VTON-MP: Multi-Pose Virtual Try-On via Appearance Flow and Feature Filtering

Feng Yu, Ailing Hua, Chenghu Du, Minghua Jiang, Xiong Wei, Tao Peng, Lijun Xu, and Xinrong Hu

***Abstract*—Multi-pose virtual try-on has become a research focus for online clothes shopping due to the fixed-pose virtual try-on methods that cannot provide a different pose try-on effect. The challenge of multi-pose virtual try-on is that the detailed information of a generated image is difficult to obtain in the pose transformation and garment distortion. To solve the issue, we propose a multi-pose virtual try-on method via appearance flow and feature filtering (VTON-MP). First, a segmentation generation network of 2D keypoints about the target pose is used to predict the body semantic distribution of the target pose. Second, the desired garment is distorted to correspond to the body posture using the appearance flow figure alignment network (AFFAN). Third, latent useless feature weights are restrained using a filtering-enhancement block (FEB), and effective appearance feature weights are enhanced. Finally, the spatial relationship of body parts in the resulting image is further optimized using spatially-adaptive instance normalization (SAIN). Compared to state-of-the-art methods of subjective and objective experiments on the MPV dataset, the proposed VTON-MP achieves the best performance in terms of SSIM, PSNR and FID. The experimental results demonstrate that the proposed algorithm can better retain image details (head, hands, arms, and trousers).**

***Index Terms*—Virtual try-on, appearance flow, semantic segmentation, instance normalization, feature filtering.**

I. Introduction

With the intellectual development of the apparel industry, garment purchases are gradually shifting online. However, the shopping experience for consumers has worsened by the mismatch, temperament, and style of garments purchased online and by many negative reviews and returns. This phenomenon imposes a considerable impact on the consumer's desire to shop. Thus, there is a growing interest in virtual try-on algorithms derived from consumer electronics via mobile terminals [1], and the algorithm attempts to achieve the visual effect of temperament and style matching for online consumers.

The state-of-the-art methods are mainly divided into two categories: 3D model reconstruction methods [2], [3], [4] and 2D image generation methods [5], [6], [7]. The 3D model reconstruction methods use computer graphic technology to reconstruct the desired body model, adjust the model's pose according to the posed 3D key points, and draw the pose as an image output after performing skinning using the garment image. The disadvantage of this type of approach is apparent: the model reconstruction process requires high computer performance. Moreover, the generated models need great precision, and the different shapes of the bodies require separate simulations or settings for each body shape. This is difficult for online virtual try-on applications.

Thus many researchers prefer to study 2D image-based virtual try-on methods. The CP-VTON [8] and ACGPN [9] are popular methods for distorting garments and transferring them onto the corresponding parts of a person. However, they can only perform fixed-pose virtual try-on tasks, i.e., distort the target in-shop garment image to the appropriate body area in a fixed-pose. However, fixed-poses do not satisfy the need for users to try on garments in multiple poses. Moreover, these types of networks are no longer effective in the face of multi-pose virtual try-on tasks, and taking multiple pictures to achieve pose changes is too cumbersome.

To realize multi-pose virtual try-on approaches, [10], [11], [12], [13], [14] proposed some advanced methods with different technologies. References [10] and [11] colored garment regions for realizing virtual try-on. References [12], [13], [14] used a garment distortion technology to realize virtual try-on. The former method needs help achieving good try-on performance under different garment styles. Thus, garment distortion technology is commonly used in virtual try-on research. Dong et al. [12] proposed a 4-stage coarse-to-fine network called the MG-VTON. First, it used the target pose to generate a semantic map [15], [16], distorted the target in-shop garment image, and then generated a coarse result. Next, it refined the garment region of the coarse result to generate the final refined result. The face obtained by the MG-VTON is blurred

Manuscript received 6 October 2022; revised 1 December 2022, 17 February 2023, and 29 June 2023; accepted 13 August 2023. Date of publication 17 August 2023; date of current version 21 February 2024. This work was supported in part by the Hubei Key Research and Development Program under Grant 2021BAA042; in part by the Wuhan Applied Basic Frontier Research Project under Grant 2022013988065212; and in part by the MIIT's AI Industry Innovation Task Unveils Flagship Projects (Key Technologies, Equipment, and Systems for Flexible Customized and Intelligent Manufacturing in the Clothing Industry). The virtual try-on technology is the key research content of above projects. This manuscript is an further version of our previous work, which appeared at the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2022) (Du, Yu et al., 2022) [DOI: 10.1109/ICASSP43922.2022.9747847]. *(Corresponding author: Minghua Jiang.)*

Feng Yu, Minghua Jiang, Tao Peng, and Xinrong Hu are with the School of Computer Science and Artificial Intelligence and the Engineering Research Center of Hubei Province for Clothing Information, Wuhan Textile University, Wuhan 430200, China (e-mail: yufeng@wtu.edu.cn; minghuajiang@wtu.edu.cn; pt@wtu.edu.cn; hxr@wtu.edu.cn).

Ailing Hua, Chenghu Du, and Xiong Wei are with the School of Computer Science and Artificial Intelligence, Wuhan Textile University, Wuhan 430200, China (e-mail: hal_wtu@163.com; duceh_lzy@163.com; wx_wh@wtu.edu.cn).

Lijun Xu is with the Department of Computer and Information Engineering, Hubei University, Hubei 430062, China (e-mail: xulijun@hubu.edu.cn).

Digital Object Identifier 10.1109/TCE.2023.3306206

because it only refined the garment region and did not further process facial details. To address this problem, Wang et al. [13] proposed another 4-stage coarse-to-fine network called the TB-VTON. Unlike the MG-VTON, the TB-VTON used a garment mask to optimize the garment region in the coarse result generation stage. In the final stage, the TB-VTON introduced a facial refinement network to improve the clarity of the resulting face. Hu et al. [14] proposed an end-to-end image-based multi-pose virtual try-on method called the SPG-VTON. Compared to the MG-VTON, the SPG-VTON generated an image mask prediction and a coarse result for the target in-shop garment; it optimized the garment region with the predicted target in-shop garment image mask. The face was refined using a face refinement network in the final stage.

In summary, the challenge of the multi-pose virtual try-on task is to learn both the distortion of the target in-shop garment image and the transformation of the human pose simultaneously. The three methods [12], [13], [14] mentioned above attempt to generate realistic images of people. However, some problems limit their application and development, including: 1) the details (e.g., logos and patterns) of the distorted garment in the presence of an arm obscuring the multi-pose virtual try-on task are not effectively retained; and 2) other characteristics parts of the image (e.g., head, hands, arms, and trousers) are not kept realistic.

To address the challenges mentioned above, we propose a novel multi-stage framework called the VTON-MP, which generates realistic multi-pose try-on results using a semantic map. In contrast to the transformation of rigid bodies, such as cars and seats [17], [18], which was achieved by using the affine transformation technique, the human body undergoes a certain level of deformation of individual parts when its pose is changed. An appearance flow [17], [19] represents the two-dimensional coordinate vectors in a source, which can be used to synthesize a target. These two-dimensional coordinate vectors specify which pixels in the input view can be used to reconstruct the target view. Appearance flow was first applied to synthesizing a rigid body by Zhou et al. [17]. Their approach can obtain an image of the desired pose when given an image of one or multiple viewpoints. Since the spatial relationships between the body pixel points and a rigid body become different after human body motion, it is not feasible to directly use appearance flow for multi-pose virtual try-on tasks. The generation of GANs has considerably contributed to image synthesis [20], [21], [22]. Pix2Pix [23] realized the image-to-image translation task and had been upgraded to the high-resolution Pix2PixHD [24] scheme, MFFN [25] also contributed a lot to the high-resolution. It is crucial to restore the image, which FFTI [26] explored this effectively. However, CNNs cannot deal with sizeable spatial deformations. Space information and pixel relationships are often lost in the encoding and decoding stages, so it could be more efficient to use both to transform a semantic map into an image of a person. The StyleGAN for style translation proposed by Karras et al. [27] utilized adaptive instance normalization (AdaIN) to maintain the generator's style control on the generated images and synthesize stunning results. However, generators for global style translation have limited abilities to synthesize sophisticated parts, such as the hands, arms, and trousers. The diffusion model [28], [29] is a parameterized Markov chain trained using variational inference to produce samples matching the data after a finite time. It has been shown to have excellent performance in the task of image generation. There is still room for improvement in sampling acceleration, likelihood maximization, and data generalization.

