MUST-GAN: Multi-level Statistics Transfer for Self-driven Person Image Generation

Tianxiang Ma[1,2], Bo Peng[2,3], Wei Wang[2], Jing Dong[2*]
[1] School of Artificial Intelligence, University of Chinese Academy of Sciences
[2] Center for Research on Intelligent Perception and Computing, CASIA
[3] State Key Laboratory of Information Security, IIE, CAS
tianxiang.ma@cripac.ia.ac.cn, {bo.peng, wwang, jdong}@nlpr.ia.ac.cn

Abstract

Pose-guided person image generation usually involves using paired source-target images to supervise the training, which significantly increases the data preparation effort and limits the application of the models. To deal with this problem, we propose a novel multi-level statistics transfer model, which disentangles and transfers multi-level appearance features from person images and merges them with pose features to reconstruct the source person images themselves. So that the source images can be used as supervision for self-driven person image generation. Specifically, our model extracts multi-level features from the appearance encoder and learns the optimal appearance representation through attention mechanism and attributes statistics. Then we transfer them to a pose-guided generator for re-fusion of appearance and pose. Our approach allows for flexible manipulation of person appearance and pose properties to perform pose transfer and clothes style transfer tasks. Experimental results on the DeepFashion dataset demonstrate our method's superiority compared with state-of-the-art supervised and unsupervised methods. In addition, our approach also performs well in the wild.

1. Introduction

Person image generation has been gaining attention in recent years, which aims to generate the person image as realistic as possible, and at the same time, to transfer the source person image to a target pose. It has great potential for applications, like virtual try-on, clothing texture editing, controllable person manipulation, and so on.

Recently, many researchers have contributed to this topic, with most of the work focusing on pose-guided person image generation [24, 35, 29, 45, 37, 18, 27, 31]. Many person image generation models [24, 35, 45, 18, 27, 31] use paired source-target images for supervised training. However, the paired images require a lot of time and workforce to collect, limiting the usage scenarios of these models. There are also some unsupervised person image generation methods [29, 3, 37], but the quality of their generated images is not fine.

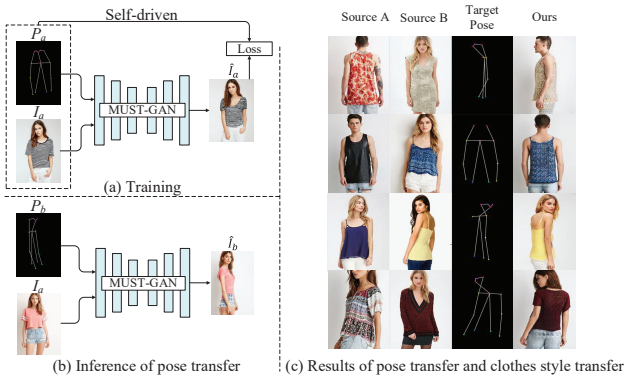

Figure 1. Self-driven person image generation and visualization of pose transfer with clothes style transfer. Our model can be trained in a self-driven way without paired source-target images and flexibly controls the appearance and pose attributes to achieve pose transfer and clothes style transfer in inference. The images in (c) show the generated results using this model for simultaneous pose and cloths style transfer. Source A is transferred to the target pose, and its clothes are replaced by source B's.

In this paper, we propose a self-driven approach for person image generation without using any paired source-target images during training, as shown in Figure 1(a). Specifically, we propose a novel multi-level statistics transfer model, which can disentangle and transfer multi-level appearance features for person images. The source image can be used as supervision information for person image generation without paired source-target training data. Our method allows for flexible manipulation of pose and appear-

* Corresponding author.

ance properties to achieve pose transfer and clothes style transfer, as shown in Figure 1(c).

Since there is no need to pair data during training for our approach, how to extract the person appearance features from input images and transfer them to a pose-guided generator for reconstruction is the key to our proposed method. It is also important to prevent the model from learning trivial solutions that directly copy all input information to generate output image. To deal with these problems, we propose a multi-level statistics transfer model, which extracts the multi-level features from the appearance encoder, and utilizes attention mechanism and attributes statistics to learn optimal representations of appearance features. Finally, the model fuses appearance features into a pose-guided generator to reconstruct the source person image.

