Pattern Recognition 165 (2025) 111626

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

FaTNET: Feature-alignment transformer network for human pose transfer

Yu Luo [a], Chengzhi Yuan [a], Lin Gao [b], Weiwei Xu [c], Xiaosong Yang [d], Pengjie Wang [a,d,*]

[a] Dalian Minzu University, No. 18 Liaohe West Road, Dalian 116000, China
[b] Institute of Computing Technology Chinese Academy of Sciences, No. 6 Kexueyuan South Road, Beijing 100080, China
[c] Zhejiang University, No. 866 Yuhangtang Road, Hangzhou 310058, China
[d] Bournemouth University, Fern Barrow, Poole, Dorset, BH12 5BB, United Kingdom

ARTICLE INFO

Keywords:
People image generation
Human pose transfer
Generative adversarial network
Transformers

ABSTRACT

Pose-guided person image generation involves converting an image of a person from a source pose to a target pose. This task presents significant challenges due to the extensive variability and occlusion. Existing methods heavily rely on CNN-based architectures, which are constrained by their local receptive fields and often struggle to preserve the details of style and shape. To address this problem, we propose a novel framework for human pose transfer with transformers, which can employ global dependencies and keep local features as well. The proposed framework consists of transformer encoder, feature alignment network and transformer synthetic network, enabling the generation of realistic person images with desired poses. The core idea of our framework is to obtain a novel prior image aligned with the target image through the feature alignment network in the embedded and disentangled feature space, and then synthesize the final fine image through the transformer synthetic network by recurrently warping the result of previous stage with the correlation matrix between aligned features and source images. In contrast to previous convolution and non-local methods, ours can employ the global receptive field and preserve detail features as well. The results of qualitative and quantitative experiments demonstrate the superiority of our model in human pose transfer.

1. Introduction

Pose-guided person image generation aims to generate photo-realistic person images using a person image and several desired poses, which has a wide range of applications in person re-identification [1], image processing [2], and video generation [3]. It is a very challenging problem due to huge spatial deformation and occlusion of characters.

Recently, Generative Adversarial Network (GAN) [4] has been successfully applied in human pose transfer. State-of-the-art human pose transfer [5–7] methods are dominated by convolutional architectures. They can be divided into two categories: direct deformation-based methods and flow/transformation-based methods. Direct deformation methods [8–12] often adopt an encoder–decoder CNN architecture, and introduce an attention module to achieve deformation task. On the other hand, flow/transformation-based methods [13,14] predict an appearance flow or transformation matrix to guide the image generation. They often warp the source image and its feature to the target pose to obtain an appearance flow or learn feature-level mapping by utilizing segmentation maps for guidance.

Despite the performance improvement achieved by previous methods, state-of-the-art methods suffer from two problems because of their convolutional architectures. *First*, Convolutional Neural Networks (CNNs) extract features in local sliding windows. They are unable to process long range dependencies, unless using very deep convolutional layers, which leads to the loss of feature resolution and intricate details. If the CNN-based encoder cannot retain more details, the incomplete features will also affect the decoding, especially in spatial transformation tasks. Thus, vanilla CNN-based models are inadequate for effectively capturing the crucial global contexts. Some arts [10,11] can obtain global context to some extent by introducing the non-local modules into CNN-based architecture. However, their query, key and value focus on different domains, resulting in low efficiency of space conversion and fuzzy results. As shown in third (PoNA) and fourth (XingGAN) columns of Fig. 1. By contrast, our transformer based method can achieve high quality results, as shown in the second column of Fig. 1.

Second, most methods encode the style image and the pose image into a latent vector, and then the network synthesizes images according to the latent vector. However, the latent code characterizes the semantic information of the image, and ignores the style feature. Consequently, some local style features are lost in the final image.

* Corresponding author at: Dalian Minzu University, No. 18 Liaohe West Road, Dalian 116000, China.
E-mail addresses: yuluo19980126@163.com (Y. Luo), orangecircle128@gmail.com (C. Yuan), gaolin@ict.ac.cn (L. Gao), xww@cad.zju.edu.cn (W. Xu), xyang@bournemouth.ac.uk (X. Yang), pengjiewang@gmail.com (P. Wang).

https://doi.org/10.1016/j.patcog.2025.111626
Received 26 August 2023; Received in revised form 21 January 2025; Accepted 19 March 2025
Available online 5 April 2025

Fig. 1. Comparing of our method with the non-local based PoNA [11] and XingGAN [10] methods. While all methods establish long-range dependencies, our proposed method achieves superior results in terms of quality.

Fig. 2. The results of feature alignment network, demonstrating alignment of the source image with the target image.

To address both problems, in this paper, we propose Feature-alignment Transformer Network (FaTNET), a transformer based person image generation framework, which consists of transformer encoder, feature alignment network and transformer synthetic network. In contrast to previous convolution and non-local methods, ours can employ the global receptive field and preserve detail features as well. We further propose feature alignment to obtain a prior image (aligned image) to prevent style features from being washed away, as shown in Fig. 2. Specifically, we first compute the pose guided matrix, then obtain the prior image by multiplying the source image with it. Since the prior image is directly warped from the source image through the pose-guided matrix, it retains the blurred style features. And our network is able to generate further image details based on it. Last but not least, we introduce convolutional layers into our transformer based framework, considering that these layers can bring strong prior of inductive bias, and make our network generalize with faster converging speed.

Specifically, our method has three steps, as shown in Fig. 3. First, we propose an encoder by introducing Swin Transformer [15] followed by depthwise convolution to extract features from source images and target pose images. Second, we propose feature alignment network with multiple cascading blocks to align source style images with target style images. Specifically, in each alignment block, we use transformer net to maintain the global relationship between tokenized features from the target pose and the source image, and guide the source features to match the target features. We also use convolutional networks to focus on local information and preserve detailed features. Benefiting from the feature alignment network, the shape and style of the image are disentangled, and a preliminary alignment feature map is obtained. Third, we propose transformer synthetic network to pay attention to correlated features from exemplar, in order to recover the fine details and predict the invisible area during decoding. Experimental results show that our method outperforms state-of-the-art methods.

The main contributions of this work are summarized as follows:

- We propose a feature-alignment pose transfer framework with multiple cascading blocks, which can align the source image with the target image in the embedded and disentangled feature space.
- We propose transformer synthetic network to maintain the features from corresponding regions of the exemplar, and generate high-quality images recurrently by warping previous stage's result with the correlation matrix between aligned features and source images.
- We incorporate transformer and convolution into our framework in an exquisite way, which can employ the global dependencies and preserve detail features as well. Experimental results on human pose transfer show the flexibility and superior performance of our person image generation method.

