Focus and retain: Complement the Broken Pose in Human Image Synthesis

Pu Ge[†], Qiushi Huang[†], Wei Xiang[‡], Xue Jing, Yule Li, Yiyong Li, Zhun Sun
Bigo Technology PTE. LTD, Singapore
{gepu,huangqiushi,xiangwei1}@bigo.sg

Abstract

Given a target pose, how to generate an image of a specific style with that target pose remains an ill-posed and thus complicated problem. Most recent works treat the human pose synthesis tasks as an image spatial transformation problem using flow warping techniques. However, we observe that, due to the inherent ill-posed nature of many complicated human poses, former methods fail to generate body parts. To tackle this problem, we propose a feature-level flow attention module and an Enhancer Network. The flow attention module produces a flow attention mask to guide the combination of the flow-warped features and the structural pose features. Then, we apply the Enhancer Network to refine the coarse image by injecting the pose information. We present our experimental evaluation both qualitatively and quantitatively on DeepFashion, Market-1501, and Youtube dance datasets. Quantitative results show that our method has 12.995 FID at DeepFashion, 25.459 FID at Market-1501, 14.516 FID at Youtube dance datasets, which outperforms some state-of-the-arts including Guide-Pixe2Pixe, Global-Flow-Local-Attn, and CocosNet.

1. Introduction

Conditional image generation and synthesis becomes a popular computer vision task recent years [29]. The term "conditional" indicates that the output is restricted according to prior knowledge, *i.e.* inputs that provide the subjects or styles. One of the ordinary conditional image synthesis task is to generate the human image with new perspective, outfit and pose, which has been widely applied to areas such as image editing [10, 50], movie making [2], person re-identification (Re-ID) [21, 34, 43, 48], virtual clothes try-on [4, 16], *etc.*

In this work, we focus on the human pose transfer problem. We receive one or more images of a person as well as a specified human pose as references, and we provide the synthesized a realistic image represents the person with the

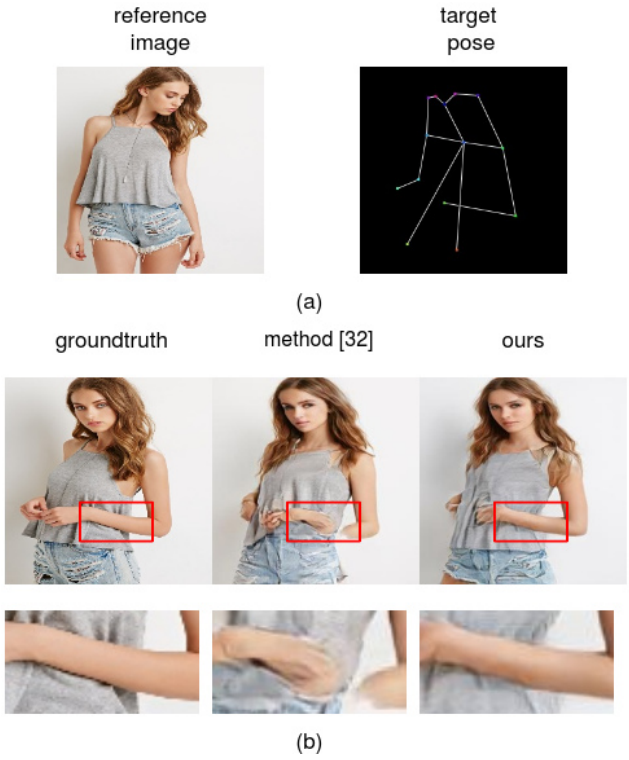

Figure 1: (a) The inputs of the task: a reference image with a target pose. (b) The outputs of the task: a generated human in the target pose. From left to right: the ground truth, results obtained using [35], results obtained using our proposed method. The overlying parts are enlarged below.

referred pose. Researchers have proposed many approaches to tackle this problem recently [25, 38, 51, 35, 45, 20]. The basic idea behind these works is to firstly provide a coarse pose transition image, then employ a Generative Adversarial Network (GAN) based generator to obtain the realistic result. The coarse pose transition is achieved with many different approaches, for instance, difference map guidance [25], feature maps deformation [38], features attention [51] and optical flow [35, 45, 20]. These approaches address the most essential part of this task, which is to preserve texture details

[†]Equal contribution, [‡] Corresponding author.

from the reference image, thus the person synthesized would appear to be the same.

Although these works can produce synthesized images with great overall qualities, their reality often suffer when the reference poses are "complicated", for instance, the arms are placed in front of one's body or the legs are crossed over, Figure 1 provides us an intuitive result. The reason for the misshapen limbs is two-fold: firstly, the difficulty of estimating how would a new pose looks like differs according to the correspondence between the referred and the target pose. Second, the deep neural network systems contain several modules with hundreds of layers, and the information about the relationship between body parts, *a.k.a* the pose, is diluted. As a result, the system is more likely to pay attention to the appearance rather than the pose itself, and it might not be that necessary for a hand to be connected to the wrist, to let the GAN discriminator consider it is realistic.

In this paper, we aim to fix the aforementioned *broken limbs* problem, while keeping the overall synthesis quality. We follow the optical flow based methods [35, 45, 20] as the baseline, and propose two modules to drive the whole system focus more on the completeness and the reality of the human body, namely, the Flow-Attention Network and the Pose-Enhancer Network.