Specifically, our method uses an appearance encoder and a pose encoder to extract features from person image and pose image, respectively. Then we introduce the MUST module to obtain multi-level statistics from the appearance encoder and use the channel attention mechanism [8] to learn the weights of each channel in multi-level feature maps. After that, we calculate the statistics of feature maps and apply a multi-layer fully connected network to learn the corresponding relationship when the statistics are transferred to the generator branch. In addition we propose a multi-level statistics matching network for pose-guided generator, which is composed of statistics matching residual blocks with AdaIN [9] and learnable skip connection. This generator module can match the scale and channel number of multi-level statistics and generate realistic person image.

Our method is self-driven by source person images throughout the training process, and no paired source-target images are used. We compare our approach with state-of-the-art supervised and unsupervised methods. Both quantitative and qualitative comparisons prove the superiority of our method. Our model allows for flexible manipulation of person appearance and pose properties to perform pose transfer and clothes style transfer tasks in the inference, and it also shows good performance in the wild.

To summarize, the main contributions of this paper are as follows:

- We propose a fully self-driven person image generation method which requires no paired source-target images for training.
- We propose a multi-level statistics transfer model that can effectively disentangle the rich appearance features from person images and allow for flexible manipulation of person appearance and pose properties.
- Our proposed model performs well compared with the state-of-the-art methods on pose transfer and clothes style transfer and also is tested in the wild for its potential applications.

2. Related Work

2.1. Image Generation

Thanks to the emergence and development of generative adversarial network [5], image generation models have been developing at a high rate in recent years. Some methods [30, 13, 14, 15] use random noise as the input of the network, and the others [28, 11, 44, 39] use conditional information as the input. Currently, the generative models with conditional inputs can generate more specific and controllable images. Pix2pix [11] can generate a street view or object of a particular shape based on different semantic or edge images. CycleGAN [44] implements image translation in an unsupervised manner through cyclic reconstruction. Many methods are devoted to improving the resolution and quality of the generated images. StyleGAN [14] utilizes progressive generation to enhance image resolution and uses AdaIN [9] embedded style code to control the style of the generated images. Pix2pixHD [39] improves the quality of the generated images using a two-stage generation from coarse to fine, and further improves the model using a multi-scale discriminator. What's more, some work focuses on improving the generative model with stacked architecture [42], attention mechanism [41], and latent representation [20, 10]. Besides, some work starts to explore the few-shot learning image translation [21, 33].

2.2. Pose-guided Person Image Generation

Pose-guided person image generation is an important sub-area of image generation that has been continuously developed in recent years. Ma et al. [24] proposed the first method for pose-guided person image generation. Siarohin et al. [35] utilized affine transformation to model the process of pose transfer. Esser et al. [3] employed a variational autoencoder [17] and combined with conditional U-Net [32] to model the shape and appearance of person images. Zhu et al. [45] introduced a progressive pose attention transfer network to build the model structure in the form of residual blocks. Liu et al. [22] warped the input images at the feature level with an additional 3D human model. Li et al. [18] added an optical flow regression module to the usual U-Net-like person image generation model to guide the pose transformation. Han et al. [6] used a flow-based method to transform the source images at the pixel level. Men et al. [27] used person parsing to divide the person image into semantic parts and fed them into a shared parameter encoder to get the style code. Besides, Ren et al. [31] considered both global optical flow and local spatial transformations to generate higher quality images. Tang et al. [38] applied a co-attention mechanism to shape and appearance branches, respectively. All of these methods supervise the train of models and require paired source-target images of person and pose. Some methods began to con-

