Pattern Recognition 124 (2022) 108451

Contents lists available at ScienceDirect

Pattern Recognition

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

Exploiting appearance transfer and multi-scale context for efficient person image generation

Chengkang Shen[a,1], Peiyan Wang[a,b,1], Wei Tang[a,*]

[a] University of Illinois at Chicago, Chicago, IL 60607, USA
[b] Purdue University, IN 47907, USA

ARTICLE INFO

Article history:
Received 22 June 2021
Revised 15 October 2021
Accepted 22 November 2021
Available online 24 November 2021

Keywords:
Person image generation
Appearance transfer
Multi-scale context
Efficient image generation

ABSTRACT

Pose guided person image generation means to generate a photo-realistic person image conditioned on an input person image and a desired pose. This task requires spatial manipulation of the source image according to the target pose. However, convolutional neural networks (CNNs) are inherently limited to geometric transformations due to the fixed geometric structures in their building modules, i.e., convolution, pooling and unpooling, which cannot handle large motion and occlusions caused by large pose transform. This paper introduces a novel two-stream context-aware appearance transfer network to address these challenges. It is a three-stage architecture consisting of a source stream and a target stream. Each stage features an appearance transfer module, a multi-scale context module and two-stream feature fusion modules. The appearance transfer module handles large motion by finding the dense correspondence between the two-stream feature maps and then transferring the appearance information from the source stream to the target stream. The multi-scale context module handles occlusion via contextual modeling, which is achieved by atrous convolutions of different sampling rates. Both quantitative and qualitative results indicate the proposed network can effectively handle challenging cases of large pose transform while retaining the appearance details. Compared with state-of-the-art approaches, it achieves comparable or superior performance using much fewer parameters while being significantly faster.

1. Introduction

Pose guided person image generation aims to transform a person image from a source pose to a target pose while retaining the appearance details. It serves as a fundamental tool for several practical applications such as image editing, video generation and data augmentation for person re-identification and action recognition [1–3]. This task is very challenging especially in case of large motion, occlusion and complex texture.

Convolutional neural networks (CNNs) [4] and their variants [5,6], trained in an adversarial fashion [7], have been widely used for image generation and translation [8–12]. However, since CNNs are composed of spatially local and translation equivariant operators, i.e., convolution, pooling and unpooling, they do not have an explicit mechanism to handle articulated body deformation. To resolve this difficult issue, two strategies have been adopted in prior person image generation approaches, i.e., parametric geometric transformation and nonparametric dense flow. For example, Siarohin et al. [13] apply an affine transformation to the features of each body part region to deal with pixel-to-pixel misalignment caused by the pose difference. However, it cannot handle occlusion or out-of-plane rotation well. Some methods [14,15] predict the dense flow field between the source and target images and apply it to warp the feature maps. However, since the flow is predicted via a CNN, it cannot account for large or non-local motion.

Large motion and occlusion still remain the two greatest challenges in person image generation. In addition, recent approaches have to use a very deep network, i.e., consisting of nine sophisticated blocks [2,16] to handle these difficulties. This unavoidably increases the model size and computational complexity.

This paper introduces a novel two-stream context-aware appearance transfer network to explicitly address the aforementioned challenges. As illustrated in Fig. 2, the network is a three-stage architecture consisting of a source stream and a target stream. The two streams respectively take as input the source pose and image, and the target pose. Each stage consists of a novel appearance transfer module, a multi-scale context module and two-stream feature fusion modules.

* Corresponding author.
E-mail addresses: cshen26@uic.edu (C. Shen), wang5035@purdue.edu (P. Wang), tangw@uic.edu (W. Tang).

[1] Chengkang Shen and Peiyan Wang contributed equally to this work. Peiyan Wang was an undergraduate student at the University of Illinois at Chicago when this work was done.

https://doi.org/10.1016/j.patcog.2021.108451
0031-3203/

The appearance transfer module is designed to address the **difficulty of large motion**. It finds the dense correspondence between the two-stream feature maps and then transfers the appearance information from the source stream to the target stream. Unlike the parametric geometric transformation or the nonparametric flow field, our appearance transfer module is inspired by the self-attention [17] and performs a *query-and-transfer* procedure. But different with the self-attention, the queries, keys and values in our module have explicit semantic meaning, and they are specially designed for pose-guided appearance transfer. Specifically, the feature vector at each spatial location in the target stream is taken as a *query* to match the *key* feature vectors in the source stream so that the corresponding appearance *values* in the source stream can be transferred to the desired location in the target stream.

The multi-scale context module is designed to address the **difficulty of occlusion**. Occlusion makes some visible content in the target image invisible in the source image. To help the target stream recover occluded pixels, it is necessary to give the network access to rich context and even a global view of the scene. For example, when an arm of a person is occluded, we can imagine what it looks like by checking the other arm of this person. When both arms are occluded, we need to understand the high-level clothing style, e.g., a suit or a set of sportswear, to reconstruct the missing information. This line of analysis motivates us to build a multi-scale context module based on the atrous spatial pyramid pooling (ASPP) [18], which was originally designed for image segmentation [19,20]. It aggregates multi-scale context via atrous convolutions with different sampling rates. Finally, the two-stream feature fusion modules allow local information exchange between the two streams to supplement the non-local appearance transfer and multi-scale context modeling.

Ablation study indicates that the proposed approach can effectively handle large pose transform, and improve the quality of the generated person images. Experimental results on two benchmark datasets, i.e., Market-1501 [21] and DeepFashion [22], show that compared with state-of-the-art methods, our approach achieves comparable or superior performance with a much smaller model size and a significantly higher inference speed. An SSIM-FPS-Parameters trade-off plot on the DeepFashion dataset is shown in Fig. 1.

The contributions of this paper are summarized below.

- We introduce a novel two-stream context-aware appearance transfer network for efficient person image generation. It progressively transfers the appearance from the source stream to the target stream guided by their dense spatial correspondence and multi-scale context.
- The proposed appearance transfer module is the first of its kind to use the target stream to query and transfer the source stream. It effectively handles the difficulty of large motion.
- The proposed multi-scale context module is the first attempt to apply atrous convolutions for contextual modeling in person image generation. Multi-scale context helps the network recover occluded pixels.
- Compared with state-of-the-art methods, our network achieves comparable or superior performance using much fewer parameters while being significantly faster. We also show our network has a great advantage when large pose transform occurs.

2. Related work

2.1. Image generation

Image generation is a basic task in computer vision. Most recently, Generative Adversarial Networks [7] based methods have been widely used for synthesizing realistic images and achieved good performance in various tasks, i.e., image-to-image translation [8,11,23–25], text-to-image translation [26–28] and image in painting [29,30]. Most image-to-image translation models are based on Conditional Generative Adversarial Networks(CGANs) [12] because it can achieved remarkable success in pixel-wise aligned image generation problems. However, pixel-wise alignment is not suitable for pose transfer due to the pose deformation.

The CocosNet introduced by Zhang et al. [31] is very related to our appearance transfer module. It translates the image via the dense correspondence between conditioned input and a given style exemplar, and the correspondence map is computed in a similar way as ours. However, our approach differs from CocosNet in several aspects. First, CocosNet is inspired by the correspondence layer proposed in Zhang et al. [32] and uses the correspondence map to warp the exemplar image. By contrast, our network is inspired by the self-attention mechanism [17]; it maintains two feature streams and employs the correspondence map to progressively transfer the appearance information from the source feature stream to the target feature stream. The different sources of inspiration also make the calculations of the correspondence map before the softmax normalization different: cosine similarity between two feature vectors with mean removed in Zhang et al. [32] versus inner product between unnormalized feature vectors in the self-attention. Second, CocosNet injects the warped exemplar image into the translation network through positional normalization and spatially-variant denormalization (SPADE) [33] to produce the translated image. By contrast, we directly feed the feature map from the target stream into the decoder to generate our final results. Third, CocosNet has only one correspondence learning module to warp the exemplar image while we find it beneficial to stack multiple appearance transfer modules to progressively transfer the appearance information from the source stream to the target stream. In addition, our network includes a multi-scale context module, which is the first attempt to apply atrous convolutions for contextual modeling in person image generation.

2.2. Person image generation

The task of pose guided person image generation was first introduced by Ma et al. [34]. Their two-stage network first generates a coarse target image and then refines it in an adversarial way. Ma et al. [35] disentangle the foreground, background and pose information, and then manipulate them to get the desired pose. The controllability of the generation process is improved, but the quality of the generated image is reduced. Esser et al. [36] combine VAE [37] and U-Net [8] to distinguish the appearance and pose of a person image. However, it is difficult to represent the appearance features as a low-dimension underlying code, which unavoidably loses information. Siarohin et al. [13] introduce deformable skip connections to transform the texture spatially. It uses a set of local affine transformations to decompose the overall articulated body deformation. However, it cannot handle occlusion or out-of-plane rotation well. The pose attention transfer network (PATN) [2] consists of an image stream and a pose stream, and it uses an attention mask to enhance the feature maps. However, it only processes features locally, and there is no explicit geometric manipulation or appearance transfer of the source image.

