Pose Guided Human Image Synthesis with Partially Decoupled GAN

Jianhan Wu   WUJIANHAN@MAIL.USTC.EDU.CN
Ping An Technology (Shenzhen) Co.,Ltd., University of Science and Technology of China

Jianzong Wang[*]   JZWANG@188.COM
Ping An Technology (Shenzhen) Co., Ltd.

Shijing Si   SHIJING.SI@OUTLOOK.COM
Ping An Technology (Shenzhen) Co., Ltd., School of economics and finance, Shanghai International Studies University

Xiaoyang Qu   QUXIAOY@GMAIL.COM
Ping An Technology (Shenzhen) Co., Ltd.

Jing Xiao   XIAOJING661@PINGAN.COM.CN
Ping An Technology (Shenzhen) Co., Ltd.

Editors: Emtiyaz Khan and Mehmet Gönen

Abstract

Pose Guided Human Image Synthesis (PGHIS) is a challenging task of transforming a human image from the reference pose to a target pose while preserving its style. Most existing methods encode the texture of the whole reference human image into a latent space, and then utilize a decoder to synthesize the image texture of the target pose. However, it is difficult to recover the detailed texture of the whole human image. To alleviate this problem, we propose a method by decoupling the human body into several parts (*e.g.*, hair, face, hands, feet, *etc.*) and then using each of these parts to guide the synthesis of a realistic image of the person, which preserves the detailed information of the generated images. In addition, we design a multi-head attention-based module for PGHIS. Because most convolutional neural network-based methods have difficulty in modeling long-range dependency due to the convolutional operation, the long-range modeling capability of attention mechanism is more suitable than convolutional neural networks for pose transfer task, especially for sharp pose deformation. Extensive experiments on Market-1501 and DeepFashion datasets reveal that our method almost outperforms other existing state-of-the-art methods in terms of both qualitative and quantitative metrics.

Keywords: Partially decoupled, GAN, Pose transfer, Attention, Human image synthesis

1. Introduction

Due to its portability and simplicity, two-dimensional (2D) human synthesis has become more and more popular in the multimedia and computer vision fields. Human image synthesis has huge potential in many commercial applications, such as virtual clothes try-on Ren et al. (2021), person re-identification (Re-ID) Zheng et al. (2015) and so on. A challenging task of 2D human synthesis is pose guided human image synthesis (PGHIS) Zhang et al.

(2022), Ren et al. (2022), which renders the photo-realistic image of arbitrary pose from a reference pose human image. Because of the considerable changes in geometry and texture, rendering adequate details during pose transfer is challenging.

Early image-based Ma et al. (2017) pose transfer methods directly combined human images, pose key points Cao et al. (2017), and the inshop clothing representations together without taking into account the correlation between them, which brings missing and blurred texture. Most of existing methods process the pose and texture separately Tang et al. (2020), Zhang et al. (2021a), Cui et al. (2021), *i.e.*, two different encoders are used to map the reference pose and the whole human texture to the high-dimensional latent space, and then a decoder is used to fuse them to generate the human image of the target pose. This allows for generating human images with complete structure, but usually does not produce very realistic images with detailed information. Because the quality of synthesized images heavily depends on the encoder model, which is inherently challenging due to the difficulty of learning the topologically complex human texture Zhang et al. (2022). Some methods make the encoder learn more details of human texture mainly by increasing the size of the model Ma et al. (2017), Lv et al. (2021), but this improves the performance of detail rendering much less than the increasement of computational cost. Inspired by this, we argue that an encoder will learn more details when it encodes a part of a human body, so we propose to structurally decouple the human image. Specifically, we decouple the human image into 8 parts, then encode these parts using a weight-sharing encoder, and finally fuse them into a texture code to guide the synthesis of a realistic human image. This not only allows the encoder to focus on enough detailed texture information, but also ensures the lightweight nature of our model because of the weight-sharing operation.

In addition, due to the inherent structure of vanilla convolutional neural networks (CNN) Zhao et al. (2022) processing one local neighborhood at a time, they can only focus on short-range information, which results in their inability to achieve efficient spatial transformation Vaswani et al. (2017). And CNN-based methods tend to have poor results when the pose transfer process involves large changes Ren et al. (2020). Attention mechanism Touvron et al. (2021), Zhang et al. (2019) computes the response of a target position as a weighted sum of all reference features, which let attention-based methods have the ability of long-range modeling. But in the pose transfer task, most of the existing methods either learn the correlation of the feature from different poses based on CNN or copy vanilla self-attention directly into image features. Few methods design a transformer module specifically for the pose transfer task. To fill this gap, we designed an efficient attention-based module (called transformer module) for pose transfer task. Specifically, the transformer module consists of two attention modules, the first one is the multi-headed self-attention module, which is responsible for capturing the correlation between the reference pose and the texture. The second is the multi-headed cross-attention module, which is used to migrate the reference texture features to the target pose. It is worthy that we introduce a Residual Fast Fourier Transform (Res FFT) Mao et al. (2021) block instead of a traditional residual block into the transformer module, which makes the model focus on more detailed information like time-domain features and frequency-domain features Wu et al. (2022b).

