SPG-VTON: Semantic Prediction Guidance for Multi-pose Virtual Try-on

Bingwen Hu, Ping Liu, *Member, IEEE,* Zhedong Zheng, and Mingwu Ren

***Abstract*—Image-based virtual try-on is challenging in fitting a target in-shop clothes onto a reference person under diverse human poses. Previous works focus on preserving clothing details (*e.g.,* texture, logos, patterns) when transferring desired clothes onto a target person under a fixed pose. However, the performances of existing methods significantly dropped when extending existing methods to multi-pose virtual try-on. In this paper, we propose an end-to-end Semantic Prediction Guidance multi-pose Virtual Try-On Network (SPG-VTON), which can fit the desired clothing into a reference person under arbitrary poses. Specifically, SPG-VTON is composed of three sub-modules. First, a Semantic Prediction Module (SPM) generates the desired semantic map. The predicted semantic map provides more abundant guidance to locate the desired clothing region and produce a coarse try-on image. Second, a Clothes Warping Module (CWM) warps in-shop clothes to the desired shape according to the predicted semantic map and the desired pose. Specifically, we introduce a conductible cycle consistency loss to alleviate the misalignment in the clothing warping process. Third, a Try-on Synthesis Module (TSM) combines the coarse result and the warped clothes to generate the final virtual try-on image, preserving details of the desired clothes and under the desired pose. In addition, we introduce a face identity loss to refine the facial appearance and maintain the identity of the final virtual try-on result at the same time. We evaluate the proposed method on the most massive multi-pose dataset (MPV) and the DeepFashion dataset. The qualitative and quantitative experiments show that SPG-VTON is superior to the state-of-the-art methods and is robust to data noise, including background and accessory changes, *i.e.*, hats and handbags, showing good scalability to the real-world scenario.**

***Index Terms*—Virtual Try-on, Multi-pose, Semantic Prediction, End-to-end**

I. Introduction

Image-based virtual try-on systems aim at fitting a target in-shop clothes into a reference person, which is a branch of the field of image synthesis. Driven by the rapid development of image synthesis [1]–[4], the topic of image-based virtual try-on has attracted more interest and has vast potential applications in virtual reality and human-computer interaction. Despite significant progress in previous works [5]–[16] for virtual try-on, the multi-pose virtual try-on has not been well studied. Concretely, for a given person image, the virtual fitting system could generate realistic images of this person in different poses while preserving the desired clothes' appearance. The multi-pose virtual fitting system is more in line with practical application scenarios. The existing works on multi-pose virtual fitting tasks are insufficient, and there are problems such as mismatch between the target clothes and the given pose, distortion of the clothes region in the try-on result, and loss of details that need to be further explored. To solve these problems, we propose a method to build a robust multi-pose virtual try-on system based on 2D images (as shown in Fig. 1).

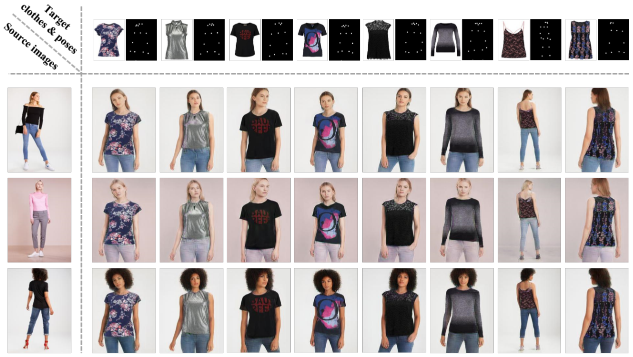


Fig. 1. Visual results of multi-pose virtual try-on by the proposed method.

Most of the previous works [5], [6], [8], [9], [11], [12] focus on swapping clothes in a fixed pose without considering the body's changing posture. However, in realistic scenarios, users would like to intuitively see the results of wearing the given clothes in different poses. Recently, MG-VTON [7] made the first attempt at a virtual fitting system guided by multiple poses. MG-VTON is a multi-stage framework that includes a human parsing network, a warping generative adversarial network, and a refinement render network. Although MG-VTON takes a significant stride in virtual fitting system construction under arbitrary posture, it still has some limitations that restricts its further applications: (1) An end-to-end mechanism is missing in the MG-VTON training process. In MG-VTON, modules designed with different purposes are utilized to generate the desired clothing deformation under

This work was supported in part by the National Natural Science Foundation of China under Grant 61773117 and Grant 61703209, and in part by the National Major Scientific Instruments and Equipments Development Project of China under Grant 61727802 *(Corresponding author: Mingwu Ren).*

B. Hu is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China, Nanjing 210092, China, and also with the ReLER Lab, AAII, University of Technology Sydney, Ultimo, NSW 2007, Australia (e-mail: hubw.sky@gmail.com).

P. Liu was with the ReLER Lab, AAII, University of Technology Sydney, NSW 2007, Australia. He is now with the Center for Frontier Artificial Intelligence Research (CFAR), Agency for Science, Technology, and Research (A*STAR), Singapore 138632 (e-mail: pino.pingliu@gmail.com).

Z. Zheng was with the ReLER Lab, AAII, University of Technology Sydney, NSW 2007, Australia. He is now with NExT++, School of Computing, National University of Singapore, Singapore 118404. (e-mail: zdzheng@nus.edu.sg).

M. Ren is with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China, Nanjing 210092, China (e-mail: renmingwu@njust.edu.cn).

Manuscript received April 19, 2005; revised August 26, 2015.

different postures, a coarse result, and a refined result step by step. Each module/step is optimized independently and cannot collaborate with others bidirectionally, which might cause a sub-optimal result. (2) In the inference process, MG-VTON requires multiple steps to generate the final results, which might cost more human interventions and cause error accumulations. (3) MG-VTON only focuses on manipulation of body parts while ignoring the other parts, for example, faces. The absence of a mechanism to process non-body parts makes their generated results unpleasant in nature and appearance.

To address the limitations mentioned above, we present a new multi-pose virtual try-on network that can work in an end-to-end manner. To make our method able to handle variations introduced by arbitrary postures, we propose to introduce semantic prior knowledge into our network, making the learning process receive additional guidance from semantic prior knowledge. We name our designed method Semantic Prediction Guidance-Virtual Try On system, *a.k.a.*, SPG-VTON. As shown in Fig. 2, the SPG-VTON consists of three major modules, including the Semantic Prediction Module (SPM), the Clothes Warping Module (CWM), and the Try-on Synthesis Module (TSM). The purpose of the designed SPM is to predict the semantic map for target images, which is utilized to provide additional spatial and semantic guidance during the learning process. Given the semantic map for source images, the in-shop clothes, the target pose, and the predicted semantic map for the target image, a coarse result and corresponding predicted clothing mask are generated by the SPM. The CWM is introduced to warp the in-shop clothes to the desired shape according to the semantic map predicted by the SPM. To alleviate the misalignment between the desired in-shop clothes and the target human posture, we propose utilizing a conductible cycle-consistency constraint in our network learning. Given the target pose, the coarse result generated by the SPM, and the warped clothes generated by the CWM, the TSM generates the final try-on image with high precision and realism. Furthermore, to make the generated results photo-realistic, we introduce a global discriminator and a local discriminator to control the global shape and local texture of the generated results; to make the generated image visually pleasant, we utilize a face identity constraint to keep the synthesis face region realistic.

Extensive experiments on the MPV dataset [7] show that our method achieves superior performance to several existing approaches [5]–[7]. It is worth noting that what we studied in this work is currently the largest dataset for the multi-pose virtual try-on task. Since all images are collected from the internet, the dataset inevitably contains unexpected ”label noise”, such as misalignment and different backgrounds. The noise existing in web collected data inevitably compromises the training process in the dataset, making it challenging to train a robust virtual try-on system based on these noisy images. The proposed method is robust to noisy data through the mutual cooperation of the introduced various losses and the end-to-end model frameworks. The specific mechanism can be summarized as follows: (1) In the training process, the end-to-end approach allows the modules to be integrated, dynamically adjusts the network parameters of each module, and encourages the model to generate better results. Meanwhile, the end-to-end generation could , in turn, correct inaccurate predicted semantic maps to guide the model with more accurate semantic information. (2) This paper introduces global-oriented losses, *i.e.*, reconstruction loss and perceptual loss, to make the generated result consistent with the ground-truth image at both the pixel level and perceptual level to resist the interference of noisy data. In addition, this paper also introduces local-oriented losses, *i.e.*, the conductible cycle consistency loss and the face identity loss, to ensure that the clothing area and facial area of the generated image retain more characteristic information of face regions. Moreover, introducing global and local adversarial loss can also ensure that the generated image is close to the real image and prevent the negative impacts from noisy input. (3) One way of compression or denoising in prior works [17]–[19] is to extract kernel information through latent representation learning. The latent representation is usually much smaller in dimension than the original input data, making it easier to control and analyze. In this work, the role of prior semantic information (human semantic map) is similar to that of latent representation. Specifically, the first process of SPM predicts the semantic map of the target image, and the second process of SPM predicts the mask of the target clothing area. In this case, SPG-VTON can accurately locate the target clothing area by combining the semantic map of the target image and the target clothing mask, which could prevent the generation of background noise. Benefiting from our designed method and exploration, we experimentally observe that the proposed method is still robust to such training noise, and demonstrates good scalability to unseen test images during inference. The main contributions of the proposed method are summarized as follows:

- We propose an end-to-end image-based multi-pose virtual try-on system called SPG-VTON, which can synthesize high-quality try-on images. Compared to previous works, the proposed method could fit the desired clothing into a reference person under arbitrary poses while preserving details of the desired clothes.
- We conduct extensive explorations and locate effective strategies for learning a robust and accurate virtual try-on network for multi-pose inputs. The novelty points of the proposed method can be summarized as follows: (1) We introduce a conductible cyclic consistency loss to alleviate the misalignment in the clothing warping process. Concretely, the conductible cycle consistency loss could match the shape of the deformed desired clothes with the target person image and maintain the characteristics of the desired clothes in the generated try-on image. (2) We introduce both global and local adversarial losses and face identity loss to refine the facial appearance and maintain the identity of the final virtual try-on result at the same time. Specifically, the role of global and local adversarial losses encourages the generated image to be close to the real image, whether it is the whole image or part of the whole image. Moreover, the role of face identity loss enforces that the identity of the generated

image remains unchanged. (3) We apply an end-to-end training strategy to boost the proposed method to generate accurate semantic maps and improve the virtual try-on results under pose transfer. Additionally, the end-to-end manner can effectively reduce human interventions and avoid error accumulations in the inference process.

