Arbitrary Virtual Try-On Network: Characteristics Preservation and Trade-off between Body and Clothing

Yu Liu, Mingbo Zhao, Senior Member, IEEE, Haijun Zhang, Senior Member, IEEE, Zhao Zhang, Senior Member, IEEE, Shuicheng Yan, Fellow, IEEE

Abstract—Deep learning based virtual try-on system has achieved some encouraging progress recently, but there still remain several big challenges that need to be solved, such as trying on arbitrary clothes of all types, trying on the clothes from one category to another and generating image-realistic results with few artifacts. To handle this issue, we in this paper first collect a new dataset with all types of clothes, *i.e.* tops, bottoms, and whole clothes, each one has multiple categories with rich information of clothing characteristics such as patterns, logos, and other details. Based on this dataset, we then propose the Arbitrary Virtual Try-On Network (AVTON) that is utilized for all-type clothes, which can synthesize realistic try-on images by preserving and trading off characteristics of the target clothes and the reference person. Our approach includes three modules: 1) Limbs Prediction Module, which is utilized for predicting the human body parts by preserving the characteristics of the reference person. This is especially good for handling crosscategory try-on task (e.g. long sleeves \leftrightarrow short sleeves or long pants \leftrightarrow skirts, etc.), where the exposed arms or legs with the skin colors and details can be reasonably predicted; 2) Improved Geometric Matching Module, which is designed to warp clothes according to the geometry of the target person. We improve the TPS based warping method with a compactly supported radial function (Wendland's Ψ -function); 3) Trade-Off Fusion Module, which is to trade off the characteristics of the warped clothes and the reference person. This module is to make the generated try-on images look more natural and realistic based on a finetune symmetry of the network structure. Extensive simulations are conducted and our approach can achieve better performance compared with the state-of-the-art virtual try-on methods.

Index Terms—Deep Learning, Virtual Try-on, Generative Adversarial Networks, Artificial Intelligence in Fashion

Manuscript received Feb. 19, 2021; revised August 26, 2021. This work is partially supported by National Key Research and Development Program of China (2019YFC1521300), supported by National Natural Science Foundation of China (61971121, 61672365) and also supported by the Fundamental Research Funds for the Central Universities of China (JZ2019HGPA0102).

Y. Liu and M. Zhao are with the School of Information Science and Technology, Donghua University, Shanghai, China. (e-mail: 2191408@mail.dhu.edu.cn, mzhao4@dhu.edu.cn).

Haijun Zhang is with the Department of Computer Science, Harbin Institute of Technology, Shenzhen, Shenzhen 518055, China (e-mail: hjzhang@hit.edu.cn).

Zhao Zhang is with the Department of Computer Science, Hefei University of Technology, Hefei 230009, China (e-mail: cszzhang@gmail.com).

Mingliang Xu is with the School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China (e-mail: iexumingliang@zzu.edu.cn).

Shuicheng Yan is with Qihoo 360, Beijing, China, and also with the National University of Singapore, Singapore 119077 (e-mail: eleyans@nus.edu.sg).

I. INTRODUCTION

In modern society, clothing plays an important role in human daily life, as proper outfits can enhance peoples' beauty and personal quality [7]. But in practice, people tend to make the final purchase decision before trying on the clothes. On the other hand, huge amounts of clothing data emerge on the Internet or social media during the past few years due to the development of the clothing e-commerce platform. Therefore, it is impossible and unfeasible for people to try on all the clothes especially on the website. Therefore, to meet the shopping needs of consumers [18], it is necessary for researchers to conducted considerable researches on imagebased virtual try-on technology [19], [56], which can easily help people to achieve image-realistic try-on effects on website or smartphone app by transforming the target clothing outfit onto a certain reference person.

During the past few years, computer vision technology has been widely utilized in the extensive applications of artificial fashion. These applications include clothes detection [34], [10], clothes parsing [29], [31], [28], [11], clothes attributions and categories recognition [47], [34], [10], clothes collocation [24], [54], [46], [20], [45], [30], [16], [3], clothes image retrieval [53], [33], [35], [14], [32], [21], [25] and fashion edit [22], [5]. These applications are all merited from recently developed technology, namely deep learning due to its powerful feature extraction ability to capture the rich mid-level image representations. Motivated by this technology, the deep learning based virtual try-on methods have been extensively studied and achieved considerable results recently. These methods have adopted the concept of GAN [12] and focus on generating the synthetic and realistic try-on images by preserving the characteristics of a clothing image (e.g., texture, logo, and embroidery) as well as warping it to an arbitrary human pose. Typical methods for virtual try-on include VITON [17], CP-VTON [47], VTNFP [52], ACGPN [51], et al. [39], [38], [42], [41], [37], [13], [56], [15]. Extensive simulation results show the effectiveness of the proposed work.

Although these researches have made some progress, it remains some ongoing challenges and limitations: 1) most benchmark datasets utilized for training virtual try-on methods mainly contain top clothes. Therefore, current related works can only focus on top clothing try-on task due to the lack of diversity in the dataset. For example, the VITON dataset [17] is a benchmark dataset used in most prior arts. It only

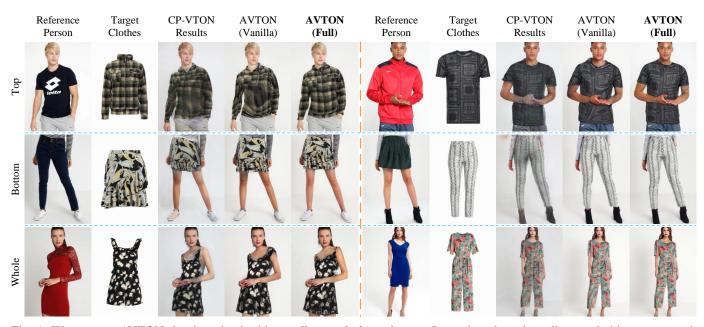


Fig. 1: We propose AVTON that is trained with an *all-type clothing dataset*. It can be adapted to all-type clothing try-on task and get image-realistic results. The types of clothes are divided into the *top, bottom, and whole*, and we apply CP-VTON [47] and AVTON on them. The CP-VTON is retained with the all-type clothing dataset, and the AVTON (Vanilla) indicates the AVTON trained without *LPM and Wendland's* Ψ -function [50], while the AVTON (Full) is just the opposite.