Therefore, VTON-MP consists of three parts: 1) a segmentation generation network (SGN), reassigning the target area of the semantic map with the target posed 2D keypoints and target in-shop garment image to ensure the correctness of the desired semantic map; 2) an appearance flow-based figure alignment network (AFFAN) is used for both garment distortion and pose transformation, and it better preserves the details (e.g., logos and patterns) of the distorted desired garment and the other characteristics of the image (e.g., hands, arms, and trousers); and 3) a filtering-enhancement-synthesis network (FESN), which fuses the distorted garment, the transferred body and the predicted semantic body map together for the final image synthesis process. We design a filtering-enhancement block (FEB) that filters the potential useless pixel code and enhances the latent target detail code, which enables the body to be refined. To enhance the body regions' structural information while decoding the target image's optimized information code, we introduce spatially adaptive instance normalization (SAIN) [30] to the deconvolution process.

Experiments show that the VTON-MP achieves substantially improved multi-pose virtual try-on image synthesis performance. The main contributions of our work are summarized as follows:

- We propose a novel image-based virtual try-on framework for multi-pose settings called the VTON-MP. For challenging poses (with an obscuring arm), it successfully transfers the body image to the target pose, synthesizes the target in-shop garment image, and generates more explicit and realistic virtual try-on images. The VTON-MP enables the generated target image to be facially clear, with all nontarget areas remaining the same as those in the original image except for the target in-shop garment image area.
- To enhance the retention of details in the trousers and arms during the multi-pose virtual try-on process, we extend the appearance flow application. This approach is applied to garment distortion and pose transformation, allowing the features of the source human image to be mapped to the target image as a whole. This method effectively addresses issues, such as the loss of garment details, and the other characteristic parts of the image (e.g., the head, hands, arms, and trousers) are kept realistic.
- We design the FESN technique to refine the obtained image globally to produce a clear and realistic image of the virtual try-on. It filters the potential useless pixel code and enhances the latent target detail code, thereby optimizing the synthesis process to preserve the original information effectively.

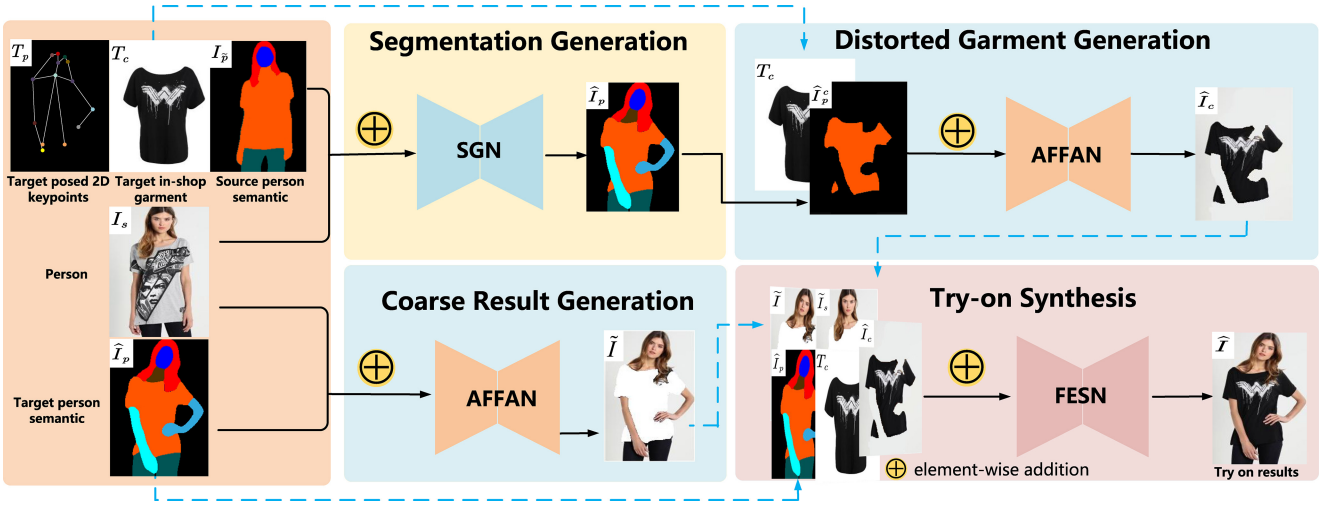

Fig. 1. The overview of the VTON-MP framework. Our framework consists of three parts: 1) the SGN is used to predict the target semantic map $\hat{I}_p$ with the target posed 2D keypoints T_p and the target in-shop garment image T_c. 2) The AFFAN is used for the distorted garment generation and transferred pose generation processes. Simultaneous garment distortion and body pose transformation are performed through the same network to retain the other characteristics of the image (e.g., hands, arms, and trousers) and allow the distorted garment to better fit the transformed target result. 3) The try-on synthesis process combines the distorted garment, transferred pose, and predicted target semantic map through the FESN for final image synthesis.

II. Methodology

Our goal is to input a source human image $I_s \in R^{3\times H\times W}$ (all H's are 256 and all W's are 192), a target in-shop garment image $T_c \in R^{3\times H\times W}$, a source human semantic map image $I_p \in R^{20\times H\times W}$, and a set of target posed 2D keypoints $T_p \in R^{18\times H\times W}$ into the VTON-MP (defined as $\mathcal{M}$) to learn to transform the source human image I_s into the image with the target pose and the target in-shop garment image. Then, we need to change the shape of the target in-shop garment T_c and the pose of the body under the target human pose T_p; finally, an output image $\hat{I} \in R^{3\times H\times W}$ is generated, where the appearance features of the generated image $\hat{I}$ should approximate the ground-truth image as much as possible. The generation process can be expressed as:

$$\hat{I} = \mathcal{M} < I_s, T_c, T_p, I_p > \tag{1}$$

The entire framework is shown in Figure 1, and we describe our architecture in more detail below.

A. Segmentation Generation Network (SGN)

The synthesis of virtual try-on images needs to be guided by the target human semantic map, so ensuring the correctness of the target human semantic map generation process directly affects the final generation effect. The task of the SGN is to generate a desired target human semantic map $\hat{I}_p \in R^{20\times H\times W}$ with the shape of the input target posed 2D keypoints T_p. We mix the upper limbs, garment, and neck in the source human semantic map I_p into a preallocated region to obtain the human semantic map $I_{\tilde{p}}$. With the target posed 2D keypoints T_p, $I_{\tilde{p}}$ is converted into $\hat{I}_{\tilde{p}}$. The preallocated region in $\hat{I}_{\tilde{p}}$ is reallocated according to garment image T_c to form the target human semantic map $\hat{I}_p$. In the SGN, we adopt the U-Net [31] as the generation network and optimize it with pixel-level cross-entropy [32].