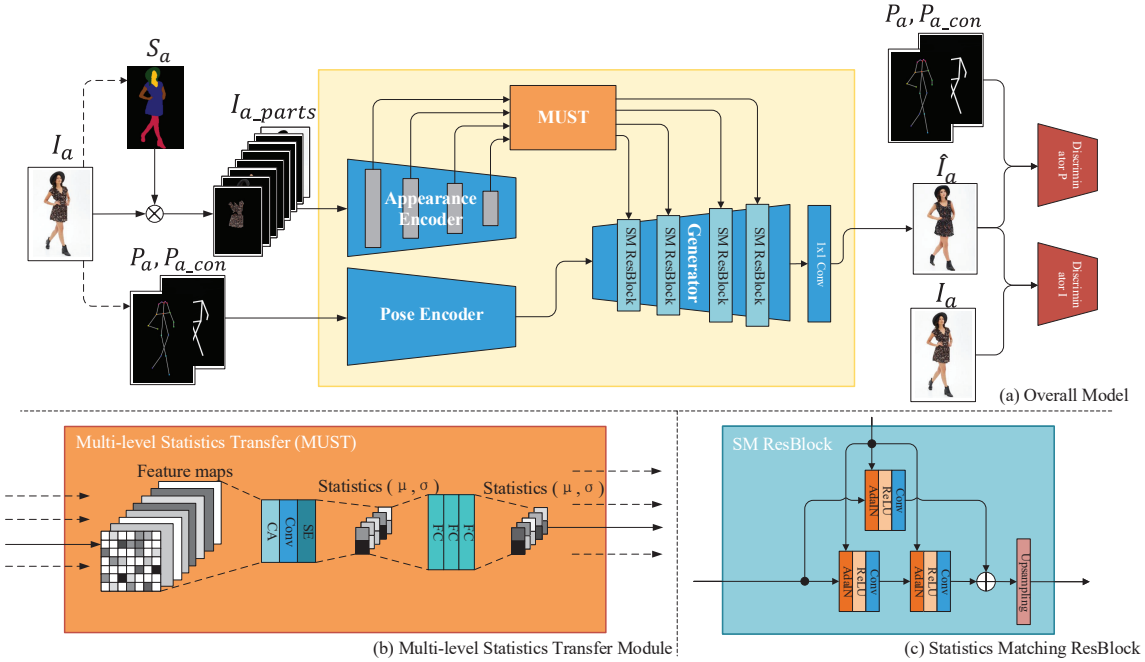

Figure 2. Overview of our MUST-GAN model for self-driven person image generation. Appearance encoder extracts the features of the person image parts I_{a_parts} by semantic segmentation map S_a. Pose encoder encodes the pose image P_a and pose connection map P_{a_con} and guides the Generator to synthesize the source posture. The MUST module disentangles and transfers multi-level appearance features, and the Generator fuses the multi-level appearance features and pose codes for reconstruction of the source person image I_a.

sider unsupervised person image generation. Pumarola et al. [29] first proposed unsupervised person image generation, which uses the CycleGAN to achieve unsupervised pose transfer, but the images generated in this way are not detailed enough. Song et al. [37] also used CycleGAN and broke down the person pose transfer into semantic parsing transformation and appearance generation, to reduce the difficulty of the direct pose transfer. However, both methods require a target pose to guide the transformation. Esser et al. [2] introduced disentangle representations of appearance and pose, which doesn't require pose annotations but needs pairs of images depicting the same object appearance. In this paper, our method can not use any target information, only rely on the input images to supervise the model training, to achieve self-driven person image generation.

3. Approach

In this paper, we want our method to effectively disentangle the person's appearance and pose properties so that the input person images and pose images are independent in the model, and the model only needs to learn the feature representation and feature fusion of the person and pose images. Thus the input images can be arbitrary, i.e., we can use only the source person image and the corresponding pose to realize self-driven person image generation. To achieve this goal, we propose a multi-level statistics transfer model, as shown in Figure 2. It contains four essential parts. Two pathway encoders for person appearance and pose, respectively, a multi-level statistics transfer network (MUST) and a multi-level statistics matching generator network.

Formally, we define I_a, I_b represent person images, and P_a, P_b represent the corresponding postures, where a and b represent source and target poses of the same person wearing the same clothes. Models with paired data supervision can be written as $\hat{I}_b = G(I_a, P_b)$, where G is the generator, $\hat{I}_b$ is the generated person image in the target pose b, and the training set is $\{I_a, P_b, I_b\}$. For the previously unsupervised person image generation models [29, 37], the formula is $\hat{I}_a = G(I_a, P_a, P_b)$, and the training set is $\{I_a, P_a, P_b\}$, where the model needs the target pose P_b as input. However, our proposed model realizes the self-driven person image generation without any target images' information, formalized as $\hat{I}_a = G(I_a, P_a)$, and the training set is $\{I_a, P_a\}$, where all supervision information is derived from the source images.