In general, the optical flow based methods predict the spatial relativity between the pose obtained from the reference image and the target pose. Thus the networks will get rough guidance, that from which patch of the reference image they could obtain the correct appearance for the generation of the new image (often by a warping). However, such guidance fails once the reference image does not contain perfect information for the warping, resulting in regions with high ambiguity, which might be synthesized as the broken limbs. To address this problem, we employ the attention mechanism [42, 40, 44] to produce a mask of the fine-grained locations of body parts. By this, we instruct the network generate the ambiguous regions, *i.e.* regions without corresponding appearances in the reference image, using the structure information of the human body.

On the other hand, the aforementioned optical flow-based methods utilize the target pose only for the coarse pose transition, and the term "realistic" is guaranteed mostly on the performance of the GAN module. However, it is difficult for a typical GAN modeling long-range dependencies such as the complicated structured pose information, the GAN loss is mostly satisfied by the generation of high-resolution appearances details that are spatially independent [44]. To address this *information dilution* problem, we propose the Enhancer Network to further emphasize the information about the structured pose information. The Enhancer Network receives both the output of the previous GAN generator and the connected joints map of the human pose as inputs and produces the refined result. We build the Enhancer Network with stacked SPADE blocks [33], such that we can "inject" the pose information while retaining the synthesized semantic information from the former generator.

Our main contributions are summarized as follows:

1. We propose a novel system to improve the *broken limb* problem that is observed in the previous pose transfer studies. The system contains two extra modules, the Flow Attention Module and the Enhancer Network, which could provide and preserve the structured pose information through the whole system.

2. We conduct the proposed method on three different datasets, and also propose a new metric for the completeness of the body. We demonstrate experimental results that are qualitatively and quantitatively superior compared to former state-of-the-art methods.

2. Related Works

2.1. Human Pose Generation

We formulate the pose transfer task as a conditional generative problem, using the conditional generative adversarial network(CGAN) [29] framework. Former researches apply U-net architecture [37] network with skip-connections to generate the target image straightforwardly [25, 26]. However, this simple approach causes feature misalignment, and the key to human pose generation becomes finding an accuracy mapping from reference pose to target pose. There also exist works that utilize the body priors to directly warp the reference to the target pose by a spatial transformer or deformable skip connections in the pixel level [14, 38]. Later, the flow-based methods are widely used in the human pose generation to transform the deformation of reference to the target pose [9, 20, 22, 35, 41]. However, flow-based methods struggle to learn the correct mapping between the source and target due to body self-occlusion and large motions. Recently, several unsupervised generative frameworks have focused on the disentanglement of object pose and appearance [24, 15, 6]. Inspired by these approaches, existing methods attempt to supplement the region where the flow is lost by the structure information of the human body, for example, the features extracted from the target pose. Li et al. [20] concatenate the flow warped features of reference images and target pose features to generate images. Ren et al. [35] learn an occlusion mask to select between flow warped features and target pose features. However, those methods are hard to handle complicated pose due to a lack of fine-grained guidance in body details.

2.2. Generative Adversarial Networks

A generative adversarial network [8] aims to synthesize realistic images by training a generator and a discriminator

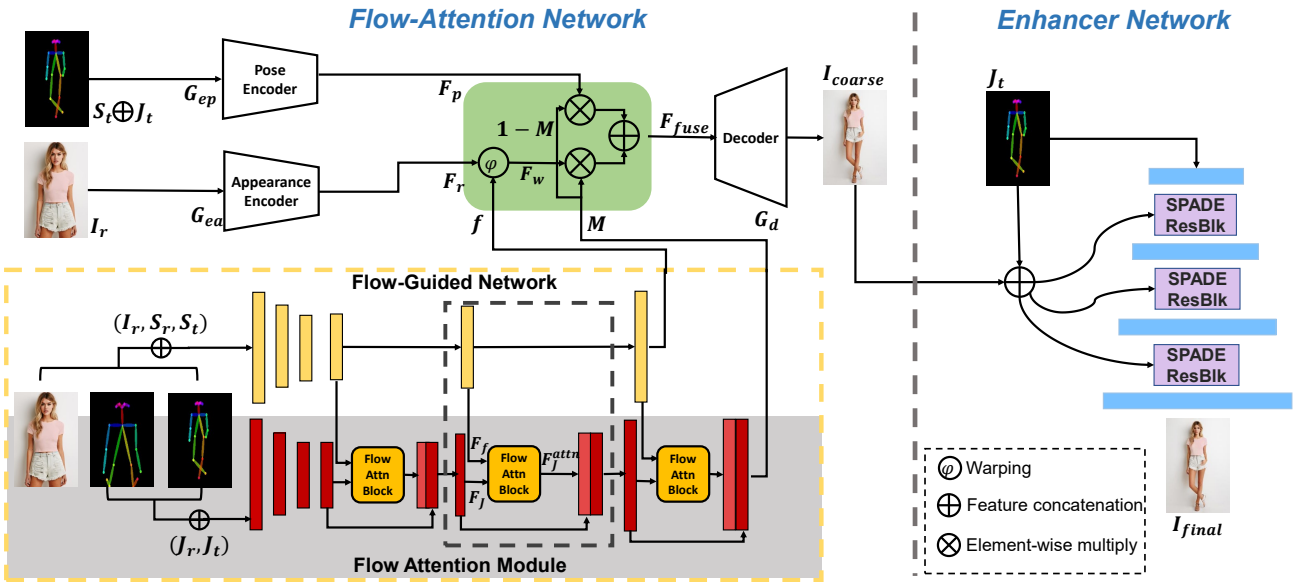

Figure 2: The overview of our system which consists of two networks: Flow-Attention Network and Enhancer Network.

