



**ORIGINAL RESEARCH**

# Robust style injection for person image synthesis

Yan Huang<sup>1</sup>  | Jianjun Qian<sup>1</sup>  | Shumin Zhu<sup>2</sup> | Jun Li<sup>1</sup> | Jian Yang<sup>1</sup><sup>1</sup>School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China<sup>2</sup>School of Fashion and Textiles, The Hong Kong Polytechnic University, Hong Kong, China**Correspondence**Jianjun Qian.  
Email: csjqian@njust.edu.cn**Funding information**

National Natural Science Foundation of China, Grant/Award Number: 62176124

**Abstract**

Person Image Synthesis has been widely used in fashion with extensive application scenarios. The point of this task is how to synthesise person image from a single source image under arbitrary poses. Prior methods generate the person image with target pose well; however, they fail to preserve the fine style details of the source image. To address this problem, a robust style injection (RSI) model is proposed, which is a coarse-to-fine framework to synthesise target the person image. RSI develops a simple and efficient cross-attention based module to fuse the features of both source semantic styles and target pose for achieving the coarse aligned features. The adaptive instance normalisation is employed to enhance the aligned features in conjunction with source semantic styles. Subsequently, source semantic styles are further injected into the positional normalisation scheme to avoid the fine style details erosion caused by massive convolution. In training losses, optimal transport theory in the form of energy distance is introduced to constrain data distribution to refine the texture style details. Additionally, the authors' model is capable of editing the shape and texture of garments to the target style separately. The experiments demonstrate that the authors' RSI achieves better performance over the state-of-art methods.

**KEYWORDS**

computer vision, image reconstruction, virtual try-on

## 1 | INTRODUCTION

With the development of computer vision (CV), Person Image Synthesis emerges and aims to generate realistic images and videos due to its extensive application scenarios, such as pose transfer [1–12], attribute manipulation [13–16], virtual try-on [17–21], person re-identification [22–26] and so on. Generative Adversarial Networks (GANs) [27] has led a series of breakthroughs in the area of pose transfer. This task is targeted at changing the texture and posture of the source person image given target poses as a condition. However, the difficulty of non-rigid deformation makes this task still challenging.

In the GAN-like framework, most methods follow the rule of the encoder–decoder scheme. Encoder captures the features of the source image and target pose which will be fused into the features of the target image later. Decoder reconstructs the target person image based on the deformed feature. Based on the aforementioned model, PATN [11] embedded large

numbers of attention-aware convolution modules to deform features. Similarly, ADGAN [3] led into several style blocks and Adaptive Instance Normalisation (AdaIN) [28] for style transfer. In this way, the synthesised person image preserves geometric shapes well. The fine texture details are unsatisfactory since the limitation of the spatial transfer ability in massive local convolution operations. To solve the above issue, flow-based methods [1, 4, 7, 29–31] are employed to predict flow fields by calculating the global correlations between sources and targets, which will instruct the model to rearrange the source features. Though flow-based scheme amends the lack of space transformation ability of Convolutional Neural Networks (CNNs), it may bring about terrible artefacts and imprecision geometric shape when the targets and sources have large differences. In refs. [2, 8, 32], the authors proposed an extra stage to obtain the target semantic segmentation and train the generator with the help of semantic segmentation results at the next stage which will enable the outfit edit.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2024 The Author(s). *CAAI Transactions on Intelligence Technology* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology and Chongqing University of Technology.

According to this way, these methods may spend much time and cost. Moreover, it is hard to predict the semantic segmentation map. Because even tiny error occurs will have an adverse impact on generated appearances.

To obtain the better style details, we seek to employ a coarse-to-fine scheme to improve the transformation between source person image to target pose. The coarse stage aims to align the source semantic style feature and pose feature. Subsequently, the fine stage utilises the cascade enhancement scheme to improve the geometric shape and semantic style details.

Based on these ideas, we propose a simple and efficient robust style injection (RSI) model for person image synthesis. First, the source image and the corresponding semantic segmentation map are fed into the style encoder to extract the features with eight semantic parts. We thus design a concise cross-attention module (CAM) to align the semantic style features with pose features. Specifically, we explore the intrinsic spatial correlation information between pose features and source style features and rearrange the semantic style features to perform the aligned results. In the cascade enhancement scheme, adaptive instance normalisation is used to improve the geometric shape in conjunction with semantic style features. We thus inject semantic style features to positional normalisation for enhancing the quality of fine style details. In training loss, Mini-batch Energy Distance is introduced to constrain the differences between the source images and the generated images from the distribution scale. Owing to the assistance of per-region style encoding, our model can edit every component of the person image to the style of reference images. Besides, we find that it is feasible to change the geometric shape and style texture of garments without introducing additional modules. Figure 1 shows some applications

of our model. In summary, the contributions are summarised as follows:

- We propose a simple CAM to align the semantic style features and pose features by exploring the intrinsic spatial correlation information behind feature maps.
- We develop the cascade enhancement scheme to improve the quality of the generated person image by combining the adaptive instance normalisation and positional normalisation with injected style features.
- We introduce the mini-batch energy distance to constrain the distribution of the source person images and the generated images, which can refine the texture style details.
- Extensive experiments on the Deepfashion dataset demonstrate that the proposed model achieves better performance and image quality over the state-of-art methods.

## 2 | RELATED WORK

Pose transfer was first proposed in 2017 and has since gained widespread attention. Large numbers of research studies combine this task with virtual try-on, fashion shopping and other scenarios, aiming to generate photo-realistic human images according to different poses. PG<sup>2</sup> [33] employed the U-net-like network to generate a coarse person image given the person image and a target pose. Based on this, the U-net-like network in an adversarial way is applied to refine the target person image. Pose GAN [34] proposed deformable skip connections and nearest-neighbour loss to improve the quality of the generated person image conditioned on the pose and the



**FIGURE 1** Left: pose transfer; Right: fashion edit (texture edit and shape edit). Given the source image and target key points, our model will generate the person image with the target pose. Besides the basic pose transfer task, our model can transfer the texture and shape of garments to the target style separately.

style. Besides, pt-GAN [35] introduced the data augmentation techniques to achieve the robustness in terms of occlusion, scale and illumination, while ADGAN [3] leveraged the semantic segmentation to encode the images into eight semantic parts, which could control the style of each part independently. Wang et al. [36] proposed a Self-supervised Correlation Mining Network to rearrange the source images in the feature space. PoNA [37] pointed out that previous works ignored the guidance function between image features and pose features. To solve this problem, PoNA designed a cross-modal block, with a pre-posed image-guided pose feature update and post-posed pose-guided image feature update.

