ShaTure: Shape and Texture Deformation for Human Pose and Attribute Transfer

Wing-Yin Yu, *Graduate Student Member, IEEE*, Lai-Man Po, *Senior Member, IEEE*, Jingjing Xiong, *Graduate Student Member, IEEE*, Yuzhi Zhao, *Graduate Student Member, IEEE*, and Pengfei Xian, *Graduate Student Member, IEEE*

***Abstract*— In this paper, we present a novel end-to-end pose transfer framework to transform a source person image to an arbitrary pose with controllable attributes. Due to the spatial misalignment caused by occlusions and multi-viewpoints, maintaining high-quality shape and texture appearance is still a challenging problem for pose-guided person image synthesis. Without considering the deformation of shape and texture, existing solutions on controllable pose transfer still cannot generate high-fidelity texture for the target image. To solve this problem, we design a new image reconstruction decoder – ShaTure which formulates shape and texture in a braiding manner. It can interchange discriminative features in both feature-level space and pixel-level space so that the shape and texture can be mutually fine-tuned. In addition, we develop a new bottleneck module – Adaptive Style Selector (AdaSS) Module which can enhance the multi-scale feature extraction capability by self-recalibration of the feature map through channel-wise attention. Both quantitative and qualitative results show that the proposed framework has superiority compared with the state-of-the-art human pose and attribute transfer methods. Detailed ablation studies report the effectiveness of each contribution, which proves the robustness and efficacy of the proposed framework.**

***Index Terms*— Human pose transfer, attribute transfer, ShaTure block, adaptive style selector module.**

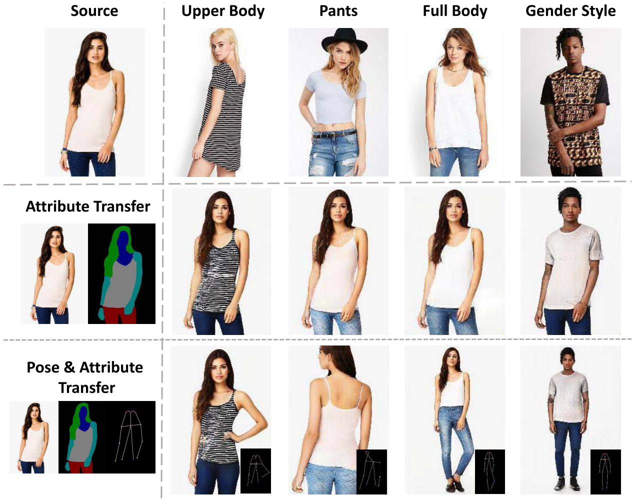

Fig. 1. Visualization of the proposed method to perform attribute transfer using a source pose or an arbitrary pose. We present 4 types of attributes including upper body, pants, full body and gender style. Please zoom in for more details.

I. Introduction

HUMAN pose and attribute transfer is a conditional image generation task that transforms a source image to a target image based on a pair of given pose heatmaps with selected attributes. Due to the spatial misalignment caused by occlusions and multi-viewpoints, maintaining original texture appearance is still a challenging bottleneck for pose-guided person image synthesis. Different from traditional pose transfer, controllability of attributes such as changing specific garment increases the difficulty for this task. From Figure 1, we visualize some examples of simple attribute transfer and combination of pose and attribute transfer. Benefitting from its commercial values, it has tremendous potential computer vision applications such as virtual try-on [1], data augmentation for person re-identification [2] or video generation [3].

There are basically two types of approaches applied in the task of human pose transfer, including prior-based and attention-based methods. For the prior-based approach [4]–[8], it consists of two separate networks for different purposes including a prior generator and an image generator. There are various kinds of prior knowledge such as coarse person image [4], [5], semantic mask [6] or optical flow [7], [8]. Notwithstanding these types of prior generative methods can leverage extra constraints to guide the transformation, non-end-to-end designs increase the training complexity of the network development. Due to the nature of specific prior information, there is no proper aggregation of shape and texture characteristics. For example, the semantic masks cannot provide reconstructive guidance of texture due to sparse distribution; the prior-based methods have the problem of large variation between the source pose and target pose resulting in misaligned shape generation and texture distortion. Apart from the prior-based approaches, the attention-based methods [9], [10] show comparative performance in quantitative accuracy. They apply the attentional mechanism to project the

Manuscript received July 9, 2021; revised January 13, 2022 and February 24, 2022; accepted February 24, 2022. Date of publication March 11, 2022; date of current version March 21, 2022. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Weiyao Lin. *(Corresponding author: Wing-Yin Yu.)*

The authors are with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong (e-mail: wingyinyu8-c@my.cityu.edu.hk; eelmpo@cityu.edu.hk; jingxiong9-c@my.cityu.edu.hk; yzzhao2-c@my.cityu.edu.hk; xian.pf@my.cityu.edu.hk).

Digital Object Identifier 10.1109/TIP.2022.3157146

source pose on the source image. Then the activated features are transformed according to the target pose. However, the visual quality should be improved due to lack of supervision on the regions without activation of pose landmarks.

In addition to pose-guided person image generation, controllable pose transfer is an extended task which allows users to change the attributes of the source image. Some applications of attribute transfer were found in face generation [11]–[13] such as changing facial expressions or customizing the scale of a particular sense organ. For human pose and attribute transfer, it supports substitution of different components of person image including garments, pants, face, or gender while generating new posture based on target pose. ADGAN [14] was the first work to solve this problem by introducing a shared encoder to embed the features of each attribute by part. Zhang *et al.* [15] introduced a prior-based model to pre-generate a semantic mask while using spatial-aware normalization to transform the per-region attributes. Although these two great works can accomplish the task of attribute and pose transfer, the deformation between the shape and texture is still an issue. Apart from disentanglement of shape and texture, it is hard to generate new content for occluded regions. In terms of completeness of garment and characteristics of face, it cannot generate fine-grained person image. Although PISE [15] tried to deal with this problem by applying a per-region normalization, it failed to generate a context-aligned texture representation compared to the original source image. It is because the most informative features of the texture are distorted during the encoding process, especially in down-sampling operation. Moreover, the multi-scale manipulation on classifying the categories of human part increases the difficulty to transfer the style feature representations for both garments and person characteristics. As a result, it is important to capture multi-scale features to generate vivid patterns from the source image.

To solve these problems, we propose a novel end-to-end framework – ShaTure to cope with the problem of human pose and attribute transfer. To achieve controllable attributes on the generated image, we firstly generate segmented human parts through an external human parser to separate the human body attributes. Different from previous methods [14], [15], we design a Style Encoder to decompose the source attributes into the style embedding without source pose and regularizations. Since the existence of multi-scale objects in the source image confuses the style codes, we introduce an Adaptive Style Selector (AdaSS) Module to facilitate the multi-scale feature extraction, so that the refined style representations can be generated. To learn the spatial correspondence between the source pose and target pose, we employ a Correspondence Encoder to acquire the geometric transformation mapping. In the reconstruction process, we propose the ShaTure Block to deform the extracted shape and texture representations in a braiding manner in order to exchange discriminative features in both feature-level space and pixel-level space. The main benefit of generating images referenced from pixel-level source is that it can provide hints of highly similar textures and patterns. For the sake of synthesizing new content, we need information from feature-level space to mutually cooperate in the generation for revealing the intrinsic characteristics of original image.

The contributions of this work can be summarized in the following three perspectives:

1) Controllable Pose-Guided Person Image Generation: We propose a novel end-to-end framework to transform a source person image to an arbitrary pose with controllable attributes. Experimental results show that the images we generate can maintain high fidelity on shape and texture transformation.

2) ShaTure Block: We design a new image reconstruction block – ShaTure Block that can decouple shape and texture in a braiding manner. The shape-and-texture-oriented architecture can preserve more details of garments and person characteristics.

3) Adaptative Style Selector (AdaSS) Module: We also develop a new bottleneck module – AdaSS Module which can enhance the feature extraction capability on multi-scale objects by self-recalibrating the feature map through channel-wise attention.

II. Related Work

In this section, we briefly introduce some related works. We discuss the fields of image generation, human pose transfer and attention mechanism.

A. Image Generation

Since the rapid development of Generative Adversarial Networks (GANs) [16] and Variational Autoencoders (VAEs) [17], there are some breakthroughs for image generative tasks. The GANs consist of a generator and a discriminator, wherein the purpose of the generator is to synthesize realistic images such that the discriminator cannot distinguish between the synthetic result and the real target. By conditioning the input source, Mirza and Osindero [18] proposed the conditional generative adversarial networks (cGANs) to generate images with some specific constraints so that they could meet the desired purpose. By further increasing the applicability of conditional generation, Isola *et al.* [19] designed the Pix2Pix framework for image-to-image translation, in which both input and output are images with flexible domains rather than latent codes. To modulate the style features with customized scale, Karras *et al.* [20] proposed an adaptive instance normalization (AdaIN) to inject style features in form of scale and bias. Based on this idea, Park *et al.* [21] suggested a spatially adaptive denormalization (SPADE) method to extend the modulation factors from vectors to tensors and use a semantic map to provide spatially contextual information. Although these kinds of non-linear normalization approaches can enhance the visual quality of the generated images, they cannot deal with the spatial misalignment problem and the sparse correspondence between the source and target landmarks for human pose transfer. Based on the success of SPADE [21], in this paper, we propose to generalize the input source from non-aligned style features to spatially aligned pixel-wise attributes so that it can optimize the fusion of shape and texture with learnable parameters.

B. Human Pose Transfer

Research on human pose transfer has recently drawn a lot of attention in the field of computer vision. In general, there are two types of architectural designs used in this challenging task, namely prior-based and attention-based approaches.

For the prior-based methods, there are two separate networks including a prior generator and an image generator with different training schemes. Ma *et al.* [4] initially proposed a pose-guided person generation network PG^2 to start this task. They proposed a prior-based model to transfer the pose by combining a coarse transferred result and a difference map but such fusion could not produce a plausible image due to unreliable texture refinement of the difference map. Yang *et al.* [5] and [22] applied a similar coarse-to-fine idea. They tried to deal with the unreliable texture refinement problem by introducing a residual texture map and a deeper feature extraction encoder. However, the result still suffers from facial representation distortion and the enhancement is limited. Instead of using residual map, Dong *et al.* [6] suggested to find a style transformation by computing the correlation between the pre-generated parsing map and the target one. Although it can find the semantic correspondence between the source image and target image, this relationship is not well defined because of inaccurate parsing generation. To address this problem, Li *et al.* [7] proposed to transfer appearance features from 3D space to 2D space by using optical flow. Based on similar operation of prior flow generation, Ren *et al.* [8] further enhanced the spatial alignment for pose transfer by leveraging a differentiable global-flow local-attention block to reassemble the inputs at the feature level. Although optical-flow transformation can perverse some details of the source image, there is no correspondence between the shape and texture for the rendering process. It leads to coarse generation quality when there is a large pose variance due to unreliable flow-warping operation.