- The qualitative and quantitative experiments on two prevailing datasets, *i.e.*, MPV [7] and DeepFashion [20], demonstrate the advantages of our method in virtual try-on, especially when given different postures with heavy variations. The ablation studies also show that the proposed method has good scalability to unseen test data and is robust to label noise in the training set.

II. Related Work

A. Pose-Guided Person Image Generation

Pose-guided person image generation is a practical yet challenging topic. In past years, Generative Adversarial Networks (GANs) [21] and various extensions [1]–[4], [22]–[26] have made significant progress in this research direction. However, due to the high variations existing in human poses and human appearance, these previous works [1]–[4], [22], [24], [25] still suffer from their limited scalability in pose-guided person image generation. To generate high-quality person images in arbitrary poses, Ma *et al.* [27] proposed a two-stage generation framework. [27] first uses the U-Net-like network to produce initial images with blur and then applies the adversarial method to refine coarse results. To further improve generated image qualities, Ma *et al.* proposes a *disentangling* strategy [28], encoding a given person image into three factors: pose, foreground, and background, which are decoded back to an image space after editing a specific factor. It is believed that manipulating those disentangled factors rather than treating them as a whole can benefit generation quality improvement. Zheng *et al.* disentangles the input pedestrian images into structure and appearance embedding, and can easily exchanges codes to generate source person with target clothes [29]. [30] also adopts the disentanglement strategy. Specifically, they use a conditional U-Net architecture that combines the appearance decomposed from a variational autoencoder [31] with a given shape to reconstruct a new image. Methods such as [27], [28], [30] focus on the *global* pose deformation between the source image and the target image, while ignoring the *local* structure of generated images. Therefore these methods have difficulty maintaining the local details of the original image, especially when there is a large pose discrepancy between the source image and the target image. Recently, some works [9], [32]–[35] have focused on the spatial deformation relationship in pose changes. [32] uses an inpainting network to estimate the coordinates in source images for elements of the body surface. Def-GAN [33] designs deformable skip connections in the generator to address the pixel-to-pixel misalignment caused by the pose differences. [9], [34], [35] employ a specific module to predict the human semantic map after the pose changes to align the source image with the target pose to enforce the module to generate high-quality images. In this work, we also introduce a semantic map prediction model to produce the semantic map under a given pose. The difference is that when our method predicts the semantic map under a given pose, the clothing region of the semantic map changes with the given clothes. In contrast, the works mentioned above for pose-guided person image generation do not involve this aspect.

B. Virtual Try-on

The image-based virtual try-on task is a particular case of person generation. The core difference is that this task aims to generate a person image while the clothes region is changed to the desired clothes. The Thin-Plate Spline (TPS) [36] transformation is a typical 2-D interpolation model that performs geometric deformation between images by controlling a set of registration points between two images. VITON [5] directly applies shape context-based matching [37] to estimate TPS transformation parameters between the mask of desired clothes and the clothes mask of the target person. Furthermore, CP-VTON [6] uses a learnable method to estimate the TPS transformation parameters via convolutional neural networks dynamically. In the image-based virtual try-on task, using convolutional neural networks to learn the TPS transformation parameters between the desired clothes and the given human image is verified as a practical approach. The pioneering works [6]–[8], [10]–[12] mainly use two ways to estimate TPS transformation parameters. One way [6], [8], [10], is to use the geometric matching network [38] to estimate the TPS transformation parameters between the target clothing and the person representation (embedding the body shape, the target pose and reserved regions of the source image). In addition, these methods directly use the pixel-wise $\mathcal{L}_1$-norm between warped clothes and the clothing area extracted from the target image to train the geometric matching network. Another way [7], [11], [12] is to apply the geometric matching network to estimate the TPS transformation parameters between the clothing or clothing mask and the clothing area mask or the body shape obtained from the predicted semantic map. Similar to the first way, these methods also use the pixel-wise $\mathcal{L}_1$-norm between warped clothes and the clothes region extracted from the target person image to train the geometric matching network.

Although the methods mentioned above can produce high-quality fitting images to a certain extent, there is still a large gap between the generated images and natural images. For example, VITON [5] and CP-VTON [6] are state-of-the-art virtual try-on approaches that adopt a multi-stage coarse-to-fine strategy to tackle the virtual try-on task of a single pose. However, neither of these two methods includes changes in human pose. In this case, these methods cannot avoid distortion and misalignment in the process of clothing deformation (such as the distortion and misalignment of texture, patterns, logos, and embroidery).

III. Method

We propose a novel image-based virtual try-on network named SPG-VTON, which focuses on multi-pose virtual try-on. Specifically, for a given source person image, a target in-shop clothes, and a target pose, the proposed method aims to

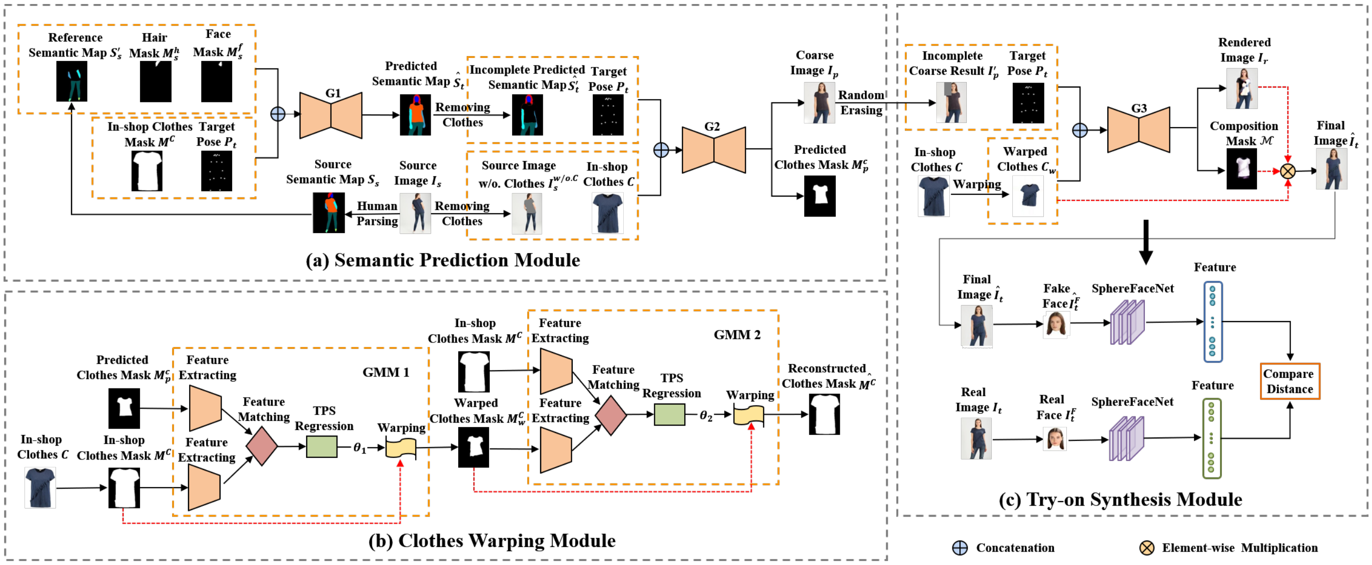


Fig. 2. The overview of our SPG-VTON. (a) The Semantic Prediction Model (SPM) consists of two processes. One is the target semantic map prediction, and the other one is the target clothes mask prediction and coarse result generation. (b) The Clothes Warping Module (CWM) warps the in-shop clothes to the shape of the clothes region of the target image, according to the Thin-Plate Spline (TPS) transformation parameters estimated between the mask of in-shop clothes and the predicted clothes mask. The CWM is composed of two Geometric Matching Modules (GMM) [38]. In specific, the GMM 1 deforms the mask of the in-shop clothes to the same shape with the predicted clothes mask, and then the GMM 2 is used to convert the warped clothes mask back to the mask of the desired clothes. Note that we do not directly apply the $\mathcal{L}_1$-norm between the warped clothes mask and the predicted clothes, and $\mathcal{L}_1$-norm between the reconstructed clothes and the original clothes mask to train GMM 1 and GMM 2. By contrary, we introduce a conductible cycle consistency loss to indirectly constrain GMM 1 and GMM 2, respectively (see III-C for details). (c) The Try-on Synthesis Module (TSM) combines the incomplete coarse results, the target pose, and the warped desired clothes to synthesis the final virtual try-on images. Also, the pre-trained SphereFaceNet [39] is applied to compare the distance between the generated face region and the ground-truth face region, which enforces the generator G_3 to generate realistic and natural faces.

generate a new person image such that the same person wears the target in-shop clothes and preserves the target pose. That is, given different poses, the proposed method can generate high-quality virtual try-on images.

The SPG-VTON consists of three sub-modules, including the Semantic Prediction Module (Section III-B), the Clothes Warping Module (Section III-C), and the Try-on Synthesis Module (Section III-D). We show the overview of SPG-VTON in Fig. 2. Concretely, the SPM has two sequential processes. The first process of SPM aims to predict the semantic map of the target image according to the source semantic map, the in-shop clothes, and the target pose. The predicted semantic map provides precise guidance to locate the region of the desired clothes and generate a coarse virtual try-on image. Then, we combine the predicted semantic map, the in-shop clothes, and the target pose as the input of the second process of SPM to generate the coarse result and the predicted clothes mask. Subsequently, the CWM warps the in-shop clothes to the desired shape according to the predicted semantic map.

A. Person Representation

The diverse clothes and human poses struggle with the performance of the virtual try-on system. During the training process, the human pose and human body semantic map are critical supervision information for understanding the human geometric structure. For training images, we apply off-the-shelf pose estimator [40] and human parser [41] to extract the human body keypoints and semantic maps, respectively. The detailed process is as follows:

Human pose embedding. Following several off-the-shelf virtual try-on methods [5]–[9], [11], we use the pose estimator [42] to extract the pose of each person image. Then we obtain the coordinates of 18 human body keypoints from each person image and convert them to an 18-channel heatmap. Each channel of the heatmap corresponds to a human pose keypoint. We use each keypoint as the center of the circle to draw a circle with a radius of 4 pixels. The values in each circle are all ones, and the parts outside the circle are all zeros. In this way, we obtain the representation of the human pose embedding.

Human semantic map. Inspired by two semantic-guided virtual try-on approaches [7], [8], we extract human semantic maps of training images by using the existing human parser [41]. Each semantic map contains 20 labels that correspond to different parts of the human body. For intractable human body parts, such as the head (including the face and hair regions), it preserves characteristic personal identity information. The extracted information provided an additional supervision signal for training our network.