The above-mentioned methods can acquire relatively accurate human shape, but the clothing texture details are not preserved to a great extent. The possible reason is CNNs are short of spatial transformation [38, 39] since CNNs repeat local operations to achieve the large receptive fields. What is more, classical neural networks often easily suffer from poor data efficiency, which results in poor generalisation capability [40]. Several methods have been proposed to tackle this issue. Flow-based models aim to compute the 2D coordinate offsets and indicate the positions of sampled source features for target points. Based on this, some scholars developed a series of flow-based methods [1, 4, 7, 31, 41] to synthesise the person image with fine texture details. Global Flow Local Attention (GFLA) [4] computed 2D flow fields to align source and target features. Based on GFLA, Dressing in Order [1] integrated global flow field into the encoder to produce spatially aligned texture feature maps, while Tang et al. [31] decomposed the overall flow field to learn local flow fields by semantic regions. Additionally, some flow-based models leverage the dense relationships between 2D image pixels and 3D body surface points to improve the generated person image [29, 30]. However, it may bring about terrible artefacts and imprecise shape when the targets and sources vary too much.

To obtain a better geometric shape, [8] utilised the predicted semantic segmentation results to train the generator for person image synthesis. Semantic segmentation is a crucial but challenging task in the fields of CV [42]. It is a pity that predicting semantic segmentation map require an expensive computation load. Moreover, it will have adverse an impact on the appearance of the generated person image if tiny errors and deviations occur. Some methods try to use the normalisation scheme to alleviate the above problem. PISE [8] proposed a spatially-aware normalisation to preserve the style details of the source person image in the target image by jointin global and local per-region encoding, while spatially adaptive warped normalisation introduced the flow field to assistant aligning the style and pose feature [7].

Additionally, attention mechanism [39] has been widely applied in the fields of natural language processing and CV, resulting in a large number of works [5, 9, 10, 43, 44] adopting attention mechanisms to compute the similarity between the source style and the target pose for fusing better target features. Chen et al. [43] led into a Progressive Multi-Attention Network, which updated features through several multi-attention transfer blocks (MATBs). Each MATB consisted of pose-conditioned

batch normalisation and the cooperative attention mechanism that acted on different levels of feature space. Zhou et al. [10] proposed a Cross-Attention-based Style Distribution Module (CASD), which consists of self- and cross-attention parts, to align the source semantic styles and pose features for person image generation. Zhang et al. [9] introduced the source-to-source task to assist the source-to-target task and constructed the dual-task pose transformer network. The pose transformer module is designed to explore the correlation between dual tasks for promoting the texture details of the synthesised person image. Based on the dual attention mechanism, Ren et al. [5] transferred the hierarchical semantic neural textures of the reference image to the target pose with the learnt spatial distribution. Bhunia et al. [45] replaced the GAN model with the diffusion model and introduced the texture diffusion module based on the attention mechanism to accurately model the correspondences between the appearance and pose information. The attention mechanism is expert in obtaining the global and local relationship. It is able to enhance regional key information, which plays a vital role in feature extraction and fusion.

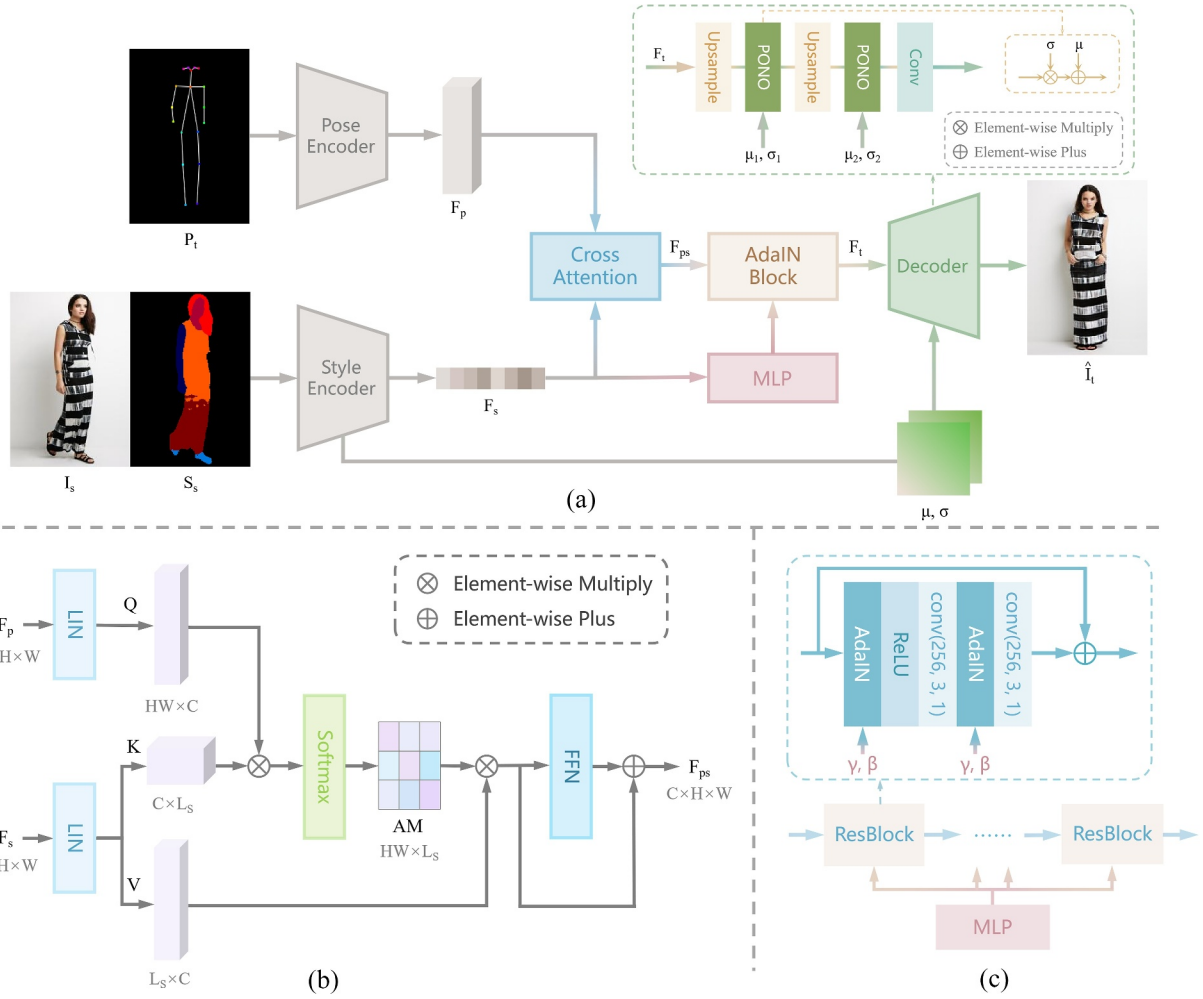
### 3 | METHOD

The overview of the proposed RSI model is shown in Figure 2. Given the source image  $I_s$ , source semantic parsing map  $S_s$  and target pose key point  $P_t$ , we aim to acquire controllable and photo-realistic person images  $\hat{I}_t$ . Specifically,  $P \in \mathbb{R}^{H \times W \times 18}$  is the corresponding heat map of 18 joints of the human body. The 18 joints are extracted by the existing pose estimation network [46].  $S \in \mathbb{R}^{H \times W \times 8}$  is the semantic parsing map of  $I$ , which divides the person into eight parts (such as face, arms, or upper clothes, etc.).

