Person image generation through graph-based and apperance-decomposed generative adversarial network

Yuling He [Equal first author, 1], **Yingding Zhao** [Equal first author, 2], **Wenji Yang** [Corresp., 2], **Yilu Xu** [2]

[1] School of Computer and Information Engineering, Jiangxi Agricultural University, NanChang, JiangXi, China

[2] School of Software, Jiangxi Agricultural University, NanChang, JiangXi, China

Corresponding Author: Wenji Yang
Email address: ywenji614@jxau.edu.cn

Due to the sophisticated entanglements for non-rigid deformation, generating person images from source pose to target pose is a challenging work. In this paper, we present a novel framework to generate person images with shape consistency and appearance consistency. The proposed framework leverages Graph Network to infer the global relationship of source pose and target pose in a graph for better pose transfer. Moreover, We decompose the source image into different attributes (e.g.,hair,clothes,pants and shoes) and combine them with the pose coding through operation method to generate a more realistic person image. We adopt an alternate updating strategy to promote mutual guidance between pose modules and appearance modules for better person image quality. Qualitative and quantitative experiments were carried out on DeepFashion dateset.The efficacy of the presented framework are verified.

Person image generation through Graph-Based and Apperance-Decomposed generative adversarial network

Yuling He[1], Yingding Zhao[2], Wenji Yang[3], Yilu Xu[4]
[1] School of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang,JiangXi, China
[2] School of Software, Jiangxi Agricultural University, Nanchang,JiangXi, China

Corresponding Author:
Wenji Yang [2]
1101 Zhimin Avenue, Nanchang, JiangXi, 330045, China
Email address: ywenji614@jxau.edu.cn

Person image generation through Graph-Based and Apperance-Decomposed generative adversarial network

Yuling He[1], Yingding Zhao[2], Wenji Yang[3], Yilu Xu[4]
[1] School of Computer and Information Engineering, Jiangxi Agricultural University, Nanchang,JiangXi, China
[2] School of Software,Jiangxi Agricultural University, Nanchang,JiangXi, China

Corresponding Author:
Wenji Yang [2]
1101 Zhimin Avenue, Nanchang, JiangXi, 330045, China
Email address: ywenji614@jxau.edu.cn

Abstract

Due to the sophisticated entanglements for non-rigid deformation， generating person images from source pose to target pose is a challenging work . In this paper ,we present a novel framework to generate person images with shape consistency and appearance consistency. The proposed framework leverages Graph Network to infer the global relationship of source pose and target pose in a graph for better pose transfer.Moreover, We decompose the source image into different attributes(e.g.,hair,clothes,pants and shoes) and combine them with the pose coding through operation method to generate a more realistic person image.We adopt an alternate updating strategy to promote mutual guidance between pose modules and appearance modules for better person image quality. Qualitative and quantitative experiments were carried out on DeepFashion dateset .The efficacy of the presented framework are verified.

Keywords: Image generation; Generative adversarial network; Graph Network; Pose Transfer

1 Introduction

Recently generating human images from source pose to target pose, which is com-monly known as pose transfer, has obtained great attentions and it is of great value in many tasks such as intelligent photo editing(Wu et al. 2017), film production(Cui & Wang 2019; Xiong et al. 2018), virtual try-on(Dong et al. 2019; Honda 2019; Kubo et al. 2018) and person re-identification(Alqahtani et al. 2019; Dai et al. 2018; Liu et al. 2019; Lv & Wang 2018). (Ma et al. 2017)first proposed this problem, where the framework transformed person images to target pose while keeping the appearance details of the source image. Then more researchers put forward pose transfer networks(Huang et al. 2020; Song et al. 2019; Tang et al. 2020; Zhu et al. 2019). The methods mentioned above are all based on convolutional neural network,which is good at extracting local relations, but it is inefficient in dealing with global inter regional relations. To obtain a greater receptive field,the traditional CNN needs to stack many layers. This may cause some problems for obtaining the global relationship between source pose and target pose, which not only increases the calculation cost, but also has incomplete problems.