at the same time. The generator tries to produce realistic images to fool the discriminator, and the discriminator tries to distinguish the images generated by the generator from the real images. Some style-based generative adversarial networks [18, 33] borrow the idea of style transfer using adaptive normalization with learned coefficients to synthesize images. Huang et al.[12] propose Adaptive Instance Normalization (AdaIN) to change the style of one image to another by normalization layers. The normalization layers adjust the mean and variance of the features of the content image by new computed mean and variance. It is adopted in many tasks such as StyleGAN [18] that can generate high-quality images. SPADE [33] is another normalization layers called spatially-adaptive normalization. It spatially modulated the activation with learned scale and bias per channel thus can hold the semantic information to generate semantic-aligned images. We use SPADE in the enhancement render module to maintain the body parts in the pose by enhancing the semantic pose information.

2.3. Optical-flow

Optical flow can describe the motion of two objects by showing the pixel displacement from one object to another. The optical flow from two images can be estimated accurately by deep neural network [5, 13, 28]. It is further used in many computer vision tasks such as video frame interpolation [30, 31], multi-view synthesis [49, 32] , image inpainting [36] etc. For the human pose transfer task, optical flow can indicate the motion from the reference pose to the target pose. Therefore, we use the optical flow to estimate the corresponding patches between the reference pose and the target pose similar to many other human pose transfer tasks [9, 20, 22, 35, 41].

3. Our Method

3.1. Overview and Notation

Inspired by the coarse-to-fine strategy [39, 3], we divide this task into two stages. The two corresponding submodules are Flow-Attention Network and Enhancer Network. The overall pose transfer framework is illustrated in Figure 2.

Given the reference image I_r and its corresponding pose S_r, together with the target pose S_t, the flow-guided network predicts the correspondence flow of S_r to S_t denoting as $f[S_r \rightarrow S_t]$ and the flow attention module generates the Flow-Attn-Mask M. Then, guided by the M, the Flow Attention Network makes feature fusion between pose features F_p of S_t and the flow warped features F_w by $f[S_r \rightarrow S_t]$ to generate coarse result I_{coarse}. We design an Enhancer Network to refine the I_{coarse} with the injection of target pose information. In our approach, we employ connected joints map pair (J_r, J_t) of reference pose and target pose as well as pose heat-map [25, 38] pair (S_r, S_t) to represent pose information.

3.2. Flow Attention Network

Overview of Flow Attention Network We first describe the overall architecture of the Flow Attention Network. As shown in Figure 2, an appearance encoder G_{ea} and pose encoder G_{ep} are applied to encode the reference image I_r and the target pose $(S_t \oplus J_t)$ (the concatenation of pose heat-map and body joint map) into the features F_r and F_p.

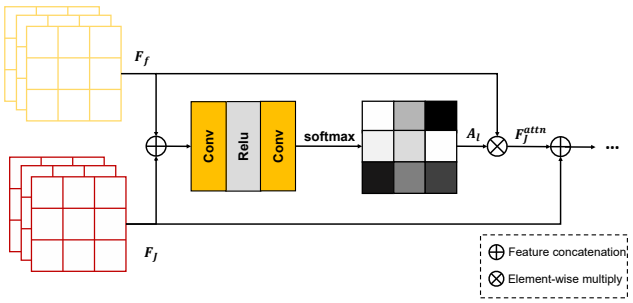

Figure 3: Details of our flow attention block.

The flow-guided network takes reference pose S_r, target pose S_t and reference image I_r as inputs, and outputs flow $f[S_r \to S_t]$. The flow $f[S_r \to S_t]$ represents where the patches of reference image should place in the target image. Then reference features F_r is warped along the flow $f[S_r \to S_t]$ to render the target pose as the local attention block proposed in [35]. However, the mapping relation $f[S_r \to S_t]$ between reference and target pose fails to be found for some body parts such as arms in complicated pose. It means that the model is unable to generate the corresponding limbs only relying on the warped features F_w. The flow attention module is designed to guide the missing region generation based on pose features F_p where the flow $f[S_r \to S_t]$ loses. This module outputs a Flow-Attn-Mask M, which is between 0 to 1 indicating the weight of pose features F_p and the flow warped features F_w. We employ the tuple $(F_w, F_p, f[S_r \to S_t], M)$ for further target image generation where $F_w \in \mathbb{R}^{HW \times C}$, $F_p \in \mathbb{R}^{HW \times C}$, $f[S_r \to S_t] \in \mathbb{R}^{H \times W}$, $M \in \mathbb{R}^{H \times W}$, H,W are feature spatial size and C is the channel-wise dimension. The M is employed to indicate feature fusion based on flow warped features F_w and pose features F_p. The combination strategy is formulated as

$$F_{fuse} = M \otimes \psi(F_r, F_p, f[S_r \to S_t]) + (1 - M) \otimes F_p \quad (1)$$

where $\otimes$ is the element-wise multiplication operation and $\psi(F_r, F_p, f[S_r \to S_t])$ denotes warping F_r along the flow $f[S_r \to S_t]$ to render the target pose by the local attention block described in [35]. Then we feed the fusion features F_{fuse} to a image decoder G_d to synthesize the coarse target image I_{coarse}.