For the attention-based methods, it normally follows Pix2Pix [19] network as the baseline to further enhance the residual block with attentional mechanism. Zhu *et al.* [9] proposed the PATN to progressively inject the pose-attentional activation to the source features. Based on similar attentional operation, Huang *et al.* [10] extended the network to an encoder-decoder architecture while proposing an adaptive normalization to normalize the appearance representation with the target pose. Tang *et al.* [23] also applied attentional operation to parallelly update the shape and appearance embeddings of source image. Notwithstanding these kinds of attentional block can obtain a higher score on quantitative evaluation, the qualitative visualization shows blurry effect on the generated images due to insufficient shape and texture guidance.

For controllable pose transfer, it is an extended task from traditional pose transfer which allows users to dynamically change the attributes of the source image. Men *et al.* [14] proposed a shared encoder to embed the features of each attribute one by one. Although it can save the computational cost during the encoding process, the global style of the generated image may not be aligned with the source image due to the lack of global context knowledge. Zhang *et al.* [15] introduced a prior-based model to pre-generate a semantic mask while normalizing per-region attributes with spatial-aware generation. Without considering the deformation of shape and texture, these methods fail to generate high quality of texture from the source image. To solve this problem, we propose an end-to-end model with a series of progressive ShaTure Blocks to deform shape and texture simultaneously so that highly plausible can be synthesized.

TABLE I

BRIEF INTRODUCTION OF SYMBOLS

Symbols	Description
I_s, I_t	Source / target person image
P_s, P_t	Source / target pose
M_s, M_t	Source / target semantic map
I_g^{l-1}, I_g^l	Previous / current generated result on layer l
A_s, A_t	Source / target decomposed attributes
A, A^k	Sub-set of k attributes selected from combined attributes A
X_s^l, X_t^l	Input source at layer l of Shape / Texture Module
S_*, T_*	Output of Shape / Texture Module at $*$ phase
I_s^l	Style features at decoder layer l
I_{ms}^l	Multi-scale style features at decoder layer l
T	Spatial transformation features
$\hat{x}_R$	Multi-scale features with receptive field of view R

C. Attention and Gate Mechanism

Different from attention-guided methods mentioned in Section 2B that use spatial attention, the gated attention methods can focus on the globally channel-wise relationship within the target feature expressions instead of being activated by the whole driving source. It is a means of highlighting the most informative components of a signal by reallocating the computational resources on the features [24]–[28]. To emphasize cross channel dependency, Hu *et al.* [28] proposed the squeeze-and-excitation operation to self-recalibrate the feature maps via gated mechanism. Apart from channel-wise dependency, Park *et al.* [29] and Woo *et al.* [30] added spatial attention during the excitation process so that location information can be involved. Based on similar idea, Li *et al.* [31] extended it to a weighted gate by splitting the features maps into two separated representations with different sizes of receptive field of view. This dynamic selection mechanism allows each neuron to adaptively optimize the receptive field size based on multiple scales of input features. We observe that it is suitable for human pose transfer task because of multi-scale distribution caused by multiple viewpoints and occlusions. To alleviate the spatial misalignment effect, we exploit the idea of dynamic kernel selection for each generative block to extract multi-scale features. It can adaptively optimize the kernel size of receptive field to manipulate the input signals.

III. PROPOSED METHOD

A. Problem Definition

Given a source person image I_s and a pair of spatially misaligned pose heatmaps including source pose P_s and target pose P_t, the task of pose-guided human synthesis is to transfer

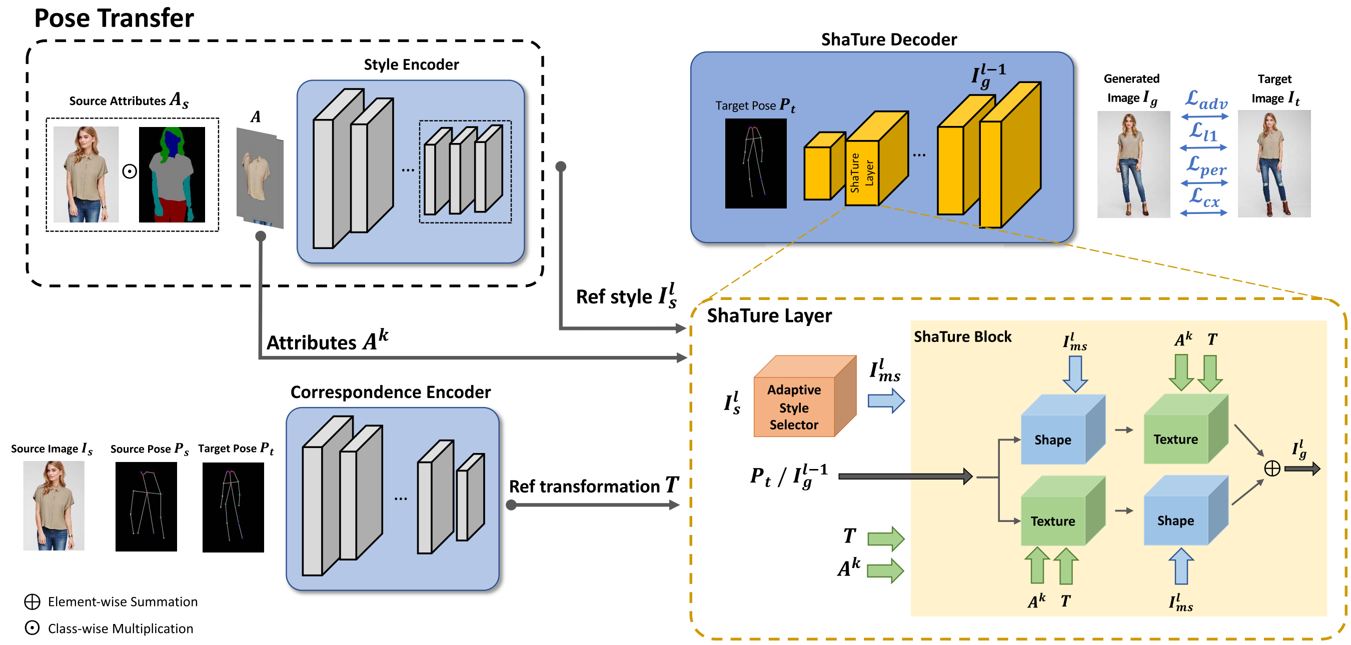

Fig. 2. Overview of the network architecture for our framework. It is built based on variational autoencoders (VAEs) architecture including two feature encoders and one generative decoder, namely Style Encoder, Correspondence Encoder and ShaTure Decoder. During training stage, we utilize the branch of pose transfer to learn the feature aggregation. In each of the Shature Layer in ShaTure Decoder, there is an Adaptive Style Selector (AdaSS) Module highlighted with orange color and a ShaTure Block highlighted with yellow color. The AdaSS Module is responsible for extracting multi-scale features I_{ms}^l from the Referenced style I_s^l. By taking the output of previous layer I_g^{l-1}, multi-scale features I_{ms}^l, selected attributes A^k and referenced transformation features T as input, the ShaTure Block can generate plausible images with a braiding approach.

as many details as I_s to a new person image I_g upon the shape of P_t. In addition to pose transfer, equipping attribute controllability requires semantic information to segment corresponding human parts. We define the semantic map as $M_s \in \mathbb{R}^{W\times H\times C}$ where C is the total number of classes of interest, W is the width and H is the height. It is spatially aligned with I_s and P_s. In this paper, we propose an end-to-end generative framework $\mathcal{F}(\cdot)$ to deal with both pose transfer and attribute editing simultaneously, which can be expressed as

$$I_g = \mathcal{F}(I_s,\ M_s,\ P_s,\ P_t\ ). \tag{1}$$

B. Network Architecture

An overview of the network architecture is shown in Figure 2. Basically, the proposed framework is built based on variational autoencoders (VAEs) architecture including two feature encoders and one generative decoder, namely Style Encoder, Correspondence Encoder and ShaTure Decoder. The objective of the Style Encoder is to encode the decomposed attributes $A_s \in \mathbb{R}^{W\times H\times 3C}$ into a latent space, where $A_s = I_s \odot M_s$ and $\odot$ denotes class-wise multiplication. The bottlenecks of the Style Encoder are three consecutive residual maps which are of the same dimension but with different receptive fields of view. In order to learn the geometric correspondence of shape and texture between the source image and the target image, we set up a Correspondence Encoder to acquire the transformation mapping T by taking I_s, P_s, P_t as input. The reference transformation mapping T is shared to each progressively generative layer in the decoder. For the decoder part, we propose a new ShaTure decoder block to jointly encapsulate the extracted shape and texture representations from the encoders in a braiding manner in order to transfer discriminative features from the original image. The detailed elaboration of the ShaTure Block will be presented in Section 3C. At the end of the Style Encoder, there is an Adaptive Style Selector (AdaSS) Module responsible for further enhancing feature extraction capability by applying squeezed self-attention operation. More information of this module will be introduced in Section 3D. We also include the inference process to perform attribute transfer which will be discussed in Section 3E.

C. ShaTure Block

To effectively explore the relationship between shape and texture, we design a braiding module – ShaTure Block to exchange discriminative features in both feature-level space and pixel-level space. The main idea of the ShaTure Block is to reconstruct a plausible person image by using the shape features from the Style Encoder while preserving pixel-level patterns from source image with corresponding spatial transformation. Although style information is embedded in the encoded features of the encoder, the fine-grained texture information is degraded due to the consecutively lossy downsampling operation. To recover as many perceptual details as the source image and be able to generate spatially aligned objects based on a given pose-guided heatmap, we propose to synthesize shape and texture in a braiding manner. As shown in Figure 3, the ShaTure Block consists of two sub-nets namely Shape Module and Texture Module. Both sub-nets take the

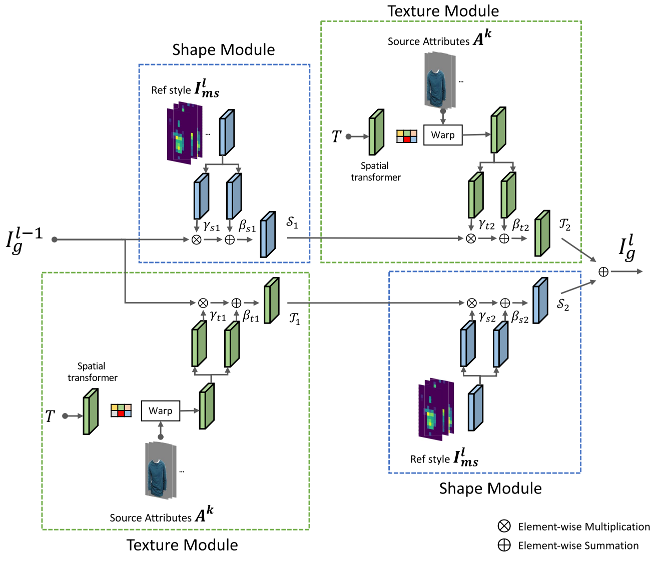

Fig. 3. Graphical structure of the ShaTure Block. It consists of two kinds of sub-nets namely Shape Module and Texture Module. Both sub-nets take the output of previous module as the major modulating signals. The Referenced multi-scale style features are generated from the Adaptive Style Selector (AdaSS) Module. There is a small spatial transformer network to produce the protocol for affine transformation with the source attributes. Finally, the Shape Module and Texture Module are connected in a parallelly braiding manner.

output of previous layer as the major modulating signals. However, they play different roles regarding reconstruction of shape and texture respectively during the generation process. There are two symmetric branches inside a ShaTure Block. For one of the branches, the Texture Module $\mathcal{T}(\cdot)$ takes the output of Shape Module $\mathcal{S}(\cdot)$ as input source for further enhancement, and vice versa. Mathematically, the relationship of a ShaTure Block can be expressed as

$$I_g^l = \mathcal{T}_2\left(\mathcal{S}_1\left(I_g^{l-1}, I_{ms}^l\right), A^k, T\right) + \mathcal{S}_2\left(\mathcal{T}_1\left(I_g^{l-1}, A^k, T\right), I_{ms}^l\right). \quad (2)$$

where I_g^{l-1} is the generated result of previous ShaTure Block, I_{ms}^l are the multi-scale reference style features computed from AdaSS Module which will be intensively discussed in Section 3D, A^k are the decomposed attributes selected from the concatenated attributes A with a sub-set of integers $K \subseteq C$ (classes of semantic mask) which are the components needed to be transferred, T is the deep spatial transformation features computed from Correspondence Encoder.

