IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 20 May 2023, accepted 19 June 2023, date of publication 27 June 2023, date of current version 5 July 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3290102

RESEARCH ARTICLE

Style-Content-Aware Adaptive Normalization Based Pose Guided for Person Image Synthesis

WEI WEI[1,2,3,4], **XIA YANG**[1], **XIAODONG DUAN**[1,2,3], **AND CHEN GUO**[4]

[1]School of Computer Science and Engineering, Dalian Minzu University, Dalian 116650, China
[2]National Ethnic Affairs Commission of the People's Republic of China Key Laboratory of Big Data Applied Technology, Dalian Minzu University, Dalian 116650, China
[3]Dalian Key Laboratory of Digital Technology for National Culture, Dalian Minzu University, Dalian 116650, China
[4]Marine Electrical Engineering College, Dalian Maritime University, Dalian 116026, China

Corresponding author: Xia Yang (yangfeixia88@163.com)

This work was supported in part by the Grant from the Foundation of Liaoning Province Education Administration of China under Grant LJYT201911, and in part by the Dalian Minzu University Special Projects for Serving National Strategy under Grant 2020fwgj075.

ABSTRACT Most of the tasks based on pose-guided person image synthesis have obtained accurate target pose, but still have not obtained reasonable style texture mapping. In this paper, we propose a new two-stage network to decouple style and content, which aims to enhance the accuracy of pose transfer and the realism of a person appearance. Firstly, we propose an Aligned Multi-scale Content Transfer Network(AMSNet) to predict the target edge map for pose content transfer in advance, which can not only preserve clearer texture content but also alleviate spatial misalignment through advancing to transfer pose information. Secondly, we propose a new Style Texture Transfer Network(STNet) to gradually transfer the source style features to the target pose to for reasonable distribution of styles. To achieve highly similar appearance texture to the source style, we use a style-content-aware adaptive normalization method. The source style features are mapped into the same latent space as aligned content images (target pose and edge), and consistency between style texture and content is enhanced through adaptive adjustment of source style and target pose. Experimental results show that the proposed model can synthesize target images consistent with the source style, achieving superior results both quantitatively and qualitatively.

INDEX TERMS Person image synthesis, pose transfer, style-content-aware adaptive normalization.

I. INTRODUCTION

Pose guided person image transformation is an image generation task that synthesizes arbitrary target poses conditioned on the person source image. It has many potential applications such as data augmentation for person re-identification [1], [2], [3], video generation [4], [5], [6] and virtual try-on [7], [8], [9], [10]. In recent years, the conversion of source images into target poses using conditional GAN has achieved significant success, such as PATN [11], XingGAN [12], ADGAN [13], PoNA [14] and so on. These methods based on the conditional GAN method insert multiple repetition modules and learn sparse correspondences between poses through neural networks to reassemble source image features to target poses. However, these methods cannot retain the features of the source style, making it difficult to predict clear and reasonable target images. To solve this problem, flow-based methods [6], [15], [16], [17] guide the source features to be warped to a reasonable target pose by predicting the correspondence of the position between the source and the target, so as to obtain more accurate and realistic texture image, but the source and target poses will face large deformations to produce noticeable artifacts.

In order to alleviate the misalignment problem caused by large pose variations, some methods [18], [19], [20], [21] introduce human parsing map to provide semantic relationships corresponding to the target pose to synthesize the target image closer to the source style. These methods can synthesize a more satisfactory target person images, but they still cannot generate realistic texture details. The aforementioned

The associate editor coordinating the review of this manuscript and approving it for publication was Davide Patti.

methods encounter three challenges to synthesize satisfactory images: 1) Insufficient information on style textures, 2) Difficulty in pose transfer, 3) Difficult to predict regions that are not present in the source image. To generate vivid target images and decouple content and style for person image synthesis, we propose a two-stage network structure that maps aligned pose information and non-aligned style features into an aligned feature space.

In the first stage, the aligned multi-scale content transfer network is utilized to learn a source-to-target edge mapping as an intermediate result. This network explicitly highlights high-frequency signals of content information, which helps to alleviate the difficulty of pose transformation while providing spatial contextual clues for character identity features and clothing features. To better preserve the source image style and predict the information of invisible regions, flow-based operations are used. These operations accurately extract the texture of the source image by assigning local feature patches to each target location. They also preserve the source image details by warping multi-scale source features at the pixel level, predict invisible regions using feature-level local warping, and obtain a coarse target image that is similar in style and very close to the target pose jointly using soft weighting. However, due to the limitations of the flow operation, such as fuzzy garment boundaries and a lack of spatial contextual relationship, image artifacts are apparent, and the clothing texture differs from the source image. To solve this problem, the authors propose the style-content-aware adaptive normalization method in the second stage. By acquiring the correspondence between the source image and the rough target image, similar style features in the source image and target content information are mapped to the same latent space. Content and style information with learnable parameters are injected to adjust the style distribution of the rough target image so that its style texture and content information can be evenly and reasonably distributed. This adjustment leads to the generation of more accurate content and detailed appearance styles. The model has the following three contributions:

1) To address the absence of content information, the authors utilize an edge map as an added constraint on the bit-pose heat map. This constraint guides the network to enhance texture details and better preserve the original content of the image.
2) The proposed method, namely the style-content-aware adaptive normalization, explicitly distributes the style features of the source image to the target bit pose. Moreover, it injects the source style layer by layer to ensure a high consistency between the image style and the source image, while minimizing output artifacts.
3) Our proposed model involves a two-stage approach of explicit perception, which effectively decouples the shape and style of clothing. Through extensive experiments using the DeepFashion dataset, we have demonstrated the effectiveness of our model. In fact, our synthesis quality has significantly improved in terms of both quantitative metrics and user studies.

II. RELATED WORKS

A. HUMAN POSE TRANSFER

Since the introduction of the Pose-Guided Image Generation (PG2) task [22], it has garnered a lot of attention from scholars worldwide. However, existing pose transfer methods, such as Def-GAN [23] and LiquidGAN [6], use rigid geometric deformation, which lacks the flexibility needed to extract accurate motion and can blur the body boundaries. To address these issues, PATN [11] employs a local attention mechanism to gradually guide the image information from the source pose to the target pose, while ADGAN [13] uses a decomposed style code in a texture encoder to obtain a style vector and the AdaIN [24] residual block to inject style features into the target pose for image synthesis. XingGAN [12] proposes a cross-attention block between style and pose to repeatedly fuse features from the target pose and source appearance. However, these methods lack alignment operations between source appearance and target pose and cannot generate ideal appearance textures. Other methods attempt to alleviate the difficulties of pose distortion, such as GFLA [16], which estimates 2D flow and occlusion masks based on source images, source, and target poses, and warps the source local patches to match the desired pose, or Li [15] and LiquidGAN [6], which use 3D models to guide geometric deformations within the foreground region. These methods generate realistic textures but cannot extract accurate motion for complex deformations or severe occlusions, resulting in significant artifacts. Recently, some researchers have used a two-stage network structure, such as PISE [20], PINET [25] and SPGNet [18], to synthesize a target resolution map with the source semantic map, source pose, and target pose as inputs to broadcast the appearance image to the corresponding semantic region. These approaches show that pose transfer of target parsing mappings has excellent potential but lacks significant texture constraints and generates smoother images. Additionally, CoCosNet [26], [27] computes dense correspondences between cross-domain images through attention-based operations, while DPTN [28] aids source-to-source reconstruction for source-to-target learning, and [29] uses unsupervised methods. CASD [30] generates very accurate target poses using cross-attention style blocks, and Scam [31] modulates the stylistic information of semantic regions based on cross-attentive methods. However, these methods still fail to retain detailed source style features.

B. IMAGE TRANSLATION

Generative adversarial networks (GANs) [32] have been successfully applied to image translation tasks by minimizing the domain discrepancies between generated images and real samples using generators and discriminators. Pix2Pix [33] is an effective method for conditional image translation based on specific input constraints. With the proposed Spatial Adaptive Normalization (SPADE) [34], the activation in the normalization layer is adjusted using a variant of Adaptive Instance Normalization (AdaIN) [24] to inject content information and synthesize a new image for a given semantic input. Similarly, StyleGAN [35] also uses Adaptive Instance Normalization (AdaIN) to achieve scale-specific control of image synthesis. A further improvement to SPADE [34] is seen in SEAN [36], which controls the per-region encoding to effectively broadcast the style to a particular region. However, these methods have relatively limited editing capabilities in human pose transformation due to sparse correspondence of keypoints, large pose and texture variations.

III. METHOD

Given the source image I_s and the target pose P_t, a realistic person image I_g with the same pose P_t and the consistency in appearance I_s is generated. Direct prediction of target images often fails to achieve the expected results due to the large variation in the field of view and poorly characterized regions brought about by self-occlusion. To reduce the complexity of the target image synthesis, the image content is pre-transferred and the constraints on the pose features and the refinement of the style texture features are strengthened. So we propose a new two-stage pose-guided person image generation model, where in the first stage the source edge map E_s, source pose P_s and target pose P_t are input to an aligned multi-scale content transfer network (AMSNet) to gradually generate the target edge map E_g, while in the second stage, the predicted target edge map E_g, source pose P_s, target pose P_t and source image I_s are used to generate a target pose image I_g that preserves the source style features through the appearance texture transfer network (STNet). The general framework of the method is shown in Fig.1, Please refer to the appendix for the detailed network structure.