contains frontal-view woman and top clothing image pairs. As a result, the trained model can only handle top clothing try-on task while cannot be adapted to work on the other bottom or whole clothing try-on task. Therefore, how to collect a diverse try-on dataset with tops, bottoms, and whole clothes and develop a method for all-type clothing try-on task is necessary and important; 2) cross-category try-on task(*e.g.* long sleeves \leftrightarrow short sleeves or long pants \leftrightarrow skirts, etc.) is another challenge in the virtual try-on system. A case in point is that when people aim to try on from long sleeves to short sleeves, some parts of people's arms will be exposed. Therefore, it is necessary to preserve the characteristics of the reference person and predict such an exposed human body when generating the image-realistic try-on results. However, most current methods [17], [47], [39], [52] mainly focus on try-on with one category by preserving the characteristics of clothes but do not consider the limb prediction when handling try-on with different categories. As a result, some methods may cause problems, e.g., the limb is covered by clothes, the color of the skin is wrongly painted and the detailed hands cannot be generated; 3) current methods are not good at trading off characteristics of the warped clothes and the reference person, such as ACGPN. Although it can preserve the characteristics of the warped clothes and reference person as much as possible, its generated images are not realistic enough, such as some artifacts near the neck regions. The reason for these results is that the final fusion module prefers to preserve the warped clothing characteristics and cannot correct these errors.

In this paper, we propose a new image-based virtual tryon method, called AVTON, to address the issues mentioned above: 1) In order to handle the all-type clothing try-on task

and increase the diversity of the dataset, we collect a new dataset from the Zalando. As shown in Fig. 1, the example in the first row represents the top clothing try-on task. This type of data is similar to the VITON dataset, but it involves more rich characteristics, such as human genders, shapes, and poses, and especially more categories in top clothing images. The second row provides the images for the bottom clothing tryon task, where the bottom clothing images include trousers, shorts, and skirts, etc., with rich textures, logos, and other details. The last row provides examples for the whole clothing try-on task, where this type of human image generally contains the entire human body. Based on such a diverse dataset, we then develop the proposed method, which can naturally handle try-on task for all-type clothes; 2) We propose a new imagebased virtual try-on method, called AVTON, to achieve arbitrary clothing try-on and image-realistic results. The proposed method contains three modules: a) Limbs Prediction Module, which is developed for predicting limbs, and keep the head and the non-target human body parts, to preserve the characteristics of the reference person. This module is especially suitable for handling cross-category try-on task, such as long sleeves \leftrightarrow short sleeves or long pants \leftrightarrow skirts, *etc.*, where the exposed arms or legs (including their skin colors and details) can be reasonably predicted. This is good to help the try-on system for formulating a realistic result in the following modules; b) Improved Geometric Matching Module, which is designed to warp clothes according to the geometry of the reference person. By carefully analyzing the basic concept of Thin-Plate Spline (TPS) [6] based methods, we argue that the selection of radial basis function is a key point to affect the performance of image warping. Motivated by this end, we then have adopted Wendland's Ψ -function [50] as the compactly

supported radial basis function. Both theoretical analysis and simulation have verified that the proposed method is able to characterize the local geometrical structure of images, which is good for warping the clothes image especially with the complex texture; c) Trade-Off Fusion Module, which is to trade off the characteristics of the warped clothes and the reference person, this module is to make the generated try-on images looks more natural and realistic based on a fine-tune symmetry of the network structure (a pair of UNet [43]). Experiments show that AVTON significantly outperforms the state-of-the-art methods for virtual try-on [17], [47], [52], [51], and can generate realistic try-on images in all-type clothing try-on task (Fig. 1).

The main contributions of this paper are summarized as follows:

- We collected a diverse dataset for handling the all-type clothing try-on task, which contains all types of clothes. Benefiting from this dataset, we are able to design an arbitrary virtual try-on method;
- In order to handle the cross-category try-on task, we propose a new limbs prediction module to preliminarily predict limbs and preserve the characteristics of the reference person. We show that it is necessary for the cross-category try-on task;
- 3) We for the first time use the Ψ -function of Wendland to warp clothes, which is compact support for the registration of images. It greatly improves the try-on quality in preserving the characteristics of the target clothes;
- We present a novel fusion network that trades off the characteristics between the warped clothes and the reference person. By adjusting characteristics and correcting errors, the try-on images learned are more realistic;
- 5) We demonstrate that our proposed method can be applied to the all-type clothing try-on task and crosscategory try-on task, and outperforms the state-of-the-art methods both qualitatively and quantitatively.

II. RELATED WORK

Image Synthesis. Generative Adversarial Network (GAN) [12] is a popular and effective tool in image synthesis [23], [26], [27], [40] and generally consists of a generator and a discriminator. The core idea of GAN is a zero-sum game, in which the generator aims to generate a realistic image and the discriminator is to distinguish it from the real one. The procedure is continuous until the discriminator is unable to judge whether the output result of the generator is true or not. By taking advantage of the GAN, researchers make great progress in image-to-image translation [23], [55], [37], photo inpainting [15], clothing translation [53], [41], *etc.* Extensive learning tasks verify the effectiveness of image synthesis.

Fashion Analysis. Benefit from the development and popularization of deep learning, fashion-related work has made great progress recently. Current works of artificial fashion mainly focus on clothing parsing [29], [31], [28], [11], clothing landmark detection [34], [10], clothes attributions and categories recognition [47], [34], [10], fashion retrieval [53], [33],

Virtual Try-On. Virtual try-on belongs to fashion image synthesis, which can be roughly divided into two categories: 3D model-based methods [1], [41], [36], [13] and 2D imagebased methods [17], [47], [52], [51]. At present, both methods are based on deep learning and make great progress. However, the dataset on the 3D model-based methods is difficult to obtain, and these methods need huge computation due to the complexity of the model; On the other hand, 2D image-based methods does not require auxiliary methods to improve the performance of the network so that it is more easily to be trained. Among all deep learning based 2D methods, VITON [17] and CP-VTON [47] are the two first works. The key idea of VITON is to exploit a Thin-Plate Spline (TPS) [6] based method to warp the clothes images with texture mapped on it, while CP-VTON has extended VITON by developing neural network layers to learn the transformation parameters of TPS so that it is trainable and can achieve more correct alignment performances. Both two methods cannot achieve satisfactory results when postural changes and complexity texture occur. VTNFP [52] and ACGPN [51] are another two popular methods that have improved the performance of the try-on task with more characteristics of clothes and human body preserved, but they still have problems that the generated images are not realistic enough, such as some artifacts near the neck regions, as the final fusion module in these methods prefers to preserve the warped clothing characteristics and cannot correct these errors. In addition, all these methods are trained via VITON dataset [17] that only contains top clothes. As a result, these methods cannot handle arbitrary try-on tasks with all types of clothes. Another challenge is the crosscategory try-on task. Though the work in [37] has proposed a variant of CycleGAN [55], namely InstaGAN, to handle this task (long sleeves \leftrightarrow short sleeves, long pants \leftrightarrow skirts, *etc.*), the final try-on results cannot be controlled due to the strategy of CycleGAN. This paper aims to address the above problems by proposing a new method for all-type clothing and crosscategory try-on tasks with more image-realistic performance.

