Arbitrary Virtual Try-on Network: Characteristics Preservation and Tradeoff between Body and Clothing

YU LIU and MINGBO ZHAO, Donghua University, China
ZHAO ZHANG, Hefei University of Technology, China
YUPING LIU, Fudan University, China
SHUICHENG YAN, The National University of Singapore, Singapore

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 article 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 preserving the human body parts by preserving the characteristics of the reference person. This is especially good for handling cross-category try-on task (e.g., long sleeves ↔ short sleeves or long pants ↔ skirts), 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 $\Psi$-function); (3) Trade-Off Fusion Module, which is to tradeoff 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 fine-tune 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.

CCS Concepts: • Computing methodologies → Computer vision;

Additional Key Words and Phrases: Deep learning, virtual try-on, generative adversarial networks, artificial intelligence in fashion

This work is supported by Graduate Student Innovation Fund of Donghua University (GSIF-DH-M-2021006), supported by National Natural Science Foundation of China (61971121, 62072151), partially supported by Anhui Provincial Natural Science Fund for the Distinguished Young Scholars (2008085J30), Open Foundation of Yunnan Key Laboratory of Software Engineering (2023SE103), CCF-Baidu Open Fund and CAAI-Huawei MindSpore Open Fund.

Authors’ addresses: Y. Liu (Corresponding author) and M. Zhao (Corresponding author), Donghua University, Shanghai, China; e-mails: 2191408@mail.dhu.edu.cn, mzhao4@dhu.edu.cn; Z. Zhang (Corresponding author), Hefei University of Technology, Hefei, China; e-mail: cszzhang@gmail.com; Y. Liu (Corresponding author), Fudan University, Shanghai, China; e-mail: liuyp@fudan.edu.cn; S. Yan, The National University of Singapore, Singapore City, Singapore.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

1551-6857/2024/01-ART123 $15.00
https://doi.org/10.1145/3636426
1 INTRODUCTION

In modern society, clothing plays an important role in human daily life, as proper outfits can enhance people’s beauty and personal quality [1–4]. 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 [5], it is necessary for researchers to conduct considerable researches on image-based virtual try-on technology [6, 7], 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 [8, 9], clothes parsing [10–13], clothes attributions and categories recognition [8, 9, 14], clothes collocation [15–22], clothes image retrieval [23–29], and fashion edit [30, 31]. 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 [32] 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 [33], CP-VTON [14], VTNFP [34], ACGPN [35], et al. [7, 36–42]. Extensive simulation results show the effectiveness of these works.

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 [33] is a benchmark dataset used in most prior arts. It only 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 is necessary and important, which can be further utilized to develop a method for handling all-type clothes try-on task; (2) cross-category try-on task (e.g., long sleeves↔short sleeves or long pants↔skirts) 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 [14, 33, 34, 36] 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 bad try-on performances, e.g., the limbs of human beings is covered by clothes, the color of the skin is wrongly painted and the hand details cannot be properly generated, may be appeared; (3) current methods are not good at trading off characteristics of the warped clothes and the reference person, such as...
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 [14] 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 [43], while the AVTON (Full) is just the opposite.

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 article, we propose a new image-based virtual try-on 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 https://www.zalando.co.uk. As shown in Figure 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 try-on task, where the bottom clothing images include trousers, shorts, and skirts, and so on, 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 image-based 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 ↔ short sleeves or long pants ↔ skirts, and so on., 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) [44] 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 [43] 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 tradeoff 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 [45]). Experiments show that AVTON significantly outperforms the state-of-the-art methods for virtual try-on [14, 33–35], and can generate realistic try-on images in all-type clothing try-on task (Figure 1).

The main contributions of this article are summarized as follows:

1. 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;
2. 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 $\Psi$-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;
4. 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 cross-category try-on task, and outperforms the state-of-the-art methods both qualitatively and quantitatively.

The rest of this article is organized as follows: In Section 2, we briefly review related work for fashion AI and GAN based virtual try-on methods; in Section 3, we will give a detailed description of the proposed Arbitrary Virtual Try-on method; in Section 4, the newly-collected dataset is introduced and extensive simulations are conducted for evaluations; the final conclusions are drawn in Section 5.

2 RELATED WORK

Image Synthesis: Generative Adversarial Network (GAN) [32] is a popular and effective tool in image synthesis [46–49] 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 [40, 46, 50], photo inpainting [42], clothing translation [23, 39], and so on. Extensive learning tasks verify the effectiveness of image synthesis.