A. ALIGNED MULTI-SCALE EDGE CONTENT TRANSFER NETWORK

The feature space of the source image I_s includes style features and content information, and it is difficult to map the features of the source image I_s to the target image I_g directly due to the intermixing of these two types of information. Therefore, the content information of the image is pre-transferred, which is used to reduce the complexity of target image synthesis and to strengthen the constraints of stylistic features. Experimentally, it is demonstrated that prior transfer of content information using pixel-level edge mapping can highlight the high-frequency signals of identity features. In order to compute the accurate transformation relationship between source and target poses and generate content-rich edge information. In this paper, we propose aligned multi-scale attention blocks that differ from previous approaches that implicitly compute the relationship between source and target poses to predict the attention mask. We explicitly compute the pose relationship between the source and target, and place the sampled source edge features at the locations indicated by the target pose according to this pose alignment relationship. This computational mechanism is able to maintain edge Features better and predict clearer and more accurate target edge information to accurately guide the appearance of the second stage of texture transfer. As shown in Fig.1, the source image I_s is used to perform edge detection to obtain the corresponding edge map E_s. And then E_s, source pose P_s and target pose P_t are input to an AMSNet to estimate a grayscale edge content map $E_g \in (0,1)$ aligned with P_t. AMSNet consists of the aligning multiscale attention decoder and three encoders. The formulation is defined as:

$$E_g = \mathbb{G}_E \left(E_s, P_s , P_t\right) \tag{1}$$

where $\mathbb{G}_E (\cdot)$ denotes AMSNet. The use of sparse key points to represent the pose information only provides limited body structure, ignoring the inter-connection and correspondence of various parts, and thus cannot handle some complex poses (e.g., legs crossed, hands on head, arms crossed, etc.). To align source and target poses better, content-rich edge information is generated. As shown in Fig.1, we propose Aligned Multi-scale Attention Blocks to compute the similarity of all pixel points between the source pose and the target pose, which guides the transfer of content information. The formulation is defined as:

$$M_{st} = soft\max\left(Conv^{\rho}\left(f_{ps}\right) \times \left(Conv^{\tau}\left(f_{pt}\right)^{T}\right)\right) \tag{2}$$

$$f_{eg} = f_{se} + \left(Conv^{\gamma}\left(f_{se}\right) \times M_{st}\right) \tag{3}$$

where $Conv^{\gamma}(\cdot)$, $Conv^{\rho}(\cdot)$ and $Conv^{\tau}(\cdot)$ denote the 1×1 convolution layer, × represents the matrix inner production. f_{se}, f_{ps} and f_{pt} represent the extracted features of the source edge map, source pose and target pose, respectively. M_{st} denotes the similarity of all spatial location correspondences, and the sum of each row is 1. The content is transferred by weighted matrix summation. In addition, to ensure less information loss during the transfer process, the source edge information is added to the edge information pixels after the pose attention calculation, which also accelerates the convergence of the model.

B. STYLE TEXTURE TRANSFER NETWORK

The main objective of the target image generation network is to transfer the texture information and style features from the source image to the target pose to obtain a highly similar appearance of the person image to the source image. This is a translation problem of mapping edge information to images conditioned on the source image, but the variation between the source image and the target image is large, and some visible regions in the target image are not visible in the source image. So in addition to predicting the target edge map, existing flow-based operations are used to warp the source image to the target image to obtain a coarse target image I_{crs}. More specifically, the P_t, I_s and P_t are stitched in the channel dimension, their correspondence is calculated, and the source image is globally and locally warped according to the correspondence, thus synthesizing images with invisible regions in style yet with realistic local details.

To conclude, a technique similar to the existing normalization is applied to regulate the style of I_{crs} by adjusting its scale and bias. However, previously used techniques tend to

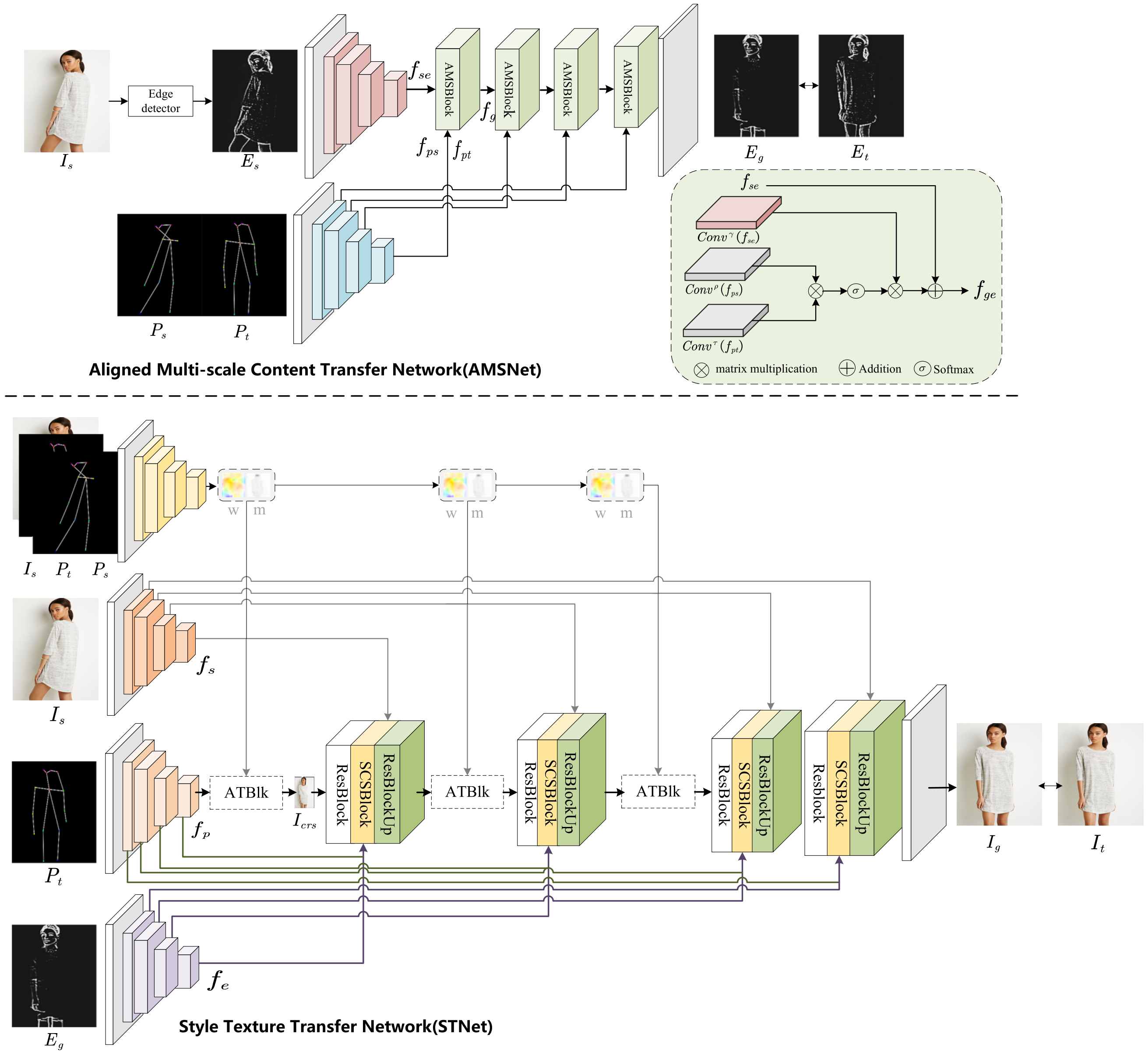

FIGURE 1. The overview of our model. We perform content transfer through the AMSNet at first, and then use the transferred edge content and target pose to guide style injection. The STNet is responsible for progressively synthesizing a realistic-looking person image using SCSBlock.

lose spatial information and texture details from the source image. To address this issue, the style-content-aware adaptive normalization method is proposed. This method dynamically adjusts the content and style activation mapping to allow for flexible demodulation of the target feature distribution, using feature maps of P_t, I_s, and E_g as input. The main advantage of this method is that it preserves the spatial contextual relationship of the source image while highlighting the high-frequency signal of the texture through information from the target edge. As a result, the clothing style appears much more realistic. Finally, the target image is generated using multiple upsampling blocks in the multi-scale feature space following the style-content-aware adaptive normalization. The formulation is defined as:

$$I_g = \mathbb{G}\left(P_t, I_s, E_g, P_s\right) \tag{4}$$

where $\mathbb{G}(\cdot)$ denotes the Style texture transfer generation network. We utilize the same encoder structure to extract pose features, edge features, and style features from input layers with different channel numbers. Specifically, we have 30 channels for pose features, 1 channel for edge features, and 3 channels for style features. The encoder consists of two down-sampled convolutional layers. To better capture the positional information of the pose, we include straight lines between points to model the pose structure. There are a total of 18 keypoints and 12 lines, which results in 30 input

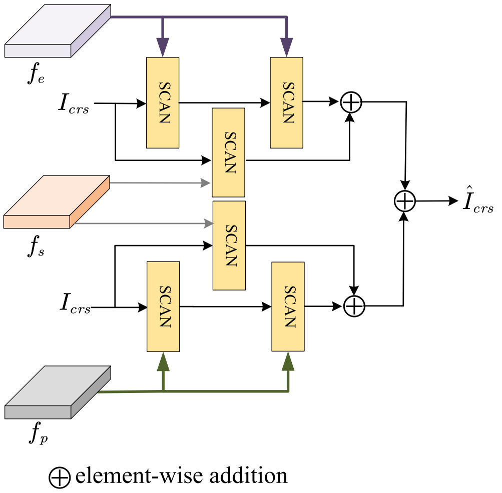

FIGURE 2. Residual structure of the SCSBlock.

channels for the pose encoder. Inspired by [16], we employ the feature deformation operation (ATBlk) to warp the source image into a coarse target image I_{crs} with an ambiguous style and target pose. To refine the style, we apply the style content adaptive normalization layer based on the appearance similarity between the source and target. The final output image is generated through multiple stacks of ResBlock, SCSBlock, and ResUpBlock in the multi-scale feature space. Overall, this method produces realistic clothing style transfer results while preserving spatial information and texture details.