To begin with, we feed  $P_t$  into the pose encoder to achieve pose features  $F_p$ . For each channel in  $S_s$ , there is a binary mask for the corresponding part. The style encoder extracts per-region style code of  $I_s$  according to the binary mask. Then, every part's style code will be concatenated together in a top-down manner to get the full style code (i.e. style features). Both style features and pose features are utilised to conduct cross-attention, which calculates the similarity between two kinds of features and finally obtains the aligned feature  $F_{ps}$ . The detailed structures of the pose encoder and style encoder are shown in Figure 3. Similar to [3], we concatenate VGG-19 [47] encoder networks with basic encoder networks. Subsequently, we introduce several cascaded AdaIN blocks with the injected semantic style features to obtain  $F_t$ . We leverage the PONO to inject the style and structure information of source images into the decoder layers so that the generated results can acquire the fine texture details that may have been lost by prior works.

#### 3.1 | Cross attention based module

In this subsection, we propose the CAM to adapt the semantic style features into the required pose features. As is shown in Figure 2b, the embedded features  $F_p$  and  $F_s$  are fed to CAM.



**FIGURE 2** (a) Overview of the robust style injection (RSI) framework. (b) The architecture of cross-attention-based module (CAM). (c) The architecture of AdaIN block.

$F_p \in \mathbb{R}^{C \times H \times W}$  represents the feature from the target pose key points. Parameters  $H$  and  $W$  are the height and width of feature maps.  $C$  denotes the number of feature channels.  $F_s \in \mathbb{R}^{C \times L_s}$  denotes the source styles obtained from source images, where  $L_s$  represents the style length of each semantic segmentation region.

To align the semantic style feature and pose feature, we conduct cross-attention to adapt the source style into the target pose. CASD [10] motivates us to employ layer instance normalisation (LIN) [48] for better performance. The pose features  $F_p$  will be utilised to calculate Queries  $Q$  after LIN.  $Q$  is composed of  $H \times W$  vectors with  $C$  dimensions. The layer instance normalisation of style features  $F_s$  is linearly projected as the Keys  $K$  and values  $V$ . Note that  $Q = F_p W_Q$ ,  $K = F_s W_K$  and  $V = F_s W_V$ . The weight matrices  $W_Q$ ,  $W_K$ , and  $W_V$  can be learnt during the training process. We thus compute the attention matrix  $AM$  based on  $K$  and  $Q$  as follows:

$$AM = \text{Softmax} \left( \frac{QK^T}{\sqrt{C}} \right) \quad (1)$$

$AM$  actually reveals the similarity between pose features and style features. We employ  $AM$  and  $V$  to align the features

between the pose and style and further fused features by Feed Forward Networks (FFN).  $F_{ps}$  can be computed as follows:

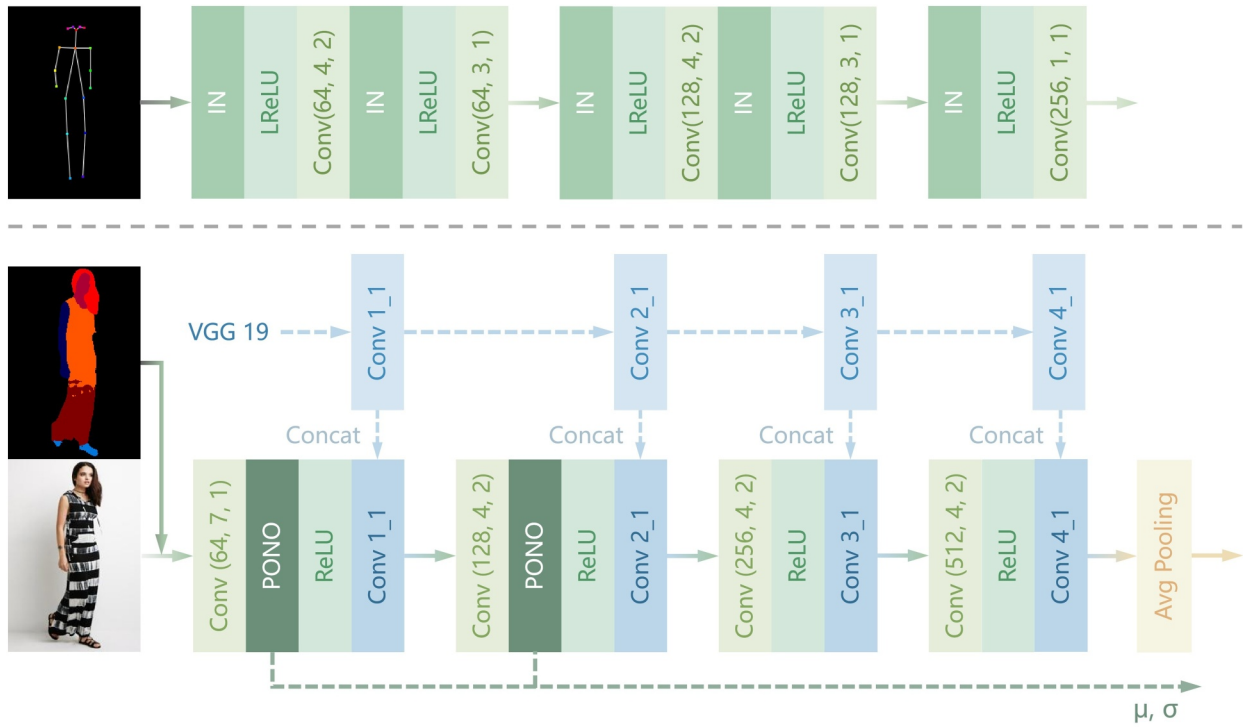
$$\begin{aligned} F_{ps} &= \text{Attention}(Q, K, V) \\ &= \text{FFN}(AM \cdot V) \\ &= \text{FFN} \left( \text{Softmax} \left( \frac{QK^T}{\sqrt{C}} \right) V \right) \end{aligned} \quad (2)$$

In CAM, for every single query in  $Q$ , we expect to seek out a weighted data item from  $V$  according to  $K$ . In this way,  $F_p$  and  $F_s$  can be fused to acquire the desired result of  $Q$ , which is referred as  $F_{ps}$ .