Most recently, Tang et al. [38] propose a cycle-in-cycle GAN, which is a cross-modal framework exploring joint exploitation of the keypoints and the image data in an interactive manner. Ren et al. [39] introduce a differentiable global-flow local-attention framework to reassemble the inputs at the feature level. Men et al. [40] propose the attribute-decomposed GAN, which means to embed human attributes into the latent space as independent codes and thus achieve flexible and continuous control of attributes via mixing and interpolation operations in explicit style representa-

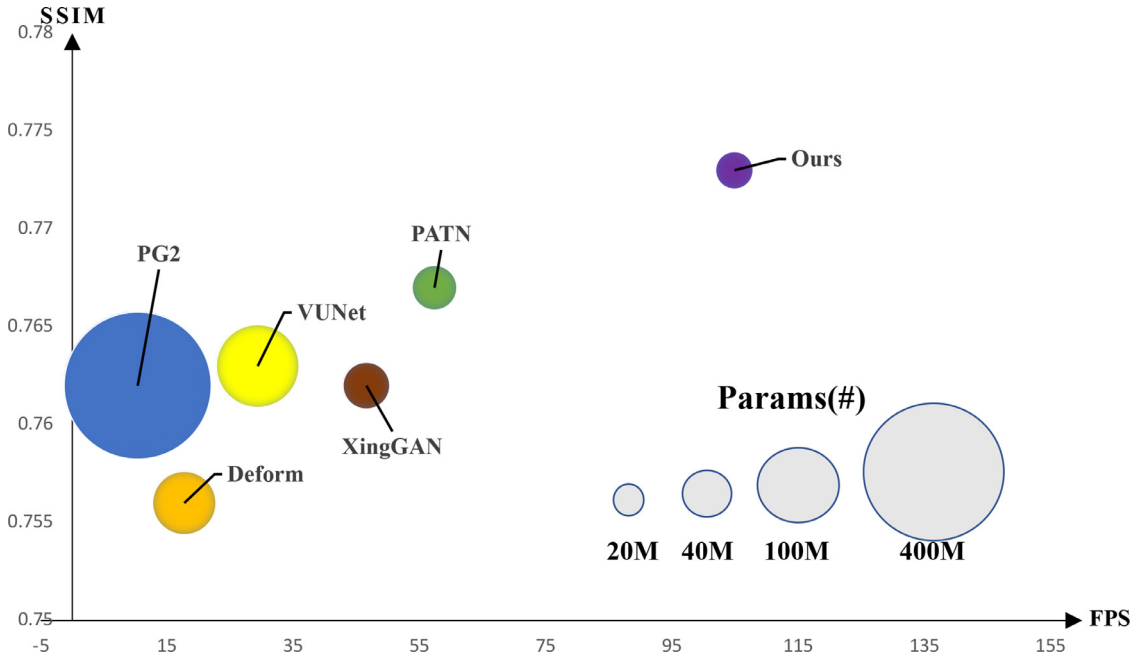

Fig. 1. Comparison with recent state-of-the-art approaches on the DeepFashion dataset. Our approach (purple) achieves superior quality, and is much more efficient. The results of PATN [2] and XingGAN [16] are reproduced by us. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

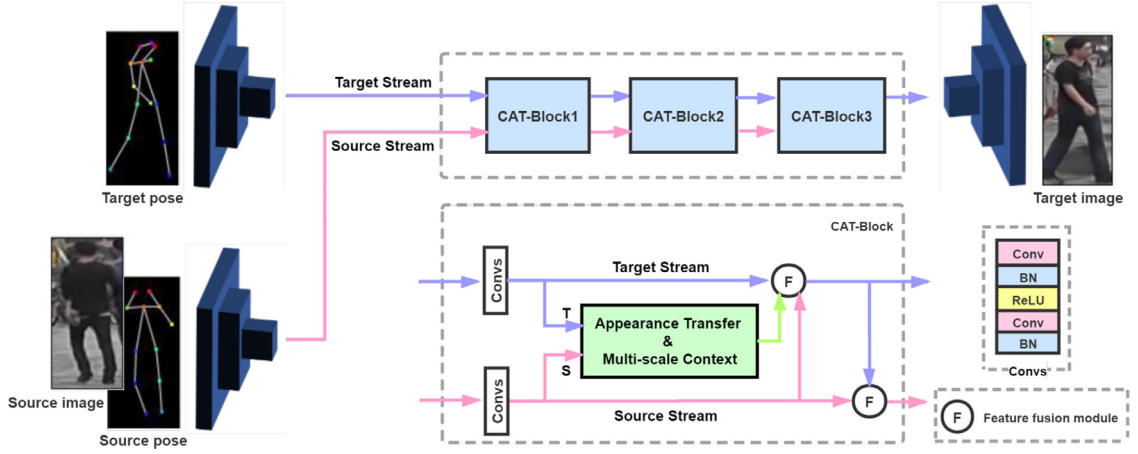

Fig. 2. An overview of the proposed network architecture. It is a three-stage architecture consisting of two streams. The target and source streams respectively take as input the target pose, and the source pose and image, and pass them through convolutional encoders. Each stage is a context-aware appearance transfer block (CAT-block). It consists of an appearance transfer module, a multi-scale context module and two-stream feature fusion modules. The target feature map from the last CAT-block passes through a convolutional decoder to generate a new person image with the same appearance as the source image but in the target pose.

tions. Huang et al. [41] introduce an appearance-aware pose stylizer, which generates human images by coupling the target pose with the conditioned person appearance progressively. Lathuiliere et al. [42] employ the local attention mechanism to select relevant information from multi-source human images for human image generation. RATE-Net [43] leverages an additional texture enhancing module to extract appearance information from the source image and estimate a fine-grained residual texture map. This helps refine the coarse estimation from the pose transfer module. Gao et al. [44] propose a portrait photo recapture system with two modules that complement each other from both intra-part and inter-part perspectives to easily transform their portraits to the desired posture.

The work most related to ours is the XingGAN recently introduced by Tang et al. [16]. It uses the self-attention to achieve bidirectional non-local communication between features of source and target poses and features of the source image. Our approach is different from XingGAN. We find the dense spatial correspondence between the source stream and target stream (like flow-based methods) to transfer the appearance information from the source image to the target image. It is a unidirectional process with explicit semantic meaning and well aligned with the task of person image generation. By contrast, XingGAN iteratively updates the shape and appearance embeddings in a non-local manner. It does not compute the correspondence between the source image/pose and the target one, nor does it perform any explicit appearance transfer between them. In addition, we model multi-scale context, which XingGAN ignores. Since our network is explicitly designed to handle large motion and occlusion, it not only generates higher-quality person images in case of large pose transform but also has a much smaller model size and a significantly higher inference speed.

Several approaches adopt DensePose [45], 3D pose [46], or human parsing [47] to generate person images since they contain more information, e.g., the body part segmentation or depth. However, the keypoint-based pose representation is much cheaper to obtain and more flexible. Therefore, we prefer to use a keypoint-based representation.

2.3. Self-attention

The self-attention [48,49] was first introduced for natural language processing. It calculates the response of a certain position in the sequence by paying attention to all positions in the same sequence. Vaswan et al. [17] prove that the machine translation model could obtain the state-of-the-art results using the self-attention. Parmar et al. [50] introduce an image transformer model that adds the self-attention to an automatic regression model for image generation. Wang et al. [51] formulate the self-attention as a non-local operation to model the spatial-temporal dependencies in video sequences. Liu et al. [52] propose a Dual Self-Attention with Co-Attention networks to model the internal dependencies of both the spatial and sequential structure respectively by using the self-attention mechanism. Wei et al. [53] propose an attention-based model (called position-aware self-attention) as well as a well-designed self-attentional context fusion layer within a neural network architecture, to explore the positional information of an input sequence for capturing the latent relations among tokens. Zhang et al. [54] propose a self-attention GAN enforcing the generator to gradually consider non-local relationships in the feature space. It can learn to find long-range dependencies within internal representations of images.

Although our appearance transfer module is inspired by the self-attention, they are significantly different. The queries, keys and values in our network are specially designed for pose-guided appearance transfer, and they are semantically different. By contrast, these items in the self-attention are obtained from the same input. As a result, the self-attention models the non-local relations within a single feature map while our network finds the spatial correspondence between the source stream and the target stream to perform appearance transfer.

3. Our approach

3.1. Overview

As illustrated in Fig. 2, our network is a three-stage architecture consisting of two streams. The input of the target stream is the target pose $\mathbf{P}_t$. The input of the source stream is the concatenation of the source pose $\mathbf{P}_s$ and the source image $\mathbf{I}_s$. Both source and target poses are represented as keypoint heatmaps. The output of the network is a generated target image $\mathbf{I}_t$ containing the same person as the source image $\mathbf{I}_s$ but in the target pose $\mathbf{P}_t$.

The network first uses two encoders to produce initial feature maps for the two streams. Each encoder consists of two downsampling convolutional layers, and they do not share weights. The initial source features contain both appearance and structure information while the initial target features contain only structure information. Then, a cascade of three context-aware appearance transfer blocks (CAT-blocks) progressively transfer the appearance from the source stream to the target stream guided by the structure information and multi-scale context. All CAT-blocks have the same architecture but do not share weights. Finally, the target feature map from the last CAT-block passes through a decoder to generate the target image. The decoder consists of two deconvolutional layers. We will detail the CAT-block in Section 3.2 and the loss function in Section 3.3.

3.2. Context-aware appearance transfer block

As shown in Fig. 2, a context-aware appearance transfer block (CAT-block) takes as input the two-stream feature maps $\mathbf{F}_s \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{F}_t \in \mathbb{R}^{C\times H\times W}$ obtained from the previous block or the encoder and outputs their updated feature maps $\mathbf{F}'_s \in \mathbb{R}^{C\times H\times W}$ and $\mathbf{F}'_t \in \mathbb{R}^{C\times H\times W}$. Here C, H and W respectively denote the channels, height and width of a feature map, and the subscripts s and t respectively indicate the source and target streams. A CAT-block consists of an appearance transfer module, a multi-scale context module and two-stream feature fusion modules, which are detailed below. Unless otherwise specified, the kernel size of a convolutional layer is set to 3×3.

Appearance transfer module The pipeline of an appearance transfer module is illustrated in Fig. 3. We first pass the two-stream feature maps $\mathbf{F}_s$ and $\mathbf{F}_t$ through convolutions and reshape the results as $\mathbf{S} \in \mathbb{R}^{C\times HW}$ and $\mathbf{T} \in \mathbb{R}^{C\times HW}$, respectively. Then we feed them into 1×1 convolution layers (implemented as matrix multiplications) to produce three matrices $\mathbf{K} \in \mathbb{R}^{\bar{C}\times HW}$, $\mathbf{V} \in \mathbb{R}^{\hat{C}\times HW}$ and $\mathbf{Q} \in \mathbb{R}^{\bar{C}\times HW}$:

$$\mathbf{K} = \mathbf{W}_k\mathbf{S} \tag{1}$$

$$\mathbf{V} = \mathbf{W}_v\mathbf{S} \tag{2}$$

$$\mathbf{Q} = \mathbf{W}_q\mathbf{T} \tag{3}$$

where $\mathbf{W}_k, \mathbf{W}_q \in \mathbb{R}^{\bar{C}\times C}$, $\mathbf{W}_v \in \mathbb{R}^{\hat{C}\times C}$ are learnable weight matrices. We set $\bar{C} = C/8$, $\hat{C} = C/2$ for memory efficiency, and it does not cause a significant performance drop. Each column of $\mathbf{K}$, $\mathbf{V}$ or $\mathbf{Q}$ is a *key*, a *value* or a *query* respectively. Our appearance transfer module means to match (target) queries to the (source) keys and then use the correspondence to transfer the relevant (source) values from the source stream to the target stream.