Flow Attention Module The Flow Attention Network generate the target image based on F_w warped by $f[S_r \to S_t]$. However, due to the body self-occlusion or large motions, the guided-flow network may fail to generate accurate $f[S_r \to S_t]$ that maps relations between reference and target pose in some body regions, resulting in losing limbs in generated result. Therefore, we design the flow attention module to generate Flow-Attn-Mask M, which indicate whether the flow warped feature F_w or target pose feature F_p

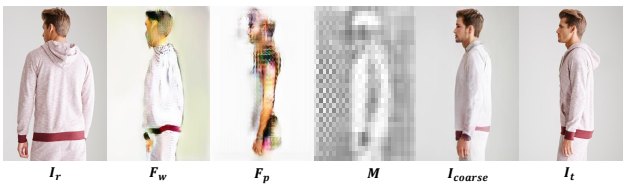

Figure 4: The visualization results of warped features F_w, pose features F_p and Flow-Attn-Mask M.

is responsible for further translation network feature fusion. Thus, we can accurately locate the flow-missing region by M, and then rely on the pose features F_p to generate new contents. The features F_f of flow-guided network roughly implicit the mapping relationship whether the information of a target domain exists in the references. However, coarse information of F_f is hard to indicate each location of body to select features between F_w and F_p. To tackle this, we design a flow attention mechanism to constrain the flow-missing region to specific body location. The flow attention module is a feed-forward CNN aims to predict a Flow-Attn-Mask M. We apply the concatenation of body joint map pair (J_s, J_t) as the input of flow attention module. Then the body joint features F_J are extracted by the encoder. As shown in Figure 3, the flownet features F_f and body joint features F_J are directly concatenated due to their spatial semantic correlation. The flow attention block $\mathcal{A}_{\text{limb}}$ consists of two fully convolutional layers and a softmax layer. We apply the flow attention block to the concatenation of F_J and F_f to obtain flow attention weights $A_l \in \mathbb{R}^{H \times W}$ (H,W are feature spatial size). Then the flow attention flow features F_f^{attn} is calculated as

$$F_f^{attn} = \mathcal{A}_{\text{limb}}(F_f, F_J) \otimes F_f = A_l \otimes F_f. \quad (2)$$

Finally, the F_J and F_f^{attn} are concatenated for next decoder layer as illustrated in Figure 2.

To further analyze the effect of Flow-Attn-Mask M, we visualize the features F_w and F_p by decoding them to images respectively using G_d (shown in Figure 4). The arm of F_w fail to generates due to lack of flow f in arm region. However, Flow-Attn-Mask M correctly indicates this flow missing region (the values of arm in M are close to 0). Thus we can obtain a limb-completeness result with the supplement of pose features F_p.

3.3. Enhancer Network

We obtain the coarse result I_{coarse} by above Flow Attention Network. However, it may obtain broken limbs with some complicated conditions because the structural pose information fades away during the Flow Attention Network generation stage. To address this problem, we present Enhancer Network by injecting the pose information to recover

the body details.

We employ the spatially-adaptive de-normalization (SPADE) block [33] to refine the generated result I_{coarse}. As illustrated in the right part of Figure 2, our Enhancer Network architecture is composed of several ResNet blocks with upsampling layers. The first layer of the Enhancer Network is fed into the target body joint map J_t. The Enhancer Network is injected with the the concatenation of coarse generated result I_{coarse} and the target pose joint map J_t. For each SPADE block, we use interpolation to convert the concatenation to the scale of corresponding feature map. Then the concatenation is encoded by two convolutional layers and outputs α and β to modulate the normalization layer in the Enhancer Network. The SPADE blocks preserve the pose structural information by projecting target body joint map J_t to different scales of layers. Finally, the Enhancer Network generate the refinement image I_{final}.

3.4. Losses

The loss functions we apply to train the total network consist of four parts: the flow loss, the image-level reconstruction loss, the face loss and the adversarial loss.

Flow Loss. The flow $f[S_r \to S_t]$ is constrained both in feature and image level. The flow loss of feature level computes the similarity between the features of I_r warped by flow $f[S_r \to S_t]$ and features of ground truth target image I_t at the VGG [39] feature level [35]. Given the reference image I_r and ground truth target image I_t, we obtain the features $\mathbf{VGG}_{19}(I_r)$ and $\mathbf{VGG}_{19}(I_t)$ of the images encoded by pre-trained VGG-19 model. We maximize the cosine similarity at each coordinate (i, j) in the feature level:

$$L_{fF} = \frac{1}{N} \sum_{i,j} \exp(-T_{(i,j)}) \tag{3}$$

$$T = \mathcal{D}_{\cos}(\mathbf{W}_\theta(\mathbf{VGG}_{19}(I_t), f[S_r \to S_t]), \mathbf{VGG}_{19}(I_t)) \tag{4}$$

where N is the number of feature map positions, $T \in \mathbb{R}^{H \times W}$, $\mathbf{W}_\theta$ denotes feature warping along the flow $f[S_r \to S_t]$, $\mathcal{D}_{\cos}(*)$ is the cosine similarity on the channel-wise dimension and i,j are the index of feature spatial dimension.