### 3.2 | Cascade style injection

In the light of the attention module, it is straightforward to obtain new pose features that integrate the style of the source image. However, it is difficult to synthesise the target pose image by using the aligned feature from CAM directly. For the





**FIGURE 3** The architecture of the pose encoder and style encoder. Target key points  $P_t$  are fed into the pose encoder to achieve pose features  $F_p$ . The style features  $F_s$  are represented by extracting per-region style features of the source image  $I_s$  according to the semantic parsing  $S_s$ .

sake of improving the aligned features, we propose the cascade style injection scheme in conjunction with AdaIN and PONO. Firstly, AdaIN blocks are used to obtain the target feature  $F_t$  for geometric shape enhancement. Based on this, PONO is further employed to improve the target feature for generating fine texture images. In this way, our model not only enables style transfer but also preserves the structure and details of the images.

It is well known that AdaIN is a good tool to perform style transfer in the feature space. AdaIN transfers the style by aligning the mean and variance of content features with those of style features. It offers flexibility and adaptability, which can accommodate various styles and generate diverse results. As demonstrated in Figure 2c, several resblocks equipped with AdaIN are used to normalise the aligned feature  $F_{ps}$ . We feed the semantic style features to Multilayer Perceptron (MLP) and compute the parameters of AdaIN. However, the deformation problem of pose-guided person image synthesis makes AdaIN incompetent. If  $F_{ps}$  is directly used to generate the target image as shown in Figure 4, we can see that AdaIN focuses on synthesising geometric shape, but the texture styles are far from the source image. To solve above problem, we introduce the PONO and inject the style features to further improve the quality of person image generation. PONO performs normalisation operations on each channel of the feature map, enabling better interaction and fusion of features across different positions in the network. It enhances the network's perception of the spatial structure. Hence, we establish a pathway between the downsample layers of the style encoder and the upsample layers of the decoder. The scale

$\sigma \in \mathbb{R}^{1 \times H \times W}$  and shift  $\mu \in \mathbb{R}^{1 \times H \times W}$  are computed from the style feature for positional normalisation. Parameters  $H, W$  are the height and width of feature maps. Because  $\mu$  and  $\sigma$  can reveal the essential structural signatures of source image to some extent.

To verify the effectiveness of  $\mu$  and  $\sigma$ , we design a pipeline to generate the images which were only affected by  $\mu$  and  $\sigma$ . From Figure 5, we can see that  $\mu$  and  $\sigma$  can preserve the appearance outline and texture details of source images. Briefly, this pathway is introduced to enhance the texture details of generated images on normalised results  $F_t$  and make up for the lost style information.

As is shown in Figure 2a, we extract positional moment information  $\mu$  and  $\sigma$  from style encoder layers and then refer them as mean value and standard deviation. Both  $\mu$  and  $\sigma$  are injected into the two layers of decoder directly. Consequently, given an input  $x$ , the normalisation result  $MS$  can be calculated by

$$MS(x) = \sigma F(x) + \mu \quad (3)$$

where  $F$  is modelled by intermediate layers.  $MS$  biases the decoder explicitly so that the activations in the decoder layers give rise to similar statistics with the corresponding layers in the encoder [49]. As a result, the PONO is applied between the downsample layers of the style encoder and the upsample layers of decoder. The fine styles of the synthesised person image is significantly improved after positional normalisation as shown in Figure 4. The possible reasons are that PONO captures feature statistic and reveals structural information



**FIGURE 4** Given the source person image, semantic segmentation map and pose skeleton, the synthesised results of our model in each step. The images for the fourth column are generated by using the aligned features of CAM  $F_{ps}$ . The aligned features are normalised by AdaIN blocks and are then used to generate the fifth column images. Based on the output of AdaIN blocks, PONO is utilised to normalise the features and synthesised the target person images as shown in the last column.

based on the output of AdaIN blocks to promote clothing shape and texture style details.

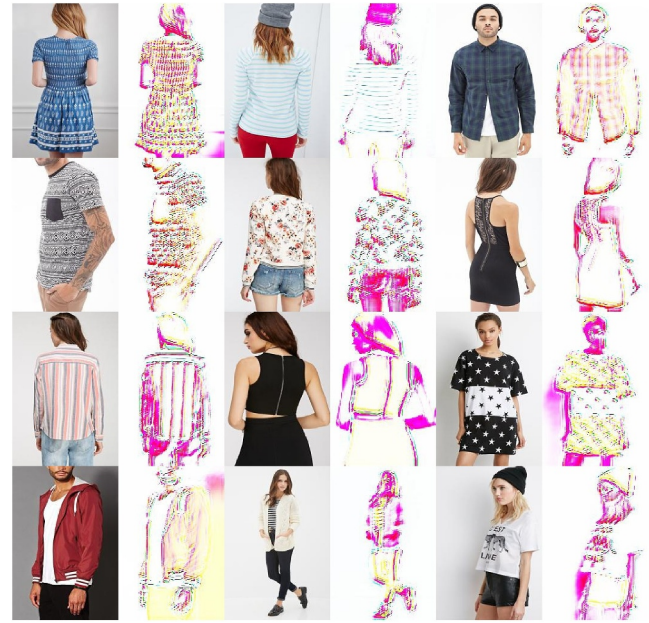
### 3.3 | Loss functions

Our full loss function is comprised of adversarial loss  $L_{adv}$ , reconstruction loss  $L_{rec}$ , perceptual loss  $L_{perc}$ , contextual loss  $L_{CX}$  and mini-batch energy loss  $L_{ME}$ . The full learning objectives are as follows:

$$L_{total} = \lambda_{adv}L_{adv} + \lambda_{rec}L_{rec} + \lambda_{perc}L_{perc} + \lambda_{CX}L_{CX} + \lambda_{ME}L_{ME} \quad (4)$$

where  $\lambda_{adv}$ ,  $\lambda_{rec}$ ,  $\lambda_{perc}$ ,  $\lambda_{CX}$ ,  $\lambda_{ME}$  denote the weights of the corresponding losses, respectively.

