IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received March 28, 2022, accepted April 10, 2022, date of publication April 14, 2022, date of current version April 21, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3167509

A Flow-Based Generative Network for Photo-Realistic Virtual Try-On

TAO WANG, XIAOLING GU, (Member, IEEE), AND JUNKAI ZHU

Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310005, China

Corresponding author: Xiaoling Gu (guxl@hdu.edu.cn)

This work was supported in part by the Zhejiang Provincial Natural Science Foundation of China under Grant LY21F020019, and in part by the National Science Foundation of China under Grant 61802100.

ABSTRACT Image-based virtual try-on systems aim at transferring the try-on clothes onto a target person. Despite making considerable progress recently, such systems are still highly challenging for real-world applications because of occlusion and drastic spatial deformation. To address the issues, we propose a novel Flow-based Virtual Try-on Network (FVTN). It consists of three modules. Firstly, the Parsing Alignment Module (PAM) aligns the source clothing to the target person at the semantic level by predicting a semantic parsing map. Secondly, the Flow Estimation Module (FEM) learns a robust clothing deformation model by estimating multi-scale dense flow fields in an unsupervised fashion. Thirdly, the Fusion and Rendering Module (FRM) synthesizes the final try-on image by effectively integrating the warped clothing features and human body features. Extensive experiments on a public fashion dataset demonstrate that our FVTN qualitatively and quantitatively outperforms the state-of-the-art approaches. The source code and trained models are available at https://github.com/gxl-groups/FVNT.

INDEX TERMS Image-based virtual try-on, image synthesis, appearance flow.

I. INTRODUCTION

As online shopping has continued to grow in popularity, virtually trying on clothes in an online fitting room has achieved much attention in recent years. A photo-realistic virtual try-on system will not only enhance the user shopping experience by fitting different clothes without changing them physically but also improve sales for retailers. This motivates many companies to develop various virtual fitting technologies, such as SenseMi,[1] triMirror,[2] etc.

Classical virtual try-on methods primarily rely on computer graphics to synthesize the try-on looks for users based on their 3D body shapes, desired poses and target clothing items [2], [28], [44], which can well control clothing deformation and material performance. However, the huge labor costs for 3D data annotation and upfront costs for scanning equipment inhibit their large-scale deployment [14].

Motivated by the rapid development of image synthesis methods [10]–[12], [19], [21], the image-based virtual try-on methods using generative models provide a more economical solution, the goal of which is to naturally warp the try-on clothes on a target person without leveraging any 3D information. Although image-based virtual try-on makes considerable progress recently, generating perceptually convincing virtual try-on images is highly challenging for the real-world scenario. The main challenges lie in: (1) Occlusion occurs in the target person. For example, the target person's arms may cross over the chest and occlude the clothing region. (2) Varying deformation exists among different human poses and shapes of the target person (e.g., limbs from non-overlapping to overlapping), which makes it extremely hard to deform the garments and well fit the posture and body shape of the target person. (3) Due to the drastic spatial deformation from source clothing to the target person, generating the try-on image that maintains the detailed visual features of the original garment such as texture and color is a non-trivial task.

The aforementioned challenges can be ultimately summed up as two key problems for tackling image-based virtual try-on. The first one is how to design a robust geometric deformation scheme to warp the source clothing for fitting the target person? Existing approaches relying on the affine

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

[1]https://viubox.com/

[2]https://www.trimirror.com/

or TPS transformation for warping the source clothing [6], [14], [36], [43] fail to generate precise appearance details because such methods cannot deal with the transformations of non-rigid objects such as clothes. Recently, flow-based methods [13], [30], [46] show advantages in learning complex non-rigid geometric deformation in comparison to the affine transformation approaches. Inspired by it, we propose a novel flow-based spatial alignment scheme for precisely capturing the clothing deformation. The second one is how to render the final try-on image by effectively fusing the contents of body parts and warped clothes? The quality of try-on look highly depends on the appearances of garments (e.g. texture, logo and color) as well as the characteristics of the target person (e.g. hair, face and arms). Previous approaches [14], [36] using a composition mask to integrate clothing and human body bring obvious boundary artifacts in the intersection regions. These approaches overlook the occluded regions and fail to synthesize the body parts flexibly.

In this paper, we propose a novel Flow-based Virtual Try-on Network (FVTN), which consists of three modules. The first one is the Parsing Alignment Module (PAM), aligning the source clothing to the posture of the target person at the semantic level. This module provides accurate spatial information for subsequent modules. The second one is the Flow Estimation Module (FEM), which learns clothing deformation by estimating multi-scale dense flow fields in an unsupervised fashion. The predicted multi-scale flows are used to establish visual correspondence between the source clothing and the target try-on clothing in the feature domain. The learned flow fields do not directly warp the source clothing at the pixel level but the feature level. This is because warping clothing at the pixel level would result in the model having difficulty extracting large motions and generating new contents [30]. The final part is the Fusion and Rendering Module (FRM), aligning the source clothing to the target person at the pixel level. By effectively integrating the warped source clothing features and the body features, the proposed FRM can generate accurate clothing appearances and fine details of the human body. Experiments on the VITON dataset [14] demonstrate that the proposed FVTN can produce photo-realistic and perceptually convincing try-on images.

The main contributions of our work can be summarized as follows:

- We propose a new flow-based generative network with three tailored modules for image-based virtual try-on.
- We design a novel spatial alignment scheme in the flow estimation module to precisely capture clothing deformation by estimating multi-scale dense flow fields in an unsupervised fashion.
- We propose a novel image synthesis network to synthesize the final try-on images by integrating the warped clothing features and the body features.
- Experimental results on VITON [14] verify that our method qualitatively and quantitatively outperforms the state-of-the-art methods.

II. RELATED WORK

A. VIRTUAL TRY-ON

Conventional approaches for virtual try-on works are based on graphics models. For instance, Sekine *et al.* [32] introduced a virtual fitting system that adjusts 2D clothing images to users by estimating their 3D body shape models from single-shot depth images. Yang *et al.* [41] computed a 3D model of a human body and outfits from a single-view image. Pons-Moll *et al.* [28] used a multi-cloth 3D model of the body and clothing for capturing a clothed person in motion and retargeting the clothing to new body shapes. Patel *et al.* [27] proposed TailorNet for estimating clothing deformation in 3D as a function of three factors: body shape, body pose and garment style. Mir [26] proposed Pix2Surf to digitally map the texture of clothing images to the 3D surface of virtual garment items, which enables 3D virtual try-on in real-time. 3D methods can generate good results for virtual try-on, but usually, they require additional 3D measurements.