Based on the above ideas, we propose a Partially Decoupled GAN (PD-GAN) for PGHIS. The structure can be seen in Figure 1. Specifically, The body part decoupling branch and the feature texture transformer branch make up the two main branches of our model. The

body part decoupling branch encodes the body parts into high-dimensional latent codes, which are then fused into features F_C that assist in generating realistic texture images. The transformer branch mainly renders the textures of the reference pose to the target pose, and we particularly design a new transformer module for the task. Our contribution can be summarized in the following three points:

- We propose an effective model named PD-GAN for pose transfer task, which consists of a human partial decoupled module and transformer module. The two modules play crucial roles for generating realistic human images.

- To improve the ability of capturing the long-range dependency of pose transfer, we design a transformer module to obtain more global information. Besides, we use the Res FFT block that can capture both time-domain and frequency-domain information to replace the traditional residual block, which brings more realistic human images generated by our method.

- We perform extensive experiments to confirm the effectiveness of our method in comparison with other baseline methods. Additionally, a comprehensive ablation research indicates the role played by each component in the increased efficacy.

2. Related Work

Pose guided human image synthesis. PGHIS was introduced by Ma et al. (2017), which concatenates the reference pose, reference image and target pose together, then uses a coarse-to-fine network to generate the target pose image with the reference texture. This often leads to texture misalignment. Some flow-based methods Tabejamaat et al. (2021), Li et al. (2019) obtain deeper correlations by estimating the correspondence between the flow of the pose and the appearance, which can alleviate the misalignment problem. However, the model size of such methods is too large and the training time is too long, which limits their applications. Some methods use a model framework with two branches controlling image texture and pose separately to solve the reference image and target image texture misalignment problem. For example, Tang et al. (2020) uses two blocks to model appearance and pose separately, Esser et al. (2018) disentangle appearance and pose by combining the Autoencoder Tang et al. (2021) and U-net Ronneberger et al. (2015). In addition, to enhance the realism of the generated images, there are methods introduce parse map to assist the generator. PISE Zhang et al. (2021a) uses a network to generate a parse mapping of the target pose based on the parse mapping of the reference image and the target pose, and then uses this target pose mapping as an input to another network to assist in generating a target pose image with the texture of the reference image. These methods can mitigate the image texture misalignment problem well, but the extracted parse mapping maps from these methods are often unreliable and the pixel-level semantic annotations of datasets are very difficult to obtain, which limits their applications.

All of these methods mentioned above encode the whole human image with an encoder, and there are few papers that decouple the person to focus on the human body parts. Unlike them, we use encoders to learn each part of the human body to improve the model's capacity to focus on details and thus enhancing PGHIS performance.
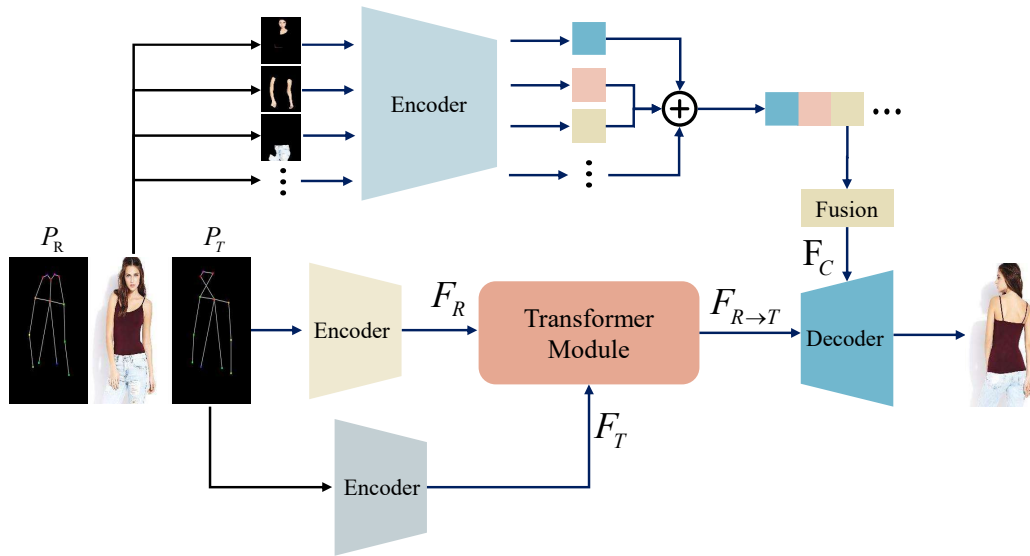
Figure 1: Our generator mainly consists of a transformer branch and a body part decoupling branch. First, the reference pose, target pose, and reference image are concatenated together and mapped into a feature F_R by an encoder, and the target pose is mapped into a feature F_T by another encoder, and then the two features are passed through the transformer module to obtain the coarse feature $F_{R\to T}$ on target pose. The body part decoupling branch uses parser map to divide the person into different attributes and then encodes these parts into a series of style features using a global texture encoder. These style features are fused into a feature F_C by a covolutional layer, and the F_C is injected into the rough features of the target pose by controlling the affine transformation parameters in the AdaIN Huang and Belongie (2017) layer, and are finally decoded to a realistic image of target pose by the decoder.