B. Appearance Flow Figure Alignment Network (AFFAN)

The desired garment T_c is distorted to the target body without the loss of details, which has been a primary challenge in virtual try-on tasks. Current methods [12], [13], [14] use the thin-plate spline (TPS) [33] algorithm to distort the garment with respect to the target posed 2D keypoints T_p. However, the images generated by this method often result in an overdistorted or underdistorted effect. Although they refine the coarse try-on result, the result is still a badly distorted appearance of the garment and involves an excessive loss of detail. As shown in Figure 2, we reference and extend a deformation network called the appearance flow garment alignment network (AFGAN) [34].

We input a pair containing the target in-shop garment image T_c and the garment semantic map $\hat{I}_P^C$ into AFFAN. The features in the garment image are extracted by encoder C, and the pose features in the semantic map are extracted by encoder P, resulting in a distorted garment $\hat{I}_p$. After completing the garment distortion step, we input the pair containing the source of the human image I_s and predicted target human semantic map $\hat{I}_p$ into the AFFAN, enabling the remaining body parts to be transferred to the target pose T_p. This procedure transfers the source human image to obtain a priori human deformation information, and the garment area is removed to obtain the coarse result $\tilde{I}$.

1) Image-Semantic Correspondence Module (ISCM): The use of a single encoder only allows mapping between the source human image and target human semantic map to be formed in a restricted feature space domain, i.e., the various parts of the source human image and target human semantic map cannot be associated, resulting in the strict limitation of the geometric transformation range corresponding to the source human image during the transformation process. Therefore, the distortion task cannot be completed. We use two different encoders, both of which consist of five convolution

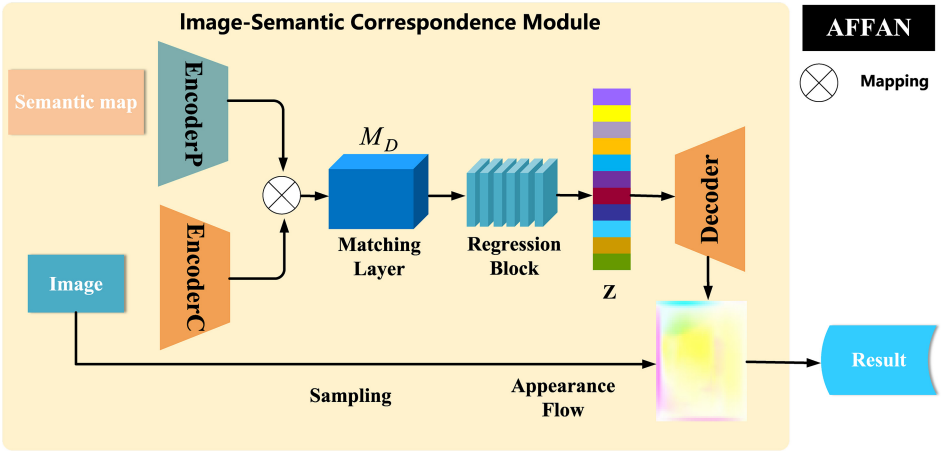

Fig. 2. The structure of the AFFAN. The mapping structure between the target human semantic map features and the source human image features generates a stable and efficient appearance flow for garment distortion and pose transfer.

layers, to extract high-level features $f_I \in \mathbb{R}^{512\times16\times12}$ from the target human semantic map and $f_P \in \mathbb{R}^{512\times16\times12}$ from the source human image. Then, we computationally map the features of both inputs to the common domain D so that we can learn reliable content within domain D. We compute a semantic image correspondence matrix $M_D \in \mathbb{R}^{16\times12\times(16\times12)}$ to express the mapping relationship between f_I and f_P in the common domain D.

$$M_D(u, v, w) = f_I(u, v)^T f_P(u_w, v_w) \tag{2}$$

where (u, v) and (u_w, v_w) denote the individual feature positions in the 16×12 dense feature maps, $w = 16(v_w-1)+u_w$ is an auxiliary index variable for (u_w, v_w), and T denotes matrix transposition.

The network further encodes matrix M_D, performs a series of convolutions, and conducts linear regression through a regression block, which consists of four convolution layers and a 786-dimensional fully connected layer. Then, a code z containing the latent appearance flow information is obtained. Finally, the code z is decoded to obtain a transformed appearance flow of the desired result image that corresponds to the semantic map, and the decoder consists of five layers of deconvolution operations.

2) Flow Interval Restriction: While the garment is distorted, some appearance flow coordinates are excessively and unevenly stretched if the distortion of the sampled garment is not restricted between adjacent flow coordinates. To prevent this problem, we introduce an appearance flow interval-restricted loss $\mathcal{L}_{fir}$ during the appearance flow learning step to maintain isometry between adjacent flow coordinates. The $\mathcal{L}_{fir}$ loss is defined as follows:

$$\begin{aligned}\mathcal{L}_{fir}\left(\hat{G}_x, \hat{G}_y\right) = &\sum_{i=-1,1}\sum_{x}\sum_{y}|\hat{G}_x(x+i, y) - \hat{G}_x(x, y)| \\ &+ \sum_{j=-1,1}\sum_{x}\sum_{y}|\hat{G}_y(x, y+j) - \hat{G}_y(x, y)|\end{aligned} \tag{3}$$

where $\hat{G}_x$ and $\hat{G}_y$ are the x and y coordinates of the grid to be mapped, respectively, and the absolute difference $|a-b|$ is used to measure the distance between two adjacent nodes a and b. $\mathcal{L}_{fir}$ denotes the distance loss between a certain point (x, y), and its adjacent points $(x+1, y)$, $(x-1, y)$, $(x, y+1)$ and $(x, y-1)$ among the grid sampling control points.

As shown in Figure 10, the introduced $\mathcal{L}_{fir}$ loss effectively avoids excessive garment distortion. It has a better performance than other losses in terms of maintaining the naturalness of the garment details.

C. Filtering-Enhancement Synthesis Network (FESN)

Once garment distortion and pose transformation have been achieved, it is critical to bring the results closer to reality during the synthesis process. The MG-VTON uses a coarse-to-fine framework, where a coarse result is first generated using Warp-GAN [35] and the target semantic map. Then, the garment is refined, which means that the rest of the body has a minimal gain relative to the garment region during the refinement stage. The TB-VTON already refines the garment region when generating the coarse results, so it designs a face refinement network to adjust for facial details. However, the details of other image regions (trousers, arms, hand, etc.) are left out.

To simultaneously refine the body regions, we design the FESN for the one-step generation of the target image in an end-to-end manner, as shown in Figure 3. Since human body motions are nonrigid body motions and the spatial relationships between adjacent pixels in the image may change after movement, the result of sampling directly through the appearance flow is coarse. It contains many redundant pixels and missing regions in the image result.

We use the source human image without the source garment $\tilde{I}_s \in R^{3\times H\times W}$(for obtaining enhancement matrix $\mathcal{M}_e$), the coarse result $\tilde{I} \in R^{3\times H\times W}$ (for obtaining suppression matrix $\mathcal{M}_s$), the target posed 2D keypoints $T_p \in R^{18\times H\times W}$ (for locating facial parts), the distorted garment $\hat{I}_c \in R^{3\times H\times W}$ (for virtual try-on purposes), and the target semantic map $\hat{I}_p \in R^{20\times H\times W}$ (for generating the region bound) as the image representation $P_t \in R^{47\times H\times W}$, which is input into the FESN. Some of the pixels in the coarse result $\tilde{I}$ are redundant or missing and cannot be directly used as a pixel basis for the input