Compared to graphics models, image-based generative models are more computationally efficient and broadly applicable. For example, VITON [14] first proposed image-based virtual try-on method, which generates warped clothes using Thin Plate Spline (TPS) transformation and maps the texture to the refined result with a composition mask. CP-VTON [36] improves VITON by using neural networks to directly learn the parameters of TPS for clothing warping, and thus achieves more accurate alignment results. CP-VTON+ [25] outperforms CP-VTON by improving the clothing warping stage and blending stage. VTNFP [43] achieves better try-on results than CP-VTON and VITON by concatenating the high-level features extracted from the body parts and the bottom garment, since CP-VTON and VITON only focus on the upper garment. ACGPN [40] synthesizes try-on images preserving both the characteristics of clothes and details of the human identity by using three modules. Xintong *et al.* proposed ClothFlow [13] for handling pose-guided synthesis and image-based virtual try-on. Similarly, ClothFlow and our proposed FVTN both learn clothing deformation by using flow-based methods. However, different from ClothFlow, we leverage an unsupervised flow training scheme relying on the photometric loss [42]. Furthermore, our FVTN uses the learned flow to warp the garments at the feature level instead of at the pixel level for extracting the large motions and generate new contents.

B. OPTICAL FLOW

Optical flow [16], [17] is the task of estimating dense pixel-to-pixel correspondence between two input images, which is widely used in many applications such as action recognition, motion tracking, video segmentation and 3D reconstruction. Optical flow has traditionally been approached as a hand-crafted optimization problem, the objective of which is defined as a trade-off between a data term and a regularization term [1]. Recently, deep learning has been shown as a promising alternative to traditional methods. FlowNet [7] is the first trainable CNN for optical flow estimation.

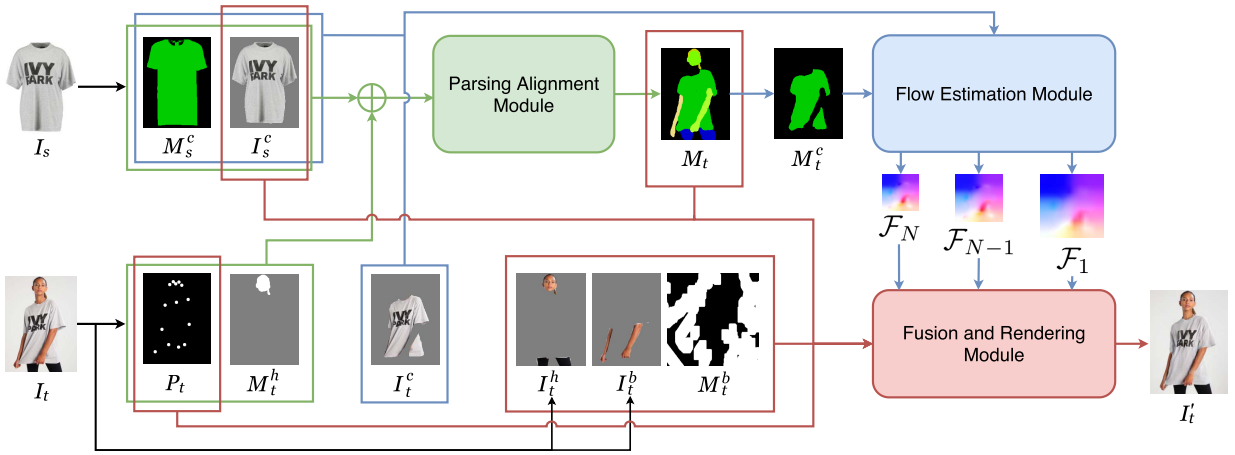

FIGURE 1. The overall network architecture of the proposed FVTN, which consists of three modules. Given the source clothing image I_s and the target person image I_t, Parsing Alignment Module (PAM) first aligns the source clothing to the posture of the target person at the semantic level by predicting the target semantic parsing map M_t. With the predicted parsing map M_t, we obtain the semantic mask of target clothing M_t^c. Based on M_s^c (the semantic mask of source clothing) and M_t^c, Flow Estimation Module (FEM) learns clothing deformation by estimating multi-scale flow fields $\{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N\}$ in an unsupervised fashion. Relying on the estimated multi-scale flow fields, Fusion and Rendering Module (FRM) renders the final try-on image I_t' by effectively integrating the features of warped clothes and features of the human body.

FlowNet2 [18] improves the flow accuracy of FlowNet by cascading several variants of it. Subsequently, Ranjan and Black introduced SpyNet [29], a compact spatial image pyramid network, which warps images at multiple scales to cope with large displacements. Recent notable contributions to end-to-end trainable optical flow include PWC-Net [34] and LiteFlowNet [18]. They proposed to use the feature warping and cost volume at multiple pyramid levels in a coarse-to-fine estimation, yielding more compact and effective networks. We draw inspiration from those coarse-to-fine flow estimation methods.

To avoid annotating labels, Meister *et al.* [24] proposed an end-to-end unsupervised learning approach by designing a bidirectional flow-based loss function. Wang *et al.* [38] further proposed an unsupervised learning framework that models occlusion and large motions. Liu *et al.* [23] proposed SelFlow that distills reliable flow estimations from non-occluded pixels using self-supervised training. Unsupervised optical flow estimation is closer to our setting. However, different from these works, we focus on learning a flow for establishing correspondence between the source clothing and target try-on clothing.

III. PROPOSED METHOD

As shown in Fig. 1, our FVTN is composed of three modules. The first one is the Parsing Alignment Module (PAM), which transfers the source clothing onto the target person at the semantic level. The proposed PAM provides accurate spatial and semantic information for subsequent modules. The second one is the Flow Estimation Module (FEM), which learns diverse spatial deformation between the source clothing and the target try-on clothing by estimating multi-scale dense flow fields in an unsupervised way. The final part is the Fusion and Rendering Module (FRM), which fuses the warped features of source clothing and the features of the human body for rendering the final try-on image. Fig. 2 illustrates details of these three modules.

Ideally, we need an image triplet $\langle I_s, I_t, I_r \rangle$ to train the FVTN, where I_s is the source clothing image, I_t is the target person image, and I_r stands for the ground-truth image. However, such a dataset is hard to obtain. Therefore, I_r is replaced with I_t to train the FVTN in our implementation.

A. PARSING ALIGNMENT MODULE (PAM)

To disentangle the generation of shape and appearance, PAM aligns the source clothing I_s to the target person I_t at the semantic level. It takes the semantic mask of source clothing M_s^c, the segmented source clothing I_s^c, the pose of the target person P_t and the binary mask of the target person's head M_t^h as input to predict the target semantic parsing map M_t. The predicted parsing map is required to retain the body parts and the pose of the target person as well as accurately show the shapes and categories of the transformed source clothing.