Transformers in image synthesis. The transformer was first introduced and widely used in NLP because it can compute in parallel and focus on global information. Many researchers have also applied it to the CV field and achieved good results in many tasks, such as image classification Touvron et al. (2021), Wu et al. (2022a), detection Zhu et al. (2021) and generation Hudson and Zitnick (2021). For generation tasks, most transformer-based GAN models are designed for unconditional generation tasks, which are not suitable to conditional generation tasks (*e.g.*, PGHIS tasks). The transformer is inherently superior to the CNN for PSGIS that requires attention to long-range dependencies, and existing methods only use the vanilla attention mechanism for PSGIS Zhu et al. (2019). Inspired by this, we specifically designed the transformer module to model complex topological structure for PSGIS.

3. Method

The framework of our end-to-end generator is depicted in Figure 1, which primarily consists of the transformer module and the body part decoupling module. The transformer module consists of some attention module, which adapts the long-range modeling capability of the transformer to alleviate the difficulties of the recovery of sharp poses. The body part decoupling module serves the decoder with Adaptive instance normalization (AdaIN) Huang and Belongie (2017) as an auxiliary role. Next, we will introduce the detailed design of every component of our model.

3.1. Body Part Decoupling Module

Most existing models process the whole body part through an encoder to obtain the body texture features, and the texture features obtained in this way tend to ignore some detailed information. To deal with this limitation, we introduce body part decoupling, an operation inspired by an intuitive idea: it is more difficult for an encoder to acquire the texture of all parts of the human body at once, so focusing on one small part of the human body at a time will allow the encoder to learn more detailed information. To implement this idea, we introduce the body part decoupling module, which first divides the human body into 8 parts using mask (respectively 'Hair', 'Upper Clothes', 'Dress', 'Pants', 'Face', 'Upper skin', 'leg' and ' background') Men et al. (2020). Multiplying the reference human image with the component mask yields the decoupling human image with component i:

$$I_R^i = I_R \odot M_i \tag{1}$$

Where the $\odot$ and M_i respectively are element-wise product and part masks. These component features are extracted by using weight-sharing encoders, and then concatenate these texture features. Finally the fused texture features are regarded as keys of the transformer module to generate a detailed image of the person. The fusion operation here is a 1×1 convolution operation, the purpose of this operation is to select the appropriate body part texture to further assist in refining the generated human image.

3.2. Transformer Module

Since the vanilla CNN-based transfer module is difficult to handle complex spatial deformations, we have designed a transformer module specifically for action transfer. As shown on the left of Figure 2, the transformer module contains two attention blocks: Multi-head Self-attention (MHSA) and Multi-head Cross-attention (MHCA). These blocks are built upon the Multi-head Attention (MHA), MHA is briefly defined as follows:

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{d_k}) \tag{2}$$

$$head_i = Attention(QW_Q^i, KW_K^i, VW_V^i) \tag{3}$$

$$MHA(Q, K, V) = concat(head_1, head_2, ..., head_t) \tag{4}$$

where Q, K, V denote query, key, value. W_Q^i, W_K^i, W_V^i denote the learnable parameters of query, key and value, t denotes the number of attention heads. From this definition, it can

be seen that the attention mechanism is computed over the whole feature variables, while it is fundamentally different from the local computation of CNN.

In our model, when Q=K=V, the MHA is used as MHSA. When Q=K, it is used as MHCA. Besides, unlike the traditional residual block, we use the Res FFT block here to focus on more detailed information. Because the attention mechanism performs well in focusing on global structural information, but its ability to focus on local detailed information is poor. FFT is a plug-and-play module that we found beneficial in the fields of image denoising and deblurring Mao et al. (2021). The structure of FFT is shown on the right side of Figure 2, the FFT process is as follows:

1. performs Real FFT2d on F to obtain the frequency domain feature, where F $\in \mathbb{R}^{H*W*C}$, H, W, and C represent the height, width, and channel of the F.