In this paper, we propose a pose transfer framework based on graph network and appearance decomposition. Inspired by(Chen et al. 2019), we map the source pose and target pose to the same interaction space. After global reasoning in the interaction space, we map the different poses back to the original independent space. Specificially,as shown in Figure2, We construct an interaction space for global reasoning, map the key points of the source pose and the target pose to the interaction space respectively, establish a fully connected graph connecting all the joint points in the space, and carry out relationship reasoning on the graph. After reasoning is completed, the updated joint points are remapped back to the original space. For appearance code, we use a VGG-based pre-trained human parser to decompose the attributes of source images. Then these attributes are input into a texture encoder to reconstruct the style code, and finally the style code and the pose code are combined to obtain the generated images. In the training process, we use a pair of conditional discriminators, which combine the conditional discriminator and appearance discriminator to improve the quality of the generated image. The performance of proposed network outperform prior works both qualitatively and quantitatively on challenging benchmarks. In total,the proposed framework has the following contributions:

- We propose a novel generative adversarial network based on graph,which can infer the global relationship between different pose.Tackling the problem that CNN needs to overlay multiple convolution layers to expand the receptive field to cover all the joint points of source pose and target pose.
- We employ the human body parser to decompose the attributes of the human body images, and fuse the attribute coding with the pose coding. Therefore, the generated images are desirable.

The remainder of this paper is structured as follows.In Section 2, the related work of this paper is introduced. Section 3, details of the proposed framework are given. Section 4 presents distinct experiments on Deepfashion dataset. Finally, a summary is given in Section 5.

2 Related Work

2.1 Person image generation

With the continuous development of computer vision technology, image generation models have been developing at a high rate in recent years. The two mainstream methods are Variational auto-encoder(VAE)(Kingma & Welling 2013; Lassner et al. 2017; Rezende et al. 2014; Sohn et al. 2015) and Generative Adversarial Networks (GANs)(Balakrishnan et al. 2018; Dong et al. 2019; Honda 2019; Si et al. 2018; Zanfir et al. 2018). The former captures the relationship between different dimensions of the data by modeling the structure of the data to generate new data. The latter generates images through mutual game between the generator and the discriminator. Since the loss used by GANs is better than VAE, GANs can generate more vivid images and is sought after by more researchers.

Aiming at the human body image generation method based on the generative adversarial network, Ma et.al first proposed PG^2(Ma et al. 2017) to achieve pose guided person body image generation, whose model is cascaded by two different generators. The first stage generates a blurry image under the target pose. The second stage improves the texture and color quality of

the image generated in the first stage. Although the second stage improves the image quality to a certain extent, it is still unable to capture the changes in image distribu-tion well, which makes the generated images lack of fine texture. To obatin better apperance texture, (Esser et al. 2018)exploited to combine VAE and U-Net to disentangle appearance and pose, using the decoupled posture information to generate pictures, and then integrate the appearance information of the source images into the generated pictures. However, it will cause the problem of feature offset caused by posture difference due to the U-Net based skip connections in the model. To tackle this problem, (Siarohin et al. 2018) introduced deformable skip connection to transfer features of various parts of the body, which effectively alleviated the problem of feature migration. In order to control the attributes flexibly, (Men et al. 2020) proposed Attribute-Decomposed GAN, which embeds the attribute codes of each part of the human body into the potential space independently, and recombines these codes in a specific order to form a complete appearance code, so as to achieve the effect of flexible control of each attribute.

2.2 Graph-based Reasoning

Graph is a data structure, which can model a group of objects (nodes) and their relationships (edges). In recent years, more and more attention has been paid to the study of graph analysis based on machine learning due to its powerful expression ability. (Kipf & Welling 2016) firstly proposed graph convolutional network which used an efficient layer-wise propagation rule that is based on a first-order approximation of spectral convolutions on graphs. In order to pay dynamical attention to the features of adjacent nodes, Graph attention networks (Veli—ikovi—ç et al. 2017) have been proposed. (Wang et al. 2020)introduced a Global Relation Reasoning Graph Convolutional Networks (GRR-GCN) to efficiently capture the global relations among different body joints. It modeled the relations among different body joints that may mitigate some challenges such as occlusion. In this paper, we introduce a graph-based reasoning in person image generation model.

3 Methods

In this section, we give a description of our network architecture. We start with some notations. $I \in R^{3\times H\times W}$ denotes the set of person images. Before training, the Human Pose Estimator (HPE) (Esser et al. 2018) is adopted to estimate the position of 18 joint points in the images. $P \in R^{18\times H\times W}$ represents a 18-channel heat map that encodes the locations of 18 joints of a human body. During the training, the model requires source images and target images (I_c , I_t) and their corresponding heat map (P_c , P_t) as input. Moreover, we adopt a VGG-based pre-trained human parser to decompose the attributes of source images. More details will be introduced below.