III. ARBITRARY VIRTUAL TRY-ON NETWORK

Our goal is to learn an arbitrary virtual try-on model that can be adapt to all-type clothing and cross-category try-on tasks and generate more realistic try-on images than prior arts. The proposed AVTON contains three modules, as shown in Fig. 2. First, the limbs Prediction Module (LPM) is utilized to predict the limbs, head, and non-target human body. This is especially useful for handling cross-category try-on task. Second, the Improved Geometric Matching Module (IGMM) uses Wendland's Ψ -function [50] to improve the TPS based method [6] for warping the clothes. Third, the Trade-Off Fusion Module (TOFM) takes the warped clothes and the human body information as input, and then generates composition mask and rendered person by a pair of UNet [43] and a fusion part. Especially, the TOFM takes advantage of GAN [12].

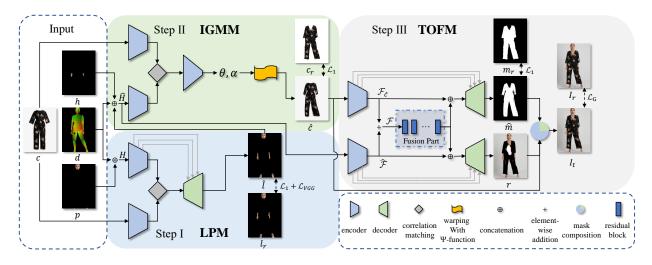


Fig. 2: An overview of our AVTON. **Step I**: Limbs Prediction Module takes the target clothes c and the human body information H as the input to predict the exposed limbs and preserve the primary human body information, and output the predicted human body \hat{l} ; **Step II**: Improved Geometric Matching Module takes the target clothes c and the refined human body information \hat{H} as input, and output the warped clothes \hat{c} ; **Step III**: Trade-Off Fusion Module firstly takes the warped clothes \hat{c} and the refined human body information \hat{H} as the input to predict the composition mask \hat{m} and the rendered person r, and then compose the outputs with the warped clothes \hat{c} to generate the try-on image I_t .

A. Limbs Prediction Module (LPM)

The purpose of designing the Limbs Prediction Module (LPM) is to predict the exposed limbs and reserve the primary human body information (*i.e.*, head, non-target human body parts). Most earlier methods generate exposed limbs in try-on steps but neglected primary human body information, which may lead to generate unreasonable color of skin and encountered occlusion issues. We thereby propose LPM to address these issues.

Given a reference person image I_r , LPM takes the target clothes c and the human body information (*i.e.*, the pose and shape information d which is extracted from the reference person I_r by DensePose [2], the head and non-target human body parts information p in the reference person I_r) as inputs to predict exposed limbs \hat{l} . In detail, the target clothes c and the human body information $H = d \oplus p$ (\oplus : concatenation) are firstly encoded as the input features. They are then formed as a single tensor by a correlation layer and input to the decoder. Finally, the exposed limbs are predicted by the decoder. The encoder and correlation layer are similar to CP-VTON's GMM step [47], while the human body information d's encoderdecoder layers are similarly to UNet [43] structure shown in Fig. 2. All this leads to preserve the primary body information.

The LPM is trained under a combination of the pixel-wise L1 loss and VGG perceptual loss between predicted result \hat{l} and ground truth l_r , where l_r includes head, non-target human body parts and exposed limbs in the reference person I_r :

$$\mathcal{L}_{\text{LPM}} = \lambda_{\text{L1}} \|\hat{l} - l_r\|_1 + \lambda_{\text{vgg}} \mathcal{L}_{\text{VGG}}(\hat{l}, l_r)$$
(1)

where

$$\mathcal{L}_{\text{VGG}}(\hat{l}, l_r) = \sum_{i=1}^{5} \lambda_i \|\phi_i(\hat{l}) - \phi_i(l_r)\|_1$$
(2)

is the VGG perceptual loss, where λ_{L1} and λ_{vgg} are the tradeoff parameters for two loss terms in Eq. 1, which all set to 1 in our experiments, and $\phi_i(l)$ denotes the feature map of limbs' image l of the *i*-th layer in the visual perception network ϕ , which is a VGG19 pre-trained on ImageNet.

B. Improved Geometric Matching Module (IGMM)

The old way of warping clothes is based on Thin-Plate Splines (TPS) [6] with $r^2 \log r$ as Radial Basis Functions (RBFs). This method yields minimal bending energy properties measured over the whole image. But since it is not a compactly supported RBFs, the deformation will cover the regions where all control points are located. It is advantageous for yielding an overall smooth deformation and preserving geometrical characteristics, but it is problematic when only a small part of the image is desired to be deformed. This will lead to unreasonable deformation when warps clothes. To address this issue, we in this work has adopted Ψ -function of Wendland [50] as RBFs. As mentioned in [8], [9], it is a more compactly support for the registration of images so that the bending region can be narrowed down when minimizing the bending energy. Here, we first give its formulation as follows:

$$\psi_{d,k}(r) \coloneqq I^k (1-r)_+^{\lfloor d/2 \rfloor + k+1}(r) \tag{3}$$

where

$$(1-r)^v_+ = \begin{cases} (1-r)^v & 0 \le r < 1\\ 0 & r \ge 1 \end{cases}$$
$$I\psi(r) \coloneqq \int_r^\infty t\psi(t) \mathrm{d}t \quad r \ge 0.$$

The equation also holds for different spatial supports α : $\psi_{\alpha}(r) = \psi(r/\alpha)$. We apply the $\psi_{\alpha,3,1}$ -function as RBFs to replace the TPS's RBFs:

$$\psi_{\alpha,3,1}(r) = (1 - r/\alpha)^4_+ (4r/\alpha + 1) \tag{4}$$

where α is a learnable parameter as same as the spatial transformation parameters θ . Inspired by CP-VTON [47], we use the same structure to learn these parameters. As shown in Fig. 2, our IGMM firstly extract high-level features of the target clothes c and the refined human body information $\hat{H} = d \oplus \hat{l} \oplus h$ (h represents hand details in reference person I_r) respectively. Then a correlation layer to combine two features into as single tensor as input to the regression network that predicts θ and α . Finally, a transformation $\Psi_{\theta,\alpha}$ based on Eq. 4 for warping the target clothes c into the result $\hat{c} = \Psi_{\theta,\alpha}(c)$.