**Mini-batch Energy Loss.** Optimal transport (OT) is a good tool to measure the differences between two arbitrary distributions. Hence, we utilise the mini-batch energy distance to constrain the generated image distribution and the real image distribution. For this task, we refer the real source images  $I_s$  as  $X$  and the real target images  $I_t$  as  $X'$ , which is different from [50]. Meanwhile, our model requires to synthesise the corresponding person images  $Y, Y'$ . The generated source image  $\hat{I}_s$  (i.e.  $Y$ ) can be achieved by feeding the source image  $I_s$ , source semantic parsing map  $S_s$  and source pose key



**FIGURE 5** We assign feature  $F_t$  the value of all one, and feed it into the decoder according to the architecture of RSI. Then, the  $\mu$  and  $\sigma$  extracted from downsample layers of the encoder are injected into the feature map so that the generated results  $\hat{I}_t$  are just affected by positional normalisation. And  $\hat{I}_t$  can present the content of  $\mu$  and  $\sigma$ .

points  $P_s$  to our model.  $Y'$  denotes the generated target image  $\hat{I}_t$ . In our model, the mini-batch energy loss is as follows:

$$\begin{aligned} L_{ME} &= 2\mathbb{E}[\mathcal{W}_M(X, Y)] - \mathbb{E}[\mathcal{W}_M(X, X')] \\ &\quad - \mathbb{E}[\mathcal{W}_M(Y, Y')] \\ &= 2\mathbb{E}[\mathcal{W}_M(I_s, \hat{I}_s)] - \mathbb{E}[\mathcal{W}_M(I_s, I_t)] \\ &\quad - \mathbb{E}[\mathcal{W}_M(\hat{I}_s, \hat{I}_t)] \end{aligned} \quad (5)$$

where  $\mathcal{W}_M(X, Y)$  is the Wasserstein distance [51] between  $X$  and  $Y$ . The formulation of Wasserstein distance can be written as follows:

$$\mathcal{W}_M(X, Y) = \inf_{T \in \mathcal{T}} \langle T, M \rangle \quad (6)$$

where  $\langle \cdot, \cdot \rangle$  denotes the Frobenius inner product,  $T$  is the transport matrix, and  $\mathcal{T}$  is the set of transport matrices. The cost function  $Cost(\cdot, \cdot)$  gives rise to a transport cost matrix  $M$ , where  $M_{ij} = Cost(x_i, y_j)$  tells that how expensive it is to transport the  $i$ th data vector  $x_i$  of  $X$  to the  $j$ th data vector  $y_j$  of  $Y$ . The cost function is defined as follows:

$$Cost(x, y) = 1 - \frac{v(x) \cdot v(y)}{\|v(x)\|_2 \|v(y)\|_2} \quad (7)$$

where  $v(\cdot)$  represents a deep neural network operation that maps the images into a learnt latent space.  $Cost(x, y)$  is the

cosine distance between the corresponding vectors  $v(x)$  and  $v(y)$ .

**Adversarial Loss.** Similar to PATN [11], the pose discriminator  $D_p$  and style discriminator  $D_s$  are employed to improve the quality of generated results in adversarial learning. To be specific, the real pose pairs  $(P_t, I_t)$  and fake pose pairs  $(P_t, \hat{I}_t)$  are fed into  $D_p$  to judge the match degree of  $\hat{I}_t$  and target pose  $P_t$ . Given the real image pairs  $(I_s, I_t)$  and fake ones  $(I_s, \hat{I}_t)$ ,  $D_s$  aims to measure the style similarity between  $\hat{I}_t$  and ground truth.

$$L_{adv} = \mathbb{E}_{P_t, I_s, I_t} [\log(D_s(I_s, I_t) \cdot D_p(P_t, I_t))] \\ + \mathbb{E}_{P_t, I_s} [\log(1 - D_s(I_s, G(P_t, I_s))) \\ \cdot (1 - D_p(P_t, G(P_t, I_t)))] \quad (8)$$

**Reconstruction Loss.** Reconstruction loss  $L_{rec}$  takes advantage of  $L_1$  distance to calculate the difference between the generated image  $\hat{I}_t$  and the target image  $I_t$  at the pixel level. This loss assists the generator to synthesise photo-realistic images.

$$L_{rec} = \|\hat{I}_t - I_t\|_1 \quad (9)$$

**Perceptual Loss.** Perceptual loss  $L_{perc}$  [52] uses the pre-trained VGG-19 network [47] to extract the features of real images  $I_t$  and then compares with the features of generated images  $\hat{I}_t$  so that it is feasible to make their features as close as possible in the aspect of the content and structure.

$$L_{perc} = \sum_i \|\phi_i(\hat{I}_t) - \phi_i(I_t)\|_1 \quad (10)$$

where  $\phi_i$  is the feature map of the  $i$ th layer of the pre-trained VGG-19 network.

**Contextual Loss.** To adapt non-aligned data, Contextual loss  $L_{CX}$  [53] is also employed to compute the cosine distance between the real image features and the generated image features.

$$L_{CX} = -\log\left(CX\left(\Phi^l(\hat{I}_t), \Phi^l(I_t)\right)\right) \quad (11)$$

$\Phi_l(\hat{I}_t)$  and  $\Phi_l(I_t)$  denote the features extracted from the layer  $l = relu\{3\_2, 4\_2\}$  by using the pre-trained VGG-19 for images  $\hat{I}_t$  and  $I_t$  respectively.

## 4 | EXPERIMENTS

**Dataset.** We conduct all experiments on the DeepFashion Inshop Clothes Retrieval Benchmark [54]. It contains 52,712 images of fashion outfit with the resolution of  $256 \times 256$ , along with corresponding pose annotations and parsing results. Following the rule of PATN [11], we split this dataset into training and testing subsets with 101,966 and 8570 pairs.

**Metrics.** In our experiments, we adopt four metrics to validate the performance of the compared methods. Peak Signal to Noise Ratio (PSNR) is one of the widely used objective measurement methods, which is applied to evaluate image quality. Structural Similarity (SSIM) [55] focuses on the luminance, contrast and structure of the image. Fréchet Inception Distance (FID) [56] lays emphasis on the diversity and quality of the generated images and calculates the distance between the real images and generated images from the distribution view. Different from the previous evaluation metrics, Learnt Perceptual Image Patch Similarity (LPIPS) [57] computes the perceptual similarity of images from the perspective of human perception.

