IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 28, 2019, accepted September 6, 2019, date of publication September 13, 2019,
date of current version September 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2941378

Inpainting-Based Virtual Try-on Network for Selective Garment Transfer

LI YU[1], YUEQI ZHONG[1,2], AND XIN WANG[1]
[1]College of Textiles, Donghua University, Shanghai 201620, China
[2]Key Laboratory of Textile Science and Technology, Ministry of Education, College of Textiles, Donghua University, Shanghai 201620, China

Corresponding author: Yueqi Zhong (zhyq@dhu.edu.cn)

This work was supported by the National Natural Science Foundation of China under Grant 61572124.

ABSTRACT Image-based garment transfer systems aim to swap the desired clothes from a model to arbitrary users. However, existing works cannot provide the capacity for users to try on various fashion articles according to their wishes, i.e., users can decide which article (e.g., tops, pants or both) to be swapped. In this paper, we propose an Inpainting-based Virtual Try-On Network (I-VTON) which allows the user to try on arbitrary clothes from the model image in a selective manner. To realize the selectivity, we reshape the virtual try-on as a task of image inpainting. Firstly, the texture from the garment and the user are extracted respectively to form a coarse result. In this phase, users can decide which clothes they hope to try on via an interactive texture control mechanism. Secondly, the missing regions in the coarse result are recovered via a Texture Inpainting Network (TIN). We introduce a triplet training strategy to ensure the naturalness of the final result. Qualitative and quantitative experimental results demonstrate that I-VTON outperforms the state-of-the-art methods on both the garment details and the user identity. It is also confirmed our approach can flexibly transfer the clothes in a selective manner.

INDEX TERMS Generative adversarial network, self-supervised learning, texture inpainting, virtual try-on.

I. INTRODUCTION

Recent researches on image synthesis [1]–[4] bring an interesting virtual try-on technology, namely image-based garment transfer [5], [6]. Given a user image and a model image, its task is to synthesize a new image of the user wearing the model's clothes. It is expected to be a promising technology for online shopping where impulse buying and fast fashion are growing trends [7], [8]. Imagine a female/male customer wants to know the compatibility of her/his own pants against the model's top, only her/his top needs to be swapped. In such scenarios, a selective garment transfer system, which can swap arbitrary clothes according to the user's wishes, is more practical, as illustrated in Fig. 1.

However, previous works on image-based virtual try-on [5], [6], [9], [10] did not take this interaction with users into consideration. Han *et al.* [9] and Wang *et al.* [10] both required in-shop clothes images, which are inconvenient for different types of clothes since prior knowledge about dressing is needed. Furthermore, the requirement of clean product images makes it somehow impractical in the real world. More recently, Raj *et al.* [5] and Wu *et al.* [6] proposed the garment transfer approaches which replaced the clothes image with the image of the model wearing the clothes. But their works are inefficient in transferring specified clothes from the model image, since they only support all clothes swapping [5] or need to train extra networks for different types of clothes [6]. Therefore, it is a challenge to make the garment transfer in a selective manner.

Moreover, most related approaches only focus on preserving the detailed garment. To eliminate the effect of the user's original clothes, these methods replaced the user texture with an abstract person representation [9], [10] or mainly utilized the texture from the model image [5], [6]. As a result, these methods often fail to generate reasonable user texture (e.g., skin and hair). Maintaining the user texture, which indicates the user identity, is equally important, especially for the selective garment transfer. The reason is that the rest region of the user should not be changed when swapping the specified clothes.

Additionally, it is difficult to prepare a dataset for the task of garment transfer, which requires each training sample contains the user image, the model image, and the image of the user wearing the model's clothes. And it would be more

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

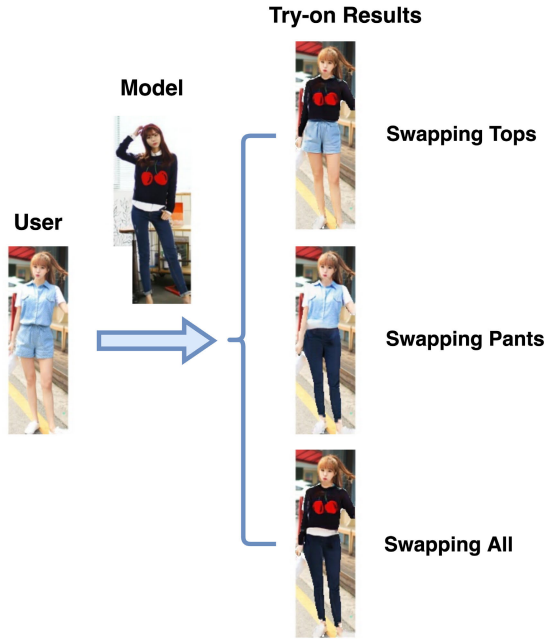


FIGURE 1. The selectivity of our method. By providing the user image and the model image, our method allows the user to choose which clothes she hopes to try-on.

complex for the selective garment transfer. Current methods solved this problem by using the self-supervised training. Raj *et al.* [5] used a single image with its augmentations as a training sample, and Wu *et al.* [6] proposed an unpair-pair joint training strategy. But the texture details in the results are still not preserved well for these methods.

To solve the aforementioned challenges, we propose an end-to-end framework called Inpainting-based Virtual Try-On Network (I-VTON) to transfer arbitrary clothes from the model to the user in a selective manner. It only requires the users to provide their own images and the choices of which clothes in the model images to try-on. We consider our task as a problem of image inpainting [11], [12] to preserve the maximum texture details, and introduce an interactive texture control mechanism to achieve the selectivity, i.e., it enables a user to try on specific regions (e.g. top, bottom).

More specifically, I-VTON includes three stages: 1) Remapping the garment texture from the model image according to the dense pose results. 2) Synthesizing a plausible human parsing of the user wearing the desired clothes via Parsing Inference Network (PIN) followed by extracting the user texture based on it. 3) Combining the two textures into a coarse result which may contain inevitable missing regions and refining it via a Texture Inpainting Network (TIN).

A triplet training strategy and a skin loss are applied to improve the final results. Qualitative and quantitative experiments demonstrate that I-VTON outperforms four state-of-the-art methods on a dataset proposed by Han *et al.* [9]. We also show the selectivity of our method by defining three swapping ways.

Our main contributions are as follows:

- To our best knowledge, this is the first selective garment transfer framework that can flexibly transfer the clothes in a selective manner.
- By converting our task into image inpainting, we introduce an interactive texture control mechanism to avoid training extra networks, and we introduce a triplet training strategy to recover the missing regions more naturally.
- We make full use of the user texture by inferring a plausible human parsing and introduce a skin loss to maintain the skin color of users.

II. RELATED WORK

A. HUMAN REPRESENTATION

Extracting human representation can be regarded as a preprocessing step for conditional human generation including the task of virtual try-on. Previous works [5], [9], [10], [13] extract human representation by either human pose estimation [14] or human parsing [15] or both. Human pose estimation provides global human structure via body joints, whereas human parsing focuses on pix-level understanding via fine-grained semantics (e.g., body parts and clothing) [16].