1) DEFORMATION OF STYLE FEATURES

Numerous academics have proposed appearance warping methods [15], [16], [19], [23] to achieve the effect of source style warping by learning global or local spatial transformations. The flow-based operation can extract vivid source textures by assigning a very local patch to each target location, but it cannot capture the complex deformation capabilities between source and target, so a combination of pixel-level and feature-level flow-based operations is used to ensure non-rigid transformation of source and target poses. The flow field of relative motion between source and target poses is calculated by sampling local source regions for each output, using average pooling to force the correlation matrix to a sparse matrix. Thus this operation can help to reconstruct vivid source textures. Specifically, the flow estimator F takes the source image I_s, source pose P_s and target pose P_t as inputs to generate the flow field $w \in \mathbb{R}^{H\times W\times 2}$ and occlusion mask $m \in \mathbb{R}^{H\times W\times 1}$ by analyzing the difference between the source image and the predicted target image. The formulation is defined as:

$$w, m = F\left(I_s, P_s, P_t\right) \tag{5}$$

where w is the 2D coordinate offset between source and target, and m denotes the presence or absence of target location information in the source image. $F(\cdot)$ is the structure of the auto-encoder, which first extracts features from the

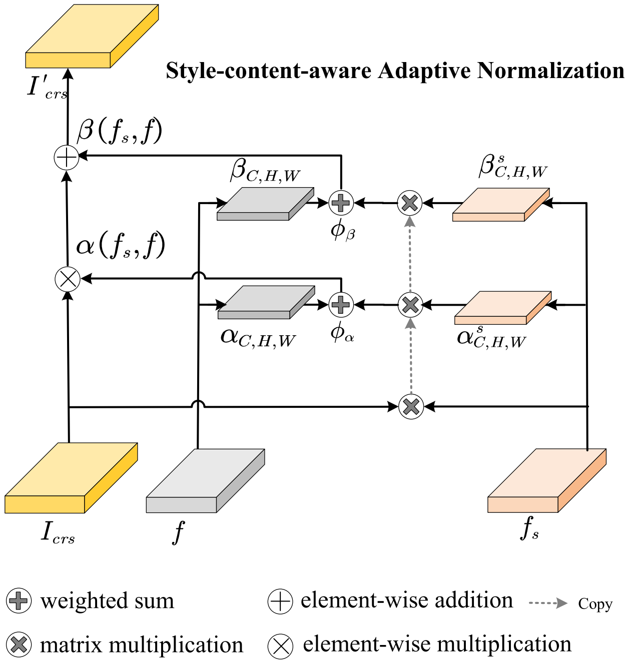

FIGURE 3. Details of the style-content-aware adapation normalization.

input information, and then decodes them into flow fields and occlusion masks based on the extracted features, preserving local and global contextual relationships through some jump connections. The flow-based approach warps the input data at the pixel level to preserve high-frequency details but may be warped by coarse flow maps and occlusion. However, warping the input data at the feature level can generate occlusion or new content. So after obtaining the flow field, the output of performing the flow-based operation using the joint feature layer and pixel level is:

$$I_{crs} = m \cdot I^p_{crs} + (1-m) \cdot I^l_{crs} \tag{6}$$

where I_{crs} denotes the output result of the flow-based operation, I^l_{crs} denotes the feature layer flow-based warping operation and I^p_{crs} denotes the pixel-level warping operation. Use the occlusion mask w with continuous values between 0 and 1 to select features between I^l_{crs} and I^p_{crs}. Where the pixel-level flow-based warping operation $I^p_{crs} = W_p\left(I_s, w\right)$ uses a bilinear difference method to sample the flow field of the input source image. Feature-level flow-based operations:

$$I^l_{crs} = W_f\left(I_s, P_t, w, m\right) \tag{7}$$

where W_f is a fully connected network that uses the source image I_s, target pose P_t, flow field w, and occlusion mask m as inputs to transform the information from the source to the target feature space by sampling the source features to obtain warped results I^l_{crs} at local locations. This operation is achieved using a local attention module similar to the one used in [16]. In order to make the appearance style of I_{crs} closer to the target image, the similarity between I_{crs} and the pre-trained VGG-19 features of the target image is increased to constrain them to be more consistent with the target features in the latent space, thus constraining them to be in the same domain. By doing so, the model can better capture the appearance transfer between the source and target images, resulting in more visually appealing and realistic results.

2) STYLE-CONTENT-AWARE ADAPTIVE NORMALIZATION

In Section III-B1 of the paper, it is mentioned that the flow-based operation leads to blurring of image boundaries, which causes loss of spatial context relationship in the source image, resulting in less realistic appearance in the resulting image. To enhance the level of style refinement in the generated image, in addition to adding similarity constraints, a 3 × 3 convolution layer is used to extract the spatial scale and bias from the source image feature map to inject the style details of the source image into the corresponding positions of the target image. But due to the misalignment of the spatial positions of the source image I_s and the ground-truth I_t, texture refinement is performed directly using the source image to extract spatial scales $\alpha(f_s, f)$ and bias $\beta(f_s, f)$, resulting in the generation of incorrectly spatially located image textures (e.g., the appearance of the side of the source image is placed in front of the body of the generated image). As depicted in Fig.3 of the paper, a potential solution to address the misalignment problem involves computing the similarity relationship between the coarse target image features I_{crs} and the source image features f_s. The similarity loss is first used to constrain the similarity between the VGG-19 features with the target pre-training such that the features are aligned with the ground-truth I_t in the same spatial domain. Then the correspondence between the features of the source image and the correlation matrix M is calculated, and the correspondence layer [37] is used to calculate the correlation matrix.

$$M(u, v) = \frac{I_{crs}(u)^T f_s(v)}{\|I_{crs}(u)\| \, \|f_s(v)\|} \tag{8}$$

where $I_{crs}(u)$ and $f_s(v)$ denote the channel-level centrated features of I_{crs} and f_s at the locations of u and v, respectively. By multiplying with the correlation matrix M, $\alpha^s \cdot M$ and $\beta^s \cdot M$, which represent the similar styles of the source image and the target image, can be transformed from the source image to the target image. Also, to obtain a more accurate pose and texture, the aligned content information (target pose and target edge mapping) is mapped to the same hidden space as the style features, and the content and style of the target image are adaptively adjusted. Suppose I'_{crs} denotes the normalized image, then it is expressed as

$$I'_{crs} = Conv\left(LReLU\left(\alpha(f_s, f)\left(\frac{I_{crs} - \mu}{\sigma}\right) + \beta(f_s, f)\right)\right) \tag{9}$$

$$\alpha(f_s, f) = \phi_\alpha\left(\alpha^s \cdot M\right) + (1 - \phi_\alpha)\,\alpha \tag{10}$$

$$\beta(f_s, f) = \phi_\beta\left(\beta^s \cdot M\right) + \left(1 - \phi_\beta\right)\beta \tag{11}$$

A set of modulation parameters α and β of the target edge map or target pose is learned by a separate convolutional neural network, and the weight of the modulation parameters is adjusted using a set of learnable weight parameters ϕ_α and ϕ_β. The modulation scale and bias are used to gradually refine and update I_{crs} to I'_{crs}. Thus, after I_{crs} has gone through the style content-aware normalization module, it not only has the target image content and details, but also retains the source image style and spatial context relationships. Finally, the spatially adaptive style content encoder is constructed from a series of style-content-aware adaptive encoding blocks as shown in Fig.2, and the output of the style content encoding block:

$$\hat{I}_{crs} = p(f_e, f_s) + q(f_t, f_s) \tag{12}$$

where $p(\cdot)$ and $q(\cdot)$ denote the style-content-aware normalization module with the residual links. I_{crs} is superimposed multiple times by ResBlock, SCSBlock, and ResUpBlock on different feature scales to generate an image close to the source image style I_g.

C. LOSS FUNCTION

In this paper, we first train AMSNet and flow generator separately. The complete image generation model is then trained end-to-end with a joint loss consisting of sampling correctness loss $\mathcal{L}_f$, bidirectional consistency loss $\mathcal{L}_{bfc}$, similarity loss $\mathcal{L}_{sim}$, reconstruction loss $\mathcal{L}_{L1}$, adversarial loss $\mathcal{L}_{adv}$, perceptual loss $\mathcal{L}_{per}$, style loss $\mathcal{L}_{style}$, contextual loss $\mathcal{L}_{cx}$ and LPIPS loss $\mathcal{L}_{LPIPS}$.

$$\begin{aligned}\mathcal{L} = {} & \lambda_f \mathcal{L}_f + \lambda_{bfc}\mathcal{L}_{bfc} + \lambda_{sim}\mathcal{L}_{sim} + \lambda_{L1}\mathcal{L}_{L1} + \lambda_{adv}\mathcal{L}_{adv} \\ & + \lambda_{per}\mathcal{L}_{per} + \lambda_{style}\mathcal{L}_{style} + \lambda_{cx}\mathcal{L}_{cx} + \lambda_{LPIPS}\mathcal{L}_{LPIPS}\end{aligned} \tag{13}$$

Among them, λ_f, λ_{sim}, λ_{bfc}, λ_{L1}, λ_{adv}, λ_{per}, λ_{style}, λ_{cx} and λ_{LPIPS} correspond to different hyperparameters to optimize and control the results, respectively.

1) SAMPLING CORRECTNESS LOSS

The pre-trained VGG can provide information about the spatial distribution of the image at the feature layer, so the VGG features of the source image v^l_{s,w^f} are wraped with the VGG features of the ground-truth v^l_t using a flow-based operation, and the sampling correctness loss minimization prediction w^f is calculated for all N locations in the feature map Ω. This loss function is expressed as

$$\mathcal{L}_f = \frac{1}{N}\sum_{l \in \Omega} \exp\left[-c\left(v^l_{s,w^f}, v^l_t\right)\right] \tag{14}$$

where $c(*)$ denotes the cosine similarity operation, but the flow field computation uses only forward mapping to supervise flow learning is sensitive to disturbances in regions with similar features, and then a backward and forward mapping check with bidirectional consistency loss is added to distinguish the locations of similar features with

