Computers & Graphics 111 (2023) 24–36

Contents lists available at ScienceDirect

Computers & Graphics

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

Technical Section

A novel Multi-scale architecture driven by decoupled semantic attention transfer for person image generation ☆

Meng Wang [a,b,*], Jiaxing Chen [a,b], Haipeng Liu [a]

[a] The Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan, 650500, China
[b] Yunnan Key Laboratory of Artificial Intelligence, Kunming, Yunnan, 650500, China

ARTICLE INFO

Article history:
Received 28 May 2022
Received in revised form 6 September 2022
Accepted 10 January 2023
Available online 14 January 2023

Keywords:
Person image generation
Pose transfer
Semantics attention
Semantics mapping
GAN

ABSTRACT

Person image generation is a challenging task aimed to transfer the person of the source image from a source pose to a target pose while preserving its style. In this paper, we proposed a Generative Adversarial Network based on Decoupled Semantic Attention Transfer (DSAT-GAN), focusing on that local semantic representations of different image styles and contents cannot be accurately decoupled and transferred. This architecture employs a novel Multi-scale Semantic Mapping Generation Network (Ms-SMGN), driven by two network modules with different semantic attention mechanism, aiming to accurately align and transfer the representations of local semantics at different spatial scales. Then, a channel-separated convolution is applied in the encoding networks instead of the traditional channel fully-connected operation, which reduces computational complexity while realizing channel semantic decoupling. Moreover, a Gram matrix-based global style loss is introduced to further enhance the consistency of high-level semantic between generated and target images. Experiments on Market-1501 and DeepFashion datasets show that DSAT-GAN has superior performance compared with other recent baselines. Additionally, this architecture can be extended to the data enhancement scenes to significantly improve the accuracy of person Re-identification.

1. Introduction

Person image generation aims to generate high-quality person images by converting a person from a source pose to a target pose while preserving appearance details of the source person image. It was first introduced into the field of computer vision by Ma et al. [1]. Since its development, it has been applied into many fields, such as person Re-identification (Re-ID) [2,3], virtual clothes try-on [4,5], video frame prediction [6,7], etc. In recent years, solutions related to this task have gradually evolved from early global prediction [8,9] to popular regional feature mapping [10,11], and the research have made great breakthroughs in source image style preservation and simple pose transfer [12,13]. However, because the human body is a non-rigid deformation structure, there are many difficulties such as how to preserve more realistic image texture details and contour details. The fundamental problems might lie in that the transferring models are prone to lose semantic feature of complex aligned regions when decoupling and coupling image styles and contents in the potential space. Therefore, this paper focuses on improving the analytical and transfer ability of the network to efficiently retain semantic representation alignment for better person detail generation.

Earlier, the literature [8,9,14] used the U-net [15] network structure for global prediction and propagation of the underlying feature related to the human body. Siarohin et al. [9] mainly employed the popular skip connection mechanism to enable the network to broadcast spatial information between encoders and decoders. In literature [14], a U-net based Variational Auto-encoder (VAE) [16] was proposed to encode spatial information for pose transfer. However, the U-net based global prediction and encoding cannot effectively address the difficulty of spatial alignment between source and target poses, which often leads to the lack of key local details in generated person images. Later, Li et al. [17] proposed a flow-based approach to estimate appearance flow to obtain a more dense correspondence between source and target poses, but it cannot reconstruct the occluded region well. In recent years, more and more researchers have devoted themselves to how to model human deformation and local feature transfer. Thus, multi-scale feature mapping and attention mechanisms have been introduced into person image generation [18,19].

Usually, traditional convolutional structures cannot effectively identify complex spatial deformations of the human body [20].

☆ This article was recommended for publication by Dr. R. McDonnell.

* Corresponding author at: The Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan, 650500, China.
E-mail addresses: wangmeng@kmust.edu.cn (M. Wang), jiaxingchen_prc@163.com (J. Chen), 42227324@qq.com (H. Liu).

https://doi.org/10.1016/j.cag.2023.01.003
0097-8493/

So multi-scale feature mapping is only capable of transferring spatial information from the source image to the synthetic image in the person image generation. Siarohin et al. [18] proposed a method for pre-computing transfer information on multi-scale feature mapping. The method performs hard-coded spatial transfer of local information of the human throughout the training process, which may cause problems such as large training parameters, low training efficiency, and misalignment of the human body structure. A pose attention network was used in literature [19] for local feature transfer to adapt to the target body topology. In addition, the approach in literature [21] fused the feature representation of appearance and pose by controlling the attention mechanism based decoder. The above methods have improved model performance, but they only consider the source image and target pose as inputs for encoding and decoding, ignoring the saliency representation and spatial alignment of human details and contours. In the past two years, the generator has connected the expected style with required posture in different layers. For example, HPT [22] and ADGAN [23] have inserted several repetitive modules with the same structure to combine style and posture features. Moreover, GFLA [24] and Intr-Flow [25] estimate the correspondence between source ad target poses to guide the propagation of appearance features. However, when these methods face large deformation, the results may produce obvious artifacts. By introducing the neural texture extraction and distribution operation based on double attention, Ren et al. [26] realized the model to extract clear and expressive neural textures representing the appearance of different semantic entities. Zhou et al. [27] used the designed cross attention matrix to express the dynamic similarity between the target pose and the source semantic style of all semantics, so that the model can route colors and textures from the source image.

Based on the above discussion, it can be seen that in order to generate vivid and realistic person images, following existing problems should be solved: (1) how to accurately preserve and align the semantic representations of different local attributes during the generation; (2) how to efficiently generate ideal target person images through a compact transfer process when there are complex point-of-view changes between source and target poses; (3) how to make reasonable inferences on the occluded regions of the source person and thus complement the spatial details of person during the process of pose transfer.

In response to above problems, we argue that ideally, in order to generate realistic person images, the network structure should be able to describe large non-rigid spatial transfer between source and target poses in an adaptive, flexible and learnable manner, and to decouple and couple the image style and content in potential space without local semantic detail degradation. Therefore, this paper focuses on the Generative Adversarial Networks based on Decoupled Semantic Attention Transfer (DSAT-GAN). The proposed DSAT-GAN consists of a Discriminative Network (DN) and a new Multiscale Semantic Mapping Generative Network (Ms-SMGN). Among them, Ms-SMGN contains a Style Representation Network (SRN), a Pose Transfer Network (PTN), and a Feature Fusion Network (FFN). In order to enable the network to retain the semantic of different attributes adaptively, different semantic attention mechanisms are designed for encoding networks (SRN and PTN), and the channel-separated convolution operation is combined to enable them to perform independent attention representation learning in the potential space, thus reasonably and efficiently complete local semantic coding. Then, the multi-scale coupling semantic representation of PTN is fused and decoded to finally reconstruct the target person image by FFN. In addition, a global style loss ($\mathcal{L}_{gs}$) based on Gram matrix is designed to constrain the high-level semantic feature of generated person images. The experimental results on the Market-1501 dataset [28] and DeepFashion dataset [29] demonstrate the superiority of DSAT-GAN.

The main contributions of this paper are as follows:

- A architecture of person image generation is proposed using a novel Multi-scale Semantic Mapping Generation Network (Ms-SMGN). This network alleviates the difficulty that semantic feature of complex aligned regions is easily lost when image styles and contents are decoupled and coupled, and thus generates more realistic person images.
- The proposed Ms-SMGN contains encoding networks (SRN and PTN) and decoding network (FFN). SRN and PTN can learn multi-scaled representations through semantic attention mechanisms with channel-separated convolutions for both image styles and contents transferring. Then, FFN can fuse these multi-scaled features obtained by PTN and finally generate integrated person details.
- A Gram matrix-based global style loss is further introduced by calculating the distances of labeled parsing mapping between the generated and target images to enhance the consistency of high-level semantic, such as the content and style consistency of each human parts.
- Numerous evaluations were performed on two open datasets Market-1501 and DeepFashion and compared with other recent baselines. The experimental results demonstrate the feasibility and superiority of the proposed architecture DSAT-GAN for the person image generation tasks.

The paper is structured as follows: recent work related to the person image generation is introduced in Section 2; the proposed method is showed specifically in Section 3; the experimental results and extensive analyses on two publicly available datasets are present in Section 4; finally, a summary in presented in Section 5.

2. Related work

2.1. Style parsing and pose encoding

Person semantic parsing provides valid a priori information for person image generation, so it is widely used. For earlier studies [19,30] that did not introduce semantic parsing, the experimental results showed generated images with inaccurate color representation and distorted style structure. To generate target parsing based on the target pose, Dong et al. [10] rendered the texture in source image into a semantic parsing map by learning feature-level mapping. With an independent generator, Han et al. [31] inferred semantic mappings and reconstructed images. Literature [10,31] both utilized person semantic parsing to assist in network training, enhancing the style representation learning capability of the model and further optimizing the semantic representation of network features. It can be seen that semantic parsing mappings are important in guiding pose transfer, especially in refining the style information in generated images.

In addition, the person 2D pose estimation technique contributes significantly to person image generation. This technology is dedicated to locating key points or joints of anatomized human body in the image. XingGAN [32] modeled style and shape information with two novel blocks, and the input of one of blocks was the connection form of the source and target pose. In the same year, in literature [11], a two-stage generative model was designed and the connection form of the source and target pose was decided as a condition to semantically guide the model. In literature [12], an end-to-end framework was proposed with a novel generator-APS at its core, which progressively encoded and coupled the target pose and conditionalized human appearance. In addition, Li et al. [19] used a simplified cascade block where

source and target pose collocation inputs allowed block to utilize pose features more efficiently.

Based on previous work, the author introduces person semantic parsing and pose estimation to Ms-SMGN networks as auxiliary inputs to address the problem of semantic information loss that occurs when image styles and contents are decoupled and coupled. A network that learns all the feature mappings of the source image directly would cause too much computational stress. Therefore, the decoupling and coupling computations are assigned to different encoding networks (SRN and PTN), which effectively reduces the loss of semantic information during computation and improves the learning efficiency of the network model.