B. Semantic Prediction Module

To precisely locate the clothes region of the generated person image and alleviate the mismatching between the target clothes and the generated human body, we introduce a Semantic Prediction Module (SPM). As shown in Fig. 2 a, SPM consists of two sequential processes and can be optimized in one step. The first process is target semantic map prediction, and the other is a target clothes mask and coarse try-on result generation. First, given a source human image I_s and its corresponding semantic map S_s, a target in-shop clothes C, and a target human pose P_t, the first process of SPM aims to predict the target human semantic map $\hat{S}_t$ conditioned on the source semantic map S_s, the target clothes C, and the target pose P_t. Second, we combine the predicted semantic map $\hat{S}_t$ with the desired clothes C, the target pose P_t, and the source image without clothes $I_s^{w/o.C}$, and then feed it into the second process of SPM to generate the coarse try-on image and the predicted clothes mask. Specifically, each process of the SPM is based on the conditional Generative Adversarial Network (cGAN) [22]. We adopt a ResNet-like network to replace the U-Net [43] structure as the generator G, and the multi-scale discriminator (PatchGAN) [3] is applied as the discriminator D. SPM contains two generators (*i.e.*, G_1 and G_2) and two discriminators (*i.e.*, D_1 and D_2). G_1 produces the predicted semantic map of the target person that makes discriminator D_1 indistinguishable from the real image. Similarly, the role of generator G_2 is to generate the realistic coarse result and the predicted clothes mask. At the same time, discriminator D_2 attempts to distinguish real images from the results generated by G_2. The network structure of generators and discriminators can be found in Table I.

Target semantic map prediction. We first define the head part as $A_s = \{M_s^f, M_s^h\}$, which is composed of the face mask M_s^f and the hair mask M_s^h. A_s means that all the masks in A are obtained from the source semantic map S_s. Next, we remove the hair, face, and clothing areas from the source semantic map S_s to obtain the reference semantic map S'_s. Finally, we binarize the desired clothes image C to obtain the mask of desired clothes M^C. As shown in Fig. 2 (a), we combine the head part A_s with the reference semantic map S'_s, the mask of desired clothes M^C, and the target pose P_t as the input of the target semantic map prediction. Therefore, the predicted semantic map can be formulated as $\hat{S}_t = G_1(S'_s, A_s, M^C, P_t)$. To encourage the generated semantic map to be indistinguishable from the ground-truth semantic map, we introduce the spatial matching adversarial loss [21] as follows:

$$\begin{aligned} L^s_{adv} = \;& \mathbb{E}[\log D_1(S_t, S'_s, A_s, M^C, P_t)] \\ & + \mathbb{E}[\log(1 - D_1(G_1(S'_s, A_s, M^C, P_t), S'_s, A_s, M^C, P_t)]. \end{aligned} \tag{1}$$

In addition, to generate high-quality target semantic maps $\hat{S}_t$, we utilize the focal loss [41] on pixel-wise segmentation. Besides, following [7], [44], we also adopt the pixel-wise $\mathcal{L}_1$-norm between the predicted semantic map and the target semantic map to push the generator G_1 to produce smoother results. Therefore, the objective function to generate the target semantic map can be formulated as:

$$L_{seg} = L^s_{adv} + \lambda_1 L^s_{fl} + L^s_{recon}, \tag{2}$$

where L^s_{fl} denotes the focal loss between $\hat{S}_t$ and S_t, L^s_{recon} denotes the pixel-wise $\mathcal{L}_1$-norm between the predicted semantic map $\hat{S}_t$ and the target semantic map S_t, and the hyper-parameter λ_1 control weights of the focal loss.

Target clothes mask prediction and coarse result generation. As shown in Fig. 2 (a), after obtaining the predicted semantic map $\hat{S}_t$ from the first process of SPM, we first remove the clothe region of $\hat{S}_t$ to obtain the incomplete predicted semantic map $\hat{S}'_t$, and then combine $\hat{S}'_t$ with the target pose P_t, the in-shop clothes C, and the source image without clothes $I_s^{w/o.C}$ as the input of the second process of SPM (*i.e.*, the input of the generator G_2). Then the generator G_2 produces both the coarse try-on image I_p and the predicted clothes mask M_p^c. To allow the generator to produce the photo-realistic coarse try-on image I_p, we also define the adversarial loss L^p_{adv} as follows:

$$\begin{aligned} L^p_{adv} = \;& \mathbb{E}[\log D_2(I_t, \hat{S}_t, C, P_t, I_s^{w/o.C})] \\ & + \mathbb{E}[\log(1 - D_2(I_p, \hat{S}_t, C, P_t, I_s^{w/o.C})]. \end{aligned} \tag{3}$$

Following several start-of-the-art virtual try-on methods [5]–[9], [11], [12], we also adopt the perceptual loss [45] to enforce the generator G_2 to synthesize photo-realistic try-on images. The perceptual loss between I_p and I_t can be defined as:

$$L^p_{perc} = \mathbb{E}[\sum_{i=0}^{5} \sigma_i \|\phi_i(I_p) - \phi_i(I_t)\|_1], \tag{4}$$

where $\phi_i(I_.)$ denotes the i-th layer feature map of the image $I_.$ in the visual perception network ϕ. We apply the pre-trained VGG-19 [46] network as ϕ. Here five activations are utilized to calculate the perceptual loss. The hyper-parameters σ_i control the weights of the i-th layer to the term in Eq. 4. We apply the pixel-wise ℓ_1 loss to guide the target clothes mask prediction and the coarse result generation. Therefore, to minimize the distance between the generated image and the ground-truth image at the pixel level, we formulate the loss function as:

$$L^p_{recon} = \mathbb{E}[\|I_p - I_t\|_1] + \mathbb{E}[\|M_p^c - M_t^c\|_1], \tag{5}$$

where M_p^c represents the predicted clothes mask, and M_t^c denotes the clothes region mask extracted from the target human image I_t. The objective function to generate the coarse try-on result and target clothes mask can be formulated as:

$$L_{prd=L^p_{adv}+\lambda_2(L^p_{recon}+L^p_{perc}),} \tag{6}$$

where λ_2 control weights of L^p_{recon} and L^p_{perc}. Then the full objective function of the SPM can be defined as:

$$\begin{aligned} L_{spm} = \;& L_{seg} + L_{prd} \\ = \;& L^s_{adv} + \lambda_1 L^s_{fl} + L^s_{recon} \\ & + L^p_{adv} + \lambda_2(L^p_{recon} + L^p_{perc}), \end{aligned} \tag{7}$$

C. Clothes Warping Module

The Clothes Warping Module (CWM) aims to fit the desired clothes into the target person according to the given pose while preserving the texture of clothes. Most existing works [6]–[8], [11] directly utilize a Geometric Matching Model (GMM) [38] to estimate the parameters of the Thin-Plate Spline (TPS) used to warp clothes. This strategy is applicable when the texture of the clothes is monotonous, and the target pose is fixed. However, when dealing with complex cases (*e.g.,* the desired clothes with complex texture and the target person under diversity poses), it might lead to misalignment between clothes and the human body, and blurred results. To address the challenges mentioned above, we introduce the conductible cycle consistency loss, which effectively aligns the desired clothing with a given pose.

As shown in Fig. 2 (b), we first apply the geometric matching network [38] to estimate the TPS transformation parameters θ_1 between the mask of desired clothes M^C and the predicted clothes mask M_p^c. The mask of desired clothes M^C is then warped using the transformation parameters θ_1 to align it with the predicted clothes mask M_p^c. We denote the operation of TPS transformation as $\mathcal{T}_\theta$, and then the warped mask of desired clothes M_w^C can be represented as $M_w^C = \mathcal{T}_{\theta_1}(M^C)$. Next, the second geometric matching network is adopted to estimate the TPS transformation parameters θ_2 between the warped mask of desired clothes M_w^C and the mask of desired clothes M^C. We use the TPS transformation parameters θ_2 to warp M_w^C back to the original mask of desired clothes M^C. We denote the output of the second geometric matching network as $\hat{M^C} = \mathcal{T}_{\theta_2}(M_w^C)$.

Conductible cycle consistency loss. A straightforward solution is to train the geometric matching network by directly applying the pixel-level $\mathcal{L}_1$-norm to encourage M_w^C to approximate M_p^c and $\hat{M^C}$ to approximate M^C to obtain the TPS transform parameters for the desired clothing deformation. However, using the self-reconstruction approach to estimate TPS parameters between the given clothes mask and the target clothes mask can only roughly align the shape of the given clothes mask with the target clothes mask. Applying the estimated parameters to warp the desired clothes directly cannot preserve the details of the clothes well. Based on these observations, we introduce the conductible cycle consistency loss for two goals. One is to match the shape of the deformed desired clothes with the target person image. The other is to maintain the characteristics of the desired clothes in the clothing region of the generated try-on image. Specifically, we adopt the estimated parameters θ_1 to warp the desired clothes C to obtain the warped clothes $C_w = \mathcal{T}_{\theta_1}(C)$, and then use the estimated parameter θ_2 to warp C_w to produce $\hat{C} = \mathcal{T}_{\theta_2}(C_w)$. Thus, we formulate the conductible cycle consistency loss as:

$$L_{cond} = \mathbb{E}[\|C_w - C_t\|_1] + \mathbb{E}[\|\hat{C} - C\|_1], \tag{8}$$

where C_t denotes the ground-truth clothes region exacted from the target image I_t. Then the full objective function of the CWM can be defined as:

$$L_{cwm} = \lambda_3 L_{cond}, \tag{9}$$

where λ_3 controls weights of the conductible cycle consistency loss.

Discussion. The existing methods [5]–[8], [11] adopt a separate clothing deformation module to calculate the TPS transformation parameters between the desired clothing and the target person image. The TPS transformation parameters are given by the pre-trained clothing deformation module, which cannot be dynamically adjusted for these parameters and leads to error stacking. In contrast to previous works [5]–[8], [11], we adopt an indirect approach to train the geometric matching network. Concretely, we first fed the mask of desired clothes and the predicted clothes mask of the target person image into the geometric matching network, then used the estimated TPS transformation parameters to warp the desired clothes. The pixel-wise $\mathcal{L}_1$ loss between the warped desired clothes and the ground-truth clothes extracted from the target person image is the constraint of the geometric matching network. Furthermore, we adopt the pixel-wise $\mathcal{L}_1$-norm between $\hat{C}$ and the desired clothes C to encourage the warped mask of desired clothes M_w^C to return to the original mask of desired clothes M^C. This indirect constraint strategy can not only achieve accurately geometric deformation between the desired clothes and the target person image under an arbitrary pose, but also preserve rich details of the clothing in the generated person image. The ablation study verifies the effectiveness of the conductible cycle consistency loss.