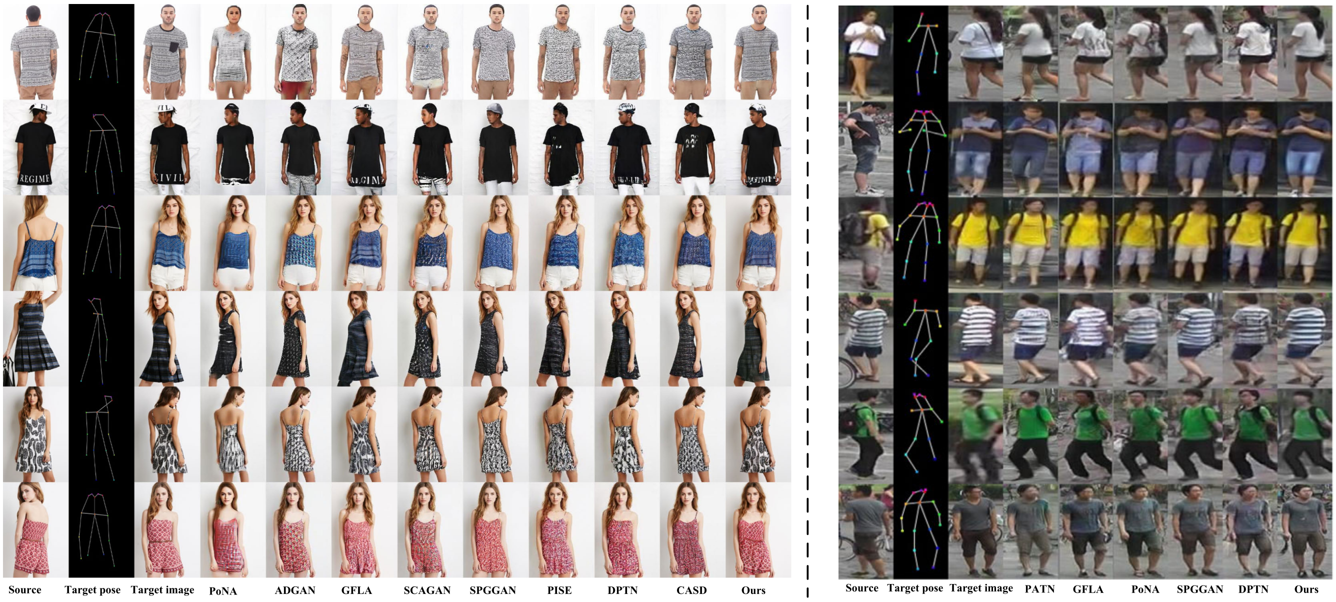

FIGURE 4. Qualitative comparisons with state-of-the-art methods on DeepFashion [38] and Market-1501 [39] respectively. From left to right are the results of PATN [11], PoNA [14], ADGAN [13], GFLA [16], SCAGAN [21], SPGNet [18], PISE [20], DPTN [28], CASD [30] and ours on DeepFashion [38], respectively. From left to right are the results of PATN [11], PoNA [14], GFLA [16], SPGNet [18], DPTN [28] and ours on Market-1501 [39], respectively.

a loss function of

$$\mathcal{L}_{bfc}=\sum_{l\in\Omega} m_l^f\cdot\eta\left(w^f(l)+w^b\left(l+w^f(l)\right)\right)+m_l^b\cdot\eta\left(w^b(l)+w^f\left(l+w^b(l)\right)\right) \tag{15}$$

where $\eta(\cdot)$ denotes the Charbonnier function, which computes the forward flow w^f and backward flow w^b and the masking mask m^f and m^b by two directions (i.e., source-to-target and target-to-source).

2) SIMILARITY LOSS

Pose-guided figure image generation requires the same pose as the target and the appearance to be consistent with the source image. Due to the large pose variation between the source and the target, it is impossible to precisely locate the style similarity between the target and the source occlusion region. Therefore, in this paper, in order to warp the source image to obtain a coarse target image I_{crs} aligned in the same domain with the target image features $\phi_i(I_t)$ extracted by pre-training VGG-19. The similarity loss is used to calculate the L_2 distance between the rough image and the target image and to constrain the style similarity between the generated image and the occluded region of the target image I_t. So as to improve the image generation qualities, which is defined as

$$\mathcal{L}_{sim}=\left\|I_{crs}-\phi_i(I_t)\right\|_2 \tag{16}$$

3) RECONSTRUCTION AND PERCEPTUAL LOSS

Reconstruction loss evaluates the consistency of the generated image I_g with ground-truth I_t at the pixel level, which is written as $\mathcal{L}_{L1}=\left\|I_g-I_t\right\|_1$. Perceptual loss is used to compute L_1 distance between features extracted from the pre-trained VGG-19 network in the multi-scale space. It is written as $\mathcal{L}_{per}=\sum_i\left\|\varphi_i\left(I_g\right)-\varphi_i(I_t)\right\|_1$, where $\varphi_i(\cdot)$ denotes the features extracted from the first layer of the pre-trained network.

4) ADVERSARIAL LOSS

We take the generated image I_g and the ground-truth I_t as the input of the image generation discriminator D, and penalize the distribution distance between them so that the distribution of the generated image I_g is getting closer to that of the target image I_t.

$$\mathcal{L}_{adv}=\mathbb{E}\left[\log\left(1-D\left(G\left(I_s,P_s,P_t,E_s\right)\right)\right)\right]+\mathbb{E}\left[\log D\left(I_t\right)\right] \tag{17}$$

5) STYLE LOSS

Style loss calculates the statistical difference in activation maps between the generated image and the target image and enhances the similarity of the color and style of the generated image to the target image.

$$\mathcal{L}_{style}=\sum_j\left\|\mathcal{G}_j^{\varphi}(I_t)-\mathcal{G}_j^{\varphi}\left(I_g\right)\right\|_2^2 \tag{18}$$

where $\mathcal{G}_j^{\varphi}(\cdot)$ is the Gram matrix of the activation mapping φ_j of the j-th layer of the VGG-19 network. Feature extraction is performed using features with the same weights.

6) CONTEXTUAL LOSS

The normalized cosine distance between feature maps is calculated by using the contextual loss proposed in [41] to

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

Model	DeepFashion							Market-1501				
	SSIM ↑	FID ↓	LPIPS ↓	PSNR ↑	R2G ↑	G2R ↑	Jab ↑	SSIM ↑	FID ↓	LPIPS ↓	PSNR ↑	Mask – LPIPS ↓
PATN [11]	0.4818	171.7202	0.4858	28.8791	-	-	-	0.1204	21.3277	0.3196	**28.0591**	0.1591
PoNA [14]	0.5093	28.9279	0.2623	30.2648	-	-	-	0.1089	25.2697	0.2949	27.9535	0.1605
ADGAN [13]	0.5679	21.3277	0.261	29.6755	17.35	35.45	0.45%	-	-	-	-	-
GFLA [16]	0.6916	12.1382	0.196	31.4324	20.58	32.23	10.52%	0.1127	23.6871	0.2817	28.0189	0.1483
PISE [20]	0.6679	19.3706	0.254	31.3816	22.71	37.08	13.76%	-	-	-	-	-
SCAGAN [21]	0.6799	19.3579	0.253	31.4624	21.83	33.57	11.54%	-	-	-	-	-
SPGNet [18]	0.6863	19.2531	0.242	31.2905	19.47	36.52	12.72%	0.1218	25.1263	0.2808	28.0432	0.1435
DPTN [28]	0.6862	18.7556	0.232	31.4354	18.36	34.09	11.48%	0.1143	21.6351	0.2713	28.0284	0.1391
CASD [30]	0.6977	14.5110	0.201	**31.6160**	23.97	40.11	20.57%	-	-	-	-	-
Ours	**0.7129**	**12.0635**	**0.195**	31.4936	**24.62**	**42.68**	**23.96%**	**0.1718**	**18.7876**	**0.2699**	28.0507	**0.1303**

TABLE 2. Comparison of model size and testing speed and floating point operations on DeepFashion dataset. "M" denotes millions, "G" denotes Giga and "fps" denotes frames per second.

Method	Params	FLOPs	Speed
PATN [11]	41.36M	124.02 G	**58.82fps**
GFLA [16]	14.04M	30.39 G	29.37 fps
PISE [20]	64.01M	122.74 G	0.53 fps
SCAGAN [21]	142.67M	517.27 G	1.34 fps
SPGNet [18]	117.13M	346.43 G	1.33 fps
DPTN [28]	**9.79M**	**26.56G**	30.82 fps
CASD [30]	58.51M	157.85 G	40.16 fps
Ours	68.41M	204.49 G	4.149 fps

measure the similarity of two spatially dislocated images.

$$\mathcal{L}_{cx} = -\log\left(CX\left(\phi_l\left(I_g\right), \phi_l\left(I_t\right)\right)\right) \tag{19}$$

where $\phi_l(\cdot)$ denotes the features extracted from the layers $l = [relu3\_2, relu4\_2]$ of pre-trained VGG-19, and CX denotes the cosine similarity measure between features.

7) LPIPS LOSS

In order to reduce distortion and learn perceptual similarity, LPIPS loss was additionally integrated, which has been shown to better preserve image quality compared to more standard perceptual losses and increase the realism of the image.

$$\mathcal{L}_{LPIPS} = \left\|\mathcal{F}\left(I_g\right) - \mathcal{F}\left(I_t\right)\right\|_2 \tag{20}$$

where $\mathcal{F}(\cdot)$ denotes the perceptual features extracted from the pre-trained VGG-16 network.

IV. EXPERIMENT

Dataset: This paper uses the DeepFashion (In-shop clothing Retrieval Benchmark) [38] and Market-1501 [39] datasets to conduct experiments. The DeepFashion dataset contains 52,712 high-quality images of people at a resolution of 256 × 256, while the Market-1501 contains 32,668 low-resolution images at 128 × 64 resolution. These images vary in terms of viewpoint, background, and illumination. We adopted the same approach to dataset segmentation as in [40]. The Extended Difference Gaussian (XDoG) edge detection method was used to extract the edge information of the images, and the Openpose pose estimator was utilized to generate pose heat maps with 18 channels, with each channel representing one pose position information.

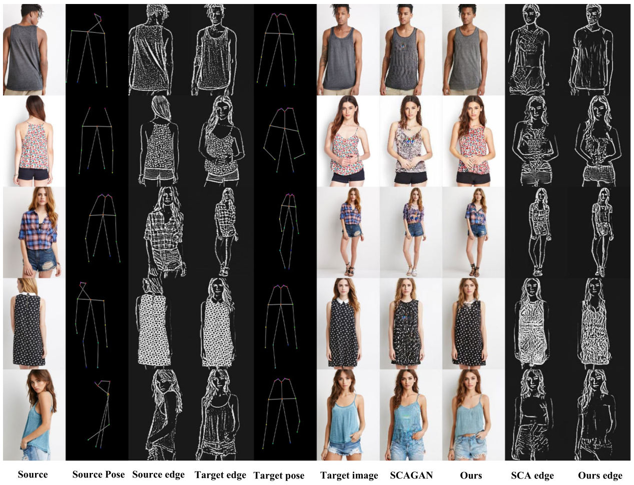

FIGURE 5. Qualitative comparison with SCAGAN [21].

TABLE 3. Quantitative comparison with SCAGAN [21] about edge image.

Method	SSIM ↑	PSNR ↑
SCAGAN [21]	0.6031	32.6996
Ours	**0.6228**	**33.3292**

Evaluation Metrics: This paper evaluates the effectiveness of the proposed approach using several image quality evaluation metrics, including SSIM [42], FID [43], LPIPS [44], and PSNR. These are commonly used evaluation methods in existing image generation tasks. SSIM and PSNR measure the quality of the generated image at the pixel level. LPIPS calculates the difference between the generated image and the real image in the perceptual domain. In addition, the authenticity of the images is measured using Fréchet Inception Distance (FID), which calculates the Wasserstein-2 distance between the generated data and the real data.

