Pose Transfer using Multiple Input Images

Anonymous PRCV 2023 submission #489

Abstract. One key challenge of pose transfer lies in its large variation and occlusion. Existing methods are basically aimed at the migration of the pose in a single input image. So these methods have difficulties predicting reasonable invisible re-gions and fail to decouple the shape and style of clothing. Although pose transfer method using single input image can generate the results with correct structure, it cannot keep the original details of the image. To solve this problem, we propose a two-stage generative model for pose transfer based on multiple input images. First, the Feature Extraction Stage, then Reposing Stage. In feature extraction stage, we extract relevant features from each input image. Then, in reposing stage, we propose a pose-conditioned transformer-based StyleGAN generator, adding residual module and fusion module at different levels of the generator. In this way, we can get the most relevant features from each input image for weighted fusion, thus improving the quality of the results. We show that our method compares favorably against those using single input image in both quantitative evaluation and visual comparison.

Keywords: Pose Transfer · Pose-guided Person Image Synthesis · Muti-source Image Generation

1 Introduction

Human pose transfer aims to synthesize a new image for the same person in a target pose, which is an active topic in person image synthesis. Human pose transfer has great application potentials in photography, image editing, video generation, virtual try-on, etc.

Recently, Generative Adversarial Networks (GANs) are widely used in image synthesis, achieving great success in human pose transfer. Many methods [16] directly learn mapping the from the original image use GANs to generate the image transformed into the target pose. These methods usually input one source image, using a two-branch (source branch and target branch) framework to transfer the feature of the source image from the source pose to the target pose. However, if only the features of single input images are used, the occluded and invisible areas will lack feature information. When the source pose and the target pose have large differences, the pose transfer method using single input image is difficult to predict a sharp and reasonable image with sparse correspondences. To deal with this problem, attention-based methods are used to provide richer appearance information. This data redundancy can possibly be exploited by the generator in order to compensate for partial occlusions, self-occlusions

or noise in the source images. Although such methods can generate results with correct structure, they cannot keep the original details and need to be supplemented by fine-tuning during testing. These methods also cannot disentangle the shape and style information (e.g., the category and texture of clothing) and fail to preserve spatial context relationships. Using person-specific abundant data the quality of the generated images can be potentially improved. For example, a training dataset specific to each target person can be recorded. Another solution is to build a full-3D model of the target person. However, these approaches lack of flexibility and need an expensive data-collection.

These human pose transfer methods using single-source image encounter the following challenges to synthesize satisfactory images: 1) lack of sufficient feature information 2) the coupling of the shape and style of clothing, 3) potential uncertainties in invisible regions, and 4) loss of spatial context relationships.

In this work we propose a new method using multiple input image to address these challenges [17]. Unlike the single input method, our method provides a set of M ($M \geq 1$) source images $\boldsymbol{X} = \{x^i\}_{i=1...M}$. These images depict the same person with the same overall appearance (e.g., the same clothes, haircut, etc.), but the shooting angle and human pose in these picture are different. Our method consists in generating a new image with an appearance similar to the general appearance pattern represented in $\boldsymbol{X}$ but in the target pose p^{target} . Our method consists of two stages: feature extraction stage and reposing stage. In feature extraction stage, we propose a feature extraction network, which takes the input image, the attitude of the input image and the target attitude as the input, to extract multi-scale features, initial transformation, occlusion map and attention map of each input image. In reposing stage, we propose a pose-conditioned transformer-based StyleGAN generator [3, 19], adding residual module and fusion module at different levels of the generator. We add the initial transformation from the feature extraction network and the residual transformation from the residual module to get the optimized transformation. At the same time, we use the attention map from the feature extraction network to fuse the features of different input images. We repeat these operations at different levels to get the final result of pose transfer.