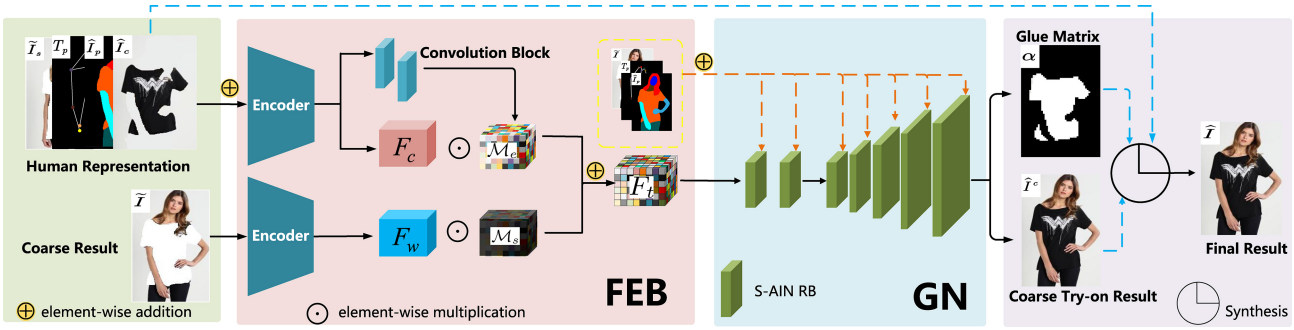

Fig. 3. General structure of the FESN. It is mainly composed of two parts: an FEB and a generative network (GN). The latent useless feature weights are restrained using the FEB, and the effective appearance feature weights are enhanced. The GN is used to decode the feature map to obtain the final target image.

image. It is essential to track valuable features [36], [37], [38], [39], [40], so we design the FEB to filter and enhance the input body representation P_t during the FESN feature extraction stage.

The encoding section consists of two encoders with the same structure, each consisting of six downsampling layers with a channel step up. We define the feature maps obtained by the content encoder as F_c, and by the distorted encoder as F_w. Then we convolve the content feature map F_c to predict an enhancement matrix $\mathcal{M}_e \in \mathbb{R}^{512\times16\times12}$, which is used to motivate the pixel feature weights represented by the predicted target image $\hat{I}$. To filter out redundant pixels in the coarse result $\tilde{I}$, we compute a suppression matrix $\mathcal{M}_s \in \mathbb{R}^{512\times16\times12}$ that suppresses the useless weights in the coarse resultant features represented by F_w. Finally, by fusing the processed F_c and F_w, a feature map F_t for all latent target image information can be obtained:

$$F_t = \mathcal{M}_e \odot F_c + \mathcal{M}_s \odot F_w \tag{4}$$

$$\mathcal{M}_s = 1 - \mathcal{M}_e \tag{5}$$

where $\odot$ denotes elementwise multiplication, $+$ denotes elementwise addition, and $-$ denotes elementwise subtraction.

Generative Network (GN): after obtaining the feature map F_t containing all target image information, it should be decoded to obtain the coarse try-on image $\tilde{I}$. The decoding section of the FESN consists of seven spatially adaptive instance normalization ResBlocks.

Spatially-Adaptive Instance Normalization ResBlock (S-AIN RB): popular image-to-image algorithms use the semantic map directly as the input of the GN for computation purposes, and the traditional normalization layers that are commonly used in GNs tend to lose spatial or semantic information contained in the semantic map. Spatial adaptive normalization (SPADE) [41] is a modification of batch normalization (BN) that has led to the proposal of a residual structure called SPADE ResBlk. To generate more realistic images, we change the BN procedure of the module to AdaIN to obtain the S-AIN RBs, as shown in Figure 4. We restructure the sequence of convolutional modules to make the model more suitable for multi-pose virtual try-on tasks. AdaIN is formulated as follows:

$$AdaIN(x_c, y) = y_c^w \left(\frac{x_c - \beta(x_c)}{\gamma(x_c)} \right) + y_c^b \tag{6}$$

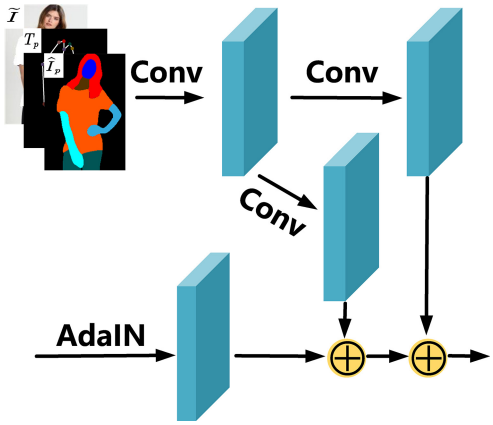

Fig. 4. The structure of the Spatially-Adaptive Instance Normalization ResBlock (S-AIN RB).

where x and y denote different features, y is the set of the source human image I_s, the target semantic map $\hat{I}_p$, the feature map, and the target posed 2D keypoints T_p, c is the number of channels, β is the mean, and γ is the standard deviation, y_c^w and y_c^b denote embeddings of feature y.

To seamlessly blend the body region and the garment, we predict a glue matrix $\alpha \in R^{1\times H\times W}$ to fuse the distorted garment $\hat{I}_c$, and the coarse try-on result $\hat{I}^c$ generated by the FESN to obtain the final virtual try-on result $\hat{I}$, where the fusion process can be expressed as follows:

$$\hat{I} = \alpha \odot \hat{I}_c + (1 - \alpha) \odot \hat{I}^c \tag{7}$$

where $\odot$ denotes the element-wise multiplication, $+$ denotes the element-wise addition, $\hat{I}^c$ is the coarse result generated by FESN.

D. Training

The following loss functions are involved in the training process and Algorithm 1 is an overview of the training process.

1) Reconstruction Loss: To guide the network and generate a visual appearance similar to that of the reference image (ground truth), the reconstruction loss $\mathcal{L}_{rec}$ is introduced to calculate the pixel-level differences. $\mathcal{L}_{rec}$ is expressed as the $\mathcal{L}_1$ distance between the generated final try-on image $\hat{I}$ and

Algorithm 1 Procedure of the Virtual Try-On Network

INPUT($\mathbf{T_c}$ $\mathbf{I_s}$ $\mathbf{T_p}$ $\mathbf{I_p}$)

$I_{\tilde{p}} \leftarrow mix(I_p) \iff$ Initialize image I_p by mixing the upper limbs, garment, and neck in the source human semantic map I_p

SGN ($\mathbf{T_c}$ $\mathbf{I_s}$ $\mathbf{T_p}$ $\mathbf{I_{\tilde{p}}}$)

$\hat{I}_{\tilde{p}} \leftarrow transfer(I_{\tilde{p}}, T_p, I_s) \iff$ Transfer $I_{\tilde{p}}$ into a semantic map $\hat{I}_{\tilde{p}}$ that matches the target pose T_p

$\hat{I}_p \leftarrow re\text{-}allocated(\hat{I}_{\tilde{p}}, T_c) \iff$ Allocate the preallocated region with $\hat{I}_{\tilde{p}}$ to obtain the target human semantic map $\hat{I}_p$

return $\hat{I}_p$

AFFAN($\mathbf{T_p}$ $\mathbf{I_s}$ $\hat{\mathbf{I}}_p$ $\hat{\mathbf{I}}_P^C$)

Extraction of the garment area semantic map $\hat{I}_P^C$ from the predicted target human semantic map $\hat{I}_p$

$\tilde{I} \leftarrow AFFAN(I_s, \hat{I}_p) \iff$ With I_s, and $\hat{I}_p$, the image $\tilde{I}$ of the human body after pose transformation is obtained, and its top area is removed to facilitate the fit of the garment to the body

$\hat{I}_{\tilde{c}} \leftarrow AFFAN(\hat{I}_P^C, T_c) \iff$ With T_c, and $\hat{I}_P^C$, obtain a distorted garment $\hat{I}_{\tilde{c}}$ that fits the target pose

$\hat{I}_c \leftarrow mul(\hat{I}_P^C, \hat{I}_{\tilde{c}})$

return $\hat{I}_c$, $\tilde{I}$

FESN($\tilde{\mathbf{I}}_s$ $\mathbf{T_p}$ $\hat{\mathbf{I}}_c$ $\hat{\mathbf{I}}_p$ $\tilde{\mathbf{I}}$)