Experiment Details: To train the aligned multiscale content edge network and the style transfer network, the Adam optimizer [45] was used with $\beta_1 = 0.0$ and $\beta_2 = 0.999$. The networks were trained end-to-end for approximately 500k iterations, using the same parameter configuration. The initial learning rate was set to 0.0001 and the batch size was 8. Weights of the learning objectives in the image generation network were set as follows: $\lambda_f = 5.0$, $\lambda_{sim} = 100$, $\lambda_{bcf} = 0.1$, $\lambda_{L1} = 5.0$, $\lambda_{adv} = 2.0$, $\lambda_{per} = 0.5$, $\lambda_{style} = 500$, $\lambda_{cx} = 0.1$, and $\lambda_{LPIPS} = 1.0$. The generator and two discriminators were alternately trained using this configuration.

A. COMPARISON WITH STATE-OF-THE-ART METHODS

1) QUANTITATIVE COMPARISON

We conducted both qualitative and quantitative comparisons with several state-of-the-art methods, including PATN [11], PoNA [14], ADGAN [13], GFLA [16], SCAGAN [21], SPGNet [18], PISE [20], DPTN [28], and CASD [30]. The quantitative results are presented in Table 1, which shows that our proposed method achieved the best scores on all metrics except for PSNR on the DeepFashion dataset. This indicates that our method generated the most structurally accurate and realistic images. Additionally, our method demonstrated the best performance on most metrics compared to other state-of-the-art methods. On the other hand, when evaluated on the Market-1501 dataset, our method achieved the best results in all metrics except for PSNR. Furthermore, the visual results show that our method generated the best images overall.

2) QUALITATIVE COMPARISON

In Fig.4, we compared the results generated by different methods. The left half of the figure shows some typical qualitative examples from the DeepFashion dataset. It is clear that our proposed method produces finer appearance textures and realistic results (e.g., the first and second rows of the penultimate row). More importantly, our method is capable of preserving the identity information and facial features of the source images (e.g., the first and third rows). Although some methods such as ADGAN [13], SCAGAN [21], and PISE [20] can generate structurally accurate images, they fail to retain complex textures in the source images due to the lack of spatial deformation blocks. The flow-based GFLA [16] method can generate realistic textures, but noticeable artifacts are produced when the pose changes greatly and occlusion is observed. DPTN [28] introduces a source-to-source task to assist source-to-target learning by sharing weights and establishing a fine-grained mapping of all pixels between source and target. However, it ignores the learning of invisible regions, resulting in severe facial distortion. SPGNet [18] and CASD [30] produce results that are very similar to the target pose, but do not retain the complex texture of the source style. In contrast, our proposed method generates target pose features that are highly consistent with the source image.

The right half of the figure shows some examples from the Market-1501 dataset. Our method can generate images of people with clearer contours, such as shoes in the second row, shoulder straps in the third row, and stripes in the fourth row. In comparison, PATN [11] and PoNA [14] use the attention mechanism to achieve high transmission accuracy, but due to the lack of sufficient feature fusion between poses and images, the generated images still suffer from missing details. In conclusion, our method retains the integrity and clearer boundaries of people and their clothing.

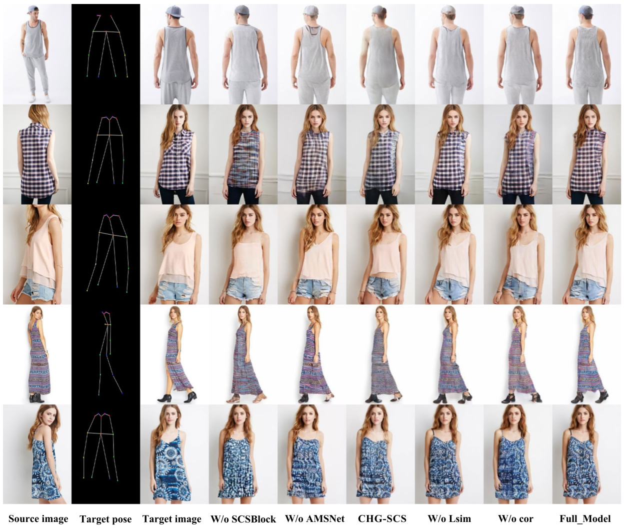

FIGURE 6. Qualitative results of ablation study.

As shown in Table 2. The comparison of the computational cost of our method with PATN [11], GFLA [16], PISE [20], SCAGAN [21], SPGNet [18], DPTN [28] and CASD [30] in terms of computational cost and speed of computation, the computational cost of our method is slightly increased and the speed of computation is slower. However, considering the high visual quality and transmission accuracy we obtained (see Table 1), this cost tends to be reasonable and acceptable.

3) USER STUDIES

Although quantitative and qualitative evaluation can provide some insights into the performance of the model, the human pose transfer task is primarily user-oriented. Therefore, we recruited 30 volunteers for an experiment to evaluate the performance of the model in terms of human perception. The evaluation included two aspects: 1) Comparison between the generated images and the real images. Thirty pairs of real and generated images were selected from the test set for evaluation, and the order of the images was randomized. Volunteers were asked to select the true or false images based on their first impression. 2) Comparison with images generated by other methods. Thirty pairs of images were selected, including source images, target poses, and images generated by eight different methods. Volunteers were asked to select the generated image that was closest to the source image and true image from the randomized order. The results of the experiment are shown in Table 1. We used the same measurement metrics as in [30]: R2G, which is the percentage of real images regarded as generated images; G2R, which is the percentage of generated images regarded as real images; and Jab, which is the percentage of images judged to be the best among all models. Higher values of these three metrics indicate better performance. From Table 1, it is clear that our model achieved the best results.

TABLE 4. Quantitative results of ablation study.

	SSIM ↑	FID ↓	LPIPS ↓	PSNR ↑
W/o SCSBlock	0.6653	13.5448	0.208	30.2165
W/o AMSNet	0.6916	12.1319	0.203	30.4324
CHG-SCS	0.6822	12.3153	0.199	31.0081
W/o$\mathcal{L}_{sim}$	0.6733	12.1390	0.197	31.4576
W/o cor	0.6998	12.2451	0.196	31.4224
Full-Model	**0.7129**	**12.0635**	**0.195**	**31.4936**

B. COMPARISON WITH SCAGAN

The SCAGAN model [21] utilizes pre-generated target edge mapping to enhance target image generation results. However, it directly injects source features and edge information into the spatial adaptive module and employs a progressive network structure to generate the final image. This approach leads to a loss of global semantic and texture information, resulting in less realistic target images. In this paper, we aim to address this issue by leveraging the region-adaptive normalized SEAN technique [34]. Specifically, we first map the source styles and rich target content information into the same latent space to complement the rough image. Next, we employ a feature-by-feature layer injection of both content and style to obtain more accurate content, while synthesizing more detailed appearance styles, resulting in more realistic target images.

Fig.5 demonstrates that the textures of the generated images are more reminiscent of the real target images when compared to SCAGAN [21]. The edge maps produced by our method (last two columns) preserve more detailed texture and pose features. Furthermore, the skin and texture of the person in the first and second rows appear more realistic, and the texture of the clothing in the second and third rows is closer to that of the target image. Hence, our method can generate precise personal images and appearance textures. Additionally, the edge maps produced by our network have a clearer definition and more comprehensive content than those of SCAGAN. This is supported by the quantitative results shown in Table 3, which highlight that our approach yields more accurate outcomes.

C. ABLATION STUDY

Ablation experiments were conducted to verify the validity of our model, and the contribution of each component was evaluated. As a result, we found that our model performs optimally when all components are integrated. This indicates that each component of the model contributes significantly towards achieving the final performance of the model.

1) W/O SCSBlock

By eliminating the SCSBlock and solely combining the target edge information with the target pose, this model may struggle to preserve the source image style and generalize the source style features. Instead, we propose an alternative approach that involves extracting the content and source style features based on correspondence, and integrating them into the same hidden space for adaptive fusion. This enables us to gradually generate the target image while also preserving the source style features, leveraging the use of residual and upsampling modules. Through this method, our model is more adept at generalization and achieving superior preservation of the source image style.

TABLE 5. Quantitative results of $\mathcal{L}_{per}$,$\mathcal{L}_{LPIPS}$ and $\mathcal{L}_{per} + \mathcal{L}_{LPIPS}$.

	SSIM ↑	FID ↓	LPIPS ↓	PSNR ↑
$\mathcal{L}_{per}$	0.6683	12.2451	0.1984	31.4224
$\mathcal{L}_{LPIPS}$	0.7025	12.0635	0.1949	30.4851
$\mathcal{L}_{per} + \mathcal{L}_{LPIPS}$	**0.7129**	**12.0638**	**0.1952**	**31.4936**

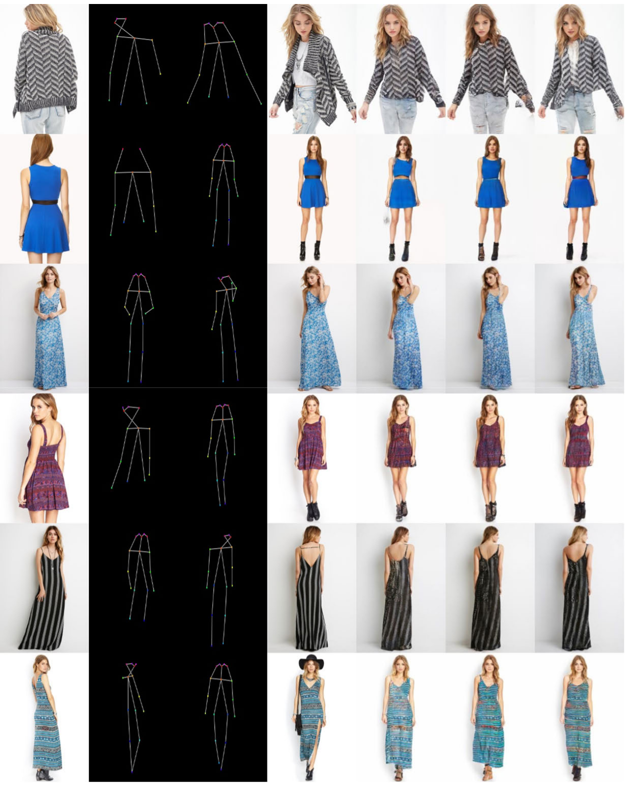

FIGURE 7. Qualitative results of $\mathcal{L}_{per}$,$\mathcal{L}_{LPIPS}$ and $\mathcal{L}_{per} + \mathcal{L}_{LPIPS}$.