2. Related works

2.1. Generative Adversarial Networks

The Generative Adversarial Network (GAN) [4] consists of a generator and a discriminator, which produce realistic images through adversarial training. Since their inception, GANs have rapidly adopted a fully convolutional architecture and have been effectively utilized in image-to-image translation [16,17], image enhancement [18,19], and image editing [20]. Recent studies have integrated the transformer module into image generation models by substituting certain components of CNNs. [21] introduced the self-attention mechanism into the GAN architecture, where the self-attention module complements convolutions and aids in modeling long-range, multi-level dependencies across image regions. Recent work [22] employed a convolutional GAN to develop a codebook of image constituents and efficiently modeled their composition with transformers within high-resolution images. There is also one work [23] that completely removed convolutions from their generative framework and only used two transformers to make one strong GAN.

2.2. Human pose transfer

An early method [24] on human pose transfer proposed a two-stage network to generate the image with the target pose. This method combined the source image, source pose, and target pose as inputs to progressively generate the target image from coarse to fine. Zhu et al. [8] proposed a progressive attention model to transfer the source image. However, it lost useful information during multiple transfers by using local attention mechanism. Some methods adopted non-local attention mechanism and used keypoints as their guidance for pose representation. Tang et al. [10] employed two generation branches to separately model the person's appearance and shape information. Li

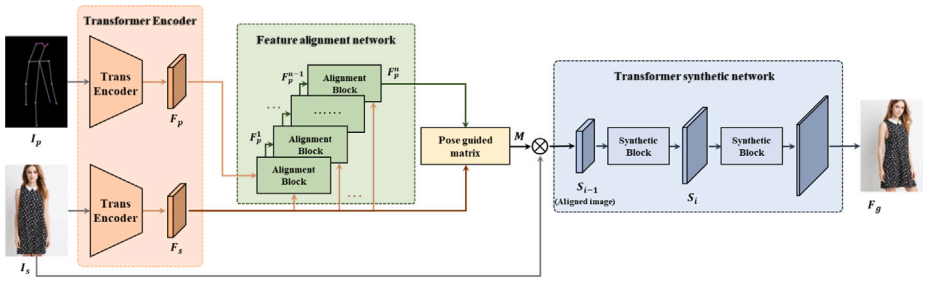

Fig. 3. Overview of our model. Our model contains two encoders that extract features from the source images I_s and target poses I_p to obtain feature maps F_s and F_p, respectively. The feature alignment network aligns F_s with F_p using n cascaded blocks, and obtains the aligned feature map F_p^n. Next, a correlation matrix M containing the target style and semantic information is then computed. Finally, the transformer synthetic network generates the final output F_g by warping the result of the previous stage with the correlation matrix based on the aligned image.

et al. [11] proposed the pose-guided non-local attention to build long-range feature dependency. They then introduced a cross-model block to better exploit both pose and image features.

These methods can model a long-range dependency scheme. But they only achieved sparse correspondence between the source and target images, making it challenging to transform the image. Han et al. [25] first generated human parsing map as semantic guidance. They estimated a dense flow to warp sources at the pixel level, which cannot perform well with large occlusion. Li et al. [13] warped the inputs at the feature level. They proposed estimating a dense and intrinsic 3D appearance flow to more effectively guide pixel transfer between poses. But their flow fields between sources and targets need additional 3D human models, which incur high computational cost. Yu et al. [14] introduced a different global-flow local-attention framework to reassemble the inputs. They did not rely on any additional information and obtained the flow fields in an unsupervised manner. However, their work is divided into two stages and is not end-to-end. Ma et al. [26] enhance texture-preserving pose transfer by exploring the complementary nature of attention and flow from a frequency perspective, transforming features from both into the wavelet domain. Li et al. [27] propose PT^2, a self-driven human pose transfer method that disentangles pose from texture at the patch level. This method removes the pose from an input image by permuting image patches and reconstructs the image by sampling from these permuted textures, achieving patch-level disentanglement.

Recently, 3D pose transfer has been attracting increasing attention, enabling the pose transformation of 3D meshes. Liu et al. [28] achieve the synthesis of arbitrary human poses based on neural scene representation and rendering. Chen et al. [29] propose an unsupervised 3D pose transfer method that employs a co-occurrence discriminator to identify the mesh's structural and pose invariance, resulting in better outcomes and improved efficiency. Wang et al. [30] introduce a zero-shot method that achieves strong generalization, enabling pose transfer for stylized 3D characters.

2.3. Transformers in computer vision

Transformers were first proposed by [31] and widely adopted for neural language processing due to multi-head self-attention and feed-forward MLP layers. Recently, more and more works used transformers to replace some or all spatial convolution layers in various computer vision tasks, yielding excellent results. Some approaches used pure transformer models for image processing. For instance, ViT [32] reshaped images into sequences of flattened 2D patches and applied a transformer architecture for image classification. However, it required large-scale training datasets. DeiT [33] introduced several training strategies to improve the data-efficiency of ViT. PVT [34] applied ViT models to the dense prediction tasks of semantic segmentation and object detection. Different from ViT, which typically has low-resolution outputs and high computational and memory cost, PVT proposed a pyramid architecture which can achieve high output resolution and low computational cost for large feature maps. T2T-Vit [35] used T2T module to assist each token in modeling local important structure information, so as to enhance the network capability. There are other works that used transformers to complement CNNs by incorporating self-attention layers to enhance backbones, enabling the encoding of distant dependencies. They then combined CNNs and transformers into an encoder–decoder architecture for object detection [36–38] and segmentation tasks [39].

3. The proposed method

We aim to generate realistic images with the target pose while preserving the original style. This requires the network to have the ability to establish long-distance dependencies, and keep the local details as well. To address this challenge, we tame transformers for human image generation. We propose a novel framework which consists of transformer encoder, feature alignment network and transformer synthetic network, as shown in Fig. 3. With these modules, we can model the long-range interaction between pixels from both the source and the target. At the same time, we combine convolution into our modules to effectively employ the inductive bias and keep the local details.

To facilitate arbitrary pose transfer, we utilize standard pose representations for guidance. Specifically, we employ 18 human keypoints extracted by the Human Pose Estimator [40] and represent the pose using a heatmap. This heatmap consists of 18 channels, each corresponding to the position of a joint in the human body.

3.1. Transformer encoder

Previous works [8–11,13,14] usually employ convolution-based encoder, which often bring the dilemma of choosing between keeping the style detail and building the global receptive field. To obtain a global receptive field, we need to perform multiple downsampling steps by employing a sufficiently deep network, which allows us to gather enough semantic information, albeit at the expense of style details. Conversely, to retain more detailed style information, we typically avoid the downsampling process. However, this results in the network not acquiring a sufficient receptive field and failing to capture enough semantic information.