2.2. Person image generation and pose transfer

The pose transfer task was introduced by Ma et al. [1]. The model training is divided into two stages. Stage-I: roughly generate an image based on the target pose. Stage-II: refine the style and content, and then generate the person image. Zhu et al. [30] proposed a progressive pose attention migration network, which consisted of multiple cascaded pose attention transfer blocks. Each block can optimize appearance and posture simultaneously by using the attention mechanism. In addition, PG^2 [33] extracted poses from person of the source video and applied the learned poses to appearance mapping to generate target subjects, and introduced a separate convolution pipeline for realistic face synthesis.

The above methods treat images and poses as latent variables and generate person images in the potential space. Recently, Siarohin et al. [18] proposed to apply a set of precomputed affine transformations on a feature map with variable resolution, which requires hard-coded spatial transfer of different body parts of a person. Similarly, Balakrishnan et al. [34] used a directly computed affine transfer. These strategies decompose the human body into several rigid parts, which leads to the loss of global pose information during training. In literature [13], without hard-coded spatial information describing the pose changes, a generator in the designed PoT-GAN network is used to manipulate the corresponding multi-scale feature mapping to allow the model to learn the transfer between two poses. GFLA [24] pretrains a network to estimate the 2D flow and occlusion mask based on source image, source and target poses. Afterwards, it uses them to warp local patches of the source to match the required pose. Without any deformation operation, PISE [35] and SPGNet [36] synthesize the target parsing maps, given the source masks, source poses and target poses as input in the first stage. Then it generates the image with the help of them in the second stage. And BGR block proposed by Tang et al. [37], aims to reason the crossing long-range relations between the source pose and the target pose in a bipartite graph, which mitigates some challenges caused by pose deformation. On the other hand, Li et al. [25] fit a 3D mesh human model onto the 2D image, and train the first stage model to predict the 3D flow, which is employed to warp the source appearance in the second stage. LiquidGAN [38] also adopts the 3D model to guide the geometry deformation within the foreground region. Although geometry-based methods generate realistic texture, they may fail to extract accurate motions, resulting in noticeable artifacts.

In the multi-scale space, Ms-SMGN sets the style features and pose features of person images as style input and content input respectively and adaptively learns information of the non-rigid space transfer related to the target pose. In addition, the designed FFN network is used to further fuse and decode feature representation in different scale spaces, and reconstruct realistic person images under the constraints of losses such as $\mathcal{L}_{gs}$, which effectively alleviates the problem of semantic information loss of different attributes.

3. Proposed DSAT-GAN model

The person image generation refers to the generation of a realistic person image I_g using a generative model $\mathcal{N}$, which follows a target pose image P_t and the person appearance of a source image I_s. The decoupling and coupling the person image style and content has never been well addressed in this task [9,14]. In this paper, we argue that the root cause is that previous studies have neglected the learning of adaptive local detail representation of attributes with independent semantics in a multi-scale space and their accurate alignment, while using existing network models may easily lose or ignore details of the person style. Therefore, the DSAT-GAN is proposed (Fig. 1), which contains a Discriminative Network (DN) and a new Multi-scale Semantic Mapping Generative Network (Ms-SMGN). Robust encoding networks (SRN and PTN) are specially designed for Ms-SMGN. Inspired by previous work [12,19,30], in this paper, unique semantic attention mechanisms for SRN and PTN are constructed, respectively, to learn local feature mappings adaptively at different scales. In this task, FFN is used for the first time to reconstruct the I_g. In addition, a new global style loss ($\mathcal{L}_{gs}$) is constructed to constrain the high-level semantic in the I_g to be consistent with the I_t. Furthermore, since the traditional convolution is to mix all input feature map channels in a fully connected manner as the basic operation, which cannot preserve the independence of decoupling and coupling of image attributes in the I_s. Thus, masked dot product and renormalization based on channel separation are used as extended operations of traditional convolution, so that the feature mapping depends only on the independent semantic representations. This dynamic feature selection mechanism is particularly important for the non-aligned person image generation [11].

Tang et al. [32] proposes two types of cross attention blocks to fuse features from the target pose and source appearance in two directions, repeatedly. Although these models design the fusion block of pose and appearance style, they lack the operation to align source appearance with the target pose. Ms-SMGN inputs the I_s, S_s and P_t independently (instead of the concatenated input of the channel dimension) to SRN and PTN to obtain the latent codes of both. Because we believe that concatenated input and cross-feature fusion can help the model to obtain and process a large amount of feature information in a unit time, but it undoubtedly increases the learning pressure of network and inevitably leads to the loss of important knowledge. We want Ms-SMGN to systematically learn the fusion of style and posture information under controlled conditions. Independent input can help the SRN accurately locate and preserve the source style information, and the PTN can pixel-level align the style information retained by the SRN with the target posture. The structure of the Ms-SMGN is depicted in the gray background of Fig. 1, and its process is expressed as $I_g = \mathcal{N}_g(I_s, S_s, P_t)$. DN is used to evaluate the authenticity of the I_g, and return the error to drive Ms-SMGN update. Based on this, the working details of SRN and PTN are also shown in Fig. 2. In addition, the decoupling calculation of multi-scale local semantic coding is realized by $\mathcal{F} = \mathcal{N}_{\text{SRN}}(S_s, I_s)$, where $\mathcal{F} = \{F^{(l)}\}_{l=1}^{L}$, and L is the maximum number of decomposition layers. Then, through $\mathcal{P} = \mathcal{N}_{\text{PTN}}(\mathcal{F}, P_t)$, the semantic encoding $\mathcal{F}$ is mapped to the feature space $\mathcal{P} = \{P^{(l)}\}_{l=1}^{L}$ through hierarchical coupling.

3.1. Style representation network (SRN)

According to the decoupling of image style and content in the potential space, most of models are difficult to accurately learn the contours of person and garments due to complex foregrounds and backgrounds involved in training samples. Also, a large amount of the high-level semantic (such as clothing texture details, etc.) is easily lost during the convolution process.

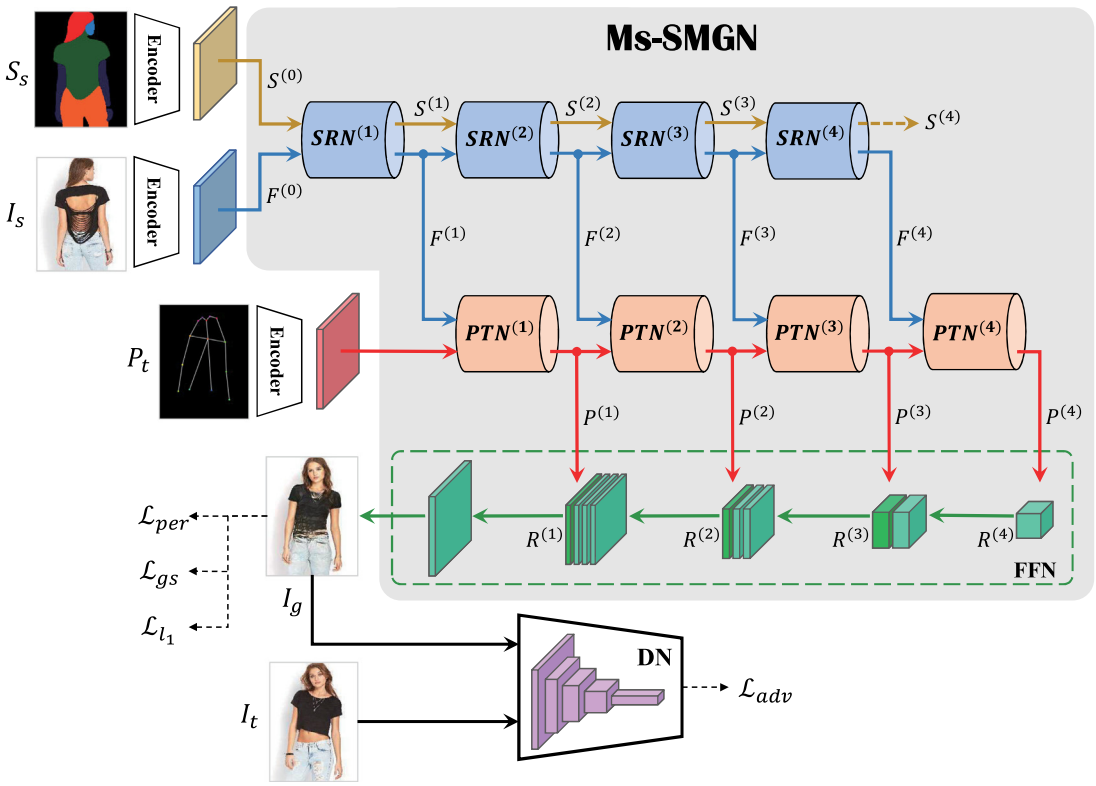


Fig. 1. The DSAT-GAN proposed in this paper. The gray background represents Ms-SMGN, which contains a SRN (blue module) with inputs I_s, S_s, a PTN (red module) with input P_t, and a FFN (green dashed box) with output I_g. The purple part represents DN. To enable network to retain the semantic of different attributes adaptively, different semantic attention mechanisms are designed for encoding networks (SRN and PTN), and the channel-separated convolution operation is combined to enable them to perform independent representation learning in the potential space, thus reasonably and efficiently complete local semantic transferring. Then, the multi-scale coupling semantic representation of PTN is fused and decoded to finally generate the target person image by FFN.