2) W/O AMSNet

This model removes AMSNet while using the source edge map as input in the Style-content-aware Adaptive Normalization. Since there is not enough conditional transformation information to perform detailed spatial alignment, it increases the difficulty of the pose transformation and reduces the additional constraints on the texture, thus increasing artifacts.

3) CHANGE THE SCSBlock (CHG-SCS)

Changing the method of using Style-content-aware Adaptive Normalization in the model from using the source edge

TABLE 6. The structure of the generator G of AMSNet.

Input	$E_s(256\times256\times1)$	$P_s(256\times256\times18)$	$P_t(256\times256\times18)$
Intermediate Layers	Conv(F=64,k=4,s=2), IN, ReLU Conv(F=64,k=3,s=1), IN, ReLU Conv(F=64,k=3,s=1), IN Conv(F=128,k=4,s=2), IN, ReLU Conv(F=128,k=3,s=1), IN, ReLU Conv(F=128,k=3,s=1), IN	Conv(F=64,k=4,s=2), IN, ReLU Conv(F=64,k=3,s=1), IN, ReLU Conv(F=64,k=3,s=1), IN Conv(F=128,k=4,s=2), IN, ReLU Conv(F=128,k=3,s=1), IN, ReLU Conv(F=128,k=3,s=1), IN	Conv(F=64,k=4,s=2), IN, ReLU Conv(F=64,k=3,s=1), IN, ReLU Conv(F=64,k=3,s=1), IN Conv(F=128,k=4,s=2), IN, ReLU Conv(F=128,k=3,s=1), IN, ReLU Conv(F=128,k=3,s=1), IN
	f_{se}^1 =Conv(F=128,k=1,s=1)	f_{ps}^1=Conv(F=128,k=1,s=1)	f_{pt}^1= Conv(F=128,k=1,s=1)
	Conv(F=256,k=4,s=2), IN, ReLU Conv(F=256,k=3,s=1), IN, ReLU Conv(F=256,k=3,s=1), IN	Conv(F=256,k=4,s=2), IN, ReLU Conv(F=256,k=3,s=1), IN, ReLU Conv(F=256,k=3,s=1), IN	Conv(F=256,k=4,s=2), IN, ReLU Conv(F=256,k=3,s=1), IN, ReLU Conv(F=256,k=3,s=1), IN
	f_{se}^0=Conv(F=256,k=1,s=1)	f_{ps}^0=Conv(F=256,k=1,s=1)	f_{pt}^0=Conv(F=256,k=1,s=1)
	AMSBlock($f_{se}^0, f_{ps}^0, f_{pt}^0$)		
	UpSample(scale factor = 2), Conv(F=128,k=4,s=2), IN, ReLU Conv(F=128,k=3,s=1), IN, ReLU, Conv(F=128,k=3,s=1), IN		
	AMSBlock($f_{se}^1, f_{ps}^1, f_{pt}^1$)		
	UpSample(scale factor = 2), Conv(F=64,k=4,s=2), IN, ReLU Conv(F=64,k=3,s=1), IN, ReLU, Conv(F=64,k=3,s=1), IN		
	UpSample(scale factor = 2), Conv(F=64,k=4,s=2), IN, ReLU Conv(F=64,k=3,s=1), IN, ReLU, Conv(F=64,k=3,s=1), IN		
	Conv(F=64,k=3,s=1), IN, ReLU, Conv(F=64,k=3,s=1), IN Conv(F=64,k=3,s=1), IN, ReLU, Conv(F=64,k=3,s=1), IN Conv(F=1,k=1,s=1), Tanh		
Output	$E_g(256\times256\times1)$		

TABLE 7. The structure of Flow encoder. *w* and *m* are used for ATBlock to wrap source image.

Input	I_s $(256\times256\times3)$	$P_s(256\times256\times18)$	$P_t(256\times256\times18)$
Intermediate Layers	Concat(I_s, P_s, P_t) IN, ReLU, Conv(F=32,k=4,s=2) IN, ReLU, Conv(F=32,k=3,s=1) IN, ReLU, Conv(F=64,k=4,s=2) IN, ReLU, Conv(F=64,k=3,s=1) IN, ReLU, Conv(F=128,k=4,s=2), IN, ReLU f_2=Conv(F=128,k=3,s=1)		
	IN, ReLU, Conv(F=256,k=4,s=2), IN, ReLU f_1=Conv(F=256,k=3,s=1)		
	IN, ReLU, Conv(F=256,k=4,s=2), IN, ReLU f_0=Conv(F=256,k=3,s=1)		
	f_0=ResBlockUp(f_0)		
	$f_{0'}$=f_0+f_1		
	$f_{0'}$=ResBlockUp($f_{0'}$)		
	w_0= Conv(F=2,k=3,s=1)		Conv(F=1,k=3,s=1) m_0=Sigmoid
	$f_{0''}$=$f_{0'}$+f_2		
	ResBlockUp($f_{0''}$)		
	w_1= Conv(F=2,k=3,s=1)		Conv(F=1,k=3,s=1) m_1=Sigmoid

map as input to using spatial-adaptive normalization methods would lead to a loss of texture details, resulting in unrealistic images.

4) W/O $\mathcal{L}_{sim}$

Training the model without using similarity loss would result in coarse generation of target images with unclear style

TABLE 8. The structure of the generator G of STNet. In SCSBlocks, the content in bracket is used as side branch to affect the main branch.

Input	$I_s(256\times256\times3)$	$P_t(256\times256\times18)$	$E_g(256\times256\times1)$
Intermediate Layers	IN, ReLU, Conv(F=64,k=4,s=2) IN, ReLU f_s^2=Conv(F=64,k=3,s=1) IN, ReLU, Conv(F=128,k=4,s=2) IN, ReLU f_s^1=Conv(F=128,k=3,s=1) IN, ReLU, Conv(F=256,k=4,s=2) IN, ReLU f_s^0=Conv(F=256,k=3,s=1)	IN, ReLU, Conv(F=64,k=4,s=2) IN, ReLU f_p^2Conv(F=64,k=3,s=1) IN, ReLU, Conv(F=128,k=4,s=2) IN, ReLU f_p^1 =Conv(F=128,k=3,s=1) IN, ReLU, Conv(F=256,k=4,s=2) IN, ReLU f_p^0 =Conv(F=256,k=3,s=1)	IN, ReLU, Conv(F=64,k=4,s=2) IN, ReLU f_e^2=Conv(F=64,k=3,s=1) IN, ReLU, Conv(F=128,k=4,s=2) IN, ReLU f_e^1=Conv(F=128,k=3,s=1) IN, ReLU, Conv(F=256,k=4,s=2) IN, ReLU f_e^0=Conv(F=256,k=3,s=1)
	I_{crs}^0=ATBlock(w_0, m_0, f_s^0, f_p^0)		
	I_{crs}^0 =ResBlock(I_{crs}^0)		
	$\hat{I}_{crs}^0$ =SCSBlock(I_{crs}^0, f_s^0, f_p^0, f_e^0)		
	$\hat{I}_{crs}^0$=ResBlockUp ($\hat{I}_{crs}^0$)		
	I_{crs}^1 =ATBlock(w_1, m_1, f_s^1, $\hat{I}_{crs}^0$)		
	I_{crs}^1=ResBlock(I_{crs}^1)		
	$\hat{I}_{crs}^1$ =SCSBlock(I_{crs}^1, f_s^1, f_p^1, f_e^1)		
	$\hat{I}_{crs}^1$=ResBlock($\hat{I}_{crs}^1$)		
	$\hat{I}_{crs}^2$ =SCSBlock($\hat{I}_{crs}^1$, f_s^2, f_p^2, f_e^2)		
	ResBlockUp($\hat{I}_{crs}^2$)		
	Conv(F = 3, K = 3, S = 1), ReLU ReflectionPad(1), Conv(F = 3, K = 3, S = 0), Tanh		
	Conv(F=1,k=1,s=1,), Tanh		
Output	$I_g(256\times256\times3)$		

textures, which affects the accuracy of similarity relationship calculation.

5) W/O COR

In this model, the corresponding relationship between source and target images is not used to extract the style patterns from the source image. Instead, the scaling and bias of the source image are calculated directly to adjust the appearance of the coarse target image.

6) FULL-MODEL

This model includes all components and achieves optimal performance in all quantitative metrics as shown in Table 2. It also achieves the best visual effects, as shown in Figure 5. The proposed Style-Content-Aware Adaptive Normalization method plays a key role, and the pre-transfer of content results in highly effective results. It can be seen from the various indicators in the table that each part of the model is effective.

7) QUALITATIVE COMPARISON

As depicted in Figure 6, the complete framework retains more appearance features, leading to a more similar and realistic image with higher resolution compared to the ground-truth image. The Style-Content-Aware Adaptive Normalization module plays a crucial role in generating images, as removing it results in generated images that do not retain the source image style. Additionally, the Aligned Multi-Scale Content Edge Transfer Network in the first stage helps in maintaining high fidelity of the face features while ensuring the continuity of the clothing. Moreover, since content consistency is well-preactivated, it facilitates better boundary reconstruction between some blurred objects. Finally, the adaptive complementary fusion of the source image style and pose content is dynamically achieved through the correspondence between the source image and target, resolving the issue of smooth appearance in shorts and jeans, hence making the clothing edges more defined.

In addition, this paper compared $\mathcal{L}_{per}$ and $\mathcal{L}_{LPIPS}$ through experiments. The experiments involved using only $\mathcal{L}_{per}$ or $\mathcal{L}_{LPIPS}$ for model training, as well as using both loss functions together for training. The quantitative and qualitative experimental findings showed that using both loss functions together produced optimal results in terms of stylistic features and image quality. Table 5 shows that the combined use of $\mathcal{L}_{per}$ and $\mathcal{L}_{LPIPS}$ produced the best results for all the metrics. In Figure 8, it can be observed that clothing waist information appears in the first row of images, clothing style in the second row, shoe information in the fourth and sixth rows, and clothing texture in the fifth row.