The main contributions of our work are summarized as follows:

- We propose a human pose transfer method using multiple input images. After obtaining the rough transformation from source to target by using neural network, the method reduces the dislocation problem between different sources through cascading residual modules and fusion modules in the generator, while retaining the details in the original image to a considerable extent.
- We propose a feature extraction network to extract multi-scale features, multi-scale initial transformation, occlusion map and attention map of each input image, making up for the lack of image information in the method using single input image.

– We design a pose-conditioned transformer-based StyleGAN generator, adding residual module and fusion module at different levels of the generator to generate high-quality images.

2 Related Work

2.1 Pose-based Person Image Generation

Ma *et al.* [12] designed a two-stage architecture to generate a realistic-looking image with target pose. This is the first attempt of pose-based person image synthesis. Further more, Ma *et al.* decomposed a person image into foreground, background and pose then used the image synthesis network to learn the respective latent features. Esser *et al.* [6] proposed a variational U-Net to disentangle the appearance and pose of portrait images. But using the common U-Net skip connections can not well handle the spatial deformations of images. Zhu *et al.* [21] employed a series of progressive pose attention blocks to transform the image. His blocks can steadily attend to different body regions to generate person images. Liu *et al.* [11] and Li *et al.* [10] warp the inputs at the feature level. Ren *et al.* [15] pointed out that Convolutional Neural Networks are limited by the lack of ability to spatially transform the inputs. So they designed a global flow local attention framework which can reasonably sample and to reassemble the inputs at the feature level. Several proposals use the UV parameterization to transfer pixel values or local features to the target pose. Recently, some pose-conditioned StyleGAN networks have been proposed. The main idea is to control the target appearance using the appearance features extracted from the source image or pose-independent UV texture to modulate the latent space. Albahar use a *spatiallyvarying* modulation for improved local detail transfer instead of *global* latent feature modulation. In addition, they train a coordinate inpainting network for completing partial correspondence field using a human body symmetry prior.

Our method builds upon pose-conditioned transformer-based StyleGAN, adding residual modules and fusion modules at different levels of its generator. In this way, we can obtain the features of multiple input images to complement the deficiencies of single input image.

2.2 Multi-source Image Generation

Sun *et al.* [18] propose a multi-source image generation approach whose goal is to generate a new image according to a target-camera position. Note that this task is different from what we address in this paper, since a human pose describes an articulated object by means of a set of joint locations, while a camera position describes a viewpoint change but does not deal with source-to-target object deformations. Stéphane Lathuilière *et al.* [8] propose a method based on a local attention mechanism which selects relevant information from different source image regions. Compared with this method, our method not only

uses attention mechanism to obtain the features of different input images, but also adds residual module and fusion module to fuse these features better. At the same time, our method focuses on the transformation features of the image, which ensures that high-quality results can be obtained even when the source pose and target pose differ greatly.

3 Our Method

In this section, we will specifically introduce our method for pose transfer based on multiple input images, mainly including feature extraction stage and reposing stage, which is shown in Fig. 1.