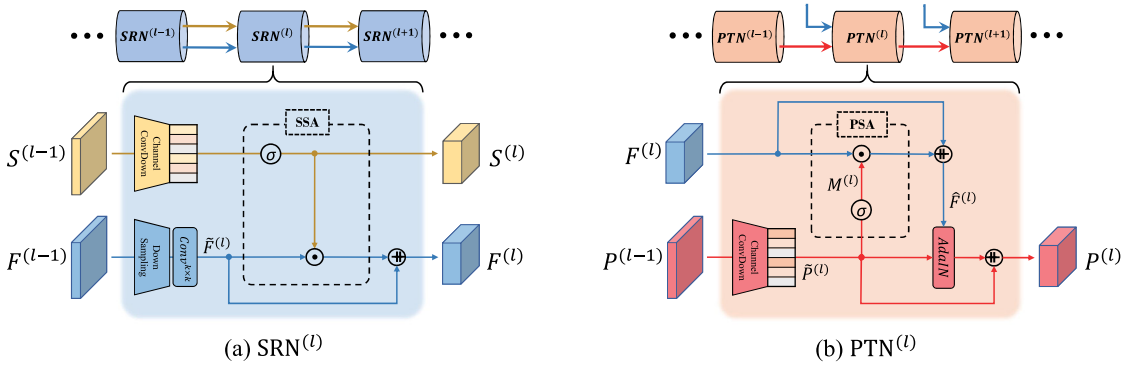


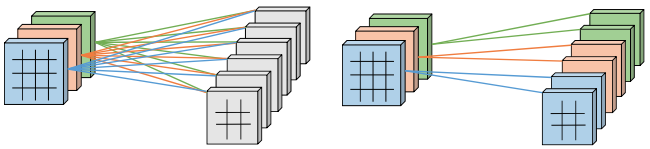
Fig. 2. Detailed structure diagrams of SRN and PTN in the lth level scale. In this figure, $\odot$ denotes element-by-element multiplication, $\oplus$ denotes element summation, and σ denotes the *sigmoid* function. The BN and ReLU layers of the convolutional block are omitted. In figure (a), according to the semantic layout of $S^{(l-1)}$, the information containing human body details and clothing details in $F^{(l-1)}$ is encoded semantically by the designed SSA, and $F^{(l)}$ is then transferred to PTN$^{(l)}$. In figure (b), based on the guidance of $P^{(l-1)}$, PTN$^{(l)}$ adaptively infers the person image region of interest by virtue of the designed PSA and uses AdaIN in the form of affine transformation to couple details between source image appearance and target pose.

Therefore, as shown in Fig. 2(a), SRN introduces S_s as auxiliary input. According to the semantic layout of S_s, the information containing human body details and clothing details in the I_s is encoded semantically by the designed Style Semantic Attention (SSA), and the encoded information $\mathcal{F}$ is then transferred to PTN.

For the SRN$^{(l)}$ of lth ($l = 1, 2, 3, 4$) layer scale, its input is $S^{(l-1)}$ and $F^{(l-1)}$ obtained in the lth layer, and the output is $S^{(l)}$ and $F^{(l)}$. The initial layer is $F^{(0)} = I_s$. Note $F^{(l-1)}$ is mapped to feature space at the lth layer scale through down-sampling and convolution, denoted as:

$$\tilde{F}^{(l)} = h_{\text{conv},\downarrow}(F^{(l-1)}), \tag{1}$$

where $h_{\text{conv},\downarrow}$ represents convolution and down-sampling (the down-sampling block decreases the width and height of the feature map by one time, and the convolution block increases the number of channels of the feature map by one time while increasing the field of perception). Then a SSA mechanism is constructed for the SRN, which motivates the network to dynamically migrate the attention of each element in $\tilde{F}^{(l)}$. Auxiliary input $S_s \in \mathbb{R}^{8\times H\times W}$ and each binary mask matrix $S_s^{(c)} \in \mathbb{R}^{H\times W}$ with $c = 1, 2, \ldots, 8$ describes eight categories, including hair, face, etc. Therefore, the biggest difference from APS [12] is that we decided to give up using the traditional convolution filters. $S^{(l-1)}$ is first down-sampled and channel-separately convoluted



(a) Traditional Convolution (b) Channel − separated Convolution

Fig. 3. Figure (a) shows the fully connected traditional convolution, and figure (b) shows the channel-separated convolution in this paper. Set the input channel number of feature map as C_{input} and for output as C_{output}, and assume C_{output} is n times of C_{input}. Then, the complexity of the channel convolution in (a) is $T_a = C_{\text{input}} \times C_{\text{output}}$, and the complexity in (b) is $T_b = C_{\text{input}} \times n = C_{\text{output}}$. It implies that the operation in (b) greatly reduces the calculation cost and better preserves the channel labeling information of S_s and P_t.

in the multi-scale space as shown in Fig. 3, which can reduce the calculation complexity while performing "channel-to-channel" feature transfer for eight attributes of S_s, thus contributing to the style decoupling of SRN$^{(l)}$. Then, a style semantic attention mask $S^{(l)}$ with a value between 0 and 1 is obtained by the *sigmoid* function to preserve salient information of feature map. $S^{(l)}$ is calculated as follows:

$$S^{(l),(c')} = h_\sigma(h^{(c)}_{\text{conv},\downarrow}(S^{(l-1),(c)})), \tag{2}$$

where $h^{(c)}_{\text{conv},\downarrow}$ represents the convolutional downsampling in channels, for instance the convolution block increases channel number of the cth channel of the input feature map by one time and maps it to the c'th channels of the output feature map while reducing the width and height by one time. Also, h_σ represents the output normalized *sigmoid* function. The normalized fraction corresponding to each pixel in $S^{(l)}$ is denoted as:

$$S^{(l),(c,h,w)} = \frac{1}{1+\exp(X^{(l),(c,h,w)})}, \tag{3}$$

where $S^{(l),(c,h,w)}$ and $X^{(l),(c,h,w)}$ represent the fractions and values corresponding to the elements of $S^{(l)}$ with coordinates of (h, w) in the cth channel, respectively. The convolution operation before the *sigmoid* function can effectively address the feature representation under different receptive fields. Being able to multiply $S^{(l)}$ with $\tilde{F}^{(l)}$ element by element means that it is possible to focus on garment types and attributes through human body layout. To further preserve and emphasize the salient spatial features of $\tilde{F}^{(l)}$, residual connections are also added to the network. Thus, $F^{(l)}$ is calculated as follows:

$$F^{(l)} = S^{(l)} \odot \tilde{F}^{(l)} + \tilde{F}^{(l)}. \tag{4}$$

As mentioned above, SRN extracts the style semantic in source image by scale and passes $\mathcal{F}$ to PTN. Especially for samples with complex backgrounds, the prior information provided by S_s to the network is undoubtedly important. In addition, as the feature space scale decreases, the semantic contained in the feature map changes from being concrete to abstract, and the SSA mechanism of SRN enables the network to transfer the attention of each element feature dynamically at different scales. The style representation learned by the network is thus also richer and more comprehensive, which makes a good pre-processing for the functional implementation of PTN.

3.2. Pose transfer network (PTN)

The complex non-rigid body structure deformation from multiple perspectives may lead to problems such as local deformation of clothing texture and distortion of body structure in the I_g These problems create difficulty for the network model to complete the coupling of image style and content. $\mathcal{F}$ being input to PTN contains both the information of appearance of the source person image and retains the texture, style and type of the garment. We would like to have P_t as another auxiliary input to build a bridge between the pose and image code. As shown in Fig. 2(b), based on the guidance of P_t, PTN adaptively infers the person image region of interest by virtue of the designed Pose Semantic Attention (PSA) and uses Adaptive Instance Normalization (AdaIN) in the form of affine transformation to couple details between the source image appearance and the target pose.

Specifically, the input of PTN$^{(l)}$ at the lth layer is $P^{(l-1)}$ and $F^{(l)}$, and the output is $P^{(l)}$. Similar to S_s, the target pose is $P_t \in \mathbb{R}^{18\times H\times W}$ and each channel is a heat map of a specific joint of a person. Therefore, $P^{(l-1)}$ is mapped to the feature space at the lth level through sub-channel convolution sampling, denoted as:

$$\tilde{P}^{(l),(c')} = h^{(c)}_{\text{conv},\downarrow}(P^{(l-1),(c)}). \tag{5}$$

The pose semantic attention (PSA) is constructed for different scale mappings. A pose semantic attention mask $M^{(l)}$ is obtained with the help of $\tilde{P}^{(l)}$ to enable the network to dynamically select appropriate regions in $F^{(l)}$ for attention migration from low to high scale layers to learn and transfer feature representations, thus effectively preventing local structural distortion of the source image clothing texture. $M^{(l)}$ is calculated as:

$$M^{(l)} = h_\sigma(\tilde{P}^{(l)}). \tag{6}$$

By multiplying $F^{(l)}$ and $M^{(l)}$ element by element, any pixel-level feature can be retained or suppressed. The residual structure can also further emphasize the useful knowledge in $F^{(l)}$ and effectively prevent semantic from being lost in the convolution process. The above process can be expressed as:

$$\hat{F}^{(l)} = M^{(l)} \odot F^{(l)} + F^{(l)}. \tag{7}$$

After that, Adaptive Instance Normalization (AdaIN) is employed for style and pose coupling. Assuming that a content input $D \in \mathbb{R}^{C\times H\times W}$ and a style input $Q \in \mathbb{R}^{C\times H\times W}$ are given, the generated content features are gradually adjusted according to the style of the source image, and it is denoted as $h_{\text{AdaIN}}(D, Q)$, and it is calculated as:

$$h_{\text{AdaIN}}(D, Q) = \gamma(Q)\left(\frac{D-\mu(D)}{\gamma(D)}\right) + \mu(Q), \tag{8}$$

where D denotes the feature representation of the target pose, and Q represents the style representation of the source image. Also, $\gamma(.)$ and $\mu(.)$ represent the independent computation for each channel and each sample in spatial dimension, and them can be expressed as:

$$\mu(D) = \frac{1}{CHW}\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W} D^{(c,h,w)}, \tag{9}$$

$$\gamma(D) = \sqrt{\frac{1}{CHW}\sum_{c=1}^{C}\sum_{h=1}^{H}\sum_{w=1}^{W}(D^{(c,h,w)}-\mu(D))^2 + \varepsilon}, \tag{10}$$

where ε is a constant added to the numerical stability constant of the small batch variance. The procedure for calculating $\gamma(Q)$ and $\mu(Q)$ is same as in Eqs. (9) and (10). In this section, $D = \tilde{P}^{(l)}$ and $Q = \hat{F}^{(l)}$. The calculation of residuals added after AdaIN also helps the network to further retain the salient feature related to the target pose in $F^{(l)}$. Therefore, $P^{(l)}$ is calculated as:

$$P^{(l)} = h_{\text{AdaIN}}(\tilde{P}^{(l)}, \hat{F}^{(l)}) + \tilde{P}^{(l)}. \tag{11}$$

The generation through the AdaIN is fast, and this method can also support the transfer of any style after the model is trained, rather than being limited to a specific style [39], which is especially beneficial for PTN to perform pose transfer on different

person image samples. Moreover, SSA and PSA efficiently help the network learn feature mappings in the potential space from low-scale to high-scale attention migration, complete decoupling of style and content, and accurately and comprehensively retain low-level and high-level semantic representations, creating good preconditions for the AdaIN stage.