2. two convolution operations and an activation function are used to extract the local detail feature information, where the convolution operation and activation function respectively are two 1×1 convolutional layers and an activation layer ReLU.

3. adopts inverse FFT2d operation to convert back to spatial (time-domain) features.

where the Real FFT2d and Inv Real FFT2d operation is the same as the literature Wu et al. (2022b). FFT block achieves attention to image details by combining frequency-domain information as residual connections with time-domain information. In summary, the transformer module can be formalized as follow:

$$F'_R = IN(F_R + FFT(F_R) + MHSA(F_R, F_R, F_R)) \tag{5}$$

$$F''_R = IN(F'_R + MHCA(F_T, F'_R, F'_R)) \tag{6}$$

$$F_{R \to T} = IN(F''_R + MLP(F''_R) + FFT(F''_R)) \tag{7}$$

Where the IN is the operation of instance normalization, MLP denotes Multi-layer Perception. After N (2 in our experiment) time transformer modules, the final output is the feature $F_{R \to T}$.

3.3. Loss Functions

To generate realistic human images with arbitrary poses, several loss functions are employed to constrain the generator and discriminator to update their parameters. Since body parts were introduced to generate more realistic images, we designed two loss functions in our experiments, for the whole images and the partial components, respectively.

Loss functions for the whole images. The loss consists of an L1, an adversarial, a perceptual and a style loss term, which can be formalized as follow:

$$\mathcal{L}_{whole} = \lambda_1 \mathcal{L}_{\ell_1} + \lambda_2 \mathcal{L}_{adv} + \lambda_3 \mathcal{L}_{per} + \lambda_4 \mathcal{L}_{style} \tag{8}$$

where λ_1, λ_2, λ_3 and λ_4 denote the weights of corresponding losses, respectively. The $\mathcal{L}_{\ell_1}$, $\mathcal{L}_{adv}$, $\mathcal{L}_{per}$, $\mathcal{L}_{style}$ can be defined as follow:

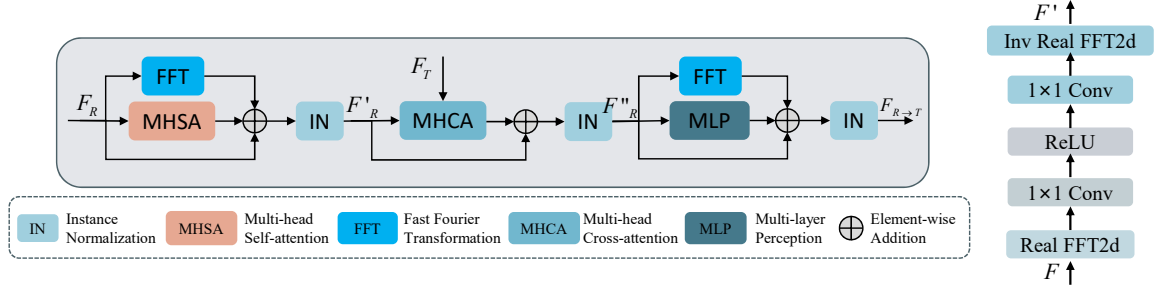$$\mathcal{L}_{\ell_1} = ||I_g - I_t||_1. \tag{9}$$

Figure 2: The structure of transformer module. It consists of two attention-based blocks: MHSA and MHCA. MHSA is responsible for establishing the correlation between pose and texture, and MHCA is responsible for rendering the reference texture into the target pose.

$$\mathcal{L}_{adv} = \mathbb{E}[log(1 - D(I_g))] + \mathbb{E}[logD(I_t)]. \tag{10}$$

$$\mathcal{L}_{per} = \sum_i ||\phi_i(I_t) - \phi_i(I_g)||_1. \tag{11}$$

$$\mathcal{L}_{style} = \sum_j ||G_j^\phi(I_t) - G_j^\phi(I_g)||_1. \tag{12}$$

where I_g, I_t stand the generated images by our model and the ground-truth images, $\mathbb{E}$ denotes the value of expectation, $\phi_i(*)$ denotes the feature of $[relui\_1]$ of the pre-trained VGG-19 model. G_j^ϕ represents the feature of $[relu2\_2, relu3\_4, relu4\_4, relu5\_2]$ of our generator. $D(*)$ represents the discriminator of our model, which is a multi-scale residual discriminator. Each of these loss functions has its role to play: By determining the pixel-level difference between the generated image and the ground truth image, the L1 loss controls the created image's overall effect. Adversarial loss with a discriminator helps the generator synthesize target pose images that resemble the texture of the reference image. Perceptual loss makes them as similar as possible at the deep feature level by mapping the generated image to the ground-truth image with a pre-trained VGG-19 network, which guides generating more ideal images at the human perceptual level. Style loss Zhang et al. (2021b) enables the generated image to be as feature-level same as feasible with the original image. **Loss functions for the partial components.** Since we decouple the human body into multiple parts, aligning the generated images with the ground-truth images at the level of these parts can help generate more detailed human images. The partial loss function can be formalized as:

$$\mathcal{L}_{par} = \sum_j ||\phi_j(C_i(I_t)) - \phi_j(C_i(I_g))||_1. \tag{13}$$

Where ϕ_j denotes the j-th activation map of the pre-trained VGG-19 network, i denotes the i-part (*e.g.*, face) cropping function according to the target poses.