More recently, Güler *et al.* [17] proposed DensePose dataset to build correspondences from RGB images to 3D surfaces of the human body base on the skinned multi-person linear model (SMPL) [18]. They also proposed DensePose-RCNN to predicate dense pose points of human images. As the dense pose provides surface-based texture information, we can remap the texture of desired clothes to align with the user image. It is noteworthy that the dense pose includes not only UV coordinates but also body segmentation. Both are used in our framework.

B. GENERATIVE ADVERSARIAL NETWORK

Image-based virtual try-on technology is essentially image synthesis for generating a new image based on the visual information of humans and clothes. Since a new framework Generative Adversarial Network (GAN) proposed in 2014 [19], many researchers [1], [3], [20], [21] demonstrated its great potential to generate photo-realistic images.

1) POSE-CONDITIONED HUMAN IMAGE GENERATION

Pose-conditioned human image generation [13], [22]–[24], which aims to generate a new person image in a different pose based on a reference image, is most related to the image-based garment transfer task [5], [6]. One of the big challenges for both tasks is how to preserve the texture details from the reference image. Ma *et al.* [22] solved it by using a two-stage coarse-to-fine strategy. Deformable GANs [13] introduced a deformable skip connection to align the feature map according to the 2D pose keypoints. Neverova [24] proposed a more accurate method to warp the texture by utilizing the dense pose results [17].

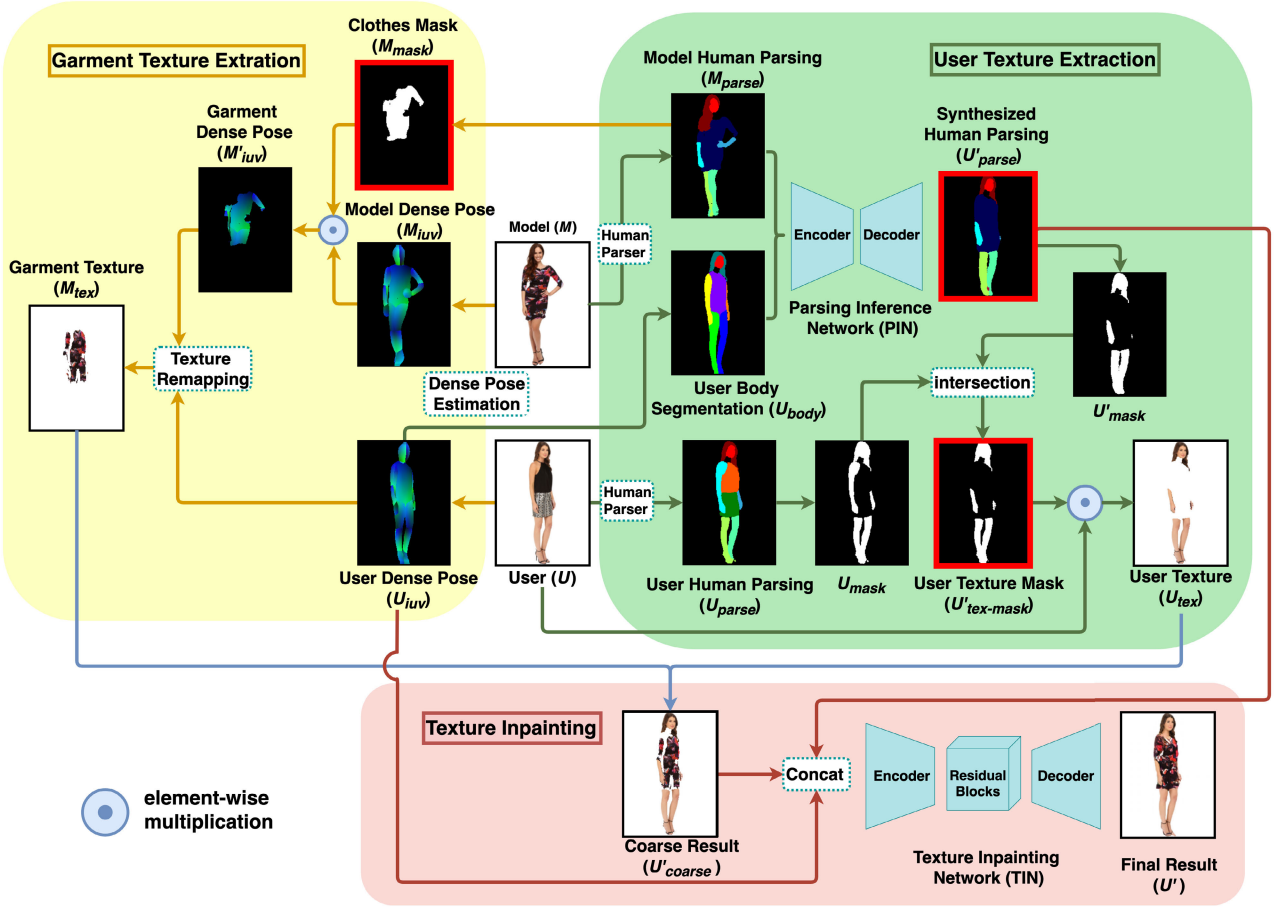


FIGURE 2. An overview of I-VTON. We extract the garment texture M_{tex} and the user texture U_{tex} separately, then refine the coarse result U'_{coarse}. Our framework consists of three modules: 1) We remapping the garment texture M_{tex} based on the dense pose results M'_{iuv} and U_{iuv}; 2) We extract the user texture U_{tex} based on the human parsing U'_{parse} and U_{parse}; 3) We refine the coarse result U'_{coarse} by Texture Inpainting Network (TIN). Note that we can alternate M_{mask}, U'_{parse}, and $U'_{tex-mask}$ (in red boxes) to control which clothes to be swapped.

2) IMAGE INPAINTING

We convert our task to image inpainting, which has attracted great interest in the computer vision community. Image inpainting is a task to naturally recover the missing pixels in an image. The traditional methods [25], [26] solve it by diffusing background data or copying image patches into missing regions, which cannot deal with more challenging cases such as humans and clothes. Recently, many related works [11], [12], [27], [28] focused on GAN based methods, which can provide more perceptually realistic results. Liu *et al.* [11] and Yu *et al.* [12] both modified the convolution layers to better handle irregular missing regions. Nazeri *et al.* [28] proposed a method to recover the image based on a plausible edge map from an edge generator.

C. IMAGE-BASED VIRTUAL TRY-ON

Early works [29]–[31] for virtual try-on applications were mainly based on computer graphics technology. Though these works achieve precise physical simulation, most works require 3D measurements. Image-based virtual try-on, which can generate photo-realistic try-on results from 2D images, is more convenient and cheaper than the 3D based approaches.