3.1 Generator

Figure 3 shows the architecture of the generator which aims to transfer the pose of the person in I_c from P_c to P_t. At the core the generator comprise two pathway, namely pose pathway and appearance pathway. The former consisted of a series of pose blocks and the latter consisted of several texture blocks.

3.1.1 Style Encoder

Due to the manifold structure composed of various human body images is very complex, it is difficult to encode the entire human body with detailed textures. Inspired by (Men et al. 2020),we decompose the source image into different components and recombine their potential code to build the complete style code. Firstly, a pretrained human body parser based on VGG is used to obtain the semantic map of the source image I_c. Then, the semantic map is mapped to a K-channels heat map $M^{K\times H\times W}$. Each channel i has a binary mask $M_i^{H\times W}$ corresponding to different components. Multiplying element-wise the source image I_c and the mask $M_i^{H\times W}$ to obtain decomposed person image with component i.

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

After that, I_c^i is input into the appearance encoder to acquire the corresponding style code F_{sty}^i.

$$F_{sty}^i = T_{enc}(I_c^i) \tag{2}$$

where T_{enc} is shared for all components and then all F_{sty}^i is concatenated to get a full style code F_{sty}. The structure of T_{enc} adopts the encoder in (Men et al. 2020).

3.1.2 Pose Encoder

In the pose pathway, the source pose P_c and target pose P_t are embedded into the latent space as the pose code F^{P_c} and F^{P_t} by pose encoder. Note that we adopt the same shape encoder for P_c and P_t which consists of N-down-sampling convolutional layers($N = 2$ in our case). That is to say two shape encoders are sharing the weights.

3.1.3 Pose Block

The pose block aims to reason the crossing longrange relations between the source pose and the target pose in a graph and output new shape codes. The main idea of this method is to map the source pose and the target pose to the graph space, then cross reasoning on the graph space, and finally map back to the original space to get the updated code. Firstly,we learn the projection function that maps source pose and target pose from coordinate space to graph space.

$$H_{source} = \theta(F^{P_c}) \in \mathbb{R}^{C\times D} \tag{3}$$

$$H_{target} = \theta(F^{P_t}) \in \mathbb{R}^{C\times D} \tag{4}$$

where function $\theta(\cdot)$ is implemented by 1×1 convolutional layer, C and D represent feature map channels and number of nodes respectively. Then we can get the new features with the cross relationship between the source pose and the target pose.

$$V = H_{source} \cdot H_{target} \tag{5}$$

After obtaining new features, we use graph convolution for interactive reasoning. In particular,let $A \in \mathbb{R}^{D\times D}$ denote the node fully-connected adjacency matrix for spreading information across nodes, and let $W \in \mathbb{R}^{D\times D}$ denote the state update function. Identity matrix I reduces the difficulty of optimization the graph convolution is formulated as:

$$Z = ((I - A)V)W \tag{6}$$

Following the principle in (Chen et al. 2019), Laplacian smoothing is used as the first step of volume product. Both A and W are adopt random initialization and updated by gradient descent.

Next, we need to map the inferred Z back to the coordinate space. Similar to the first step, we adopt the projection matrix H_{target} and linear projection $\varphi(g)$ to formulate.

$$\hat{F}^{P_t} = \varphi(H_{target} g Z) \tag{7}$$

3.1.4 Texture Block

The texture blocks aims to transfer pose and texture simultaneously and interactively. Firstly, we compute attention mask M_t by two convolutional layers. Mathematically,

$$M_t = \sigma(Conv(\hat{F}^{P_c})) \tag{8}$$

After getting the attention mask, the appearance code is updated by:

$$F_{sty}^{i} = F_{sty}^{i-1} M_t + F_{sty}^{i-1} \tag{9}$$

The pose code is updated by:

$$F^{P_c} = Conv(\hat{F}^{P_c}) \| F_{sty}^{i} \tag{10}$$

where ‖ means connecting along the depth axis.

3.2 Generator

The primary focus of the decoder is to generate a new image by decoding codes. We finally take the texture code to generate a new person image. According to standard practice,the decoder generates the generated image I_g via N deconvolutional layers.