To sum up, the total loss of our image generator is:

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_\ell + \lambda_2\mathcal{L}_{adv} + \lambda_3\mathcal{L}_{per} + \lambda_4\mathcal{L}_{style} + \lambda_5\mathcal{L}_{par} \tag{14}$$

where λ_1, λ_2, λ_3 and λ_4 denote the weights of corresponding losses, respectively, which are the hyper-parameters of our model.

4. Experiment Setup

4.1. Dataset

We conduct experiments on DeepFashion Liu et al. (2016) dataset and Market 1501 Zheng et al. (2015) dataset. The DeepFashion dataset contains 52712 high-quality human models in different poses and clothes images, which are with a clean background. In our experiments, we split this dataset into 110426 pairs of images according to the processing of the PISE method. Among them, 101966 pairs of images are used for training and the others are used for testing. For comparison with other methods, we cropped all images to 256×176. Market 1501 dataset contains 32668 images of campus people with complex background images of size 128×64. We divide this dataset into 275632 pairs with the same setting as Zhang et al. (2022), of which 263632 pairs are used as the training set and 12,000 pairs are used as the test set.

4.2. Implementation Details

The image size of DeepFashion (256×176), Market-1501 (128×64), batch size (64), the number of body part (8), total training iteration (400000) are the basic setup. The Adam optimizer is adpoted to train our model in our experiment. The learning rate ($1 \times e - 5$), head(h=2), number of transformer module (N=2), λ_1 (2), λ_2 (0.25), λ_3 (200), λ_4 (2.5) and λ_5 (0.5) are the network training hyperparameters (and the default settings we use). And we use 4 Tesla-V100 with 16G memory to experiment.

4.3. Metrics

To evaluate our generated images from different aspect, we employ three typical metrics for quantitative evaluation: peak signal-to-noise ratio (PSNR) Hore and Ziou (2010), Frechet Inception Distance (FID) Heusel et al. (2017) , Learned Perceptual Image Patch Similarity (LPIPS) Zhang et al. (2018). PSNR calculates the error between the generated images and the ground-truth images at the pixel value level, which discriminates the overall color accuracy of the generated image. FID calculates the difference at the distribution level, which discriminates the realism of the generated image. LPIPS calculates the difference at the feature level, which estimates the rationality of the generated image at the deep features.

5. Experimental Results

In this section, we perform extensive experiments to compare qualitatively and quantitatively with the state-of-the-art methods in recent years, and sufficient ablation experiments are performed to support the soundness of our model design.
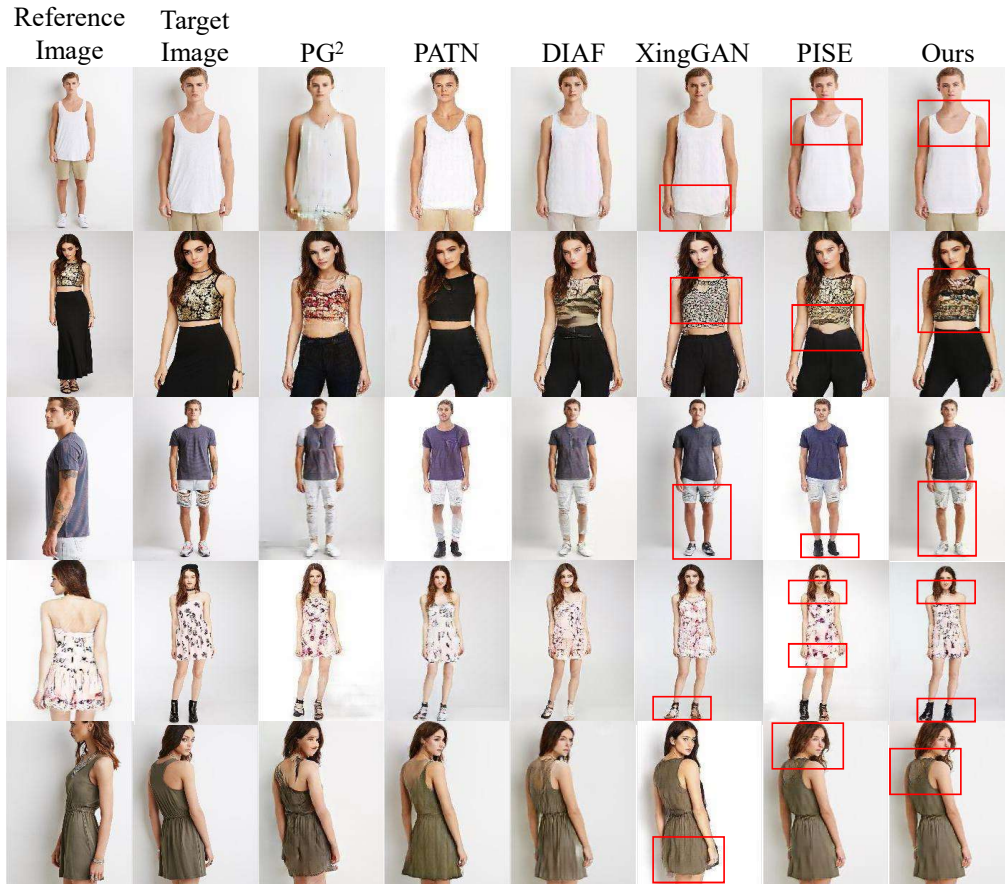
Figure 3: Qualitative comparison results on DeepFashion dataset. Our method yields more realistic human images, especially in terms of detail. As you can see from the red box that is outlined in the figure.