However, for the human transfer task, the output image is a rearranged group of input image pixels, which requires the network to have the ability to not only retain the source style information, but also establish long-distance dependencies on images. In light of this, we propose a transformer encoder which combines Swin Transformer with depthwise convolution. First, we introduce Swin Transformer [15] as a backbone network to extract features from the source image I_s and pose image I_p. We then adopt depthwise convolution in MBConv [41] to follow the Swin Transformer layer and leverage the shortcut to retain the source image features. They together form a downsampling module. The encoded result F_s contains the style information of the source image, and F_p contains the semantic information of the target pose.

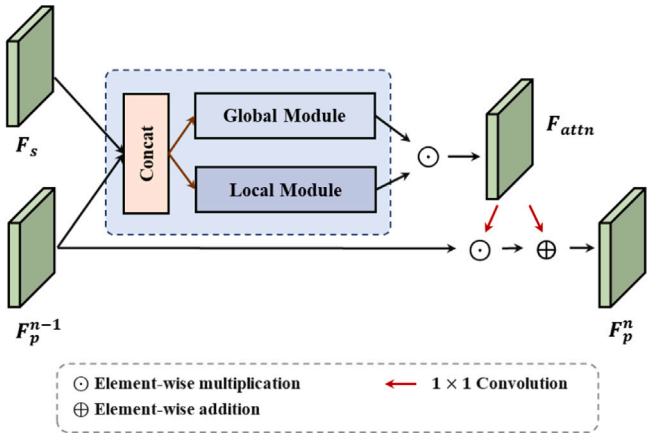

Fig. 4. The last block of our feature alignment network.

3.2. Feature alignment network

Pose transfer involves moving patches from the conditional pose to the target pose and establishing dependencies among these patches. In this context, the pose guides the transfer by determining where to extract the conditional patch and where to place the target patch, while maintaining the relationships among patches. In light of this, we propose a feature alignment network, which combines transformer layers with convolution to model both global dependencies among patches and the local features within each patch.

In Fig. 4, we present the last block of our feature alignment network. Image features can be gradually transferred from source poses to target poses through multiple alignment blocks. Among them, each block uses style features to render pose features to complete the task of pose-guided transfer. The input of the first block is the style feature map F_s and the pose feature map F_p. When outputting, F_p is updated to $F_p^i (i = 1)$, and in the tth block, F_p^{t-1} is updated to F_p^t. Through the update of F_p by n blocks, we get the final output F_p^n and send it to the decoder to generate the final result.

Since our alignment blocks share the same structure, we only take the last block as an example to illustrate how they work, as shown in Fig. 4. Our alignment block mainly consists of global module and local module. The global module focuses on establishing global dependencies, while local module focuses on local features and generates an attention mask, where the value of the attention mask is between 0 and 1, indicating the importance of each position. Then, the results of the two nets are multiplied element by element to obtain attention map, F_{attn} as shown in Eq. (1):

$$F_{attn} = Glob(Cat(F_p^{n-1}, F_s)) \odot S(Loc(Cat(F_p^{n-1}, F_s))), \tag{1}$$

where, $Glob$ means the global module, Loc means the local module, Cat means concatenation, S means Sigmoid and $\odot$ means element-wise multiplication. Global module and local module share a common input, which is obtained by concatenating F_s and F_p^{n-1}.

The local module is a residual convolutional layer, which contains both a BN layer [42] and a ReLU. The global module consists of two transformer layers, and a BN layer [42]. It splits the feature map into patches, and then uses multi-head self-attention to establish relationships among patches. Taking inspiration from [43], we modulate normalized features F_p^{n-1} with the scale and bias, which are predicted with two fully connected layers from F_{attn}. F_p^{n-1} will be updated as F_p^n, value at site ($c \in C^i$, $h \in H^i$, $w \in W^i$) as Eq. (2):

$$F_p^n = \gamma_{c,h,w}^i (F_{attn}) \times \frac{F_{c,h,w}^i - \mu_c^i}{\sigma_c^i} + \beta_{c,h,w}^i (F_{attn}) \tag{2}$$

where, C, H and W denote the channel, height and width of the tensor. $\gamma_{c,h,w}^i (F_{attn})$ and $\beta_{c,h,w}^i (F_{attn})$ are the learned modulation parameters. $F_{c,h,w}^i$ is the activation preceding the ith normalization layer F_p^{n-1}. σ_c^i and $\beta_{c,h,w}^i$ are the mean and standard deviation of the activations in channel c.

To further demonstrate the motivation and effectiveness of our feature alignment network, Fig. 5 visualizes the updated pose feature map F_p^n generated by the feature alignment network when provided with the style feature map F_s and the pose feature map F_p. The resulting feature map F_p^n demonstrates that the pose feature map can align the style features with the target pose.

3.3. Transformer synthetic network

For image synthesizing, it is crucial to keep the network focusing on the correct regions, as in this way the generated image will have fewer visible artifacts and fewer errors in the prediction [44]. Our transformer synthesis network aims to produce high-quality images by referencing the correct corresponding regions in the exemplar and aligning with concentrated semantics from target poses. Fig. 6 gives an example of one block of our transformer synthetic network.

In order to constrain the network to retain the correct corresponding regions, we match the aligned exemplar generated by the feature alignment network and the semantic map with the correspondence layer proposed in [45] to obtain the correlation matrix $M \in \mathbb{R}^{N \times N}$. N is the size of the matrix, $N = 4096$. Each element is a pairwise feature correlation, as shown in Eq. (3):

$$M(u, v) = \frac{\hat{F}_p^n(u)^T \hat{F}_s(v)}{\|\hat{F}_p^n(u)\| \|\hat{F}_s(v)\|}. \tag{3}$$

where, $\hat{F}_p^n(u)$ and $\hat{F}_s(v)$ denote the channel-wise centralized features of F_p^n and F_s in positions u and v, i.e., $\hat{F}_p^n(u) = F_p^n(u)$ - mean($F_p^n(u)$) and $\hat{F}_s(v) = F_s(v)$ - mean($F_s(v)$).