3.3 Discriminators

The main purpose of the discriminator is to promote the generator to generate a more realistic image by distinguishing the generated image from the real image. In the training process, we adopte pose discriminator D_P and texture discriminator D_t to identify the shape consistency and appearance consistency. The discriminators are implemented by Resnet Discriminator, each discriminator is independently trained, and all the discriminators can be analyzed and optimized separately.

3.4 Loss function

3.4.1 Adversarial Loss

The goal of adversarial loss is to guide the images generated by the generator to be close to the real images. This goal is achieved by the min-max confrontation process between the generator and the discriminator. The discriminator needs to maximize the probability of correctly determining the distribution of real images and false image. The task of the generator is to identify minimize the probability of the generated images being identified as false images, the two continue to fight, and ultimately achieve Nash equilibrium. In this paper, an adversarial loss function with D_P and D_t is used to help the generator optimize the generation parameters and synthesize the human body images in the target pose. The formula for adversarial loss in this paper is as follows:

$$\begin{aligned} \max_{G} \min_{D} L_{adv} = & E_{I_t, I_c, P_t} \left\{ \log \left[D_t(I_c, I_t) \cdot D_p(P_t, I_t) \right] \right\} \\ & + E_{I_t, I_g, P_t} \left\{ \log \left[\left(1 - D_t(I_c, I_g)\right) \cdot \left(1 - D_p(P_t, I_g)\right) \right] \right\} \end{aligned} \tag{11}$$

where $\log[D_t(I_c, I_t) \cdot D_p(P_t, I_t)]$ represents the probability that the discriminator will distinguish the real image as real data. $\log[(1 - D_t(I_c, I_g) \cdot (1 - D_p(P_t, I_g))]$ represents the probability that the discriminator will judge the generated image as a false image.

3.4.2 Reconstruction Loss

The goal of reconstruction loss is to improve the similarity between the original image and the generated images, avoid significant distortion of colors, and accelerate the convergence process. This paper uses L1 reconstruction loss to calculate the pixel difference between the generated source image $\hat{I}_c$ and the source image I_c. The formula is as follows:

$$L_{\text{pixel-rec}} = \| I_g - I_c \|_1 \tag{12}$$

3.4.3 Perceptual loss

Because we often use MSE loss function, the output images will be smoother (losing the details / high frequency part), so we can enhance the image details by choosing the perceptual loss function. The preceputal loss is computed as (Ma et al. 2018)：

$$L_{per} = \frac{1}{W_j H_j C_j} \sum_{x=1}^{W_j} \sum_{y=1}^{H_j} \sum_{z=1}^{C_j} \| \phi_j(I_g)_{x,y,z} \tag{13}$$

where ϕ_j is the output feature of the *j*-th layer in the VGG19 network, and W_j, H_j, C_j are the spatial width, height and depth of ϕ_j, respectively.

4 Experiments

4.1. Datasets and Details

In this paper, we use dataset DeepFashion(Liu et al. 2016)for performance evaluation. DeepFashion contains 52,712 images with the resolution of 256 × 256. Before training, we use Human Pose Estimator (HPE) to remove noisy images from the dataset in which human body can't be detected by HPE. Here we select 37,258 images for training and 12,000 images for testing. In particular, the test sets do not contain the person identities in the training sets in order to objectively evaluate the generalization ability of the network. In addition, we implement the proposed framework in Pytorch framework using two NVIDIA Quadro P4000 GPUs with 16GB memory. The generator contains 9 cascaded residual blocks. To optimize the network parameters, we adopt Rectified Adam(RAdam), which can not only have the advantages of Adam's fast convergence but also possess the advantages of SGD. We train our network for about 120k iterations. The learning rate is initially set 1×e-5 and linearly decayed to zero after 60k iterations. The batch size for DeepFashion is set 1. We alternatively train the generator and discriminator with the above configuration.

4.2 Metrics

Inception score (IS)(Barratt & Sharma 2018; Salimans et al. 2016) and Structure Similarity (SSIM)(Wang et al. 2004) are the most commonly used indicators to evaluate the quality of generated images. Inception score uses the Inception Net V3 network to evaluate the quality of the generated imagea from two aspects: image clarity and diversity. Structure Similarity is a perception-based calculation model that measures the similarity of two images from three

aspects: brightness, contrast, and structure. However, IS only rely on the generated image itself for judgment, ignoring the consistency between the generated imageit and the real image. What's more,based on this, Fréchet Inception Distance (FID)(Heusel et al. 2017)is adopted to measure the realism of the generated image. This method first converts both the generated image and the real image into a feature space, and then calculates the Wasserstein-2 distance between the two images. In addition to the above-mentioned objective evaluation indicators, a User Study was also conducted, and subjective indicators were formed by collecting volunteers' evaluation of the generated images.