3.1. Pose Encoder

In the pose pathway, we utilize the trained person pose estimation method [1] to get person pose joints estimate and

construct the pose joints as 18-channel heat maps P_a. There are connections between body joints, such as arms and legs, while person postures represented by heat maps lack such connections. Therefore, we introduce a pose connection map P_{a_con}, including the trunk and limbs, which is spliced with the joint heat maps along the channel dimension, to help the model generate a more accurate structure of the person. For the pose encoder network, we use a down-sampling convolutional neural network with Instance Normalization to encode the joint heat maps and pose connection maps into a high-dimensional space to guide the generator network.

3.2. Appearance Encoder

For the person images pathway, in order to the preliminary disentanglement of the person image, we use the person parsing [4] to obtain the semantic segmentation maps S_a of the input person image. Then we merge the semantic maps into eight classes as in [27] and element-wise multiply them with the person image to obtain I_{a_parts}. It enables complex person appearance to be segmented into several parts to facilitate feature extraction and transfer of the network later. It is also useful for the task of clothes style transfer. The appearance encoder's purpose is to extract rich and robust features at different levels to serve the later MUST module. Therefore, we use the VGG [36] model trained in the COCO dataset [19] as the appearance encoder. The advantage of this is that robust and rich image features can be obtained at the early stage of model training. Because the VGG model is trained on COCO, its generalization is better, and it facilitates the application of our model in the wild.

3.3. Multi-level Statistics Transfer

Our model is trained without using paired source-target images, as we find that when effectively disentangled and transfer person appearance and pose features, the source image itself can provide supervisory information to guide person image generation. We call this self-driven person image generation.

To achieve effective decoupling and transfer of appearance attributes, we propose a multi-level statistics transfer network, as shown in Figure 2(b). We select features extracted from the appearance encoder at four levels from shallow to deep. And first, we utilize the channel attention layer to learn adaptive weights for the features at each level and then reduce the number of channels of the feature maps to a size suitable for the generator network through the convolution layer. After that, we extract the statistics(mean and variance) of each feature map. Because the statistics are the global information, the feature map's structural information is masked, while the statistics can represent the style information. This method enables the disentanglement of appearance and pose in the person image. The use of multi-level features allows the extracted person appearance

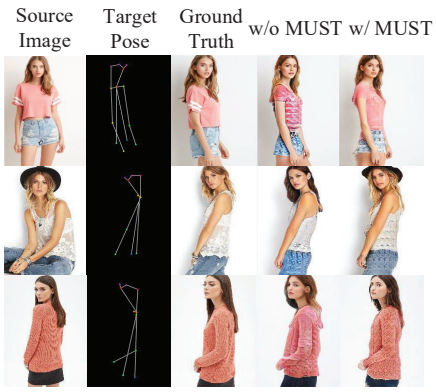

Figure 3. The effect of the MUST module in our model.

attributes to have low-level color and texture information and high-level style information.

The effect of this module is shown in Figure 3. From the results, we can see that using the MUST allows the model to acquire and transfer more accurate appearance attributes. The images generated without the MUST are very different in color and texture from the source image. More specific quantitative comparisons are discussed in the ablation study experiment. The MUST can be simply expressed as

$$s_i = SE\left(Conv_i\left(CA\left(f_i\right)\right)\right), i \in 1, 2, 3, 4, \quad (1)$$

where f_i is the feature maps, CA is the channel attention layer, and SE is the statistics extraction operation. After this, we utilize a multi-layer fully connected network to transform the extracted attributes statistics so that the network learns the mapping of the statistics in the generator. So the complete MUST network can be represented as

$$s_i = Trans\left(SE\left(Conv_i\left(CA\left(f_i\right)\right)\right)\right), i \in 1, 2, 3, 4, \quad (2)$$

where the $Trans$ represents feature transformation of a multi-layer fully connected network.

3.4. Multi-level Statistics Matching Generator

Since we have extracted attributes statistics at multiple levels by MUST, we need to map the statistics to the pose-guided generator and correctly match the features' size and channel at each level. Therefore, we propose a multi-level statistics matching generator network, which is composed of four statistics matching residual blocks, as shown in Figure 2(c). First of all, the attributes statistics parameters obtained from MUST are applied to the generator through the AdaIN [9], which normalizes the network features and adjusts the distribution of feature maps according to the input statistics parameters. Secondly, we use a multi-level residual blocks network as the generator's backbone. And we utilize a learnable skip connection to implement the residual structure even if the numbers of input and output channels are different. The bilinear upsampling layer is used to

increase the size of the feature map gradually. Finally, the person image is reconstructed by 1×1 convolutional layer, which integrates each channel's features.