For the Shape Module, it aims to recover the whole shape and style from source image based on the target pose P_t or previous layer I_g^{l-1}. We exploit a spatially adaptive image translation approach based on the SPADE [21] denormalization unit. We firstly apply a shared convolutional layer to extract intermediate features corresponding to specific resolution. It is then manipulated as a scaling and bias factor respectively. Instead of using batch normalization [32], we apply instance normalization [33] to calculate the mean and variance across the spatial location. It can provide better visual and appearance in-variance compared to batch normalization. The activated features of Shape Module at space ($n \in N, c \in C^i$, $h \in H^i, w \in W^i$) can be expressed as:

$$\mathcal{S}_* = \delta\left(\gamma\left(X_s^l\right) \cdot \frac{f_{n,c,h,w}^i - u_{nc}^i}{\sigma_{nc}^i} + \beta\left(X_s^l\right)\right). \quad (3)$$

where δ is the LeakyReLU function [34], u_{nc}^i and σ_{nc}^i are the mean and stand deviation of $f_{n,c,h,w}^i$ which is a function of either a previous layer I_g^{l-1} or a demodulated transformation result $\mathcal{T}$. The learnable parameters scale $\gamma\left(X_s^l\right)$ and bias $\beta\left(X_s^l\right)$ are convoluted by input source X_s^l which is the target pose P_t at layer l for the Shape Module. As shown in Figure 3, there are three independent weights involved in the computation. We firstly utilize a shared convolution to generalize the input source to some features with a constant number of channels. Based on the shared features followed by a LeakyReLU function [34], the scale γ_{s*} and bias β_{s*} features are generated by another two weights separately.

For the Texture Module, we further investigate the enhancement of discriminative pattern synthesis by making use of the deep transformation features T. It can maintain part of texture information of the source image by directly transforming features in pixel-level space. However, the ability of generating new content for segmented attributes being adopted to a new pose is limited. To solve this problem, we generalize the input source of Shape Module mentioned above from style features to warped target attributes A^k. The spatially adaptive modulation operation can generate corresponding shape while preserving the context of the warped attributes. In addition to being effective for pose transfer, it contributes a more robust attribute transferring ability. We use X_t^l to denote the input source of Texture Module at layer l which can be expressed as:

$$X_t^l = A^k \mathcal{H}_t^l(T). \quad (4)$$

where $\mathcal{H}_t^l(\cdot)$ is a simplified version of spatial transformer network [35] at layer l to be served as an affine parametric transformation. It localizes the geometric feature representations by using two convolutional neural layers. We also use a fully connected network as a grid generator to generate a transformation grid indicating the affine flow to warp a source pixel to a target location. Since directly transforming the source image in pixel-level space cannot generate new content information, we further render the warped attributes by applying Equation 3 on X_t^l to optimize the completeness. The generation of learnable parameters scale γ_{t*} and β_{t*} for Texture Module is similar to the computation of Shape Module by changing X_{ms}^l to X_t^l as the input source.

D. Adaptive Style Selector Module

To synthesize high-quality images, we generate the image from the source pose to a transformed image with a progressive decoder. During the process of reconstruction, we exploit the reference style encoded from the source image. Similar to the architecture of U-Net [36], one of the solutions is to directly utilize the style features with different resolutions from the layers of Style Encoder. However, based on our observation, the style features with different resolutions in the

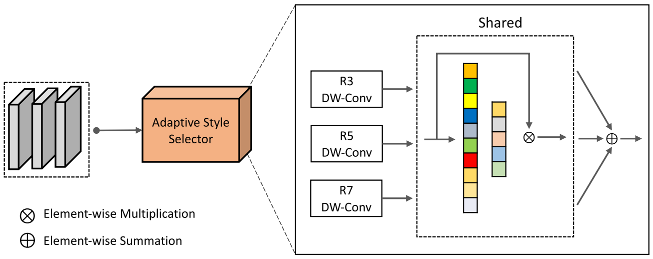

Fig. 4. Graphical structure of the adaptive style selector. There are 3 depth-wise convolutional blocks with different sizes of receptive field of view. The shared gate consists of a global average pooling layer and a small fully-connected network. The joined style codes are highlighted by attentional self-recalibration.

Style Encoder contain sparse semantic information leading to confused synthesis in the image reconstruction layers, especially some shallow encoding layers which make it hard to generalize the representative style codes. We believe that the style embedding itself from the Style Encoder is variant to the up-sampling level of reconstruction layers. It should be projected to a common space where the style features are consistent. Although a vanilla encoder can embed a source image into some style codes at the end of the most down-sampled layers, it cannot provide informatively multi-scale features to be adaptively customized for each generative layer. To solve this problem, inspired by SENet [28], we propose an Adaptive Style Selector (AdaSS) Module to strengthen the ability of dynamic reconstruction for each decoding block. The idea behind the AdaSS Module is to adaptively extract informative features from the source styles by emphasizing cross-channel correspondences between the bottlenecks of the style encoder. With such attentional aggregation, the network can allocate the attention toward the most suitable style features to corresponding reconstruction layers.

To enforce contextual knowledge, we design a deep residual bottleneck at the end of the encoder. Referring to Figure 4, it comprises three blocks of feature representations with the same resolution but different receptive fields. The objective of enlarging the field of view is to enable the network to encapsulate more contextual information with different scales of kernel. Moreover, we also adopt the depth-wise/group separable convolution [37], [38] to reduce the number of parameters by partitioning convolution into a depth-wise kernel and a point-wise kernel. We sequentially apply point-wise convolutions with an 1×1 kernel on the last style feature representation for t repetitions. To expand the receptive field, we stack the depth-wise convolutions with some 3×3 kernels along with group g for each point-wise convolution. It can get a receptive field of size $(2t + 1) \times (2t + 1)$. We take the reference style I_s^l extracted from the Style Encoder as input. For simplicity, we denote the style features I_s^l as $x \in \mathbb{R}^{w \times h \times c}$, then the output of the separable convolution can be represented as:

$$\hat{x}_R = V_R\left(\sum_{t=1}^{T} U^t\left(x\right)\right). \tag{5}$$

where $V_R\left(\cdot\right)$ indicates depth-wise convolution, $U^t\left(\cdot\right)$ is the t^{th} stacked point-wise convolutions, the R represents the kernel size which is linearly proportional to repetitions $R = 2T + 1$. We set the maximum of receptive field of view to 7×7, i.e. $T = [1, 2, 3]$.

Once we have the multi-scale representations, the next step is to dynamically fuse them according to the level of reconstruction layers. Inspired by the squeeze-and-excitation [28] operation, we introduce a shared attentional aggregation to generate channel-wise statistics. To highlight the channel-wise dependency, we obtain a master activation scalar using a global average pooling to downscale the dimension followed by two fully-connected (FC) layers. Finally, we aggregate the gated style features by a weighted summation. The formulation can be represented as:

$$F\left(\hat{x}_R\right) = \sigma\left(W_2 \delta\left(W_1 Avgpool\left(\hat{x}_R\right)\right)\right) \tag{6}$$

$$I_{ms}^l = \sum_{R=\{3,5,7\}} \hat{x}_R \otimes F\left(\hat{x}_R\right) \tag{7}$$

where σ denotes sigmoid function, δ refers to ReLU [39] function, W_* are the FC layers and the $\otimes$ operator indicates Hadamard product. Since the W_* are learnable functions, it can provide discriminative responses with respect to level of reconstruction layers. The shared property also serves as a common agent to intrinsically introduce dynamics on multiple channels with the same judging criteria.

E. Objective Function

To synthesize photo-realistic person images and maintain details of the source images, we mainly focus on pixel-level and feature-level loss functions. Following similar training strategy of the existing pose transfer frameworks [9], [10], [14], we formulate the objective function with four terms including an adversarial loss $\mathcal{L}_{adv}$, a L1 loss $\mathcal{L}_1$, a perceptual loss $\mathcal{L}_{per}$ and a contextual loss $\mathcal{L}_{cx}$ as follows:

$$\mathcal{L}_{full} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_1\mathcal{L}_1 + \lambda_{per}\mathcal{L}_{per} + \lambda_{cx}\mathcal{L}_{cx} \tag{8}$$

where λ_{adv}, λ_1, λ_{per} and λ_{cx} are the corresponding hyperparameters to optimize the performance.

1) Adversarial Loss: We utilize the adversarial loss $\mathcal{L}_{adv}$ to maintain consistency of style and texture by leveraging two independent discriminators, D_s and D_c. The D_* consists of two down-sampling convolutional layers followed by three convolutional blocks for enhancement of discriminative capability. The joint adversarial loss terms are formulated as:

$$\begin{aligned}\mathcal{L}_{adv} = \; & \mathbb{E}_{I_s,I_t,I_g}\left[\log\left(D_s\left(I_s, I_t\right)\right) + \log\left(1 - D_s\left(I_s, I_g\right)\right)\right] \\ & + \mathbb{E}_{I_t,P_t,I_g}\left[\log\left(D_c\left(P_t, I_t\right)\right) + \log\left(1 - D_c\left(P_t, I_g\right)\right)\right],\end{aligned} \tag{9}$$

where D_s and D_c are the visual style discriminator and pose content discriminator; $(I_s, I_t) \in \mathbb{I}_{real}$, $P_t \in \mathbb{P}_{real}$, $I_g \in \mathbb{I}_{fake}$ indicate samples from the distribution of real person image, real pose heatmap and generated person image.