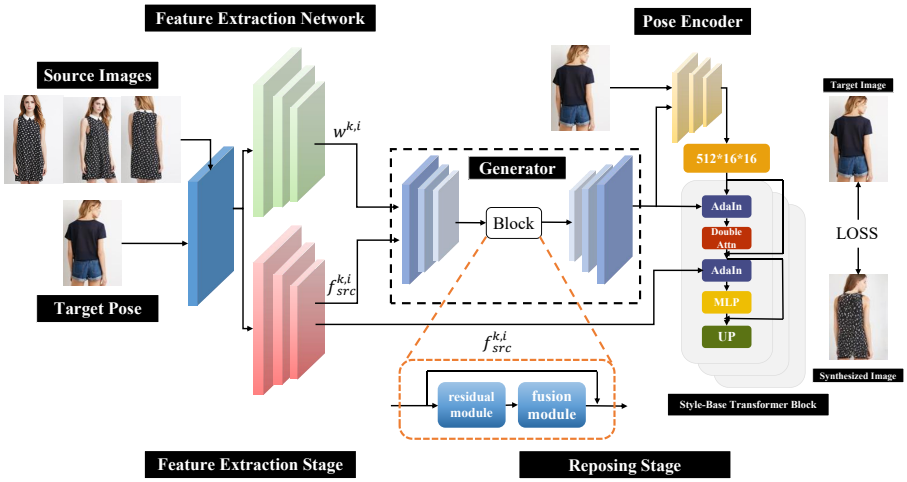

Fig. 1. Method overview. In our Human Pose Transfer method, we propose a two-stage network including feature extraction stage and reposing stage. In feature extraction stage, we extract the multi-scale features of each input image and the initial transformation, occlusion map and attention map of each input image at multi-scale. In reposing stage, we propose a pose-conditioned transformer-based StyleGAN generator, adding residual module and fusion module at different levels of the generator.

3.1 Feature Extraction Stage

The feature extraction stage includes image feature extraction network and transform feature extraction network. Image feature extraction network extracts multi-scale features of each input image. Then we put the input image, the attitude of the input image and the target attitude into the trained transform feature extraction network, and extract the initial transformation, occlusion map and

attention map of each input image at multiple scales. The initial transformation is responsible for roughly transforming the input image features to the target pose. The occlusion map encodes the visible and invisible areas. The attention map selects more important inputs from different input sources. Due to the limitation of the article, we will display the specific details of the stage in our appendix.

3.2 Reposing Stage

We show an overview of our reposing stage in Fig. 2. We propose a pose-conditioned transformer-based StyleGAN generator and add residual module and fusion module at different levels of the generator. In this stage, we perform weighted fusion on the features of different input images obtained by feature extraction network to finally obtain the target image. Due to the limitation of the article, in addition to the content of this section, we will display the specific details of the stage in our appendix.

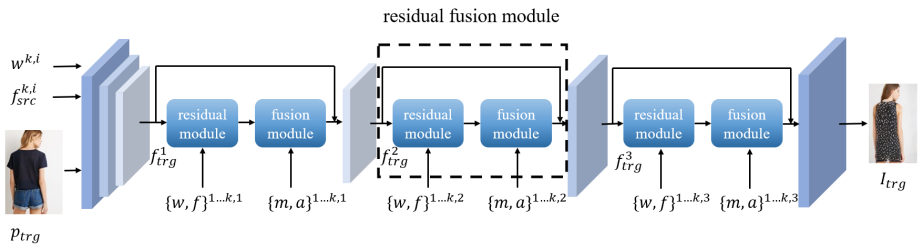

Fig. 2. Overview of reposing stage. We propose a pose-conditioned transformer-based StyleGAN generator and add residual module and fusion module at different levels of the generator.

Pose Encoder We use pose-Encoder E_{pose} to encode the target pose P_{trg} represented by improved generator and produce the pose feature E. It is a fully convolutional network, which is composed of several residual and downsampling blocks. This is a simple and flexible encoder.

$$E = E_{pose}(P_{trg}) \tag{1}$$

Residual Prediction Through the residual module, we can get the optimal transformation from the source pose to the target pose. According to this optimal transformation, we transform the features of each input image to reduce the dislocation problem between different inputs. We use differentiable bilinear sampling W to transform image features $f_{src}^{k,i}$ with the initial transformation $w^{k,i}$ obtained from feature extraction.

$$f_{w}^{k,i} = W(f_{src}^{k,i}, w^{k,i}) \tag{2}$$

Based on the features $f_w^{k,i}$ after initial transformation and target features f_{trg}^i, we estimate the residual transformation $r^{k,i}$ through the residual estimation network R^i.

$$r^{k,i} = R^i(f_{trg}^i, f_w^{k,i}) \tag{3}$$

Finally, we add the residual transformation $r^{k,i}$ and the original transformation $w^{k,i}$ to improve the accuracy of the transformation from the source to the target and the alignment after the transformation from different sources.