To achieve this goal, we first obtain a correspondence map $\mathbf{D} \in \mathbb{R}^{HW\times HW}$ by applying a softmax normalization to each row of $\mathbf{Q}^T\mathbf{K}$:

$$\mathbf{D}_{ij} = \frac{\exp(\mathbf{Q}_i^T\mathbf{K}_j)}{\sum_{j=1}^{HW}\exp(\mathbf{Q}_i^T\mathbf{K}_j)} \tag{4}$$

where $\mathbf{D}_{ij}$ is the (i, j)th element of $\mathbf{D}$, $\mathbf{Q}_i$ is the ith column of $\mathbf{Q}$, $\mathbf{K}_j$ is the jth column of $\mathbf{K}$. $\mathbf{D}_{ij}$ is a *soft* correspondence score between the ith query, i.e., the ith position in the target feature map, and the jth key, i.e., the jth position in the source feature map. We can interpret the ith row of $\mathbf{D}$ as a probability distribution of each key matching the ith query. The correspondence map serves as the basis of appearance transfer.

Then we retrieve the value for the ith query as a linear combination of the columns of $\mathbf{V}$ weighted by the ith row of $\mathbf{D}$. A matrix $\mathbf{W}_o \in \mathbb{R}^{C\times\hat{C}}$ is multiplied to the retrieved values to increase their dimension:

$$\mathbf{A} = \mathbf{W}_o\mathbf{V}\mathbf{D}^T \tag{5}$$

where $\mathbf{A} \in \mathbb{R}^{C\times HW}$ is the appearance information to be transferred from the source stream to the target stream. During the *query-and-transfer* process, the source appearance is aligned with the target pose. Since the alignment is non-local, our appearance transfer module can handle large motion.

After a scaled residual connection $\mathbf{A}' = \sigma\mathbf{A} + \mathbf{T}$ (σ is a learnable scalar) and a subsequent convolutional layer, the appearance transfer module outputs a feature map whose shape is $C \times H \times W$.

Multi-scale context module Not all content of the target image can be found in the source image because of occlusion. To help the target stream recover occluded pixels, it is necessary to give the network access to richer context and even a global view of both streams.

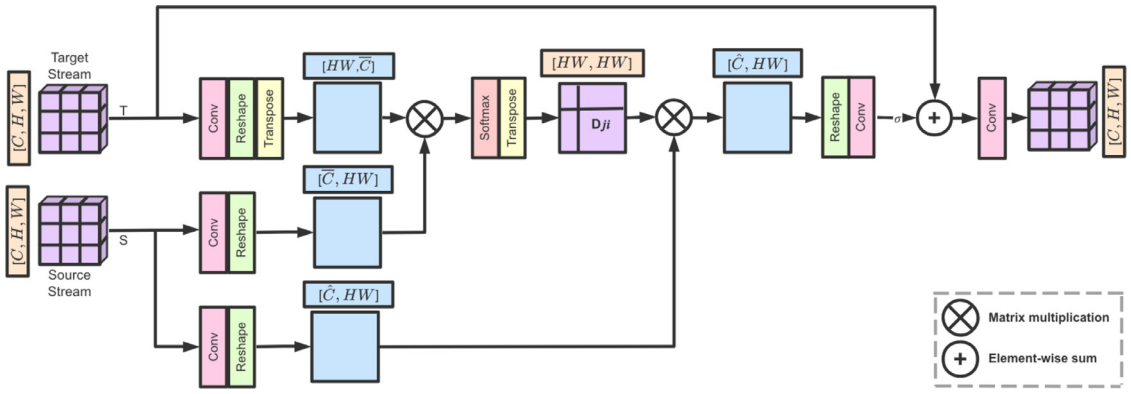

Fig. 3. Illustration of the proposed appearance transfer module. It calculates the correspondence between the feature vector at each location in the target stream and the feature vector at each location in the source stream. Then the correspondence is used to transfer the appearance information from the source stream to the target stream.

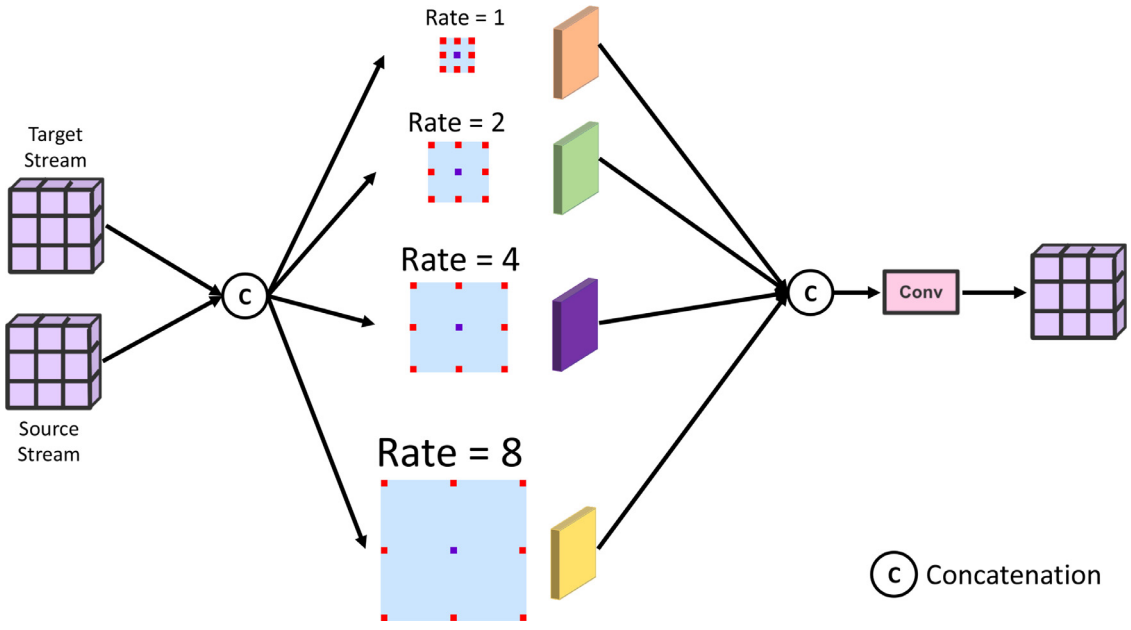

Fig. 4. Illustration of the multi-scale context module.

To this end, we build a multi-scale context module based on the atrous spatial pyramid pooling (ASPP) [18] widely used in image segmentation. Its pipeline is illustrated in Fig. 4. We first concatenate the two-stream feature maps $\mathbf{F}_s$ and $\mathbf{F}_t$, and pass them through several parallel atrous convolutions with different sampling rates r = 1, 2, 4, 8, respectively:

$$\mathbf{H}_r = \text{AstrousConv}_r(\text{Concat}(\mathbf{F}_s, \mathbf{F}_t)) \tag{6}$$

where AstrousConv_r is the atrous convolution with a sampling rate r. Then we concatenate the output feature maps enriched with different scales of context $\{\mathbf{H}_r\}$ and pass them through a convolutional layer.

$$\mathbf{H} = \text{Conv}(\text{Concat}(\{\mathbf{H}_r\}) \tag{7}$$

where $\mathbf{H} \in \mathbb{R}^{C\times H\times W}$ is the output of the multi-scale context module.

Two-stream feature fusion modules As shown in Fig. 2, the features in the target stream are updated by fusing the features in the source stream, the transferred appearance and multi-scale context. The features in the source stream are updated by fusing the new target features. We find that a simple fusion module composed of a concatenation operation and a subsequent convolutional layer (output channels set to C) works well. The two-stream feature fusion modules are important as they allow local information exchange between the two streams, which supplements the non-local appearance transfer and multi-scale context modeling.

3.3. Loss function

The full loss function is:

$$\mathcal{L} = \arg\min_G \max_D \alpha_g \mathcal{L}_{GAN} + \alpha_1 \mathcal{L}_1 + \alpha_p \mathcal{L}_p \tag{8}$$

where $\mathcal{L}_{GAN}$, $\mathcal{L}_1$ and $\mathcal{L}_p$ respectively denote the adversarial loss, the ℓ_1-norm loss and the perceptual loss, and α_g, α_1 and α_p represent their respective weights. $\mathcal{L}_1$ calculates the ℓ_1-norm distance between the generated image $\mathbf{I}_t$ and the ground truth target image $\mathbf{I}_{gt}$: $\ell_1 = ||\mathbf{I}_{gt} - \mathbf{I}_t||_1$. The perceptual loss $\mathcal{L}_p$ has been widely used for image generation and translation [9,10,13,36] as it helps generate more realistic and smoother images. It is defined as:

$$\mathcal{L}_p = \frac{1}{W_\rho H_\rho C_\rho} ||\phi_\rho(\mathbf{I}_{gt}) - \phi_\rho(\mathbf{I}_t)||_1 \tag{9}$$

where ϕ_ρ is the output of the conv1_2 layer from the VGG-19 model [55] pretrained on ImageNet [56], and W_ρ, H_ρ, C_ρ are the

width, height and depth of ϕ_ρ, respectively. We adopt the adversarial loss introduced in Zhu et al. [2]. It consists of an appearance discriminator and a shape discriminator to determine the possibility that the generated image contains the same person in the input image and the degree to which the generated image is aligned with the target pose.

