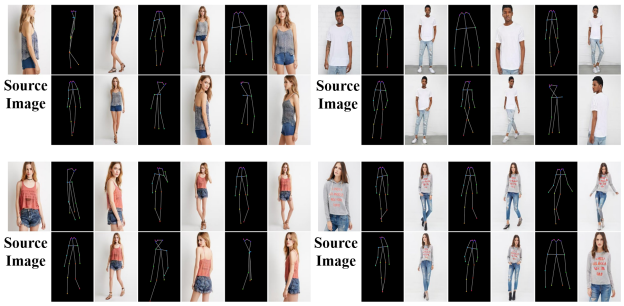

Figure 8: The results of generated person images in random target poses on the DeepFashion dataset.

Conclusion

We present a structure-aware appearance flow based approach to generate realistic person images conditioned on the source appearances and target poses. We decompose the task of learning the overall appearance flow field into learning different local flow fields for different human body parts, which can simplify the learning and model the pose change of each part more precisely. Besides, we carefully design different modules within our framework to capture the local and global semantic correlations of features inside and across human parts respectively. Both qualitative and quantitative results demonstrate the superiority of our proposed method over the state-of-the-art methods. Moreover, the results of ablation study and visualization verify the effectiveness of our designed modules.

Acknowledgments

This work is supported by the National Key R&D Program of China (2018YFB1004300), NSF China (No. 61772462, No. U1736217) and the 100 Talents Program of Zhejiang University.

References

Cao, Z.; Simon, T.; Wei, S.-E.; and Sheikh, Y. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7291–7299.

Chen, L.-C.; Papandreou, G.; Schroff, F.; and Adam, H. 2017. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587* .

Esser, P.; Sutter, E.; and Ommer, B. 2018. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 8857–8866.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Han, X.; Hu, X.; Huang, W.; and Scott, M. R. 2019. Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE International Conference on Computer Vision*, 10471–10480.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* 37(9): 1904–1916.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in neural information processing systems*, 6626–6637.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Johnson, J.; Alahi, A.; and Fei-Fei, L. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

Kim, S.-W.; Kook, H.-K.; Sun, J.-Y.; Kang, M.-C.; and Ko, S.-J. 2018. Parallel feature pyramid network for object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 234–250.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* .

Li, Y.; Huang, C.; and Loy, C. C. 2019. Dense intrinsic appearance flow for human pose transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3693–3702.

Li, Y.; Zhang, X.; and Chen, D. 2018. Csrnet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1091–1100.

Liu, W.; Piao, Z.; Min, J.; Luo, W.; Ma, L.; and Gao, S. 2019. Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision*, 5904–5913.

Liu, Z.; Luo, P.; Qiu, S.; Wang, X.; and Tang, X. 2016. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1096–1104.

Ma, L.; Jia, X.; Sun, Q.; Schiele, B.; Tuytelaars, T.; and Van Gool, L. 2017. Pose guided person image generation. In *Advances in neural information processing systems*, 406–416.

Ma, L.; Sun, Q.; Georgoulis, S.; Van Gool, L.; Schiele, B.; and Fritz, M. 2018. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 99–108.

Men, Y.; Mao, Y.; Jiang, Y.; Ma, W.-Y.; and Lian, Z. 2020. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5084–5093.

Ren, Y.; Yu, X.; Chen, J.; Li, T. H.; and Li, G. 2020. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7690–7699.

Ren, Y.; Yu, X.; Zhang, R.; Li, T. H.; Liu, S.; and Li, G. 2019. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE International Conference on Computer Vision*, 181–190.

Ronneberger, O.; Fischer, P.; and Brox, T. 2015. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, 234–241. Springer.

Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. 2015. Imagenet large scale visual recognition challenge. *International journal of computer vision* 115(3): 211–252.

Siarohin, A.; Sangineto, E.; Lathuiliere, S.; and Sebe, N. 2018. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3408–3416.

Wang, X.; Girshick, R.; Gupta, A.; and He, K. 2018. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 7794–7803.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to

structural similarity. *IEEE transactions on image processing* 13(4): 600–612.

Yu, F.; and Koltun, V. 2015. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122* .

Zhang, R.; Isola, P.; Efros, A. A.; Shechtman, E.; and Wang, O. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 586–595.

Zheng, L.; Shen, L.; Tian, L.; Wang, S.; Wang, J.; and Tian, Q. 2015. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, 1116–1124.

Zhou, T.; Tulsiani, S.; Sun, W.; Malik, J.; and Efros, A. A. 2016. View synthesis by appearance flow. In *European conference on computer vision*, 286–301. Springer.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.

Zhu, Z.; Huang, T.; Shi, B.; Yu, M.; Wang, B.; and Bai, X. 2019. Progressive pose attention transfer for person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2347–2356.