**Training Details.** We resize all the images from Deepfashion into  $256 \times 176$  for both training and testing. The number of semantic parsing part is 8, including background, pants, hair, skirt, face, upper clothes, arms, and legs. Our method is implemented in the PyTorch framework and trained on 2 NVIDIA TITAN RTX GPUs with the batch size of 12. We adopt the Adam optimiser [58] with  $\beta_1 = 0.5$ ,  $\beta_2 = 0.999$  for 1000 epochs, which is similar to ADGAN [3]. The initial learning rate is 0.001 and linearly decayed to 0 after 500 epochs. Concerning the learning objectives, the weights of five loss terms are set as  $\lambda_{adv} = 5$ ,  $\lambda_{rec} = 1$ ,  $\lambda_{perc} = 1$ ,  $\lambda_{CX} = 0.1$  and  $\lambda_{ME} = 1$ .

### 4.1 | Comparison with state-of-the-art methods

We compare the proposed method with the current state-of-the-art methods on Deepfashion benchmark, including ADGAN [3], SCAGAN [6], SPGNet [2], FHPT [59], TCN [60], DPTN [9], CASD [10], PISE [8], and PoNA [37]. The codes and well-trained models of the above methods are released by the corresponding authors. We test the released models and calculate several indicators on the generated images. For CASD [10] is trained on  $256 \times 256$  images, we generate the  $256 \times 256$  results and resize them into  $256 \times 176$  for evaluation. For PISE [8], the official results are  $256 \times 256$  images with blank on both sides, and we resize them into  $256 \times 176$  for evaluation. Some results from the published papers are also reported, which are marked with\*. In this subsection, we conduct quantitative and qualitative comparisons for the pose transfer task on these state-of-the-art methods. Besides, user study is also performed to verify the quality of generated images from a human perspective.

**Quantitative Comparison.** The quantitative results are listed in Table 1. The four metrics indicate the excellent image quality and similarity with ground truth. From Table 1, we can see that our model achieves the best performance on three of four metrics among all compared methods, which demonstrate the advantages of our model in improving the quality of the synthesised target person image.

**Qualitative Comparison.** Figure 6 shows some visual results of five selected methods, which we obtain the generated images. ADGAN, SCAGAN, SPGNet, PISE and PoNA



seem to transfer the human body shape and face from the source person image to the target pose while failing to preserve the texture style details. The possible reason is CNNs are short of spatial transformation since CNNs repeat local

**TABLE 1** Quantitative comparisons of each method on deepfashion.

Model	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$
ADGAN [3]	18.5647	0.7485	16.22	0.234
SCAGAN [6]	18.7984	0.7587	14.48	0.227
SPGNet [2]	18.5867	0.7759	12.70	0.210
FHPT <sup>a</sup> [59]	-	0.7740	14.61	0.218
TCN <sup>a</sup> [60]	18.7079	0.7742	12.36	0.208
CASD [10]	19.5329	0.7880	<b>11.37</b>	0.193
DPTN [9]	19.1772	0.7747	13.06	0.196
PISE <sup>a</sup> [8]	-	-	13.61	0.206
PoNA <sup>a</sup> [37]	-	0.7750	-	-
Ours	<b>19.6437</b>	<b>0.7919</b>	12.13	<b>0.182</b>

Note: Best in **bold**.

<sup>a</sup>Denotes the results from original paper.

operations to achieve the large receptive fields. SCAGAN employs the prior edge content transfer scheme to alleviate the spatial misalignment but fails to solve all problems. The generated target results are still unsatisfactory. Both CASD and DPTN perform better than ADGAN, SCAGAN and SPGNet in most cases, while the quality of generated target images is not satisfactory. In the first row of Figure 6, the upper clothes of the generated images by ADGAN, SCAGAN, and CASD have serious artefacts. In the fourth row of Figure 6, there are varying degrees of colour distortion on generated target images by ADGAN, SCAGAN, SPGNet and CASD. The shoulder straps of upper clothes generated by DPTN are missing. Additionally, we also find that our model achieves better results in predicting the invisible parts (e.g. the trousers and shoes in the last row). Compared with the above methods, our model can synthesise more clear face and realistic clothing textures (e.g. the first and fourth rows). The generated human body and clothes are more detailed and vivid (e.g. the second, third and last rows).

**User Study.** In addition, we also evaluate the image quality by using the subjective perception and evaluation of generated images from the view of users. Hence, we randomly select 20 groups of images from the test set and invite 35 volunteers to fill in the questionnaire. Specifically, we provide



**FIGURE 6** Qualitative comparisons with state-of-the-art methods on Deepfashion, including ADGAN [3], SCAGAN [6], SPGNet [2], CASD [10], DPTN [9], PISE [8], and PoNA [37].



the ground truth of the target image and the corresponding generated images of six methods for the volunteers. They are supposed to consider and select the image that is closest to the ground truth. In the description of the questionnaire, we do not give the definition of ‘closest’, which should be decided by the subjective of each individual. For fairness, all methods were unseen in the option, so the volunteers did not know which method generated which option. The evaluation results are listed in Table 2. We adopt Jab as the evaluation metric, which means the percentage that the image is judged as the best one (i.e. the closest to the ground truth). Each volunteer only gave one choice in each question, so we counted the proportion of times that this option has been selected in 20 questions. Higher values indicate better performance. The statistics indicate that our model significantly outperforms the compared methods (e.g. 20.9% higher than the second best one in Jab).

## 4.2 | Ablation study

In this subsection, we conduct an ablation study on the DeepFashion dataset to show the role of each component in our model. We thus remove the key component alternatively from our full model to introduce four variants as follows (w/o CAM, w/o AdaIN, w/o PONO, w/o ME loss).

**Impact of Cross-Attention (w/o CAM).** We remove the proposed cross-attention-based module to test and verify the effectiveness of our model. As vividly shown in Table 3, the cross-attention-based module has a great impact on the experimental results in the ablation study, which proves the necessity of this feature-aligned module.

**Impact of Adaptive Instance Normalisation (w/o AdaIN).** The AdaIN block is removed during the training process. From Table 3 and Figure 7, we can see that AdaIN is an indispensable part of our model, which helps to generate the exact geometric shape and lays the appearance foundation

for the detail injection. The results become blurry and irregular in shape when AdaIN is missing.

**Impact of Positional Normalisation (w/o PONO).** This model removes the positional normalisation. From Table 3 and Figure 7, we can see that PONO plays a vital role in generating fine texture details. It is obvious that the texture details and clothing shape generated by the model without PONO are



FIGURE 7 The qualitative results of the ablation study on four variants.



FIGURE 8 More texture details with the ablation study of positional normalisation.

TABLE 2 The quantitative comparisons of the user study on several state-of-the-art methods.