5.1. Qualitative Comparison

As shown in Figures 3 and 4, we have compared the human images generated by our method with several recent state-of-the-art methods on two datasets, these methods are: PG^2 Ma et al. (2017), PATN Zhu et al. (2019), DIAF Li et al. (2019), XingGAN Tang et al. (2020), PISE Zhang et al. (2021a), SPIG Lv et al. (2021), whose results were generated from the generated images and pre-trained models provided by the corresponding authors. For a fair comparison, we standardized the image size to 256×176 for all methods on the DeeepFashion dataset and 128×64 on Market-1501 dataset. The generated images by PISE and SPIG are better than the other baseline methods, including details such as clothes and colors. However, they are not as effective as they could be when transforming large movements (the fourth rows). Our method is designed to focus on more local detail information by decoupling the human body parts, and we have designed attention-based methods to handle the complex texture alignment. So our method shows superiority both in the overall

Figure 4: Qualitative comparison results on Market1501 dataset. Although the dataset itself is not well trained, our method still produces better results compared to other baseline methods. As you can see from the red box that is outlined in the figure.

rendering effect and in the local detail generation effect. Specifically, our method generates human images with more appropriate collar size and clothing color (the first rows and the second rows), which indicates that our method pays more attention to details. In addition, for large pose transfer (the third rows and the fifth rows), our method generates reasonable character textures for each compared to other methods, which reflects the role of attention mechanism in the overall rendering effect.

5.2. Quantitative Comparison

We compared the proposed method with several state-of-the-art methods including PG^2 Ma et al. (2017), PATN Zhu et al. (2019), DIAF Li et al. (2019), XingGAN Tang et al. (2020), PISE Zhang et al. (2021a), SPIG Lv et al. (2021). To show fair competition, we uniformly crop the image size to 256×176 for the DeepFashion dataset and 128×64 for the Market-1501 dataset. And the evaluation results of these models are calculated based on their pre-trained models, the results are shown in Table 1. The PISE and SPIG models gets more satisfied scores in three quantitative metrics than other baseline models, suggesting that the two-stage model and the region-adaptive normalization approach are at work. But it is obvious that our method gets the best score on the LPIPS metric on both two datasets, which indicates that the images generated by our method contain the most details and are the most consistent with human perception. And in PSNR and FID metrics, our method is also quite comparable with other models. For example, our method gets the best

Table 1: Quantitative comparative results of image quality and model size with baseline methods. Our method achieves the best score for both PSNR and LPIPS on the DeepFashion dataset, while the FID are comparable than baseline methods. And on the Market-1501 dataset, the FID and LPIPS of our method are the best.

Model	DeepFashion			Market1501			Model
	PSNR ↑	FID ↓	LPIPS ↓	PSNR ↑	FID ↓	LPIPS ↓	Size ↓
PG² Ma et al. (2017)	17.53	49.56	0.2928	14.17	86.03	0.3619	437.09 M
DSC Siarohin et al. (2018)	18.09	21.27	0.2440	14.31	27.01	0.3029	82.08 M
PATN Zhu et al. (2019)	18.25	20.75	0.2536	14.26	22.68	0.3194	41.36 M
DIAF Li et al. (2019)	16.90	14.88	0.2388	14.20	32.88	0.3059	49.58 M
XingGAN Tang et al. (2020)	17.92	39.32	0.2928	14.45	22.52	0.3058	42.77 M
PISE Zhang et al. (2021a)	18.52	**11.51**	0.2080	—	—	—	64.01 M
SPIG Lv et al. (2021)	18.56	12.93	0.2128	**14.47**	23.11	0.2827	117.13 M
PD-GAN (Ours)	**18.64**	13.13	**0.2070**	14.13	**20.16**	**0.2827**	**18.52** M

score on the PSNR metric on the DeepFashion dataset and the best on the FID metric on the Market-1501 dataset. Moreover, our model size is the smallest of any other models because we use the weight-sharing body part decoupling encoder, which is very important in application.