V. CONCLUSION

In this paper, a two-stage person image synthesis model is proposed to handle the challenging pose transfer task. The first stage generates target edge maps that can significantly

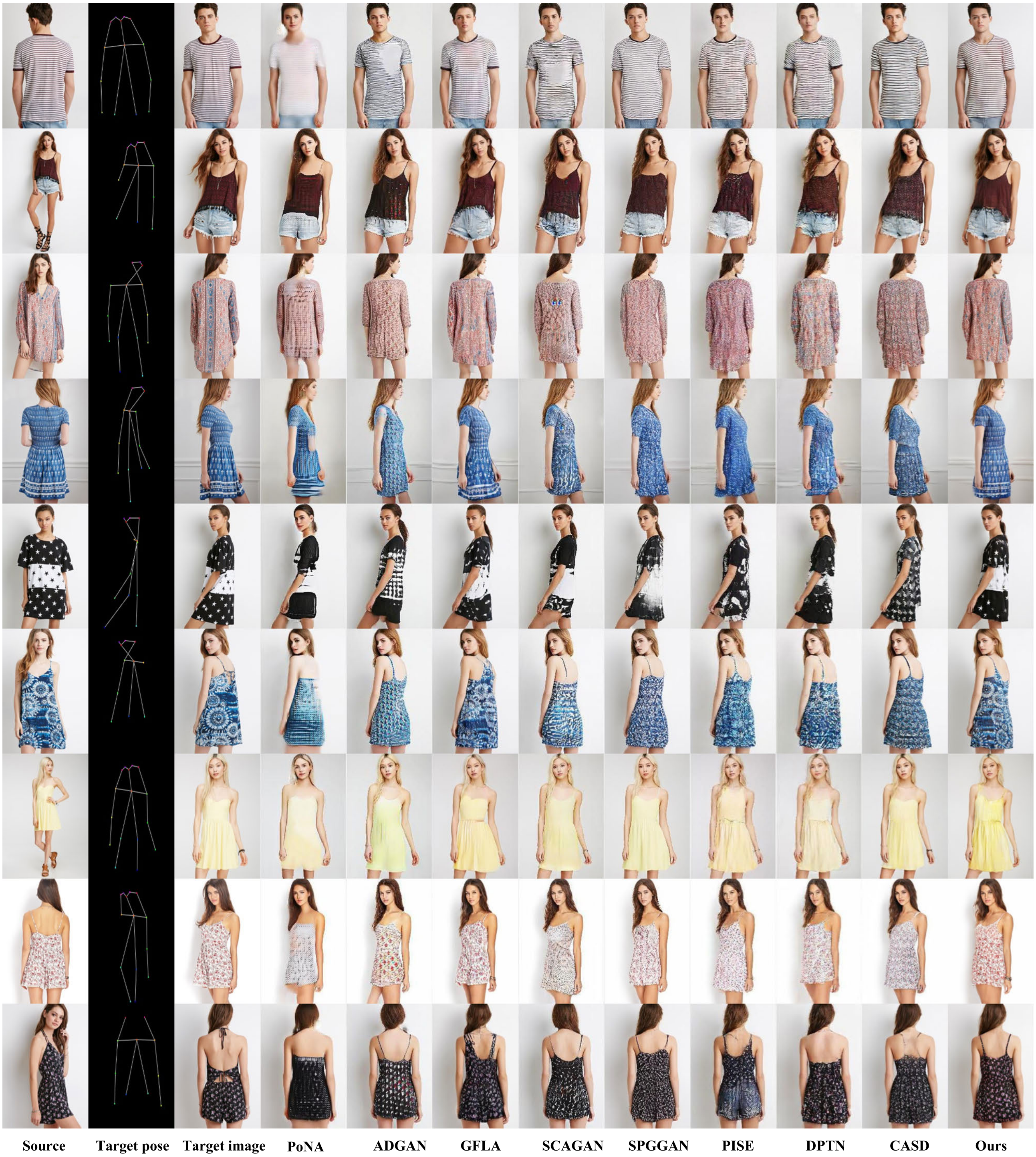

FIGURE 8. Qualitative comparison between our method and other state-of-the-arts on DeepFashion dataset. The target ground truths and the synthesized results from each models are listed in rows.

highlight important content information (pose and texture) to eliminate the difficulty of pose transfer, acting on the appearance translation in the second stage. In the second stage, based on the predicted edge map, a new style-content-aware adaptive normalization method is used to generalize and arrange the style and content information in the same latent space for features to accomplish the realistic character image generation task more effectively. Experiments show that the framework proposed in this paper can significantly improve the supervised and unsupervised perceptual metrics of the existing state-of-the-art, while generating finer-grained features of clothing textures and characters.

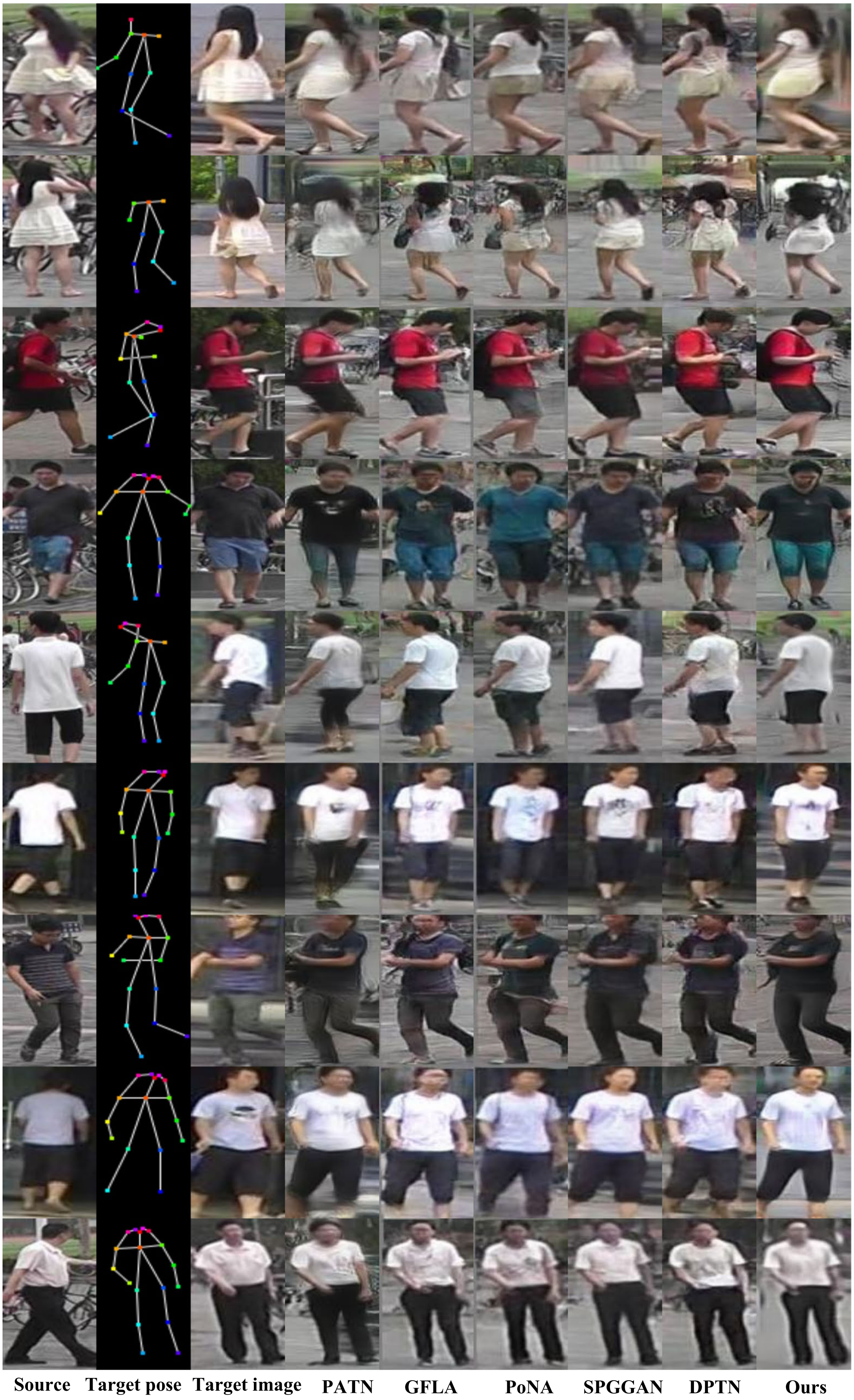

FIGURE 9. Qualitative comparison between our method and other state-of-the-arts on Market-1501 dataset. The target ground truths and the synthesized results from each models are listed in rows.