Model	ADGAN	SCAGAN	SPGNet	CASD	DPTN	Ours
Jab $\uparrow$	3.7%	7.5%	7.1%	25.4%	10.0%	<b>46.3%</b>

Note: Jab means the percentage that the image is judged as the best one. Best in **bold**.

TABLE 3 The quantitative comparisons of the ablation study on the deepfashion dataset.

Model	PSNR $\uparrow$	SSIM $\uparrow$	FID $\downarrow$	LPIPS $\downarrow$
w/o CAM	19.2019	0.7765	12.72	0.1960
w/o AdaIN	18.9440	0.7662	16.55	0.2063
w/o PONO	19.2878	0.7801	12.45	0.1908
w/o ME	19.3900	0.7823	12.18	0.1874
Ours	<b>19.6437</b>	<b>0.7919</b>	<b>12.13</b>	<b>0.1820</b>

Note: Best in **bold**.



**FIGURE 9** Some examples of texture and shape edit on source images according to reference images.

inferior to the ones in the full model. In order to observe more clearly, we enlarge the local texture details to demonstrate the effects with PONO, which is shown in Figure 8.

**Impact of Mini-batch Energy Loss (w/o ME loss).** In the training stage, we remove the mini-batch energy loss defined in Equation (5). The performance of our model degrade without ME loss as listed in Table 3. In the sixth column of Figure 7, we also find that the edge between sleeveless shirt and skirt is confused. The colour of a sleeveless shirt is distortion. In our opinions, ME loss helps our model to refine the texture style details.

To sum up, both quantitative and qualitative results empirically demonstrate that our full model achieves a better performance and higher image quality against other variants of our models.

### 4.3 | Fashion edit

In this subsection, we divide the image into eight parts (background, pant, hair, skirt, face, upper cloth, arm and



**FIGURE 10** More examples of fashion edit with different regions: upper clothes and dresses.



leg), following the common semantic segmentation manner. Owing to the assistance of per-region style encoding, our model can also edit every component of person images to the style of reference images without further training.

**Texture Edit.** Similar to ADGAN [3], we can achieve the controllable person image by changing the texture of the targeted semantic region. Hence, we replace the specific part of the style code according to the reference image for texture edit. Given a reference image, it is feasible to edit the texture of different semantic regions by our model as shown in Figure 9. More results are displayed in Figure 10.

**Shape Edit.** Here, we further provide the visual results of shape edit to show the merits of our model. We find that it is easy to change the shape of upper clothes as long as we replace the style code of arms. For the same reason, our model is capable of editing the clothing shape of pants. As a result, both texture and clothing shape of the targeted region can be edited by changing the style code of the corresponding parts as shown in Figure 9.

#### 4.4 | Limitations

Here, we also provide some failure cases to demonstrate the limitations of our model as shown in Figure 11. From Figure 11, we find that the performance of our model degrades when body parts overlap and key points are missing. Additionally, it is difficult to preserve the complete graphics or texts of garments by just using the encoder–decoder scheme. In the future, it is also interesting to investigate

the above problems to improve the quality of person image synthesis.

## 5 | CONCLUSION

In this paper, we propose an RSI model for person image synthesis and fashion edit. Our model aligns the style features and target pose features by using the proposed cross-based attention model. Subsequently, we combine the AdaIN and PONO together to develop the cascade style injection scheme, which enhances the human shape and improves the texture style details for the target person image. Additionally, mini-batch energy loss is introduced to make the feature distribution of the generated images approximate to that of the real ones and further refine the texture style details. The quantitative and qualitative results show that our model can improve the quality of the synthesised person image over previous works. Besides, the ablation study demonstrates the indispensable role of each component.

## ACKNOWLEDGEMENTS

The authors would like to thank the editor and the anonymous reviewers for their critical and constructive comments and suggestions. This work was supported by the National Science Fund of China under Grant No. 62176124.

## CONFLICT OF INTEREST STATEMENT


The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The dataset is publicly available.

## ORCID

Yan Huang  <https://orcid.org/0009-0003-1498-7440>

Jianjun Qian  <https://orcid.org/0000-0002-0968-8556>

## REFERENCES

1. Cui, A., McKee, D., Lazebnik, S.: 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, pp. 14638–14647 (2021)
2. Lv, Z., et al.: Learning semantic person image generation by region-adaptive normalization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10806–10815 (2021)
3. Men, Y., et al.: Controllable person image synthesis with attribute-decomposed gan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5084–5093 (2020)
4. Ren, Y., et al.: Deep spatial transformation for pose-guided person image generation and animation. *IEEE Trans. Image Process.* 29, 8622–8635 (2020). <https://doi.org/10.1109/tip.2020.3018224>
5. Ren, Y., et al.: Neural texture extraction and distribution for controllable person image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13535–13544 (2022)
6. Yu, W.-Y., et al.: Spatial content alignment for pose transfer. In: 2021 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2021)



**FIGURE 11** Some examples of failure cases: missing part key points (in the first row); body overlap (in the second row); graphics or texts (in the last row).