$F_c \leftarrow Encoder\_A(\tilde{I}_s, T_p, \hat{I}_c, \hat{I}_p) \iff$ Feature extraction is performed on $\hat{I}_c$, $\tilde{I}_s$, T_p, and $\hat{I}_p$ by the content encoder to obtain feature map F_c

$F_w \leftarrow Encoder\_B(\tilde{I}) \iff$ Feature extraction is performed on $\tilde{I}$ by the distorted encoder to obtain feature map F_w

$\mathcal{M}_s \leftarrow 1\text{-}\mathcal{M}_e \quad \mathcal{M}_e \leftarrow Conv(F_c)$

$F_t = \mathcal{M}_e \odot F_c + \mathcal{M}_s \odot F_w \iff$ With F_c, F_w, their processed feature maps are fused by the enhancement matrix $\mathcal{M}_e$ and suppression matrix $\mathcal{M}_s$ to obtain F_t

With $F_t, \tilde{I}, \hat{I}_p, T_p$, the GN is decoded to obtain a preliminary try-on image $\hat{I}^c$ and a predicted glue matrix α

With α, the garment is seamlessly blended with the body area to obtain the final virtual try-on image $\hat{I}$

return $\hat{I}$

the reference image $\bar{I}$:

$$\mathcal{L}_{rec}\left(I_s, \bar{I}, \hat{I}_p^c\right) = \lambda_{rec}\|\hat{I} - \bar{I}\|_1 + \|\hat{I}_p^c - \alpha\|_1 \tag{8}$$

2) Perceptual Loss: It is a practical approach that extracting the feature maps of the generated image $\hat{I}$ and the reference image $\bar{I}$, which from some layers of the pre-trained VGG [42] network, for feature matching in guiding the real image generation process. Therefore, we introduce a perceptual loss $\mathcal{L}_{per}$, which can be expressed as:

$$\mathcal{L}_{per}\left(\hat{I}, \bar{I}\right) = \sum_{i=0}^{n} \alpha_i \|\phi_i\left(\hat{I}\right) - \phi_i\left(\bar{I}\right)\|_1 \tag{9}$$

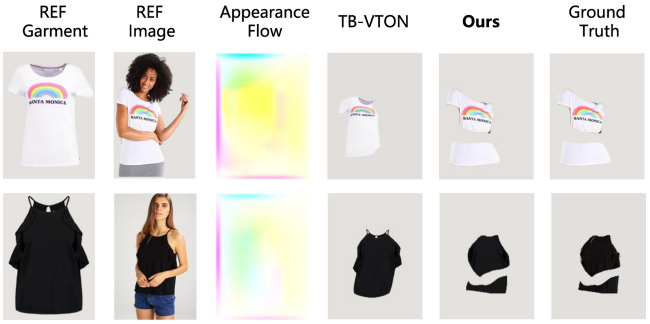

Fig. 5. Visual comparison between the warpage effects of the AFFAN and TB-VTON [13] (based on the TPS algorithm).

where $\phi_i(\cdot)$ denotes the feature map of the i-th (i = 0, 1, 2, 3, 4) layer from the pre-trained VGG-19 network ϕ, and α_i controls the weight of the loss for each layer.

3) Adversarial Loss: To generate more natural and realistic result images, we introduce an adversarial loss in the training phase to guide the generation of heads, arms, hands, etc. This loss function can be expressed as follows:

$$\begin{aligned}\mathcal{L}_{adv} = &\mathbb{E}_{x,y}\left[\log D(x, y)\right] \\ &+ \mathbb{E}_{x,z}\left[\log(1 - D(x, G(x, z)))\right]\end{aligned} \tag{10}$$

where x denotes the input data, y is the ground-truth image, and z indicates input noise sampled from the standard normal distribution.

4) Contextual Loss: To find the similarity metric between the generated image and the target image features, we introduce the contextual loss [43]. In contrast to the pixel-level loss, the contextual loss focuses on the similarity between the image and the feature distribution of the target image; it can learn to generate a more natural, less distorted texture and a more reasonable output image. This loss can be expressed as follows:

$$\mathcal{L}_{CX} = -log\left(CX\left(\phi_i\left(\hat{I}\right), \phi_i\left(\bar{I}\right)\right)\right) \tag{11}$$

where ϕ_i denotes the feature maps extracted from layer $i = ReLU\{3\_2, 4\_2\}$ of the pretrained VGG19 network for the input image, and CX denotes the contextual and semantic similarity metric between the matched features.

5) Overall Loss: The framework $\mathcal{M}$ needs to be trained separately for the AFFAN and FESN. The total loss $\mathcal{L}_{total}^{AFFAN}$ of the AFFAN consists of $\mathcal{L}_{rec}$, $\mathcal{L}_{per}$, and $\mathcal{L}_{fir}$. It can be expressed as:

$$\mathcal{L}_{total}^{AFFAN} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{fir}\mathcal{L}_{fir} \tag{12}$$

The total loss $\mathcal{L}_{total}^{FESN}$ of the FESN consists of $\mathcal{L}_{rec}$, $\mathcal{L}_{per}$, $\mathcal{L}_{CX}$, and $\mathcal{L}_{adv}$, and this function can be expressed as:

$$\mathcal{L}_{total}^{FESN} = \lambda_{rec}\mathcal{L}_{rec} + \lambda_{per}\mathcal{L}_{per} + \lambda_{CX}\mathcal{L}_{CX} + \lambda_{adv}\mathcal{L}_{adv}. \tag{13}$$

III. Experiments

A. Dataset

The dataset used in the experiments is the MPV dataset collated by Dong et al. [12], it consists of 14,754 pairs of garments, and female models in all directions with a resolution

Fig. 6. A visual comparison among the posture transformation results produced by our method and other state-of-the-art methods shows that our method yields the best representation of the red dashed boxes. Zoom in to display more details.

of 256 × 192. The MPV dataset is divided into a training set and a test set, containing 12,410 and 2,340 pairs of images, respectively.

B. Implementation Details

All experiments are carried out on 5 Tesla V100 devices with 32 G RAM. By default, the learning rate is 1.0×10^{-4} for the generator, and the discriminator (the structure of the discriminator is same as that in Pix2PixHD, as shown in Figure 1 (e)) is linearly reduced to 0 over the second half of epochs with a batch size of 4. The framework is trained with the adaptive moment estimation (ADAM) optimizer [44], and we set $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The time overheads for modules SGN, AFFAN, and FESN are 0.002s, 0.937s, and 0.249s.

During the training stage of the framework, the hyperparameters of the loss function in the appearance flow figure alignment network (AFFAN) are set to $\lambda_{rec} = \lambda_{per} = \lambda_{fir} = 1$, and the hyperparameters of the loss function in the Filtering-enhancement synthesis network (FESN) are set to $\lambda_{rec} = 10$, $\lambda_{per} = 5$, and λ_{CX}=$\lambda_{adv} = 1$.

C. Qualitative Results

We conduct a qualitative comparison between our method and the state-of-the-art benchmark TB-VTON method.

First, to verify the effectiveness of the AFFAN in the proposed framework, we compare the garment distortion effects of the TB-VTON [13] and VTON-MP using visualization. The TPS algorithm is used by the TB-VTON to distort the target in-shop garment images, where the distortion degree is controlled by the 50 sampling parameters θ. Some of the distorted garments are overdistorted or underdistorted due to the sampling characteristics of uneven local stretching. The region with the garment and covered arms is not processed further, resulting in the garment covering the arms, causing an unnatural effect. As shown in Figure 5, the AFFAN is remarkably effective in controlling garment distortion by learning the appearance flow change between the predistortion and postdistortion garments. By learning the feature mapping between the garment semantic map $\hat{I}_p^c$ and the source garment, the distorted garment logos, patterns, and other pixels are naturally sampled via the flow field. Since overlapping parts of the garment and arms do not exist in the semantic map, redundant pixels are not mapped. The results show strong performance where the effect is close to the ground truth.