To learn the module above, we make some derivations and experiments to study the size of spatial supports α , which shows a significant positive correlation between spatial warping range and α , and we concluded that the best α should meet the condition: $\alpha \ge D$, where D is the maximum distance among control points in Delaunay triangles [4]. In our experiments, we set $D = \sqrt{a^2 + b^2}$ (a and b is the vertical distance and the horizontal distance among nearest control points, respectively). Consequently, the final $\alpha = \lambda_{\alpha}\hat{\alpha} + d$, where $\hat{\alpha}$ is the sigmoid's output of the regression network and λ_{α} is set to 6 in our experiments.

To train the module, we conducted it under the pixel-wise L1 loss between the warped clothes \hat{c} and ground truth c_r , where c_r is the clothes worn on the reference person in I_r :

$$\mathcal{L}_{\text{IGMM}} = \|\hat{c} - c_r\|_1 = \|\Psi_{\theta,\alpha}(c) - c_r\|_1$$
(5)

C. Trade-Off Fusion Module (TOFM)

Try-on image synthesis from the warped clothes and the reference person is a many-to-one mapping problem. It aims at not only preserving the characteristics of the warped clothes and the reference person, but also trading off them to make images realistic. One of the common methods [17] is to produce a composition mask for fusing UNet [43] rendered person with warped clothes and finally to produce a refined result. Although it can refine the course try-on image, it lacks preserving characteristics of the warped clothes. Another common method [47] is to utilize an UNet to render a person image and predict a composition mask simultaneously, and then synthesizing the try-on image by fusing the rendered person and the warped clothes via the composition mask. But this way failed to preserve characteristics of reference person on account of using a single UNet structure, this structure prefers to preserve characteristics of the warped clothes. And other methods [52], [51] have some problems with trading off characteristics between the warped clothes and the reference person (e.g., artifacts near the neck regions).

In this paper, we adopt a GAN [12] based method for generating the realistic try-on image. In detail, we formulate the generator G in GAN by a pair of UNet, where the inputs are the warped clothes \hat{c} and refined human body information \hat{H} , while the outputs are the generated composition mask \hat{m} and rendered person r. However, making the characteristics of the warped clothes and the reference person contribute equally when generating the try-on image is still problematic, as this will cause occlusion and artifact problems. To make images more realistic, we add a fusion part in our try-on module (Fig. 2 residual blocks) for handling the above problems. Specifically, in the proposed module, the warped clothes' features $\mathcal{F}_{\hat{c}}$ and the refined human body features are firstly summed element-wisely, so that the fused features can be obtained, i.e. $\mathcal{F}(\mathcal{F}_{\hat{c}} + \hat{\mathcal{F}} \to \mathcal{F})$. Then the fused features \mathcal{F} are concatenated with $\mathcal{F}_{\hat{c}}$ and $\hat{\mathcal{F}}$, and decoded respectively to get predicted mask \hat{m} and predicted rendered person r. Finally, similarly to CP-VTON [47], \hat{c} and r are fused together using \hat{m} to synthesize try-on image I_t :

$$I_t = \hat{m} \odot \hat{c} + (1 - \hat{m}) \odot r \tag{6}$$

where \odot represents element-wise matrix multiplication. Additionally, we use the multi-scale discriminators D that is similar to pix2pixHD [48].

At the training phase, the generator's loss is the combination loss scheme of CP-VTON and pix2pixHD, it includes L1 loss, VGG perceptual loss (Eq. 2) and LSGAN loss:

$$\mathcal{L}_{G} = \lambda_{L1} \| I_t - I_g \|_1 + \lambda_{vgg} \mathcal{L}_{\text{VGG}}(I_t, I_g) + \lambda_{mask} \| \hat{m} - m_r \|_1 + \lambda_{lsgan} (D(\hat{c}, \hat{H}, I_t) - 1)^2$$
(7)

while the discirminator's loss is:

$$\mathcal{L}_D = ((D(\hat{c}, \hat{H}, I_t))^2 + (D(\hat{c}, \hat{H}, I_g) - 1)^2)/2 \qquad (8)$$

where I_g is the ground truth image, and $I_g = I_r$ in training stage, m_r is the mask of c_r . In our experiments, we set λ_{L1} , λ_{vqg} and λ_{mask} to 10, while set λ_{lsqan} to 1.

IV. EXPERIMENTS

A. Datasets

Experiments are conducted on two datasets: the VITON dataset [17] that is used in VITON and CP-VTON *et al.* [17], [47], [52], [51], and the self-collected Zalando dataset. In this paper we will call them **VITON-Dateset** and **Zalando-Dataset** respectively.

VITON-Dataset. It contains 16253 frontal-view woman and top clothing image pairs, which is split into a training set and a testing set with 14221 and 2032 pairs respectively. To further research and comparison, we use the method of ACGPN [51] to score the complexity of each woman image and divide VITON-Dataset into Easy, Medium, and Hard levels.

Zalando-Dataset. It contains 34928 frontal-view human (include man and woman) and clothing (include top, bottom, and whole) image pairs, it was split into a training set and testing set with 32746 and 2182 respectively. The training set contains 19185 tops, 10587 bottoms, and 2974 whole clothes, and the testing set contains 1310 tops, 692 bottoms, and 180 whole clothes. We also score the complexity of each human image except the lower human body image.

Note that we use the DensePose [2] to extract the pose and shape information, and we use the SCHP [29] to obtain the human parses for the head, hand detail, and non-target human body parts.



Fig. 3: **On the VITON-Dataset**. Qualitative comparisons of VITON [17], CP-VTON [47], VTNFP [52], ACGPN [51] and AVTON in easy to hard levels (from top to bottom). Our method preserves more characteristics of the reference person with the LPM, and it also preserves more characteristics of the targets clothes with the IGMM. What is more, AVTON can generate more realistic try-on images with the TOFM, which is good at trading off characteristics of the warped clothes and the reference person.

B. Implementation Details

The experiments are conducted on VITON-Dataset and Zalando-Dataset respectively, and the results totally independent of each other.

Training. Followed by steps in Fig. 2, we first train the LPM and then use the LPM's trained results to train IGMM, finally train the TOFM with the trained results of LPM and IGMM. On the VITON-Dataset training setup, each module is trained for 400K steps with batch size 4, while on the Zalando-Dataset training setup, each module is trained for 800K steps with batch size 4. Both training setups use Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and learning rate is fixed at 0.0001. Additionally, the resolution for all input and output images is 256×192 , and we use one NVIDIA 2080Ti GPU in our experiments.