Existing works in image-based virtual try-on can be divided into two groups according to the type of input. The first group takes a user image and an in-shop clothes image as the input (user and article). The second group takes a user image and a model image as the input (user and model). Both groups aim to synthesize a new image of the user wearing the clothes from the product image or the model image.

1) USER AND ARTICLE

Jetchev and Bergmann [32] proposed a conditional analogy GAN for swapping clothes by fusing Conditional GAN [33] and Cycle GAN [20]. Due to the convolutional layer lacking the capability in dealing with large spatial misalignment, the output is always blurry. To preserve the detail of garment texture, Han *et al.* [9] and Wang *et al.* [10] both geometrically deform the clothes image to match the user image for preserving the garment characteristics. However, these methods still have not solved the problem of matching different types of clothes with the user, especially for dresses and skirts.

2) USER AND MODEL

Recently, Raj *et al.* [5] proposed SwapNet to transfer the garment from the model to the user without in-shop clothes image. In the warping stage of SwapNet, similar to Fashion-GAN [34], they trained a network to infer the plausible human parsing of the user in desired clothes. However, the method only supports all clothes swapping with a low output resolution (128× 128). Wu *et al.* [6] proposed M2E-Try On Net (M2E-TON) by converting the task into pose-conditioned human generation [13], [24]. They first generated an image of the model in the user's pose, then refined it and added the user identity based on the dense poses [17]. They did not consider the mismatch of the skin color between the user and the output. Furthermore, extra networks are required for different types of clothes.

III. INPAINTING-BASED VIRTUAL TRY-ON NETWORK

We propose an Inpainting-based Virtual Try-On Network (I-VTON) which can flexibly transfer the selected clothes from the model image according to the user's wishes. Given a user image U and a model image M, our goal is to synthesize a photo-realistic image U', which depicts the user wearing the desired clothes extracted from M while maintaining the original pose. We extract the garment texture M_{tex} (see Section III-A) and the user texture U_{tex} (see Section III-B) separately, then combine them together as a coarse result U'_{coarse} and recover the missing texture to generate the final output U' (see Section III-C). Furthermore, we achieve the selectivity via an interactive texture control mechanism (see Section III-D). The overall pipeline is illustrated in Fig. 2.

A. GARMENT TEXTURE EXTRACTION

We can obtain the garment texture according to the model human parsing M_{parse}, but the desired clothes are not aligned with the user image U. We solve this problem by utilizing the dense poses of the model and the user (denoted as M_{iuv} and U_{iuv} respectively). The desired clothes mask M_{mask} can be directly extracted from M_{parse}. If a pixel in M belongs to the desired clothes, its mask value is set to 1; otherwise 0. We obtain the desired clothes dense pose as

$$M'_{iuv} = M_{mask} \odot M_{iuv}, \tag{1}$$

where $\odot$ indicates the element-wise multiplication, M_{iuv} is the model dense pose, M'_{iuv} is the desired garment dense pose.

In DensePose [17], the human body is subdivided into 24 body parts and the UV fields for each body part are defined according to the SMPL model [18]. The texture maps are obtained by mapping the pixels in 2D human images according to the UV points. These maps can be regarded as "charts" to build the association between the 2D human image and the 3D body surface, as shown in Fig. 3. Though U and M are in different human poses, body shape and perspective, they can be mapped onto the common texture maps. Thus we can remap the texture maps of M according to the UV coordinates of U to "warp" the desired clothes of M.

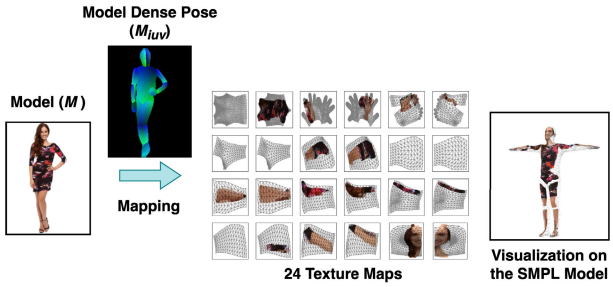


FIGURE 3. Visualizing the 24 texture maps on an SMPL model.

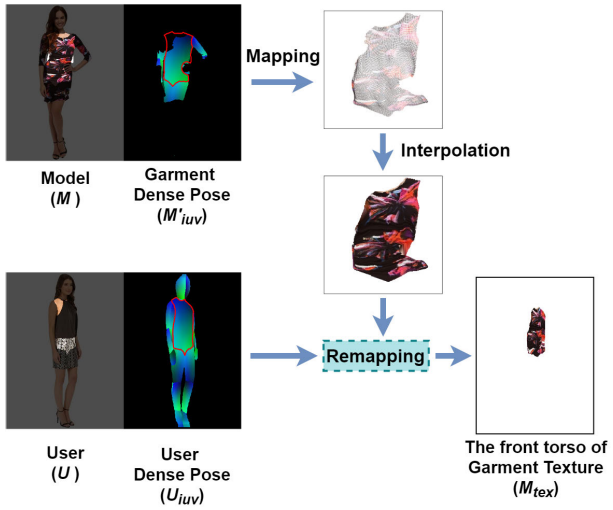


FIGURE 4. The pipeline of texture remapping for the front torso. The red circles in M'_{iuv} and U_{iuv} indicate the front torso regions which are defined by DensePose.

More specifically, we map the pixels of M onto 24 common texture maps according to M'_{iuv} and then obtain the texture of the desired clothes M_{tex} by extracting the corresponding pixels from each texture map according to U_{iuv}.

Take font torso for example, as shown in Fig. 4, firstly, we map the RGB pixels of the torso in M onto a 256×256 texture map according to the UV coordinates in the torso part of M'_{iuv}. Secondly, since the texture map consists of discrete pixel points, we create a 2D Delaunay triangulation of these points [35] and perform linear barycentric interpolation [36] on each triangle to get a continuous texture map. Finally, we retrieve the corresponding RGB pixels from the texture map via the UV coordinates in the torso part of U_{iuv}.

It is inevitable that some points in M_{tex} may have no corresponding pixels which are regarded as missing regions. Additionally, we found there is a mirror-symmetric part for each body part except torso in the dense pose. Therefore, we retrieve the pixels from both the original and symmetric texture maps which contribute to a smaller missing area of M_{tex}, as shown in Fig. 5.