2) L1 Loss: To enforce pixel-level supervision, we employ pixel-wise L1 loss to minimize the least absolute deviations between the generated image and the ground truth.

3) Perceptual Loss: To minimize the distance in feature-level space, we apply the standard perceptual loss [40] in our network. It aims to enhance the visual quality by increasing the similarity of feature matching in a large-scale classification network. Precisely, it computes the pixel-wise L1 difference of a selected layer $\theta_\ell(\cdot)$ from a VGG-19 [41] model pre-trained in ImageNet [42]. It is defined as:

$$\mathcal{L}_{per} = \frac{1}{C_\ell H_\ell W_\ell} \sum_{c,h,w} \|\theta_\ell(I_g)|_{c,h,w} - \theta_\ell(I_t)|_{c,h,w}\|_1. \quad (10)$$

where C_ℓ is the number of channel, H_ℓ and W_ℓ are the height and width of the feature maps in particular layer ℓ respectively.

4) Contextual Loss: To maximize the similarity between two non-aligned images with similar context space, we exploit the contextual loss [43] to allow spatial alignment according to contextual correspondence during the deformation process. It makes use of the normalized Cosine distance between two feature maps to measure the similarity of two non-aligned features. It is formulated as:

$$\mathcal{L}_{cx} = -\frac{1}{C_\ell H_\ell W_\ell} \sum_{c,h,w} log\left[CX\left(\delta_\ell\left(I_g\right)|_{c,h,w}, \delta_\ell\left(I_t\right)|_{c,h,w}\right)\right], \quad (11)$$

where $\ell = relu\{3_2, 4_2\}$ layers from a pre-trained VGG-19 [41] model $\theta(\cdot)$, the $CX(\cdot)$ function is the similarity measurement defined in [43].

IV. Experiments and Results

In this section, we describe the implementation details of the proposed framework. Firstly, we introduce the dataset used in all experiments. Secondly, we define some evaluation metrics that can quantify the images generated from our framework. Moreover, we compare our proposed method with other state-of-the-art methods to verify the superiority quantitatively and qualitatively. We also conduct comprehensive ablation studies on each proposed component to show the efficacy of our contributions. Finally, we provide some visualizations on the feature maps to demonstrate the roles of Shape Module and Texture Module.

A. Datasets and Metrics

1) Dataset: In order to demonstrate the effectiveness of our proposed framework, we conducted experiments of both pose transfer and attribute editing on the commonly used benchmarks – *In-shop Clothes Retrieval Benchmark DeepFashion* [44]. It is a large-scale dataset that can provide high-resolution person images for these two tasks. It consists of 52,712 in-shop clothing items with a wide diversity of garments, poses, viewpoints, and occlusion scenarios. In order to filter out the noisy samples, we removed the samples of which pose heatmaps were unable to be detected by human pose estimator (HPE) [45]. Finally, we sampled a total of 101,966 training pairs and 8,570 testing pairs. The person identities are not overlapped for training and testing pairs to ensure the generalization ability. All the images were resized to 256×176 dimension during both training and testing phases. For the task requiring controllable attributes, the segmentation masks were pre-generated by the Look Into Person (LIP) [46] human parsing algorithm in order to facilitate the training process. We projected the masks into 7 categories of human parts including head, upper clothes, pants, shorts, arms, legs and 1 class for background.

2) Metrics: For general image generation tasks, Inception Score (IS) [47] and Structural Similarity (SSIM) [48] are two widely used evaluation metrics to quantify the perceptual performance and image quality. The IS is used to measure the global shape consistency by part of an image classifier. To quantify the structural similarity, the SSIM index is used to achieve this goal by applying co-variance and means. We also employ two supervised perceptual metrics including Fréchet Inception Distance (FID) [49] and Learned Perceptual Image Patch Similarity (LPIPS) [50] to consolidate the visual quality assessment in terms of perceptual distance between the generated images and real images. The FID is used to measure the reconstruction error between the generated images and the source images by computing the perceptual distances. Similar to the FID, the LPIPS is targeted at evaluating the Wasserstein-2 distance between the distributions of the generated samples and real samples.

B. Experiment Setting

We implement our proposed solution with the public framework PyTorch. We use Adam optimizer [51] with momentum $\beta_1 = 0.5$ and $\beta_2 = 0.999$ to train our model for 400 epochs in total. The learning rate is initially set to $1e - 4$ which is linearly decayed to 0 after 200 epochs. The negative slope of LeakyReLU [34] is set to 0.2. The weighting hyper-parameters λ_{adv}, λ_1, λ_{per} and λ_{cx} are set to 5, 1, 1 and 0.1. All models are trained and tested on a NVIDIA GeForce RTX 2080 Ti GPU with 11GB memory. The batch size is set to 4.

C. Experiment Result on Pose Transfer

To illustrate the effectiveness of the proposed network, we compare it with several state-of-the-art methods on the task of pose transfer. These methods include PG^2 [4], Def-GAN [52], RATE-Net [5], PATN [9], APS [10], XingGAN [23], ADGAN [14], GFLA [8] and PISE [15]. Without attribute controllability, this task requires the ability to transfer the style and characteristics of the original image from a source pose to a target pose. We present both quantitative and qualitative results for this task. The generated images of other methods are inferenced by the pre-trained models from their public repositories.

1) Quantitative Result: As shown in Table II, our method outperforms the current state-of-the-art methods with promising improvement for all evaluation metrics in the validation set of DeepFashion [44]. It gets the best scores in IS, SSIM, FID and LPIPS.

More specifically, our method has a better pose transferring ability compared to attention-based methods including PATN [9], APS [10] and XingGAN [23]. Noted that although some attention-based methods [23] can get a good score in IS, the performance of other metrics and visual quality need to be

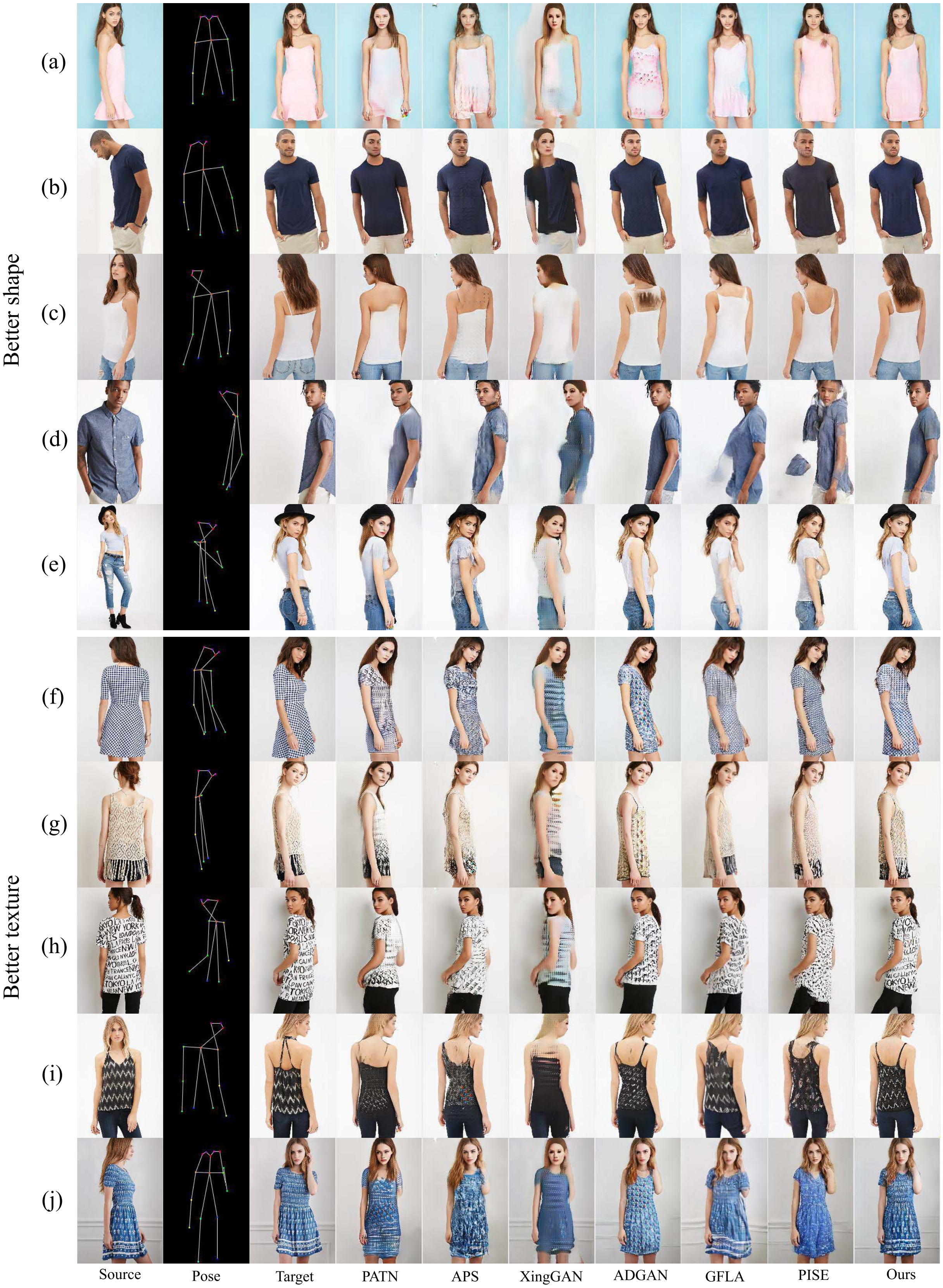

Fig. 5. Qualitative comparisons on the performance of pose transfer with some state-of-the-art methods on the DeepFashion benchmark. From left to right are the results of PATN [9], APS [10], XingGAN [23], ADGAN [14], GFLA [8], PISE [15] and our method. We compare it with respect to the quality of shape (a-e) and texture (f-j). The shape is defined as the global view of the model in the generated images while the texture factor focuses on the local details of the clothing items. Please zoom in for more details.