We apply ℓ_1 loss to calculate the flow loss in image level, which is is defined as

$$L_{fI} = \mathcal{D}_{\ell_1}(\mathbf{D}_\theta(I_r, f[S_r \to S_t]), I_t) \tag{5}$$

where $\mathbf{D}_\theta(I_r, f[S_r \to S_t])$ denotes warping image I_r along the flow $f[S_r \to S_t]$ to align it with the target image I_t, and $\mathcal{D}_{\ell_1}(*)$ is the ℓ_1 loss. Then the flow loss is formulated as

$$L_{flow} = L_{fI} + L_{fF} \tag{6}$$

Image-Level Reconstruction Loss. We use ℓ_1 loss and perceptual loss [17] to minimize the distance between the generated image and the target image. Since the network generates I_{coarse} and the enhanced image I_{final}, the loss is calculated between I_{coarse}, I_t and I_{final}, I_t:

$$\begin{aligned} L_I = {} & \mathcal{D}_{\ell_1}(I_{coarse}, I_t) + \mathcal{D}_{\ell_1}(I_{final}, I_t) \\ & + \mathcal{D}_{\ell_1}(\mathbf{VGG}_{19}(I_{coarse}), \mathbf{VGG}_{19}(I_t)) \\ & + \mathcal{D}_{\ell_1}(\mathbf{VGG}_{19}(I_{final}), \mathbf{VGG}_{19}(I_t)) \end{aligned} \tag{7}$$

where $\mathrm{VGG}_{19}(*)$ is the features extracted from pre-trained VGG-19 model.

Face Loss. To strengthen the realism of the face generation, we add a separate face loss to the face region. On the basis of the generated full image, we determine the face region according to six face pose keypoints(nose, neck, left ear, right ear, left eye and right eye). With the same as the loss of full image, face loss is composed of ℓ_1 loss and perceptual loss which formulated as

$$\begin{aligned} L_{face} = {} & \mathcal{D}_{\ell_1}(\mathbf{G}_\theta(I_{final}), \mathbf{G}_\theta(I_t)) \\ & + \mathcal{D}_{\ell_1}(\mathbf{VGG}_{19}(\mathbf{G}_\theta(I_{final})), \mathbf{VGG}_{19}(\mathbf{G}_\theta(I_t))) \end{aligned} \tag{8}$$

where $\mathbf{G}_\theta(*)$ is the face region.

Adversarial Loss. The adversarial loss is calculated by the image discriminator $\mathbf{D}$ and the face discriminator $\mathbf{D_f}$, we use the loss format of LS-GAN [27]:

$$\begin{aligned} L_{adv} = {} & (\mathbf{D}(I_{coarse}) - 1)^2 + (\mathbf{D}(I_{final}) - 1)^2 \\ & + (\mathbf{D_f}(\mathbf{G}_\theta(I_{final})) - 1)^2 \end{aligned} \tag{9}$$

Total Loss. The final loss is the weighted sum of those losses as follow:

$$L_{full} = \lambda_f L_{flow} + \lambda_I L_I + \lambda_{face} L_{face} + \lambda_{adv} L_{adv} \tag{10}$$

where λ_f, λ_I, λ_{face} and λ_{adv} denote the weights of corresponding losses.

4. Experimental Results

4.1. Datasets

DeepFashion Dataset. The DeepFashion Dataset [23] contains 52,712 model images and 7,982 personal identities. We split the dataset as [51] and select 101,966 pairs for training and 8,570 pairs for testing. The images in the DeepFashion are high-quality with a clean background and retrieved from models with in-shopping clothes, which are simple in the pose. To evaluate the performance on complicated poses, we select 101 pairs of test datasets in hard pose (legs raise, arms overlap torso, *etc.*) to evaluate the human body completeness of generated images by PRE (introduced in Sec 4.2).

Market-1501 Dataset. The Market-1501 Dataset [47] contains 32,668 images and 1,501 personal identities. The images in the Market-1501 dataset are low-resolution (128×64)

Table 1: Results of the quality of the generated human images using different metrics.

	DeepFashion				Market-1501				Youtube-dance			
	FID	LPIPS	PRE	JND	FID	LPIPS	MLPIPS	JND	FID	LPIPS	PRE	JND
Vid2Vid[41]	16.692	0.256	12.97	4.96%	56.635	0.348	0.187	2.50%	44.803	0.278	5.274	1.22%
GuideP2P[1]	15.658	0.245	12.06	8.04%	50.848	0.433	0.202	2.19%	18.176	0.159	4.745	4.96%
CocosNet[45]	16.493	0.238	11.23	9.48%	36.804	0.304	0.164	10.03%	17.517	0.155	4.562	2.43%
GFLA[35]	13.234	0.224	12.97	23.92%	47.425	0.338	0.202	19.61%	15.898	0.165	4.707	8.10%
Ours	**12.995**	**0.220**	**10.38**	**26.88%**	**25.459**	**0.292**	**0.151**	**24.67%**	**14.516**	**0.148**	**4.228**	**12.91%**

with wild backgrounds. We split the datasets as [51] and select 263,632 training pairs and 12,000 testing pairs.

YouTube-Dance Dataset. We collect dance videos from the YouTube website and extract frames from videos to build the YouTube-Dance Dataset. The YouTube-dance dataset contains 70,000 images and 1,400 personal identities for training. The test dataset is composed of 2,700 images with 1,350 personal identities. We randomly selected 70000 pairs for training and 2,700 pairs for testing. The images in the YouTube-dance dataset vary in terms of the poses. Therefore, we apply it to evaluate the body completeness of generated images. In order to focus on the generation of body parts, we remove the background to eliminate interference. Considering that the current segmentation methods are state-of-the-art, we segment each image to retain human foreground.