D. Try-on Synthesis Module

After producing the coarse try-on result I_p by SPM, we first randomly erase a part of I_p to obtain the incomplete coarse result $I_p^{'}$, and then we combine $I_p^{'}$, and the deformed desired clothes C_w obtained by CWM and the target pose P_t. Next, we directly fed them into the TSM to generate the final virtual try-on image $\hat{I}_t$. As shown in Fig. 2 (c), we use the generator G_3 to produce a rendered result I_r and a composition mask $\mathcal{M}$ at the same time. The final virtual try-on result can be formulated as follows:

$$\hat{I}_t = C_w \odot \mathcal{M} + I_r \odot (1 - \mathcal{M}), \tag{10}$$

where $(I_r, \mathcal{M}) = G_3(I_p^{'}, C_w, P_t)$. $\odot$ represents the element-wise matrix multiplication, and the clothes part in the final try-on image can be denoted as $\hat{C}_t = C_w \odot \mathcal{M}$. The value of each element in $\mathcal{M}$ is between 0 and 1. Following several start-of-the-art virtual try-on methods [6], [8], [11], [12], we apply both the self-reconstruction loss and the perceptual loss [45] to enforce the generated image $\hat{I}_t$ to approximate the target image I_t. We define the full reconstruction loss as:

$$L_{recon}^t = \alpha_1 \mathbb{E}[\|\hat{I}_t - I_t\|_1]) + \alpha_2 \mathbb{E}[\|1 - \mathcal{M}\|_1], \tag{11}$$

where we adopt the second term in Eq. 11 as the regularization to constrain the generation of composition mask $\mathcal{M}$. We set $\alpha_1 = 2$ and $\alpha_2 = 0.5$. Additionally, the perceptual loss between $\hat{I}_t$ and I_t can be formulated as follows:

$$L_{perc}^t = \mathbb{E}[\sum_{i=0}^{5} \sigma_i \|\phi_i(\hat{I}_t) - \phi_i(I_t)\|_1]. \tag{12}$$

Global and local adversarial loss. In the real scenario, it is difficult to obtain images of the same person wearing the desired clothes in arbitrary poses. Therefore, we cannot retain parts outside the clothing area following many existing virtual try-on methods based on a fixed pose. In this case, to generate the photo-realistic try-on image, we employ a global adversarial loss and a local adversarial loss in TSM. The diagram of global and local discriminators is shown in Fig. 3. Specifically, we apply the global adversarial loss L_{adv}^g to enforce the generator G_3 to synthesize sharp virtual try-on images with global consistency. Furthermore, the local adversarial loss L_{adv}^l is adopted to refine the face area of the final result with local consistency. The global and local adversarial loss can be formulated as follows:

$$\begin{aligned} L_{adv}^g = \; & \mathbb{E}[\log D_g(I_t, C_t, P_t, I_s^{w/o.C})] \\ & + \mathbb{E}[\log(1 - D_g(\hat{I}_t, \hat{C}_t, P_t, I_s^{w/o.C})], \end{aligned} \tag{13}$$

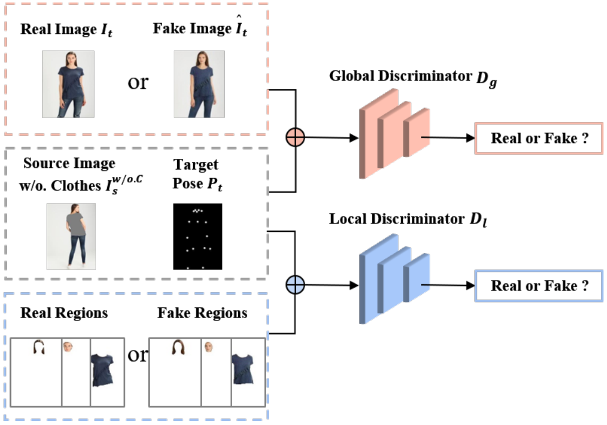


Fig. 3. The diagram of global and local discriminators. More details can be found at Table I.

$$\begin{aligned} L^l_{adv} = & \ \mathbb{E}[\log D_l(I_t^f, I_t^h, C_t, P_t, I_s^{w/o.C})] \\ & + \mathbb{E}[\log(1 - D_l(\hat{I_t^f}, \hat{I_t^h}, \hat{C}_t, P_t, I_s^{w/o.C})], \end{aligned} \tag{14}$$

where we extract the face region I_t^f and the hair region I_t^f from the ground-truth image I_t. In the same way, we also extract the face region $\hat{I_t^f}$, and the hair region $\hat{I_t^h}$ from $\hat{I}_t$. D_g denotes the global discriminator, and D_l denotes the local discriminator. D_g and D_l share the same structure.

Face identity loss. After adopting the global and local adversarial loss, our method can produce high-quality fitting images while making the face region look natural. Furthermore, to enforce that the face region cropped from the generated images remains similar to the face region of the related ground-truth images, we introduce the face identity loss.

$$L_{id} = \mathbb{E}[\|\mathcal{F}(\hat{I_t^F}) - \mathcal{F}(I_t^F)\|_1], \tag{15}$$

where $\mathcal{F}$ denotes the pre-trained SphereFaceNet [39]. $\hat{I_t^F}$ represents the face region extraction guided by the predicted semantic map, including the face part and the hair part. Then we formulate the full objective function of the TSM as:

$$L_{tsm} = \lambda_4(L^t_{recon} + L^t_{perc}) + L^g_{adv} + L^l_{adv} + \lambda_5 L_{id}, \tag{16}$$

where the hyper-parameter λ_4 controls weights of the reconstruction loss L^t_{recon} and the perceptual loss $L_{perc}(\hat{I}_t, I_t))$, and λ_5 controls weights of the face identity loss L_{id}.

E. Optimization

Objective function. Taking all of the above loss functions into consideration, we formulate the total objective function as:

$$\begin{aligned} L_{total} = & \underbrace{L^s_{adv} + \lambda_1 L^s_{fl} + L^s_{recon} + L^p_{adv} + \lambda_2(L^p_{recon} + L^p_{perc})}_{\text{SPM}} \\ & + \underbrace{\lambda_3 L_{cond}}_{\text{CWM}} + \underbrace{\lambda_4(L^t_{recon} + L^t_{perc}) + L^g_{adv} + L^l_{adv} + \lambda_5 L_{id}}_{\text{TSM}}. \end{aligned} \tag{17}$$

Our ultimate goal is to solve:

$$G^*, \mathcal{T}^* = \arg \min_{G_1, G_2, \mathcal{T}_{\theta_1}, \mathcal{T}_{\theta_2}, G_3} \max_{D_1, D_2, D_g, D_l} L_{total}. \tag{18}$$

End-to-end training. In this article, we divide the training process of the proposed method into three steps: (1) We first separately train the semantic prediction network to obtain the preliminary semantic prediction map, and this process corresponds to the first process of the semantic prediction module. (2) Then, we use the semantic prediction map generated from the fixed pre-trained semantic prediction network as guidance for the subsequent steps, including rough result generation, clothing area prediction, target clothing deformation, and refined try-on generation. (3) Finally, we jointly train three sub-modules of the proposed method to synthesize the final virtual try-on image. This step can alleviate the impact of inaccurate semantic maps and improve the quality of the generated results.

Differences from MG-VTON [7]. There are significant differences between MG-VTON. (1) MG-VTON adopts a multi-stage framework, and each stage independently implements a distinct task. In particular, MG-VTON applies different modules to achieve the desired clothing deformation, coarse result generation, and final result refinement separately. Instead, we employ end-to-end training to encourage the generator to produce realistic virtual try-on results. Specifically, we dynamically update the parameters of each process including the desired clothing deformation, the coarse result generation, and the final result refinement. (2) Both MG-VTON and SPG-VTON use a semantic prediction model to predict the target semantic map (*i.e.*, the first process of SPM in SPG-VTON and the conditional human parsing network in MG-VTON). However, unlike MG-VTON, SPM in SPG-VTON contains one extra clothes mask prediction. Therefore, SPG-VTON can accurately locate the target clothing area by combining the predicted target semantic map, and the predicted target clothes mask. (3) MG-VTON divides the coarse result generation and the composition mask production into two independent steps. In contrast to this technique, we concatenate the coarse result, the warped clothes, and the target pose as the input of G_3 to produce a rendered result and a composition mask at the same time. Then, we combine the warped clothes, the rendered result, and the composition mask to synthesize the final result. These differences make our method generate images with improved qualities in both qualitative and quantitative evaluations, which is demonstrated in Fig. 5, TABLE II and TABLE III.

IV. Experiments

A. Dataset

We perform experiments on the MPV dataset [7], which is the largest multi-pose virtual fitting dataset available. The MPV dataset consists of 35,687 person images and 13,524 clothes images collected from the internet, with a resolution of 256 × 192. For each in-shop item of clothes, the dataset contains multiple images of the same person wearing the given in-shop clothes in different poses. The MPV dataset contains 62,780 three-tuples, including 52,236 training sets and 10,544

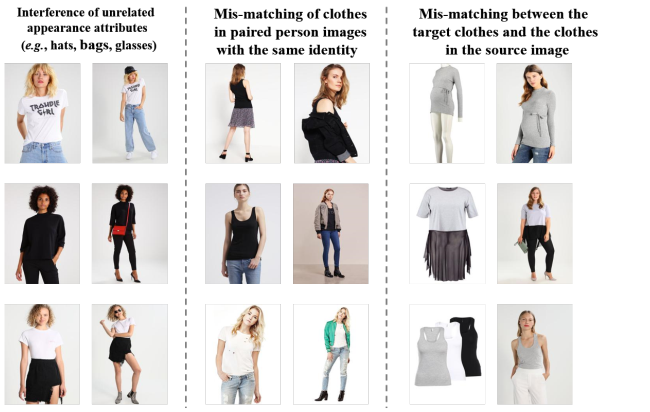

Interference of unrelated appearance attributes (e.g., hats, bags, glasses)
Mis-matching of clothes in paired person images with the same identity
Mis-matching between the target clothes and the clothes in the source image

Fig. 4. Examples of noisy images in the training set and the test set. We observe the three typical kinds of noise existing in the training set and the test set, including interference of unrelated appearance attributes (*e.g.*, hats, bags, glasses), mismatching of clothes in paired person images with the same identity, and mismatching between the target clothes and the clothes in the source image. These noise demands strong robust ability of the proposed algorithm during both training and testing.

test sets. Each input data point is a three-tuple composed of two human images and one clothes image. The two person images wear the clothes in the clothes image but with different poses. Moreover, we also conduct experiments on the DeepFashion dataset (*In-shop Clothes Retrieval Benchmark*) [20] to verify the effectiveness of the proposed method. Following the setting in MG-VTON [7], we collect 10,000 pairs (the same person in different poses) from DeepFashion, and randomly select in-shop clothes from the test set of the MPV dataset.