Testing. We use the same steps as the training stage to test modules. On the VITON-Dataset testing stage, we test our modules in easy, medium, and hard cases respectively, and evaluate the results qualitatively and quantitatively. On the Zalando-Dataset testing process, we evaluate the qualitative and quantitative results in the top, bottom, and whole clothing cases respectively, and perform ablation experiments. Note that we also analyze the qualitative results in the easy, medium, and hard cases respectively (there are complexity labels in Fig. 4 and Fig. 5).

C. Qualitative Results

On the VITON-Dataset. Fig. 3 shows visual comparison of our proposed method with VITON [17], CP-VTON [47], VTNFP [52] and ACGPN [51] in different complexity levels. To save a lot of work to reproduce them (VTNFP has no official code), we refer to the results from the paper of ACGPN. Note we validated the official code of ACGPN qualitatively and quantitatively and got the same results with ACGPN.

Among these results, VITON is the worst method in preserving characteristics, such as the blurred hands (*i.e.* the red box of the first row), the disappeared pattern (*i.e.* the red box of the second row), and so on. CP-VTON has a slight improvement on characteristics-preserving, but it still failed to keep the non-target parts (*i.e.* the green box of the second row) and deal with large occlusions (*i.e.* the green box of the fourth row). In a word, VITON and CP-VTON are bad at preserving characteristics of target clothes and reference person (mentioned in Section III-A and Section III-B).

In comparison to VITON and CP-VTON, VTNFP preserves more characteristics on account of using segmentation representation to preserve the non-target parts, but it does not contain enough details. For example, the thin shoulder straps become wider in the try-on image (*i.e.* the green box of the first row) and the hands are unclear (*i.e.* the red box of the third row). This happens because of the unawareness of the semantic layout and the relationship within the layout. ACGPN performs better than VTNFP, it can preserve hand detail but also failed to generate try-on details. It is because ACGPN uses TPS to warp clothes (mentioned in Section III-B), and it uses a simple UNet to fuse the features, which makes it difficult to trade off characteristics of the target clothes and the reference person (mentioned in Section III-C). Such as the sleeves of



Fig. 4: On the Zalando-Dataset. Qualitative comparisons of CP-VTON and AVTON in different types of clothes. Our AVTON adapts successfully in all-type clothing try-on task.

error lengths (*i.e.* the green box of the third row) that caused by TPS [6], and the artifacts near the neck (*i.e.* the red box of the fourth row, as the inner collar is not ignored) that caused by preserving too many characteristics of the warped clothes.

However, AVTON does better both in preserving characteristics and trading off characteristics. Benefited from the LPM, it predicts limbs firstly and then provides limbs' information to the TOFM, it helps solve occlusion problems(*e.g.* the blue box of the fourth row, the arm is clearer than others). What's more, the IGMM warps clothes more reasonable to preserve styles and patterns (*e.g.* the blue box of the third row, sleeves are the same length and patterns are clear), and the TOFM makes the try-on image more realistic due to characteristics' trade-off (*e.g.* the yellow box of the fourth row, the inner collar should be ignored). In a word, AVTON can generate more realistic try-on images than VITON, CP-VTON, VTNFP, and ACGPN.

On the Zalando-Dataset. For a fair comparison, we retrained CP-VTON on Zalando-Dataset and selected the besttrained model. We put on different types of clothes for a person (Fig. 4) and put on one type of clothes for person of different complexity levels(Fig. 5). It is evident from the results that CP-VTON is not suitable for the all-type clothing try-on task, as CP-VTON uses TPS to warp clothes that we motioned in Section III-B. And there are some defects such as the significant color difference of arm (Medium image of the third row in Fig. 4), as CP-VTON uses a single UNet to generate limbs in the try-on step that we motioned in Section III-A. However, AVTON can deal with these issues, it is benefited from the LPM, IGMM, and TOFM, where the LPM can predict the reasonable color of skin, the IGMM can warp clothes accurately, and the TOFM can make results more realistic.

D. Quantitative Results

We employ Structure SIMilarity (SSIM) [49] to measure the similarity between try-on images and groundtruths, and Inception Score (IS) [44] to measure the visual quality of try-on

images. Specially, we measure SSIM in different complexity levels on the VITON-Dataset, and measure it in different types of clothes on the Zalando-Dataset.

On the VITON-Dataset. Table I shows a quantitative comparison of our AVTON with VITON [17], CP-VTON[47], VTNFP [52] and ACGPN [51]. In our experiments, AVTON obtains a significant lead in all these metrics over the baseline methods. Specifically, in terms of SSIM, our method exceeds the best baseline method (*i.e.* ACGPN) by 0.010, 0.014, and 0.023 for all complexity levels. And our method surpasses the best baseline method by 0.195 in terms of IS.

Methods		IC			
	Mean	Easy	Medium	Hard	IS
VITON	0.783	0.787	0.779	0.779	2.650
CP-VTON	0.745	0.753	0.742	0.729	2.757
VTNFP	0.803	0.810	0.801	0.788	2.784
ACGPN	0.845	0.854	0.841	0.828	2.829
AVTON (Vanilla)	0.813	0.820	0.810	0.798	2.880
AVTON (w/o Ψ)	0.819	0.826	0.816	0.805	2.983
AVTON (w/o LPM)	0.856	0.861	0.852	0.849	2.859
AVTON (Full)	0.859	0.864	0.855	0.851	3.024

TABLE I: **On the VITON-Dataset**. SSIM and IS results of five methods, while SSIM are measured in different complexity levels. AVTON (Vanilla), AVTON (w/o Ψ) and AVTON (w/o LPM) are for ablation study.

On the Zalando-Dataset. As shown in Table II, it presents a quantitative comparison of our AVTON with CP-VTON, where CP-VTON is retrained that mentioned in Section IV-C. Compared to the SSIM of CP-VTON, AVTON increase it by 0.053, 0.088 and 0.020 for all-type clothes. And AVTON outperforms CP-VTON by 0.420 in terms of IS.

E. Ablation Study

We evaluate the effectiveness of the LPM and Ψ (Wendland's Ψ -function). Similarly to the quantitative comparisons, we use SSIM and IS. As shown in Table I and Table II, Wendland's Ψ -function plays an important role, where AV-TON (Full) surpasses the AVTON (w/o Ψ) by 0.046 and 0.008



Fig. 5: On the Zalando-Dataset. Qualitative comparisons of CP-VTON and AVTON in different complexity levels. Our AVTON adapts successfully in cross-category try-on task and arbitrary shape and pose person.