As shown in Figure 6, we visually compare the pose transfer images produced by the AFFAN, the baseline TB-VTON [13] method and our method. We also visually compare the TB-VTON and VTON-MP in terms of a virtual try-on, as shown in Figure 7. The TB-VTON [13] results exhibit a high level of quality transfer. Nevertheless, because the TPS algorithm is used to distort the garment, the garment region

Fig. 7. A visual comparison between the virtual try-on results of our method and the TB-VTON [13], with the regions possessing improved details indicated by red dashed boxes. Zoom in to display more details.

Fig. 8. A comparison of our method with other state-of-the-art methods regarding the retention of detail in a target in-shop garment, trousers, and faces, which shows that our method yields the best representation of the red dashed box. Zoom in to display more details.

is unnaturally distorted. Although the TB-VTON [13] refines explicitly refines the face, the lack of information regarding the positioning of the facial organs (eyes, nose, etc.) also leads to a certain lack of facial detail. The AFFAN directly samples the source human images, which effectively preserves the appearance of the image features. However, the results of some images are abnormal because the excess pixels are not filtered after sampling. To further demonstrate the effectiveness of our method, we conducted comparative experiments with other state-of-the-art methods on the retention of details in a target in-shop garment, trousers, and faces, as shown in Figure 8; it can be clearly seen that the details in the three red box areas are better retained.

Our overall framework introduces the AFFAN to sample the appearance flow of the desired garment. Since the garment semantic map is oriented for mapping-based sampling, the distorted garments can be seamlessly aligned with the body regions while retaining many natural details. The filtering-enhancement block (FEB) is used to separately enhance and filter some features of the AFFAN results. The face is more realistically and naturally reproduced by adding the target posed 2D keypoints T_p as the positioning basis of the facial part of the input. Spatially-adaptive instance normalization (SAIN) allows the generated image to maintain the original spatial structure characteristics, avoiding further refinement operations and generating convincing and reasonable results in one step. As shown in Figure 8, we visually compare the pose transfer images produced by the baseline MG-VTON [12], TB-VTON [13], SPG-VTON [14] methods and our method. From the visual point of view, the VTON-MP is excellent for preserving the garment details in the generated image. The color and form of the hair, the facial features, etc., are highly consistent with those of the original image.

In summary, we use the AFFAN to distort the source garment, which effectively prevents overdistortion, underdistortion, covered arms, etc. We use the AFFAN to generate coarse transfer results $\tilde{I}$ to provide a priori human body information for the FESN. In the FESN, we design the FEB to maintain the appearance features of the source human images by filtering out the redundant pixels in the coarse results $\tilde{I}$ and enhancing the necessary features derived from the human representations P_t and coarse results $\tilde{I}$. In addition, SAIN has good performance in terms of constructing the spatial layout structures in the generated images. The visual results show that the VTON-MP performs well in the multi-pose virtual try-on

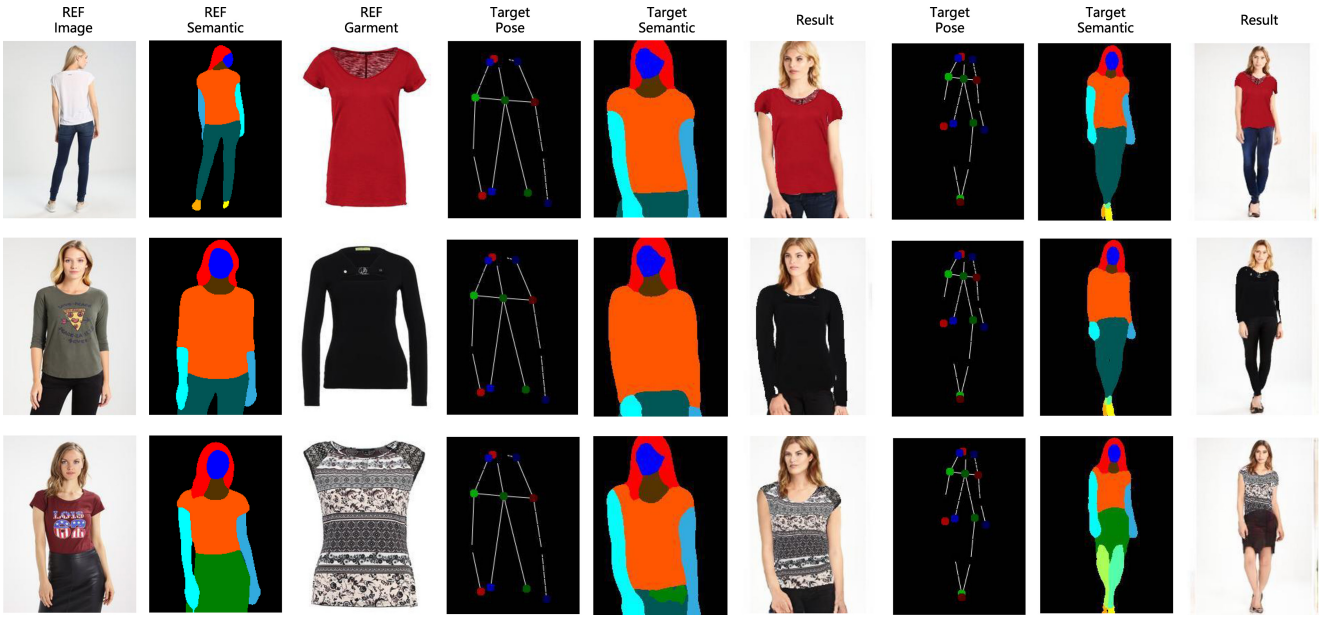

Fig. 9. The visual results obtained in the multi-pose virtual try-on scenario. Two target poses, three reference human bodies, and three reference garments are included.

Fig. 10. Ablation experiments conducted with the AFFAN in terms of garment deformation.

task, maintaining better details regarding the appearances of garments, the head, hands, arms, trousers, etc.

Multi-pose virtual try-on is completed by selecting two human poses to perform a visualized try-on effect experiment. As shown in Figure 9, the accurate prediction of the target semantic map corresponding to the target pose through the provided target pose is the basis for generating real multi-pose try-on images. Furthermore, the FESN is used to retain the original details of the source human body and has a very ideal effect.

D. Quantitative Results

We use the structural similarity index measure (SSIM) [45], inception score (IS) [46], fréchet inception distance (FID) [47], learned perceptual image patch similarity (LPIPS) [48], and peak signal-to-noise ratio (PSNR) to quantitatively evaluate the state-of-the-art networks. To verify the influence of the background on the quantitative evaluation, we calculate the mask-SSIM, mask-IS, mask-PSNR, mask-LPIPS, and mask-FID by masking the background. The IS metric is used to assess the sharpness and diversity of the generated image; a higher IS value indicates a sharper result and a greater variety of colors and effects. The SSIM is a metric for fully evaluating the referenced image quality; it measures image similarity in terms of brightness, contrast, and structure. A higher value indicates less image deformation. The LPIPS is used to measure the difference between two images, with a lower value indicating that the two images are more similar and vice versa. The PSNR is used to assess the quality of the resulting image obtained after compression in comparison with the original image. A higher PSNR value represents less deformation and higher quality when the image is compressed. Because the IS only considers the quality of the generated samples and does not consider the influence of the actual data, the FID is used to calculate the distance between the actual samples and the generated samples in the feature space. Therefore, a lower FID value means that the images have higher quality and diversity, and a lower LPIPS value means that it is more similar to the real image, as shown in Table I. Our method can achieve convincing scores in the quantitative assessment by adding details, such as garment aspects and human faces. AFFAN produced the highest calculated IS value and TB-VTON [13] produced the lowest LPIPS value due to the preservation of background integrity. However, in terms of other indicators, our method is able to achieve the best performance and the best LPIPS metrics in the absence of background interference.