4.3 Quantitative and qualitative comparison

Since the judgment of the generated image is more subjective, We compare our method with several stare-of-the-art methods including PATN(Zhu et al. 2019)、ADGAN(Men et al. 2020)、PISE(Zhang et al. 2021). The qualitative comparison results are shown in Figure 4. In terms of visual ef-fects, our method achieves excellent performance. Our method avoids a lot of noise, such as the images in the first line of the figure, and other methods appear white noise points on the clothes, but our generated image has no noise and perfectly presents the style of the clothes in the source image. In addition,our method shows better details than other methods in hair and face, and is closer to the real image. For more details, zoom in on Figure 4.

In order to verify the effectiveness of our method, we conducted experiments on four benchmarks. In order to get a more fair comparison, we reproduce PATN、ADGAN、PISE and test it with the test set in this paper. The results of comparison and the advantages of this method are clearly shown in Table 1. Our method is superior to other methods in SSIM and mask SSIM, which verifies the effectiveness of the Graph-based generative adversarial network and maintains the consistency of the structure in the pose conversion process. Although the IS value is slightly lower than that of ADGAN, the FID value is comparable, indicating that our generated images are very close to the real images.

4.4 User Study

Human subjective judgment is a very important indicator for generating images. This article relies on the questionnaire star website to do a difference test. In the experiment, 100 volunteers were asked to select the more realistic image from the generated images and the real images within one second. In order to ensure the confidence, following the rules in (Ma et al. 2018), we randomly select 55 real images and 55 generated images for out-of-order processing, and then pick out 10 of them for volunteer practice, and the remaining 100 for evaluation and judgment. Each image was compared 3 times by different volunteers. The results are shown in Table 2. The images generated by the method in this paper have achieved significant effectsremarkable results in human subjective evaluation.

R2G means the percentage of real images being rated as the generated w.r.t. all real images. G2R means the percentage of generated images rated as the real w.r.t. all generated images. The results of other methods are drawn from their papers.

4.5 Ablation Study

As shown in Figure 5 and Table 3, the evaluation results of different versions of our proposed method are shown. We first compare the results using appearance decomposi-tion to the results without using it. We remove the appearance decomposition part from the model, use the encoder similar to the PATN to encode the source image directly, and then transfer it to the generation network directly. By comparison, we find that the appearance decomposition module in our method can effectively improve the performance of the generator. It describes the spatial layout of the region level through the partition mapping, so as to guide the image generation with higher-level structural constraints. Then, we verify the role of graph-based global reasoning. In the pose pathway,we replace the graph-based global reasoning with the method used in (Zhu et al. 2019), which use the super position of convolution layer to expand the receptive field gradually for pose transfer. From the Table 3, the graph-based global reasoning module can get higher SSIM value, which shows that the module can improve the structural consistency of the image.In addition, we also verify the influence of each objective function on the generated results. It can be seen that adding these objective functions together can effectively improve the performance of the generator.

5 Conclusion

In this paper, a generation model based on appearance decomposition and graph-based global reasoning is proposed for pose guided image generation. The task of pose transfer is divided into pose path and appearance path. We use graph network for global reasoning and appearance decomposition for texture synthesis simultaneously. Through several comparative experiments on Deepfashion dataset, our model shows superior performance in terms of subjective visual authenticity and objective quantitative indicators.

References

Alqahtani H, Kavakli-Thorne M, and Liu CZ. 2019. An introduction to person re-identification with generative adversarial networks. *arXiv preprint arXiv:190405992*.

Balakrishnan G, Zhao A, Dalca AV, Durand F, and Guttag J. 2018. Synthesizing images of humans in unseen poses. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p 8340-8348.

Barratt S, and Sharma R. 2018. A note on the inception score. *arXiv preprint arXiv:180101973*.

Chen Y, Rohrbach M, Yan Z, Shuicheng Y, Feng J, and Kalantidis Y. 2019. Graph-based global reasoning networks. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p 433-442.