5.3. Ablation Study

We train several ablation models to prove our hypotheses and validate the effect of our improvements. Below are the many alternatives for eliminating the corresponding components from the full model:

The model without Transformer module (w/o Trans). This model removes the Transformer module. In this way, the model becomes a pure CNN-based model, which lacks the ability of long-range dependency modeling.
The model without Fast Fourier Transformation block (w/o FFT). This model removes the FFT module from the transformer module, which means that the ability to focus on low frequency information is missing and the ability to focus on global information is weaken.
The model without the part of face (pants, hair and so on) in the body decoupled part module (w/o Face (Pants, Hair)). This model removes the face (pants, hair and so on) part in the body decoupled part module.
Full model (Full). We use the PD-GAN model.

The results on DeepFashion dataset are shown in Figure 5 and Tabel 2. As shown in Figure 5, we can see that (1) w/o Trans method has the most severe results and fails to generate clear and complete images, which indicates transformer module plays an important role in building realistic human images. (2) Compare with w/o Trans, w/o FFT gets more

| Reference Image | Target Image | w/o Trans | w/o FFT | w/o Face | w/o Pants | w/o Hair | Full |

Figure 5: Qualitative results of image quality on ablation study. Without transformer module has the greatest impact on the model to produce images, in addition, the absence of the FFT module also reduces the quality of the images. Also missing each part of the human body will have an impact on the clarity of the human image generation.

sharp human images though gets worse results than the full model, which demonstrates FFT block with focusing on low frequency information can help generate detailed human images. (3) The quality of the images generated by w/o Face, w/o Pants and w/o Hair model is poor in the corresponding part than the full model, which means that each part of the human body part decoupled module will have an impact on the clarity of the human image generation. (4) As shown in Table 2, our Full model not only has satisfactory textural appearances but also gets the best scores of any other ablation model.

5.4. Human Image Pose Transfer

Human image pose transfer is to generate a target image with a target pose, given a reference image and a target pose. As shown in Figure 6, our model can generate realistic human images with details depending on different poses.

6. Conclusion and Discussion

In this paper, we have proposed a novel model PD-GAN for PGHIS by explicitly controlling the pose of the reference human images. Specifically, we decouple the body parts to help the generation of realistic human images. For PGHIS, we design a new transformer module to handle human images with complex topology, which not only can capture the long-range dependency, but also focus on the details of human images using the FFT block. Extensive experiments show that the results of PD-GAN are qualitatively and quantitatively better

Table 2: Quantitative results of image quality on ablation study. w/o Trans: without transformer module, w/o FFT: Transformer module without FFT block, w/o Face: lack of the part of face in the body decoupled part module, w/o Pants: lack of the part of pants in the body decoupled part module, w/o Hair: lack of the part of hair in the body decoupled part module, Full: our complete model.

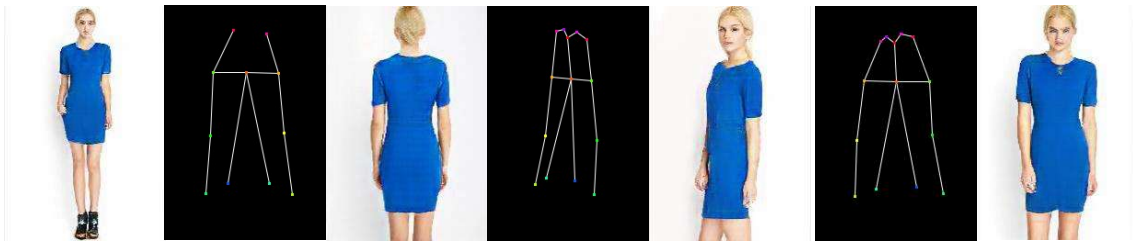|	w/o Trans	w/o FFT	w/o Face	w/o Pants	w/o Hair	Full
PSNR ↑	13.39	13.45	13.67	13.79	13.88	14.13
FID ↓	19.22	19.42	19.46	19.89	19.77	20.16
LPIPS ↓	0.3122	0.3021	0.3010	0.2991	0.2973	0.2827

Figure 6: Our results of human image synthesis in different poses.

than or competitive with existing state-of-the-art methods. Otherwise, the size of PD-GAN is the smallest, which is a big highlight in the application.