3.5. Discriminator

There are two commonly used discriminators for pose-guided person image generation methods [29, 45, 27], D_I for person images and D_P for pose images. Similarly, we utilize both discriminators. But for D_P, we add a pose connection map input P_{a_con}, which provides the interrelationship of the pose joints to produce a more accurate posture of the person image. We splice it with the original 18-channel joint heat maps P_a along channel dimension. Both discriminators use residual blocks and downsampling convolutional layers similar to the method [45].

3.6. Loss Functions

The purpose of our model's overall loss functions is to make the generated person image conform to the input person image in appearance and match the input pose image in posture. The specifics are as follows, where I_a and P_a in all formulas represent the input person image and pose image, P_{a_con} and P_a are merged together as P_a, $G(I_a, P_a)$ represents the generated person image.

Adversarial loss. We use the lsgan [26] loss as the adversarial loss. Both discriminator D_I and D_P are used to help the generator G to synthesize realistic person image in a particular pose and keep the input person image's appearance. Adversarial loss is defined as:

$$\begin{aligned}\mathcal{L}_{\text{adv}} =& \mathbb{E}_{I_a,P_a}\left[\log\left(D_I\left(I_a\right)\cdot D_P\left(P_a,I_a\right)\right)\right]+\\ & \mathbb{E}_{I_a,P_t}\left[\log\left(\left(1-D_I\left(G\left(I_a,P_a\right)\right)\right)\right.\right.\\ & \left.\left.\cdot\left(1-D_P\left(P_a,G\left(I_a,P_a\right)\right)\right)\right)\right].\end{aligned} \tag{3}$$

Reconstruction loss. The reconstruction loss aims to make the generated person image match the input person image at the pixel level. Reconstruction loss is calculated using the L1 loss function with the following formula:

$$\mathcal{L}_{\text{rec}} = \left\|G\left(I_a,P_a\right)-I_a\right\|_1. \tag{4}$$

Perceptual loss. Unlike reconstruction loss, the perceptual loss constrains the generated image at a higher feature level. This loss function comes from the style transfer method [12]. Similar to it, we use the trained VGG19 network to extract the features of the generated and input person images, respectively, and compute L1 loss for the features at the specified level l with the following equation,

$$\mathcal{L}_{\text{perc}} = \left\|\phi^l\left(G\left(I_a,P_a\right)\right)-\phi^l\left(I_a\right)\right\|_1. \tag{5}$$

Style loss. To further improve the similarity between the generated person image and the input image in terms of appearance attributes such as texture and color, we use a style loss introduced by method [12]. This loss function calculates the statistic error between activation maps of the input image and the generated image with the Gram matrix. The specific formula is as follows,

$$\mathcal{L}_{\text{style}} = \sum_l \left\|\mathbb{G}\left(\phi^l\left(G\left(I_a,P_a\right)\right)\right)-\mathbb{G}\left(\phi^l\left(I_a\right)\right)\right\|_1, \tag{6}$$

where ϕ^l is the lth activation layer of the trained VGG network, and $\mathbb{G}$ is the Gram matrix.

The overall loss function is,

$$\begin{aligned}\mathcal{L}_{\text{overall}} =& \lambda_{\text{adv}}\mathcal{L}_{\text{adv}}+\lambda_{\text{rec}}\mathcal{L}_{\text{rec}}+\lambda_{\text{perc}}\mathcal{L}_{\text{perc}}\\ &+\lambda_{\text{style}}\mathcal{L}_{style}\end{aligned} \tag{7}$$

where λ_{adv}, λ_{rec}, λ_{perc}, λ_{style} are the weights of the corresponding loss functions.

4. Experiments

4.1. Implementation Details

In this section, we specify the implementation details of our method.