3.3. Feature fusion network (FFN)

It is not enough for the generation module of the network model to decode only the semantic feature under the high-scale layer, because they are too abstract. We believe that the key issue to improve the image quality of I_g is whether the foregrounds and backgrounds of the source image is well preserved during the decoding process. Therefore, in the person image generation, the feature fusion method [40] among target detection tasks is borrowed to optimize the reconstructed network. As shown in the green dotted box in Fig. 1, Feature Fusion Network (FFN) integrates and decodes the semantic of different scale spaces of $\mathrm{PTN}^{(l)}$ one by one to generate a realistic person image I_g.

FFN is used to perform the deconvolution upsampling and stitching operations on $P^{(l)}$ $(l = 1, 2, \ldots, L)$ one by one with the following equation:

$$R^{(l)} = \begin{cases} h_{\mathrm{conv},\uparrow}(P^{(l)}), & l = L \\ h_{\mathrm{conv},\uparrow}(h_{\mathrm{concat}}(P^{(l)}, R^{(l+1)})), & l \leq L, \end{cases} \tag{12}$$

where $h_{\mathrm{conv},\uparrow}$ represents the deconvolution upsampling, h_{concat} represents the tensor concatenating in the channel dimension. Finally, the channel information is collated and normalized using a convolutional block with a convolutional kernel of 1, denoted as:

$$I_g = h_\sigma(h^{1\times1}_{\mathrm{conv}}(R^{(l)})), l = 1. \tag{13}$$

The high-level network has a relatively large receptive field and strong semantic representation capability, but the feature map has low resolution, weak geometric representation and lack of spatial geometric feature details. The low-level network has a smaller receptive field, strong geometric representation. Although it has high resolution, its semantic representation capability is weak [13]. FFN integrates the feature under different receptive fields, which improves the learning ability of the generative network at different scales and the image quality of I_g.

3.4. Hybrid loss function

The images with smooth textures and blurred human contours can be generated using earlier global prediction methods based on U-net networks [8,9,14]. In addition to shortcomings of the network itself, their loss functions do not differentiate semantic with different attributes. In this way, even if the semantic feature is preserved during the network learning process, generated person images are not accurately aligned with target images at the pixel level (including style and content alignment). Therefore, the method proposed in this paper uses the constructed global style loss $\mathcal{L}_{gs}$ to focus on constraining the high-level semantic feature of I_g and further constrain the low-level semantic feature and overall feature of I_g in conjunction with other losses.

3.4.1. Loss function of global style

The patch style loss [8] is used in many studies to enhance the texture around the corresponding posture joints between I_g and I_t. However, the spatial misalignment between two leads to different textures and shapes of garments in different poses. The patch style loss does not sufficiently take into account the texture of the main part. To solve the above problem, a semantic global style loss function $\mathcal{L}_{gs}$ is designed to better preserve high-level semantic feature such as texture and style of the garment.

First, the person resolution mapping $S_t \in \mathbb{R}^{8\times H\times W}$ of I_t is introduced. Its structural properties are the same as those of S_s. By multiplying the image with the component mask $S_t^{(c)}$, the style layout under the cth channel label after the person image is decomposed can be obtained. The decomposition process can be described as:

$$\begin{cases} X_t^{(c)} = I_t \odot S_t^{(c)} \\ X_g^{(c)} = I_g \odot S_t^{(c)}, \end{cases} \quad c \in \{1, 2, \ldots, 8\}. \tag{14}$$

The kth $(k = 1, 2, \ldots, K)$ layer activation map of the VGG-19 model [41] pre-trained on ImageNet [42] is used to obtain high-level texture feature of $X^{(c)}$, and it is denoted as $\phi_k(X^{(c)})$. Then, the texture similarity is measured using the Gram matrix [43] (defined as the inner product between $X^{(c)}$ vectorized feature maps), and it is denoted as $\mathscr{F}_{\mathrm{Gram}}(.)$. The above process can be expressed as:

$$\mathscr{F}_{\mathrm{Gram}}(\phi_k(X^{(c)})) = [\phi_k(X^{(c)})][\phi_k(X^{(c)})]^T. \tag{15}$$

Between I_t and I_g, $\mathcal{L}_{gs}$ can be defined as the Euclidean distance of the Gram matrix under the same style labels for both:

$$\mathcal{L}_{gs} = \sum_{k=1}^{K}\sum_{c=1}^{8} \left\| \mathscr{F}_{\mathrm{Gram}}(\phi_k(X_t^{(c)})) - \mathscr{F}_{\mathrm{Gram}}(\phi_k(X_g^{(c)})) \right\|_2^2. \tag{16}$$

In the training, the k corresponding to [*relu*5_2] layer of VGG-19 is used. Because Mahendran et al. [20] revealed that the early layers of the convolutional network tend to obtain low-level feature, such as edge and local shape geometry information; the higher layers tend to obtain high-level feature, such as stylized texture details of the image.

Guided by the semantic layout, $\mathcal{L}_{gs}$ can not only focus on body shape to mitigate the effect of pixel-level context, but also enhance the consistency of high-level semantic between generated and target images.

3.4.2. Other loss functions

In addition to the designed $\mathcal{L}_{gs}$, other loss functions are also important for the network training effect, and they each limit the properties of I_g from different perspectives.

Adversarial loss. Using the Discriminative Network (DN) (the network structure is shown in Fig. 4) to penalize the difference in distribution between the generated (fake) image I_g and the target (real) image I_t. The adversarial loss $\mathcal{L}_{adv}$ is defined as:

$$\mathcal{L}_{adv} = \mathbb{E}[\log(1 - D(I_g))] + \mathbb{E}[\log D(I_t)]. \tag{17}$$

Reconstruction loss. The reconstruction loss $\mathcal{L}_{l_1}$ is used to induce the generated image I_g to be similar to the target image I_t at the pixel level, and it is defined as:

$$\mathcal{L}_{l_1} = \frac{1}{HW} \left\| I_g - I_t \right\|_1, \tag{18}$$

where W and H represent the width and height of the image pixel, and $\|.\|_1$ denotes the $L1$-norm distance for all the pixels.

Perceptual loss. Using $\mathcal{L}_{l_1}$ loss to measure error by pixel is ill-considered because the L1 paradigm inherently leads to blurring effects and possible loss of high-frequency information of the image [15,34]. To address this problem, a perceptual loss $\mathcal{L}_{per}$ is added to improve the reconstructive performance of the model. Perceptual loss was originally introduced for style transfer and image super-resolution, and then it was more popular in image generation [18,30,34]. To generate more textured and realistic person images, $\mathcal{L}_{per}$ is defined as the distance between the feature representation of the generated and real image:

$$\mathcal{L}_{per} = \frac{1}{C_k H_k W_k} \left\| \phi_k(I_g) - \phi_k(I_t) \right\|_1, \tag{19}$$

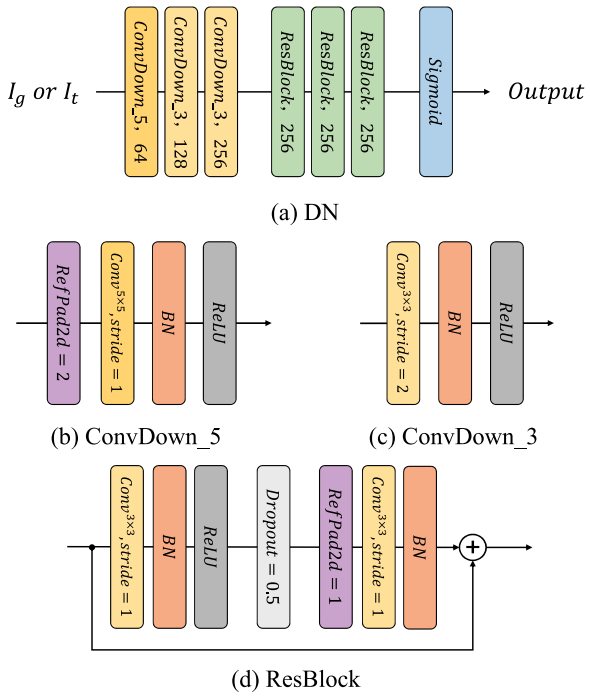


Fig. 4. The specific structure of DN. Figure (a) shows the overall architecture of DN, including three convolution downsampling blocks, three residual blocks and a sigmoid normalization function. The feature representation can be retained more effectively by using a multi-level residual structure. Figure (b) and (c) respectively describe the detailed structure of the convolution downsampling blocks with convolution kernels of 5 and 3, and figure (d) describes the internal structure of the residual blocks.

where H_k, W_k and C_k respectively, represent the width, height and channel width of the kth layer activation map of the VGG-19 model. Unlike $\mathcal{L}_{gs}$, the k corresponding to [$relu$1_2] layer of VGG-19 is used to obtain the underlying feature of I_g.

In summary, the total loss function of the model can be rewritten as

$$\mathcal{L}_{full} = \alpha \mathcal{L}_{l_1} + \arg\min_{G}\max_{D} \lambda \mathcal{L}_{adv} + \eta \mathcal{L}_{per} + \xi \mathcal{L}_{gs}, \tag{20}$$

where α, λ, η and ξ represent the weights corresponding to the four loss functions in $\mathcal{L}_{full}$. The whole training procedure is shown in Algorithm 1. Ms-SMGN pays particular attention to the perception of pose transfer between source and target person images, and accomplish the generation of person images in an end-to-end manner without relying on any hard-coded or pre-computed information. In addition, the model in this paper better learns the semantic knowledge of images with different attributes, which better realize the decoupling and coupling of styles and contents of person image in this task.