E. Ablation Study

To demonstrate the necessity and effectiveness of each module in the proposed framework, we perform ablation experiments on the MPV dataset to separately remove the FEB, SAIN, $\mathcal{L}_{Rec}$ loss, and $\mathcal{L}_{CX}$ loss and record the corresponding IS, SSIM, PSNR, and FID values, as shown in Table II. We learn that the whole model obtains the highest SSIM score, the highest PSNR score, and the lowest FID score. We also note that the IS scores achieved without

TABLE I
QUANTITATIVE EVALUATION BETWEEN OUR METHOD AND OTHER STATE-OF-THE-ART METHODS ON THE MPV TEST SET

	IS↑	mask-IS↑	SSIM↑	mask-SSIM↑	PSNR↑	mask-PSNR↑	FID↓	mask-FID↓	LPIPS↓	mask-LPIPS↓
TB-VTON	2.7974±0.1267	2.6854±0.1428	0.6726	0.6913	16.2522	17.5957	22.2014	18.7796	0.263	0.173
AFFAN	**2.9895±0.1468**	3.0284±0.1712	0.6988	0.7398	16.6161	17.7960	31.0222	19.8827	0.301	0.176
Ours	2.9818±0.1831	**3.0509±0.1483**	**0.7842**	**0.7927**	**20.1536**	**20.7856**	**11.9856**	**10.9005**	0.273	**0.143**
Real	3.0410±0.1435	3.0564±0.1821	1	1	N/A	N/A	0	0	0	0

Fig. 11. The ablation study results obtained by the VTON-MP. The experiment consists of separately removing the FEB, removing the SAIN method, removing the $\mathcal{L}_{Rec}$ loss, and removing the $\mathcal{L}_{CX}$ loss.

the contextual loss are higher than those obtained with the contextual loss. The contextual loss is based on feature similarity, which ignores the spatial positions of the features. In contrast, the VTON-MP prevents excessive uneven stretching to preserve the naturalness of the garments during distortion, enabling it to focus on the spatial relationships between body parts after garment distortion and pose transformation; this results in a low global IS score. The qualitative visual results obtained with and without the contextual loss are also provided in Figure 11.

In terms of garment distortion, the qualitative visual results obtained with and without the image-semantic correspondence module (ISCM) or the flow interval restriction are also provided in Figure 10. It is clear that compared with the TB-VTON [13], our method retains better garment details in cases with covered arms. The distorted garment is more suitable with respect to posture and is more similar to the ground truth. In Figure 10, we can see that without the ISCM, the distortion task cannot be completed. Due to the various parts of the source human image, the semantic map cannot be associated. This results in a strict limitation on the geometric transformation range corresponding to the source human image during the transformation process, which prevents the completion of the distortion task. In addition, we can see that without the flow interval restriction, although the details, such as garment patterns are retained, the details are overstretched and look unnatural.

Additionally, the other characteristic parts of the image (e.g., head, hands, arms, and trousers) can be retained. After removing each module, we compare the resulting visual effects, as shown in Figure 11. It can be seen in the figure that without the SAIN module, the face's characteristics are badly distorted with incomplete or blurred generation results for the hands, arms, and other parts. This indicates that SAIN is crucial for enhancing the structural information of various body regions. Moreover, without the FEB module, the facial

TABLE II
ABLATION STUDY OF OUR FRAMEWORK

	IS↑	SSIM↑	PSNR↑	FID↓
w/o **FEB**	2.9587±0.1894	0.7737	19.5557	13.5334
w/o **SAIN**	2.9151±0.1974	0.7561	18.8569	19.6676
w/o $\mathcal{L}_{Rec}$	2.9542±0.1934	0.7785	19.6299	13.0465
w/o $\mathcal{L}_{CX}$	**3.0205±0.1300**	0.7707	19.3102	14.1618
Ours	2.9818±0.1831	**0.7842**	**20.1536**	**11.9856**
Real	3.0410±0.1435	1	N/A	0

contours are not smooth, and the details of the facial features are unnatural. This means that it is essential to use the appearance flow to filter the potentially useless sampled pixel weights from the original image feature map for enhancing the latent target detail weights. Furthermore, it can be seen in the figure that without the reconstruction loss, the image's characteristics generate blurred and unclear results. The reconstruction loss is introduced to generate a clear image that is close to the actual image. The results in Figure 11 show that the image generated by the whole model is the most complete version and is closest to the actual image with respect to other body parts.

The qualitative and quantitative evaluations conducted above demonstrate that our method performs better than the existing state-of-the-art methods in the multi-pose virtual try-on task. Generally, by adding the FEB, the feature weights without gains are effectively suppressed, and the features of valuable pixels are enhanced. SAIN can control the generation of detailed spatial relationships, such as those between human faces and arms. $\mathcal{L}_{Rec}$ effectively enhances the pixel-level details of the generated results. $\mathcal{L}_{CX}$ can maintain the realism of the generated images by comparing the feature contexts of the generated images and the ground truth.

IV. CONCLUSION

In this paper, a novel multi-stage framework is proposed to implement a multi-pose virtual try-on, which is called VTON-MP. The VTON-MP algorithm introduces an appearance flow to distort the garment and transfer pose. This allows the garment to be distorted in a way that retains more detail under the condition of the covered body. Additionally, the pose is allowed to be transformed in a way that better retains detail in the non-target garment area, e.g., heads, hands, arms, trousers, and garment features (patterns and logos). Second, we introduce an ISCM and a flow interval restriction, which keep the garment and the human body closer to reality. Finally, we designed a FEB and introduced SAIN to refine the generated image and make it clearer and more realistic. Experimental results show that VTON-MP achieves better than state-of-the-art methods in quantitative and qualitative evaluation performance.

Our approach has some limitations. Due to the diversity of garments, it is not currently possible to adapt to all types of clothing, such as types of V-neck dresses, and intricate decorative items. Then the virtual try-on scene is relatively simple and clean, and this is something that needs to be progressed. In the future, we will further improve the multi-adaptability of our approach and explore high-performance, low-volume multi-pose virtual try-on solutions.

ACKNOWLEDGMENT

The image analysis is the key technology of the manuscript, which is the research content of Open project of engineering research center of Hubei province for clothing information (No. 2022HBCI01).