We use a human parser [9] to compute the parsing map with 20 semantic labels for I_s and I_t. Each parsing map is represented as a one-hot tensor with 20 channels. On the other hand, we use a state-of-the-art pose estimator [3] to estimate the pose of the target person. Following [36], P_t is represented as 18-channels heat maps that each one encodes one joint of a human body.

In this module, we simply adopt a conditional generative adversarial network [37], in which a U-Net structure is used as the generator while a discriminator is utilized to distinguish generated parsing map from the ground-truth parsing map. The overall objective function for PAM is formulated as:

$$\mathcal{L}_{\text{PAM}} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{seg}\mathcal{L}_{seg} \tag{1}$$

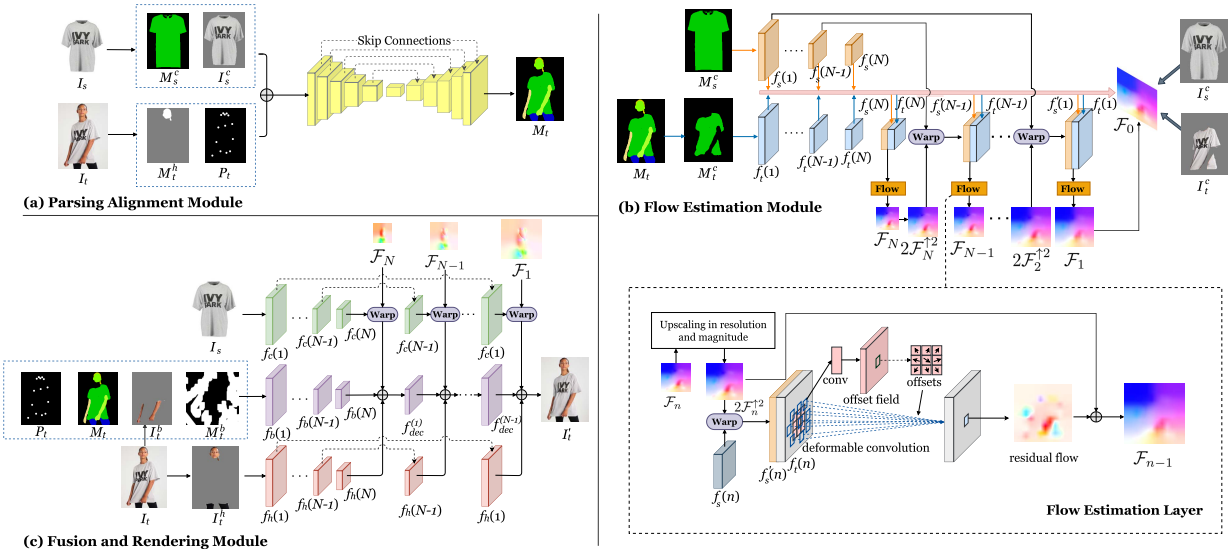

FIGURE 2. (a) Parsing Alignment Module (PAM). (b) Flow Estimation Module (FEM). (c) Fusion and Rendering Module (FRM).

where $\mathcal{L}_{adv}$ is the adversarial loss [37] and $\mathcal{L}_{seg}$ is the pixel-wise cross-entropy loss. λ_{adv} and λ_{seg} are the trade-off parameters for these two loss terms, which are set to 0.2 and 1, respectively, in our experiments.

The pixel-wise cross-entropy loss $\mathcal{L}_{seg}$ constrains pixel-level accuracy during semantic parsing map generation, which is defined as:

$$\mathcal{L}_{seg}(M_t, \tilde{M}_t) = -\frac{1}{HW}\sum_{m=1}^{HW}\sum_{c=1}^{C}\tilde{M}_t \log(M_t) \tag{2}$$

where H, W and C are height, width and the number of channels of the parsing map, respectively. M_t is the generated parsing map and $\tilde{M}_t$ is the ground-truth.

B. FLOW ESTIMATION MODULE (FEM)

As we've discussed, building a robust clothing deformation model is crucial for image-based virtual try-on. Early methods of image-based virtual try-on [14], [25], [36], [43] warp clothes by computing a Thin Plane Spline (TPS) transformation. However, because of its low degree of freedom, TPS transformation can only model limited geometric transformations and is inflexible to achieve complex and non-rigid deformation [46]. Considering that flow-based methods can capture complex non-rigid geometric deformation [13], [30], [46], we design an unsupervised flow-based clothing deformation scheme without using explicit correspondence annotation.

With the predicted parsing map M_t, we first obtain the semantic mask of target clothing M_t^c. FEM takes M_s^c and M_t^c as input to predict multi-scale dense flow fields for establishing visual correspondence between the source clothing and the target try-on clothing in the feature domain. To deal with the drastic spatial deformation existing between the source clothing and the target try-on clothing, we estimate the flow fields in an iterative manner, where the flow is first estimated at low resolution followed by upsampled and refined at high resolution.

Specially, we deploy a two-stream weight-sharing Feature Pyramid Network (FPN) to extract two feature pyramids from M_s^c and M_t^c, that is, $\{f_s(1), f_s(2),\ldots, f_s(N)\}$ and $\{f_t(1), f_t(2),\ldots, f_t(N)\}$, respectively, where N corresponds to the lowest spatial resolution (in our case $N = 5$) and 1 corresponds to the highest spatial resolution. The extracted multi-scale features will be used to estimate the flow from the source clothing to the target one in an unsupervised way. Beginning with the lowest spatial resolution, after concatenating $f_s(N)$ and $f_t(N)$, a flow estimation layer initially infers a coarse flow $\mathcal{F}_N$. Formally,

$$\mathcal{F}_N = \mathrm{Def}\big(f_s(N), f_t(N)\big) \tag{3}$$

where Def denotes the deformable convolution [4] layer. We replace the standard convolution with the deformable convolution in the flow estimation layer for improving the network's ability to handle drastic spatial deformation, since the standard convolution is limited by the lack of ability to spatially transform the inputs [4].