Cui Y, and Wang W. 2019. Colorless Video Rendering System via Generative Adversarial Networks. 2019 IEEE International Conference on Artificial Intelligence and Computer Applications (ICAICA): IEEE. p 464-467.

Dai P, Ji R, Wang H, Wu Q, and Huang Y. 2018. Cross-modality person re-identification with generative adversarial training. IJCAI. p 2.

Dong H, Liang X, Shen X, Wang B, Lai H, Zhu J, Hu Z, and Yin J. 2019. Towards multi-pose guided virtual try-on network. Proceedings of the IEEE/CVF International Conference on Computer Vision. p 9026-9035.

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

Heusel M, Ramsauer H, Unterthiner T, Nessler B, and Hochreiter S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *arXiv preprint arXiv:170608500*.

Honda S. 2019. Viton-gan: Virtual try-on image generator trained with adversarial loss. *arXiv preprint arXiv:191107926*.

Huang S, Xiong H, Cheng Z-Q, Wang Q, Zhou X, Wen B, Huan J, and Dou D. 2020. Generating Person Images with Appearance-aware Pose Stylizer. *arXiv preprint arXiv:200709077*.

Kingma DP, and Welling M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:13126114*.

Kipf TN, and Welling M. 2016. Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:160902907*.

Kubo S, Iwasawa Y, and Matsuo Y. 2018. Generative adversarial network-based virtual try-on with clothing region.

Lassner C, Pons-Moll G, and Gehler PV. 2017. A generative model of people in clothing. Proceedings of the IEEE International Conference on Computer Vision. p 853-862.

Liu J, Li W, Pei H, Wang Y, Qu F, Qu Y, and Chen Y. 2019. Identity preserving generative adversarial network for cross-domain person re-identification. *IEEE Access* 7:114021-114032.

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

Lv J, and Wang X. 2018. Cross-dataset person re-identification using similarity preserved generative adversarial networks. International Conference on Knowledge Science, Engineering and Management: Springer. p 171-183.

Ma L, Jia X, Sun Q, Schiele B, Tuytelaars T, and Van Gool L. 2017. Pose guided person image generation. *arXiv preprint arXiv:170509368*.

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

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

Rezende DJ, Mohamed S, and Wierstra D. 2014. Stochastic backpropagation and approximate inference in deep generative models. International conference on machine learning: PMLR. p 1278-1286.

Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, and Chen X. 2016. Improved techniques for training gans. *arXiv preprint arXiv:160603498*.

Si C, Wang W, Wang L, and Tan T. 2018. Multistage adversarial losses for pose-based human image synthesis. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p 118-126.

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

Sohn K, Lee H, and Yan X. 2015. Learning structured output representation using deep conditional generative models. *Advances in neural information processing systems* 28:3483-3491.

Song S, Zhang W, Liu J, and Mei T. 2019. Unsupervised person image generation with semantic parsing transformation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. p 2357-2366.

Tang H, Bai S, Zhang L, Torr PH, and Sebe N. 2020. Xinggan for person image generation. European Conference on Computer Vision: Springer. p 717-734.

Veli—ìkovi—ç P, Cucurull G, Casanova A, Romero A, Lio P, and Bengio Y. 2017. Graph attention networks. *arXiv preprint arXiv:171010903*.

Wang R, Huang C, and Wang X. 2020. Global relation reasoning graph convolutional networks for human pose estimation. *IEEE Access* 8:38472-38480.

Wang Z, Bovik AC, Sheikh HR, and Simoncelli EP. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13:600-612.

Wu X, Xu K, and Hall P. 2017. A survey of image synthesis and editing with generative adversarial networks. *Tsinghua Science and Technology* 22:660-674.

Xiong W, Luo W, Ma L, Liu W, and Luo J. 2018. Learning to generate time-lapse videos using multi-stage dynamic generative adversarial networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p 2364-2373.

Zanfir M, Popa A-I, Zanfir A, and Sminchisescu C. 2018. Human appearance transfer. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. p 5391-5399.

Zhang J, Li K, Lai Y-K, and Yang J. 2021. PISE: Person Image Synthesis and Editing with Decoupled GAN. *arXiv preprint arXiv:210304023*.

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

Table 1(on next page)

Quantitative comparisons with the state-of-the-art methods on DeepFashion.* denotes the results tested on our test set.

Table 1. Quantitative comparisons with the state-of-the-art methods on DeepFashion.* denotes the results tested on our test set.