B. USER TEXTURE EXTRACTION

For extracting more user texture, we firstly generate a plausible human parsing of U' (denoted as U'_{parse}), which contains

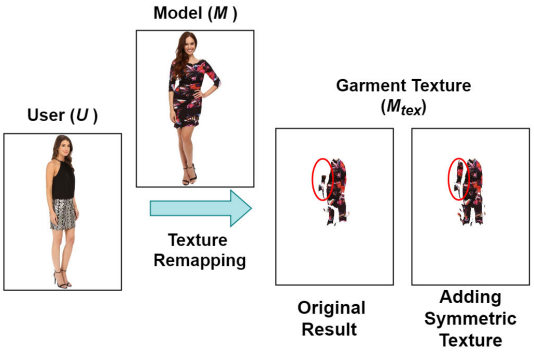


FIGURE 5. The visual comparison of whether adding symmetric texture maps.

not only the target garment information but also the user information (e.g., face and limbs). Inspired by SwapNet [5], we use the body segmentation of U (denoted as U_{body}) and the human parsing of M (denoted as M_{parse}) as input, to infer U'_{parse} via Parsing Inference Network (PIN). Therefore, PIN is a dual-path auto-encoder, As shown in Fig. 5. To eliminate the effect of hair class in M_{parse}, we copy both face and hair classes from U_{parse} into U_{body}. Note that the dense pose result includes body segmentation, i.e., we directly obtain U_{body} from U_{iuv}.

Now the user texture can be obtained by utilizing the user parsing U_{parse} and the synthesized parsing U'_{parse}. More specifically, we obtain two binary masks (denoted as U_{mask} and U'_{mask} respectively) by merging the parsing classes needed to be preserved (e.g., face and arms) from U_{parse} and U'_{parse} separately. The user texture mask $U'_{tex-mask}$ is the intersection of U_{mask} and U'_{mask}:

$$U'_{tex-mask} = U_{mask} \cap U'_{mask}, \tag{2}$$

where $\cap$ indicates the intersection operation.

The reason we extract the mask from both U_{parse} and U'_{parse} instead of only U_{parse} is that U_{parse} implicates the user's original garment information which will affect the final output. As shown in Fig. 2, for a task to swap a half-sleeve dress (target article) from the model to a user wearing a sleeveless top, if we only use the mask from U_{parse}, the arm texture will overlap with the sleeve texture. As a result, the final output would contain a sleeveless dress that is inconsistent with the target article. Hence the intersection operation is critical to make full use of the user texture.

C. TEXTURE INPAINTING NETWORK

After obtained the desired garment texture M_{tex} (as mentioned in Section III-A) and the user texture mask $U'_{tex-mask}$ (as mentioned in Section III-B), we merge them into a coarse try-on result as

$$U'_{coarse} = U \odot U'_{tex-mask} + M_{tex} \odot (1 - U'_{tex-mask}), \tag{3}$$

where $\odot$ indicates the element-wise multiplication and U is the user image.

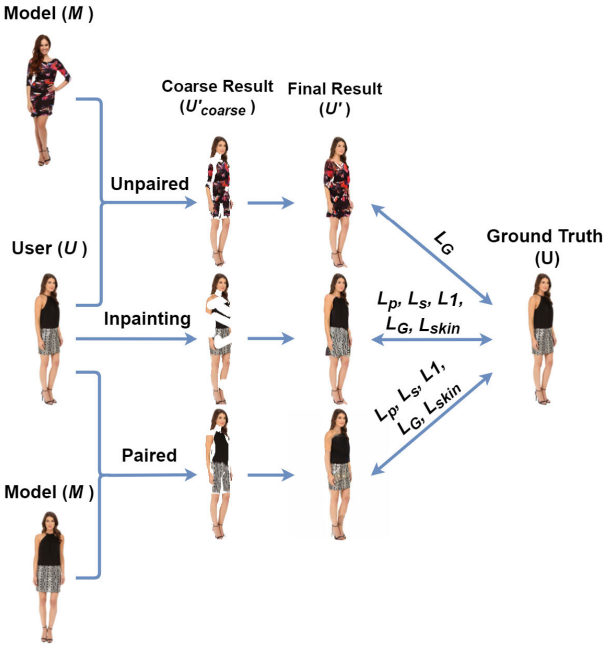


FIGURE 6. Our triplet training strategy which contains the unpaired sample, the inpainting sample, and the paired sample.

The result U'_{coarse} has approximated the final output U' except some missing regions which are inevitable. The task of refining U'_{coarse} can be considered as image inpainting for irregular holes [11], [12]. Therefore, we propose Texture Inpainting Network (TIN) to recover the missing regions in U'_{coarse}. The input of the network is the concatenation of U_{iuv}, U'_{parse} (encoded as an RGB image), and U'_{coarse}. As shown in Fig. 2, TIN is an auto-encoder with several residual blocks [37] (we use 6 blocks in our experiments). Since we obtain the mask of missing regions U'_{hole} in Section III-A, we use partial convolutions [11] for TIN with U'_{hole}. We also add the self-attention module [3] and spectral normalization [38] to further improve the results.

Inspired by the unpair-pair joint training [6], we propose a triplet training strategy that contains the unpaired, paired, and inpainting training samples, as shown in Fig. 6.

1) UNPAIRED SAMPLE

The unpaired sample is the same as the input during the testing stage, which includes two images of the different people wearing different clothes in different poses. Obviously, there is no ground truth in this case. Therefore, we only use a hinge version of GAN loss [39]:

$$\begin{aligned} \mathcal{L}_G &= -\mathbb{E}[D(U')], \\ \mathcal{L}_D &= -\mathbb{E}[\min(0, -1 + D(U))] \\ &\quad - \mathbb{E}[\min(0, -1 - D(U'))], \end{aligned} \tag{4}$$

where D is a discriminator, U is the user image, U' is the result generated by TIN, $\mathcal{L}_G$ is the generator loss, $\mathcal{L}_D$ is the discriminator loss.

2) PAIRED SAMPLE

The paired sample includes two images of a person wearing identical clothes in different poses; one is U, and the other is M. Since the ground truth U exists in this case, we use perceptual loss [40] and style loss [41], which can measure semantic differences via an ImageNet-pretrained VGG network [42], written as follows:

$$\mathcal{L}_p = \sum_{i=1}^{5} \|\phi_i(U) - \phi_i(U')\|_2, \tag{5}$$

$$\mathcal{L}_s = \sum_{i=1}^{5} \|\mathcal{G}_{\phi_i}(U) - \mathcal{G}_{\phi_i}(U')\|_F^2, \tag{6}$$

where ϕ_i is the i-th selected feature map in VGG-19 network (we use conv1_2, conv2_2, conv3_2, conv4_2, and conv5_2 in our implementation), $\mathcal{G}_{\phi_i}$ represents the Gram matrix of the feature map ϕ_i, U is the user image, U' is the final result, $\mathcal{L}_p$ is the perceptual loss which calculates the Euclidean distance between the feature maps of U and U', $\mathcal{L}_s$ is the style loss which calculates the Frobenius norm of the difference between the Gram matrices of U and U'.