TABLE II

COMPARISON OF POSE TRANSFER PERFORMANCE WITH SEVERAL STATE-OF-THE-ART METHODS ON TEST DATASET OF DEEPFASHION

Methods	IS↑	SSIM↑	FID↓	LPIPS↓
PG^2 [4]	3.202	0.773	-	-
Def-GAN [52]	3.362	0.760	18.457	0.233
RATE-Net [5]	3.125	0.774	14.611	0.218
PATN [9]	3.209	0.773	19.816	0.253
APS [10]	3.295	0.775	15.017	0.178
XingGAN [23]	3.475	0.726	37.816	0.224
ADGAN [14]	3.364	0.772	13.224	0.176
GFLA [8]	3.402	0.770	15.271	0.189
PISE [15]	3.404	0.767	11.802	0.163
Ours	**3.487**	**0.777**	**11.434**	**0.159**

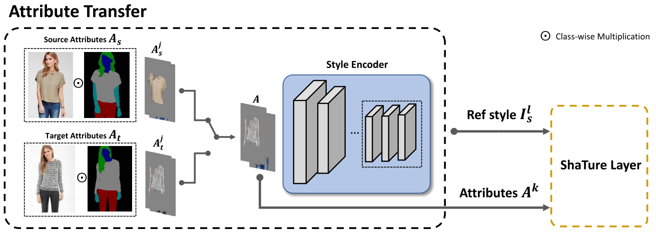

Fig. 6. Implementation of Style Encoder when performing attribute transfer. The branch of attribute transfer is activated in testing stage only. If j is specified, source attribute A_s^j is replaced by target attribute A_t^j to form the final encoding attribute A.

enhanced. Due to lack of supervision on sufficient numbers of pose landmarks, it cannot transfer specific details of the image although the shape consistency is able to be maintained. For our method, it achieves higher brands for all evaluation metrics especially IS and SSIM scores compared to them. It quantifies that our method can strike a balance between the shape and texture consistency under the same conditions.

In addition to attention-based approaches, our method also gets obvious gains compared to attribute-transfer methods such as ADGAN [14] and PISE [15]. We observe that these kinds of methods can get a better performance on some supervised perceptual metrics including FID and LPIPS. It is due to the reason that there is assistance of warped parsing attributes to provide pixel-level guidance. Such prior knowledge is helpful during the reconstruction process. Among three of the frameworks supporting controllable attributes, our method gets the best scores for all the metrics especially in terms of FID and LPIPS. It demonstrates that the proposed method can generate highly realistic images with less reconstruction error in feature space. With lower LPIPS score, it indicates that the feature quality of our synthesized images is highly competitive to the real data from a global distribution perspective. It represents that our method can perform very well in pose transfer with minimum distribution distance which is crucial in domain transferring tasks.

2) Qualitative Result: Apart from quantitative measurement, we also deliver qualitative comparison to those recent state-of-the-art methods including PATN [9], APS [10], XingGAN [23], ADGAN [14], GFLA [8] and PISE [15]. From Figure 5, our method can produce images with more photo-realistic quality than other approaches. We focus on two major factors that can affect the synthesis quality, namely shape and texture. For a better visual appearance, the shape is defined as the global view of the model in the generated images while the texture factor focuses on the local details of the clothing items.

As shown in Figure 5(a-e), we demonstrate that our method can generate some images with better shapes for models. For Figure 5(a), the dress on the model generated from our method contains the minimum artifacts. For Figure 5(b), we can transfer more characteristics on the head areas including facial attributes or hair style than other solutions. We believe that such variances are benefited from the ability of multi-scale encapsulation of our Adaptive Style Selector (AdaSS) Module. To synthesize the occluded regions, our model can predict a more reasonable hair shape based on little hints on the shoulder as shown in Figure 5(c). Compared to two methods with controllable attributes such as ADGAN [14] and PISE [15], our method can also produce a better visual quality in the situation of a large pose variation. As shown in Figure 5(d-e), the images generated from PISE [15] fail to be reconstructed because of the sparse correspondence between the parsing masks, like the head and hat. Although ADGAN [14] can generate reasonable images, there are some artifacts on the regions with ambiguous objects like hands and back in Figure 5(d) and rare posture in Figure 5(e). On the other hand, our method can overcome these challenges. It shows that our method performs well in shape reconstruction.

To compare the performance in terms of texture synthesis, we also provide some samples in Figure 5(f-j) to illustrate the superiority of our method in this perspective. Under the same pose variation settings, we have the best texture reconstruction ability which can transfer as much detailed texture as the source image. As shown in Figure 5(f), the plaid patterns on the dress are successfully transferred, in contrast with the thick-line styles transferred by XingGAN [23]. Apart from singular pattern, there are some combinations of regular symbols. For example, the diamond shapes and ribbons on the upper cloth of the model in Figure 5(e) are well maintained in our generated images while others can keep the color style only. For some special symbols and unique patterns, it is hard to transfer to a new pose without distortion due to lossy interpolation during downsampling encoding. However, our method is still able to accomplish this goal with a good result. For example, in Figure 5(h-i), there are some tailor-made characters or irregular patterns on the shirts of the models. Previous approaches can only capture the basic shape and distribution of those details while our method can sustain a large proportional appearance of those texture information with minimum distortions. Although GFLA [8] can sometimes recover some patterns, the visual quality is downgraded due to the blurry effect caused by unreliable flow-warping. Compared to the methods with controllable attributes, ADGAN [14] and PISE [15], our method retains plausible texture transferred from the source image like the global style of the jumpsuit in Figure 5(j). We believe that it is the effort of our ShaTure Block that can synthesize the shape and texture in a

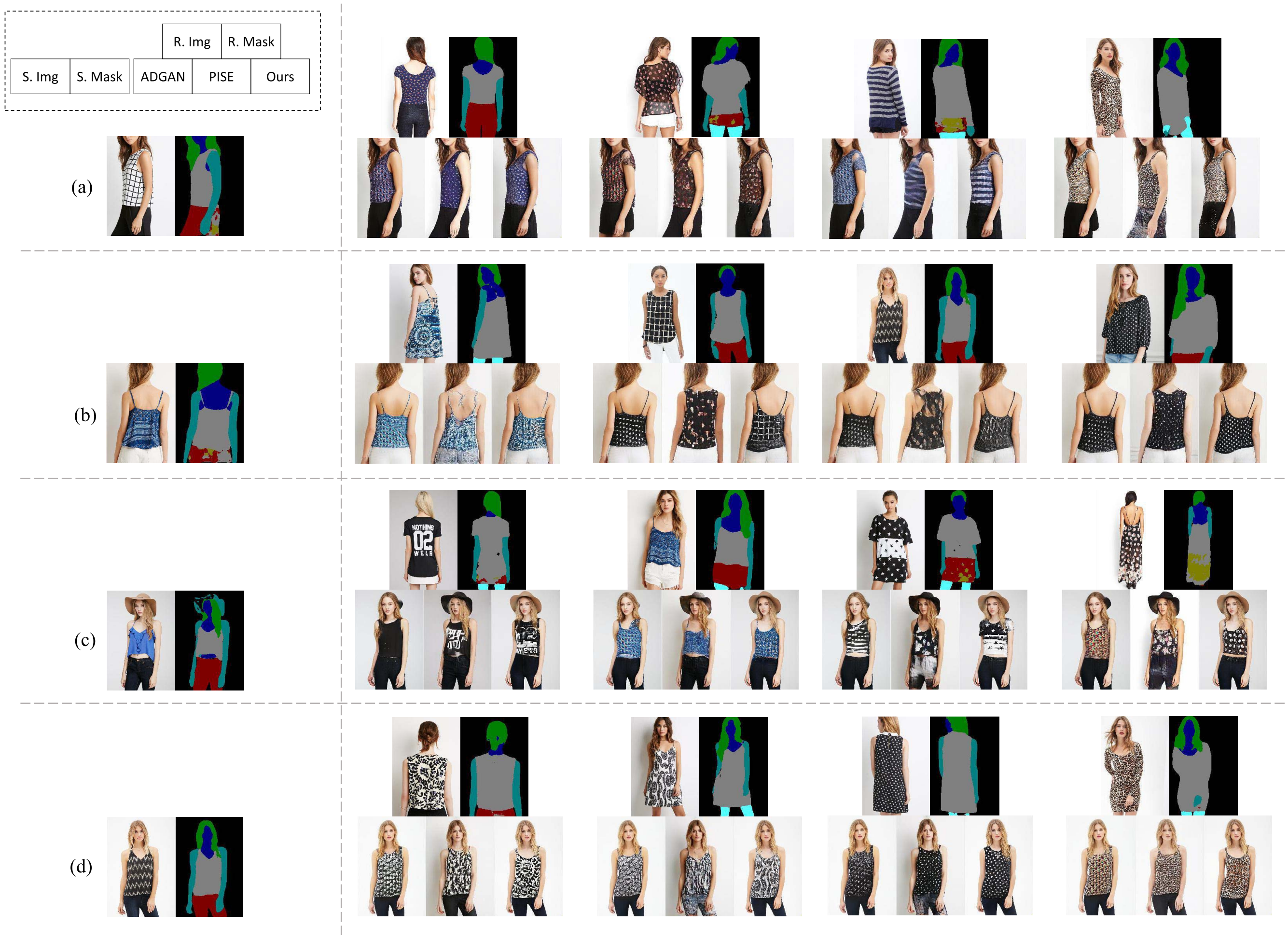

Fig. 7. Qualitative comparisons on the performance of attribute transfer with some state-of-the-art methods on the DeepFashion benchmark. The overall layout is indicated at the upper-left corner. We present the source and target attribute images with their parsing masks. From left to right are the results of ADGAN [14], PISE [15] and our method. In order to highlight the difference of the texture, we illustrate the result of upper-body clothes as the source attribute since there are more variations for this garment compared to others. We demonstrate different angles of view including side (a), back (b) and front (c-d) view. Please zoom in for more details.

braiding manner so that a photo-realistic image can be well reconstructed.

D. Experiment Result on Attribute Editing

Apart from pose transfer, our method can also support attribute transfer in which the style of the source image is able to be customized by some referenced images while maintaining the ability of pose transfer. Once the Style Encoder is fully trained, it is supposed to be able to project each attribute into the style code space. Therefore, we can perform attribute transfer during the testing stage by simply replacing the source attributes with the target attributes. We illustrate the implementation of Style Encoder for performing attribute transfer in Figure 6. The decomposition of source attributes with the segmentation mask is the key element during the encoding process. We use the same Style Encoder in both training and testing stages. Instead of encoding all attributes from source image A_s as the final attributes A, we replace some attributes from some referenced images with defined index j. If j is specified, the source attribute A_s^j is replaced by target attribute A_t^j to form the final encoding attribute A. Noted that the index j is different from the subset k where the index j is defined by the users during testing stage but the subset k attributes are hard-coded to be warped in ShaTure Module. Multiple attributes are also supported by specifying more than 1 index.