Model	DeepFashion			
	IS	**SSIM**	**Mask-SSIM**	**FID**
PATN	3.209	0.774	/	/
ADGAN	3.364	0.772	/	/
PISE	/	/	/	13.61
*PATN	3.054	0.7748	0.9275	20.374
*ADGAN	3.196	0.7736	0.9267	13.457
*PISE	**3.233**	0.7776	0.9281	13.286
Ours	3.1825	**0.7916**	**0.9328**	**12.649**

Table 2(on next page)

User study results

1 **Table 2.**User study results

Indicator	**DeepFashion**			
	PATN	**ADGAN**	**Def-GAN**	**Ours**
R2G	19.14	23.49	12.42	22.84
G2R	31.78	38.67	24.61	**39.45**

2

Figure 1

The results of our method in the pose transfer task.

Figure 2

Illustration of our idea.

The red and blue nodes represent the source pose and the target pose respectively (not all the key points are shown in the figure for convenience). The nodes are mapped from the original space to the interactive space to form a connected graph for reasoning. Then the nodes are projected back to the original space for further processing.

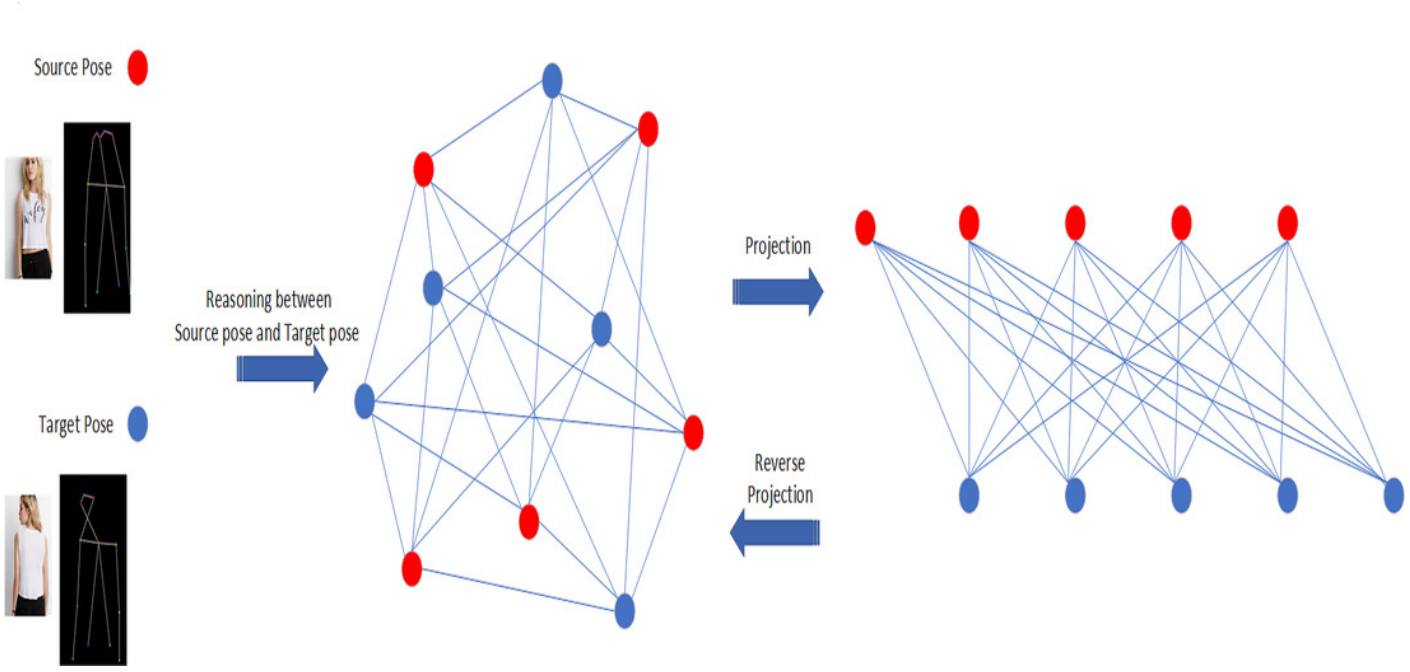

Figure 3

Structure of our proposed method

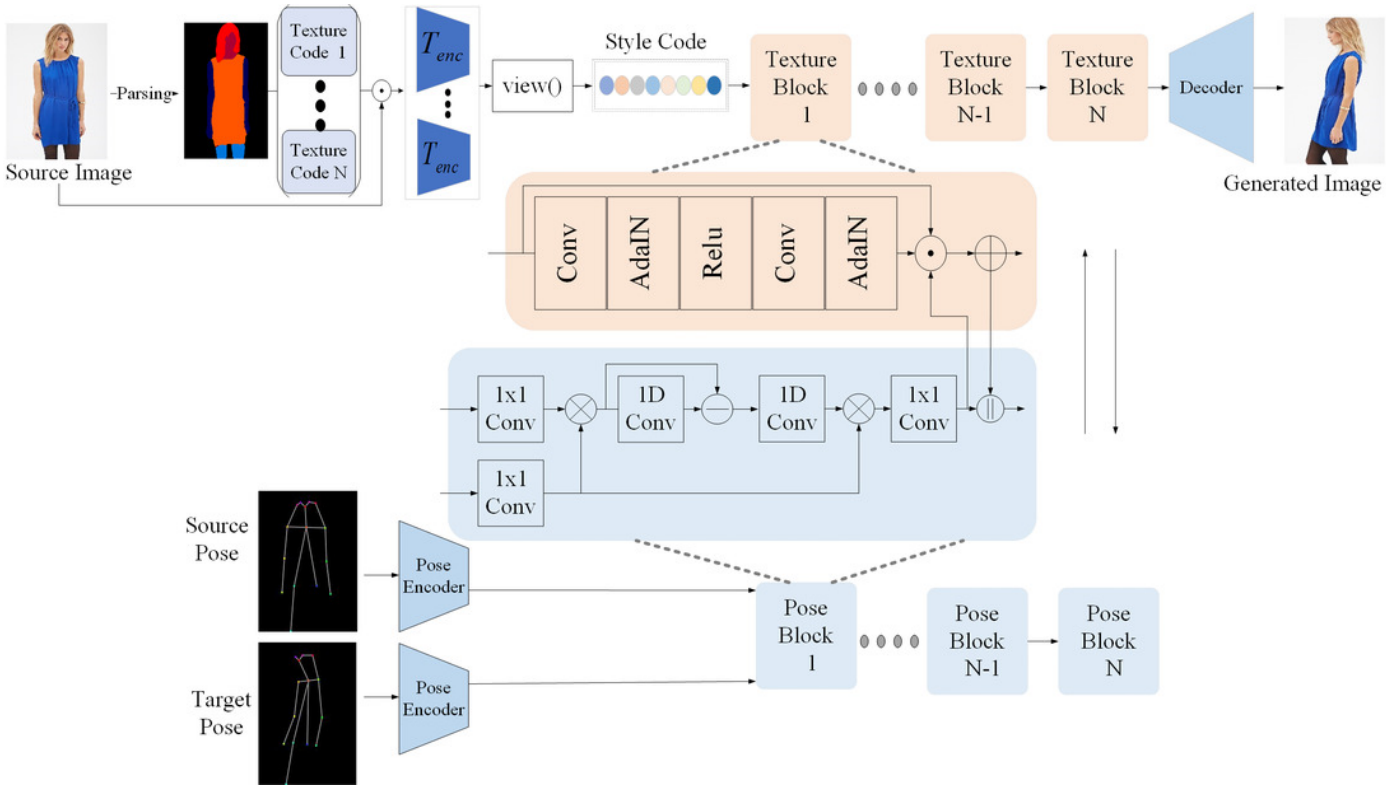

Figure 4

Qualitative comparison with the state-of-the-art methods on DeepFashion.Our results are shown in the last column.

Figure 5

The qualitative results of ablation study.

Table 3(on next page)

The evaluation results of ablation study.

Table 3. The evaluation results of ablation study.

Model	DeepFashion			
	IS	**SSIM**	**Mask-SSIM**	**FID**
w/o decomposition	3.128	0.781	0.930	14.862
w/o graph reasoning	3.025	0.778	0.929	17.306
w/o L_{adv}	3.168	0.776	0.932	13.394
w/o $L_{\text{pixel-rec}}$	3.164	0.774	0.931	12.672
w/o L_{per}	3.178	0.785	**0.933**	14.862
Full	**3.183**	**0.7916**	**0.933**	**12.649**