4. Experiment

Datasets We use two challenging person image datasets: Market-1501 [21] and DeepFashion [22]. The resolution of images in DeepFashion is higher (256×256) than that in Market-1501 (128×64). We employ OpenPose [58] to detect human body joints. Both the source and target poses consist of an 18-channel heatmap encoding the positions of 18 human body joints. There are 263,632 pairs of training images in Market-1501, and 101,966 pairs in DeepFashion. Their testing sets contain 12,000 pairs and 8570 pairs, respectively. Note the person identities of the training set do not overlap with those of the testing set.

Evaluation metrics We follow [2,13,34] and adopt Structure Similarity (SSIM) [59], Inception Score (IS) [60], and their masked versions, i.e., Mask-SSIM and Mask-IS, as the evaluation metrics. We also use other common metrics such as Learned Perceptual Image Patch Similarity (LPIPS) [61] and Fréchet Inception Distance (FID) [62]. LPIPS and FID calculate the perceptual distance between the generated images and ground truth images in the feature space w.r.t. each pair of samples and the global distribution, respectively. Moreover, we adopt the PCKh score proposed in Zhu et al. [2] to assess the shape consistency.

Implementation details Our method is implemented in PyTorch using two NVIDIA GeForce RTX 2080 Ti GPUs. The Adam optimizer [63] is adopted to train the proposed model for around 90k iterations with $\beta_1 = 0.5$, $\beta_2 = 0.999$. The learning rate is fixed as 0.0001 in the first 60k iterations and then linearly decayed to 0 in the last 30k iterations. We use 3 CAT-blocks in the generator for both datasets. For the hyper-parameters, $(\alpha_g, \alpha_1, \alpha_p)$ are set as (5, 1, 1) for DeepFashion and (5, 4, 4) for Market-1501, respectively. Instance normalization [64] is applied for both datasets. The batch size is set as 7 for DeepFashion and 32 for Market-1501. Dropout [65] is only used in the CAT-blocks, and the dropout rate is set to 0.5. Leaky ReLU [66] is applied after every convolution or normalization layer in the discriminators, and its negative slope coefficient is set to 0.2.

4.1. Comparison with state-of-the-art methods

Quantitative and qualitative results We compare the proposed network with several state-of-the-art methods such as DPIG [35], VUnet [36], Deform [13], PATN [2], BTF [57], C2GAN [38], ADG [40], XingGAN [16] and APS [41]. Table 1 shows the quantitative results measured by SSIM, IS, Mask-SSIM, Mask-IS, and PCKh metrics. Our network achieves the best performance under most metrics on the

Fig. 5. Qualitative comparison on Market-1501.

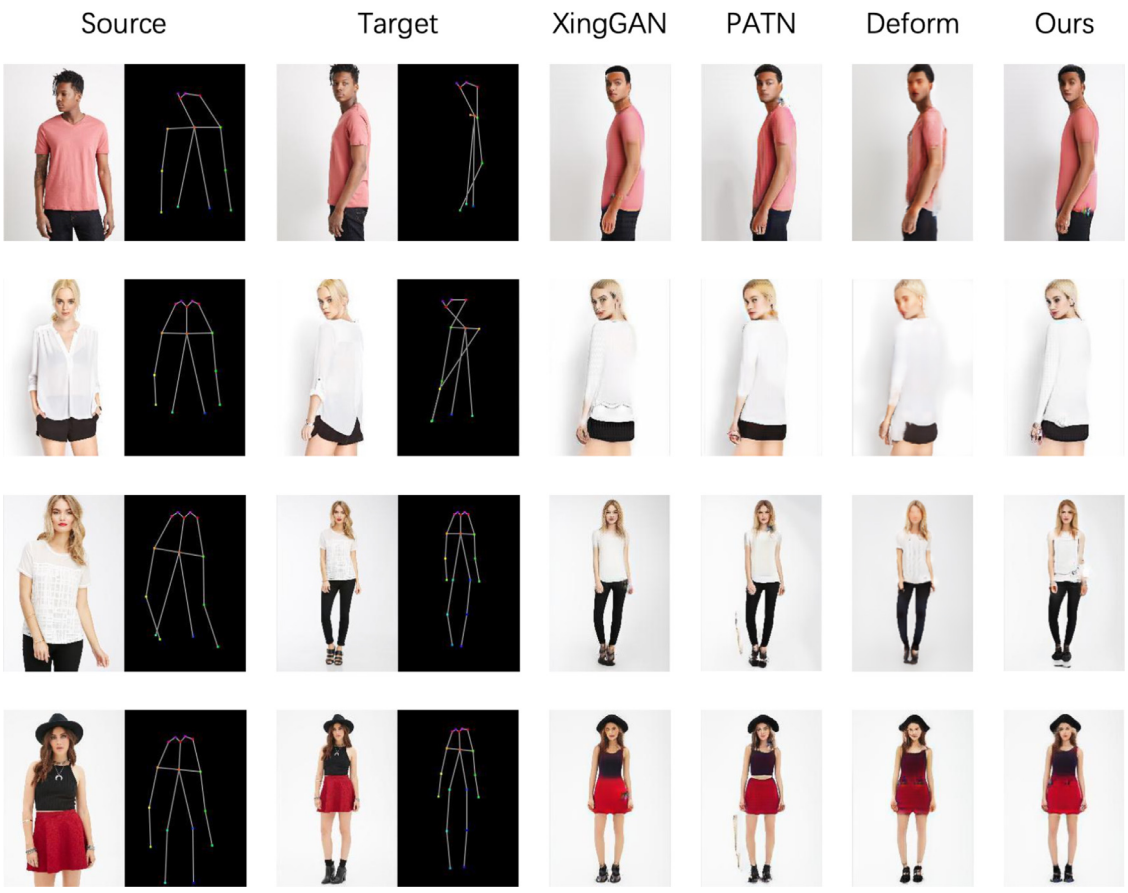

Fig. 6. Qualitative comparison on DeepFashion.

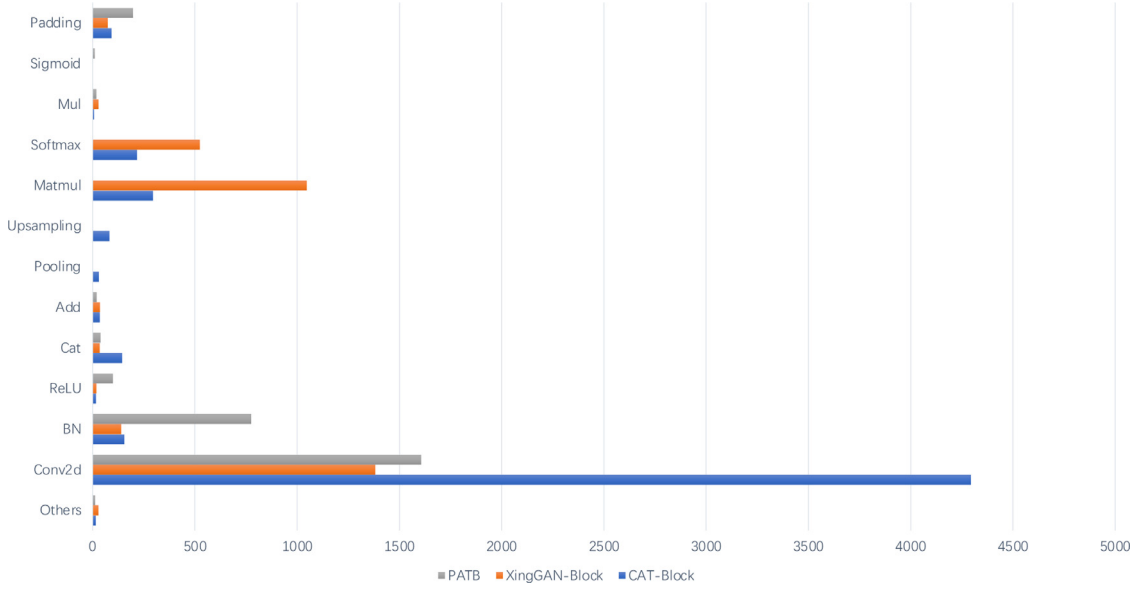

Fig. 7. We use PyTorch Profiler to obtain the GPU running time (μs) of each operation in a CAT-block, a XingGAN-block and a PATB. The PATB is a building block of PATN. The results are collected on the same image from DeepFashion. Conv2d, BN, Mul, Cat, Matmul, Pooling, and Upsampling represent convolution, batch normalization, multiplication, concatenation, matrix multiplication, average pooling and bi-linear upsampling, respectively.

two datasets. Figs. 5 and 6 show that the appearance or texture generated by the proposed method is more consistent and appealing than the others.

User study We conducted user study with 30 volunteers to give an instant judgment (real/fake) about each image within a second. R2G means the percentage of real images rated as generated w.r.t. all real images. G2R means the percentage of generated images rated as real w.r.t. all generated images. Our R2G and G2R scores are respectively 42.32 and 75.68. By contrast, the two scores of PATN are respectively 32.23 and 63.47. These results indicate the images generated by our network are more realistic.