1) Comparison With the State-of-the-Arts: For the task of pose transfer supporting controllable attributes, we make an extensively comparative experiment compared to the attribute-transfer methods such as ADGAN [14] and PISE [15]. As mentioned in Section 3C, the A^k are selected according to the attribute that needs to be transferred. Based on the semantic segmentation result containing 8 classes in total, we focus on only 3 important attributes to be transferred including face, upper clothes and pants. Therefore, we set $K = \{1, 3, 5\}$ during the training stage. When performing attribute transfer in testing stage, users can swap the referenced components with the source components respectively. In this comparison, we show 4 source images associated with 4 driving reference images for each example in Figure 7. They are with different scales, poses, viewpoints and clothing styles. Based on our observation, there are a few visual variations of the

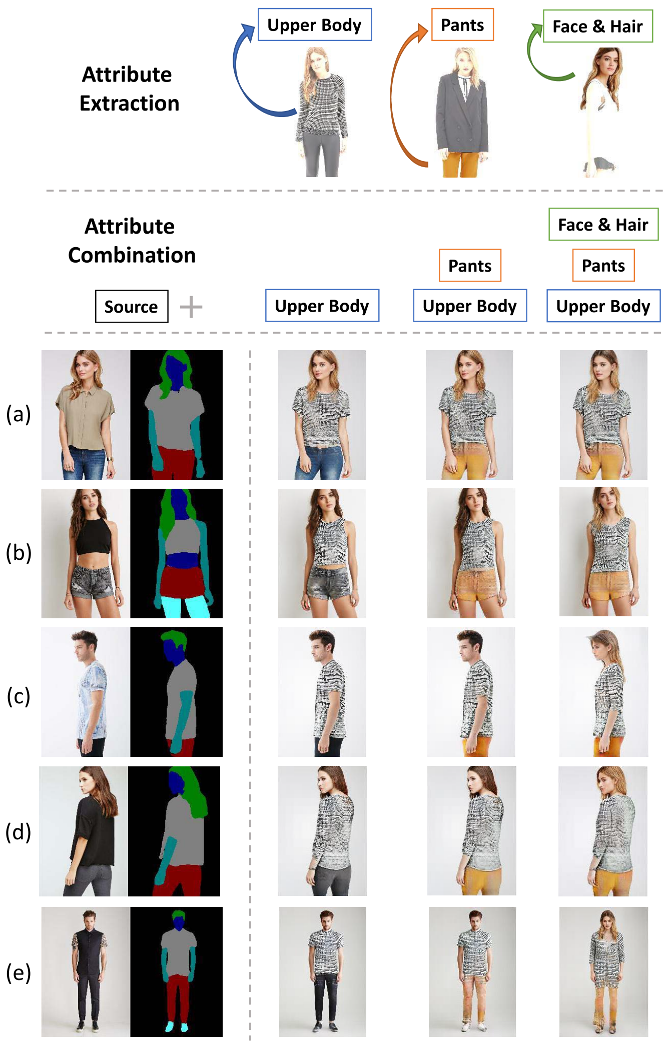

Fig. 8. Visualization of controllable pose-guided person image generation. To demonstrate the controllability of our method, we provide some results on multiple attributes transfer for different angles of view including front-upper-body (a-b), side-upper-body (c-d) and front-full-body (e) view. We take the source image as the attribute base. By extracting target attributes from other images, we can obtain the corresponding style codes. We demonstrate the attribute transferring result for upper body, pants and face&hair combination. There are multiple view angles for the source images in order to show the robustness.

lower-body clothing items such as pants and shorts in the dataset. We conduct the texture transferring experiments on upper-body clothing items with the original pose in order to better distinguish the visual difference. Noted that although it targets at evaluating the vividness of the texture in priority, the quality of shape should not be neglected.

Firstly, we compare the visual results when there is a rare posture in the image. As shown in Figure 7(a), although the facial appearance is partially cropped, our method can still synthesize a plausible lip and recover the original skin color. In general, our method can handle better for those repeated patterns such as the polka dots and strips. Although the PISE [15] can transfer the basic outline of the strips, it is obvious that the whole appearance is blended by unreliable flow-warping, just like GFLA [8]. We believe that the AdaSS Module provides positive effect to extract features in a wider field of view so that the completeness of the texture can be maintained.

Secondly, we want to investigate the transferring performance when there is no facial expression. It narrows the perceptual evaluation range down to the garments only. As illustrated in Figure 7(b), our method can reasonably follow some texture guidance of the reference images like the camisoles of the model. More specifically, the generated results of ADGAN [14] are very similar although the source texture is totally different. Being too general for the style codes is the drawback of decoupling attributes by a shared encoder. By encoding the attributes in group instead of separately, the result shows that our method can build a connection among the decoupled attributes during encoding stage so that a better visual quality can be generated.

Thirdly, it is important to keep as many details as the source image except the target attributes during the texture transfer. In Figure 7(c), our method can also transfer more texture information to the target image. There is a highlight that our method can preserve the shape as well. For example, it can successfully retain the hat with original shape and color while other two approaches fail to do it.

Last but not least, we focus on some texture with decorative patterns. As presented in Figure 7(d), we demonstrate different continuous patterns such as chiffons, polka dots which are hard to be synthesized. Benefited to the spatially aware normalization in our ShaTure Block, our method can retain more unique patterns at closely geometric location. Meanwhile, it fine-tunes the whole representation to fit the target shape. Compared to other approaches, our method can perform better texture transfer under the same settings.

2) Multiple Attribute Transfer: In order to illustrate the attribute controllability of our method, we present more results on attribute transfer with multiple attributes. As shown in Figure 8, the target attributes are required to be extracted through a segmentation mask. To demonstrate the diversity, we provide three kinds of attributes including upper body, pants and face&hair. Combining face with hair as one attribute can easily highlight the difference after transferring. We also randomly select several source images with different view angles such as front view, side view, back view and fully front view. We take the source image as the attribute base. By replacing target attributes from the reference images, we can obtain the corresponding style codes. For example, the examples in the last column represent the result of swapping the upper-body, pants and face&hair attributes extracted from other candidate images with the source image.

It is observed that the target attributes are successfully transferred to the source image while the visual quality of shape and texture can be maintained. The ability of interpolation for progressive attribute combination is also demonstrated. In general, our method can handle multiple attributes editing independently. In Figure 8(a-c), the models with same gender can preserve the style of the referenced upper-body garment and the referenced pants. The facial characteristics, contour of head and hair color can also be well transferred. Apart from the same gender, our method can also support cross gender style. We visualize the ability of changing the characteristics

TABLE III

QUANTITATIVE RESULT OF ABLATION STUDY ON THE SHATURE BLOCK AND ADAPTIVE STYLE SELECTOR COMPARED TO THE BASELINE NETWORK

Modules			IS↑	SSIM↑	FID↓	LPIPS↓
Baseline	ShaTure	AdaSS				
✓			3.273	0.769	16.014	0.178
✓	✓		3.387	$\underline{0.773}$	$\underline{13.756}$	$\underline{0.167}$
✓		✓	$\underline{3.477}$	0.770	15.521	0.175
✓	✓	✓	**3.487**	**0.777**	**11.434**	**0.159**

of other gender in Figure 8(d-e). In this setting, our method can not only naturally transfer the texture of garments but also adaptively synthesize the appearance to suit the dressing style for the genders.

E. Ablation Study

In order to prove the effectiveness and robustness of our network, we conduct a series of experiments on our main contributions including ShaTure Block and AdaSS Module. As shown in Table III, we examine our full model starting from a baseline model composing of a Style Decoder and a SPADE-liked [21] decoder only. Based on this setting, we replace the SPADE [21] normalization block with our ShaTure Block. With the assistance of style and texture enhancement, the performance is boosted up to 14% of FID score and 6% of LPIPS score. From another point of view, the FID and LPIPS score can further be enhanced by 16% and 4% with the help of AdaSS Module. It can certify the effectiveness of the ShaTure Block and AdaSS Module with minimum reconstruction error for image generation. Moreover, the AdaSS Module contributes a lot of effort for the improvement of IS score which is up to 6% increment. It can be interpreted that the multi-scale style representations are significant to maintain the global shape consistency. Lastly, by comparing the full module with each setting, the superiority of the metric scores demonstrates the efficacy of the proposed ShaTure Block and AdaSS Module. By digging into the details, we also provide a comprehensive ablation study on the elements of each suggested unit.

1) Analysis on ShaTure Block: The objective of the ShaTure Block is to progressively generate good-looking images by exploiting shape features and texture representations. By comparing the quantitative and qualitative results in Table IV and Figure 9, the full model can surpass its variants. We divide the quantitative results into 3 categories including Module analysis, Parallel design and Cascaded design in Table IV. The details of the settings are described as follows:

a) Module analysis: To recover a reasonable image, the shape of the person should be sharp and clear. Without a doubt, we have an obvious difference between the settings *W/o the Shape Module* and the *Full Model*. Moreover, the visual quality is not up to standard where many artifacts on the images are easily identified. We also compare the settings with similar parameters in *Parallel design – All Texture Module*, the large margin on the metrics indicates great importance of Shape Module in the ShaTure Block. Based on this observation, we can conclude that the Shape Module can provide significant contribution on the overall vividness of the generated images.

TABLE IV

QUANTITATIVE RESULT OF ABLATION STUDY ON DIFFERENT VARIANTS OF SHATURE BLOCK. TO SIMPLIFY THE NOTATIONS, WE USE S TO REPRESENT SHAPE MODULE AND T TO REPRESENT TEXTURE MODULE. THE STRUCTURE IS A GRAPHICAL REPRESENTATION TO ILLUSTRATE THE ARCHITECTURE OF CORRESPONDING DESIGN

Methods	Structure	IS↑	SSIM↑	FID↓	LPIPS↓
Module Analysis					
W/o Shape Module	$\begin{bmatrix} T \\ T \end{bmatrix}$	3.141	0.754	22.245	0.200
W/o Texture Module	$\begin{bmatrix} S \\ S \end{bmatrix}$	3.376	0.770	15.199	0.172
W/o STN	$\begin{bmatrix} S & T \\ T & S \end{bmatrix}$	3.389	0.775	12.529	0.163
Shared STN in Block	$\begin{bmatrix} S & T \\ T & S \end{bmatrix}$	3.442	$\underline{0.776}$	11.860	$\underline{0.160}$
Parallel Design					
W/o Braiding	$\begin{bmatrix} S & S \\ T & T \end{bmatrix}$	3.397	$\underline{0.776}$	12.006	0.161
With 1 Braiding Block	$\begin{bmatrix} S \\ T \end{bmatrix}$	$\underline{3.444}$	0.769	14.573	0.174
With 3 Braiding Blocks	$\begin{bmatrix} S & T & S \\ T & S & T \end{bmatrix}$	3.356	$\underline{0.776}$	11.728	0.161
All Shape Module	$\begin{bmatrix} S & S \\ S & S \end{bmatrix}$	3.441	**0.777**	12.226	0.161
All Texture Module	$\begin{bmatrix} T & T \\ T & T \end{bmatrix}$	3.279	0.771	13.659	0.169
Cascaded Design					
Double@Shape First	$\begin{bmatrix} S & S & T & T \end{bmatrix}$	3.428	0.775	12.285	0.163
Double@Texture First	$\begin{bmatrix} T & T & S & S \end{bmatrix}$	3.406	0.775	12.085	0.163
Twisting@Shape First	$\begin{bmatrix} S & T & S & T \end{bmatrix}$	3.358	$\underline{0.776}$	$\underline{11.446}$	0.161
Twisting@Texture First	$\begin{bmatrix} T & S & T & S \end{bmatrix}$	3.369	$\underline{0.776}$	12.105	0.162
Full Model	$\begin{bmatrix} S & T \\ T & S \end{bmatrix}$	**3.487**	**0.777**	**11.434**	**0.159**

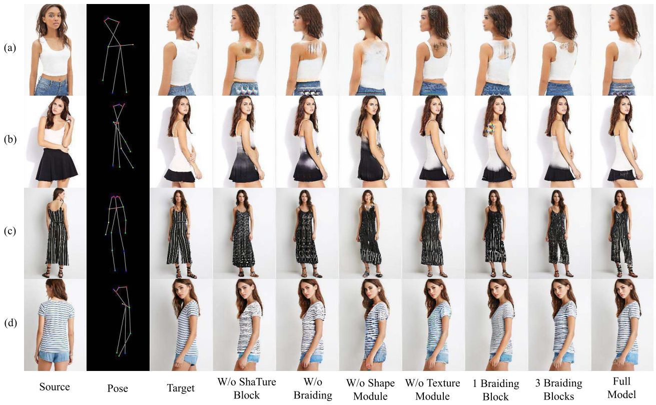

Fig. 9. Qualitative results of the ablation study on the ShaTure Block for different poses (a-d). Please zoom in for more details.