At a higher spatial resolution, the flow estimation layer gets a refined flow $\mathcal{F}_{N-1}$ by computing a residue flow $\mathcal{R}_{N-1}$ and adding the upscaled flow field $\mathcal{F}_N^{\uparrow 2}$, as illustrate in Fig. 2(b). $\mathcal{R}_{N-1}$ is computed based on the upscaled flow field $\mathcal{F}_N^{\uparrow 2}$, $f_s(N-1)$ and $f_t(N-1)$. Formally,

$$\mathcal{F}_{N-1} = \mathcal{R}_{N-1} + 2\mathcal{F}_N^{\uparrow 2} \tag{4}$$

$$\mathcal{R}_{N-1} = \mathrm{Def}\Big(\mathcal{W}\big(f_s(N-1), 2\mathcal{F}_N^{\uparrow 2}\big), f_t(N-1)\Big) \tag{5}$$

where $\mathcal{W}$ is warping operation with bilinear interpolation when the flow field falls into a sub-pixel coordinate. This allows end-to-end training via stochastic gradient

descent [47]. Note that the resolution of $\mathcal{F}_N$ is upsampled with bilinear interpolation and its value is doubled. Such process will be repeated until inferring the finest flow $\mathcal{F}_1$ from the two pyramial features with the highest spatial resolution $f_s(1)$ and $f_t(1)$. Finally, $\mathcal{F}_1$ is upsampled and its value is doubled to $\mathcal{F}_0 = 2\mathcal{F}_1^{\uparrow 2}$. Inspired by [35], the flow estimation layers share weights between iterations for speeding up model training and reducing the amount of model parameters.

Since we do not have the ground-truth flow, we leverage an unsupervised flow training scheme relying on the photometric loss [42] with the clothing images. The overall objective function for FEM is formulated as:

$$\mathcal{L}_{\mathsf{FEM}} = \lambda_{phot}\mathcal{L}_{phot} + \lambda_{TV}\mathcal{L}_{TV} + +\lambda_{perc_1}\mathcal{L}_{perc_1} \tag{6}$$

where $\mathcal{L}_{phot}$ is a multi-scale photometric loss, $\mathcal{L}_{TV}$ is a flow regularization loss and $\mathcal{L}_{perc_1}$ is the perceptual loss [20]. λ_{phot}, λ_{TV} and λ_{perc_1} are the trade-off parameters for these three loss terms, which are set to 5, 2 and 1, respectively, in our experiments.

The multi-scale photometric loss $\mathcal{L}_{phot}$ sums the photometric loss between the source clothing regions and the target one at multiple scales for fast convergence [46], which is defined as:

$$\mathcal{L}_{phot}(I_t^c, I_s^c, \mathcal{F}) = \sum_{i=0}^{N} ||\rho(I_t^c(i) - \mathcal{W}(I_s^c(i), \mathcal{F}_i))||_1 \tag{7}$$

where $\rho(x) = (x^2 + \epsilon^2)^\alpha$ is a penalty function for mitigating the effects of outliers [42]. I_t^c and I_s^c is the segment image with the clothing regions of the target clothing and the source clothing, respectively. And i represents the spatial resolution of images and flows. Note that $I_t^c(i)$, $I_s^c(i)$ and $\mathcal{F}_i$ have the same spatial resolution.

The flow regularization loss $\mathcal{L}_{TV}$ is a total variation-based (TV) smoothness penalty term to regularize the flow prediction, which is defined as:

$$\mathcal{L}_{TV}(\mathcal{F}_N, \mathcal{R}) = ||\Delta\mathcal{F}_N||_1 + \sum_{i=1}^{N-1} ||\Delta\mathcal{R}_i||_1 \tag{8}$$

Unlike previous methods [13], [46] that regularize the multi-scale flows, we apply smoothness loss on the coarse flow $\mathcal{F}_N$ and the multi-scale residue flows.

In order to preserve realistic details and textures of source clothing, we add the perceptual loss between I_t^c and the warped source clothing segment image (i.e., $W(I_s^c) = \mathcal{W}(I_s^c, \mathcal{F}_0)$). Specifically, the perceptual loss $\mathcal{L}_{perc_1}$ models the distance between I_t^c and $W(I_s^c)$ in a feature space, which is defined as:

$$L_{perc_1}(I_t^c, W(I_s^c)) = \sum_{i=0}^{N_l} \beta_i ||\phi_i(I_t^c) - \phi_i(W(I_s^c))||_1 \tag{9}$$

where N_l is the number of chosen layers. And $\phi_i(I_t^c)$ denotes the feature map of image I_t^c at the i-th layer in a VGG-19 [33] network pre-trained on ImageNet [5]. β_i is the hyper-parameter that controls the contributions of different layers and is set by following [14].

C. FUSION AND RENDERING MODULE (FRM)

Going beyond the clothing deformation model, it is another great challenge to render the final try-on image by fusing the contents of the human body and warped clothes. FRM accepts the segmented source clothing I_s^c, the body parts of the target person I_t^b, the head region of the target person I_t^h, the parsing map of the target person M_t and the pose of the target person P_t as input to synthesize the photo-realistic try-on image I_t'.

Specifically, FRM adopts three encoders of the same architecture, i.e., $\mathsf{ENC_c}$, $\mathsf{ENC_b}$ and $\mathsf{ENC_h}$, to encode the features for the source clothing, the body of the target person, the head of the target person, respectively. Note that the three encoders do not share weights during training. $\mathsf{ENC_c}$ extracts source clothing features from I_s^c through N downsampling layers. Formally,

$$\mathsf{ENC_c}(I_s^c) = \{f_c(1), f_c(2), \ldots, f_c(N)\} \tag{10}$$

where $f_c(n), n = 1, \ldots, N$ denotes the extracted clothing features after n downsampling layers.

$\mathsf{ENC_h}$ extracts the head features of the target person separately from the body parts to enhance the siginals of facial features and hair features. Formally,

$$\mathsf{ENC_h}(I_t^h) = \{f_h(1), f_h(2), \ldots, f_h(N)\} \tag{11}$$

where $f_h(n), n = 1, \ldots, N$ denotes the extracted head features after n downsampling layers.

Likewise, $\mathsf{ENC_b}$ extracts the body features of the target person conditioned on I_t^b, M_t and P_t,

$$\mathsf{ENC_b}(I_t^b, M_t, P_t) = \{f_b(1), f_b(2), \ldots, f_b(N)\} \tag{12}$$

where $f_b(n), n = 1, \ldots, N$ denotes the extracted body features after n downsampling layers. With the removed clothing and head regions, the arms and legs are represented as body parts.

Next, the final try-on result I_t' is generated through N decoding blocks, where each decoding block accepts the concatenated warped clothing features, body features and head features. Formally,

$$\begin{aligned} f_{\text{dec}}^{(1)} &= \mathsf{DEC}\big(f_b(N), \mathcal{W}(f_c(N), \mathcal{F}_N), f_h(N)\big) \\ f_{\text{dec}}^{(2)} &= \mathsf{DEC}\big(f_{\text{dec}}^{(1)}, \mathcal{W}(f_c(N-1), \mathcal{F}_{N-1}), f_h(N-1)\big) \\ &\ldots \\ f_{\text{dec}}^{(N)} &= \mathsf{DEC}\big(f_{\text{dec}}^{(N-1)}, \mathcal{W}(f_c(1), \mathcal{F}_1), f_h(1)\big) \end{aligned} \tag{13}$$

where the clothing feature $f_c(n)$ is warped via the predicted flow $\mathcal{F}_n$ from previous module. Note that $f_c(n)$ and $\mathcal{F}_n$ is ensured to have the same spatial resolution. Tanh function is applied after $f_{\text{dec}}^{(N)}$ to generate the normalized image, i.e. $I_t' = tanh(f_{\text{dec}}^{(N)})$.