The correlation matrix contains the target style and semantic information, and we use the matrix in each part of the synthesis phase to obtain a correlated exemplar field to constrain the network to generate the correct images. According to CoCosNet [44], we use correlation matrix M to warp the source image I_s and obtain the aligned image $r_{I_s \to I_p}$, as shown in Eq. (4):

$$r_{I_s \to I_p}(u) = \sum_v softmax(\alpha M(u, v)) \cdot I_s(v). \tag{4}$$

where α is the coefficient that regulates the sharpness of the softmax function, with a default value set to 100, following the practice established by CoCosNet [44]. CoCosNet aims to translate images from a source domain to a target domain using an input image and an exemplar image, which closely aligns with our objective. However, different sizes of correlation matrices are required in each synthesis phase, and in order for the matrix to function at each stage of the synthesis, we need to use a pooling layer to adjust its size. Unfortunately, the correlation matrix will lose the spatial context due to the pooling layer. To address this problem, we extract the spatial context relation of features. We define the output of each stage of the transformer synthetic network as S_i. Specifically, we convert the feature S_{i-1} to the feature vector, which is obtained by the Swin-Transformer module and the 1 × 1 conv module. We use the Swin Transformer module, which consists of two Swin-Transformer layers, to focus on regional features in a global way to model the context of feature S_{i-1}. Then we extract the spatial information of the feature S_{i-1} through the 1 × 1 conv, which is subsequently flattened into a feature vector. Meanwhile, the feature guidance matrix matches the corresponding feature size through average pooling, allowing for matrix multiplication with the feature vector. By multiplying the vector with the correlation matrix, we can obtain a correlated exemplar field E. The correlated exemplar field E preserves both the spatial context of the image and the correlated pixel matrix, as shown in Eq. (5):

$$E = Avg(M) \times Conv(Tran(S_{i-1})). \tag{5}$$

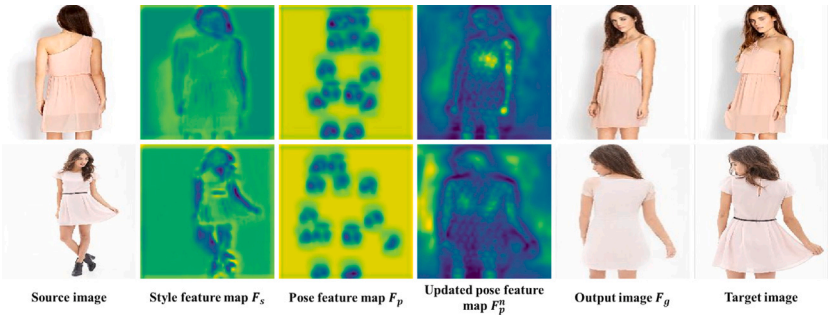

Fig. 5. Visualization of the updated pose feature map, which is sent to the decoder to produce the final result.

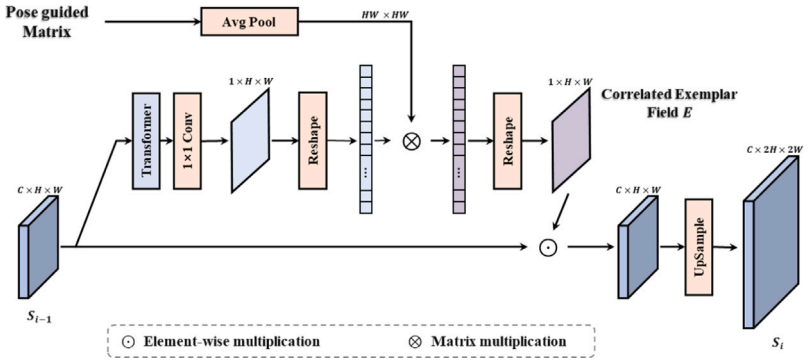

Fig. 6. One block of our transformer synthetic network.

where Avg means average pooling, $Tran$ means Swin Transformer module. Finally, we apply the correlated exemplar field to each stage of the synthetic network, enabling the network to focus on the correlated pixels and synthesize high-quality images. Please note that the input source and pose images are processed in our transformer encoder module, resulting in style and pose feature maps with smaller spatial dimensions. Consequently, here the features are progressively upsampled to align with the sizes of both the input image and the ground-truth image, ensuring there are no resolution issues. The output of each stage can be expressed by Eq. (6):

$$S_i = Up(S_{i-1} \otimes E). \tag{6}$$

where $\otimes$ means element-wise multiplication, and Up means upsampling.

Visformer [46] has proved the importance of embedding layer for transformer to preserve the features of patches. We go further by introducing consecutive small convolution layers to replace original single large convolution layer in Swin Transformer Block. In contrast to the original design, ours can better preserve the information of each patch and significantly improve the ability of transformer in terms of modeling and establishing long-range dependencies among patches. Among them, we set the size of the convolution kernel in Embedding to 4, 2, and 1, respectively, and the stride to 4, 2, and 1, respectively.

3.4. Objective function

Our total loss includes a conditional adversarial term, an $\mathcal{L}1$ term, a perceptual term and a contextual term, as:

$$\mathcal{L}_{total} = \lambda_a \mathcal{L}_{adv} + \lambda_{L1} \mathcal{L}_{L1} + \lambda_p \mathcal{L}_{per} + \lambda_{CX} \mathcal{L}_{CX}, \tag{7}$$

where λ_a, λ_{L1}, λ_p, λ_{CX} are hyperparameters to balance the objectives and further details are presented in subsection of Implementation Details.

Conditional adversarial loss. We utilize two discriminators as described in [8]: an appearance discriminator D_A to assess the likelihood that F_g contains the same individual as I_s (appearance consistency), and shape discriminator D_P to judge how well F_g is aligned with the target pose I_p (shape consistency). The conditional adversarial loss [4] is defined as:

$$\begin{aligned}\mathcal{L}_{adv} =& \mathbb{E}_{I_p \in I, (I_s, I_t) \in \mathcal{P}} \{\log[D_A(I_s, I_t) \cdot D_P(I_p, I_t)]\} + \\ & \mathbb{E}_{I_p \in I, I_s \in \mathcal{P}, F_g \in \mathcal{P}_g} \{\log[(1 - D_A(F_g, I_s)) \cdot (1 - D_P(F_g, I_p))]\}.\end{aligned} \tag{8}$$

Note that I, $\mathcal{P}$ and $\mathcal{P}_g$ represent the distribution of person poses, real images and generated images, respectively.

L1 loss. We adopt $\mathcal{L}_{L1}$ loss [47] to compute the pixel-wise differences between the generated image F_g and the ground truth I_t, defined as:

$$\mathcal{L}_{L1} = \|F_g - I_t\|_1 \tag{9}$$

Perceptual loss. We adopt perceptual loss [48] by obtaining the $\mathcal{L}_1$ distance between activation maps of the pretrained VGG-19 network, computed as:

$$\mathcal{L}_{per} = \|\phi_l(F_g) - \phi_l(I_t)\|_1 \tag{10}$$

where we select ϕ_l as the activation following the $relu4\_2$ layer in the VGG-19 network, as this layer predominantly captures high-level semantics.

Contextual loss. We adopt the contextual loss proposed by [49,50] to match the statistics between F_g and I_t. The contextual loss can guide spatial deformations w.r.t. the target, and lead to less texture distortion and more reasonable output. We compute the contextual loss as:

$$\mathcal{L}_{CX} = -\log(CX(\phi_l(F_g), \phi_l(I_t))) \tag{11}$$

Following the practice of [49,50], we calculate the contextual loss by computing the similarity metric (CX) between the feature maps of the generated image F_g, denoted as $\phi_l(F_g)$, and the feature maps of the target image I_t, denoted as $\phi_l(I_t)$. Here, l refer to $relu\{3\_2, 4\_2\}$ of the pretrained VGG19 network.