In addition to a sharp shape, the texture appearance is another significant factor during the reconstruction process. From the statistics in Table IV, the full model can obtain 2-3% improvement for the FID and IS score. By inspection on the images, it is clear that the distribution of strips on the skits and the color of T-shirt in Figure 9(c-d) are different from the source images. We also compare the settings with similar parameters in *Parallel design – All Shape Module* where the increment on evaluation metrics is also obvious. On the other hand, the full model can well preserve the texture completeness with higher fidelity. It verifies the functionality of the Texture Module inside the ShaTure Block as well.

The formulation of the spatial transformation network (STN) inside the Texture Module is also an important element to generate plausible texture of the garments.

TABLE V

QUANTITATIVE RESULT OF ABLATION STUDY ON DIFFERENT VARIANTS OF ADAPTIVE STYLE SELECTOR MODULE

Methods	IS↑	SSIM↑	FID↓	LPIPS↓
W/o R3	3.453	**0.777**	11.654	0.160
W/o R5	3.408	0.776	11.855	0.161
W/o R7	3.377	0.776	12.391	0.161
W/o R3+R5	3.385	0.775	12.253	0.162
W/o R3+R7	3.382	0.775	13.030	0.163
W/o R5+R7	3.355	0.771	13.016	0.168
W/o Shared Gate	3.429	0.776	12.005	0.163
W/o DW Conv	3.381	0.776	12.203	0.161
With Gate Group 4	3.431	0.776	11.631	0.160
U-Net Architecture	3.378	0.775	11.726	0.162
Full Model	**3.487**	**0.777**	**11.434**	**0.159**

The objective of the STN is to provide the affine transformation matrix to warp the source attributes to the target position based on the transformation relationship. Based on the decrements of IS and SSIM score of setting *W/o STN*, we can observe that the attributes should be spatially aligned with the target pose in order to mimic the texture characteristics. The setting *Shared STN* in Block illustrates that an independent STN for each Texture Module is more effective than the setting with a dependent STN. We believe that some little variations of transformation are sensitive to the evaluation metrics while it is hard to be identified by inspection.

b) Parallel design vs. cascaded design: The objective of twisting the Shape Module and Texture Module in a parallelly braiding manner is to exchange discriminative features for both feature-level space and pixel-level space. It enables 3-4% enhancement on the IS and FID score for our module. By visualizing the generated images, it has as many artifacts as the model *w/o ShaTure Module* shown in Figure 9(a-b). It shows that such braiding operation is effective on synthesizing photo-realistic images.

We provide a testing on the hyper-parameters of the numbers of braiding blocks. We observe that 2 braiding blocks are the optimal choice with satisfactory results and efficiency. Since there are only 2 modules needed to be braided, the combination of 4 modules inside a ShaTure Block is reasonable to fulfill the fine-tuning purpose.

Apart from a parallelly braiding design, we also evaluate the performance of the Shature Decoder using a sequentially cascaded design in each ShaTure Block. As illustrated in the Table IV, the parallelly braiding design is superior to all kinds of cascaded design. In general, the cascaded design is able to obtain satisfactory results in SSIM and LPIPS metrics but improvement of IS and FID scores is required. It represents the great fine-tuning ability of overall generation for the parallelly braiding design helps strike a balance between the shape and texture synthesis.

2) Analysis on Adaptive Style Selector Module: The motivation of the AdaSS Module is to adaptively extract multi-scale features from the source style by emphasizing cross-channel correspondence. As illustrated in Table V and Figure 10, the AdaSS Module has the merit of producing an accurate shape and a plausible texture appearance. The details of the settings are described as follows:

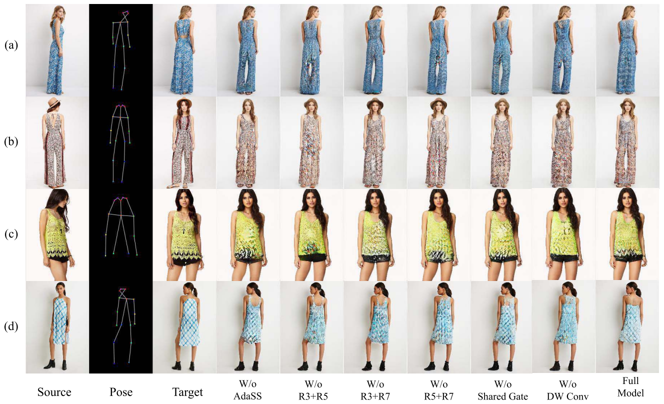

Fig. 10. Qualitative results of the ablation study on the Adaptive Style Selector Module for different poses (a-d). Please zoom in for more details.

a) W/o rate #: We decouple the size of the receptive field of view into different combinations. We believe that the larger the receptive field of view is set, the more local informative representation the network can get. From the numerical result of experiment, the overall performance of the network is increased from 1-12% with fewer artifacts when the receptive field of view is enlarged to 7. It verifies that a larger kernel size is able to provide positive gains to the pose transfer.

b) W/o shared gate: The objective of the shared gate is to dynamically enforce channel-wise correspondence to the style codes. It can also obtain 1-4% enhancement for the IS and FID score. Visually, the images generated from this setting have the worst shape such as Figure 10(d) which cannot produce a reasonable skirt. It shows that the shared gate can provide positive effect to network, especially in terms of shape reconstruction.

c) W/o DW & with group # Conv: As discussed in [37], [38], the objective of the depth-wise/group separable convolution is to reduce the number of parameters by partitioning convolution into a depth-wise kernel and a point-wise kernel. Apart from the network optimization, it indicates that there are 3-6% increments for the IS and FID scores as well. It proves the efficacy of the gated/grouped convolution in the AdaSS Module.

d) U-net architecture: In this setting, we investigate the effectiveness of the source of style codes. Instead of extracting multi-scale style features at the end of the Style Encoder, we can utilize the features of different resolutions in the implementation similar to the architecture of U-Net [36]. However, the performance is obviously inferior to some extent in all the evaluation metrics. It is because there is no consistent contextual information to generally represent the style codes inside the shallow layers. On the other hand, our AdaSS Module can effectively formulate the multi-scale style codes from a consistent feature space by self-recalibrating the same feature source through channel-wise attention.

F. Feature Map Visualization

To illustrate the role of the Shape Module and Texture Module in the ShaTure Block, we provide some visualizations of warped attributes and feature maps in Figure 11. Although

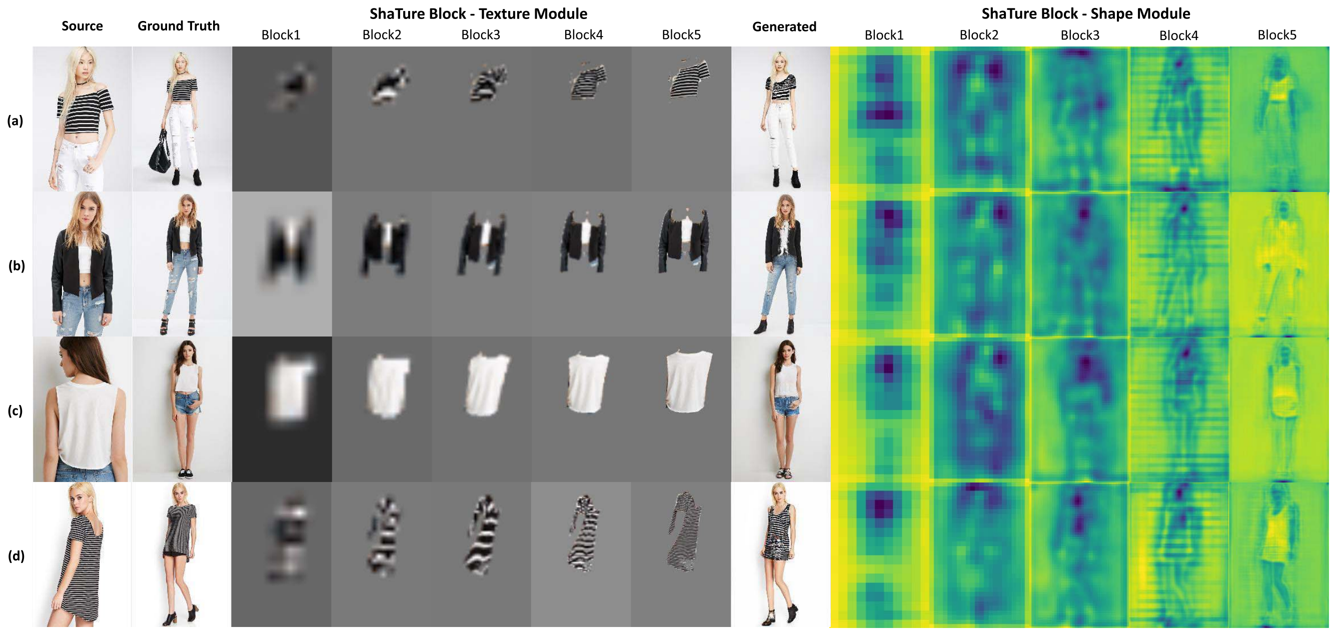

Fig. 11. Visualization of some warped attributes of Texture Module and shape features of Shape Module in ShaTure Block. For the Texture Module, we utilize the affine transform matrix in the module to warp the source attributes so that it can clearly show the effectiveness of the spatial transformation. The visualization of feature maps in Shape Module can demonstrate the generation process of target shape. All images and features are bilinearly scaled to final generation resolution where Block1 is the lowest resolution and Block5 is the highest. We demonstrate different angles of view including front (a-b), back (c) and slide (d) view. Please zoom in for more details.

there are two Shape Modules and Texture Modules in a ShaTure Block, we observe that there is no obvious visual difference between them. Therefore, we pick the first unit as the sampled module. Due to different resolutions for each block, we use bilinearly up-sampling method to resize the images for visualization purpose.