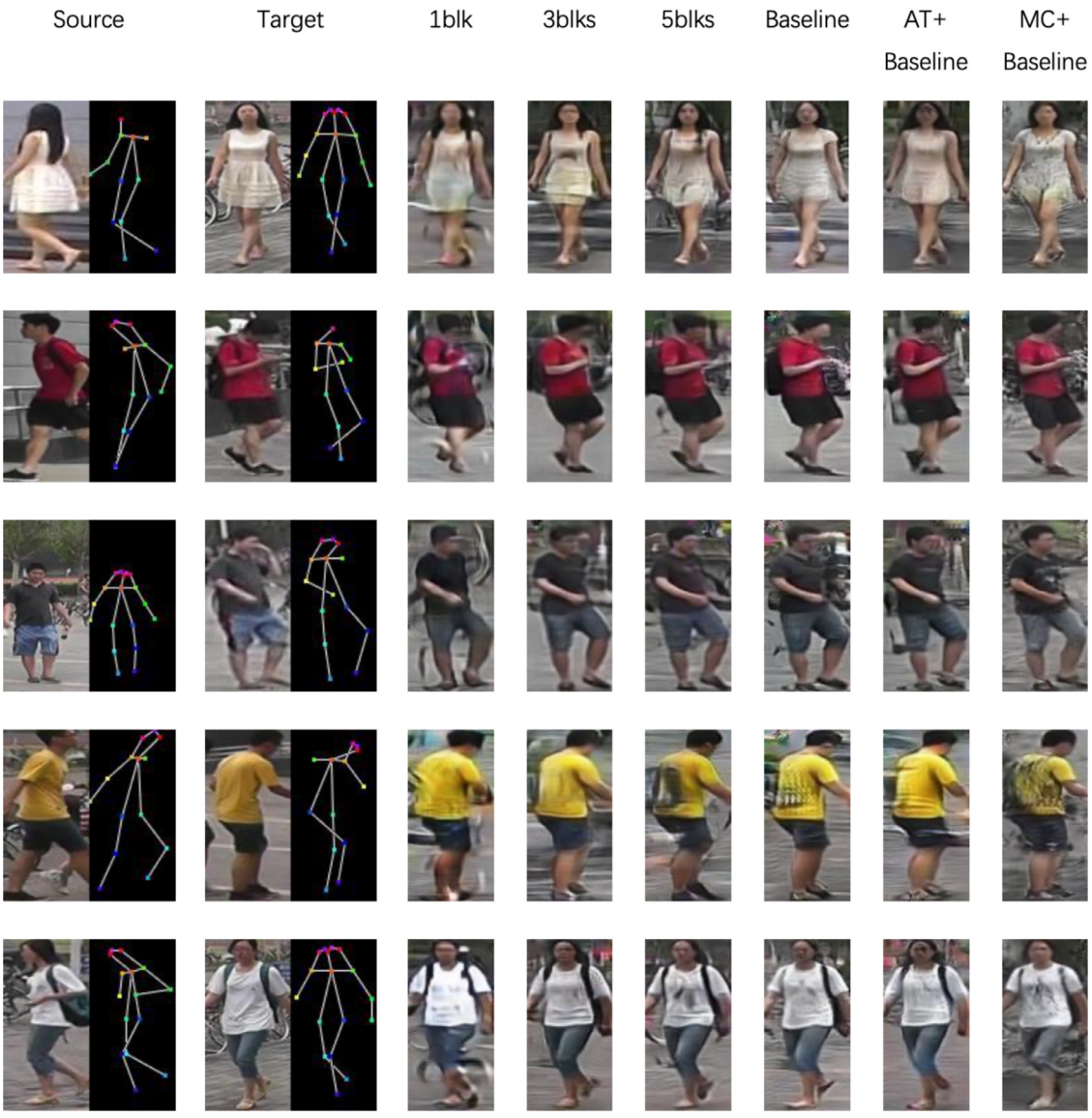

Fig. 8. Ablation study of our network on Market-1501. Columns 1 and 2: input and ground truth output and their pose maps. Columns 3–5: ablation study on the number of CAT-blocks of our full network. Columns 6–8: ablation study on effectiveness of the appearance transfer module (AT) and the multi-scale context module (MC).

Table 1
Quantitative results on Market-1501 and DeepFashion. (∗) denotes the results reproduced by us using the code released by the authors. For SSIM, IS, Mask-SSIM, Mash-IS and PCKh, higher values indicate better performance. For FID and LPIPS, lower values indicate better performance.

Method	#Blocks	Market-1501							DeepFashion				
		SSIM	IS	Mask-SSIM	Mask-IS	PCKh	FID	LPIPS	SSIM	IS	PCKh	FID	LPIPS
DPIG [35]	–	0.099	3.483	0.614	3.491	–	–	–	0.614	3.228	–	–	–
VUNet [36]	–	0.266	2.965	0.793	3.549	0.92	21.214	0.321	0.763	3.440	0.93	23.836	0.264
Deform [13]	–	0.290	3.185	0.805	3.502	–	29.035	0.299	0.756	3.439	–	26.283	**0.233**
PATN [2]	9	0.311	3.323	0.811	3.773	0.94	–	–	0.773	3.209	0.96	–	–
BFT [57]	–	–	–	–	–	–	–	–	0.767	3.220	–	–	–
C2GAN [38]	–	0.282	3.349	0.811	3.510	–	–	–	–	–	–	–	–
ADG [40]	–	–	–	–	–	–	–	–	0.772	3.364	–	–	–
APS [41]	–	0.312	3.132	0.808	3.729	0.94	–	–	0.775	3.295	0.96	–	–
XingGAN [16]	9	0.313	**3.506**	0.816	3.872	0.93	–	–	**0.778**	**3.476**	0.95	–	–
PATN* [2]	9	0.301	3.344	0.805	3.773	0.94	22.657	0.319	0.767	3.209	0.96	21.563	0.249
XingGAN* [16]	9	0.305	3.425	0.806	**3.883**	0.93	22.307	0.302	0.762	3.209	0.95	33.414	0.282
Ours	**3**	**0.322**	3.318	**0.816**	3.780	**0.94**	**20.455**	**0.298**	0.773	3.216	**0.96**	**19.628**	0.251
Real Data	–	1.000	3.890	1.000	3.706	1.00	4.854	0.000	1.000	4.053	1.00	7.785	0.000

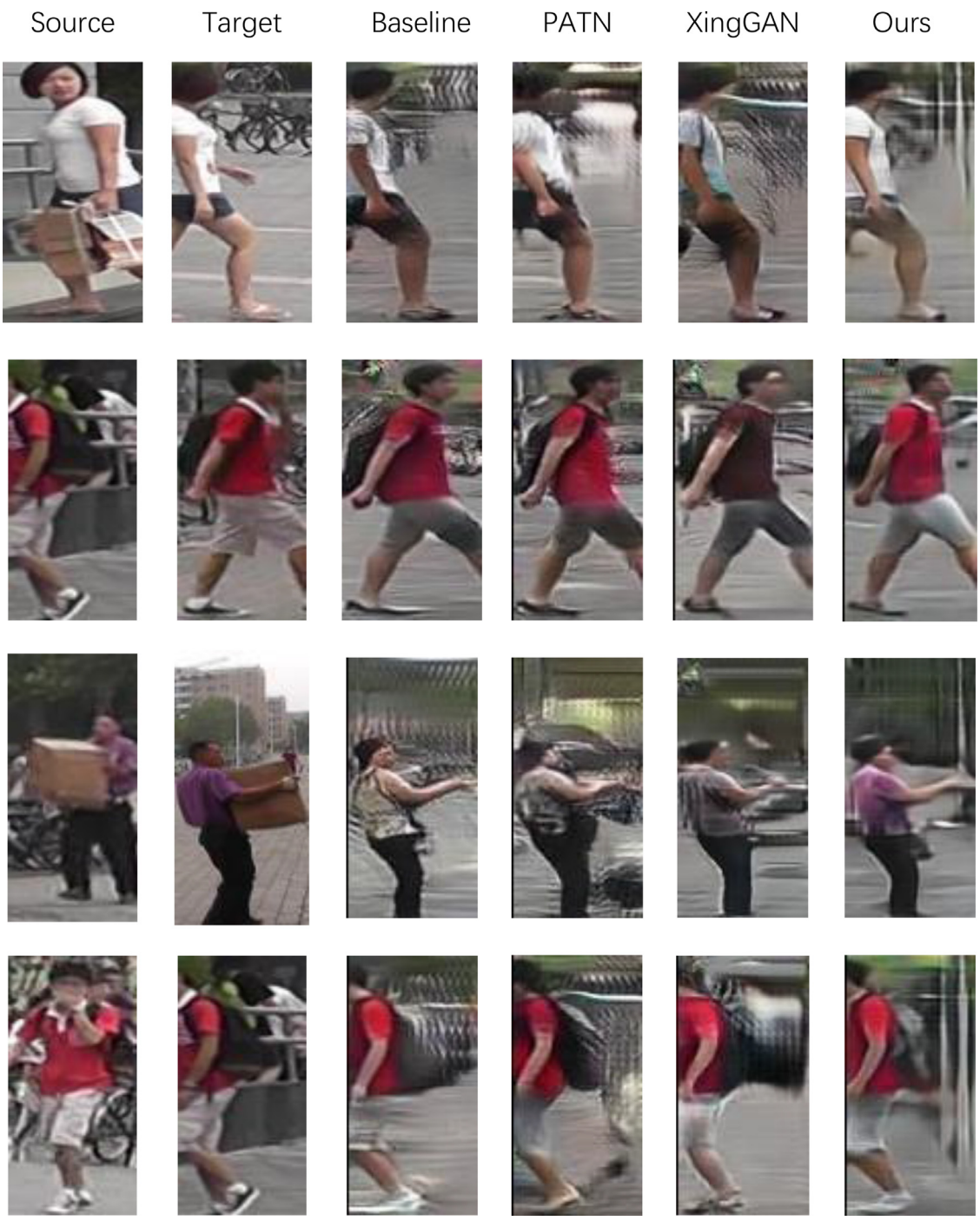

Fig. 9. Results obtained on cases of large pose transform. The scaled cosine distances between source and target poses are greater than 0.7.

Comparison of model sizes and inference speeds Table 3 reports the GPU running time of a single block, the encoder, the decoder, and the whole network of our approah, PATN and XingGAN obtained via Pytorch Profiler. We can see our single block is slower than those of PATN and XingGAN. But our overall network is the fastest because it contains only 3 blocks instead of 9 blocks in PATN and XingGAN, and our encoder and decoder are the fastest. We also measure the overall inference speeds of these networks via timer functions (put before and after the networks) instead of the profiling tool. The results are shown in Table 2 and Fig. 1. We can see that our network is significantly faster than PATN and XingGAN, and its model size is much smaller. It is worth noting the inference speeds of all networks in Table 3 measured by PyTorch Profiler are slower than those in Table 2 measured via timer functions through all other settings are the same. This is because the profiling tool brings extra overhead; GPU running time measured via timer functions more accurately reflects the speed of a network in practice.