Besides, to enhance the robustness of the mapping in FRM, we introduce a body mask M_t^b for I_t^b to randomly remove some regions of the human body,

$$I_t^b = (1 - M_t^b) * I_t^b \tag{14}$$

where M_t^b is sampled from the Irregular Mask Dataset [22]. Without M_t^b, FRM tends to learn an identity mapping for the body parts (i.e., arms and legs). For example, when transferring a long-sleeve garment to the target person in a short-sleeve one, the arm parts should be rendered with clothing textures instead of retaining the original arms. In the opposite case, when transferring a short-sleeve garment to the target person in a long-sleeve one, the arm parts should be synthesized instead of retaining the original clothing textures. By introducing M_t^b, FRM can adaptively determine the generation or preservation of the body parts.

The overall objective function for FRM is formulated as:

$$\mathcal{L}_{\mathsf{FRM}} = \lambda_{L1}\mathcal{L}_{L1} + \lambda_{perc_2}\mathcal{L}_{perc_2} + \lambda_{sty}\mathcal{L}_{style} \tag{15}$$

where $\mathcal{L}_{L1}$ is the reconstruction loss, $\mathcal{L}_{perc_2}$ is the perceptual loss [20] and $\mathcal{L}_{style}$ is the style loss [8]. λ_{L1}, λ_{perc_2} and λ_{sty} are the trade-off parameters for these three loss terms, which are set to 1, 1 and 400, respectively, in our experiments.

The reconstruction loss is the $L1$ loss between the synthesized image I_t' and the ground-truth image I_r, which is defined as:

$$\mathcal{L}_{L1}(I_t', I_r) = ||I_t' - I_r||_1 \tag{16}$$

The perceptual loss $\mathcal{L}_{perc_2}$ models the distance between the synthesized image I_t' and the ground-truth image I_r in a feature space, i.e., $L_{perc_2}(I_t', I_r) = \sum_{i=0}^{N_l} \beta_i ||\phi_i(I_t') - \phi_i(I_r)||_1$.

$\mathcal{L}_{style}$ is the style loss that matches style information between the synthesized image and the ground-truth image, which is defined as:

$$L_{style}(I_t', I_r) = \sum_{i=1}^{N_l} \gamma_i ||\psi_i(I_t') - \psi_i(I_r)||_1 \tag{17}$$

where $\psi_i(I_t')$ denotes the Gram matrix [8] of image I_t' at the i-th layer in a VGG-19 [33] network pre-trained on ImageNet [5]. γ_i is the hyper-parameter that controls the contributions of different layers and is set by following [13].

IV. EXPERIMENT

A. IMPLEMENTATION DETAILS

1) DATASET

We conduct the experiments on the VITON dataset [14] to evaluate the proposed FVTN on virtual try-on task. We follow [36] to split 16,253 image pairs into a training set and a validation set with 14,221 and 2,032 pairs, respectively. Each image pair includes a front-view woman image and a top clothing image. Resolution for all images in VITON is 256 × 192. Note that the images of the validation set are rearranged into unmatched pairs as the test set.

2) TRAINING DETAILS

We train three modules of the proposed FVTN separately. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$ in all experiments. PAM is trained with a minibatch of size 16 and a learning rate of 0.0001 for 15 epochs. FEM is trained with a minibatch of size 8 and a learning rate of 0.00005 for 20 epochs. In FEM, the FPN consists of five encoding layers where each layer is a convolution layer with kernel 3 and stride 2 followed by one residual block. Besides, we use a deformable convolution with kernel 3 and stride 1 as the flow estimation layer. FRM is trained with a minibatch of size 8 and a learning rate of 0.0001 for 20 epochs. In FRM, three encoders share the same architecture, consisting of five downsampling layers where each layer contains two convolution layers with kernel 3 and stride 2 and with kernel 3 and stride 1, respectively.

TABLE 1. Quantitative comparison results of our proposed method and the baselines on the virtual try-on task. A higher score of SSIM/IS is better. A lower score of FID/LPIPS is better.

	SSIM ↑	FID ↓	LPIPS ↓	IS ↑
CP-VTON [36]	0.820	14.60	0.1816	2.76
CP-VTON+ [25]	0.829	13.10	0.1751	**2.77**
ClothFlow [13]	0.838	11.90	0.1595	2.61
ACGPN [40]	**0.855**	10.62	0.1535	2.68
w.o/iterat	0.847	8.38	0.1343	2.61
w.o/deconv	0.848	8.37	0.1360	2.66
w.o/multi	0.848	8.16	0.1327	2.68
w.o/fwarp	0.849	8.04	0.1309	2.69
w.o/mask	0.848	8.43	0.1347	2.69
Our Method	0.850	**7.96**	**0.1292**	2.70

3) EVALUATION METRICS

We adopt four widely used evaluation metrics to evaluate the quality of the synthesized images. Inception Score (**IS**) [31] is used to measure the quality and diversity of the generated images. Structural Similarity (**SSIM**) [39] is used to measure the similarity between the generated images and ground-truth images. Fréchet Inception Distance (**FID**) [15] is used to measure the realism of the generated images by computing the Wasserstein-2 distance between distributions of the generated images and ground-truth images. Learned Perceptual Image Patch Similarity (**LPIPS**) [45] is used to measure how similar are two images by computing the distance between the generated images and generated images at the perceptual domain.

B. EVALUATIONS

We mainly perform visual comparison of our method with recent proposed virtual try-on networks [13], [25], [36], [40].