We provide the feature maps to demonstrate the density distribution for Shape Module. We generate the density map by summing the produced output features across channel dimension. It is obvious that the feature maps can represent the contours of target shape generated from coarse to fine fashion. For Texture Module, we show some warped attributes according to the affine transform matrix so that the ability of texture alignment and generation is illustrated. Although the visual quality of warped features is degraded due to bilinearly down-sampling operation, we can clearly distinguish the difference of warped attributes after the affine transformation. Since there is an independent spatial transformation network inside the Texture Module, the spatial transformation is also invariant to each ShaTure Block as well.

Based on the visualization of the ShaTure Block, we can observe that the disentanglement between the shape and texture information is effectively aligned with our theoretical principle. The evolution of Shape Module indicates that the network can globally synthesize the overall shape in lower resolution stage while refining the details on the human parts in higher resolution stage. We believe that it is the contribution of Shape Decoder to provide semantic signals to preserve the reconstruction of human-body shape. On the other hand, the distorted high frequency signals can be maintained by Texture Module. The texture patterns are spatially sensitive to spatial locations. Corresponding spatial texture features can be captured by spatially adaptive modulation operation. With the aligned shape and texture features, we can produce plausible images effectively.

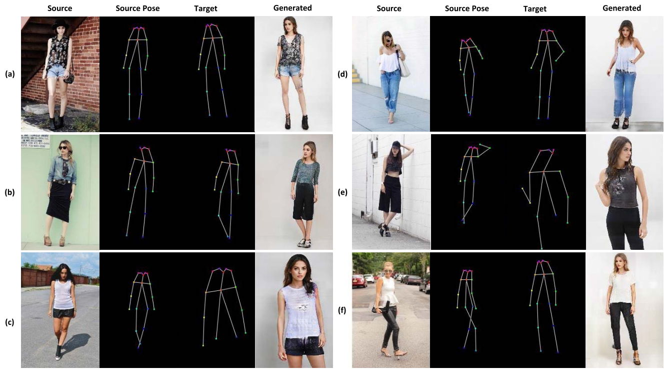

Fig. 12. Visualization of some pose transfer results on in-the-wild images. The images (a-f) are inferenced directly by the model pre-trained on DeepFashion dataset instead of training from scratch.

G. In-the-Wild Pose Transfer

Producing high-fidelity pose transfer results for in-the-wild images is an interesting application for both academic and industry practitioners. Due to limited high-quality training dataset for paired human images, we use the pre-trained model on DeepFashion to perform inference on some randomly selected images from a human parsing dataset – ATR [53]. In Figure 12, It is observed that most of the images can successfully perform pose transfer with natural shapes and vivid texture compared with the source images. Since there is limitation of DeepFashion dataset on background transfer, it is normal that the backgrounds of the generated images are highly similar to the style DeepFashion. We believe that this limitation can be alleviated when the training samples include in-the-wild settings.

V. Future Work

To achieve efficient pose and attributes transfer results in multi-terminal scenario, it is suggested to combine our method with video coding for machine (VCM) [54] techniques and lossless compression for key-point sequence [55] to jointly encode and transmit the massive visual data (images) and semantic data (skeleton, attributes) among devices. Through collaborative compression with feedback mechanism [54] and adaptive selection of prediction modes to minimize spatial and temporal redundancies [55], it can overcome the computational limitation for developing real-time style-transfer products.

VI. Conclusion

In conclusion, we propose a novel end-to-end framework to accomplish the task of human pose and attribute transfer. There are two main contributions of the suggested network architecture including the ShaTure Block and Adaptive Style Selector (AdaSS) Module. The goal of the ShaTure Block is to solve the problem of spatial misalignment by decoupling style codes and warped texture representation in a braiding manner. It can interchange discriminative features in both feature-level space and pixel-level space. We further propose an AdaSS Module to enhance the multi-scale feature extraction capability through channel-wise attention with dynamical self-recalibration of the feature maps. By emphasizing the aggregation of cross-channel correspondences with a larger field of view, the network can allocate the attention to the most appropriate style features to the corresponding reconstruction layers. Both quantitative and qualitative experimental results prove the effectiveness of the suggested ShaTure Block and AdaSS Module. The proposed framework can also achieve the state-of-the-art performance in all metrics which demonstrates the promising robustness and generality in the task of human pose and attribute transfer.

References

[1] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7543–7552.

[2] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by GAN improve the person re-identification baseline *in vitro*," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3754–3762.

[3] J. Walker, K. Marino, A. Gupta, and M. Hebert, "The pose knows: Video forecasting by generating pose futures," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 3332–3341.

[4] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," 2017, *arXiv:1705.09368*.

[5] L. Yang *et al.*, "Region-adaptive texture enhancement for detailed person image synthesis," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Jul. 2020, pp. 1–6.

[6] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-GAN for pose-guided person image synthesis," 2018, *arXiv:1810.11610*.

[7] Y. Li, C. Huang, and C. C. Loy, "Dense intrinsic appearance flow for human pose transfer," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3693–3702.

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

[9] Z. Zhu, T. Huang, B. Shi, M. Yu, B. Wang, and X. Bai, "Progressive pose attention transfer for person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2347–2356.

[10] S. Huang *et al.*, "Generating person images with appearance-aware pose stylizer," 2020, *arXiv:2007.09077*.

[11] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen, "Attgan: Facial attribute editing by only changing what you want," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5464–5478, Nov. 2019.

[12] W. Shen and R. Liu, "Learning residual images for face attribute manipulation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4030–4038.

[13] G. Zhang, M. Kan, S. Shan, and X. Chen, "Generative adversarial network with spatial attention for face attribute editing," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 417–432.

[14] Y. Men, Y. Mao, Y. Jiang, W.-Y. Ma, and Z. Lian, "Controllable person image synthesis with attribute-decomposed GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5084–5093.

[15] J. Zhang, K. Li, Y.-K. Lai, and J. Yang, "Pise: Person image synthesis and editing with decoupled gan," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 7982–7990.

[16] I. J. Goodfellow *et al.*, "Generative adversarial networks," 2014, *arXiv:1406.2661*.

[17] D. P Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[18] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*.

[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1125–1134.

[20] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4401–4410.

[21] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 2337–2346.

[22] L. Yang *et al.*, "Towards fine-grained human pose transfer with detail replenishing network," *IEEE Trans. Image Process.*, vol. 30, pp. 2422–2435, 2021.

[23] H. Tang, S. Bai, L. Zhang, P. H. S. Torr, and N. Sebe, "XingGAN for person image generation," in *Proc. Eur. Conf. Comput. Vis.* Springer, 2020, pp. 717–734.

[24] B. A. Olshausen, C. H. Anderson, and D. C. Van Essen, "A neurobiological model of visual attention and invariant pattern recognition based on dynamic routing of information," *J. Neurosci.*, vol. 13, no. 11, pp. 4700–4719, Nov. 1993.

[25] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[26] V. Mnih *et al.*, "Recurrent models of visual attention," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2204–2212.

[27] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[28] J. Hu, L. Shen, and G. Sun, "Squeeze- and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[29] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*.

[30] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[31] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 510–519.

[32] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[33] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.

[34] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Icml*, 2010.

[35] M. Jaderberg, K. Simonyan, A. Zisserman, and K. Kavukcuoglu, "Spatial transformer networks," 2015, *arXiv:1506.02025*.

[36] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, 2015, pp. 234–241.

[37] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[38] A. G. Howard *et al.*, "MobileNets: Efficient convolutional neural networks for mobile vision applications," 2017, *arXiv:1704.04861*.

[39] B. Xu, N. Wang, T. Chen, and M. Li, "Empirical evaluation of rectified activations in convolutional network," 2015, *arXiv:1505.00853*.

[40] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, 2016, pp. 694–711.

[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.

[42] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[43] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 768–783.

[44] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering robust clothes recognition and retrieval with rich annotations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1096–1104.

[45] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 7291–7299.

[46] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 932–940.

[47] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," 2016, *arXiv:1606.03498*.

[48] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[49] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," 2017, *arXiv:1706.08500*.

[50] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[51] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[52] A. Siarohin, E. Sangineto, S. Lathuiliere, and N. Sebe, "Deformable GANs for pose-based human image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3408–3416.

[53] X. Liang *et al.*, "Human parsing with contextualized convolutional neural network," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jan. 2015, pp. 1386–1394.

[54] L. Duan, J. Liu, W. Yang, T. Huang, and W. Gao, "Video coding for machines: A paradigm of collaborative compression and intelligent analytics," *IEEE Trans. Image Process.*, vol. 29, pp. 8680–8695, 2020.

[55] W. Lin *et al.*, "Key-point sequence lossless compression for intelligent video analysis," *IEEE Multimedia Mag.*, vol. 27, no. 3, pp. 12–22, Jul. 2020.

Lai-Man Po (Senior Member, IEEE) received the B.S. and Ph.D. degrees in electronic engineering from the City University of Hong Kong, Hong Kong, in 1988 and 1991, respectively.

He has been with the Department of Electronic Engineering, City University of Hong Kong, since 1991, where he is currently an Associate Professor with the Department of Electrical Engineering. He has authored over 150 technical journals and conference papers. His research interests include image and video coding with an emphasis on deep learning-based computer vision algorithms. He is a member of the Technical Committee on Multimedia Systems and Applications and the IEEE Circuits and Systems Society. He was the Chairperson of the IEEE Signal Processing Hong Kong Chapter in 2012 and 2013. He also served on the Organizing Committee of the IEEE International Conference on Acoustics, Speech and Signal Processing in 2003 and the IEEE International Conference on Image Processing in 2010. He was an Associate Editor of *HKIE Transactions* from 2011 to 2013.

Jingjing Xiong (Graduate Student Member, IEEE) received the B.S. degree in mechanical design, manufacturing, and automation from Xiangtan University, Hunan, China, in 2015, and the M.S. degree in artificial intelligence and pattern recognition from the Shenyang Institute of Automation, Chinese Academy of Sciences, Liaoning, China, in 2018. She is currently pursuing the Ph.D. degree in electrical engineering with the City University of Hong Kong, Hong Kong. Her research interests are in image segmentation, deep learning, and computer vision.

Yuzhi Zhao (Graduate Student Member, IEEE) received the B.Eng. degree in electronic information from the Huazhong University of Science and Technology, Wuhan, China, in 2018. He is currently pursuing the Ph.D. degree with the Department of Electronic Engineering, City University of Hong Kong. His research interests include image processing, deep learning, and machine learning.

Wing-Yin Yu (Graduate Student Member, IEEE) received the B.Eng. degree in information engineering from the City University of Hong Kong in 2019. He is currently pursuing the Ph.D. degree with the Department of Electrical Engineering. His research interests are deep learning and computer vision.

Pengfei Xian (Graduate Student Member, IEEE) received the B.Eng. degree in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 2017. He is currently pursuing the Ph.D. degree in electrical engineering with the City University of Hong Kong. His research interests include instance and semantic segmentation on images and videos, deep learning, and computer vision.