Fig. 7 provides detailed comparison of running time cost by each operation in a CAT-block, a XingGAN-block, and a PATB (a block of PATN). It indicates a CAT-block is slower than a PATB and a XingGAN-block mainly because it spends more time on computing convolution and matrix multiplication. The convolution is the core operation in our multi-scale context module to access rich context

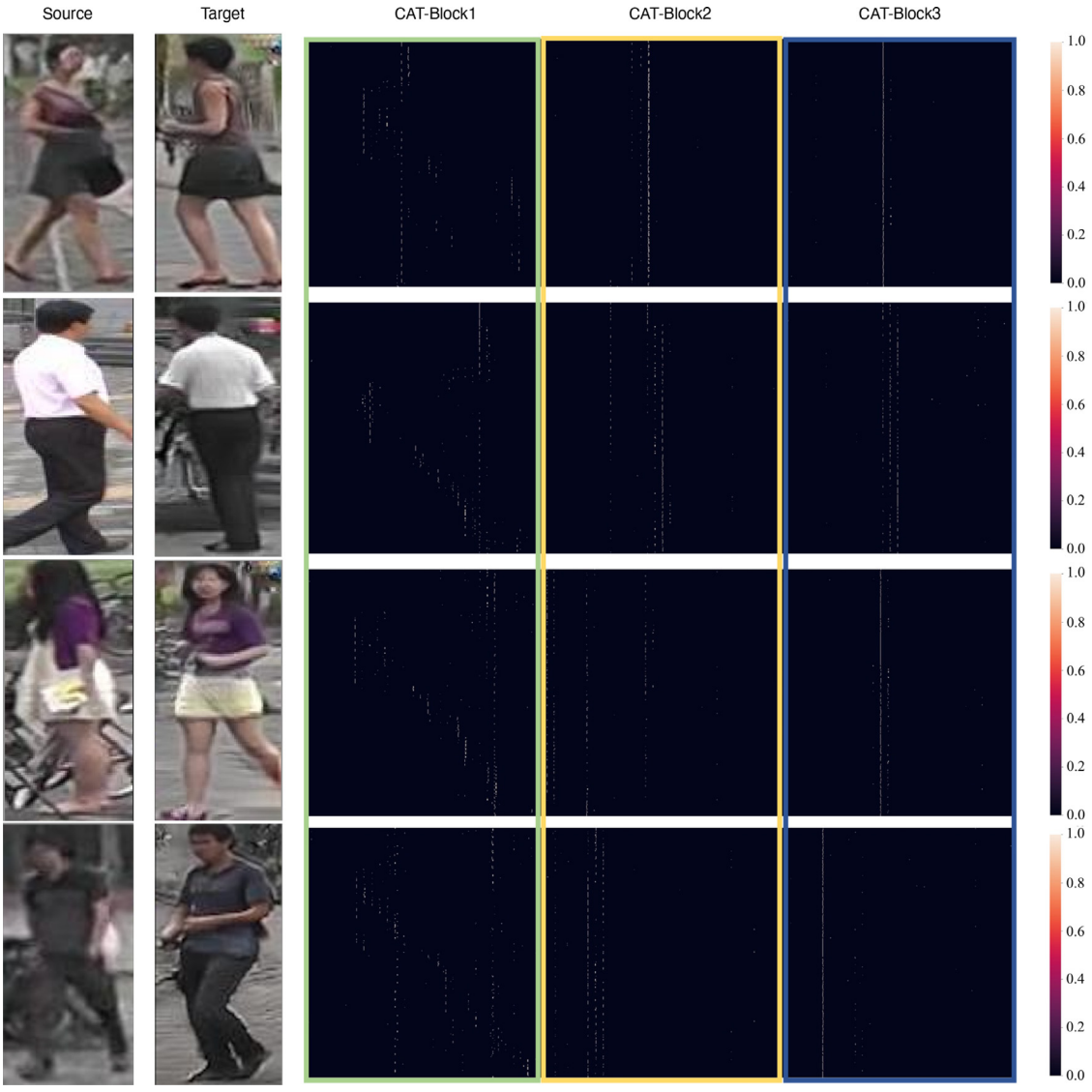

Fig. 10. Visualization of correspondence map in each CAT-block.

Table 2
Comparison of model sizes and inference speeds on DeepFashion. (∗) denotes the results reproduced by us using the code released by the authors.

Method	#Parameters	Speed
PG2 [34]	437.09 M	10.36 fps
Deform [13]	82.08 M	17.74 fps
VUNet [36]	139.36 M	29.37 fps
PATN* [2]	41.36 M	57.38 fps
XingGAN* [16]	44.85 M	46.59 fps
Ours	**29.13 M**	**104.89 fps**

Table 3
We use Pytorch Profiler to obtain the GPU running time (μs) comparison among a single block, the encoder, the decoder and the whole network of PATN, XingGAN and our approach. The results are collected on the same image from DeepFashion. "The First Block" means we obtain the GPU running time of the first building block of a network, i.e., a CAT-block, a XingGAN-block or a PATB. Note the profiling tool brings extra overhead and increases the running time.

Methods	The first block	Encoder	Decoder	Overall
PATN	**2777.46**	5902.48	3288.63	38631.67
XingGAN	3312.99	3522.05	10320.68	51124.76
Ours	5384.73	**3262.31**	**2555.53**	**24213.10**

and even a global view of the scene to address the difficulty of occlusion; the matrix multiplication is the core operation in our appearance transfer module to calculate the correspondence map between the source and target streams. However, our model only uses 3 CAT-blocks while XingGAN and PATN both use 9 blocks. Overall, our model uses fewer parameters and runs much faster than XingGAN and PATN.

4.2. Ablation study

In this section, we perform ablation studies to analyze the impact of each component in our model on performance. We conduct extensive ablation studies on Market-1501 datasets to evaluate different components of our network.

Effect of each module The results of the ablation study are shown in Table 4. The SSIM score compares the first-order and second-order statistics between patches in two images to measure their local structure similarity. The IS score uses a pre-trained image classifier to evaluate the quality of a generated image from the semantic perspective. It is worth noting the Mask-IS scores of all four methods in this table are higher than that of the ground truth. This means all these methods perform as well as the ground truth under this metric, and the Mask-IS scores have been satu-

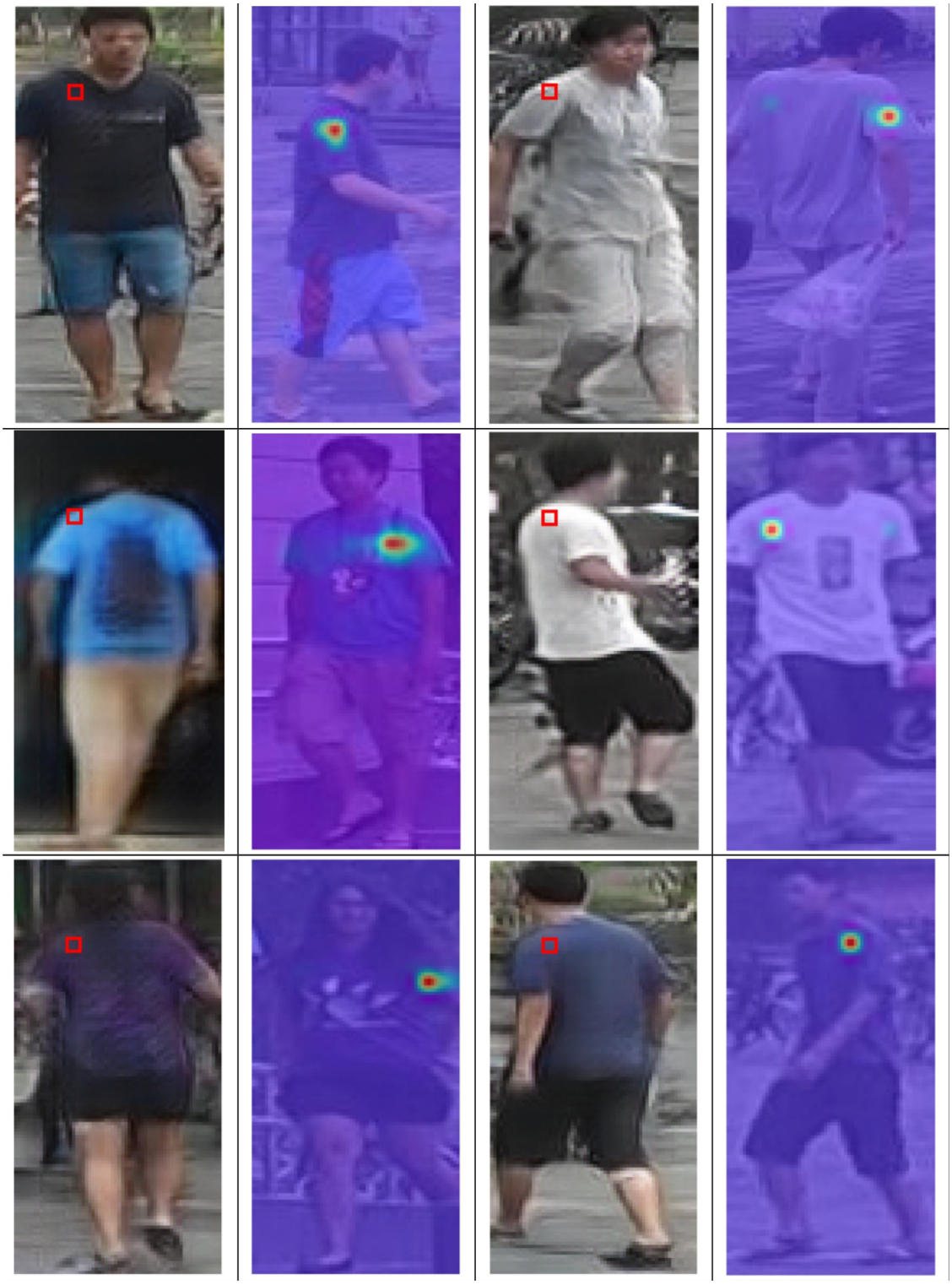

Fig. 11. Columns 1 and 3: the output target person images. Columns 2 and 4: the input person images. After selecting the shoulder location of the target feature map, we visualize which part of the source feature map is involved in the appearance transfer according to the correspondence map of the first CAT-block.

rated. Thus, we will focus on the performance comparison under the other three metrics below.