Fig. 7. Image generation results conditioned by various poses on the DeepFashion dataset and Market-1501 dataset.

4. Experiments

4.1. Experimental setup

4.1.1. Dataset

We conduct our experiments on person re-identification dataset Market-1501 [51] and DeepFashion In-shop Clothes Retrieval Benchmark [52]. Images in Market-1501 are low resolution (128 × 64) and the images differ in aspects such as viewpoints, backgrounds, and illumination. DeepFashion consists of high-quality model images in fashion clothes, which has clean backgrounds. We follow [37] to split the data and collect 263,632 training pairs and 12,000 testing pairs from Market-1501 and 101,966 training pairs and 8,570 testing pairs from DeepFashion. Note that the person identities in the test set are different from those in the training set.

4.1.2. Metrics

We use Learned Perceptual Image Patch Similarity (LPIPS) proposed by [53] to calculate the reconstruction error between the generated image and the ground-truth image in the perceptual level. Meanwhile, we use the Fréchet Inception Distance [54] (FID) to measure the realism of the generated images. Additionally, we use the Peak Signal-to-Noise Ratio (PSNR) to measure the pixel-level error between the generated image and the ground-truth image.

4.1.3. Implementation details

We train our model using 256 × 256 images for the Fashion dataset and 128 × 64 for the Market-1501 dataset. We adopt the Adam [55] solver with $\beta_1 = 0.5$, $\beta_2 = 0.999$. For the learning rate, we set 0.0001 and 0.0002 respectively, the generator and discriminator following the TTUR [54]. It is important to note that the L1 loss and adversarial loss provide the overall optimization direction for the model, while the perceptual loss and CX loss are used to refine detailed performance. If the values of the perceptual loss and CX loss are too large, it may lead to training instability or even crashes. Therefore, we set smaller hyperparameters for these losses. Based on these and our empirical exploration, the weights for the loss terms are set to λ_a=10, λ_p=0.0001, λ_{L1}=10 and λ_{CX}=0.001. Our method is implemented in PyTorch using an NVIDIA RTX3090 GPUs with 24 GB memory. Besides, in the feature alignment network, we use two alignment blocks.

4.2. Comparison with state-of-the-arts

We conduct qualitative comparisons on Market-1501 and DeepFashion datasets with several stare-of-the-art methods including Pose-Attn [8], BiGraph [9], XingGAN [10], PoNA [11], GFLA [14], PISE [56], Pose2Pose [57] and PT^2 [27]. Fig. 7 shows the image generation results of our method conditioned by various poses on the DeepFashion and Market-1501 datasets, and Figs. 8 and 9 show some qualitative comparisons. For the Fashion dataset, PATN and BiGrapth fail to generate sharp images and cannot predict complex textures due to the lack of global receptive fields in these models. XingGAN and PoNA use the attention module to obtain the global receptive field, which makes the non-local module focus on different domains, i.e., style maps and segmentation maps. However, it is difficult to generate reasonable images. GFLA uses the flow-based method, and PISE synthesizes a target semantic segmentation map, respectively, to preserve detailed textures in the source image. PT^2 proposes a permuting textures method to reconstruct the original image from the permuted input, showing outstanding performance. However, they fail to predict some fine textures and shapes for the invisible regions of the source image. Meanwhile, their methods require two-stage training or additional semantic segmentation maps. Pose2Pose uses dense multi-scale attention to improve performance, but there are still many wrong textures. It can generate realistic images with correct poses and vivid details. Since our research aims to generate images that blend the style of the source image with the target pose, our model lacks a structural design to preserve facial detail features. This may lead to some deviations in eye direction for certain characters (e.g., the first row). Even so, most of our results (e.g., the second, third, and fifth rows) show no such issues, and the generated images have higher quality compared to previous methods. For the Market-1501 Dataset, although it has a low resolution and complex background, we can generate more natural and sharper images. Compared with recent methods PoNA, GFLA and PT^2, our method can restore more details and less artifacts.

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

Model	DeepFashion			Market-1501		
	FID ↓	LPIPS ↓	PSNR ↑	FID ↓	LPIPS ↓	PSNR ↑
Pose-Attn	23.70	0.2520	15.20	37.99	0.3187	14.25
PoNA	24.20	0.2594	13.41	26.42	0.2859	14.64
XingGAN	44.81	0.2920	16.38	37.61	0.3050	14.42
BiGraph	24.41	0.2493	17.05	37.32	0.3042	**15.03**
GFLA	14.82	0.2311	16.98	28.05	**0.2809**	14.33
PISE	13.63	0.2326	16.74	–	–	–
Pose2Pose	14.59	0.2654	15.75	–	–	–
PT^2	28.94	0.2415	16.51	22.14	0.3140	14.08
Ours	**13.11**	**0.2120**	**17.64**	**22.02**	0.2819	14.99

Table 1 gives quantitative results of our model compared with several stare-of-the-art methods: Pose-Attn [8], PoNA [11], XingGAN [10], BiGraph [9], GFLA [14], PISE [56], Pose2Pose [57] and PT^2 [27]. Because PISE and Pose2Pose do not provide pre-trained model on the Market-1501 dataset, we only compare with them on the DeepFashion dataset. For DeepFashion, our results get the best FID score, which means our generated images are more realistic. Besides, we adopt LPIPS to compute the similarity in perceptual level and PSNR to measure the error in pixel level. Our method has the best results in terms of both LPIPS and PSNR, which indicates that our results have less error in pixel level and are more consistent in shape and texture with the target images. For Market-1501, the quantitative results demonstrate that the images generated by our method are closer to real images in shape and texture and our metrics outperform most other methods, even though the conditioned images are of lower resolution with significant changes in pose and background.

4.3. Ablation study

We train several ablation models to verify our assumptions and the effectiveness of each component.

Fig. 8. The qualitative comparisons with several state-of-the-art models on DeepFashion dataset, including PATN [8], BiGraph [9], XingGAN [10], PoNA [11], GFLA [14], PISE [56], Pose2Pose [57], PT^2 [27] and ours, respectively.