Noise in the training set and the test set. For the virtual try-on task, the input three-tuple consists of two paired person images and one in-shop clothes image. When one of the following conditions occurs, the input three-tuple can be regarded as "noise data." (1) One person image contains attributes that do not exist in the other person image, and these attributes are limited to glasses, bags, hats, scarves, necklaces, gloves, coats, upper clothes, and background; (2) The given clothes are different from the clothes in the paired person images. Since MPV does not provide labels for the noise data, in this case, to accurately obtain the noise data in the training set, we manually filter it and obtain 8,740 sets of noise data, accounting for approximately 16.73% of the training set. We show three typical kinds of noise in both the training set and the test set in Fig. 4, including interference of unrelated appearance attributes (*e.g.*, hats, bags, glasses), mismatching of clothes in paired person images with the same identity, and mismatching between the target clothes and the clothes in the source image. It is challenging to train a robust virtual fitting system based on these noisy images.

B. Evaluation Metrics

Structural SIMilarity (SSIM). SSIM [47] is widely used to evaluate the similarity of generated images in GAN-based methods. In this work, we adopt the SSIM metric to measure the similarity between the generated image and the real image. Higher scores indicate that the generated image is closer to the ground-truth image.

TABLE I
THE NETWORK ARCHITECTURE OF OUR GENERATORS AND DISCRIMINATORS, WHERE K, S, P, AND A DENOTE THE KERNEL SIZE, STRIDE SIZE, PADDING SIZE, AND THE ACTIVATION FUNCTION, RESPECTIVELY. **IN** REPRESENTS THE INPUT CHANNELS, AND **OUT** DENOTES THE OUTPUT CHANNELS.

Generator G.

Layer	Input	Output	K & S & P & A
Conv1	**IN** ×256×192	64×256×192	3×3, 1, 1, ReLU
Conv2	64×256×192	128×256×192	3×3, 1, 1, ReLU
Conv3	128×256×192	128×128×96	3×3, 2, 1, ReLU
Resblock	128×128×96	128×128×96	3×3, 3×3, 1, 1, ReLU
Conv4	128×128×96	256×64×48	3×3, 2, 1, ReLU
Resblock	256×64×48	256×64×48	3×3, 3×3, 1, 1, ReLU
Conv5	256×64×48	512×32×24	3×3, 2, 1, ReLU
Resblock × 4	512×32×24	512×32×24	3×3, 3×3, 1, 1, ReLU
Upsample	512×32×24	512×64×48	-
Conv6	512×64×48	256×64×48	3×3, 1, 1, ReLU
Upsample	256×64×48	256×128×92	-
Conv7	256×128×92	128×128×92	3×3, 1, 1, ReLU
Upsample	128×128×92	128×256×192	-
Conv8	128×256×192	64×256×192	3×3, 1, 1, ReLU
Conv9	64×256×192	**OUT** ×256×192	3×3, 1, 1, ReLU
Conv10	**OUT** ×256×192	**OUT** ×256×192	3×3, 1, 1, Tanh
Conv11	**OUT** ×256×192	**OUT** ×256×192	1×1, 1, 0, None

Discriminator D.

Layer	Input	Output	K & S & P & A
Conv1	**IN** ×256×192	32×256×192	3×1, 1, 0, LeakyReLU
Conv2	32×256×192	32×256×192	3×3, 1, 1, LeakyReLU
Conv3	32×256×192	32×128×96	3×3, 2, 1, LeakyReLU
Conv4	32×128×96	32×128×96	3×3, 2, 1, LeakyReLU
Conv5	32×128×96	64×64×48	3×3, 2, 1, LeakyReLU
Conv6	64×64×48	1 ×64×48	1×1, 1, 0, None

Inception Score (IS). IS [48] is a general metric used to estimate the quality of the synthesis image. In this work, we apply IS to evaluate the quality of generated images by our method. Notably, all generated images used to calculate IS have no corresponding ground-truth images. To evaluate the quality of specific regions (*i.e.*, clothing regions, face regions) in the final virtual try-on image, we also calculate Mask-SSIM and Mask-IS for the extracted regions.

C. Implementation Details

Architecture. Here we provide details about the network architecture of three sub-modules in SPG-VTON. In specific, generators G_1, G_2, and G_3 adopt the same structure, which is a ResNet-like architecture. The generated result of G_2 and G_3 is a 4-channel tensor that could be split into a 1-channel mask and a 3-channel RGB image. We show the detailed network structure of generators and discriminators in TABLE I. In addition, all the discriminators in the proposed method apply the multi-scale structure from pix2pixHD [3]. Additionally, we use instance normalization [49] both in generators and discriminators. We employ ReLU and LeakyReLU activation functions in the generator and discriminator, respectively. Following existing works [29], [50], we adopt LSGAN [51] for all adversarial losses in our method, and a gradient punishment strategy [52] is also used to stabilize the training process.

Setting. In this work, we use the Adam optimizer [53] to optimize generators and discriminators in SPG-VTON, and set the initial learning rate to 0.0002, weight decay to 0.0005, and exponential decay rates $(\beta_1, \beta_2) = (0, 0.999)$. For training, we set hyper-parameters $\lambda_i = 10,\ (i = 1, 2, 4, 5)$, and $\lambda_3 = 20$. Additionally, we set the batch size of semantic map prediction

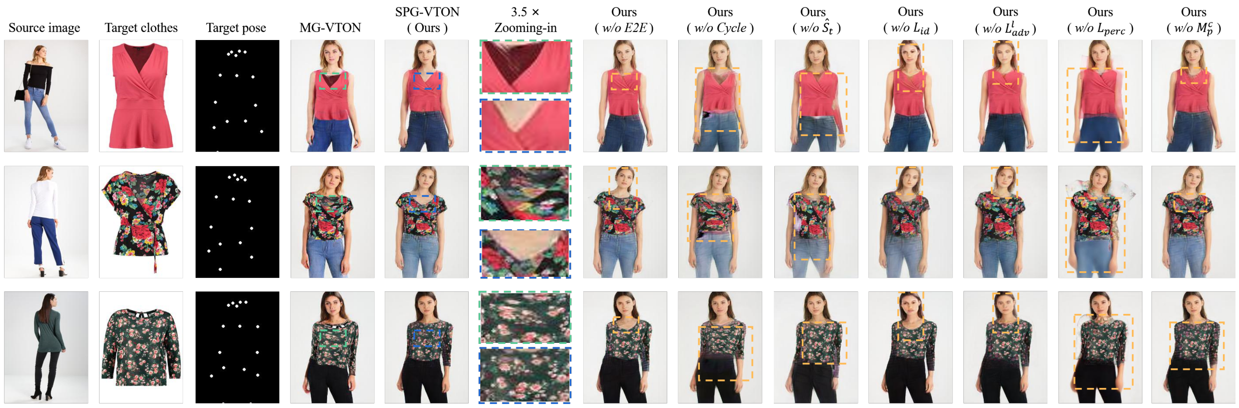


Fig. 5. Visualized comparison between MG-VTON [7] and several variants of our method on the MPV dataset. To make the comparison clearer, we use the green dashed box and the yellow dashed box to cut out the local areas of generated images by MG-VTON and SPG-VTON, respectively. Then, they are enlarged to 3.5 times the original size.

TABLE II
QUANTITATIVE RESULTS. COMPARISON RESULTS IN TERMS OF SSIM AND IS ON BOTH MPV AND DEEPFASHION. ↑ DENOTES THAT HIGHER SCORES ARE BETTER.

Method	MPV		DeepFashion
	SSIM ↑	IS ↑	IS ↑
Real data	1.000	3.391 ± 0.024	3.332 ± 0.126
VITON [5]	0.639	2.394 ± 0.205	2.302 ± 0.116
CP-VTON [6]	0.705	2.519 ± 0.107	2.459 ± 0.212
MG-VTON [7]	0.744	3.154 ± 0.142	3.030 ± 0.057
SPG-VTON (Ours)	**0.752**	**3.243** ± **0.127**	**3.124** ± **0.027**

TABLE III
SPG-VTON *v.s.* MG-VTON [7]. QUANTITATIVE COMPARISON RESULTS IN TERMS OF MASK-SSIM AND MASK-IS ON THE PART OF MPV. ↑ DENOTES THAT HIGHER SCORES ARE BETTER.

Method	Mask-SSIM ↑		Mask-IS ↑	
	Face	w/o Clothes	Face	Clothes
Real data	1.000	1.000	1.848 ± 0.137	3.711 ± 0.237
MG-VTON [7]	0.717 ± 0.094	0.722 ± 0.045	1.419 ± 0.061	3.270 ± 0.409
SPG-VTON (Ours)	**0.737** ± **0.091**	**0.745** ± **0.042**	**1.584** ± **0.053**	**3.660** ± **0.327**

(first process of SPM, generator G_1) to 16, and the batch size of subsequent steps (generators G_2 and G_3, geometric matching models GMM 1 and GMM 2) to 8, and the batch size for joint training is set to 8. In end-to-end training, we train the semantic map prediction network for $70k$, and then with the fixed pre-trained semantic map prediction network, the subsequent network is trained for $70k$. Finally, we jointly train the whole model for $100k$.

D. Quantitative Results

We compare the proposed method with several start-of-the-art virtual try-on methods, including VITON [5], CP-VTON [6], and MG-VTON [7]. VITON and CP-VTON adopt a coarse-to-fine strategy to tackle the virtual try-on task of the single pose, and neither of these two methods includes the change of human pose. To make a fair comparison, we first enrich the input of VITON and CP-VTON by adding the target pose. We report the quantitative results based on the SSIM and IS metrics (higher scores are better) to evaluate the realism of the synthesized virtual try-on images. As shown in TABLE II, our method achieves the maximum SSIM scores, and the maximum IS score on the MPV dataset. In addition, our method also obtains the highest IS scores on the DeepFashion dataset. The results verify the effectiveness of the proposed method on generating high-fidelity virtual try-on images.