$$\hat{f}_w^{k,i} = W(f_{src}^{k,i}, w^{k,i} + r^{k,i}) \tag{4}$$

Feature Fusion We use the feature fusion module to fuse the most relevant features obtained from each input image with weight to obtain the final target feature. After the residual transformation prediction, all the input features are transformed to the target attitude. We repair each input feature and the target feature f_{trg}^i of the current layer according to the occlusion map, and sum the different input features according to the attention map. The sum result is used as the input target feature f_{trg}^{i+1} of the next level.

$$f_{trg}^{i+1} = G^i(\sum_{i=1}^{K} a^{k,i} \otimes (\hat{f}_w^{k,i} \otimes (1 - m^{k,i}) + f_{trg}^i \otimes m^{k,i})) \tag{5}$$

$\otimes$ represents pixel by pixel multiplication, G^i is the layer i of the generator, and f_{trg}^i is the target feature of the layer i. $a^{k,i}$ are attention maps, normalized by softmax function, including

$$\sum_{i=1}^{K} a^{k,i} = 1 \tag{6}$$

Pose-conditioned Transformer-based StyleGAN The original StyleGAN [13] takes a constant tensor as input on which convolutions are performed. A separate latent noise vector that controls the generated image is passed through a mapping network, and its output w is used to modulate the weights of the style-blocks (In original StyleGAN, it is the convolution layers). We change the constant input and use the tensor E of the dimensions $16 \times 16 \times 512$ which encoded by Pose-Encoder E_{pose} as the input to the style-blocks. And the style-block weights [2] are demodulated using the the appearance encoding Z to finally reconstruct the result image $\hat{I}_{trg}$. The original style-block consists of several 3×3 convolution layers, upsample layers and Adaptive Instance Normalization blocks. But Convolutional Neural Networks (CNNs) are inefficient to spatially transform the inputs. Because CNNs calculate the outputs with a particular form of parameter sharing. This mechanism results to the equivariance to transformation. And, attention mechanism makes nerworks to take use of non-local information, which gives networks abilities to build long-term correlations. It has been proved to be efficient in image generation task. Also, a few works attempt to propose transformer-based GANs. *StyleSwin* has achieved success on high-resolution

synthesis using transformers. Inspired by this work [5], we also adopt transformer as the basic sytle block which computes the multi-head self-attention locally. Instead of using a 3×3 Convolution kernels, we use *doubleattention* which allows a single transformer block to simultaneously attend to the context of the local and shifted windows.

3.3 Training Losses

In addition to StyleGAN adversarial loss L_{adv}, we train our model to minimize the following reconstruction losses:

ℓ_1 **loss** We minimize the ℓ_1 loss foreground human regions of the synthesized image $\hat{I}_{trg}$ and of the ground truth target I_{trg}. The reconstruction ℓ_1 loss is written as

$$L_{\ell_1} = \parallel \hat{I}_{trg} \cdot M_{trg} - I_{trg} \cdot M_{trg} \parallel \tag{7}$$

where M_{trg} is the human foreground mask.

Perceptual loss The perceptual loss calculates ℓ_1 distance between activation maps of a pre-trained network of the synthesized image $\hat{I}_{trg}$ and of the ground truth target I_{trg}. It can be written as

$$L_{perc} = \Sigma \parallel \phi(\hat{I}_{trg}) - \phi(I_{trg}) \parallel_1 \tag{8}$$

where ϕ is the activation map of the middle layer of a pre-trained network. In our network, we use the VGG16 model, and calculate the output feature maps of the middle layers, and then compare the difference to get the loss.