Dataset. We validate our method on the DeepFashion: Inshop Clothes Retrieval Benchmark [23], which contains a large number of model's person images in different poses and different appearances and is widely used in person image generation methods. This dataset includes 52,712 person images with a resolution of 256×256. Since our model can be trained without paired source-target images, the cross pose pairs provided in the dataset are not required. Therefore, we randomly select 37,344 images from the entire dataset as training data. To validate applications such as pose transfer and clothes style transfer during the inference phase, we use the test data pairs of 8570 images constructed in pose transfer model [45]. And there is no overlap between the training set and this test set.

Metrics. Many previous methods of person image generation have used Structure Similarity(SSIM) [40] and Inception Score(IS) [34] as quantitative evaluation metrics for generated images, which were first introduced to the person image generation task by Ma et al. [24]. They are used to evaluate the similarity of the generated person images to Ground Truth and the realism and diversity of the generated images. To further verify the effect of the model in this paper, we also use Fréchet Inception Distance(FID) [7] and Learned Perceptual Image Patch Similarity(LPIPS) [43] metrics to assess the realism and consistency of generated images, which have recently been applied to the evaluation of person image generation models [31].

Network architecture. The model proposed in this paper uses end-to-end training with an Auto-Encoder-like structure. Specifically, the pose encoder uses a downsampling convolutional neural network for encoding pose information. The appearance encoder is a trained VGG network on

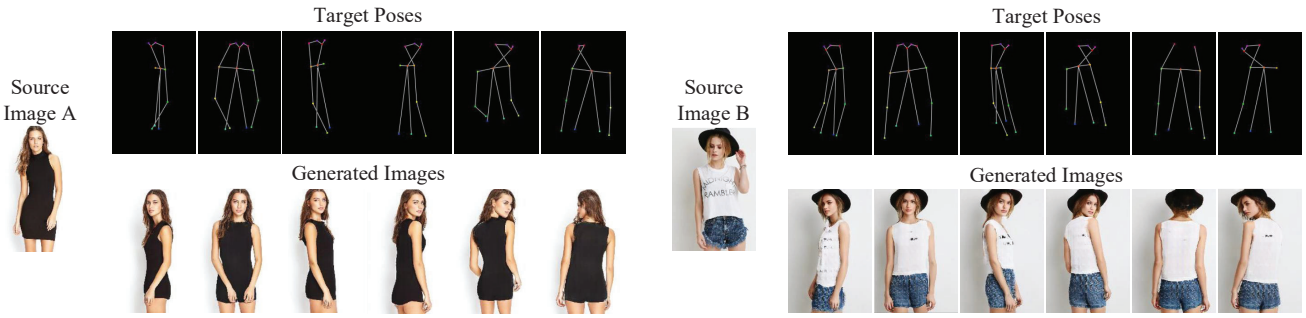

Figure 4. The results of our method in the pose transfer task.

the COCO dataset. The multi-level statistics matching generator utilizes a residual blocks network, bilinear upsampling, and 1×1 convolutional layer to generate a realistic person image. The multi-level statistics transfer network uses the structure, as shown in Figure 2(b).

Training details. Our method is implemented in PyTorch using an NVIDIA TITAN RTX GPU with 24GB memory. Pose estimation uses OpenPose [1], and person parsing utilizes the method [4]. We group the person semantic segmentation maps extracted by [4] into eight parts, i.e., upper clothes, pants, skirt, hair, face, arm, leg, and background same with [27]. We adopt Adam [16] optimizer with the learning rate as 10^{-4} for Generator and 4×10^{-4} for Discriminator (TTUR [7]). The weights for the overall loss function are set to $\lambda_{\text{adv}} = 5$, $\lambda_{\text{rec}} = 1$, $\lambda_{\text{perc}} = 1$, $\lambda_{\text{style}} = 150$ for all experiments.

4.2. Pose Transfer

One of the most important applications of person image generation is the pose transfer. Given the source person image and the target pose, it is required to generate the source person image in the target pose. The results of our model in the pose transfer task are shown in Figure 4. From the results, we can see that our model enables pose transfer well and maintains source images' appearance information. In addition, we compare our method with a number of state-of-the-art methods, including methods [35, 13, 18, 27, 31] with paired source-target images supervision, and unsupervised method [29, 25, 3, 37].