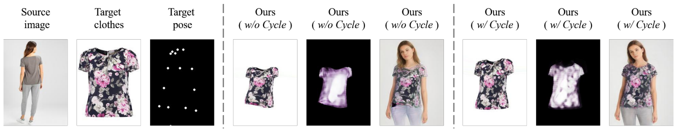


Fig. 6. Qualitative visual results with/without the cycle consistency constraint.

Comparison with MG-VTON [7]. The multi-pose virtual try-on task aims to fit the desired clothes onto the target person image, according to the source image and the given pose. Unlike the pose fixed virtual try-on issue, the multi-pose virtual try-on task is much more challenging to synthesize the whole try-on image, which does not preserve the original body parts. Since the existing methods (such as VITON [5] and CP-VTON [6]) are based on the fixed pose and cannot find the spatial deformation relationship in pose change, directly applying these methods is inappropriate. MG-VTON is the first method to address the multi-pose virtual fitting issue. Hence, to further verify the effectiveness of this method, we conduct additional experiments to compare our method with MG-VTON on the MPV dataset. Consequently, we evaluate both the global metrics SSIM and IS and the local indicators Mask-SSIM and Mask-IS, which estimate the local (*i.e.*, face region and clothes region) similarity and local realism between the generated image and the ground-truth image. Without the official code of MG-VTON, we calculate the results of MG-VTON from the try-on images provided by the original author. Note that these try-on images are generated from part of the test set. To make a fair comparison, we obtain our results from the same test images. Each test sample consists of a three-tuple, including source images, target poses, and desired clothes. The source image and target pose in each tuple in the test set have the same identity, so we calculate the Mask-SSIM value between the generated image and the image corresponding to the target pose. Specifically, we calculate Mask-SSIM

TABLE IV

QUANTITATIVE RESULTS OF DIFFERENT VARIANTS OF SPG-VTON ON THE MPV DATASET. WE HIGHLIGHT THE **BEST** AND THE **SECOND-BEST** PERFORMANCES. ↑ DENOTES THAT HIGHER SCORES ARE BETTER. THE VARIANT OURS W/ S_t REPRESENTS THAT OUR METHOD USES THE GROUND-TRUTH SEMANTIC MAP S_t OF THE TEST IMAGES TO REPLACE THE PREDICTED SEMANTIC MAP $\hat{S}_t$.

Method	SSIM ↑	IS ↑	Mask-SSIM ↑		Mask-IS ↑	
			Face	Clothes	Face	Clothes
Real data	1.000	3.391 ± 0.024	1.000	1.000	2.255 ± 0.030	5.094 ± 0.084
Ours (*full*)	0.752 ± 0.084	3.243 ± 0.127	0.801 ± 0.127	0.872 ± 0.076	1.982 ± 0.024	5.057 ± 0.064
w/o *E2E*	0.747 ± 0.087	2.968 ± 0.031	0.795 ± 0.130	0.869 ± 0.081	1.976 ± 0.013	4.854 ± 0.019
w/o *Cycle*	0.739 ± 0.086	2.745 ± 0.036	0.794 ± 0.127	0.867 ± 0.080	1.801 ± 0.016	4.841 ± 0.096
w/o $\hat{S}_t$	0.736 ± 0.088	2.647 ± 0.017	0.794 ± 0.129	0.861 ± 0.083	1.856 ± 0.009	4.796 ± 0.027
w/o L_{id}	0.732 ± 0.085	2.909 ± 0.069	0.780 ± 0.135	0.866 ± 0.080	1.969 ± 0.045	4.601 ± 0.070
w/o L^l_{adv}	0.700 ± 0.079	**3.358 ± 0.022**	0.786 ± 0.132	0.866 ± 0.080	1.830 ± 0.017	4.546 ± 0.047
w/o L_{perc}	0.724 ± 0.086	**3.474 ± 0.052**	0.798 ± 0.129	0.866 ± 0.081	**2.056 ± 0.015**	**5.373 ± 0.077**
w/o M^c_p	0.747 ± 0.086	2.845 ± 0.012	0.800 ± 0.127	0.869 ± 0.079	1.866 ± 0.016	4.958 ± 0.047
w/o *Noise*	**0.756 ± 0.087**	3.254 ± 0.019	**0.802 ± 0.127**	**0.874 ± 0.079**	1.989 ± 0.017	5.066 ± 0.057
w/ S_t	**0.788 ± 0.080**	3.165 ± 0.028	**0.893 ± 0.081**	**0.923 ± 0.063**	**2.070 ± 0.013**	**5.186 ± 0.071**

TABLE V

QUANTITATIVE RESULTS OF THE EFFECTIVENESS OF FIVE HYPER-PARAMETERS. WE HIGHLIGHT THE **BEST** PERFORMANCES. ↑ DENOTES THAT HIGHER SCORES ARE BETTER.

λ	SSIM ↑	IS ↑	Mask-SSIM ↑		Mask-IS ↑	
			Face	Clothes	Face	Clothes
$\lambda_1 = 0$	0.714 ± 0.078	2.983 ± 0.042	0.781 ± 0.131	0.852 ± 0.086	1.796 ± 0.006	4.916 ± 0.090
$\lambda_2 = 0$	0.554 ± 0.105	**3.985 ± 0.052**	0.784 ± 0.136	0.859 ± 0.081	**2.121 ± 0.032**	4.771 ± 0.048
$\lambda_3 = 0$	0.739 ± 0.085	2.581 ± 0.016	0.792 ± 0.130	0.869 ± 0.078	1.836 ± 0.018	4.151 ± 0.024
$\lambda_4 = 0$	0.420 ± 0.015	3.526 ± 0.059	0.775 ± 0.136	0.857 ± 0.083	2.028 ± 0.013	4.858 ± 0.042
$\lambda_5 = 0$	0.732 ± 0.085	2.909 ± 0.069	0.780 ± 0.135	0.866 ± 0.080	1.969 ± 0.045	4.601 ± 0.070
Ours (*full*)	**0.752 ± 0.084**	3.243 ± 0.127	**0.801 ± 0.127**	**0.872 ± 0.076**	1.982 ± 0.024	**5.057 ± 0.064**

TABLE VI

QUANTITATIVE RESULTS. WE HIGHLIGHT THE **BEST** PERFORMANCES. ↑ DENOTES THAT HIGHER SCORES ARE BETTER.

λ_3	SSIM ↑	IS ↑	Mask-SSIM ↑		Mask-IS ↑	
			Face	Clothes	Face	Clothes
$\lambda_3 = 0$	0.739 ± 0.085	2.581 ± 0.016	0.792 ± 0.130	0.869 ± 0.078	1.836 ± 0.018	4.151 ± 0.024
$\lambda_3 = 1$	0.747 ± 0.087	2.865 ± 0.017	0.795 ± 0.131	0.869 ± 0.081	1.905 ± 0.019	4.869 ± 0.069
$\lambda_3 = 5$	0.745 ± 0.087	2.929 ± 0.081	0.794 ± 0.131	0.869 ± 0.081	**2.008 ± 0.022**	4.749 ± 0.097
$\lambda_3 = 10$	0.744 ± 0.086	2.959 ± 0.033	0.793 ± 0.129	0.868 ± 0.082	2.003 ± 0.020	4.849 ± 0.097
$\lambda_3 = 20$	**0.752 ± 0.084**	**3.243 ± 0.127**	**0.801 ± 0.127**	**0.872 ± 0.076**	1.982 ± 0.024	**5.057 ± 0.064**
$\lambda_3 = 40$	0.746 ± 0.087	2.918 ± 0.026	0.795 ± 0.140	0.868 ± 0.081	1.963 ± 0.017	4.809 ± 0.087
$\lambda_3 = 80$	0.746 ± 0.086	2.910 ± 0.028	0.794 ± 0.131	0.869 ± 0.080	1.952 ± 0.030	4.934 ± 0.012

of the facial area and the area unrelated to clothes between the generated image and the original image that provides the target pose. As shown in TABLE III, the proposed method achieves higher Mask-SSIM scores and Mask-IS scores than MG-VTON, suggesting that our method is better at fitting desired clothes onto the target human image.

E. Qualitative Results

We present visualized comparisons between MG-VTON and several variants of our method on the MPV dataset in Fig. 5. We observe that our method achieves higher-quality try-on results than MG-VTON. In particular, our method generates realistic and natural face regions while preserving the details of desired clothes.

F. Ablation Studies

To study the effectiveness of each component in our method, we compare seven variants of SPG-VTON on the MPV dataset as follows: (1) w/o *E2E*: our method without end-to-end training. Specifically, we train our method in a two-stage manner. We first separately train the semantic prediction network to obtain the preliminary semantic prediction map. Next, we use the semantic prediction map generated from the fixed pre-trained semantic prediction network to guide the subsequent steps; (2) w/o *Cycle*: our method uses $\mathcal{L}_1$ loss between M^C_w and M^c_p to replace the conductible cycle consistency loss in the CWM. In this case, $L_{cond} = \mathbb{E}[\|C_w - C_t\|_1]$; (3) w/o $\hat{S}_t$: our method without the predicted semantic map $\hat{S}_t$. In the training process, we use the predicted target body shape M^b_p to replace the predicted semantic map $\hat{S}_t$; (4) w/o L_{id}: our method removes the face identity loss. In this case, $L_{tsm} = \lambda_4(L^t_{recon} + L^t_{perc}) + L^g_{adv} + L^l_{adv}$; (5) w/o L^l_{adv}: our method without the local adversarial loss. In this case, $L_{tsm} = \lambda_4(L^t_{recon} + L^t_{perc}) + L^g_{adv} + \lambda_5 L_{id}$; (6) w/o L_{perc}: our method without the perceptual loss L^p_{perc} and the perceptual loss L^t_{perc}. In this case, $L_{spm} = L^s_{adv} + \lambda_1 L^s_{fl} + L^s_{recon} + L^p_{adv} + \lambda_2 L^p_{recon}$, and $L_{tsm} = \lambda_4 L^t_{recon} + L^g_{adv} + L^l_{adv} + \lambda_5 L_{id}$;