7. Zhang, J., et al.: Controllable person image synthesis with spatially-adaptive warped normalization. *arXiv preprint arXiv:2105.14739* (2021)
8. Zhang, J., et al.: Pise: person image synthesis and editing with decoupled gan. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7982–7990 (2021)
9. Zhang, P., et al.: Exploring dual-task correlation for pose guided person image generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7713–7722 (2022)
10. Zhou, X., et al.: Cross attention based style distribution for controllable person image synthesis. In: *European Conference on Computer Vision*, pp. 161–178. Springer (2022)
11. Zhu, Z., et al.: Progressive pose attention transfer for person image generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2347–2356 (2019)
12. Ma, L., et al.: Multi-scale cross-domain alignment for person image generation. *CAAI Trans. Intell. Technol.* 9(2), 374–387 (2023). <https://doi.org/10.1049/cit2.12224>
13. Wang, T., et al.: High-fidelity gan inversion for image attribute editing. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11379–11388 (2022)
14. Wang, Q., et al.: Coarse-to-fine attribute editing for fashion images. In: *CAAI International Conference on Artificial Intelligence*, pp. 396–407. Springer (2021)
15. Zhang, G., et al.: Generative adversarial network with spatial attention for face attribute editing. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 417–432 (2018)
16. He, Z., et al.: Attgan: facial attribute editing by only changing what you want. *IEEE Trans. Image Process.* 28(11), 5464–5478 (2019). <https://doi.org/10.1109/tip.2019.2916751>
17. Han, X., et al.: Viton: an image-based virtual try-on network. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7543–7552 (2018)
18. Hsiao, W.-L., et al.: Fashion++: minimal edits for outfit improvement. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5047–5056 (2019)
19. Issenhuth, T., Mary, J., Calauzènes, C.: End-to-end learning of geometric deformations of feature maps for virtual try-on. *arXiv* (2019). preprint *arXiv:1906.01347*
20. —, Do not mask what you do not need to mask: a parser-free virtual try-on. In: *European Conference on Computer Vision*. Springer, pp. 619–635 (2020)
21. Yang, H., et al.: Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7850–7859 (2020)
22. Ge, Y., et al.: Fd-gan: pose-guided feature distilling gan for robust person re-identification. *Adv. Neural Inf. Process. Syst.* 31 (2018)
23. Sun, W., Liu, F., Xu, W.: Unlabeled samples generated by gan improve the person re-identification baseline. In: *Proceedings of the 2019 5th International Conference on Computer and Technology Applications*, pp. 117–123 (2019)
24. Wei, L., et al.: Person transfer gan to bridge domain gap for person re-identification. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 79–88 (2018)
25. Wang, G., et al.: Spatial-temporal person re-identification. *Proc. AAAI Conf. Artif. Intell.* 33(01), 8933–8940 (2019). <https://doi.org/10.1609/aaai.v33i01.33018933>
26. Wang, Y., et al.: Multi-granularity re-ranking for visible-infrared person re-identification. *CAAI Trans. Intell. Technol.* 8(3), 770–779 (2023). <https://doi.org/10.1049/cit2.12182>
27. Goodfellow, I., et al.: Generative adversarial networks. *Commun. ACM* 63(11), 139–144 (2020). <https://doi.org/10.1145/3422622>
28. Huang, X., Belongie, S.: Arbitrary style transfer in real-time with adaptive instance normalization. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1501–1510 (2017)
29. Grigorev, A., et al.: Coordinate-based texture inpainting for pose-guided human image generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12135–12144 (2019)
30. Albahar, B., et al.: Pose with style: detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Trans. Graph.* 40(6), 1–11 (2021). <https://doi.org/10.1145/3478513.3480559>
31. Tang, J., et al.: Structure-aware person image generation with pose decomposition and semantic correlation. *Proc. AAAI Conf. Artif. Intell.* 35(3), 2656–2664 (2021). <https://doi.org/10.1609/aaai.v35i3.16369>
32. Chen, J., et al.: Exploring kernel-based texture transfer for pose-guided person image generation. *IEEE Trans. Multimed.* 25, 7337–7349 (2022). <https://doi.org/10.1109/tmm.2022.3221351>
33. Ma, L., et al.: Pose guided person image generation. *Adv. Neural Inf. Process. Syst.* 30 (2017)
34. Siarohin, A., et al.: Deformable gans for pose-based human image generation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3408–3416 (2018)
35. Karmakar, A., Mishra, D.: A robust pose transformational gan for pose guided person image synthesis. In: *National Conference on Computer Vision, Pattern Recognition, Image Processing, and Graphics*, pp. 89–99. Springer (2019)
36. Wang, Z., et al.: Self-supervised correlation mining network for person image generation. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7703–7712 (2022)
37. Li, K., et al.: Pona: pose-guided non-local attention for human pose transfer. *IEEE Trans. Image Process.* 29, 9584–9599 (2020). <https://doi.org/10.1109/tip.2020.3029455>
38. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* 521(7553), 436–444 (2015). <https://doi.org/10.1038/nature14539>
39. Vaswani, A., et al.: Attention is all you need. *Adv. Neural Inf. Process. Syst.* 30 (2017)
40. Li, Z., et al.: Knowledge-guided semantic transfer network for few-shot image recognition. *IEEE Transact. Neural Networks Learn. Syst.*, 1–15 (2023). <https://doi.org/10.1109/tnnls.2023.3240195>
41. Ma, L., et al.: Fda-gan: flow-based dual attention gan for human pose transfer. *IEEE Trans. Multimed.* 25, 930–941 (2021). <https://doi.org/10.1109/tmm.2021.3134157>
42. Li, Z., et al.: Ctnet: context-based tandem network for semantic segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 44(12), 9904–9917 (2021). <https://doi.org/10.1109/tpami.2021.3132068>
43. Chen, B., et al.: Pman: progressive multi-attention network for human pose transfer. *IEEE Trans. Circ. Syst. Video Technol.* 32(1), 302–314 (2021). <https://doi.org/10.1109/tcsvt.2021.3059706>
44. Tang, H., et al.: Xinggan for person image generation. In: *European Conference on Computer Vision*, pp. 717–734. Springer (2020)
45. Bhunia, A.K., et al.: Person image synthesis via denoising diffusion model. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5968–5976 (2023)
46. Cao, Z., et al.: Realtime multi-person 2d pose estimation using part affinity fields. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7291–7299 (2017)
47. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014)
48. Xu, W., et al.: Drb-gan: a dynamic resblock generative adversarial network for artistic style transfer. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6383–6392 (2021)
49. Li, B., et al.: Positional normalization. *Adv. Neural Inf. Process. Syst.* 32 (2019)
50. Salimans, T., et al.: Improving gans using optimal transport. *arXiv preprint arXiv:1803.05573* (2018)
51. Genevay, A., Peyré, G., Cuturi, M.: Sinkhorn-autodiff: tractable wasserstein learning of generative models. *Working Papers* (2017)
52. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: *European Conference on Computer Vision*, pp. 694–711. Springer (2016)
53. Mechrez, R., Talmi, I., Zelnik-Manor, L.: The contextual loss for image transformation with non-aligned data. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 768–783 (2018)

54. Liu, Z., et al.: Deepfashion: powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1096–1104 (2016)
55. Wang, Z., et al.: Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.* 13(4), 600–612 (2004). <https://doi.org/10.1109/tip.2003.819861>
56. Heusel, M., et al.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Adv. Neural Inf. Process. Syst.* 30 (2017)
57. Zhang, R., et al.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595 (2018)
58. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014)
59. Yang, L., et al.: Towards fine-grained human pose transfer with detail replenishing network. *IEEE Trans. Image Process.* 30, 2422–2435 (2021). <https://doi.org/10.1109/tip.2021.3052364>
60. Zhang, P., et al.: Lightweight texture correlation network for pose guided person image generation. *IEEE Trans. Circ. Syst. Video Technol.* 32(7), 4584–4598 (2021). <https://doi.org/10.1109/tcsvt.2021.3131738>

**How to cite this article:** Huang, Y., et al.: Robust style injection for person image synthesis. *CAAI Trans. Intell. Technol.* 10(2), 402–414 (2025). <https://doi.org/10.1049/cit2.12361>