4.2.1 Qualitative Comparison

The results of the qualitative comparison are shown in Figure 5. We compare the generated images of our method with several state-of-the-art approaches, including Def-GAN [35], PATN [13], Intr-Flow [18], ADGAN [27] and GFLA [31]. All results are obtained using source code and the trained model released by the authors. We can see from the figure that the person image generated by our method

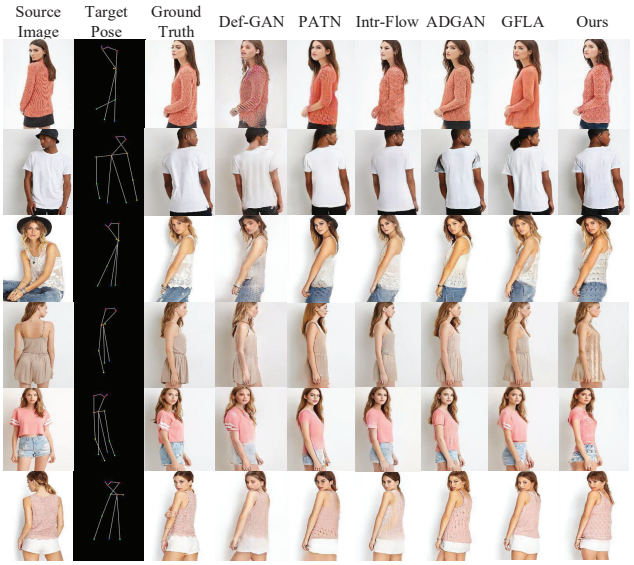

Figure 5. The qualitative comparison with state-of-the-art methods.

is comparable to that of SOTA methods. The preservation of the source person image's appearance properties is also perfect, e.g., in the third line of the figure, our approach is able to maintain the hat feature in the source person image, while other methods do not. It is also worth noting that all the methods compared here require using complete paired source-target images to train the models. In contrast, our method's training process is more demanding because our model is self-driven, relying only on information from the source images to supervise the model. No source-target pairs of person and pose images are used during the entire training process.

4.2.2 Quantitative Comparison

First, we quantitatively compare our method with the unsupervised person image generation methods, as shown in

Figure 6. The results of our method in the clothes style transfer task.

Model	IS↑	SSIM↑	FID↓	LPIPS↓
UPIS [29]	2.971	0.747	-	-
BodyROI7 [25]	3.228	0.614	-	-
VU-Net [3]	3.087	**0.786**	23.583	0.2637
E2E [37]	3.441	0.736	19.248	0.2546
Ours	**3.692**	0.742	**15.902**	**0.2412**

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

Model	IS↑	SSIM ↑	FID↓	LPIPS↓
Def-GAN [35]	3.141	0.741	18.197	0.2330
PATN [13]	3.213	0.770	24.071	0.2533
Intr-Flow [18]	3.251	**0.794**	16.757	**0.2131**
ADGAN [27]	3.329	0.771	18.395	0.2383
GFLA [31]	3.635	0.713	**14.061**	0.2341
Ours	**3.692**	0.742	15.902	0.2412

Table 2. Quantitative comparison with supervised state-of-the-art methods on DeepFashion.

Table1. The results show that our method achieves SOTA on IS, FID, and LPIPS metrics. UPIS cannot calculate the FID and LPIPS metrics because the source code is not available. BodyROI7 uses a different test protocol from other methods and therefore cannot calculate the valid metrics of FID and LPIPS. However, comparing the other two metrics still shows the superiority of our method. What's more, it is worth noting that UPIS and E2E require the target pose as input, while our approach requires no target information and is entirely self-driven training.

To further validate our method's performance, we also compare with the supervised methods, as shown in Table 2. Similarly, our method performs best on the IS metrics. For SSIM and LPIPS scores, all methods are relatively close, and our approach is also at a high level for FID. Although our approach does not score as high as GFLA, which explicitly models the source-to-target transformation process, our method even exceeds most supervised models. And it is important that our method is trained in a self-driven manner without source-target pairs supervision.

4.3. Ablation Study

In this section, we perform an ablation experiment to validate the effect of our proposed MUST module, channel attention mechanism and pose connection map(PCM) on the overall model. For the MUST module's ablation, we remove the MUST module from the model, use the original encoder similar to that in [9], and directly transfer the attributes statistics from the last level of the appearance encoder to the generator network. For the ablation of CA and PCM, we directly remove them from the model. The results of the experiment are shown in Tabel 3. From the table, we can see that the MUST module improves the model under all metrics very much. The channel attention mechanism in the MUST also plays an important role, which enhances the model under SSIM, FID and LPIPS metrics. The PCM has also improved our model on these metrics. The ablation experiments demonstrate the ability of MUST to extract and transfer appearance attributes. And the introduction of channel attention mechanism and pose connection map further enhance the realism and consistency of the images generated by our model under the evaluation metrics in this experiment.