4.2. Metrics

Fréchet Inception Distance (FID) [11] and Learned Perceptual Image Patch Similarity (LPIPS) [46] are two commonly-used evaluation metrics in the human pose transfer task in recent years. The FID score compares the statistics of generated samples to real samples, using the activations of the Inception-v3 network. If a generated image contains meaningful objects, the divergences between its activation distribution and the real images' would be small. This distance helps measure the semantic realism of generated image samples. The LPIPS score, on the other hand, calculates the distance of the target image and generated results by features at the perceptual level. If the generated image is similar to the target image, the distance of features would be small. Therefore, we apply this metric to measure the similarity between the generated image and target image in the feature domain. To alleviate the interference of the backgrounds in the Market-1501 Dataset, we also calculate Mask-LPIPS which was applied in previous work [25]. We also conduct a subjective test based on metric Just Noticeable Difference (JND) to evaluate the subjective quality of generate images.

To evaluate the human body completeness of generated images, we design a pose reconstruction error (PRE) to measure it. If an image is well-generated, the pose information obtained by the pose estimation network should be close to the origin pose used for image generation. Therefore, we first obtain the poses of the generated images through AlphaPose [7] and calculate the ℓ_1 loss of pose coordinates between ground truth and the estimation pose of generated results. The PRE metric is defined as

$$\mathrm{PRE} = \frac{1}{K}\sum_{i}|\tilde{p}_i - p_i| \tag{11}$$

Here, $\tilde{p}_i$ and p_i are the i-th keypoint of ground truth pose and the generated image pose, and K is the number of pose key-points.

4.3. Implementation

Our flow-guided translation network is built based on the auto-encoder structures. The basic component of G_{ea}, G_{ep} and G_d is residual block. For the flow-guided network, we utilize dilated Convolutions to enlarge the receptive fields. In the Flow Attention Network, we make features fusion between F_w and F_p by M at resolutions of 32×32 and 64×64. The Enhancer Network consists of four SPADE residual blocks with up-sampling layers. In our experiments, we train the Flow-Attention Network first. Then we freeze the Flow-Attention Network and train the Enhancer Network. We apply the Adam [19] optimizer with learning rate 10^{-4}. The setting of batch size is 8 while training.

4.4. Comparison with Sota Methods

Our proposed approach is compared with several state-of-the-art methods including Guided-Pix2Pix [1], Global-Flow-Local-Attn [35], CocosNet [45] and Few-Shot-Vid2vid [41].

The quantitative comparison results are shown in Table 1. It shows that our proposed approach outperforms others in terms of metric FID and LPIPS, which means our model can generate realistic images and detailed textures. Furthermore, the metric PRE demonstrates that our method can preserve better body completeness. Note that the FID and LPIPS of Few-Shot-Vid2Vid in the Youtube-Dance Dataset are much higher than other methods since our implementation is different from Few-Shot-Vid2Vid. We only use the key points because of the inaccuracy of the UV map in dance frames.