We also apply pixel-wise L1 loss [1] and GAN loss (4) for TIN. Furthermore, a skin loss $\mathcal{L}_{skin}$ is introduced to ensure that the final output maintains the skin color of the user. It is defined as

$$\mathcal{L}_{skin} = \|U - U'\|_1 \odot U_{skin}, \tag{7}$$

where $\odot$ indicates the element-wise multiplication and U_{skin} indicates the mask of the exposed skin (e.g., neck and arms) in U. The total loss $\mathcal{L}_{TIN}$ for the paired sample is the sum of the above losses:

$$\mathcal{L}_{TIN} = \lambda_p\mathcal{L}_p + \lambda_s\mathcal{L}_s + \lambda_1\mathcal{L}_{L1} + \lambda_G\mathcal{L}_G + \lambda_{skin}\mathcal{L}_{skin}, \tag{8}$$

where $\mathcal{L}_p$ is the perceptual loss, $\mathcal{L}_s$ is the style loss, $\mathcal{L}_{L1}$ is the pixel-wise L1 loss, $\mathcal{L}_G$ is the generator loss, $\mathcal{L}_{skin}$ is the skin loss. λ_p, λ_s, λ_1, λ_G and λ_{skin} are the weights for each loss.

3) INPAINTING SAMPLE

For the inpainting sample, only a single image (i.e., U and M are the same image) is required. We draw some random irregular holes on the user image U as U'_{coarse} and regard the original image U as the ground truth. We use the same objective function (8) as the paired sample. The inpainting samples ensure that TIN learns to fill the missing regions in U'_{coarse} naturally (see Section IV-F).

D. INTERACTIVE TEXTURE CONTROL

The interactive texture control mechanism achieves the garment swapping in a selective manner. Since M_{tex} and U_{tex} are extracted separately, the desired clothes (e.g., tops) in M can be regarded as M_{tex} and the rest of the clothes (e.g., pants) in U can be regarded as U_{tex}.

As illustrated in Fig. 7, we define three types of swapping ways: Top (for the upper clothes in M), Bottom (for the lower clothes in M), and All (for all clothes in M). To achieve these,

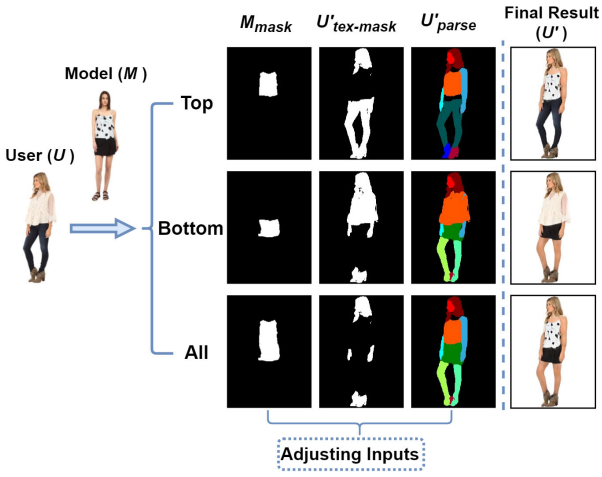


FIGURE 7. Three types of swapping ways for the interactive texture control mechanism. The intermediate results M_{mask}, $U'_{tex-mask}$, U'_{parse} are automatically adjusted according to the user's selection.

we alternate the desired cloth mask M_{mask}, the user texture mask $U'_{tex-mask}$, and the synthesized human parsing U'_{parse} based on the classes of human parsing. If the user only hopes to try on the top of the model (refer to the first row of Fig. 7, M_{mask} can be obtained according to the tops class in M_{parse}. $U'_{tex-mask}$ can be obtained by only removing the top class for both U_{parse} and U'_{parse} (i.e., set all as 1 except the tops and the background); The categories of pants, legs, and shoes in U_{parse} also need to be copied to U'_{parse} as the input of TIN.

IV. EXPERIMENTS

A. DATASETS

We perform our experiments on LookBook dataset [43] and a subset of Multi-View Clothing (MVC) dataset [44]. All these datasets contain groups of images of the same person wearing the same clothing in different poses, and each group includes at least two images. All images are with a size of 256×256. For LookBook dataset, we select 6334 groups for training and 300 groups for testing with totally 58064 human images. For MVC dataset, we select 69249 human images labeled as dresses, tops or sweaters categories and then split them into 19000 groups for training and 598 groups for testing. We also use Zalando dataset [9] to compare our results with other methods.

For paired samples, we randomly select two images from the same group; for unpaired samples, we randomly select two images from two different groups respectively; for inpainting samples, we randomly select one image and then draw some irregular holes via a mask generation algorithm [12].

B. IMPLEMENTATION DETAILS

we used CE2P [45] as our human parser and DensePose-RCNN [17] as our dense pose estimator. Adam optimizer [46] with $\beta_1 = 0.5$, $\beta_2 = 0.99$ and a learning rate of 0.0002 were