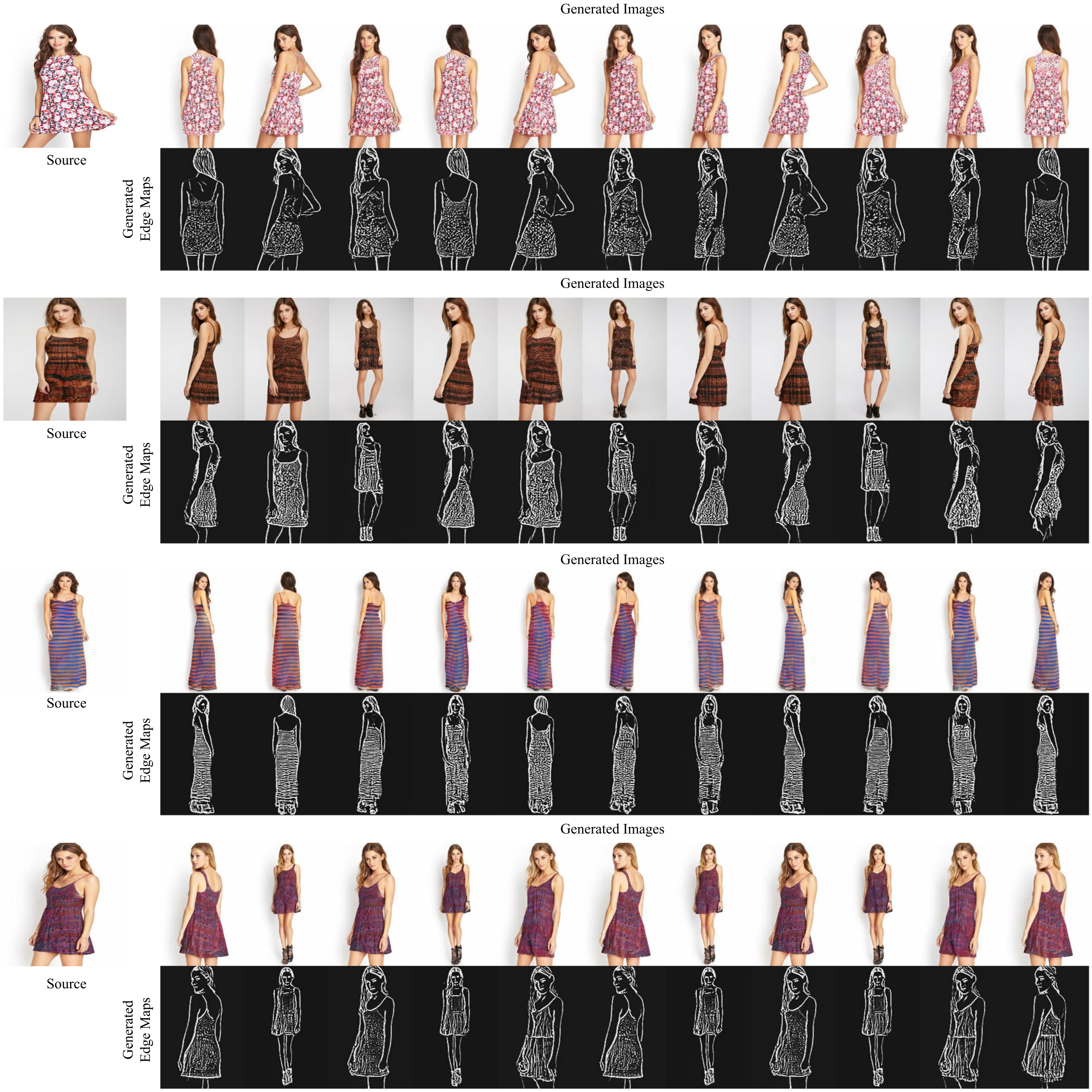

FIGURE 10. Given the source image, our model is able to transfer the pose as required. The synthesized person and visualization of the generated target edge maps are shown.

While the model is able to generate good results, there are still some artifacts in the generated images. Additionally, the model has poor generalization capability and limited generation capability for target poses that have significant pose variations, such as transitioning from standing to sitting. It has been demonstrated through experiments that utilizing real edge map features can effectively generate people images that are indistinguishable from the target images. Therefore, improving the accuracy of the target edge map can effectively enhance the quality of image generation. Furthermore, it has been found that utilizing a distributed balanced dataset can also effectively alleviate these issues and improve the model's generalization ability by modeling the correlation of deformation in adjacent regions. This will be the focus of subsequent work.

APPENDIX I.

In this section, we provide the details of the network structure. Table 6,7,8 are the network structures of the generator G of the AMSNet, Flow encoder, and the generator G of STNet, respectively. F, K, and S respectively represent the output dimension, convolution kernel size and stride. IN represents instance normalization.

APPENDIX II.

In Fig. 8 and 9, we provide additional qualitative comparisons between our approach and other state-of-the-art approaches on the DeepFashion [38] and Market-1501 [39] datasets, respectively.(e.g. PATN [11], PoNA [14], ADGAN [13], GFLA [16], SCAGAN [21], SPGNet [18], PISE [20], DPTN [28], CASD [30]) Results show that our method can generate more consistent appearance and pose with the target.

APPENDIX III.

We also provide more visualization results of the generated edging maps in Fig.10. It is clear that AMSNet can accurately predict the target edging map regardless of diverse pose and viewpoint changes, revealing the effectiveness of the proposed Style-content-aware Adaptive Normalization module.

REFERENCES

[1] 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.

[2] L. Wei, S. Zhang, W. Gao, and Q. Tian, "Person transfer GAN to bridge domain gap for person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 79–88.

[3] Y. Ge, "FD-GAN: Pose-guided feature distilling GAN for robust person re-identification," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 31, 2018, pp. 1–12.

[4] C. Yang, Z. Wang, X. Zhu, C. Huang, J. Shi, and D. Lin, "Pose guided human video generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 201–216.

[5] T.-C. Wang, M.-Y. Liu, A. Tao, G. Liu, J. Kautz, and B. Catanzaro, "Few-shot video-to-video synthesis," 2019, *arXiv:1910.12713*.

[6] W. Liu, Z. Piao, J. Min, W. Luo, L. Ma, and S. Gao, "Liquid warping GAN: A unified framework for human motion imitation, appearance transfer and novel view synthesis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 5903–5912.

[7] D. Morelli, M. Fincato, M. Cornia, F. Landi, F. Cesari, and R. Cucchiara, "Dress code: High-resolution multi-category virtual try-on," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 2230–2234.

[8] N. Pandey and A. Savakis, "Poly-GAN: Multi-conditioned GAN for fashion synthesis," *Neurocomputing*, vol. 414, pp. 356–364, Nov. 2020.

[9] S. Ishikawa and T. Ikenaga, "Image-based virtual try-on system with clothing extraction module that adapts to any posture," *Comput. Graph.*, vol. 106, pp. 161–173, Aug. 2022.

[10] S. Choi, S. Park, M. Lee, and J. Choo, "VITON-HD: High-resolution virtual try-on via misalignment-aware normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14131–14140.

[11] 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. 2342–2351.

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

[13] Y. Men, Y. Mao, Y. Jiang, W. 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. 5083–5092.

[14] K. Li, J. Zhang, Y. Liu, Y. Lai, and Q. Dai, "PoNA: Pose-guided non-local attention for human pose transfer," *IEEE Trans. Image Process.*, vol. 29, pp. 9584–9599, 2020.

[15] 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. 3688–3697.

[16] 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.

[17] Y. Ren, Y. Wu, T. H. Li, S. Liu, and G. Li, "Combining attention with flow for person image synthesis," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 3737–3745.

[18] Z. Lv, X. Li, X. Li, F. Li, T. Lin, D. He, and W. Zuo, "Learning semantic person image generation by region-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10801–10810.

[19] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-GAN for pose-guided person image synthesis," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–11.

[20] J. Zhang, K. Li, Y. Lai, and J. Yang, "PISE: Person image synthesis and editing with decoupled GAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7978–7986.

[21] W. Yu, L. Po, Y. Zhao, J. Xiong, and K. Lau, "Spatial content alignment for pose transfer," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2021, pp. 1–6.

[22] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[23] A. Siarohin, E. Sangineto, S. Lathuilière, 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.

[24] X. Huang and S. Belongie, "Arbitrary style transfer in real-time with adaptive instance normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1510–1519.

[25] J. Zhang, X. Liu, and K. Li, "Human pose transfer by adaptive hierarchical deformation," *Comput. Graph. Forum*, vol. 39, no. 7, pp. 325–337, Oct. 2020.

[26] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5142–5152.

[27] X. Zhou, B. Zhang, T. Zhang, P. Zhang, J. Bao, D. Chen, Z. Zhang, and F. Wen, "CoCosNet v2: Full-resolution correspondence learning for image translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11460–11470.

[28] P. Zhang, L. Yang, J. Lai, and X. Xie, "Exploring dual-task correlation for pose guided person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7703–7712.

[29] S. Liu, H. Guo, K. Zhu, J. Wang, and M. Tang, "Unsupervised cycle-consistent person pose transfer," *Neurocomputing*, vol. 453, pp. 502–511, Sep. 2021.

[30] X. Zhou, M. Yin, X. Chen, L. Sun, C. Gao, and Q. Li, "Cross attention based style distribution for controllable person image synthesis," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 161–178.

[31] N. Dufour, D. Picard, and V. Kalogeiton, "Scam! transferring humans between images with semantic cross attention modulation," in *Proc. Eur. Conf. Comput. Vis.*, Tel Aviv, Israel: Springer, Oct. 2022, pp. 713–729.

[32] I. Goodfellow, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, 2020.

[33] P. Isola, J. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[34] T. Park, M. Liu, T. Wang, and J. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2332–2341.

[35] 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. 4396–4405.

[36] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "SEAN: Image synthesis with semantic region-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5103–5112.

[37] M. He, D. Chen, D. Liao, V. Sander, Pedro, and L. Yuan, "Deep exemplar-based colorization," *ACM Trans. Graph. (TOG)*, vol. 37, no. 4, pp. 1–6, 2018.

[38] 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.

[39] L. Zheng, L. Shen, L. Tian, S. Wang, J. Wang, and Q. Tian, "Scalable person re-identification: A benchmark," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1116–1124.

[40] Z. Zhu, T. Huang, M. Xu, B. Shi, W. Cheng, and X. Bai, "Progressive and aligned pose attention transfer for person image generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 4306–4320, Aug. 2022.

[41] 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 Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 586–595.

[42] 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.

[43] 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," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–38.

[44] 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.

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

WEI WEI received the B.S. and M.S. degrees in computer science and technology from the Shenyang University of Technology, in 2002 and 2005, respectively, and the Ph.D. degree in transportation information engineering and control from Dalian Maritime University, Dalian, China, in 2013. Currently, he is an Assistant Professor with the School of Computer Science and Engineering, Dalian Minzu University, Dalian. His research interests include machine learning and artificial intelligence.

XIA YANG received the B.S. degree in computer science and technology from Tangshan College, Hebei, China, in 2017. She is currently a Postgraduate Student with Dalian Minzu University, Dalian, China. Her main research interests include deep learning and computer vision.

XIAODONG DUAN received the B.S. degree in computer science and technology from Nankai University, Tianjin, China, in 1985, and the M.S. and Ph.D. degrees in applied mathematics and computer software and theory from Northeastern University, Shenyang, China, in 1988 and 2001, respectively. He is the Director of the Dalian Key Laboratory of Digital Technology for National Culture, Dalian Minzu University, Dalian, China. His research interests include pattern recognition and data mining.

CHEN GUO received the B.E. degree in automatic control from Chongqing University, China, in 1982, and the M.Sc. and Ph.D. degrees in marine engineering automation from Dalian Maritime University, Dalian, China, in 1985 and 1991, respectively. He is a Professor and the Director of the Institute of Ship Automation and Simulator, Dalian Maritime University, Dalian, China. His current research interests include intelligent control and marine system automation and simulation.