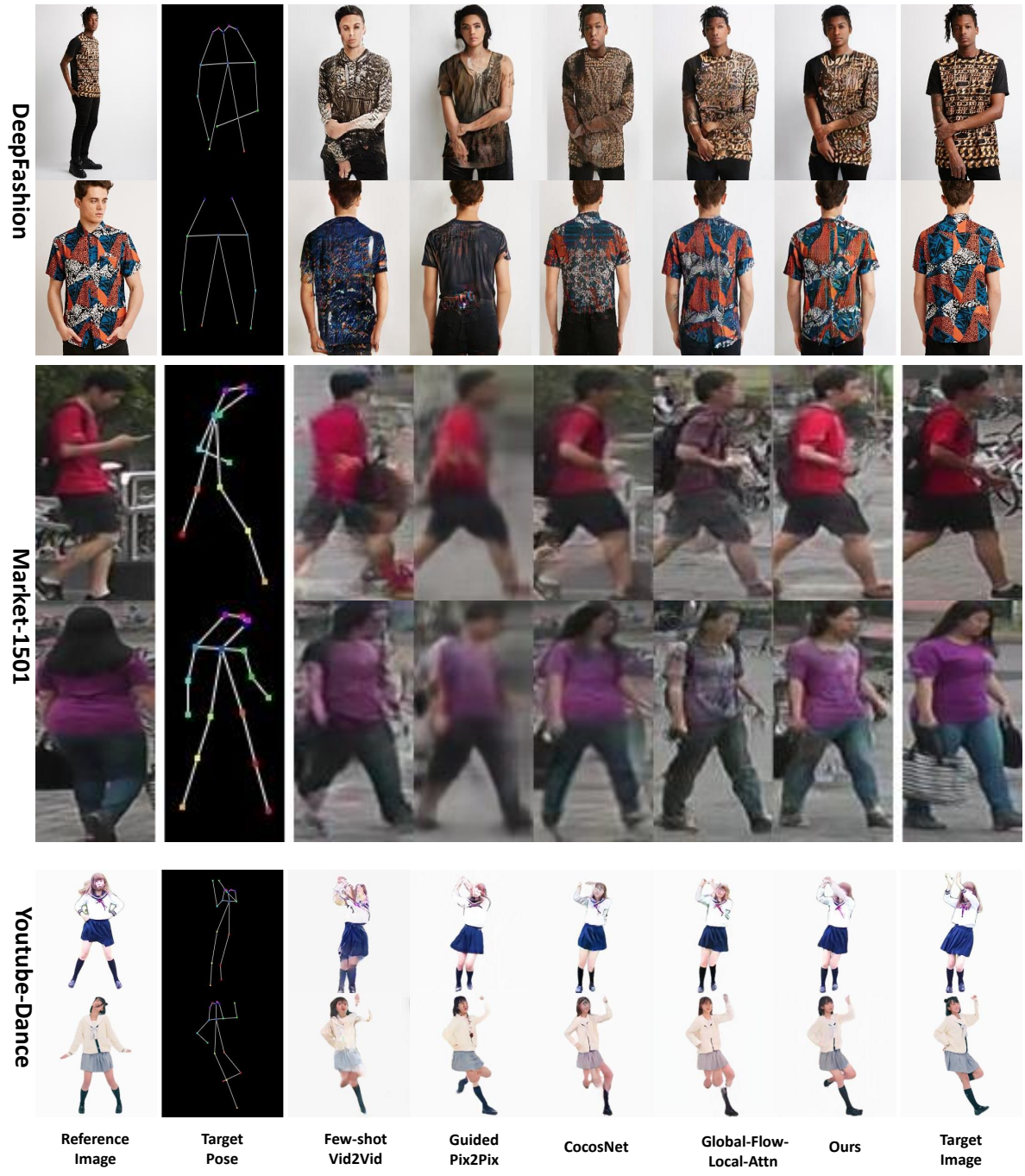

Figure 5: The qualitative results compared with state-of-the-art methods.

In our subjective evaluation, we random select 200 images for each dataset and model. 5 volunteers are required to choose the more realistic image from the data pair of ground-truth and generated images. Table 1 shows that our model achieves the best result in three dataset.

Further visualization comparison results of different methods in DeepFashion, Market-1501, and Youtube-Dance are shown in Figure 5. For the DeepFashion Dataset with high-quality images, Guided-Pix2Pix and Few-Shot-Vid2Vid are insufficient to represent textures or color of the clothes. Meanwhile, the generated arms of comparison methods in row 1 of Figure 5 produce distortions when arms and clothes

Table 2: Quantitative results of the ablation study that excludes different modules from the overall system.

	FID	LPIPS	PRE
w/o Flow Attn Module	16.014	0.166	4.826
w/o Face Loss	15.451	0.151	4.386
w/o Enhancement Network	14.831	0.149	4.302
Full	**14.516**	**0.148**	**4.228**

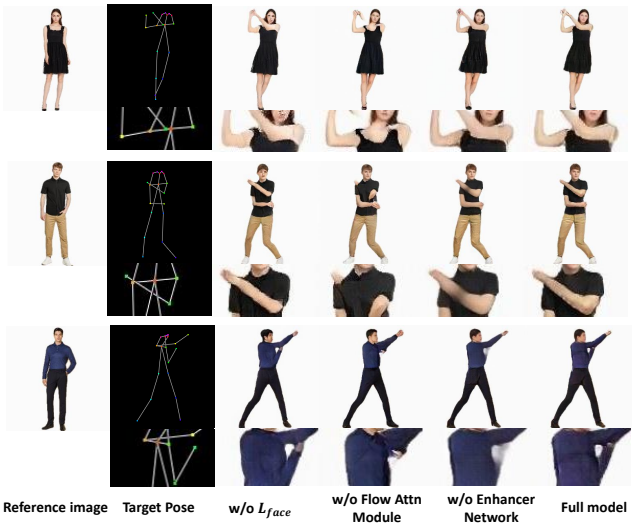

Figure 6: The visualization of ablation results shown in Table 2.

are overlapping. It can be seen that our model can generate complete and realistic arms. Furthermore, as shown in row 2 of Figure 5, our method can preserve the details of clothes even with complex patterns. Compared to the DeepFashion Dataset, the images in the low-resolution Market-1501 Dataset vary in terms of the pose and background. We can observe our method can generate images with the correct pose and fewer artifacts than other competitors. The images in Youtube-Dance Dataset provide complicated poses, and we can use it to evaluate the completeness of the generated body in complicated poses. When arms or legs are raised, it can be seen that our model can generate complete limbs. However, the limbs of the human body in comparison methods fail to generate in these complicated conditions.

4.5. Ablation Study

In this section, we perform ablation studies to further analyze our model. We evaluate the contributions of each component by removing it from the full model. To validate the performance of generating complete limbs in complicated poses, we conduct ablation studies in the test dataset of Youtube-Dance.

The quantitative results of removing each component

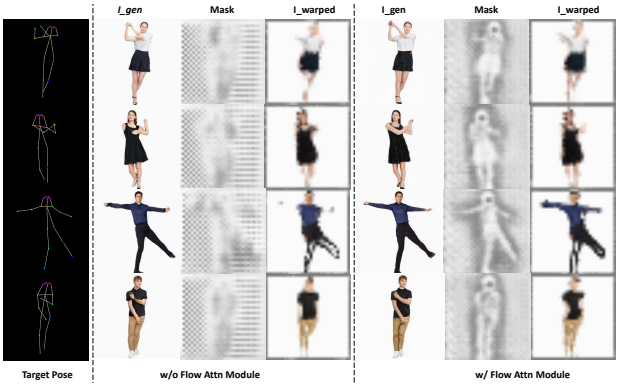

Figure 7: The generated masks and warped images obtained from flow attention module.

are displayed in Table 2. We also show the visualization results of the ablation studies in Figure 6. Column 3 of Figure 6 shows the visualization results of the generated results without face enhancement. With face loss, L_{face}, and discriminator, the face region is focused, which facilitates the face synthesis clearer. We can observe that our model can generate realistic faces even with low-quality reference images of Youtube-Dance.