Limitation. Although our method produces satisfactory results in most cases, it still fails when the materials (*e.g.*, the stool in 7) in the datasets are extremely lacking. We show some failure cases in Figure 7, inconsistencies and blur appeared in these images. We believe that training the model on a larger dataset would alleviate this situation where the person and object in images are confused. Furthermore, applying this model to more tasks (*e.g.*, virtual try-on) is a future research direction for us, and we are committed to working on a generalized human image synthesis model.
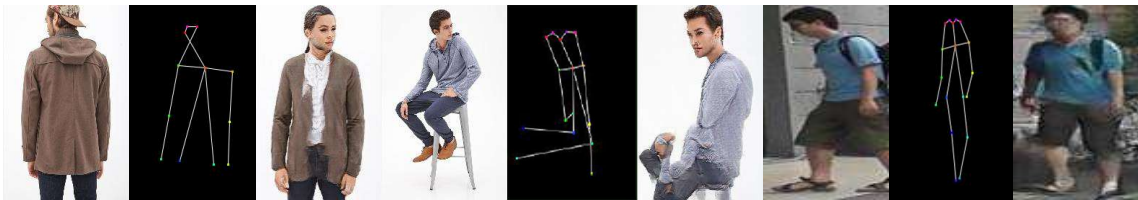

Figure 7: Failure cases. Our method generates less satisfactory images for very rare items in the dataset (*e.g.*, stools), or when there are obvious interferences.

7. Acknowledgement

References

Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7291–7299, 2017.

Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14638–14647, 2021.

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

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 (NeurIPS)*, 30, 2017.

Alain Hore and Djemel Ziou. Image quality metrics: Psnr vs. ssim. In *2010 20th international conference on pattern recognition*, pages 2366–2369. IEEE, 2010.

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 (ICCV)*, pages 1501–1510, 2017.

Drew A Hudson and C. Lawrence Zitnick. Generative adversarial transformers. *International Conference on Machine Learning (ICML)*, 2021.

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 (CVPR)*, pages 3688–3697, June 2019.

Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *CVPR*, pages 1096–1104, 2016.

Zhengyao Lv, Xiaoming Li, Xin Li, Fu Li, Tianwei Lin, Dongliang He, and Wangmeng Zuo. Learning semantic person image generation by region-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10806–10815, June 2021.

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 (NeurIPS)*, pages 405–415, 2017.

Xintian Mao, Yiming Liu, Wei Shen, Qingli Li, and Yan Wang. Deep residual fourier transformation for single image deblurring. *arXiv preprint arXiv:2111.11745*, 2021.

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

Bin Ren, Hao Tang, Fanyang Meng, Runwei Ding, Ling Shao, Philip HS Torr, and Nicu Sebe. Cloth interactive transformer for virtual try-on. *arXiv preprint arXiv:2104.05519*, 2021.

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 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7687–7696, 2020.

Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *CVPR*, pages 13535–13544, 2022.

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.

Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3408–3416, June 2018.

Mohsen Tabejamaat, Farhood Negin, and Francois Bremond. Guided flow field estimation by generating independent patches. *British Machine Vision Conference (BMVC)*, 2021.

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

Huaizhen Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, Zhen Zeng, Edward Xiao, and Jing Xiao. TGAVC: Improving autoencoder voice conversion with text-guided and adversarial training. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU2021)*, pages 938–945. IEEE, 2021.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning, (ICML)*, pages 10347–10357. PMLR, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6000–6010, 2017.

Jianhan Wu, Shijing Si, Jianzong Wang, and Jing Xiao. Augmentation-induced consistency regularization for classification. *arXiv preprint arXiv:2205.12461*, 2022a.

Jianhan Wu, Shijing Si, Jianzong Wang, and Jing Xiao. Improving human image synthesis with residual fast fourier transformation and wasserstein distance. *arXiv preprint arXiv:2205.12022*, 2022b.

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

Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7982–7990, 2021a.

Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7713–7722, 2022.

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 (CVPR)*, pages 586–595, 2018.

Xulong Zhang, Jianzong Wang, Ning Cheng, Edward Xiao, and Jing Xiao. CycleGEAN:cycle generative enhanced adversarial network for voice conversion. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU2021)*, pages 930–937. IEEE, 2021b.

Botao Zhao, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. nnspeech: Speaker-guided conditional variational autoencoder for zero-shot multi-speaker text-to-speech. In *ICASSP2022*, pages 4293–4297. IEEE, 2022.

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 (ICCV)*, pages 1116–1124, 2015.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *International Conference on Learning Representations (ICLR)*, 2021.

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 (CVPR)*, pages 2342–2351, 2019.