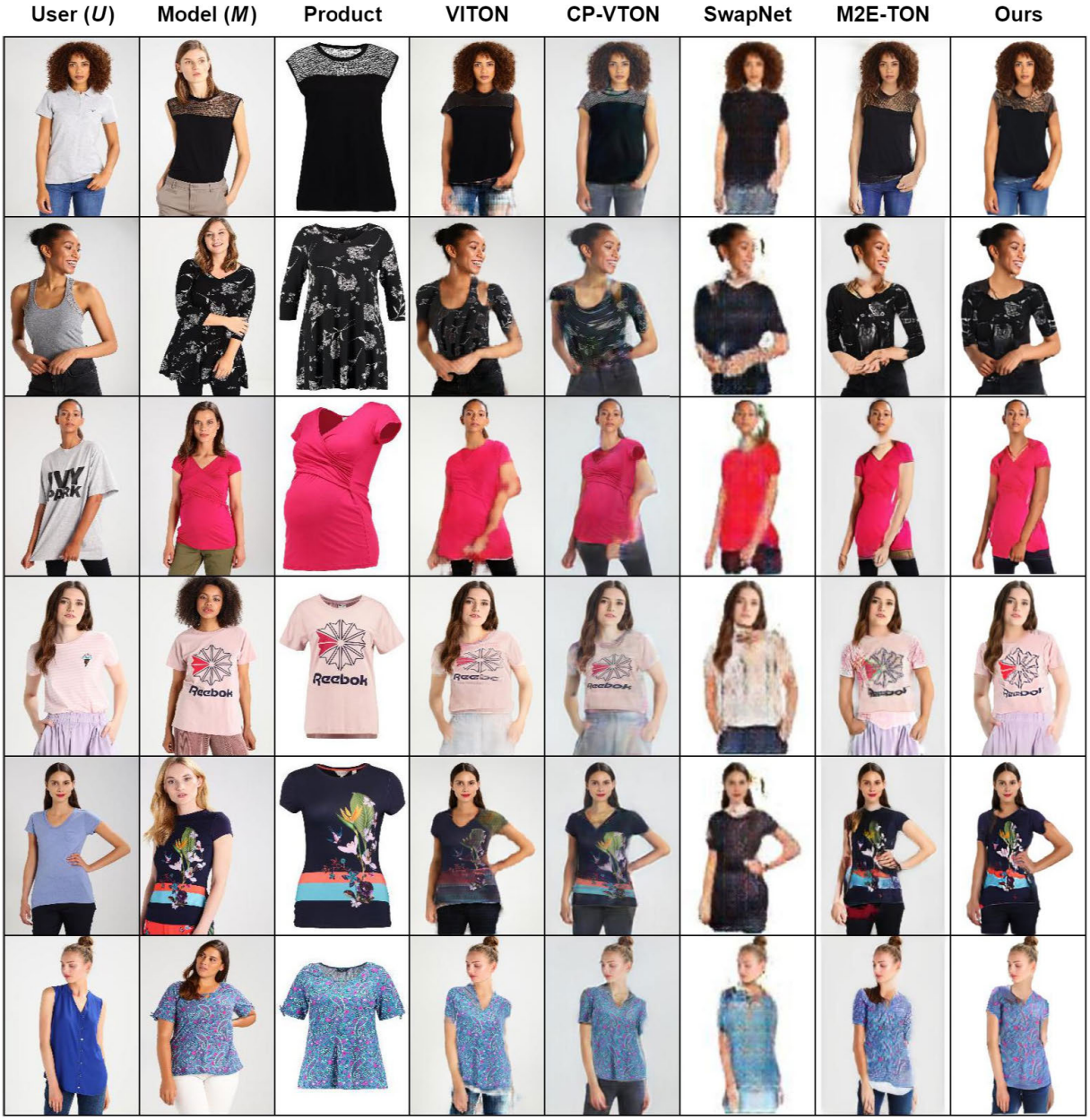


FIGURE 8. Visual comparisons with baselines for try-on results. VITON and CP-VTON cannot deal with the desired clothes with lace (the first row) and the occlusion of the user's arm (the third row). SwapNet fails to preserve the details of the garments (the fourth row). M2E-TON fails to maintain the skin color of the user's arms (the second row) and incorrectly preserve the model's hair (the last row).

used for both PIN and TIN. 3×3 kernels were adopted for all convolution layers, transpose convolution was used for all upsampling layers. instance normalization [4] was chosen for all normalization layers. The weights of loss for TIN (8) were, $\lambda_p = 0.04$, $\lambda_s = 0.03$, $\lambda_1 = 7$, $\lambda_G = 0.05$ and $\lambda_{skin} = 15$.

Since the images in LookBook dataset are street photos, we removed the background when training TIN. Otherwise, there would be no ground truth for paired training. We also randomly clipped the limbs for the coarse result U'_{coarse} to ensure that TIN can generate more realistic limbs. More specifically, for the paired samples, some texture of arms and legs in U'_{coarse} were removed before feeding it into TIN, which is similar to the inpainting samples (see Section III-C).

C. BASELINES

Since there is no previous work to tackle the task of selective garment transfer, we choose four image-based virtual try-on methods that are most related to our work. These methods are VITON [9], CP-VTON [10], SwapNet [5], and M2E-TON [6].

VITON and CP-VTON are the state-of-the-art methods by using a user image and an in-shop clothes image as the

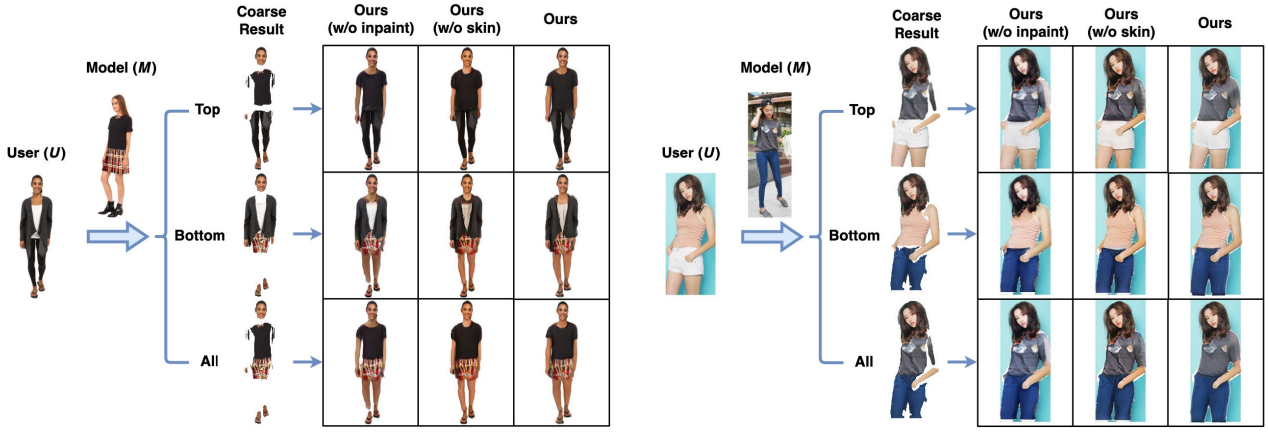


FIGURE 9. An ablation study for different swapping ways. Ours(w/o skin) fails to keep the skin of the user's neck and legs (the second row in the left case). Ours(w/o inpaint) cannot fill the missing sleeve naturally (the last row in the right case).

input. They both warp the clothes image to align with the user image but differ in the way of obtaining the parameters of the thin-plate spline (TPS) [47] transformation. The former estimates the parameters by shape context matching [48], whereas the latter by training a convolutional network [49].

SwarpNet and M2E-TON are the state-of-the-art methods by taking a user image and a model image as the input. SwapNet first synthesizes the human parsing of the user in desired clothes and then generates detailed texture according to the synthesized parsing. M2E-TON first transfers the model image to the user's pose and then fits the desired clothes to the user via a Fitting Network.

Since VITON and CP-VTON require in-shop clothes images, we evaluate all methods on the testing set of Zalando dataset [9]. Note that SwapNet, M2E-TON and our method have not fine-tuned on this dataset.

D. QUALITATIVE RESULTS

We present visual comparisons with the baselines in Fig. 8. VITON and CP-VTON always fail to handle the occlusion of the arms, the desired clothes with lace and the inner side of the clothes. Additionally, the generated garments sometimes are distorted due to the effect of the user's original clothes. SwapNet cannot preserve the details of the garments since it uses the downsampling textures as input. M2E-TON fails to maintain the skin colors of the user since it only utilizes the arm texture from the model image. And M2E-TON incorrectly preserves the model's hair in try-on results. Compared with the aforementioned baselines, our framework well preserves the garment details (e.g., textures and styles) and the user identity (e.g., skin color and hair). Our framework only focuses on the desired clothes when remapping the garment texture and make full use of the user texture according to human parsing, thus more reasonable and realistic try-on results are generated.

E. QUANTITATIVE RESULT

We use Inception Score (IS) [50] and Fréchet Inception Distance (FID) [51], which are common metrics for related works, to evaluate the quality of try-on results. IS correlates closely with the human judgment of image quality and FID can capture the similarity between the generated images and real data. As IS seems not a suitable metric for the human generation task [9], [13], we also adopt Detection Score (DS) [13] to verify the performance of these methods. DS is computed by averaging the person-class detection scores of SSD [52]. It can be regarded as a measure of how much person-like for the generated images.