Style loss the style loss calculates the statistic error between the activation maps. It can be written as

$$L_{style} = \Sigma \parallel G^{\phi}(\hat{I}_{trg}) - G^{\phi}(I_{trg}) \parallel_1 \tag{9}$$

Face identity loss We use MTCNN to detect, crop and align faces from the generated image $\hat{I}_{trg}$ and ground truth target I_{trg}. When a face is detected, we maximize the cosine similarity between the pretrained SphereFace features of the generated face and the ground truth target face. It can be written as

$$L_{face} = 1 - \left(\frac{SF\left(\hat{I}_{trg}\right)^{\top} SF\left(I_{trg}\right)}{\max\left(\left\| SF\left(\hat{I}_{trg}\right)\right\|_2 \cdot \left\| SF\left(I_{trg}\right)\right\|_2, \epsilon\right)} \right) \tag{10}$$

Parsing loss Like attention map and occlusion map, we need to draw the outline of the subject first, and then consider the texture details. So we propose *parsingloss*, it can be written as

$$L_{parsing} = \parallel P(\hat{I}_{trg}) - P(I_{trg}) \parallel_1 \tag{11}$$

Therefore, our final loss is:

$$L_{total} = \ell_1 + L_{perc} + L_{style} + L_{face} + L_{parsing} + L_{adv} \tag{12}$$

4 Experiments

4.1 Network Implementation and Training Details

Implementation details We implement our model with PyTorch. We use ADAM optimizer with a learning rate of $ratio \cdot 0.002$ and beta parameters $(0, 0.99^{ratio})$. We set the generator ratio to $\frac{4}{5}$ and discriminator ratio to $\frac{14}{15}$. **Training** We train our model using 512×512 images for the extend-pose datasets. When resize the images, we first set the image height resize to 512. We then fill or crop the image width to 512. We first train our model by focusing on generating the foreground. We apply the reconstruction loss and the adversarial loss only on the foreground. We set the batch size to 1 and train for 50 epochs. We use 6 Style-Blocks in our reposing model, each block has 1 or 2 transformer-based blocks.

4.2 Comparison with previous work

Quantitative Comparison In Table 1 we show quantitative comparisons with both state-of-the-art single-source methods and multi-source methods. We use the DeepFashion dataset for evaluation and comparison. We split the datasets with the same method as that of [15].

Table 1. Quantitative comparison with the methods using single input image on the DeepFashion dataset

Method	Input Number	PSNR	SSIM	FID	LPIPS
PATN [22]	1	17.2156	0.7134	21.8568	0.195
GFLA [15]	1	17.7133	0.7541	18.8519	0.175
VU-Net [7]	1	15.1275	0.6142	23.6736	0.264
Posewithstyle [1]	1	18.1428	0.7651	9.0002	0.143
DiOr [4]	1	18.1428	0.7651	13.1077	0.229
NTED [14]	1	18.1428	0.7651	8.6838	0.173
CASD [20]	1	18.1428	0.7651	11.3732	0.194
Ours	1	**18.1821**	**0.7683**	**7.4804**	**0.134**
Attention-based U-Net [9]	2	18.2856	0.7664	7.4742	0.129
	3	18.2910	0.7691	7.4613	0.124
	5	18.3063	0.7732	7.4446	0.121
	7	18.3207	0.7786	7.4318	0.120
	10	18.3261	0.7869	7.4279	0.118
Ours	2	**18.4005**	**0.7693**	**7.4532**	**0.126**
	3	**18.4161**	**0.7742**	**7.4210**	**0.122**
	5	**18.4215**	**0.7784**	**7.4008**	**0.120**
	7	**18.5142**	**0.7833**	**7.3974**	**0.118**
	10	**18.5671**	**0.8015**	**7.3625**	**0.115**

We report the human foreground peak signal-to-noise ratio (PSNR), structural similarity index measure (SSIM), learned perceptual image patch similarity (LPIPS) and Frechet Inception Distance (FID). Actually, PSNR/SSIM often do not correlate well with perceived quality, particularly for synthesis tasks. And, we report these metrics only for completeness. We compared some of human pose transfer methods using single input image such as VU-Net [7], PATN [22],

GFLA [15], Posewithstyle [1], Dior [4], NTED [14], and CASD [20]. In terms of multi-source methods, we chose Attention-based U-Net [9] as the comparative object.