Methods	SSIM				IS
Wiethous	Mean	Тор	Bottom	Whole	15
CP-VTON	0.758	0.732	0.792	0.812	3.556
AVTON (Vanilla)	0.809	0.775	0.868	0.824	3.967
AVTON (w/o Ψ)	0.811	0.776	0.876	0.818	3.971
AVTON (w/o LPM)	0.813	0.779	0.874	0.827	4.023
AVTON (Full)	0.819	0.785	0.880	0.832	3.976

TABLE II: **On the Zalando-Dataset**. SSIM and IS results of two methods, while SSIM are measured in different types of clothes. AVTON (Vanilla), AVTON ($w/o \Psi$) and AVTON (w/o LPM) are for ablation study.

in terms of the mean of SSIM, respectively. Here we show the visual comparison in Fig. 6, it is obvious that Wendland's Ψ -function can warp clothes locally and smoothly, while TPS warped clothed globally that affects other parts. Additionally,

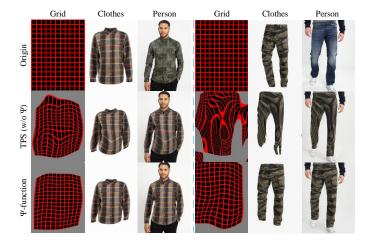


Fig. 6: Comparison of TPS and Wendland's Ψ -function.

the LPM has a great impact on IS, compared to that of AVTON (w/o LPM), it is increased by 0.165 (Table I). This of course increases IS score, as shown in Fig. 7, the results without the LPM have broken arms. Especially in Table II,

the IS of AVTON (Full) is lower than AVTON (w/o LPM). It can be explained that the testing set contains Bottom and Whole cases, most of them do not have occlusion problems. Therefore, the prediction error caused by LPM can be avoided, and the IS of the results without the LPM is higher. To sum up, the LPM is necessary for the cross-category try-on task in our experiments.

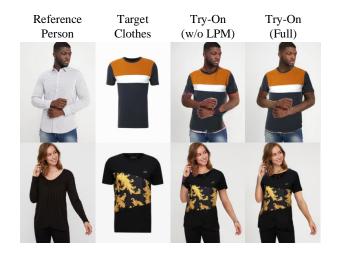


Fig. 7: Comparison of the results with and without the LPM.

F. User Study

As shown in Table III and Table IV, we conduct two user studies on the VITON-Dataset and Zalando-Dataset. Among these two studies, we compare the ACGPN [51] and our proposed method AVTON on the VITON-Dataset from easy, medium, and hard cases, respectively, and we compare the CP-VTON [47] and AVTON on the Zalando-Dataset from the top, bottom, and whole cases, respectively. Note that we only choose the ACGPN as the baseline, as it is by far the best baseline in the VITON-Dataset based methods.

Methods	Mean	Easy	Medium	Hard
ACGPN	40.9%	39.0%	40.6%	43.0%
AVTON (Full)	59.1%	61.0%	59.4%	57.0%

TABLE III: **On the VITON-Dataset**. User study results between ACGPN and our proposed method AVTON in different complexity levels.

Specifically, we invite 40 volunteers to complete the experiment. In each study, each volunteer is assigned 50 image pairs in a case and asked to select the most realistic image between two virtual try-on results. Both two studies show that the AVTON has better performance than other methods in alltype clothing and cross-category try-on tasks.

Methods	Mean	Тор	Bottom	Whole
CP-VTON	14.7%	14.1%	15.3%	14.6%
AVTON (Full)	85.3%	85.9%	84.7%	85.4%

TABLE IV: **On the Zalando-Dataset**. User study results between CP-VTON and our proposed method AVTON in different types of clothes.

G. Arbitrary Clothing Collocation

In real life, a single clothing virtual try-on cannot meet people's needs, and clothing collocation can improve people's favor of virtual try-on. Hence, We conduct an extra experiment to show the results of arbitrary clothing collocation (Fig. 8). During the experiment, we first try on tops with our AVTON and get the intermediate results, then try on bottoms based on the intermediate results, and finally get the clothing collocation results. It can be seen from the experimental results that due to the characteristics-preserving function of the LPM and IGMM, the target clothes and the reference person characteristics can still be retained after two try-on steps, and benefit from the characteristics trade-off function of TOFM, the final try-on images are natural and realistic.

V. CONCLUSION

In this paper, we propose a novel virtual try-on network, named AVTON, which aims at handling all-type clothing tryon task (tops, bottoms, and whole clothes) and cross-category try-on task (e.g. long sleeves \leftrightarrow short sleeves or long pants \leftrightarrow skirts, *etc.*). Following extensive simulation results, we can observe: 1) the Limbs Prediction Module can well predict the human body parts by preserving the characteristics of the reference person. This is especially good for handling crosscategory try-on task (e.g. long sleeves \leftrightarrow short sleeves or long pants \leftrightarrow skirts, *etc.*), where the exposed arms or legs with the skin colors and details can be reasonably predicted; 2) the developed Improved Geometric Matching Module can achieve better warped performance than TPS based method, which can well characterize the clothes feature according to the geometry of target person. 3) while the Trade-Off Fusion Module (TOFM) can fuse both the information from render person as well as warped clothes to generate a realistic tryon image. Extensive simulations based on varied try-on tasks

have been conducted. Simulation results based on Quantitative, qualitative evaluation and user study illustrate the great superiority of our AVTON over the state-of-the-art methods.