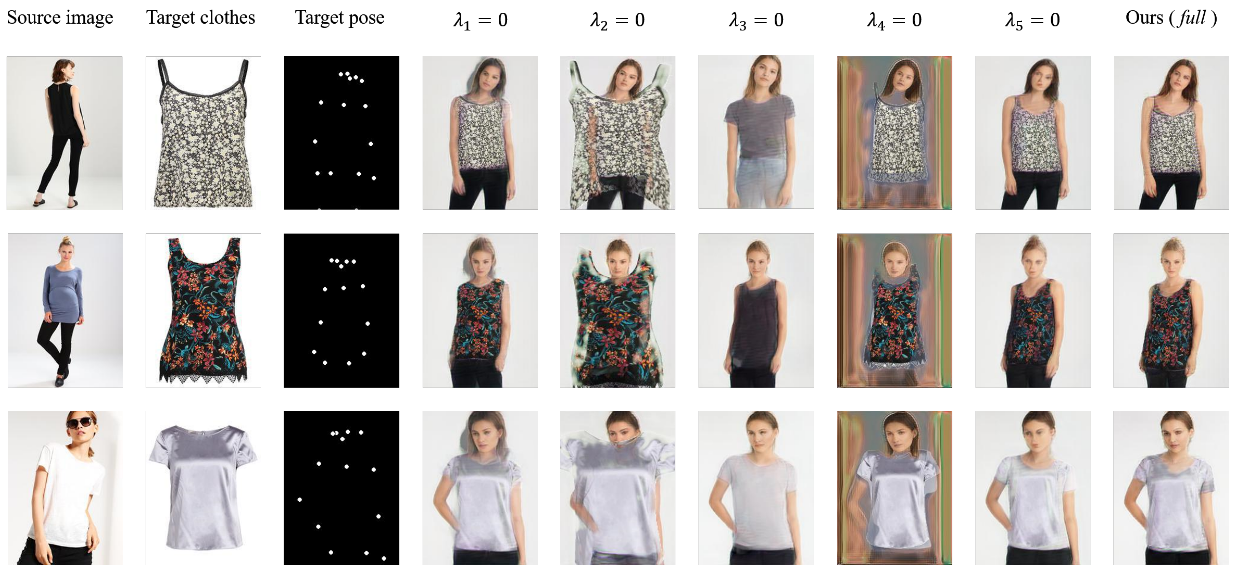


Fig. 7. Qualitative visual comparison of the full model with the variants obtained by setting each of the five hyper-parameters to zero.

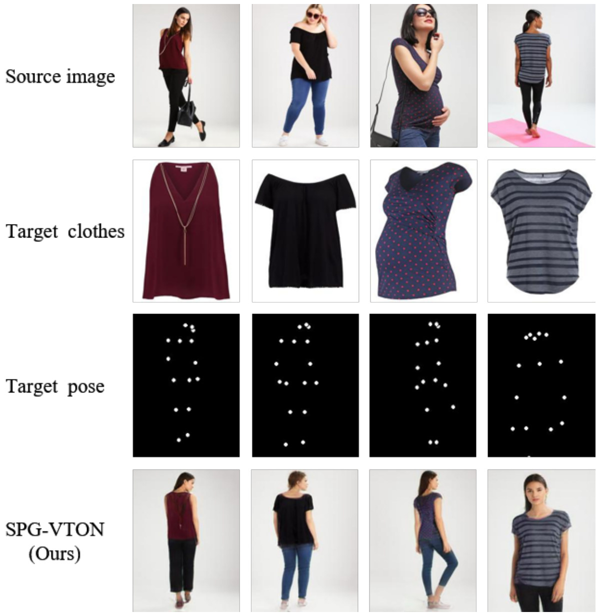


Fig. 8. Examples of try-on results based on noisy images in the test set. For real-world applications, bags, sunglasses and Yoga mat are not desirable in the generated results. In this work, we view these objects as noise for the virtual try-on, and our method is still robust for such obstacles. The model has not seen the source inputs, which are all from the test set.

(7) w/o M_p^c: our method without the predicted clothes mask M_p^c. In the training process, we use the ground-truth clothes mask M_t^c to replace the predicted clothes mask M_p^c. Accordingly, the TPS transformation parameters $theta_1$ are calculated between the mask of desired clothes M^C and the ground-truth clothes mask M_t^c; (8) w/o *Noise*: we use filtered noise free data to train the proposed method; (9) w/ S_t: our method uses the ground-truth semantic map S_t of the test images to replace the predicted semantic map $\hat{S}_t$.

We show the qualitative results of several variants in Fig. 5. We observe that without end-to-end training, our method can still produce plausible results. However, some artifacts existed near the hair regions and the neck regions in the generated images, suggesting that the end-to-end training manner could alleviate the impact of incorrect semantic maps on the generated results. We also find that without the cycle consistency constraint, the clothes regions in the generated images are deformed, which shows that the conductible cycle consistency loss could alleviate the mismatching between the desired clothes and the target person image. In Fig. 6, we show the ablation study on the effect of conductible cycle consistency loss L_{cond}. Additionally, we observe artifacts in the non-clothing regions of the generated results after replacing the predicted target semantic map with the predicted target body shape. This result indicates that without the guidance of the predicted semantic map, our method cannot accurately distinguish between regions of the human body. Moreover, we note that the facial regions in the generated results look less realistic after removing the global and local adversarial loss and the face identity loss, which suggests that introducing the global and local adversarial loss and the face identity loss can encourage the model to generate realistic and natural face regions. Additionally, we observe that removing the perceptual loss causes the model to generate blurred results with artifacts, suggesting that introducing the perceptual loss can synthesize sharper and clearer images. In addition, we find that when our method does not employ the predicted clothes mask, some artifacts appear in the generated image, and distortion exists in the clothing region. These results show that the process of target clothes mask prediction can encourage the model to accurately locate the clothing region and can alleviate the distortion between the desired clothes and the target pose.

We present the quantitative ablation study results of nine variants in TABLE IV. We observe that the full model obtains higher SSIM scores and Mask-SSIM scores than seven different ablation methods except for the training set without noisy data and apply the ground-truth semantic map to replace the predicted semantic map. We also note that without the perceptual loss, IS/Mask-IS scores are higher than those of the whole model. This is due to the perceptual loss with deeper supervision (leveraging the activation of deeper layers), which focuses on semantic patterns. Therefore, the full model with the perceptual loss will improve the SSIM matching quality to the ground-truth in terms of multiple scales, but fail to hold better global matching IS. We surmise that this alternative is still open to the readers whether in considering to use the perceptual loss and that choices depend on the application, which emphasizes the pair matching performance or the global comparison with the whole dataset. The qualitative visual results with/without the perceptual loss are also provided in Fig. 5. Additionally, we find that all indicators are decreased when our method does not implement the predicted clothes mask, suggesting that combining global semantic information with local semantic information could guide the proposed method to generate higher quality fitting results. More importantly, ablation studies also show that the conductible cycle consistency loss could match the shape of the deformed desired clothes with the target person image and maintain the characteristics of the desired clothes in the generated try-on image. We present the qualitative visual results with/without the cycle consistency constraint in Fig. 5 and Fig. 6. The ablation studies also point out that using real semantic maps to replace predicted semantic maps at test time yields higher metric scores, suggesting that the semantic maps obtained by the semantic prediction networks still have a gap with real semantic maps. However, we find that all metrics increase with an end-to-end training manner. In particular, the IS/Mask-IS scores were superior. This result indicates that adopting an end-to-end training manner can enhance the accuracy of predicted semantic maps and lead the model to generate realistic fitting images.

Impact of the introduced five hyper-parameters. To clarify the role of the five hyper-parameters introduced in the objective function (*i.e.*, Eq. 17), we first show the qualitative visual results for the full model with five variants in Fig. 7. We observe that when $\lambda_1 = 0$, our method can still generate plausible fitting images, but artifacts appear. The main reason is that when $\lambda_1 = 0$, the focal loss is invalid. Therefore, the quality of the semantic map generated by SPM only maintains the rough body outline but cannot precisely locate the position of each body component. In addition, when $\lambda_2 = 0$, our method cannot generate reasonable fitting images. The main reason is that when $\lambda_2 = 0$, both L^p_{recon} and L^p_{perc} are invalid, which leads to the inability of SPM to generate coarse results and predicted clothing masks, which are critical inputs for TSM and CWM, respectively. Meanwhile, when $\lambda_3 = 0$, we find that the details of the desired clothes are entirely lost in the generated images, which indicates that the conductible cycle consistency loss plays a vital role in the process of clothes deformation. Additionally, when $\lambda_4 = 0$, we notice that even though the details of the facial and clothing regions in the generated image are well maintained, the entire quality of the generated image is poor due to the lack of global constraints. We also observe that when $\lambda_5 = 0$, the face identity loss is invalid, resulting in less realistic facial regions in the generated results. Furthermore, we report the quantitative results of the effectiveness of five hyper-parameters in TABLE V. The full model obtains the highest SSIM/Mask-SSIM scores and the highest Mask-IS scores for the clothing regions in the generated images. Combining the results of qualitative and quantitative experiments, we consider it necessary to introduce these five hyper-parameters. In addition, to determine the optimal value of the hyper-parameter λ_3, we conduct additional experiments and present the quantitative results in TABLE VI. We first set the values of λ_3 to 0, 1, 5, 10, 20, 40, and 80, and then the experimental results show that when $\lambda_3 = 20$, the highest scores are obtained for all indicators except for the Mask-IS scores for facial regions.

Impact of noisy images in the training set and the test set. We conduct quantitative experiments to verify the impact of noisy data in the training set. As shown in Table IV, we observe that all indicators (*i.e.*, SSIM/Mask-SSIM, IS/Mask-IS) of the proposed method trained by noisy data are slightly lower than the variant trained with noise-free data. These results confirm that noisy training data cause the performance of our method to decrease, but the degree of decrease is still within an acceptable range. Notably, the test set still contains noisy data. Moreover, we provide qualitative visual results based on noisy images (*i.e.*, source images with interference from attributes unrelated to the fitting task, such as bags, glasses, and complicated backgrounds). As shown in Fig. 8, we observe that the proposed method is robust to such training noise and shows good scalability to the unseen test images during inference.

V. Conclusion

In this paper, we propose a novel multi-pose virtual try-on framework (SPG-VTON) based on semantic prediction guidance, which focuses on producing photo-realistic try-on results while fitting the desired clothes onto an arbitrary pose of the same person. SPG-VTON consists of three sub-modules, including the semantic map prediction module, the clothes warping module, and the try-on synthesis module. On the one hand, we introduce a conductible cycle consistency loss that can alleviate the mismatching between the desired clothes and the target image. On the other hand, we also apply a face identity loss to make the face region of the final virtual try-on image look natural and to preserve the identity of the source image. Extensive qualitative and quantitative experiments demonstrate that the proposed method outperforms previous state-of-the-art methods and has good scalability to the training data noise as well as the unseen test images during inference. In the future, we will continue to explore the application of the proposed method to new fields, such as vehicle appearance design [54] and language-based cloth generation [55], and new modalities [56], [57].