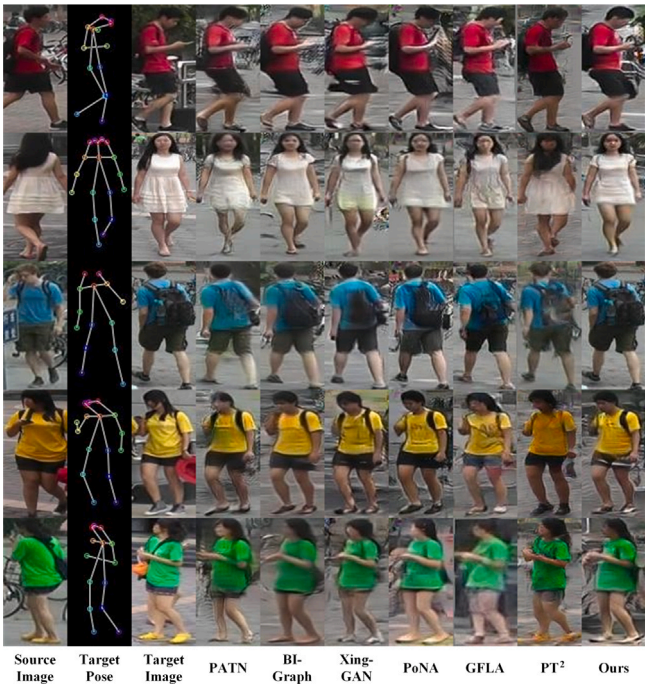

Fig. 9. The qualitative comparisons with several state-of-the-art models on Market-1501 dataset, including PATN [8], BiGraph [9], XingGAN [10], PoNA [11], GFLA [14], PT^2 [27] and ours, respectively.

Global-Enc: Our encoder consists of transformer layers and convolution layers. We use only transformer layers as the backbone of the encoder, so that the network can obtain a global receptive field to verify the effectiveness of Transformer.

Local-Enc: We use only convolution layers as the backbone of the encoder to make the network pay attention to local information, to evaluate the performance of convolutional backbone and Transformer backbone.

Without Feature alignment network (w/o FA): In this model, we remove feature alignment network from our Full Model **(w/o FA-F)**, to verify the effectiveness of the FA model on the final image. In addition, we removed the global module **(w/o FA-G)** and local module **(w/o FA-L)** from the alignment module separately to compare their effects in more detail.

Table 2
Quantitative results of ablation study.

Model		FID ↓	LPIPS ↓	PSNR ↑
w/o FA	w/o FA-F	13.92	0.2156	17.39
	w/o FA-G	13.77	0.2142	17.49
	w/o FA-L	13.54	0.2128	17.57
Global Enc		19.24	0.2270	17.02
Local Enc		14.05	0.2135	17.46
w/o SN		14.60	0.2147	17.45
w/o DA		18.27	0.2188	17.40
w/o DP		18.43	0.2224	17.35
w/o CX		18.86	0.2214	17.40
Ours		**13.11**	**0.2120**	**17.64**

Without Transformer synthetic network (w/o SN): In this model, we remove transformer synthetic network from our Full Model, to verify the effectiveness of the SN model on the final image.

Without the appearance discriminator (w/o DA): We do not use the appearance discriminator in the training phase to judge how likely F_g contains the same person as I_s.

Without the shape discriminator (w/o DP): We do not use the shape discriminator in the training phase to judge how well F_g is aligned with the target pose I_p.

Without CX loss (w/o CX): We remove contextual loss from our full training loss, to verify the effectiveness of the contextual loss.

Full models (Ours): We use our proposed framework in this model.

Quantitative results on DeepFashion test images are shown in Table 2. Global-Enc ignores local details and Local-Enc ignores context features. Compared with the Global-Enc and Local-Enc, our encoders produce better results than pure convolutions and pure transformers. In the ablation experiment, we compared only Transformer (Global-Enc) and only convolution (Local-Enc) as the backbone network. The experiment proved that adding Transformer as the backbone network will make the network more powerful. However, without adding Transformer, our network can also achieve a 14.05 FID score, higher than other previous methods. In light of this, we can say that the network

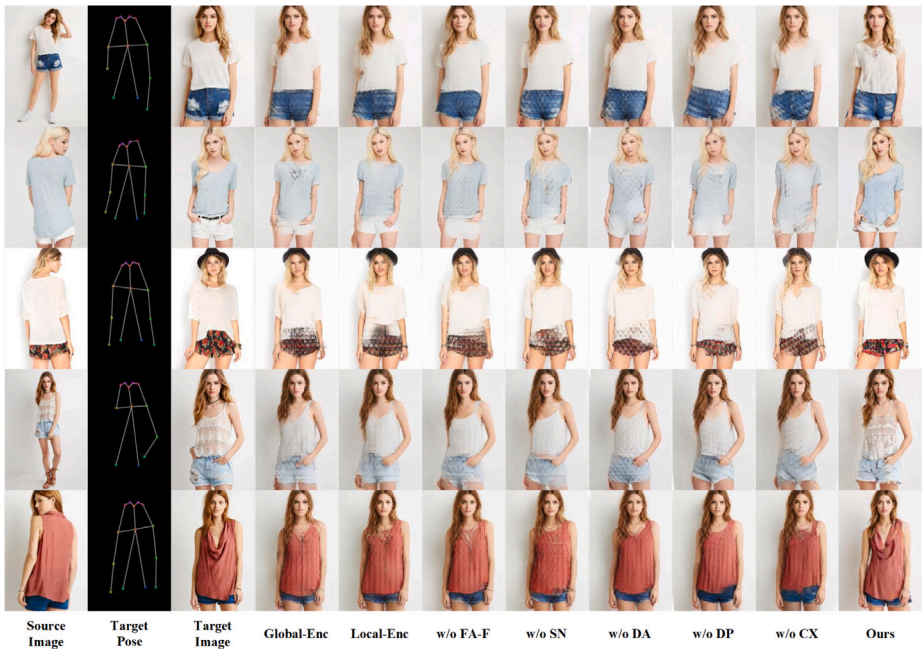

Fig. 10. Qualitative results of ablation study.

benefits more from our proposed feature alignment module and synthetic module. Besides, FA aligns the features of the semantic map and exemplar with the correspondence layer, which can better provide prior information. The global module and the local module of the feature alignment network can consider the global semantic information and the local texture information, respectively, in order to obtain more accurate correspondence information, and quantitative results show their effectiveness. In the synthetic network, the global receptive field provided by SN is also important. The appearance discriminator and the shape discriminator ensure the consistency of the generated image with the target image, and the CX loss can lead to more reasonable results with less texture distortion, as demonstrated by FID and LPIPS.

Fig. 10 shows some intuitive visual results of ablation study. It can be seen that the Global-Enc and Local-Enc can generate correct structures. However, they only focus on global or local information, and their texture details are not satisfactory. As shown in the second row, fourth, and fifth columns. Without the FA-F module, the resulting image is partially blurred. Without SN module, our model cannot preserve more details and generate sharper images, as shown in the third and fifth rows, and seventh column. In addition, without using the D_A discriminator, the D_P discriminator, or the CX loss during the training phase, the synthesized images will produce rough or incorrect textures, as shown in the eighth, ninth and tenth columns.