The flow attention module is further analyzed in Figure 7. We visualize the Flow-Attn-Mask M and the warped results through flow f of reference images I_r. We also show the mask generated only based on flow features. Compared with the mask without a flow attention module, the Flow-Attn-Mask can generate as a human form to indicate each location of the body. It can be seen that the arms fail to be generated when encountering complex dance poses without Flow-Attn-Mask (such as the torso and the arms overlap). When the flow does not find the corresponding relationship between target and reference, the Flow-Attn-Mask M indicates the pose feature F_p to generate that region (the values of M in arms region are close to 0). Therefore, our model can generate complete limbs even with body self-occlusion.

The Enhancer Network is applied to refine some cases which have a color mixture of clothes and body or the incorrect occlusion relationship of clothes and body. with the injection of pose joint map, we further enhance the structural pose information to recover some artifacts.

5. Conclusion

In this study, we aim to complement the generation results of former optical-flow based methods, in which the limbs might be broken under complicated pose conditions. We propose a flow-attention network and an enhancer network to provide the necessary details for generator to synthesize the detailed limbs properly. The experimental results over three different datasets demonstrate the effectiveness of our proposed method.

References

[1] Badour AlBahar and Jia-Bin Huang. Guided image-to-image translation with bi-directional feature transformation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9016–9025, 2019.

[2] Jason Antic. Deoldify.(2019). *GitHub: github. com/jantic/DeOldify*, 2019.

[3] 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.

[4] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1161–1170, 2019.

[5] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766, 2015.

[6] Aysegul Dundar, Kevin J Shih, Animesh Garg, Robert Pottorf, Andrew Tao, and Bryan Catanzaro. Unsupervised disentanglement of pose, appearance and background from images and videos. *arXiv preprint arXiv:2001.09518*, 2020.

[7] Hao-Shu Fang, Shuqin Xie, Yu-Wing Tai, and Cewu Lu. Rmpe: Regional multi-person pose estimation. In *ICCV*, 2017.

[8] 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.

[9] Artur Grigorev, Artem Sevastopolsky, Alexander Vakhitov, and Victor Lempitsky. Coordinate-based texture inpainting for pose-guided human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 12135–12144, 2019.

[10] Zhenliang He, Wangmeng Zuo, Meina Kan, Shiguang Shan, and Xilin Chen. Attgan: Facial attribute editing by only changing what you want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019.

[11] 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.

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

[13] Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.

[14] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al. Spatial transformer networks. In *Advances in neural information processing systems*, pages 2017–2025, 2015.

[15] Tomas Jakab, Ankush Gupta, Hakan Bilen, and Andrea Vedaldi. Unsupervised learning of object landmarks through conditional image generation. In *Advances in neural information processing systems*, pages 4016–4027, 2018.

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

[17] 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.

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

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

[20] 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.

[21] Jinxian Liu, Bingbing Ni, Yichao Yan, Peng Zhou, Shuo Cheng, and Jianguo Hu. Pose transferrable person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4099–4108, 2018.

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

[23] 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.

[24] Dominik Lorenz, Leonard Bereska, Timo Milbich, and Bjorn Ommer. Unsupervised part-based disentangling of object shape and appearance. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10955–10964, 2019.

[25] 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.

[26] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 99–108, 2018.

[27] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802, 2017.

[28] Simon Meister, Junhwa Hur, and Stefan Roth. Unflow: Unsupervised learning of optical flow with a bidirectional census loss. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

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

[30] Simon Niklaus and Feng Liu. Context-aware synthesis for video frame interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1710, 2018.

[31] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive convolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 670–679, 2017.

[32] Eunbyung Park, Jimei Yang, Ersin Yumer, Duygu Ceylan, and Alexander C Berg. Transformation-grounded image generation network for novel 3d view synthesis. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 3500–3509, 2017.

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

[34] Xuelin Qian, Yanwei Fu, Tao Xiang, Wenxuan Wang, Jie Qiu, Yang Wu, Yu-Gang Jiang, and Xiangyang Xue. Pose-normalized image generation for person re-identification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 650–667, 2018.

[35] 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.

[36] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 181–190, 2019.

[37] 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.

[38] 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.

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

[40] 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.

[41] Ting-Chun Wang, Ming-Yu Liu, Andrew Tao, Guilin Liu, Jan Kautz, and Bryan Catanzaro. Few-shot video-to-video synthesis. *arXiv preprint arXiv:1910.12713*, 2019.

[42] 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.

[43] Longhui Wei, Shiliang Zhang, Wen Gao, and Qi Tian. Person transfer gan to bridge domain gap for person re-identification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 79–88, 2018.

[44] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363, 2019.

[45] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5143–5153, 2020.

[46] 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.

[47] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proceedings of the IEEE international conference on computer vision*, pages 1116–1124, 2015.

[48] Zhedong Zheng, Liang Zheng, and Yi Yang. Unlabeled samples generated by gan improve the person re-identification baseline in vitro. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3754–3762, 2017.

[49] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016.

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[51] 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.