References

[1] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *CVPR*, 2017, pp. 1125–1134. 1, 3

[2] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *ICCV*, 2017, pp. 2223–2232. 1, 3

[3] 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 *CVPR*, 2018, pp. 8798–8807. 1, 3, 5, 8

[4] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *CVPR*, 2019, pp. 4401–4410. 1, 3

[5] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "Viton: An image-based virtual try-on network," in *CVPR*, 2018, pp. 7543–7552. 1, 2, 3, 4, 5, 6, 9

[6] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *ECCV*, 2018, pp. 589–604. 1, 2, 3, 4, 5, 6, 9

[7] 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 *ICCV*, 2019, pp. 9026–9035. 1, 2, 3, 4, 5, 6, 7, 8, 9

[8] R. Yu, X. Wang, and X. Xie, "Vtnfp: An image-based virtual try-on network with body and clothing feature preservation," in *ICCV*, 2019, pp. 10 511–10 520. 1, 3, 4, 5, 6

[9] X. Han, X. Hu, W. Huang, and M. R. Scott, "Clothflow: A flow-based model for clothed person generation," in *ICCV*, 2019, pp. 10 471–10 480. 1, 3, 4, 5

[10] T. Issenhuth, J. Mary, and C. Calauzènes, "End-to-end learning of geometric deformations of feature maps for virtual try-on," *arXiv:1906.01347*, 2019. 1, 3

[11] N. Zheng, X. Song, Z. Chen, L. Hu, D. Cao, and L. Nie, "Virtually trying on new clothing with arbitrary poses," in *ACM MM*, 2019, pp. 266–274. 1, 3, 4, 5, 6

[12] Y. Han, Z. Ruimao, G. Xiaobao, L. Wei, Z. Wangmeng, and L. Ping, "Towards photo-realistic virtual try-on by adaptively generating-preserving image content," in *CVPR*, 2020. 1, 3, 5, 6

[13] 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 Transactions on Neural Networks and Learning Systems*, 2021. 1

[14] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, and P. Luo, "Parser-free virtual try-on via distilling appearance flows," in *CVPR*, 2021, pp. 8485–8493. 1

[15] X. Gao, Z. Liu, Z. Feng, C. Shen, K. Ou, H. Tang, and M. Song, "Shape controllable virtual try-on for underwear models," *arXiv:2107.13156*, 2021. 1

[16] Z. Xie, X. Zhang, F. Zhao, H. Dong, M. C. Kampffmeyer, H. Yan, and X. Liang, "WAS-VTON: Warping Architecture Search for Virtual Try-on Network," in *ACM MM*, 2021. 1

[17] Z. Lu and L. Wang, "Noise-robust semi-supervised learning via fast sparse coding," *Pattern Recognition*, vol. 48, no. 2, pp. 605–612, 2015. 2

[18] Y. Qi, L. Qin, J. Zhang, S. Zhang, Q. Huang, and M.-H. Yang, "Structure-aware local sparse coding for visual tracking," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3857–3869, 2018. 2

[19] Z. Zhou, J. Li, Y. Quan, and R. Xu, "Image quality assessment using kernel sparse coding," *IEEE Transactions on Multimedia*, vol. 23, pp. 1592–1604, 2020. 2

[20] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "Deepfashion: Powering robust clothes recognition and retrieval with rich annotations," in *CVPR*, 2016, pp. 1096–1104. 3, 8

[21] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *NeurIPS*, 2014, pp. 2672–2680. 3, 5

[22] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv:1411.1784*, 2014. 3, 5

[23] Z. Zheng, L. Zheng, and Y. Yang, "Unlabeled samples generated by gan improve the person re-identification baseline in vitro," in *ICCV*, 2017, pp. 3754–3762. 3

[24] H. Dong, X. Liang, Y. Zhang, X. Zhang, Z. Xie, B. Wu, Z. Zhang, X. Shen, and J. Yin, "Fashion editing with multi-scale attention normalization," *arXiv:1906.00884*, 2019. 3

[25] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen, "Stgan: A unified selective transfer network for arbitrary image attribute editing," in *CVPR*, 2019, pp. 3673–3682. 3

[26] Z. Huang, Z. Zheng, C. Yan, H. Xie, Y. Sun, J. Wang, and J. Zhang, "Real-world automatic makeup via identity preservation makeup net." in *IJCAI*, 2020, pp. 652–658. 3

[27] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," in *NeurIPS*, 2017, pp. 406–416. 3

[28] L. Ma, Q. Sun, S. Georgoulis, L. Van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *CVPR*, 2018, pp. 99–108. 3

[29] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang, and J. Kautz, "Joint discriminative and generative learning for person re-identification," in *CVPR*, 2019, pp. 2138–2147. 3, 8

[30] P. Esser, E. Sutter, and B. Ommer, "A variational u-net for conditional appearance and shape generation," in *CVPR*, 2018, pp. 8857–8866. 3

[31] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv:1312.6114*, 2013. 3

[32] A. Grigorev, A. Sevastopolsky, A. Vakhitov, and V. Lempitsky, "Coordinate-based texture inpainting for pose-guided image generation," *arXiv:1811.11459*, 2018. 3

[33] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable gans for pose-based human image generation," in *CVPR*, 2018, pp. 3408–3416. 3

[34] H. Dong, X. Liang, K. Gong, H. Lai, J. Zhu, and J. Yin, "Soft-gated warping-gan for pose-guided person image synthesis," in *NeurIPS*, 2018, pp. 474–484. 3

[35] S. Song, W. Zhang, J. Liu, and T. Mei, "Unsupervised person image generation with semantic parsing transformation," in *CVPR*, 2019, pp. 2357–2366. 3

[36] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE transactions on pattern analysis and machine intelligence*, vol. 11, no. 6, pp. 567–585, 1989. 3

[37] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE transactions on pattern analysis and machine Intelligence*, vol. 24, no. 4, pp. 509–522, 2002. 3

[38] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *CVPR*, 2017, pp. 6148–6157. 3, 4, 5, 6

[39] W. Liu, Y. Wen, Z. Yu, M. Li, B. Raj, and L. Song, "Sphereface: Deep hypersphere embedding for face recognition," in *CVPR*, 2017, pp. 212–220. 4, 7

[40] P. Li, Y. Xu, Y. Wei, and Y. Yang, "Self-correction for human parsing," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 4

[41] 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 *CVPR*, 2017, pp. 932–940. 4, 5

[42] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017, pp. 7291–7299. 4

[43] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241. 5

[44] Y. Yan, J. Xu, B. Ni, W. Zhang, and X. Yang, "Skeleton-aided articulated motion generation," in *Proceedings of the 25th ACM international conference on Multimedia*, 2017, pp. 199–207. 5

[45] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European conference on computer vision*. Springer, 2016, pp. 694–711. 5, 6

[46] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015. 5

[47] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 8

[48] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training gans," in *NeurIPS*, 2016, pp. 2234–2242. 8

[49] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," *arXiv:1607.08022*, 2016. 8

[50] B. Hu, Z. Zheng, P. Liu, W. Yang, and M. Ren, "Unsupervised eyeglasses removal in the wild," *IEEE Transactions on Cybernetics*, 2020, doi:10.1109/TCYB.2020.2995496. 8

[51] X. Mao, Q. Li, H. Xie, R. Y. Lau, Z. Wang, and S. Paul Smolley, "Least squares generative adversarial networks," in *ICCV*, 2017, pp. 2794–2802. 8

[52] L. Mescheder, A. Geiger, and S. Nowozin, "Which training methods for gans do actually converge?" *arXiv:1801.04406*, 2018. 8

[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv:1412.6980*, 2014. 8

[54] Z. Zheng, T. Ruan, Y. Wei, Y. Yang, and T. Mei, "Vehiclenet: Learning robust visual representation for vehicle re-identification," *IEEE Transactions on Multimedia (TMM)*, 2020, doi:10.1109/TMM.2020.3014488. 12

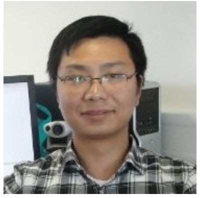
[55] Z. Zheng, L. Zheng, M. Garrett, Y. Yang, M. Xu, and Y.-D. Shen, "Dual-path convolutional image-text embeddings with instance loss," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 2, pp. 1–23, 2020, doi:10.1145/3383184. 12

[56] Y. Yang, Y. Zhuang, and Y. Pan, "Multiple knowledge representation for big data artificial intelligence: framework, applications, and case studies," *Frontiers of Information Technology & Electronic Engineering*, vol. 22, no. 12, pp. 1551–1558, 2021. 12

[57] L. Zhu, H. Fan, Y. Luo, M. Xu, and Y. Yang, "Temporal cross-layer correlation mining for action recognition," *IEEE Transactions on Multimedia (TMM)*, pp. 1–1, 2021. 12

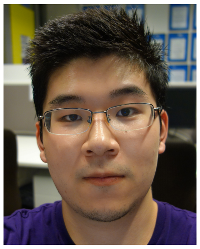
Bingwen Hu is currently pursuing the Ph.D. degree with Nanjing University of Science and Technology, Nanjing, China.

He is also a Joint Ph.D. Student with the University of Technology Sydney. His research interests include computer vision, image processing, and pattern recognition.



Ping Liu received the bachelor's degree in electrical engineering from the Wuhan University of Technology, Wuhan, China, in 2005, the master's degree from the Huazhong University of Science and Technology, Wuhan, in 2008, and the Ph.D. degree in computer science and engineering from the University of South Carolina, Columbia, SC, USA, in 2015. From 2018 to 2020, he was a Research Staff with the Center for Artificial Intelligence, University of Technology Sydney, Ultimo, NSW, Australia.

He is currently a Scientist with the Center for Frontier AI Research (CFAR), A*STAR, Singapore. His research interests include computer vision, machine learning, and deep learning.



Zhedong Zheng received the Ph.D. degree from the University of Technology Sydney, Australia, in 2021 and the B.S. degree from Fudan University, China, in 2016. He is currently a postdoctoral research fellow at NExT++, School of Computing, National University of Singapore. He was an intern at Nvidia Research (2018) and Baidu Research (2020). His research interests include robust learning for image retrieval, generative learning for data augmentation, and unsupervised domain adaptation.

Mingwu Ren received the Ph.D. degree in pattern recognition and intelligent system from Nanjing University of Science and Technology (NUST), Nanjing, Jiangsu, China, in 2001.

He is currently a Professor with the School of Computer Science and Engineering, NUST. His research interests include computer vision, image processing, and pattern recognition.