5. Conclusion and discussion

In this paper, we explore a transformer based method for the human pose transfer task. We find that combining transformer and convolution in a reasonable way will improve the performance of the model, while providing aligned feature regions in the embedded and disentangled feature space when synthesizing the final image will also improve the quality of the final image. Our method first generates aligned images with the target pose by the feature alignment network, and then generates high-quality images gradually by the transformer synthetic network. Experimental results demonstrate that our model can generate realistic images with vivid details by paying attention to both global and local information. In addition, the ablation study also verifies the effectiveness of each designed component.

However, our method has some shortcomings. First, our model requires more computing time and memory, compared to convolution based methods. Since our method employs the Transformer structure, it has higher computational requirements than CNN-based methods and requires more inference time compared to models like XingGAN [10], although it is still within the same order of magnitude. In the future, we aim to improve the efficiency of our method by introducing novel lightweight Transformer structures from other spatial deformation tasks, such as facial animation.

Second, our method might encounter cross-domain problems if the domains differ significantly. For example, training on DeepFashion data and testing on the Market-1501 dataset can lead to suboptimal outputs. There are significant differences between these two datasets: DeepFashion typically features images of models against plain white backgrounds with clear human poses, while Market-1501 contains images with complex street backgrounds, where subjects often blend with the surroundings and exhibit unclear and ambiguous poses. In our current experimental setup, this disparity makes it challenging for our encoder to extract accurate features, negatively impacting the generation of the pose guidance matrix and leading to suboptimal outputs.

To address this challenge, in our future work, we aim to enhance the model's generalization capability through three specific research directions. First, we propose employing domain adaptation techniques, which are widely used in object detection [58,59]. By introducing feature adversarial training between source and target domain datasets, this approach enables the encoder to learn domain-invariant, generalizable representations, thereby improving the model's generalization capability across diverse datasets. We plan to alternately use DeepFashion and Market-1501 as the source and target domains for joint training, extracting intermediate features for adversarial learning. The objective is to capture shared features between the two datasets, achieving superior cross-domain generalization. Second, we intend to implement more effective data augmentation strategies, such as background replacement and pose transformation, to further enhance the model's robustness in cross-domain scenarios. Finally, we plan to leverage language models to tackle this challenge. Language models have demonstrated significant effectiveness in pose-guided image generation tasks [60] and other visual applications [61]. We aim to utilize the semantic information extracted by language models as a global condition, embedding it into various layers of the model for feature fusion. This approach will guide the model to focus on human pose details and clothing textures, reducing the impact of background or pose variations on performance.

CRediT authorship contribution statement

Yu Luo: Writing – original draft, Software, Methodology, Data curation, Formal analysis. **Chengzhi Yuan:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Data curation. **Lin Gao:** Writing – review & editing, Supervision, Resources, Formal analysis. **Weiwei Xu:** Supervision, Writing – review & editing. **Xiaosong Yang:** Supervision, Writing – review & editing. **Pengjie Wang:** Writing – review & editing, Project administration, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work is supported by a Research funding of the European Union's Horizon 2020 research and innovation programme (Ref.: Marie Skłodowska-Curie grant agreement No 900025), a Natural Science Foundation of Liaoning (Ref.: 2023-MS-133) and a Research funding of Liaoning Provincial Education Department, China (Ref.: LJ242412026018). Yu Luo completed all experiments in the original submission, and Chenzhi Yuan completed all experiment in the revision submission.

Data availability

Data will be made available on request.

References

[1] Q. Zhou, H. Fan, H. Yang, H. Su, S. Zheng, S. Wu, H. Ling, Robust and efficient graph correspondence transfer for person re-identification, IEEE Trans. Image Process. 30 (2019) 1623–1638.

[2] H. Yue, J. Yang, X. Sun, F. Wu, C. Hou, Contrast enhancement based on intrinsic image decomposition, IEEE Trans. Image Process. A Publ. IEEE Signal Process. Soc. 26 (99) (2017) 3981–3994.

[3] J. Walker, K. Marino, A. Gupta, M. Hebert, The pose knows: Video forecasting by generating pose futures, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017, pp. 3332–3341.

[4] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Adv. Neural Inf. Process. Syst. 27 (2014).

[5] H. Shi, L. Wang, N. Zheng, G. Hua, W. Tang, Loss functions for pose guided person image generation, Pattern Recognit. 122 (2022) 108351.

[6] C. Han, X. Yu, C. Gao, N. Sang, Y. Yang, Single image based 3D human pose estimation via uncertainty learning, Pattern Recognit. 132 (2022) 108934.

[7] S. Yang, W. Yang, Z. Cui, Searching part-specific neural fabrics for human pose estimation, Pattern Recognit. 128 (2022) 108652.

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

[9] H. Tang, S. Bai, P.H.S. Torr, N. Sebe, Bipartite graph reasoning GANs for person image generation, 2020, CoRR abs/2008.04381, arXiv:2008.04381.

[10] H. Tang, S. Bai, L. Zhang, P.H. Torr, N. Sebe, Xinggan for person image generation, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 717–734.

[11] K. Li, J. Zhang, Y. Liu, Y.-K. Lai, Q. Dai, PoNA: Pose-guided non-local attention for human pose transfer, IEEE Trans. Image Process. 29 (2020) 9584–9599.

[12] L. Yang, P. Wang, C. Liu, Z. Gao, G. Wen, Towards fine-grained human pose transfer with detail replenishing network, IEEE Trans. Image Process. PP (99) (2021) 1.

[13] Y. Li, C. Huang, C.C. Loy, Dense intrinsic appearance flow for human pose transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3693–3702.

[14] Y. Ren, G. Li, S. Liu, T.H. Li, Deep spatial transformation for pose-guided person image generation and animation, IEEE Trans. Image Process. 29 (2020) 8622–8635.

[15] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[16] P. Isola, J.-Y. Zhu, T. Zhou, A.A. Efros, Image-to-image translation with conditional adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1125–1134.

[17] J.-Y. Zhu, T. Park, P. Isola, A.A. Efros, Unpaired image-to-image translation using cycle-consistent adversarial networks, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2017, pp. 2223–2232.

[18] Y. Jiang, X. Gong, D. Liu, Y. Cheng, C. Fang, X. Shen, J. Yang, P. Zhou, Z. Wang, Enlightengan: Deep light enhancement without paired supervision, IEEE Trans. Image Process. 30 (2021) 2340–2349.

[19] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.

[20] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, T.S. Huang, Generative image inpainting with contextual attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 5505–5514.

[21] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in: K. Chaudhuri, R. Salakhutdinov (Eds.), Proceedings of the 36th International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 97, PMLR, 2019, pp. 7354–7363, URL https://proceedings.mlr.press/v97/zhang19d.html.