4. Experimental results and discussion

In this section, the decoupling and coupling of content and style of source image by Ms-SMGN and the preservation of high-level semantic such as clothing texture details are evaluated and discussed in a multi-scale space. Compared with existing models, verify whether the proposed method addresses the problems such as poor prediction of occluded regions of the source image and loss of the local semantic, and the comparative performance for non-rigid pose transfer of large person details. In this section, we will demonstrate the effectiveness and superiority of Ms-SMGN with various rigorous and consistent experimental settings.

Algorithm 1: Training for the overall framework.

Input : Source person image I_s, source parsing mask S_s, target person image I_t, target pose P_t.
Target parsing mask S_t for $\mathcal{L}_{gs}$.
Freeze VGG-19 for $\mathcal{L}_{per}$ and $\mathcal{L}_{gs}$.
Ensure: DSAT-GAN model.
while *no abnormality* **do**
  Input S_s, I_s and P_t into Ms-SMGN;
  Get $S^{(0)}$, $F^{(0)}$ and $P^{(0)}$ by Encoder;
  for $l = 1; l <= L; l = l + 1$ **do**
    Input $S^{(l-1)}$ and $F^{(l-1)}$ into SRN$^{(l)}$;
    Get $S^{(l)}$ and $F^{(l)}$ by SRN$^{(l)}$ in Eqs. (1)–(4);
    Input $F^{(l)}$ and $P^{(l-1)}$ into PTN$^{(l)}$;
    Get $P^{(l)}$ by PTN$^{(l)}$ in Eqs. (5)–(11);
    Input $P^{(l)}$ into FFN;
  end
  Get fake person image I_g from FFN in Eqs. (12)–(13);
  Train discriminator with real person image I_t;
  Train discriminator with real person image I_g;
  Update DN with $\mathcal{L}_{adv}$ in Eq. (17);
  Update Ms-SMGN with $\mathcal{L}_{full}$ in Eq. (20).
end

4.1. Experimental setup

4.1.1. Datasets and evaluation metrics

The method proposed in this paper is validated on two benchmarks, including datasets of Market-1501 [28] and DeepFashion [29]. **The Market-1501** contains 12,936 images for training and 19,732 images for testing. These images capture 1501 different people from six different surveillance cameras. All images in this dataset have 128 × 64 pixels, and these images include different viewpoints, poses, backgrounds, and illumination. **The DeepFashion (*In-shop Clothes Retrieval Benchmark*)** contains 52,712 images of in-store clothing. All images of people are of 256 × 256 pixels and they contains approximately 200,000 cross or cross-scale person pairs and 7982 clothing images.

The evaluation of pose based person image generation tasks is still an open problem. The Structural Similarity (SSIM) [44], Initial Score (IS) [45], Fréchet Initial Distance (FID) [46], and Learning Perceptual Image Patch Similarity (LPIPS) [47] are used for qualitative experiments and quantitative evaluation. On the Market-1501 dataset, pose transfer is quite challenging due to the complex backgrounds. In order to reduce the influence of background, variants of SSIM and IS, named Mask-SSIM and Mask-IS, were proposed in literature [1]. These two evaluation metrics are purposefully applied to the model evaluation experiments on the Market-1501 dataset.

Specifically, IS calculates the statistics of the generated images by KL divergence, which can be expressed as:

$$\mathscr{F}_{\mathrm{IS}}(G) = \exp(\mathbb{E}_{x\sim p_g} D_{\mathrm{KL}}(p(y|x) \parallel p(y))), \tag{21}$$

where p_g represents the distribution of the image x. $D_{KL}(.)$ represents the KL divergence, and $p(y|x)$ and $p(y)$ represent the conditional distribution and the edge distribution, respectively. SSIM evaluates the similarity of two images (I_g and I_t) in terms of luminance, contrast, and structure. The function is shown below.

$$\mathscr{F}_{\mathrm{SSIM}}(G) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C_1)(\sigma_x^2 + \sigma_y^2 + C_2)}, \tag{22}$$

where x and y represent the real and pseudo images, and μ and σ represent the mean and variance, respectively. The constants C_1

Table 1
The influence of different combinations of weights on experimental results.

Datasets	Weights				Evaluation metrics		
	α	λ	η	ξ	FID↓	SSIM↑	Mask-SSIM↑
Market-1501 [28]	20	20	20	60	13.977	0.304	0.787
	20	10	10	80	14.057	0.325	0.792
	15	10	5	100	13.903	0.311	0.807
	10	5	5	100	**13.267**	**0.334**	**0.825**
	5	10	10	100	13.843	0.316	0.810
DeepFashion [29]	20	20	20	60	19.102	0.744	–
	20	10	10	80	18.618	0.785	–
	10	15	15	80	18.505	0.796	–
	20	15	5	80	**18.396**	**0.804**	–
	30	20	10	60	18.800	0.751	–

and C_2 are introduced to avoid the instability once $\mu_x^2+\mu_y^2$ is very close to zero. LPIPS calculates the reconstruction error between the generated image I_g and the real image I_t in the perceptual domain. Accordingly, the LPIPS evaluation process is illuminated in the following equation:

$$\mathscr{F}_{\text{LPIPS}}(I_g, I_t) = \sum_l \frac{1}{H_l W_l} \sum_{h,w} \|\omega_l \odot (y^{(l),(h,w)} - y_0^{(l),(h,w)})\|_2^2, \tag{23}$$

where $y^{(l)}$ and $y_0^{(l)}$ represents the feature stack of the extracted lth layer, and $y^{(l)}, y_0^{(l)} \in R^{H_l \times W_l \times C_l}$. The proportion of activated channels is $\omega_l \in R^{C_l}$. The distance d(., .) of Wasserstein-2 is calculated between the mean of Gaussian distribution (μ, σ) and the mean of Gaussian distribution (μ_w, σ_w) by FID, where (μ, σ) and (μ_w, σ_w) are respectively obtained from I_t and I_g. This can be denoted as:

$$\begin{aligned}\mathscr{F}_{\text{FID}}(I_t, I_g) &= \text{d}((\mu, \sigma), (\mu_w, \sigma_w)) \\ &= \|\mu - \mu_w\|_2^2 + \text{Tr}(\sigma + \sigma_w - 2\sqrt{\sigma\sigma_w}).\end{aligned} \tag{24}$$

Then, with the help of the above two publicly available datasets and six evaluation metrics, we believe that the reliability and superiority of the proposed DSAT-GAN can be verified through a series of qualitative and quantitative tests.

4.1.2. Experimental configuration details

Before training, $S_s \in \mathbb{R}^{8\times H\times W}$ is extracted from the source image I_s with the help of a Part Grouping Network (PGN) [48]. Eight categories (including hair, face, top, upper body skin, dress, pants, legs and background) are described by each binary mask matrix $S_s^{(c)} \in \mathbb{R}^{H\times W}$ with $c = 1, 2, \ldots, 8$. Both S_t and S_s are obtained with the same method and attributes. Furthermore, based on existing literature [11,20], a Human Pose Estimation (HPE) [49] is used to extract 18 human joints from the source image I_s and they are defined as $P_t \in \mathbb{R}^{18\times H\times W}$. Thus, the pose labels are represented in the form of a matrix of 18 channels, each channel is the heat map of a specific human joint.

The methods proposed in this paper are all implemented using the popular Pytorch framework. For DeepFashion, the image are of 256 × 256 pixels. For Market-1501, the pixels of samples are artificially changes as 128 × 128. It is worth noting that the maximum number of decomposition layers for network structure applied in this paper is $L = 4$. In order to adapt network to the receptive field of feature map at different scales, different convolutional kernel sizes are set for the convolutional blocks contained in SRN, PTN and FFN at different scales, respectively, i.e. $K^{(l)} \triangleq \{k^{(1)} = 7, k^{(2)} = 5, k^{(3)} = 3, k^{(4)} = 3\}_{l=1}^4$. l represents the lth layer scale. The channel dimensions corresponding to these four scale spaces are $C^{(l)} \triangleq \{18 \times 2^{l-1}\}_{l=1}^4$; the height and width dimensions are $H^{(l)} \times W^{(l)} \triangleq \{h \times w \times 2^{-2l}\}_{l=1}^4$, where the value of $h \times w$ is kept consistent with sample image pixels in dataset. Referring to recent work [12,19,30], the Adam optimizer [12] is used to train the designed method, $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The initial learning rate of the model is 2×10^{-4} and the ratio of the learning rate of generation network to that of DN is 1×10^{-1}. After it is trained for 200 epochs, the learning rate decays linearly to 0. The batchsize is set as: the Market-1501 is 16 and DeepFashion is 8. In addition, we show the influence of different combinations of loss weights on the experimental results in Table 1. The weight parameters in the table are all integer type, and the weight sum O_w that is expressed as $O_w = \sum(\alpha, \lambda, \eta, \xi)$ is set as a fixed value ($O_w = 120$ in this paper). Our scheme selects the optimal combination of weights. So, the normalized weights $[\alpha, \lambda, \eta, \xi]$ of loss functions $[\mathcal{L}_{l_1}, \mathcal{L}_{adv}, \mathcal{L}_{per}, \mathcal{L}_{gs}]$ are set to [0.083, 0.042, 0.042, 0.833] for the Market-1501 dataset, and set to [0.166, 0.125, 0.042, 0.667] for the DeepFashion dataset to achieve the best training effect.

4.2. Comparative assessment with recent work

The proposed ground method is compared with recent state-of-the-art methods, including PATN [30], APS [12], PoNA [19], ADGAN [23], PISE [35], SPGNet [36] and PoT-GAN [13]. For a fair comparison, the above pre-trained models are downloaded and the quantitative performance is re-evaluated on the test set.