The AT module consistently improves the baseline under all three unsaturated metrics, i.e., SSIM, IS and Mask-SSIM. This means the AT module helps generate more realistic images from both the structure and semantic perspectives, thanks to its capability to find the dense correspondence between the source and target feature maps, and transfer the appearance information from the source stream to the target stream. The MC module increases the IS score of the baseline significantly. This means it helps generate images more like persons due to the multi-scale context modeling. Meanwhile, it achieves comparable performance with the baseline un-

Table 4
Ablation study on the appearance transfer (AT) module and the multi-scale context (MC) module. The **highest score** and the second highest score under each metric are highlighted. Note the Mask-IS scores of all four methods are higher than that of the ground truth real data. This means all these methods perform as well as the ground truth under this metric, and the Mask-IS scores have been saturated.

Method	Market-1501			
	SSIM	IS	Mask-SSIM	Mask-IS
Baseline	0.308	3.245	0.803	**3.783**
+ AT	0.315	3.301	0.811	3.775
+ MC	0.307	**3.370**	0.810	3.751
+ AT + MC	**0.322**	3.318	**0.816**	3.780
Real Data	1.000	3.890	1.000	3.706

Table 5
Ablation study on the number of CAT-blocks.

#CAT-blocks	Market-1501			
	SSIM	IS	Mask-SSIM	Mask-IS
1	0.281	**3.879**	0.798	3.676
3	**0.322**	3.318	**0.816**	**3.780**
5	0.317	3.285	0.815	3.757

der the SSIM metric and better performance under the Mask-SSIM metric. Our full model combines the AT module and MC module. It consistently outperforms the baseline under all three unsaturated metrics, and achieves the overall best performance among all methods. Concretely, the SSIM, IS and Mask-SSIM scores of the baseline are increased from 0.308, 3.245 and 0.803 to 0.322, 3.318 and 0.816, respectively.

Fig. 8 shows qualitative comparison of these ablation models. We have the following observations. (1) Compared with the baseline, the AT module alone helps generate cleaner images and more consistent appearance with the target, but it does not improve obviously on texture generation. (2) Compared with the baseline, the MC module alone helps generate more detailed texture, but sometimes too much texture is generated. (3) The full model combines the advantages of the AT module and the MC module, and avoids their respective limitations. Overall, it generates images with the most realistic texture and the most consistent foreground and background appearance with the target.

In addition, as we will show in Section 4.3, our full model achieves significant improvement over the baseline and state-of-art methods in the challenging scenario of large pose transform.

Effect of the number of CAT-blocks To further analyze the generation process, we conduct experiments by setting the number of CAT-blocks to 1, 3, 5, respectively. Quantitative and qualitative results are respectively shown in Table 5 and Fig. 8. We observe that the proposed generator works best and efficiently when it consists of 3 CAT-blocks. Increasing or decreasing the number of CAT-blocks may result in slightly worse quantitative and qualitative performance. Based on these observations and the visualization of correspondence maps in three blocks in Fig. 10, we find our network can achieve the pose transfer progressively. When there is only one CAT-block, the network starts to learn the correspondence between the source and target, but the network has not established a complete correspondence. This is why the results with only one CAT-block network are not good enough. The correspondences of more regions are learned in the second CAT-block, where these regions finish the appearance transfer procedure. In the third block, the correspondences of only fewer features need to be established for appearance transfer. In sum, 3 CAT-blocks enable the generator to transfer the necessary appearance information from the source stream to the target stream to generate the desired person image. However, it is worth noting that using 5 CAT-blocks still generates high-quality images, but the number of parameters increases significantly, and the inference is much slower. Therefore, we have used 3 CAT-blocks as the default setting in all experiments.

Table 6
Experiments on three subsets of testing cases obtained by setting different thresholds of scaled cosine distances between the source and target poses. The higher the threshold, the larger the pose transform. SSIM scores are reported. The baseline is constructed by removing the appearance transfer module and multi-scale context module from our network.

Threshold	0.7	0.5	0.3
Baseline	0.243	0.273	0.296
PATN [2]	0.234	0.261	0.276
XingGAN [16]	0.256	0.282	0.298
Ours	**0.263**	**0.294**	**0.314**

4.3. Experiment on large pose transform

This experiment means to verify whether the proposed approach can effectively handle large pose transform. It is the major challenge in the task of person image generation because it causes large motion and severe occlusions. We measure the degree of pose transform by calculating the cosine distance between the source and target pose vectors. It is defined as $1 - \mathbf{u}^T\mathbf{v}/(\|\mathbf{u}\|_2\|\mathbf{v}\|_2)$ for two vectors $\mathbf{u}$ and $\mathbf{v}$. Among all testing cases in the Market-1501 dataset, the maximum distance is 0.9623, and the minimum distance is 0.141. We linearly scale all distances so that the scaled maximum and minimum distances are 1 and 0, respectively. Then we create three subsets of testing cases by setting thresholds of the scaled distance to be 0.3, 0.5, 0.7, respectively. Note the higher the threshold, the larger the pose transform. The SSIM scores obtained by different approaches are shown in Table 6. Qualitative results are shown in Fig. 9. We observe that the proposed networks achieves the best performance on cases of large pose transform.

4.4. Visualization of the correspondence map

Fig. 10 visualizes the correspondence map in each CAT-block. We can interpret the ith row of a correspondence map as a probability distribution of each element in the source feature map matching the ith element in the target feature map. To gain more insights about how the proposed appearance transfer module works, we visualize which areas of the source feature map are involved to produce a feature vector in the target feature map in Fig. 11. We identify the location in the target feature map corresponding to a shoulder and visualize which part of the source feature map the network is paying attention to by reshaping the corresponding row in the correspondence map of the first CAT-block. As visualized in the attention maps in Fig. 11, the areas of the source feature map that are involved in the appearance transfer belong to the same semantic part as the selected target location. This visualization experiment verifies that our proposed method can find meaningful correspondences between the source stream and target stream (like flow-based methods) to transfer the appearance information from the source image to the target image.

4.5. Failure cases analysis

Fig. 12 illustrates failure cases obtained by our method. We also include images generated by some states of the art. Our results are imperfect in some significant challenging scenarios. For example, in the first two rows of Fig. 12, the target poses miss a few body joints in the lower body, which makes all models confusing. As a result, the output images miss some texture details in the corresponding areas. The case in the third row is tough because the pose transform is large and the source image lacks texture information in the man's backpack. So the bag in the output image con-

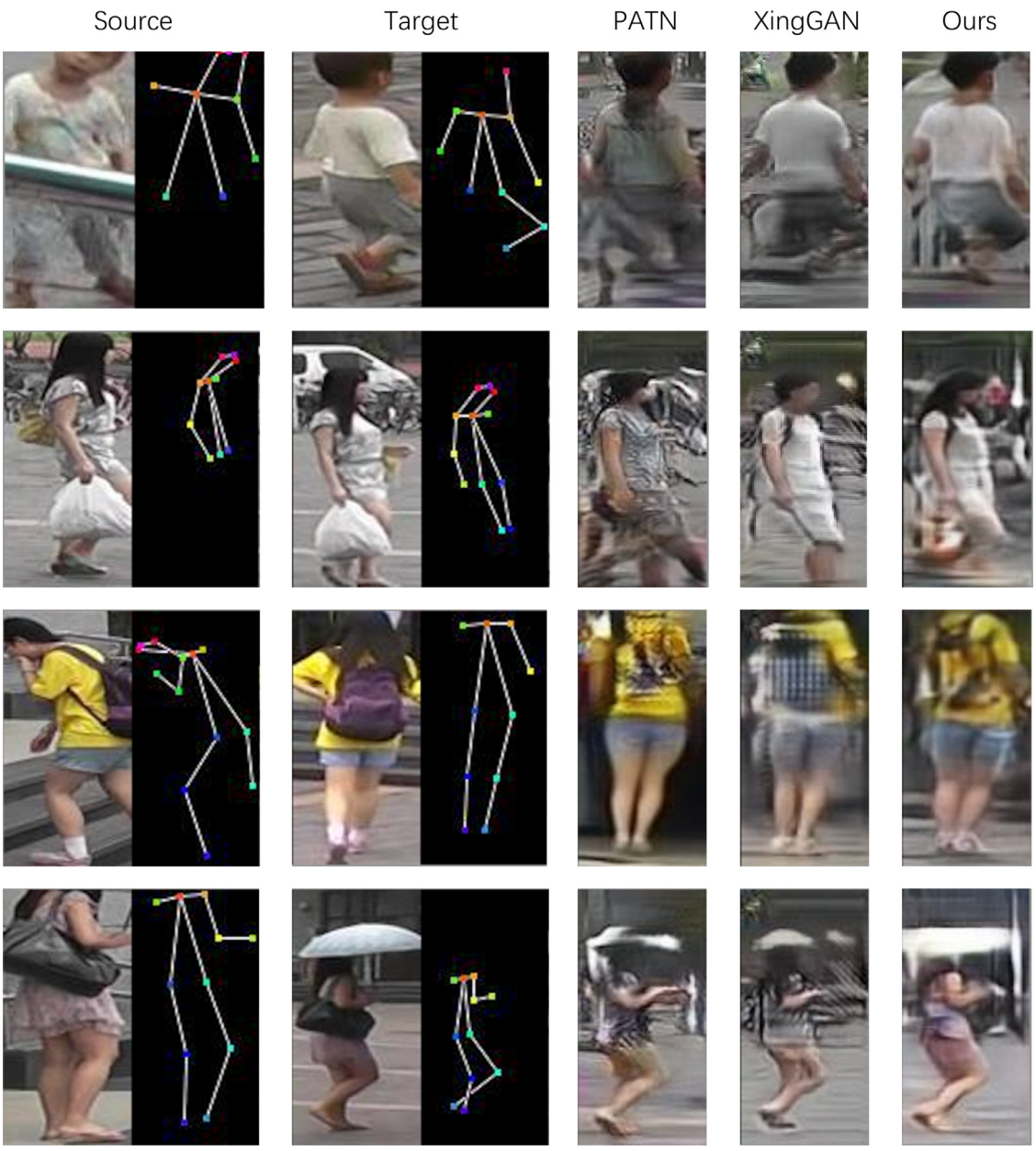

Fig. 12. Failure cases on Market-1501.

tains mixed color artifacts. In the last row, the umbrella is occluded in the source image, making the generation of it in the target image difficult and causing some artifacts.