Concerning the comparison with single-source methods, our method reports the highest performance according to both the PSNR and the SSIM metrics and the lowest performance according to both the FID and the LPIPS metrics on the DeepFashion dataset when we use 10 source images. When we employ fewer images, our method still perform better compared with methods using single input images.

Qualitative Comparison Fig.3 shows some images obtained using the DeepFashion dataset. We compare our results with the images generated by

Fig. 3. A qualitative comparison with single-input image. The first column shows the source images. The second column is a single input image and the target poses are given in the third column. In column 4, we compared the generation results.

two single-source methods and one muti-source method. The upper half of Fig. 3 uses two input images, while the lower half uses three input images. The mul-

tiple input images are shown in the first column. The second column is a single input image and the target pose is in the third column. In the the last column, we compared the generation results of GFLA, Posewithstyle and Attention-based U-Net, two pose transfer methods using single input image and one using multiple input images with our method. We present more results in our appendix.

The qualitative results confirm the quantitative evaluation since we clearly obtain better images when we increase the number of source images. The images become sharper and with more details and contain less artifacts.

Our method can achieve more realistic effort in the synthesis of face, clothing texture and limbs. As shown in figure 4, the image synthesized by our approach achieves better effectiveness on the sleeve of clothes and the human face.

Fig. 4. The detail comparison of our method (Right) with Posewithstyle (Left).

Concerning the comparison with the methods using single input image, we observe that our model better preserves the details of the source images. In general, we obtain higher-quality details and less artefacts.

4.3 Ablation Study

To validate the effectiveness of the proposed components, we conduct ablation studies. We first describe the compared methods, obtained by amputating important parts of our method.

- No Residual: In this baseline version of our method, we use generator without residual moudle.
- No Fusion: We use the same generator of our method without adding fusion moudle.
- Baseline: We use the baseline architecture instead of adding the transformer-base double attention component.
- Full: This is the full-pipeline as described in our method.

Table 2 shows a quantitative evaluation. First, we notice that our method without residual moudle performs poorly on both datasets. This is particularly evident with the SSIM scores. This confirms the importance of source-target alignment.

When using only two source images, No residual, No fusion, Baseline and Full perform similarly to each other on the DeepFashion dataset. However, when we dispose of more source images we clearly observe the benefit of using our proposed approach.

Table 2. Quantitative ablation study on the DeepFashion dataset

Model Configuration	Input Number	PSNR	SSIM	FID	LPIPS
Single input	1	18.1821	0.7683	7.4804	0.134
Full	2	**18.4005**	**0.7693**	**7.4532**	**0.126**
Baseline	2	18.3698	0.7521	7.4736	0.129
No Fusion	2	18.3592	0.7486	7.4775	0.131
No Residual	2	18.3561	0.7443	7.4829	0.133
Full	5	**18.4215**	**0.7784**	**7.4008**	**0.120**
Baseline	5	18.7748	0.7542	7.4385	0.124
No Fusion	5	18.7736	0.7517	7.4521	0.127
No Residual	5	18.7624	0.7361	7.4639	0.132

5 Conclusion

To solve the problems of single-source pose transfer method, we propose a human pose transfer method using multiple input image. Specifically, we transfer human pose conditioned on a target pose and a set $\boldsymbol{X}$ of source images. This makes it possible to exploit multiple and possibly complementary images. We propose a two-stage network including feature extraction stage and reposing stage. Our core technical novelties lie in 1) feature extraction network: extraction of image features and preliminary transformation of different input images at multi-scale 2) residual and fusion moudle: reducing dislocation between different sources and retraining details in the original drawing 3) reposing model: spatial modulation of a pose-conditioned transformer-based StyleGAN generator. This method is an important supplement to the single input pose transfer method, and can improve the robustness of pose transfer. It can be shown that our method is capable of synthesizing realistic-image in the reference target pose.