TABLE 1. Quantitative comparisons of different methods. Inception Score (IS), Fréchet Inception Distance (FID) and Detection Score (DS) are automatic measures to evaluate the quality of images. The user study shows the preferences of the users for different methods.

Method	IS	FID	DS	User Study
VITON	2.718 ± 0.079	41.80	99.65	49.76%
CP-VTON	2.577 ± 0.091	**23.60**	99.82	52.13%
SwapNet	2.631 ± 0.071	114.50	98.74	14.22%
M2E-TON	2.510 ± 0.110	33.28	99.86	58.85%
Ours (w/o inpaint)	$\mathbf{2.729 \pm 0.088}$	40.91	99.85	-
Ours (w/o skin)	2.437 ± 0.074	34.47	99.86	-
Ours	2.708 ± 0.078	29.48	**99.88**	**76.29%**
Real Data	3.086 ± 0.083	0	99.90	-

However, the above automatic measures cannot judge whether the desired garment has been successfully transferred to the user image. Therefore, we perform a user study to compare our method with the baselines. Following the related works [9], [10], we conduct pairwise A/B tests on Amazon Mechanical Turk (MTurk). For each job, workers are given a user image, a model image, and two try-on results synthesized by different methods. Then the workers are asked to pick the better result which should preserve more garment details and user identity. We randomly collected 500 comparisons from the try-on results (all in 256×192) and the workers can fulfill these tasks with unlimited time.

The quantitative comparisons are summarized in Table 1. Higher scores are better for all evaluation metrics except FID. According to the table, SwapNet has higher IS even though its visual results are worse than others. Compared with IS, DS is more coincident with the qualitative results. And CP-VTON

FIGURE 10. **(a) The performance of our method for perspective change, large pose change and occlusions by fashion accessories. (b) Failure cases of our method.**

achieves the lowest FID. One of the reasons may be that it warps and copies the clothes images to make the distribution of its results greatly close to real data even though the results are unrealistic. Our method achieves the second-lowest FID, the highest DS and human evaluation score which indicate our framework outperforms the baselines.

F. ABLATION STUDY

We analyze the effectiveness of the triplet training strategy (Section III-C) and the skin loss (7) via two experiments.

- Ours (w/o inpaint): we remove the inpainting samples when training TIN.
- Ours (w/o skin): we remove skin loss when training TIN.

As illustrated in Fig. 9, without inpainting samples, TIN sometimes fails to recover the missing regions naturally and keep the color of the desired clothes. One of the reasons is that the inpainting sample has more reliable ground truth than the paired sample. TIN also fails to maintain the skin color of the user without skin loss which enforces a more accurate constraint for the skin. Our full-pipeline can transfer clothes in different swapping ways with the best performance. As Table 1 shows, without inpainting samples or skin loss, FID would rise by more than 16% and DS would decrease slightly, which also indicate our full-pipeline can generate more realistic human images.

G. LIMITATIONS AND FUTURE WORK

The architecture of I-VTON is dual-path which enables it to generate try-on results by interchanging the input images between users and models, As shown in Fig. 10 (a). U to M means users try on the clothe from models and M to U means models try on the clothes from users.

I-VTON can handle limited perspective changes, as shown in the first row of Fig. 10 (a). For the transfer between the frontal-view and back-view, it cannot generate correct results since the provided information is insufficient. We utilize the texture from both the front and back of the desired clothes in our implementation, which only works when both sides of the clothes are similar, as shown in the second row of Fig. 10 (a).

As shown in the third row of Fig 10 (a), the small occlusions caused by fashion accessories such as hats, sunglasses, and bags have a slight effect on our method. And our method can still generate reasonable results for large pose change, as illustrated in the last row of Fig. 10 (a). However, the final result may appear some white edges in the background when the region of the generated results is smaller than the user wearing the original clothes. One possible solution is to train another network to fill the missing region in the background.

Our framework is sensitive to the accuracy of the human parser and the dense pose estimator. Fig. 10 (b) shows some failure cases of our method. It always fails to generate the

FIGURE 11. More qualitative results for different swapping ways. For each side, the first two columns are the model images and user images, the next three columns are the final results for three swapping ways (see Section III-D). We test our method on three public datasets. The first six rows in the left side are testing on Zalando dataset; the last four rows in the right side are testing on LookBook dataset; the rest are testing on MVC dataset.

clothes with line patterns due to the prediction error of the dense pose. As shown in the second row, the shirt's logo is distorted in both the coarse and final results. It is noteworthy that TIN can recover the distorted texture from the coarse result to some extent, as shown in the first row.

It also typically fails if the clothes region in the human parsing is wrongly classified as some body parts (e.g., head, arms). The third row of Fig. 10 (b) shows that the user's sleeve is wrongly preserved since it was recognized as the arm class in the user human parsing. The last row shows that the accuracy of the human parsing for models is also crucial.

According to the limitation of our method, we aim to find a more robust texture extraction approach to reduce the failure cases in our future work. We would also extend our method to swap more fashion items such as shoes and scarfs.

V. CONCLUSION

We propose a novel framework named Inpainting-based Virtual Try-On Network (I-VTON), which enables a user to try on different pieces of clothing in a selective manner. With the aid of an interactive texture control mechanism, our method flexibly swaps the selected clothes without training extra network. We show our framework can generate reasonable try-on results via extracting texture properly. We also validate the effectiveness of the inpainting samples and the skin loss. Quantitative and qualitative comparisons show that our I-VTON outperforms the baselines for both the garment details and the user identity.

APPENDIX
MORE TRY-ON RESULTS

Fig. 11 shows more try-on results of our method for three swapping ways on Zalando dataset [9], Multi-View Clothing (MVC) dataset [44] and LookBook dataset [43].

REFERENCES

[1] 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.*, Jul. 2017, pp. 1125–1134.

[2] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays, "TextureGAN: Controlling deep image synthesis with texture patches," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8456–8465.

[3] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," 2018, *arXiv:1805.08318*. [Online]. Available: https://arxiv.org/abs/1805.08318

[4] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6924–6932.

[5] A. Raj, P. Sangkloy, H. Chang, J. Hays, D. Ceylan, and J. Lu, "SwapNet: Image based garment transfer," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2018, pp. 679–695.

[6] Z. Wu, G. Lin, Q. Tao, and J. Cai, "M2E-try on net: Fashion from model to everyone," 2018, *arXiv:1811.08599*. [Online]. Available: https://arxiv.org/abs/1811.08599

[7] Y.-J. Rhee, "Online impulse buying behavior with apparel products: Relationships with apparel involvement, website attributes, and product category/price," Ph.D. dissertation, Virginia Tech, Blacksburg, VA, USA, 2006.

[8] V. Bhardwaj and A. Fairhurst, "Fast fashion: Response to changes in the fashion industry," *Int. Rev. Retail, Distrib. Consum. Res.*, vol. 20, no. 1, pp. 165–173, Feb. 2010.

[9] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis, "VITON: An image-based virtual try-on network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7543–7552.