5. Conclusion

This paper introduces a novel two-stream context-aware appearance transfer network for person image generation. It features an appearance transfer module to handle large motion and a multi-scale context module to handle occlusion. Experimental results show that our network is both effective and efficient. It has a great advantage on cases of large pose transform.

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.

Acknowledgment

This work was supported in part by Wei Tang's startup funds from the University of Illinois at Chicago and the National Science Foundation (NSF) award CNS-1828265.

References

[1] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, D. Lin, Pose guided human video generation, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 201–216.
[2] 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 Conference on Computer Vision and Pattern Recognition, 2019, pp. 2347–2356.
[3] X. Qian, Y. Fu, T. Xiang, W. Wang, J. Qiu, Y. Wu, Y.-G. Jiang, X. Xue, Pose-normalized image generation for person re-identification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 650–667.
[4] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[5] O. Ronneberger, P. Fischer, T. Brox, U-Net: convolutional networks for biomedical image segmentation, in: International Conference on Medical Image Computing and Computer-Assisted Intervention, Springer, 2015, pp. 234–241.

[6] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[7] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[8] 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.
[9] 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 Conference on Computer Vision and Pattern Recognition, 2017, pp. 4681–4690.
[10] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: European Conference on Computer Vision, Springer, 2016, pp. 694–711.
[11] 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 International Conference on Computer Vision, 2017, pp. 2223–2232.
[12] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv preprint arXiv:1411.1784 (2014).
[13] A. Siarohin, E. Sangineto, S. Lathuiliere, N. Sebe, Deformable GANs for pose-based human image generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3408–3416.
[14] X. Han, X. Hu, W. Huang, M.R. Scott, Clothflow: a flow-based model for clothed person generation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 10471–10480.
[15] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, S. Gao, Liquid warping GAN: a unified framework for human motion imitation, appearance transfer and novel view synthesis, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5904–5913.
[16] H. Tang, S. Bai, L. Zhang, P.H. Torr, N. Sebe, Xinggan for person image generation, arXiv preprint arXiv:2007.09278 (2020).
[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.
[18] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.
[19] X. Lian, Y. Pang, J. Han, J. Pan, Cascaded hierarchical atrous spatial pyramid pooling module for semantic segmentation, Pattern Recognit. 110 (2021) 107622.
[20] C. Peng, J. Ma, Semantic segmentation using stride spatial pyramid pooling and dual attention decoder, Pattern Recognit. 107 (2020) 107498.
[21] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, Q. Tian, Scalable person re-identification: a benchmark, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1116–1124.
[22] 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 Conference on Computer Vision and Pattern Recognition, 2016, pp. 1096–1104.
[23] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, J. Choo, Stargan: unified generative adversarial networks for multi-domain image-to-image translation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8789–8797.
[24] W. Xu, K. Shawn, G. Wang, Toward learning a unified many-to-many mapping for diverse image translation, Pattern Recognit. 93 (2019) 570–580.
[25] R. Li, W. Cao, Q. Jiao, S. Wu, H.-S. Wong, Simplified unsupervised image translation for semantic segmentation adaptation, Pattern Recognit. 105 (2020) 107343.
[26] Y. Dong, Y. Zhang, L. Ma, Z. Wang, J. Luo, Unsupervised text-to-image synthesis, Pattern Recognit. 110 (2021) 107573.
[27] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, D.N. Metaxas, Stackgan: text to photo-realistic image synthesis with stacked generative adversarial networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5907–5915.
[28] T. Xu, P. Zhang, Q. Huang, H. Zhang, Z. Gan, X. Huang, X. He, Attngan: fine–grained text to image generation with attentional generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1316–1324.
[29] D. Pathak, P. Krahenbuhl, J. Donahue, T. Darrell, A.A. Efros, Context encoders: feature learning by inpainting, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 2536–2544.
[30] X. Huang, Y. Li, O. Poursaeed, J. Hopcroft, S. Belongie, Stacked generative adversarial networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 5077–5086.
[31] 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.
[32] 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.
[33] 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.
[34] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, L. Van Gool, Pose guided person image generation, in: Advances in Neural Information Processing Systems, 2017, pp. 406–416.
[35] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, M. Fritz, Disentangled person image generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 99–108.
[36] P. Esser, E. Sutter, B. Ommer, A variational U-Net for conditional appearance and shape generation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 8857–8866.
[37] D.P. Kingma, M. Welling, Auto-encoding variational Bayes, arXiv preprint arXiv:1312.6114 (2013).
[38] H. Tang, D. Xu, G. Liu, W. Wang, N. Sebe, Y. Yan, Cycle in cycle generative adversarial networks for keypoint-guided image generation, in: Proceedings of the 27th ACM International Conference on Multimedia, 2019, pp. 2052–2060.
[39] Y. Ren, X. Yu, J. Chen, T.H. Li, G. Li, Deep image spatial transformation for person image generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 7690–7699.
[40] 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.
[41] S. Huang, H. Xiong, Z.-Q. Cheng, Q. Wang, X. Zhou, B. Wen, J. Huan, D. Dou, Generating person images with appearance-aware pose stylizer, arXiv preprint arXiv:2007.09077 (2020).
[42] S. Lathuilière, E. Sangineto, A. Siarohin, N. Sebe, Attention-based fusion for multi-source human image generation, in: The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 439–448.
[43] L. Yang, P. Wang, X. Zhang, S. Wang, Z. Gao, P. Ren, X. Xie, S. Ma, W. Gao, Region-adaptive texture enhancement for detailed person image synthesis, in: 2020 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2020, pp. 1–6.
[44] C. Gao, S. Liu, R. He, S. Yan, B. Li, Recapture as you want, arXiv preprint arXiv:2006.01435 (2020).
[45] R. Alp Güler, N. Neverova, I. Kokkinos, Densepose: dense human pose estimation in the wild, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7297–7306.
[46] Y. Li, C. Huang, C.C. Loy, Dense intrinsic appearance flow for human pose transfer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3693–3702.
[47] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, J. Yin, Soft-gated warping-GAN for pose-guided person image synthesis, in: Advances in Neural Information Processing Systems, 2018, pp. 474–484.
[48] J. Cheng, L. Dong, M. Lapata, Long short-term memory-networks for machine reading, arXiv preprint arXiv:1601.06733 (2016).
[49] A.P. Parikh, O. Täckström, D. Das, J. Uszkoreit, A decomposable attention model for natural language inference, arXiv preprint arXiv:1606.01933 (2016).
[50] N. Parmar, A. Vaswani, J. Uszkoreit, Ł. Kaiser, N. Shazeer, A. Ku, D. Tran, Image transformer, arXiv preprint arXiv:1802.05751 (2018).
[51] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 7794–7803.
[52] Y. Liu, X. Zhang, Q. Zhang, C. Li, F. Huang, X. Tang, Z. Li, Dual self-attention with co-attention networks for visual question answering, Pattern Recognit. 117 (2021) 107956.
[53] W. Wei, Z. Wang, X. Mao, G. Zhou, P. Zhou, S. Jiang, Position-aware self-attention based neural sequence labeling, Pattern Recognit. 110 (2021) 107636.
[54] H. Zhang, I. Goodfellow, D. Metaxas, A. Odena, Self-attention generative adversarial networks, in: International Conference on Machine Learning, 2019, pp. 7354–7363.
[55] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556 (2014).
[56] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al., Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[57] B. AlBahar, J.-B. Huang, Guided image-to-image translation with bi-directional feature transformation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9016–9025.
[58] 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.
[59] Z. Wang, A.C. Bovik, H.R. Sheikh, E.P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Trans. Image Process. 13 (4) (2004) 600–612.
[60] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, Improved techniques for training GANs, in: Advances in Neural Information Processing Systems, 2016, pp. 2234–2242.
[61] 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 Conference on Computer Vision and Pattern Recognition, 2018, pp. 586–595.
[62] 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, arXiv preprint arXiv:1706.08500 (2017).
[63] D.P. Kingma, J. Ba, Adam: a method for stochastic optimization, arXiv preprint arXiv:1412.6980 (2014).
[64] D. Ulyanov, A. Vedaldi, V. Lempitsky, Instance normalization: the missing ingredient for fast stylization, arXiv preprint arXiv:1607.08022 (2016).

[65] G.E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, R.R. Salakhutdinov, Improving neural networks by preventing co-adaptation of feature detectors, arXiv preprint arXiv:1207.0580 (2012).
[66] A.L. Maas, A.Y. Hannun, A.Y. Ng, Rectifier nonlinearities improve neural network acoustic models, in: Proc. ICML, vol. 30, 2013, p. 3.

Chengkang Shen is currently a Ph.D. student in the School of Electronic Science and Engineering at Nanjing University. He obtained his M.S degree from the University of Illinois at Chicago, and B.E degree from Nanjing University of Aeronautics and Astronautics. His research interests include computer vision and deep learning.

Peiyan Wang is currently a graduate student at Purdue Univerisity in the department of Electrical and Computer Engineering. During undergraduate study, she studied ECE at the University of Illinois at Chicago as an exchange student, and her research mainly focuses on Computer Vision. She got B.E degree from Yanshan University.

Wei Tang received his Ph.D. degree in Electrical Engineering from Northwestern University, Evanston, Illinois, USA in 2019. He received the B.E. and M.E. degrees from Beihang University, Beijing, China, in 2012 and 2015 respectively. He is currently an Assistant Professor in the Department of Computer Science at the University of Illinois at Chicago. His research interests include computer vision, pattern recognition and machine learning.