REFERENCES

[1] D. Jo and G. J. Kim, "ARIoT: Scalable augmented reality framework for interacting with Internet of Things appliances everywhere," *IEEE Trans. Consum. Electron.*, vol. 62, no. 3, pp. 334–340, Aug. 2016.
[2] Q. Ma et al., "Learning to dress 3D people in generative clothing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6468–6477.
[3] A. Mir, T. Alldieck, and G. Pons-Moll, "Learning to transfer texture from clothing images to 3D humans," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7021–7032.
[4] H. Zhu et al., "Deep Fashion3D: A dataset and benchmark for 3D garment reconstruction from single images," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 512–530.
[5] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin, "FW-GAN: Flow-navigated warping GAN for video virtual try-on," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2019, pp. 1161–1170.
[6] X. Han, X. Hu, W. Huang, and M. R. Scott, "ClothFlow: A flow-based model for clothed person generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, p. 9.
[7] C. Du et al., "VTON-HF: High fidelity virtual try-on network via semantic adaptation," in *Proc. IEEE 33rd Int. Conf. Tools Artif. Intell. (ICTAI)*, 2021, pp. 224–231.
[8] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 589–604.
[9] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo, "Towards photo-realistic virtual try-on by adaptively generating preserving image content," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7847–7856.
[10] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, J. Liu, and W.-H. Cheng, "FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 275–283.
[11] C.-L. Chou, C.-Y. Chen, C.-W. Hsieh, H.-H. Shuai, J. Liu, and W.-H. Cheng, "Template-free try-on image synthesis via semantic-guided optimization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4584–4597, Sep. 2022.
[12] H. Dong et al., "Towards multi-pose guided virtual try-on network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 9026–9035.
[13] J. Wang, T. Sha, W. Zhang, Z. Li, and T. Mei, "Down to the last detail: Virtual try-on with fine-grained details," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 466–474.
[14] B. Hu, P. Liu, Z. Zheng, and M. Ren, "SPG-VTON: Semantic prediction guidance for multi-pose virtual try-on," *IEEE Trans. Multimedia*, vol. 24, pp. 1233–1246, 2022.
[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder–decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[16] I. Demir et al., "DeepGlobe 2018: A challenge to parse the earth through satellite images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, 2018, pp. 172–181.
[17] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 286–301.
[18] K. Lv, H. Sheng, Z. Xiong, W. Li, and L. Zheng, "Pose-based view synthesis for vehicles: A perspective aware method," *IEEE Trans. Image Process.*, vol. 29, pp. 5163–5174, 2020.
[19] M. Zhai, X. Xiang, R. Zhang, N. Lv, and A. El Saddik, "Optical flow estimation using dual self-attention pyramid networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 10, pp. 3663–3674, Oct. 2020.

[20] M. Yuan and Y. Peng, "Bridge-GAN: Interpretable representation learning for text-to-image synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 11, pp. 4258–4268, Nov. 2019.
[21] Y. Xia, W. Zheng, Y. Wang, H. Yu, J. Dong, and F.-Y. Wang, "Local and global perception generative adversarial network for facial expression synthesis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1443–1452, Mar. 2022.
[22] K. Yang, D. Liu, Z. Chen, F. Wu, and W. Li, "Spatiotemporal generative adversarial network-based dynamic texture synthesis for surveillance video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 359–373, Jan. 2022.
[23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1125–1134.
[24] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8798–8807.
[25] Y. Chen, R. Xia, K. Yang, and K. Zou, "MFFN: Image super-resolution via multi-level features fusion network," *Vis. Comput.*, to be published.
[26] Y. Chen, R. Xia, K. Zou, and K. Yang, "FFTI: Image inpainting algorithm via features fusion and two-steps inpainting," *J. Vis. Commun. Image Represent.*, vol. 91, Mar. 2023, Art. no. 103776.
[27] 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.*, 2019, pp. 4401–4410.
[28] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 2256–2265.
[29] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 6840–6851.
[30] J. Wang et al., "Neural pose transfer by spatially adaptive instance normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 5830–5838.
[31] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent.*, vol. 9351, 2015, pp. 234–241.
[32] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 932–940.
[33] J. Duchon, "Splines minimizing rotation-invariant semi-norms in Sobolev spaces," in *Constructive Theory of Functions of Several Variables*. Berlin, Germany: Springer, Apr./May 1976, pp. 85–100.
[34] C. Du et al., "VTON-SCFA: A virtual try-on network based on the semantic constraints and flow alignment," *IEEE Trans. Multimedia*, vol. 25, pp. 777–791, 2022.
[35] Y. Shi, D. Deb, and A. K. Jain, "WarpGAN: Automatic caricature generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 10762–10771.
[36] D. K. Park, H. S. Yoon, and C. S. Won, "Fast object tracking in digital video," *IEEE Trans. Consum. Electron.*, vol. 46, no. 3, pp. 785–790, Jun. 2000.
[37] Y. Chen et al., "Image super-resolution reconstruction based on feature map attention mechanism," *Appl. Intell.*, vol. 51, no. 7, pp. 4367–4380, 2021.
[38] J. Zhang, J. Sun, J. Wang, Z. Li, and X. Chen, "An object tracking framework with recapture based on correlation filters and Siamese networks," *Comput. Elect. Eng.*, vol. 98, Mar. 2022, Art. no. 107730.
[39] J. Zhang, W. Feng, T. Yuan, J. Wang, and A. K. Sangaiah, "SCSTCF: Spatial-channel selection and temporal regularized correlation filters for visual tracking," *Appl. Soft Comput.*, vol. 118, Mar. 2022, Art. no. 108485.
[40] R. Xia, Y. Chen, and B. Ren, "Improved anti-occlusion object tracking algorithm using unscented Rauch–Tung–Striebel smoother and kernel correlation filter," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 8, pp. 6008–6018, 2022.
[41] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 2337–2346.
[42] K. Simonyan and A. Zisserman. "Very deep convolutional networks for large-scale image recognition." Sep. 2014. [Online]. Available: https://arxiv.org/abs/1409.1556
[43] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 768–783.
[44] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[45] 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.
[46] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. NIPS*, vol. 29, 2016, pp. 2234–2242.
[47] 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. 6627–6638.
[48] 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.

Feng Yu received the Ph.D. degree from the School of Computer Science and Technology, Huazhong University of Science and Technology. He is currently an Associate Professor with the School of Computer Science and Artificial Intelligence, Wuhan Textile University. His research interests include machine vision algorithms, artificial intelligence applications, and clothing intelligent manufacturing.

Ailing Hua received the B.E. degree in software engineering from Wuhan Textile University in 2019, where she is currently pursuing the master's degree with the School of Computer Science and Artificial Intelligence. Her research interests include image processing and machine learning.

Chenghu Du received the Bachelor of Engineering degree in computer science and technology from the Wuhan Institute of Technology in 2019, and the master's degree from the School of Computer and Artificial Intelligence, Wuhan Textile University in 2022. He is currently pursuing the Doctoral degree with the School of Science and Artificial Intelligence, Wuhan University of Technology. His research interests include image processing, computer vision, and deep learning.

Minghua Jiang received the Ph.D. degree from the School of Computer Science and Technology, Huazhong University of Science and Technology. He is currently the Vice-Chancellor with Wuhan Textile University, where he is also a Professor with the School of Computer Science and Artificial Intelligence. His research interests include computer system architecture, artificial intelligence application, and clothing intelligent manufacturing.

Xiong Wei received the Ph.D. degree in computer architecture from the National University of Defense Technology in 2011, where he was a Postdoctoral Researcher of Computer Architecture in 2011. He is currently an Associate Professor and the Vice Dean of the School of Computer Science and Artificial Intelligence, Wuhan Textile University. His research interests include storage architecture, GPU, and parallel algorithm.

Tao Peng received the M.Sc. and Ph.D. degrees in computer science from the Huazhong University of Science and Technology in 2006 and 2011, respectively. He is currently an Associate Professor with the School of Computer Science and Artificial Intelligence, Wuhan Textile University. His research interests include data mining, pattern recognition, and network security.

Lijun Xu received the Ph.D. degree from the School of Computer Science and Technology, Huazhong University of Science and Technology. She is currently working as a full-time Teacher of Software Engineering with the School of Computer and Information Engineering, Hubei University. Her research interests include machine vision, digital twinning, image processing and analysis, machine learning, medical big data analysis, and auxiliary diagnosis.

Xinrong Hu was born in 1973. She received the Ph.D. degree from the Institute for Pattern Recognition and Artificial Intelligence, Huazhong University of Science and Technology in 2008. She is currently a Professor and the Dean of the School of Computer Science and Artificial Intelligence, Wuhan Textile University. Her research interests include image processing, virtual reality technology, and computer vision.