4.2.1. Qualitative results

Firstly, DSAT-GAN is compared qualitatively with recent methods. It can be seen in Fig. 5 that the most realistic person images can be generated through DSAT-GAN. For some test images from the Market1501 and DeepFashion datasets, with other methods, there are problems such as image blurring and loss of texture detail. It may because the geometric difference information between the source and target pose are not fully utilized. Specifically, for the low-resolution images in Market-1501, as shown in Fig. 5, the network in this paper can obtain the best agreement of appearance details with the source images, for example: the book bag and bag strap are preserved in the first and third rows, and the handkerchief is preserved in the fourth row. The method in this paper also can accurately generate the desired target pose, for instance the leg details in the first row.

For the high-resolution image in DeepFashion, as shown in Fig. 6, by using the method in this paper, more details of the garment texture can be obtained, e.g. the tattoo in the first row and shoulder straps in the third row are accurately preserved. In addition, it succeeds in preserving better body shapes, i.e., arms and legs. And DSAT-GAN also makes reasonable predictions for the obscured parts of the source image, such as the face in the first row and the front style of clothes in the fourth row.

To further validate the performance of proposed model, it was also compared separately with models of other attention mechanisms, several person images in different poses under the same sample are obtained, and the results are shown in Fig. 7. The results in PATN [30] and APS [12] show overly smooth clothes and distorted faces. Some of the results in PoNA [19] also show

Table 2
Quantitative results on Market-1501 and DeepFashion datasets.

	Market-1501 [28]						DeepFashion [29]			
Model	IS↑	Mask-IS↑	SSIM↑	Mask-SSIM↑	FID↓	LPIPS↓	IS↑	SSIM↑	FID↓	LPIPS↓
Real Data	3.903	3.732	1.000	1.000	0.000	0.0000	3.996	1.000	0.000	0.0000
PATN [30]	3.379	3.587	0.297	0.818	20.073	0.3049	3.209	0.773	20.739	0.2533
APS [12]	3.478	3.533	0.318	0.783	21.976	0.2986	3.325	0.686	22.490	0.2641
PoNA [19]	3.416	**3.705**	0.313	0.819	18.712	0.3008	3.365	0.775	18.529	0.2238
ADGAN [23]	3.496	3.692	0.311	0.814	19.584	0.2755	3.480	0.771	18.547	0.2330
PISE [35]	3.581	3.615	0.315	0.804	15.933	0.2551	3.596	0.780	19.640	0.2254
SPGNet [36]	3.622	3.618	0.323	0.811	16.359	0.2681	3.498	0.763	21.550	0.2212
PoT-GAN [13]	3.673	3.605	0.309	0.800	14.528	0.2642	3.582	0.795	20.148	**0.2052**
Ours	**3.781**	3.626	**0.334**	**0.825**	**13.267**	**0.2418**	**3.673**	**0.804**	**18.396**	0.2117

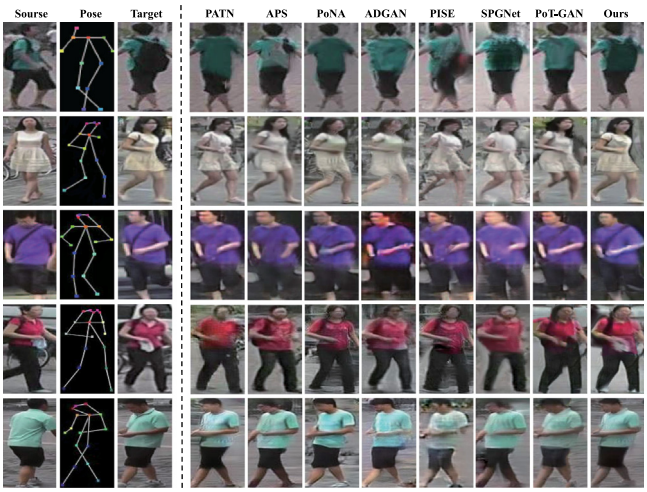


Fig. 5. Qualitative results on the Market-1501 dataset. The compared methods include PATN [30], APS [12], PoNA [19], ADGAN [23], PISE [35], SPGNet [36] and PoT-GAN [13].

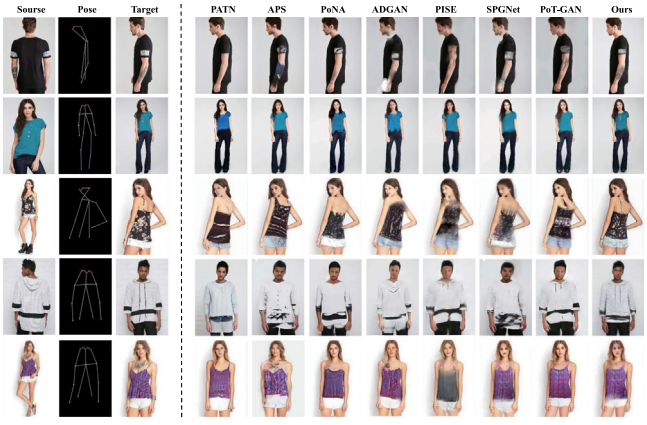


Fig. 6. Qualitative results on the DeepFashion dataset. The compared methods include PATN [30], APS [12], PoNA [19], ADGAN [23], PISE [35], SPGNet [36] and PoT-GAN [13].

distorted backgrounds and figure images, with inconsistencies in texture and style between source and generated images. The occurrence of the above problems is attributed to the fact that none of the earlier studies [12,19,30] introduced human semantic parsing as an auxiliary input, resulting in generated images with inaccurate color representation and distorted style structure. It also ignores the fact that traditional convolution is not always effective for decoupling and coupling of image attributes in person image generation. Compared with previous studies, the network in this paper can generate more realistic and natural person

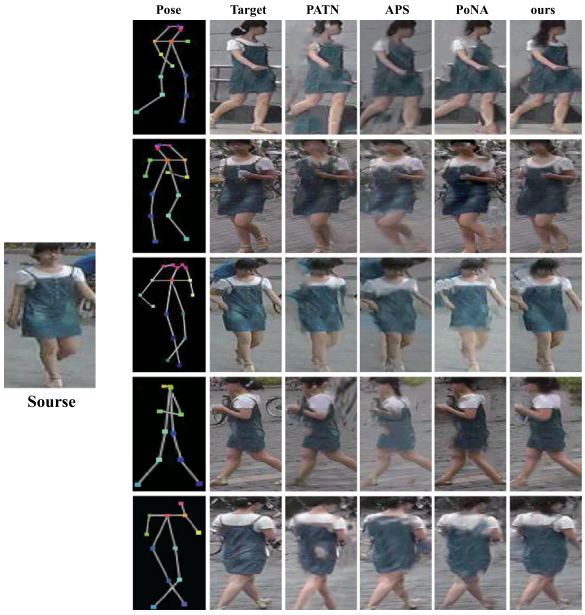


Fig. 7. Qualitative results for the same sample in different poses on the Market-1501 dataset. The compared methods include PATN [30], APS [12] and PoNA [19].

images while maintaining the consistency of the style and texture of the generated image and the source image.

4.2.2. Quantitative results

The results of the quantitative comparison between DSAT-GAN and recent methods are shown in Table 2. It is clear that it outperforms these methods on most metrics. Intuitively, for Market-1501, the proposed method can exhibit the most natural pose and the images are more realistic. This indicates that SRN and PTN are able to learn complex feature mappings with channel-separated convolutions and semantic attention mechanisms to decouple and couple image style and content, thus the generated images have better source image appearance and identity preservation. For DeepFashion, the best results can be obtained on IS, SSIM and FID. Although the score of LPIPS is slightly higher than that in the PoT-GAN [13], more details of clothing textures can be captured while preserving body shape by using the proposed method. The experimental results show that the method achieves a good balance between realism and similarity of images. This is because the global style loss $\mathcal{L}_{gs}$ constrains the high-level semantic feature of the generated person image, which strengthens the consistency of the high-level feature such as the content and style of each component between generated and target image. In addition, the method in this paper achieves the lowest FID score on the DeepFashion and outperforms other methods on the Market-1501. Presumably, because the model is

more reliable in discriminating complex backgrounds, and part of the reason is the semantic attention mechanisms it contains. On the DeepFashion, DSAT-GAN performs quite well compared with the latest methods. Because the DeepFashion consists of a large number of samples with full-body and half-body images of people sharing the same appearance. When a half-body image is transformed to a full-body image, the extracted appearance is usually insufficient to infer a full-body person image due to the lack of information about the lower half of the body. Experiments show that the method in this paper successfully solves the problem. In other words, DSAT-GAN is both a "manipulator and a "creator due to the design and introduction of Ms-SMGN.

4.3. Module performance analysis

4.3.1. Ablation experiments

In order to analyze each component of the overall framework proposed in this paper, the following experiments with different configurations are designed. All applied method follow the same discriminator network structure, which is explained below:

- **Baseline**: Baseline framework is an auto-encoder–decoder with convolution networks applied to replace Ms-SMGN. Also, I_s and P_t are directly spliced together as inputs in the model.
- **w/o SRN**: To verify the validity of SRN, input S_s is dropped, the SRN module is eliminated, and I_s and P_t are directly used as the input of PTN for training.
- **w/o PTN**: To verify PTN, PTNs are all replaced with auto-encoder convolutional network for training.
- **w/o FFN**: To illustrate the importance of integrating semantic in different scale spaces, the convolutional upsampling block and fully connected layers are used to decode $P^{(4)}$ for the highest scale layer only and the image quality of I_g is observed.
- **w/o** $\mathcal{L}_{gs}$: Attention perception loss [21] is used to replace the global style loss in the model, and the training effect of the model is observed.
- **Full model**: The entire framework proposed in this paper is evaluated in these configurations.

The corresponding images generated according to the different configurations are shown in Figs. 8 and 9. The full model can effectively use the appearance and pose features of the source images to guide the pose transfer. Removing each component of Ms-SMGN results in performance degradation, a degree of color distortion and unreal detail.