REFERENCES

- T. Alldieck, M. Magnor, W. Xu, C. Theobalt, and G. Pons-Moll. Video based reconstruction of 3d people models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8387– 8397, 2018.
- [2] R. Alp Güler, N. Neverova, and I. Kokkinos. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7297–7306, 2018.
 [3] Z. Cui, Z. Li, S. Wu, X.-Y. Zhang, and L. Wang. Dressing as a whole:
- [3] Z. Cui, Z. Li, S. Wu, X.-Y. Zhang, and L. Wang. Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks. In *The World Wide Web Conference*, pages 307–317, 2019.
- [4] B. Delaunay, S. Vide, A. Lamémoire, and V. De Georges. Bulletin de l'académie des sciences de l'urss. Classe des sciences mathématiques et naturelles, 6:793–800, 1934.
- [5] 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. *CoRR*, *abs/1906.00884*, 2019.
- [6] J. Duchon. Splines minimizing rotation-invariant semi-norms in sobolev spaces. In *Constructive theory of functions of several variables*, pages 85–100. Springer, 1977.
- [7] J. Entwistle. The fashioned body: Fashion, dress and social theory. John Wiley & Sons, 2015.
- [8] M. Fornefett, K. Rohr, and H. S. Stiehl. Elastic registration of medical images using radial basis functions with compact support. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, volume 1, pages 402–407. IEEE, 1999.
- [9] M. Fornefett, K. Rohr, and H. S. Stiehl. Radial basis functions with compact support for elastic registration of medical images. *Image and* vision computing, 19(1-2):87–96, 2001.
- [10] Y. Ge, R. Zhang, X. Wang, X. Tang, and P. Luo. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and reidentification of clothing images. In *Proceedings of the IEEE conference* on computer vision and pattern recognition, pages 5337–5345, 2019.
- [11] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017.
- [12] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.
- [13] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black. Drape: Dressing any person. ACM Transactions on Graphics (TOG), 31(4):1– 10, 2012.
- [14] X. Guo, H. Wu, Y. Gao, S. Rennie, and R. Feris. The fashion iq dataset: Retrieving images by combining side information and relative natural language feedback. *arXiv preprint arXiv:1905.12794*, 2019.
 [15] X. Han, Z. Wu, W. Huang, M. R. Scott, and L. S. Davis. Compatible
- [15] X. Han, Z. Wu, W. Huang, M. R. Scott, and L. S. Davis. Compatible and diverse fashion image inpainting. arXiv preprint arXiv:1902.01096, 2019.
- [16] X. Han, Z. Wu, Y.-G. Jiang, and L. S. Davis. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1078–1086, 2017.
 [17] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-
- [17] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An imagebased virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018.
- [18] G. Häubl and V. Trifts. Consumer decision making in online shopping environments: The effects of interactive decision aids. *Marketing science*, 19(1):4–21, 2000.
- [19] S. Hauswiesner, M. Straka, and G. Reitmayr. Virtual try-on through image-based rendering. *IEEE transactions on visualization and computer graphics*, 19(9):1552–1565, 2013.
- [20] R. He, C. Packer, and J. McAuley. Learning compatibility across categories for heterogeneous item recommendation. In 2016 IEEE 16th International Conference on Data Mining (ICDM), pages 937–942. IEEE, 2016.
- [21] M. Hosseinzadeh and Y. Wang. Composed query image retrieval using locally bounded features. In *Proceedings of the IEEE/CVF Conference* on Computer Vision and Pattern Recognition, pages 3596–3605, 2020.
- [22] W.-L. Hsiao, I. Katsman, C.-Y. Wu, D. Parikh, and K. Grauman. Fashion++: Minimal edits for outfit improvement. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5047–5056, 2019.

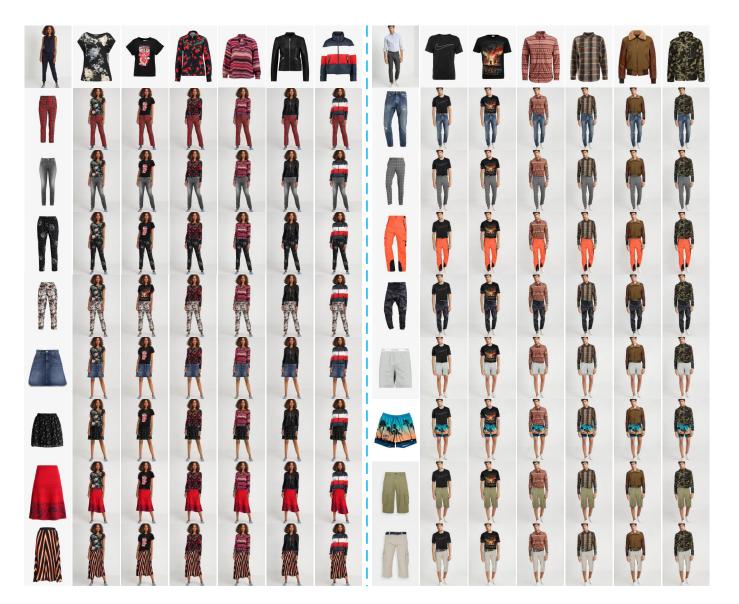


Fig. 8: The clothing collocation results of two parts. The upper left image of each part represents the reference person, the upper column of each part represents the target tops, and the left column of each part represents the target bottoms. It can be seen that our AVTON can match clothes arbitrarily.

- [23] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE* conference on computer vision and pattern recognition, pages 1125– 1134, 2017.
- [24] T. Iwata, S. Wanatabe, and H. Sawada. Fashion coordinates recommender system using photographs from fashion magazines. In *IJCAI*, volume 22, page 2262. Citeseer, 2011.
- [25] S. Jandial, A. Chopra, P. Badjatiya, P. Chawla, M. Sarkar, and B. Krishnamurthy. Trace: Transform aggregate and compose visiolinguistic representations for image search with text feedback. arXiv preprint arXiv:2009.01485, 2020.
- [26] T. Karras, T. Aila, S. Laine, and J. Lehtinen. Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196, 2017.
- [27] T. Karras, S. Laine, and T. Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference* on computer vision and pattern recognition, pages 4401–4410, 2019.
- [28] J. Li, J. Zhao, Y. Wei, C. Lang, Y. Li, T. Sim, S. Yan, and J. Feng. Multiple-human parsing in the wild. arXiv preprint arXiv:1705.07206, 2017.
- [29] P. Li, Y. Xu, Y. Wei, and Y. Yang. Self-correction for human parsing. arXiv preprint arXiv:1910.09777, 2019.
- [30] Y. Li, L. Cao, J. Zhu, and J. Luo. Mining fashion outfit composition

using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia*, 19(8):1946–1955, 2017.

- [31] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan. Deep human parsing with active template regression. *IEEE transactions* on pattern analysis and machine intelligence, 37(12):2402–2414, 2015.
- [32] Y.-L. Lin, S. Tran, and L. S. Davis. Fashion outfit complementary item retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3311–3319, 2020.
- [33] J. Liu and H. Lu. Deep fashion analysis with feature map upsampling and landmark-driven attention. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.
 [34] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering
- [34] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.
- [35] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang. Fashion landmark detection in the wild. In *European Conference on Computer Vision* (ECCV), October 2016.
- [36] A. Mir, T. Alldieck, and G. Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7023– 7034, 2020.
- [37] S. Mo, M. Cho, and J. Shin. Instagan: Instance-aware image-to-image

translation. arXiv preprint arXiv:1812.10889, 2018.