1) QUANTITATIVE RESULTS

Table 1 reports the quantitative comparison between our approach and the baselines. Except for SSIM and IS, our method significantly outperforms all baselines on FID and LPIPS. The IS metric provides a proxy to evaluate the performance but it is not a good measurement of how well the model is performing our task. Although our method gets the second-highest SSIM score, the FID and LPIPS

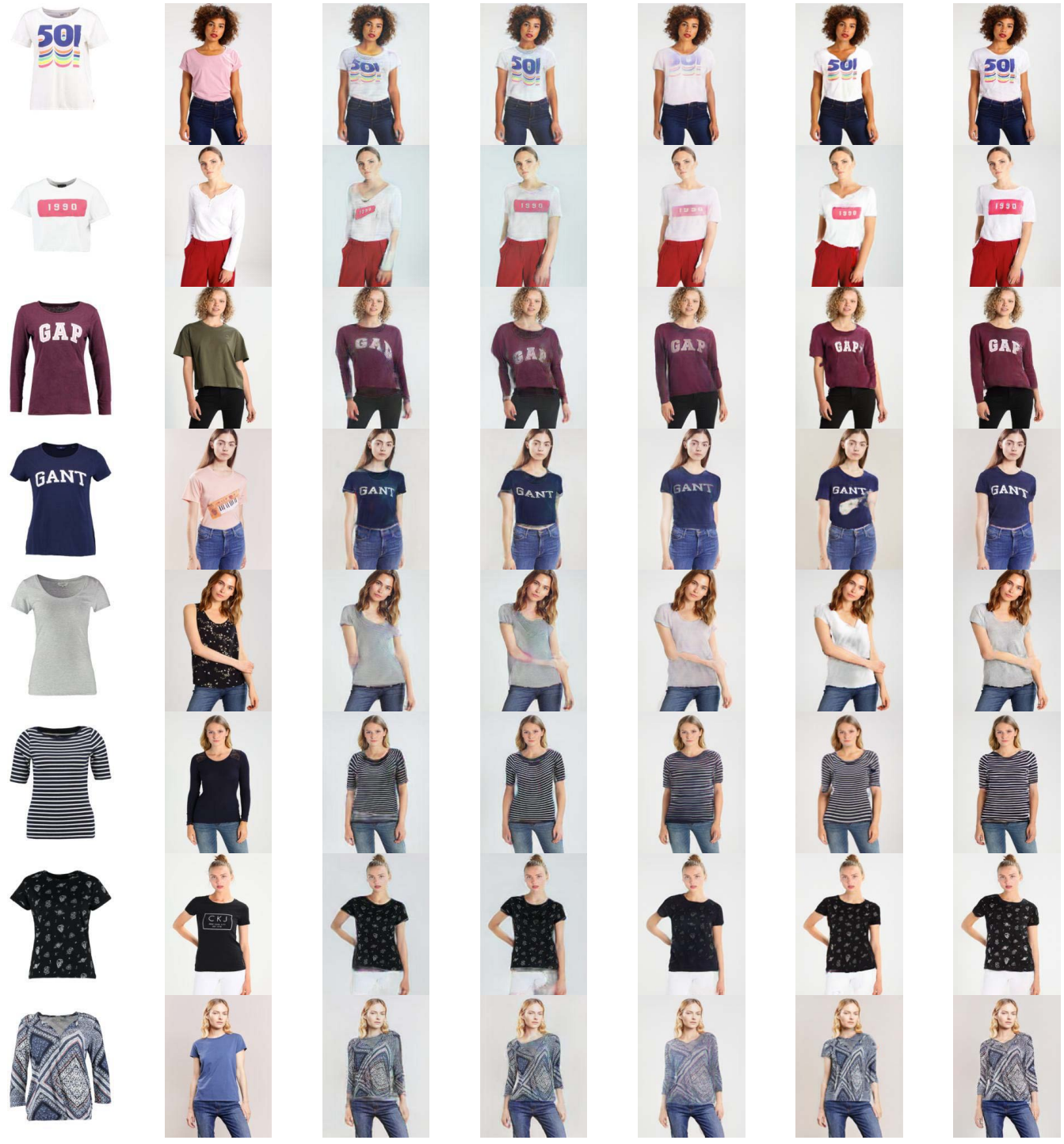

FIGURE 3. **Comparison visual results of different approaches on the virtual try-on task.**

more accurately reflect the similarity between the synthesized images and the ground-truth images.

2) QUALITATIVE RESULTS

Figure 3 presents a visual comparison of the evaluated methods, where the first column is the garment image, the second column is the target person image and the other columns are the synthesized virtual try-on images with different approaches. We observe that CP-VTON produces the worst visual effects, which fails to handle clothing deformation and generate the fine details of the human body. This observation verifies the inefficiency of