Specifically, compared with the full model, there are always problems such as distorted characters, distorted styling and smooth clothing texture by using baseline. The results of w/o SRN show that the SRN module can help the network to distinguish the foreground and background of source image and extract the style features of the person and clothes accurately. In addition, the results of w/o PTN show that PTN can effectively prevent the distortion of person and the distribution of person styles after pose transfer is more reasonable. The results of w/o FFN are even worse. The entire image is inaccurate and grainy in style color, with blurred outlines of people's faces and clothes. This indicates that the semantic contained in the feature map of high-scale layer is too abstract to support the network decoding to generate normal person images. Finally, it can be clearly seen that because the results using w/o $\mathcal{L}_{gs}$ show smooth clothes textures and hair textures of people, $\mathcal{L}_{gs}$ can constrain the high-level feature of I_g and enrich the texture details and character details of images.

Table 3 shows the quantitative results of the ablation experiments. Compared with the full model, the evaluation results of w/o SRN and w/o PTN show that the model performance has

Fig. 8. Qualitative results of ablation experiments on the DeepFashion dataset.

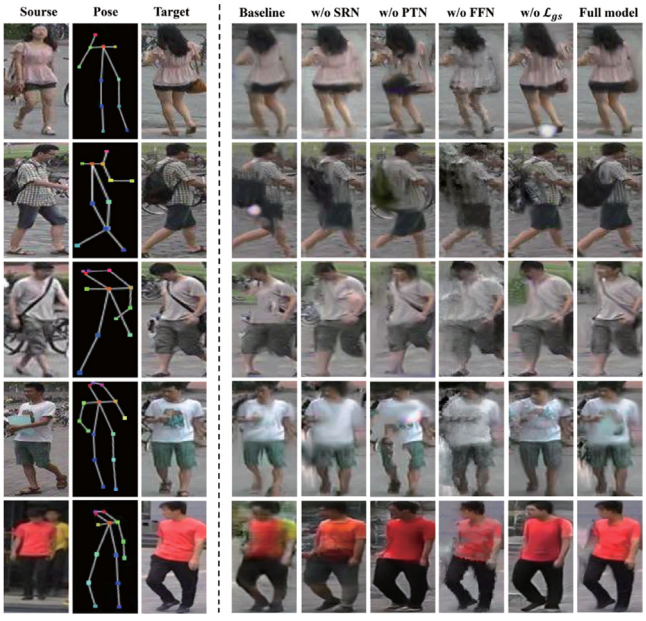


Fig. 9. Qualitative results of ablation experiments on the Market-1501 dataset.

declined, especially SSIM, Mask-SSIM and FID. This shows that the designed SRN and PTN effectively improve the semantic expression ability and reconstruction ability of the model. Moreover, the SSIM and mask SSIM of w/o PTN are lower than those of other models, and the FID of w/o SRN is higher than that of other models. This further shows that PTN has made more contributions to the authenticity and beauty of the generated images, while SRN helps the network to accurately learn the human body contour and clothing texture details. In addition, combining the quantitative and qualitative results of w/o FFN and w/o $\mathcal{L}_{gs}$, we can determine that FFN consolidates the semantic integration ability of the network, and $\mathcal{L}_{gs}$ also makes the model performance better with feature constraints.

In conclusion, the optimal results can be obtained by using Full model on Market-1501 and DeepFashion datasets. It implies that the introduction of semantic attention mechanisms and multi-scale style loss $\mathcal{L}_{gs}$ can significantly improve the visual quality and

Table 3
Quantitative results of ablation experiment.

	Market-1501 [28]						DeepFashion [29]			
Model	IS↑	Mask-IS↑	SSIM↑	Mask-SSIM↑	FID↓	LPIPS↓	IS↑	SSIM↑	FID↓	LPIPS↓
Real Data	3.903	3.732	1.000	1.000	0.000	0.0000	3.996	1.000	0.000	0.0000
Baseline	3.187	3.499	0.301	0.714	47.035	0.3211	3.272	0.716	43.932	0.2133
w/o SRN	3.410	3.615	0.327	0.793	26.829	0.2987	3.318	0.781	30.408	0.2125
w/o PTN	3.580	3.581	0.321	0.740	25.931	0.3075	3.394	0.737	29.311	0.2170
w/o FFN	3.356	3.544	0.329	0.744	26.376	0.3050	3.316	0.789	27.861	0.2121
w/o $\mathcal{L}_{gs}$	3.632	3.572	0.329	0.802	20.808	0.2720	3.453	0.788	21.330	0.2154
Full model	**3.781**	**3.626**	**0.334**	**0.825**	**13.267**	**0.2418**	**3.673**	**0.804**	**18.396**	**0.2117**

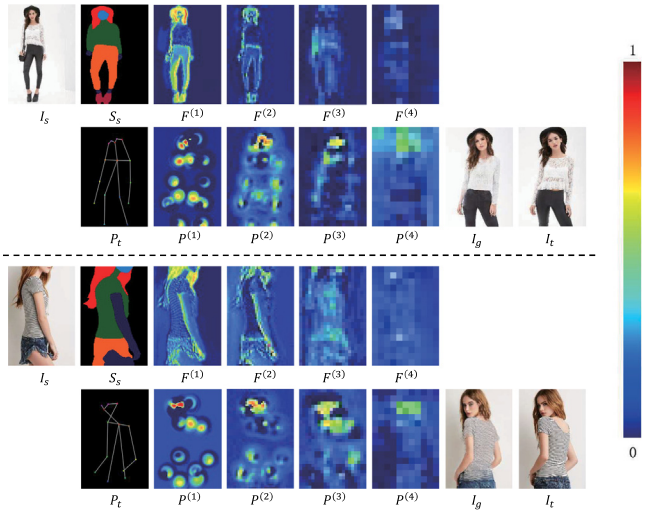


Fig. 10. Semantic attention migration heat map on the DeepFashion dataset. The heat maps at different scales are scaled to the same size for easier observation.

quantization results of the output images, and further demonstrates effectiveness of the semantic attention mechanisms in decoupling and coupling style and content.

4.3.2. Semantic attention analysis

In addition, to more intuitively demonstrate impact of the semantic attention mechanisms at different scales, both $F^{(l)}$ and $P^{(l)}$ with $l = 1, 2, 3, 4$ are visualized as heat maps at each scale using the Deepfashion dataset (Fig. 10).

Intuitively, the feature selection (attention mask) varies from coarse to fine in the different scale spaces. Feature selection focuses on the background and large regions in the lower scale layers, while it pays more attention to the foreground and small regions in the higher scale layers. Specifically, feature regions where $F^{(l)}$ is aligned with S_s have greater weight in the low-scale layer, but attention gradually migrates and focuses on the important parts such as head, hands, and crotch in high-scale layer. This suggests that SSA helps the network dynamically select features to decouple style and content from coarse to fine, from concrete to abstract. Meanwhile, as the scale layer rises, feature regions where $P^{(l)}$ with $l = 1, 2, 3, 4$ is aligned with P_t are gradually coupled with the style semantic in $F^{(l)}$, and the feature selection gradually migrates from initial 18 joint to important parts of the body, such as head and hands. This shows that PSA enables the feature selection and feature deformation to be realized together with different importance at different scale layers.

4.4. Application to person re-ID

Benefiting from the development of deep learning, the person Re-identification (Re-ID) has made great progress in recent years [50–52]. In this paper, the model of person image generation can obtain images of the same person in different poses, which helps to expand the relevant dataset of Re-ID to solve the difficulty of lacking training data.

To illustrate the performance of the proposed model, Market-1501 dataset [28] in the field of Re-ID and two-advanced backbone networks (ResNet-50 [53] and Inception-v2 [54]) are used as testbeds, and the mean Average Precision (mAP) is used as a metric to test the performance of Re-ID. First, a simplified training set D_ρ is obtained by selecting a portion ρ of the Market-1501 training set to realize the random extraction of images in the training set. Note that at least one relevant pose image is kept for each pedestrian sample to ensure the diversity of pedestrian samples. Then, using person image generation models (including PATN [30], APS [12], PoNA [19], ADGAN [23], PISE [35], SPGNet [36], PoT-GAN [13] and our DSAT-GAN), new person images are generated based on the pose of the preserved and missing images. Finally, the simplified training set D_ρ is combined with all the generated images to obtain the enhanced training set. Four simplified training sets and four enhanced training sets are obtained by selecting 20% $\sim$ 80% of the Market-1501 training set as ρ at 20% intervals. The training and testing protocols used [38] are followed to train Resnet-50 and Inception-v2.

Table 4 shows the Re-ID test results using the simplified training set (denoted "None") and the enhanced training sets corresponding to the different generation models. For a fair comparison, the images for all methods are generated with same settings (source image and target pose). With the same model and parameters, Re-ID performance relies on the realism of the generated images and the texture consistency of the same identity. By adding more generated images with different poses, the performance of Re-ID is improved. Compared with other existing methods for person image generation in the table, the optimal result can be obtained through the method in this paper. The most accurate Re-ID estimation is obtained with enhanced dataset using the method of this paper, indicating that this method generates more realistic images with consistent textures. This is because the designed robust encoding networks (SRN and PTN) decouple and couple the style and content of the source image accurately and efficiently, and then FFN fuses the feature representations under different receptive fields, which greatly improves the learning ability and the reconstruction ability of the target image of generation network.

5. Conclusion

In this paper, DSAT-GAN is proposed to better solve the problem that semantic information of different attributes is easily lost when styles and contents of the image are decoupled and coupled. It contains DN and a novel transferring framework Ms-SMGN with a Style Representation Network (SRN), a Pose Transfer Network (PTN), and a Feature Fusion Network (FFN). To enable the networks to retain semantic with different attributes adaptively, two different semantic attention mechanisms (SSA and

Table 4
Comparison results of mAP in Re-ID. (.)% represents the proportion of part ρ.