[10] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang, "Toward characteristic-preserving image-based virtual try-on network," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 589–604.

[11] G. Liu, F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro, "Image inpainting for irregular holes using partial convolutions," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 85–100.

[12] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Free-form image inpainting with gated convolution," 2018, *arXiv:1806.03589*. [Online]. Available: https://arxiv.org/abs/1806.03589

[13] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe, "Deformable gans for pose-based human image generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3408–3416.

[14] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 7291–7299.

[15] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 932–940.

[16] X. Liang, K. Gong, X. Shen, and L. Lin, "Look into person: Joint body parsing & pose estimation network and a new benchmark," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 871–885, Mar. 2018.

[17] R. A. Güler, N. Neverova, and I. Kokkinos, "Densepose: Dense human pose estimation in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7297–7306.

[18] M. Loper, N. Mahmood, J. Romero, G. Pons-Moll, and M. J. Black, "SMPL: A skinned multi-person linear model," *ACM Trans. Graph.*, vol. 34, no. 6, pp. 248:1–248:16, Oct. 2015.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[20] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2223–2232.

[21] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," 2018, *arXiv:1809.11096*. [Online]. Available: https://arxiv.org/abs/1809.11096

[22] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. van Gool, "Pose guided person image generation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 406–416.

[23] L. Ma, Q. Sun, S. Georgoulis, L. van Gool, B. Schiele, and M. Fritz, "Disentangled person image generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 99–108.

[24] N. Neverova, R. A. Guler, and I. Kokkinos, "Dense pose transfer," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 123–138.

[25] M. Bertalmio, G. Sapiro, V. Caselles, and C. Ballester, "Image inpainting," in *Proc. 27th Annu. Conf. Comput. Graph. Interact. Techn.*, 2000, pp. 417–424.

[26] C. Barnes, E. Shechtman, A. Finkelstein, and D. Goldman, "PatchMatch: A randomized correspondence algorithm for structural image editing," *ACM Trans. Graph.*, vol. 28, no. 3, p. 24, 2009.

[27] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang, "Generative image inpainting with contextual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5505–5514.

[28] K. Nazeri, E. Ng, T. Joseph, F. Qureshi, and M. Ebrahimi, "Edgeconnect: Generative image inpainting with adversarial edge learning," 2019, *arXiv:1901.00212*. [Online]. Available: https://arxiv.org/abs/1901.00212

[29] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black, "DRAPE: Dressing any person," *ACM Trans. Graph.*, vol. 31, no. 4, pp. 1–35, 2012.

[30] M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama, "Virtual fitting by single-shot body shape estimation," in *Proc. Int. Conf. 3D Body Scanning Technol.*, 2014, pp. 406–413.

[31] 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, p. 73, 2017.

[32] N. Jetchev and U. Bergmann, "The conditional analogy GAN: Swapping fashion articles on people images," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 2287–2292.

[33] M. Mirza and S. Osindero, "Conditional generative adversarial nets," 2014, *arXiv:1411.1784*. [Online]. Available: https://arxiv.org/abs/1411.1784

[34] S. Zhu, R. Urtasun, S. Fidler, D. Lin, and C. C. Loy, "Be your own prada: Fashion synthesis with structural coherence," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2017, pp. 1680–1688.

[35] C. B. Barber, D. P. Dobkin, and H. Huhdanpaa, "The Quickhull algorithm for convex hulls," *ACM Trans. Math. Softw.*, vol. 22, no. 4, pp. 469–483, 1996.

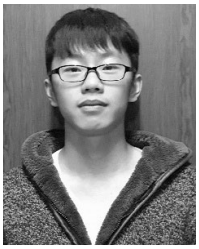
[36] M. Schindler and E. Chen, "Barycentric coordinates in olympiad geometry," *Olympiad Articles*, pp. 1–40, Jul. 2012.

[37] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

[38] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," 2018, *arXiv:1802.05957*. [Online]. Available: https://arxiv.org/abs/1802.05957

[39] J. H. Lim and J. C. Ye, "Geometric GAN," 2017, *arXiv:1705.02894*. [Online]. Available: https://arxiv.org/abs/1705.02894

[40] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 694–711.
[41] L. Gatys, A. S. Ecker, and M. Bethge, "Texture synthesis using convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 262–270.
[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*. [Online]. Available: https://arxiv.org/abs/1409.1556
[43] D. Yoo, N. Kim, S. Park, A. S. Paek, and I. S. Kweon, "Pixel-level domain transfer," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 517–532.
[44] K.-H. Liu, T.-Y. Chen, and C.-S. Chen, "MVC: A dataset for view-invariant clothing retrieval and attribute prediction," in *Proc. ACM Int. Conf. Multimedia Retr.*, 2016, pp. 313–316.
[45] T. Ruan, T. Liu, Z. Huang, Y. Wei, S. Wei, and Y. Zhao, "Devil in the details: Towards accurate single and multiple human parsing," 2018, *arXiv:1809.05996*. [Online]. Available: https://arxiv.org/abs/1809.05996
[46] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: https://arxiv.org/abs/1412.6980
[47] F. L. Bookstein, "Principal warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, Jun. 1989.
[48] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, Apr. 2002.
[49] I. Rocco, R. Arandjelovic, and J. Sivic, "Convolutional neural network architecture for geometric matching," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 6148–6157.
[50] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen, "Improved techniques for training GANs," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2234–2242.
[51] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6626–6637.
[52] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 21–37.



LI YU was born in Zhejiang, China, in 1995. He received the B.S. degree in textile engineering from Zhejiang Sci-Tech University, Zhejiang, in 2017. He is currently pursuing the M.S. degree in digital textile engineering with Donghua University, Shanghai, China. His research interests include deep learning and computer vision, especially for the fashion industry.

YUEQI ZHONG received the Ph.D. degree from Donghua University, in 2001. He is regarded as a specialist in the area of virtual clothing and virtual human body. He joined the faculty of the College of Textiles, Donghua University, in October 2005, on completion of his Postdoctoral Research at The University of Texas at Austin. He is currently a Professor with the College of Textiles, Donghua University. His research interests include virtual clothing, online sizing, fit evaluation, and virtual human body toward E-commerce. He was granted the National Natural Science Foundation of China (NSFC) funding many times to support his research work on digitalizing the physical world in the cyberspace. He was also the PI of many projects granted at the Level of Province and Department. In 2014, he was a recipient of the Nationwide Prize for his contribution to the textile and apparel industry. His patents on solving the problem of "virtual reality towards online dressing" won him the prize of Shanghai Science and Technology Award, in 2013.

XIN WANG received the B.S. degree from the Lanzhou University of Technology, in 2015. He is currently pursuing the Ph.D. degree with the College of Textiles, Donghua University, under the supervision of Prof. Y. Zhong. His research interests include the recognition and understanding of fashion images, recommendation system of fashion products, and developing methods to combine both visual and textual information.