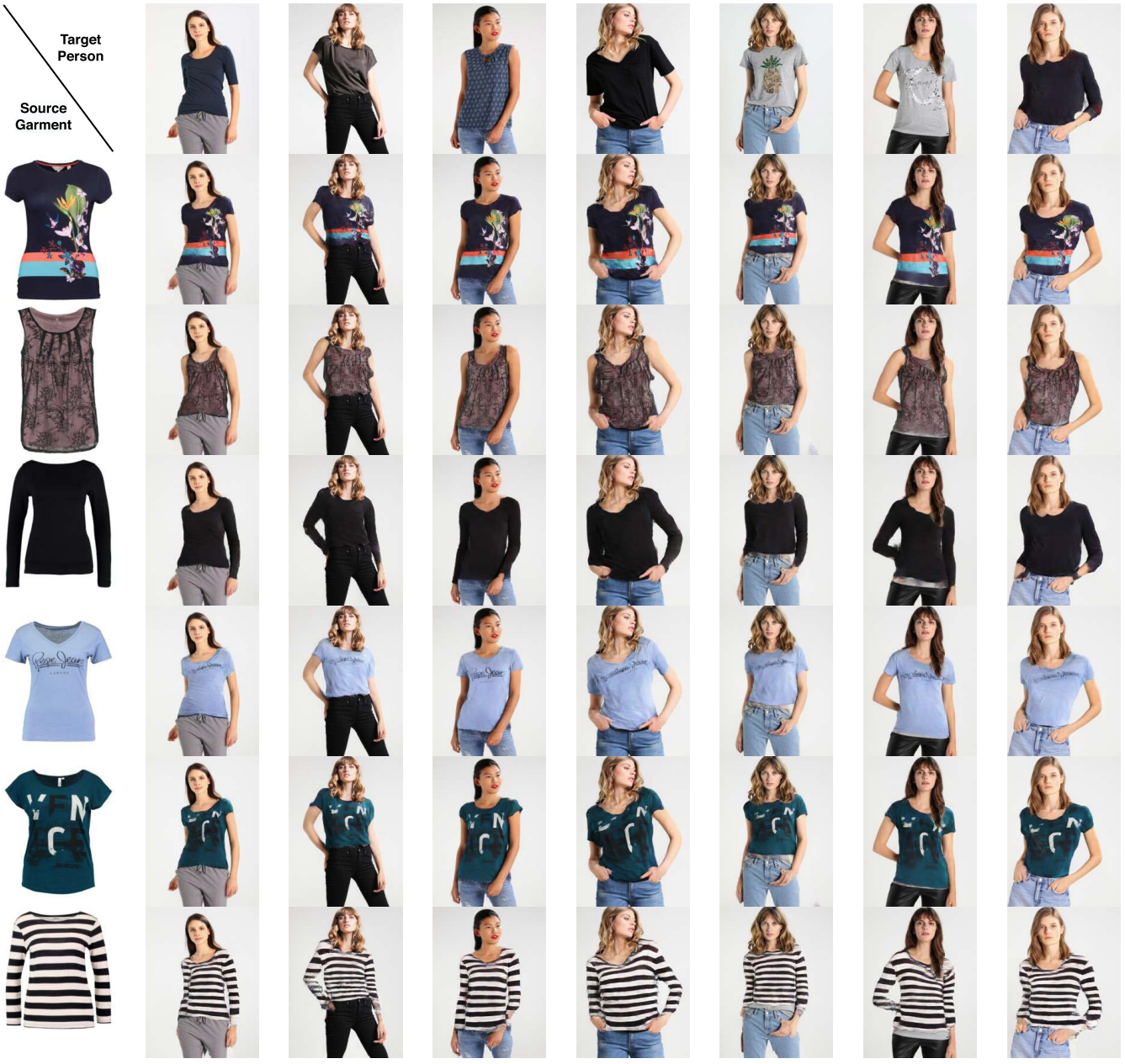

FIGURE 4. More visual try-on results of the proposed FVTN.

TPS for modeling the highly non-rigid transformation like clothing. What's more, the composition mask of CP-VTON cannot flexibly synthesize the body parts. By improving the clothing warping stage and the blending stage, CP-VTON+ gets better visual results than CP-VTON, especially in the case of body occlusion. Compared with CP-VTON and CP-VTON+, ClothFlow achieves much better try-on results, especially in the clothing regions. Such comparison demonstrates the advantages of the flow-based methods during learning clothing deformation. However, ClothFlow gets unsatisfactory human body parts because of its simple try-on image rendering model. ACGPN outperforms these three methods because of its proposed second-order spatial transformation constraint and inpainting module. However, we notice visible high-frequency artifacts in the collar regions of try-on images generated by ACGPN. By contrast, our proposed FVTN generates more perceptually convincing synthetic results which warp the garments more naturally and align them with the human body more accurately.

Figure 4 displays more synthetic examples of our FVTN in which different target persons under arbitrary poses virtually try on various garments. Notably, our method preserves the fine-grained appearance details of the garments (such as the

FIGURE 5. **Ablation study: synthesis quality evaluation on the virtual try-on task.**

logos and clothing textures) along with the body parts under complex posture changes and occlusion. Our FVTN can handle different types of body shapes because PAM provides accurate spatial information by aligning the source clothing to the posture of the target person at the semantic level. To sum up, these quantitative and qualitative results verify the effectiveness of our FVTN.

3) ABLATION STUDY

We train several ablation experiments on the VITON dataset to assess the contribution of each component of the FVTN. Several variants of the FVTN are trained: (1) *w.o/iter*, the model using a direct flow estimation scheme instead of an iterative flow estimation scheme in FEM. (2) *w.o/deconv*, the model using the standard convolution instead of deformable convolution in the flow estimation layer. (3) *w.o/multi*, the model using conventional single-scale photometric loss in FEM. (4) *w.o/fwarp*, the model accepting the warped clothing as input instead of the warped feature in FRM. (5) *w.o/mask*, the model accepting the body parts of the target person I_t^b without the body mask M_t^b as input in FRM.

TABLE 2. User study results on the VITON dataset. The results indicate the proportion of images that human subjects regard our method are better and more realistic than the compared method.

Model	CP-VTON	CP-VTON+	ClothFlow	ACGPN
VITON	87.1%	83.1%	69.8%	59.7%

FIGURE 6. Failure cases of our method on the virtual try-on task.

The quantitative evaluation results are shown in Table 1. From the results, we find that our model outperforms all the variants on all metrics and all components improve performances in different degrees. The qualitative evaluation results are visualized in Fig. 5. Consistent with the quantitative evaluations, our model surpasses all the variants with the highest quality of visual results. Besides, we obtain the following observations: (1) *w.o/mask* cannot adaptively determine the generation or preservation of the body parts. (2) *w.o/iter* generates very poor appearances of the garments. (3) *w.o/deconv*, *w.o/multi* and *w.o/fwarp* all bring visual artifacts in the synthesized clothing regions. Those findings prove the necessity of each component of our FVTN.

4) USER STUDY

We further evaluate the image quality of our synthesized images via a human subjective study involving 20 participants. Following [40], given two generated images, each participant is asked to choose a better and more realistic image meeting three criteria: (1) how well the target clothing characteristics of the source clothing image are preserved; (2) how photo-realistic the whole image is; (3) how good the whole person seems. Our FVTN achieves significantly better human evaluation scores on the image-based virtual try-on, as shown in Table 2. The results of the user study are consistent with those of qualitative and quantitative experiments, which demonstrate the effectiveness of the proposed FVTN.

5) FAILURE CASES AND LIMITATIONS

Fig. 6 displays two failure cases of our proposed FVTN on the virtual try-on task. The example of the top row is caused by the rarely-seen human poses while another example is due to the viewpoint transformation from the front view to the back view. Besides, our FVTN relies on human segmentation of different body parts to enable the learning procedure. Thus, wrong segmentation would lead to highly-unrealistic try-on images.

V. CONCLUSION

In this work, we propose a novel Flow-based Virtual Try-on Network (FVTN), which aims at generating photo-realistic try-on results. We present three tailored modules, i.e., Parsing Alignment Module (PAM), Flow Estimation Module (FEM) and Fusion and Rendering Module (FRM). Specifically, we design an unsupervised flow-based spatial alignment scheme in FEM to precisely capture clothing deformation. We propose an image synthesis network in FRM to synthesize the try-on look by integrating information from the warped clothing and the human body. The results clearly show the great superiority of our proposed FVTN in terms of quantitative metrics, visual quality and user study.