[22] P. Esser, R. Rombach, B. Ommer, Taming transformers for high-resolution image synthesis, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 12873–12883.

[23] Y. Jiang, S. Chang, Z. Wang, Transgan: Two pure transformers can make one strong gan, and that can scale up, Adv. Neural Inf. Process. Syst. 34 (2021).

[24] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, Pose guided person image generation, Adv. Neural Inf. Process. Syst. 30 (2017).

[25] X. Han, X. Hu, W. Huang, M.R. Scott, Clothflow: A flow-based model for clothed person generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 10471–10480.

[26] L. Ma, T. Gao, H. Shen, K. Huang, Freqhpt: Frequency-aware attention and flow fusion for human pose transfer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 3491–3496.

[27] N. Li, K.J. Shih, B.A. Plummer, Collecting the puzzle pieces: Disentangled self-driven human pose transfer by permuting textures, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 7126–7137.

[28] L. Liu, M. Habermann, V. Rudnev, K. Sarkar, J. Gu, C. Theobalt, Neural actor: Neural free-view synthesis of human actors with pose control, ACM Trans. Graph. 40 (6) (2021) 1–16.

[29] H. Chen, H. Tang, H. Shi, W. Peng, N. Sebe, G. Zhao, Intrinsic-extrinsic preserved gans for unsupervised 3d pose transfer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 8630–8639.

[30] J. Wang, X. Li, S. Liu, S. De Mello, O. Gallo, X. Wang, J. Kautz, Zero-shot pose transfer for unrigged stylized 3d characters, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 8704–8714.

[31] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, 2020, arXiv preprint arXiv:2010.11929.

[33] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, H. Jégou, Training data-efficient image transformers & distillation through attention, in: Proceedings of the 38th International Conference on Machine Learning, PMLR, 2021, pp. 10347–10357.

[34] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, L. Shao, Pyramid vision transformer: A versatile backbone for dense prediction without convolutions, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 568–578.

[35] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z.-H. Jiang, F.E. Tay, J. Feng, S. Yan, Tokens-to-token vit: Training vision transformers from scratch on imagenet, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 558–567.

[36] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, S. Zagoruyko, End-to-end object detection with transformers, in: Proceedings of the European Conference on Computer Vision, Springer, 2020, pp. 213–229.

[37] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, J. Dai, Deformable detr: Deformable transformers for end-to-end object detection, 2020, arXiv preprint arXiv:2010.04159.

[38] C. Chi, F. Wei, H. Hu, Relationnet++: Bridging visual representations for object detection via transformer decoder, Adv. Neural Inf. Process. Syst. 33 (2020) 13564–13574.

[39] H. Wang, Y. Zhu, H. Adam, A. Yuille, L.-C. Chen, Max-deeplab: End-to-end panoptic segmentation with mask transformers, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 5463–5474.

[40] Z. Cao, T. Simon, S.-E. Wei, Y. Sheikh, Realtime multi-person 2d pose estimation using part affinity fields, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 7291–7299.
[41] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: Efficient convolutional neural networks for mobile vision applications, 2017, arXiv preprint arXiv:1704.04861.
[42] S. Ioffe, C. Szegedy, Batch normalization: Accelerating deep network training by reducing internal covariate shift, in: F. Bach, D. Blei (Eds.), Proceedings of the 32nd International Conference on Machine Learning, in: Proceedings of Machine Learning Research, vol. 37, PMLR, Lille, France, 2015, pp. 448–456, URL https://proceedings.mlr.press/v37/ioffe15.html.
[43] T. Park, M.-Y. Liu, T.-C. Wang, J.-Y. Zhu, Semantic image synthesis with spatially-adaptive normalization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 2337–2346.
[44] P. Zhang, B. Zhang, D. Chen, L. Yuan, F. Wen, Cross-domain correspondence learning for exemplar-based image translation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5143–5153.
[45] B. Zhang, M. He, J. Liao, P.V. Sander, L. Yuan, A. Bermak, D. Chen, Deep exemplar-based video colorization, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 8052–8061.
[46] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, Q. Tian, Visformer: The vision-friendly transformer, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 589–598.
[47] H. Zhao, O. Gallo, I. Frosio, J. Kautz, Loss functions for image restoration with neural networks, IEEE Trans. Comput. Imaging 3 (1) (2017) 47–57, http://dx.doi.org/10.1109/TCI.2016.2644865.
[48] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 694–711.
[49] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, Z. Lian, Controllable person image synthesis with attribute-decomposed gan, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 5084–5093.
[50] R. Mechrez, I. Talmi, L. Zelnik-Manor, The contextual loss for image transformation with non-aligned data, in: Proceedings of the European Conference on Computer Vision, Springer, 2018, pp. 800–815.
[51] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: A benchmark, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2015, pp. 1116–1124.
[52] Z. Liu, P. Luo, S. Qiu, X. Wang, X. Tang, Deepfashion: Powering robust clothes recognition and retrieval with rich annotations, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 1096–1104.
[53] R. Zhang, P. Isola, A.A. Efros, E. Shechtman, O. Wang, The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[54] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, S. Hochreiter, Gans trained by a two time-scale update rule converge to a local nash equilibrium, Adv. Neural Inf. Process. Syst. 30 (2017).
[55] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, 2014, arXiv preprint arXiv:1412.6980.
[56] J. Zhang, K. Li, Y.-K. Lai, J. Yang, Pise: Person image synthesis and editing with decoupled gan, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 7982–7990.
[57] P. Roy, S. Bhattacharya, S. Ghosh, U. Pal, Multi-scale attention guided pose transfer, Pattern Recognit. 137 (2023) 109315, http://dx.doi.org/10.1016/j.patcog.2023.109315.
[58] F. Rezaeianaran, R. Shetty, R. Aljundi, D.O. Reino, S. Zhang, B. Schiele, Seeking similarities over differences: Similarity-based domain alignment for adaptive object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 9184–9193, http://dx.doi.org/10.1109/ICCV48922.2021.00907.
[59] H. Li, R. Zhang, H. Yao, X. Zhang, Y. Hao, X. Song, X. Li, Y. Zhao, L. Li, Y. Chen, DA-Ada: Learning domain-aware adapter for domain adaptive object detection, 2024, arXiv:2410.09004.
[60] Y. Ma, Y. He, X. Cun, X. Wang, S. Chen, Y. Shan, X. Li, Q. Chen, Follow your pose: Pose-guided text-to-video generation using pose-free videos, 2024, arXiv:2304.01186.
[61] W. Wang, Z. Chen, X. Chen, J. Wu, X. Zhu, G. Zeng, P. Luo, T. Lu, J. Zhou, Y. Qiao, et al., Visionllm: Large language model is also an open-ended decoder for vision-centric tasks, Adv. Neural Inf. Process. Syst. 36 (2024).