- [38] A. Neuberger, E. Borenstein, B. Hilleli, E. Oks, and S. Alpert. Image based virtual try-on network from unpaired data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5184–5193, 2020.
 [39] N. Pandey and A. Savakis. Poly-gan: Multi-conditioned gan for fashion
- [39] N. Pandey and A. Savakis. Poly-gan: Multi-conditioned gan for fashion synthesis. *Neurocomputing*, 414:356–364, 2020.
- [40] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2337– 2346, 2019.
- [41] G. Pons-Moll, S. Pujades, S. Hu, and M. J. Black. Clothcap: Seamless 4d clothing capture and retargeting. ACM Transactions on Graphics (TOG), 36(4):1–15, 2017.
- [42] A. Raj, P. Sangkloy, H. Chang, J. Hays, D. Ceylan, and J. Lu. Swapnet: Image based garment transfer. In *European Conference on Computer Vision*, pages 679–695. Springer, 2018.
- [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*, pages 234–241. Springer, 2015.
- [44] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. *Advances in neural information processing systems*, 29:2234–2242, 2016.
- [45] Y.-S. Shih, K.-Y. Chang, H.-T. Lin, and M. Sun. Compatibility family learning for item recommendation and generation. In *Thirty-Second* AAAI Conference on Artificial Intelligence, 2018.
- [46] A. Veit, B. Kovacs, S. Bell, J. McAuley, K. Bala, and S. Belongie. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4642–4650, 2015.
- [47] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang. To-ward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018.
 [48] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro.
- [48] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.
 [49] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image
- [49] 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*, 13(4):600–612, 2004.
- [50] H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. Advances in computational Mathematics, 4(1):389–396, 1995.
- [51] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7850–7859, 2020.
 [52] R. Yu, X. Wang, and X. Xie. Vtnfp: An image-based virtual try-on
- [52] R. Yu, X. Wang, and X. Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10511– 10520, 2019.
- [53] W. Zeng, M. Zhao, Y. Gao, and Z. Zhang. Tilegan: category-oriented attention-based high-quality tiled clothes generation from dressed person. NEURAL COMPUTING & APPLICATIONS, 2020.
- [54] M. Zhao, Y. Liu, X. Li, Z. Zhang, and Y. Zhang. An end-to-end framework for clothing collocation based on semantic feature fusion. *IEEE MultiMedia*, 27(4):122–132, 2020.
- [55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings* of the IEEE international conference on computer vision, pages 2223– 2232, 2017.
- [56] S. Zhu, S. Fidler, R. Urtasun, D. Lin, and C. C. Loy. Be your own prada: Fashion synthesis with structural coherence. In *International Conference* on Computer Vision (ICCV), October 2017.



Yu Liu received the B.E degree from Donghua University in 2019. He is currently a Master candidate in Donghua University, Shanghai, P. R. China. His research interests include deep learning, pattern recognition, image synthesis.



Mingbo Zhao (SM'19) received the Ph.D. degree in computer engineering from the Department of Electronic Engineering, City University of Hong Kong, Kowloon, Hong Kong S.A.R., in Jan. 2013. After that he also worked in City University of Hong Kong as a Postdoctoral researcher. He is currently a full professor in Donghua University, Shanghai, P. R. China. He has authored or co-authored over 50 technical papers published at prestigious international journals and conferences, including the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA

ENGINEERING, the IEEE TRANSACTIONS ON IMAGE PROCESSING, the IEEE TRANSACTIONS ON INDUSTRIAL INFORMATICS, the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS, the Pattern Recognition, Neural Networks, Knowledge based Systems, His current research interests include pattern recognition and Machine Learning. He is serving as an Associate Editor (AE) for IEEE ACCESS and Leader Guest Editor for Neural Computing and Application.



Haijun Zhang (SM'19) received the B.Eng. and master's degrees from Northeastern University, Shenyang, China, and the Ph.D. degree from the Department of Electronic Engineering, City University of Hong Kong, Hong Kong, in 2004, 2007, and 2010, respectively. He was a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada, from 2010 to 2011. Since 2012, he has been with the Shenzhen Graduate School, Harbin Institute of Technology, Harbin, China, where he is currently

a Full Professor of computer science. His current research interests include multimedia data mining, machine learning, and computational advertising. Dr. Zhang is also an Associate Editor of Neurocomputing, Neural Computing and Applications, and Pattern Analysis and Applications.



Zhao Zhang (SM'17) received the Ph.D. degree from the City University of Hong Kong, Hong Kong, in 2013. He was a Visiting Research Engineer with the National University of Singapore, Singapore, from February to May 2012. He visited the National Laboratory of Pattern Recognition, Chinese Academy of Sciences, from September to December 2012. From October 2013 to October 2018, he was an Associate Professor with Soochow University, Suzhou, China, where he was with the School of Computer Science and Technology. He is currently

with the School of Computer and Information, Hefei University of Technology, Hefei, China. His current research interests include multimedia data mining and machine learning, image processing, and computer vision. Dr. Zhang is a member of the Association for Computing Machinery (ACM). He has been serving as a Senior PC Member or the Area Chair of ECAI, BMVC, PAKDD, and ICTAI, and a PC Member for more than ten popular prestigious conferences, including CVPR, ICCV, IJCAI, AAAI, ACM MM, ICDM, CIKM, and SDM. He has served or is serving as an Associate Editor (AE) for IEEE Trans. on Image Processing and Signal Processing.



Shuicheng Yan (F'17) is currently the Vice President and the Chief Scientist with Qihoo 360 Technology Co., Ltd., Beijing, China, where he is the Head of the 360 Artificial Intelligence Institute. He is also a tenured Associate Professor with the National University of Singapore, Singapore. He has authored or coauthored about 500 high-quality technical articles, with Google Scholar citation over 25 000 times and the H-index of 70. His current research interests include computer vision, machine learning, and multimedia analysis. Dr. Yan is a fellow of

the International Association for Pattern Recognition and the Distinguished Scientist of the Association for Computing Machinery (ACM). His team received seven times winner or honorable-mention prizes in five years over the PASCAL Visual Object Classes and the ImageNet Large Scale Visual Recognition Challenge competitions that are the core competitions in the field of computer vision, along with over ten times best (student) paper awards, especially a Grand Slam at the ACM Multimedia, the top conference in the field of multimedia, including the Best Paper Award, the Best Student Paper Award, and the Best Demo Award. He was a Thomson Reuters Highly Cited Researcher from 2014 to 2016.