4.4. Clothes Style Transfer

Our proposed method can manipulate the appearance and pose attributes of the person images separately. So the model can achieve clothes style transfer, i.e., the clothes style of the source person changes to that of the target person. The performance results are shown in Figure 6. In this experiment, we transfer the target person's upper clothes style to the source person by replacing the upper clothes part in I_{a_parts} of source image. From the results, we can see that our approach can accurately transfer the colors, textures, and styles in clothes while not changing the source person's identity attribute.

Further to validate the flexibility of the model, we per-

Figure 7. The test results of our method in the wild.

Ours	IS ↑	SSIM ↑	FID ↓	LPIPS ↓
w/o MUST	3.375	0.736	21.928	0.2840
w/o CA	**3.729**	0.737	17.537	0.2450
w/o PCM	3.649	0.724	16.963	0.2554
Full Model	3.692	**0.742**	**15.902**	**0.2412**

Table 3. The evaluation results of ablation study.

form both pose transfer and clothes style transfer at the same time, and the results are shown in Figure 1(c). In the figure, source A's pose is transferred to the target, while its clothes style is replaced by source B's. We can see that the pose transfer and clothes style transfer are both achieved very accurately and realistically. The identity of the generated images remain the same as the source persons, proving that our method can disentangle the appearance and pose well and manipulate them flexibly.

4.5. Testing in The Wild

The person image generation models are trained and tested on the known datasets such as DeepFashion. Still, there are few methods to test the trained models in the wild. In this experiment, we test the performance of our trained model in the wild, and the results are shown in Figure 7. The source images in the figure are obtained from the web. And we use the person parsing method to replace the background of the source images with a clean one so that the complex image background does not affect our approach. As can be seen from the results, our method performs well in maintaining the color and texture of the source person appearance when completing the pose transfer.

5. Conclusion

This paper presents a novel multi-level statistics transfer model to realize a self-driven person image generation. The method tackles the hard problem of disentanglement and transfer of appearance and pose attributes in the absence of paired source-target training data. The extraction and transfer of multi-level statistics enable the low-level color and texture information and high-level style information of person images to be well used. Both qualitative and quantitative comparisons demonstrate the superiority of our method. Our approach allows for flexible manipulation of appearance and pose properties to perform pose transfer and clothes style transfer tasks. Finally, our approach also shows the robustness in the wild, which demonstrates the application potential of our model.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC) under Grants 61772529, 61972395, 61902400, and Beijing Natural Science Foundation under Grant 4192058.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017.

[2] Patrick Esser, Johannes Haux, and Bjorn Ommer. Unsupervised robust disentangling of latent characteristics for image synthesis. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2699–2709, 2019.

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

[4] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017.

[5] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, pages 2672–2680, 2014.

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

[7] 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. *Advances in neural information processing systems*, pages 6626–6637, 2017.

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

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

[10] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 172–189, 2018.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

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

[13] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *International Conference on Learning Representations*, 2018.

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

[15] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.

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

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

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

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. *European conference on computer vision*, pages 740–755, 2014.

[20] Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. *Advances in neural information processing systems*, pages 700–708, 2017.

[21] Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsupervised image-to-image translation. *Proceedings of the IEEE International Conference on Computer Vision*, pages 10551–10560, 2019.

[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. *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. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016.

[24] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. *Advances in neural information processing systems*, pages 406–416, 2017.

[25] 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.

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

[27] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5084–5093, 2020.

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

[29] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8620–8628, 2018.

[30] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[31] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020.

[32] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2015.

[33] Kuniaki Saito, Kate Saenko, and Ming-Yu Liu. Coco-funit: Few-shot unsupervised image translation with a content conditioned style encoder. *arXiv preprint arXiv:2007.07431*, 2020.

[34] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. *Advances in neural information processing systems*, pages 2234–2242, 2016.

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

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

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

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

[39] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[40] 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.

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

[42] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[43] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

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

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