	ResNet-50 [53]					Inception-v2 [54]				
Aug. Model	20%	40%	60%	80%	100%	20%	40%	60%	80%	100%
None	28.7	53.0	59.4	64.1	68.4	15.9	33.4	41.8	50.1	53.3
PATN [30]	51.9	53.1	62.0	65.8	68.4	42.5	44.7	44.5	49.6	53.3
APS [12]	54.3	58.5	60.2	64.3	68.4	43.3	46.1	46.8	48.8	53.3
PoNA [19]	52.0	58.5	62.6	66.4	68.4	42.6	46.5	47.0	51.0	53.3
ADGAN [23]	52.6	57.3	63.0	65.5	68.4	43.3	45.8	47.0	50.4	53.3
PISE [35]	53.5	59.0	62.7	65.6	68.4	44.2	44.8	45.3	50.9	53.3
SPGNet [36]	55.2	57.8	**65.1**	66.2	68.4	44.0	45.8	47.5	49.8	53.3
PoT-GAN [13]	54.3	58.1	63.4	66.7	68.4	43.5	46.2	48.4	51.2	53.3
Ours	**57.6**	**60.1**	64.4	**67.2**	68.4	**45.6**	**48.3**	**49.2**	**52.1**	53.3

PSA) are designed for the encoding networks (SRN and PTN), respectively. Combined with channel-separated convolutions, they can perform independent attention representation learning in potential space, so that local semantic coding can be done reasonably and efficiently. Then, the multiscale decoupled semantic representation of PTN is fused and decoded by FFN to finally generate target person image. In addition, a global style loss $\mathcal{L}_{gs}$ based on the Gram matrix is also designed, and it can constrain the low-level and high-level semantic features of the generated person images combined with other loss functions. Experiments on the Market-1501 and DeepFashion datasets show that this method exhibits better performance in both qualitative and quantitative results. Moreover, our method can effectively alleviate the problem of insufficient training data in the Re-ID task. In the future, we expect to extend our research to video frame prediction and person Re-ID.

CRediT authorship contribution statement

Meng Wang: Conceptualization, Methodology, Validation, Writing – review & editing, Supervision, Funding acquisition. **Jiaxing Chen:** Methodology, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Haipeng Liu:** Validation, Resources, Supervision, Project administration.

Declaration of competing interest

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

Data availability

The authors are unable or have chosen not to specify which data has been used.

Acknowledgments

The work is supported by National Natural Science Foundation of China (62062048). This work is also supported by Faculty of Information Engineering and Automation, Kunming University of Science and Technology, China.

References

[1] Ma L, Jia X, Sun Q, Schiele B, Tuytelaars T, Van Gool L. Pose guided person image generation. Adv Neural Inf Process Syst 2017;30.
[2] Zhang W, He X, Lu W, Qiao H, Li Y. Feature aggregation with reinforcement learning for video-based person re-identification. IEEE Trans Neural Netw Learn Syst 2019;30(12):3847–52.
[3] Zhang W, He X, Yu X, Lu W, Zha Z, Tian Q. A multi-scale spatial-temporal attention model for person re-identification in videos. IEEE Trans Image Process 2019;29:3365–73.
[4] Dong H, Liang X, Shen X, Wang B, Lai H, Zhu J, et al. Towards multi-pose guided virtual try-on network. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 9026–35.
[5] Yang H, Zhang R, Guo X, Liu W, Zuo W, Luo P. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 7850–9.
[6] Villegas R, Yang J, Zou Y, Sohn S, Lin X, Lee H. Learning to generate long-term future via hierarchical prediction. In: International conference on machine learning. PMLR; 2017, p. 3560–9.
[7] Yang C, Wang Z, Zhu X, Huang C, Shi J, Lin D. Pose guided human video generation. In: Proceedings of the European conference on computer vision. 2018, p. 201–16.
[8] Pumarola A, Agudo A, Sanfeliu A, Moreno-Noguer F. Unsupervised person image synthesis in arbitrary poses. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 8620–8.
[9] Siarohin A, Lathuilière S, Sangineto E, Sebe N. Appearance and pose-conditioned human image generation using deformable gans. IEEE Trans Pattern Anal Mach Intell 2019;43(4):1156–71.
[10] Dong H, Liang X, Gong K, Lai H, Zhu J, Yin J. Soft-gated warping-gan for pose-guided person image synthesis. Adv Neural Inf Process Syst 2018;31.
[11] Zhang J, Liu X, Li K. Human pose transfer by adaptive hierarchical deformation. In: Computer graphics forum, vol. 39, Wiley Online Library; 2020, p. 325–37.
[12] Huang S, Xiong H, Cheng ZQ, Wang Q, Zhou X, Wen B, et al. Generating person images with appearance-aware pose stylizer. 2020, arXiv preprint arXiv:2007.09077.
[13] Li T, Zhang W, Song R, Li Z, Liu J, Li X, et al. Pot-GAN: Pose transform GAN for person image synthesis. IEEE Trans Image Process 2021;30:7677–88.
[14] Esser P, Sutter E, Ommer B. A variational U-Net for conditional appearance and shape generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 8857–66.
[15] Isola P, Zhu JY, Zhou T, Efros AA. Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 1125–34.
[16] Kingma DP, Welling M. Auto-encoding variational bayes. 2013, arXiv preprint arXiv:1312.6114.
[17] Li Y, Huang C, Loy CC. Dense intrinsic appearance flow for human pose transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 3693–702.
[18] Siarohin A, Sangineto E, Lathuiliere S, Sebe N. Deformable gans for pose-based human image generation. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 3408–16.
[19] Li K, Zhang J, Liu Y, Lai YK, Dai Q. Pona: Pose-guided non-local attention for human pose transfer. IEEE Trans Image Process 2020;29:9584–99.
[20] Mahendran A, Vedaldi A. Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, p. 5188–96.
[21] Ren Y, Yu X, Chen J, Li TH, Li G. Deep image spatial transformation for person image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 7690–9.
[22] Yang L, Wang P, Liu C, Gao Z, Ren P, Zhang X, et al. Towards fine-grained human pose transfer with detail replenishing network. IEEE Trans Image Process 2021;30:2422–35.
[23] Men Y, Mao Y, Jiang Y, Ma WY, Lian Z. Controllable person image synthesis with attribute-decomposed gan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 5084–93.
[24] Ren Y, Yu X, Chen J, Li TH, Li G. Deep image spatial transformation for person image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020, p. 7690–9.
[25] Li Y, Huang C, Loy CC. Dense intrinsic appearance flow for human pose transfer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 3693–702.
[26] Ren Y, Fan X, Li G, Liu S, Li TH. Neural texture extraction and distribution for controllable person image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022, p. 13535–44.

[27] Zhou X, Yin M, Chen X, Sun L, Gao C, Li Q. Cross attention based style distribution for controllable person image synthesis. 2022, arXiv preprint arXiv:2208.00712.
[28] Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q. Scalable person re-identification: A benchmark. In: Proceedings of the IEEE international conference on computer vision. 2015, p. 1116–24.
[29] Liu Z, Luo P, Qiu S, Wang X, Tang X. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 1096–104.
[30] Zhu Z, Huang T, Shi B, Yu M, Wang B, Bai X. Progressive pose attention transfer for person image generation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019, p. 2347–56.
[31] Han X, Wu Z, Wu Z, Yu R, Davis LS. Viton: An image-based virtual try-on network. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 7543–52.
[32] Tang H, Bai S, Zhang L, Torr PH, Sebe N. Xinggan for person image generation. In: European conference on computer vision. Springer; 2020, p. 717–34.
[33] Chan C, Ginosar S, Zhou T, Efros AA. Everybody dance now. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 5933–42.
[34] Balakrishnan G, Zhao A, Dalca AV, Durand F, Guttag J. Synthesizing images of humans in unseen poses. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 8340–8.
[35] Zhang J, Li K, Lai YK, Yang J. Pise: Person image synthesis and editing with decoupled gan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 7982–90.
[36] Lv Z, Li X, Li X, Li F, Lin T, He D, et al. Learning semantic person image generation by region-adaptive normalization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021, p. 10806–15.
[37] Tang H, Bai S, Torr PHS, Sebe N. Bipartite graph reasoning gans for person image generation. 2020, arXiv preprint arXiv:2008.04381.
[38] Liu W, Piao Z, Min J, Luo W, Ma L, Gao S. Liquid warping gan: A unified framework for human motion imitation, appearance transfer and novel view synthesis. In: Proceedings of the IEEE/CVF international conference on computer vision. 2019, p. 5904–13.
[39] Huang X, Belongie S. Arbitrary style transfer in real-time with adaptive instance normalization. In: Proceedings of the IEEE international conference on computer vision. 2017, p. 1501–10.
[40] Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 2117–25.
[41] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, arXiv preprint arXiv:1409.1556.
[42] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis 2015;115(3):211–52.
[43] Gatys LA, Ecker AS, Bethge M. Image style transfer using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 2414–23.
[44] Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: From error visibility to structural similarity. IEEE Trans Image Process 2004;13(4):600–12.
[45] Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X. Improved techniques for training gans. Adv Neural Inf Process Syst 2016;29.
[46] Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Adv Neural Inf Process Syst 2017;30.
[47] Zhang R, Isola P, Efros AA, Shechtman E, Wang O. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 586–95.
[48] Gong K, Liang X, Li Y, Chen Y, Yang M, Lin L. Instance-level human parsing via part grouping network. In: Proceedings of the European Conference on Computer Vision. 2018, p. 770–85.
[49] Cao Z, Simon T, Wei SE, Sheikh Y. Realtime multi-person 2D pose estimation using part affinity fields. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017, p. 7291–9.
[50] Pala P, Seidenari L, Berretti S, Del Bimbo A. Enhanced skeleton and face 3D data for person re-identification from depth cameras. Comput Graph 2019;79:69–80.
[51] Protopapadakis E, Voulodimos A, Doulamis A, Doulamis N, Stathaki T. Automatic crack detection for tunnel inspection using deep learning and heuristic image post-processing. Appl Intell 2019;49(7):2793–806.
[52] Acharya UR, Fujita H, Oh SL, Hagiwara Y, Tan JH, Adam M, et al. Deep convolutional neural network for the automated diagnosis of congestive heart failure using ECG signals. Appl Intell 2019;49(1):16–27.
[53] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 770–8.
[54] Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z. Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 2818–26.