References

1. Albahar, B., Lu, J., Yang, J., Shu, Z., Shechtman, E., Huang, J.B.: Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. ACM Transactions on Graphics (TOG) **40**(6), 1–11 (2021)
2. Andressa A. Bertolazzo Andressa A. BertolazzoDepartment of Chemistry, The University of Utah, S.E.S.L.C.U.U.S.b.A.A.B., Bertolazzo, A.A., Bertolazzo, M.B.A.A., Debdas Dhabal Debdas DhabalDepartment of Chemistry, The University of Utah, S.E.S.L.C.U.U.S.b.D.D., Dhabal, D., Dhabal, M.B.D., Valeria Molinero * Valeria MolineroDepartment of Chemistry, The University of Utah, S.E.S.L.C.U.U.S.e.b.V.M., Molinero, V., Emailprotected, E., emailprotected: Polymorph selection in zeolite synthesis occurs after nucleation. Journal of physical chemistry letters (13-4) (2022)
3. Chen, Y., Shen, C., Wei, X.S., Liu, L., Yang, J.: Adversarial posenet: A structure-aware convolutional network for human pose estimation. In: IEEE Computer Society (2017)
4. Cui, A., Mckee, D., Lazebnik, S.: Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing (2021)
5. Cui, Y., Li, M., Zhu, N., Cheng, Y., Su, S.L., Chen, J., Gao, Y., Zhao, J.: Bi-based visible light-driven nano-photocatalyst: the design, synthesis, and its application in pollutant governance and energy development. Nano Today **43**, 101432– (2022)

6. Esser, P., Sutter, E., Ommer, B.: A variational u-net for conditional appearance and shape generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8857–8866 (2018)
7. Esser, P., Sutter, E., Ommer, B.: A variational u-net for conditional appearance and shape generation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8857–8866 (2018)
8. LathuilièÍre, S., Sangineto, E., Siarohin, A., Sebe, N.: Attention-based fusion for multi-source human image generation. In: Workshop on Applications of Computer Vision (2019)
9. LathuilièÍre, S., Sangineto, E., Siarohin, A., Sebe, N.: Attention-based fusion for multi-source human image generation (2019)
10. Li, Y., Huang, C., Loy, C.C.: Dense intrinsic appearance flow for human pose transfer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3693–3702 (2019)
11. Liu, W., Piao, Z., Min, J., Luo, W., Ma, L., Gao, S.: Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5904–5913 (2019)
12. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Gool, L.V.: Pose guided person image generation. CoRR **abs/1705.09368**, 406–416 (2017)
13. Messelmani, T., Morisseau, L., Sakai, Y., Legallais, C., Goff, A.L., Leclerc, E., Jellali, R.: Liver organ-on-chip models for toxicity studies and risk assessment. Lab on a Chip **22** (2022)
14. Ren, Y., Fan, X., Li, G., Liu, S., Li, T.H.: Neural texture extraction and distribution for controllable person image synthesis. arXiv e-prints (2022)
15. Ren, Y., Yu, X., Chen, J., Li, T.H., Li, G.: Deep image spatial transformation for person image generation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7690–7699 (2020)
16. Siarohin, A., Lathuiliere, S., Sangineto, E., Sebe, N.: Appearance and pose-conditioned human image generation using deformable gans. IEEE Transactions on Pattern Analysis and Machine Intelligence (4) (2021)
17. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv e-prints (2014)
18. Sun, S.H., Huh, M., Liao, Y.H., Zhang, N., Lim, J.J.: Multi-view to novel view: Synthesizing novel views with self-learned confidence. In: European Conference on Computer Vision (2018)
19. Zhang, J., Siarohin, A., Tang, H., Chen, J., Sangineto, E., Wang, W., Sebe, N.: Controllable person image synthesis with spatially-adaptive warped normalization (2021)
20. Zhou, X., Yin, M., Chen, X., Sun, L., Gao, C., Li, Q.: Cross attention based style distribution for controllable person image synthesis (2022)
21. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: 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)
22. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: 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)