REFERENCES

[1] Q. Chen and V. Koltun, "Full flow: Optical flow estimation by global optimization over regular grids," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 4706–4714.
[2] W. Chen, H. Wang, Y. Li, H. Su, Z. Wang, C. Tu, D. Lischinski, D. Cohen-Or, and B. Chen, "Synthesizing training images for boosting human 3D pose estimation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 479–488.
[3] Y. Chen, Z. Wang, Y. Peng, Z. Zhang, G. Yu, and J. Sun, "Cascaded pyramid network for multi-person pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7103–7112.
[4] J. Dai, H. Qi, Y. Xiong, Y. Li, G. Zhang, H. Hu, and Y. Wei, "Deformable convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 764–773.
[5] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.
[6] H. Dong, X. Liang, X. Shen, B. Wang, H. Lai, J. Zhu, Z. Hu, and J. Yin, "Towards multi-pose guided virtual try-on network," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9025–9034.
[7] A. Dosovitskiy, P. Fischer, E. Ilg, P. Hausser, C. Hazirbas, V. Golkov, P. V. D. Smagt, D. Cremers, and T. Brox, "FlowNet: Learning optical flow with convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2758–2766.
[8] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.
[9] K. Gong, X. Liang, Y. Li, Y. Chen, M. Yang, and L. Lin, "Instance-level human parsing via part grouping network," in *Proc. ECCV*, in Lecture Notes in Computer Science, vol. 11208, 2018, pp. 805–822.
[10] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. NIPS*, 2014, pp. 2672–2680.
[11] X. Gu, J. Yu, Y. Wong, and M. S. Kankanhalli, "Toward multi-modal conditioned fashion image translation," *IEEE Trans. Multimedia*, vol. 23, pp. 2361–2371, 2021.
[12] Y. Guo, Q. Chen, J. Chen, Q. Wu, Q. Shi, and M. Tan, "Auto-embedding generative adversarial networks for high resolution image synthesis," *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2726–2737, Nov. 2019.
[13] X. Han, W. Huang, X. Hu, and M. Scott, "ClothFlow: A flow-based model for clothed person generation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10470–10479.
[14] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7543–7552.
[15] 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. NIPS*, 2017, pp. 6626–6637.

[16] B. K. P. Horn and B. G. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, nos. 1–3, pp. 185–203, Aug. 1980.
[17] P. Hu, G. Wang, and Y.-P. Tan, "Recurrent spatial pyramid CNN for optical flow estimation," *IEEE Trans. Multimedia*, vol. 20, no. 10, pp. 2814–2823, Oct. 2018.
[18] E. Ilg, N. Mayer, T. Saikia, M. Keuper, A. Dosovitskiy, and T. Brox, "FlowNet 2.0: Evolution of optical flow estimation with deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1647–1655.
[19] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.
[20] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. ECCV*, in Lecture Notes in Computer Science, vol. 9906, 2016, pp. 694–711.
[21] C. Lassner, G. Pons-Moll, and P. V. Gehler, "A generative model of people in clothing," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 853–862.
[22] G. Liu, A. F. Reda, J. K. Shih, T. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. ECCV*, in Lecture Notes in Computer Science, vol. 11215, Sep. 2018, pp. 89–105.
[23] P. Liu, M. Lyu, I. King, and J. Xu, "SelFlow: Self-supervised learning of optical flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4571–4580.
[24] S. Meister, J. Hur, and S. Roth, "UnFlow: Unsupervised learning of optical flow with a bidirectional census loss," in *Proc. AAAI*, 2018, pp. 7251–7259.
[25] M. Minar, T. Tuan, H. Ahn, P. Rosin, and Y. Lai, "CP-VTON+: Clothing shape and texture preserving image-based virtual try-on," in *Proc. CVPRW*, 2020, pp. 1–4.
[26] 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. (CVPR)*, Jun. 2020, pp. 7021–7032.
[27] C. Patel, Z. Liao, and G. Pons-Moll, "TailorNet: Predicting clothing in 3D as a function of human pose, shape and garment style," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7363–7373.
[28] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black, "ClothCap: Seamless 4D clothing capture and retargeting," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 73:1–73:15, 2017.
[29] A. Ranjan and M. J. Black, "Optical flow estimation using a spatial pyramid network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2720–2729.
[30] Y. Ren, X. Yu, J. Chen, T. H. Li, and G. Li, "Deep image spatial transformation for person image generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 7687–7696.
[31] T. Salimans, I. J. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. NIPS*, 2016, pp. 2226–2234.
[32] M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama, "Virtual fitting by single-shot body shape estimation," in *Proc. 5th Int. Conf. 3D Body Scanning Technol.*, Oct. 2014, pp. 406–413.
[33] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, pp. 1–14, Sep. 2014.
[34] D. Sun, X. Yang, M.-Y. Liu, and J. Kautz, "PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8934–8943.
[35] Z. Teed and J. Deng, "RAFT: Recurrent all-pairs field transforms for optical flow," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, in Lecture Notes in Computer Science, vol. 12347, 2020, pp. 402–419.
[36] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *Proc. ECCV*, in Lecture Notes in Computer Science, vol. 11217, 2018, pp. 607–623.
[37] 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/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.
[38] Y. Wang, Y. Yang, Z. Yang, L. Zhao, P. Wang, and W. Xu, "Occlusion aware unsupervised learning of optical flow," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4884–4893.
[39] 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.
[40] 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. (CVPR)*, Jun. 2020, pp. 7847–7856.
[41] S. Yang, T. Amert, Z. Pan, K. Wang, L. Yu, T. L. Berg, and M. C. Lin, "Detailed garment recovery from a single-view image," *CoRR*, vol. abs/1608.01250, pp. 1–13, Aug. 2016.
[42] J. J. Yu, A. W. Harley, and K. G. Derpanis, "Back to basics: Unsupervised learning of optical flow via brightness constancy and motion smoothness," in *Proc. ECCV Workshops*, in Lecture Notes in Computer Science, vol. 9915, 2016, pp. 3–10.
[43] R. Yu, X. Wang, and X. Xie, "VTNFP: An image-based virtual try-on network with body and clothing feature preservation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10510–10519.
[44] M. Yuan, I. R. Khan, F. Farbiz, S. Yao, A. Niswar, and M.-H. Foo, "A mixed reality virtual clothes try-on system," *IEEE Trans. Multimedia*, vol. 15, no. 8, pp. 1958–1968, Dec. 2013.
[45] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.
[46] H. Zheng, L. Chen, C. Xu, and J. Luo, "Unsupervised pose flow learning for pose guided synthesis," *CoRR*, vol. abs/1909.13819, pp. 1–12, Sep. 2019.
[47] T. Zhou, S. Tulsiani, W. Sun, J. Malik, and A. A. Efros, "View synthesis by appearance flow," in *Proc. ECCV*, in Lecture Notes in Computer Science, vol. 9908, 2016, pp. 286–301.

TAO WANG is currently pursuing the master's degree with Hangzhou Dianzi University. His research interests include computer vision and deep learning.

XIAOLING GU (Member, IEEE) received the Ph.D. degree in computer science from Zhejiang University, in 2017. She is currently an Associate Professor at the School of Computer Science and Technology, Hangzhou Dianzi University. She has published several top-tier conferences and journal papers, such as SIGIR, ACM Multimedia, and the IEEE TRANSACTIONS ON MULTIMEDIA. Her current research interests include computer vision, machine learning, and fashion data analysis.

JUNKAI ZHU is currently pursuing the master's degree with Hangzhou Dianzi University. His research interests include computer vision and deep learning.