Fashion Artificial Intelligence (AI): Benefiting from the development and popularization of deep learning, fashion-related work has made great progress recently due to its powerful feature extraction ability to grasp the extensive mid-level image representations. According to the real-world application area in computer vision, current works of fashion AI can be generally divided into the following stages: (1) detection stage, which is to find where the clothes is in a multi-media data. These applications include clothes detection and segmentation [8, 9], clothing parsing [10–13, 51, 52], clothing landmark detection [8, 9, 25]; (2) Recognition and Analysis stage, which is to discover the detailed information and relationships for clothes. Typical applications
include clothes attributions and categories recognition [8, 14, 53], clothes Re-Identity (retrieval) and recommendation [24–29], clothes image captioning [54], and fashion compatibility and collocation [15–22]; (3) generation and human interaction stage, which is related higher stage including clothes template generation [23], virtual try-on [14, 33], fashion IQ (Q&A systems) [26, 55], et al. Recently, due to successful utilization of GAN in different learning areas, GAN based virtual try-on task is one of the popular tasks of fashion AI that has been widely studied and has made great achievements [14, 33–35, 39, 41, 56, 57].

Virtual Try-On: Virtual try-on belongs to fashion image synthesis, which can be roughly divided into two categories: 3D model-based methods [39, 41, 56, 57] and 2D image-based methods [14, 33–35]. 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 [33] and CP-VTON [14] are the two first works. The key idea of VITON is to exploit a TPS [44] 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 [34] and ACGPN [35] 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. Recently, SPG-VTON [58] has further proposed an end-to-end semantic prediction guidance multi-pose virtual try-on network, which aims to generate the try-on image by preserving the desired clothing onto a reference person under arbitrary poses simultaneously. In addition, all these methods are trained via VITON dataset [33] 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 cross-category try-on task. Though the work in [40] has proposed a variant of CycleGAN [50], namely InstaGAN, to handle this task (long sleeves ↔ short sleeves, long pants ↔ skirts, etc), the final try-on results cannot be controlled due to the strategy of CycleGAN. This article aims to address the above problems by proposing a new method for all-type clothing and cross-category try-on tasks with more image-realistic performance.

3 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 Figure 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 [43] to improve the TPS based method [44] 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 a composition mask and rendered person by a pair of UNet [45] and a fusion part. Especially, the TOFM takes advantage of GAN [32].

3.1 Limbs Prediction Module (LPM)

The purpose of designing the LPM is to predict the exposed limbs and preserve the primary human body information (i.e., head, hand details, and non-target human body parts). Most earlier methods...
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 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 \( I_t \).

generate exposed limbs in try-on steps but neglected primary human body information, which may generate unreasonable color of skin and occlusion. This can usually happen when handling cross-category try-on task, i.e., long sleeve \( \rightarrow \) short sleeve, or long pant \( \rightarrow \) short pant, as some limbs originally covered by the clothes will be exposed. We thereby propose LPM to address these issues.

Given a reference person image \( I_r \), LPM takes the target clothes \( c \), the human body information (i.e., the pose and shape information \( d \) extracted by DensePose [59], the head and non-target human body parts information \( p \) as mentioned in Section 4.2) as inputs to predict exposed limbs \( \hat{l} \). In detail, the target clothes \( c \) and the human body information \( H = d \oplus p \) \((\oplus: \text{concatenation})\) are firstly encoded as the input features. They are then formed as a single tensor by a correlation layer and input to the decoder, which is to calculate the feature correlation by \( M = \hat{M}_c \hat{M}_H \). \( \hat{M}_c \) and \( \hat{M}_H \) are the normalized encoder features of body information \( H \) and clothes information \( c \), respectively. Finally, the exposed limbs are predicted by the decoder. The encoder and correlation layer are similar to CP-VTON’s GMM step [14], while the human body information \( d \)’s encoder-decoder layers are similarly to UNet [45] structure shown in Figure 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 \( I_r \):

\[
L_{\text{LPM}} = \lambda_{\text{L1}} \| \hat{l} - l_r \|_1 + \lambda_{\text{vgg}} L_{\text{VGG}}(\hat{l}, l_r),
\]

where

\[
L_{\text{VGG}}(\hat{l}, l_r) = \sum_{i=1}^{5} \lambda_i \| \phi_i(\hat{l}) - \phi_i(l_r) \|_1,
\]

is the VGG perceptual loss, where \( \lambda_{\text{L1}} \) and \( \lambda_{\text{vgg}} \) are the tradeoff parameters for two loss terms in Equation (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 \( \phi \), which is a VGG19 pre-trained on ImageNet.

### 3.2 Improved Geometric Matching Module (IGMM)

The old way of warping clothes is based on TPS \([44] \) with \( r^2 \log r \) as RBFs. This method yields minimal bending energy properties measured over the whole image. However 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 \( \Psi \)-function of Wendland \([43] \) as RBFs. As mentioned in \([60, 61] \), 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:

\[
\psi_{d,k}(r) = \int_{(1-r)^{d/2}k+1}^{r} z^{k} (1-r)^{d/2} \, dz, \tag{3}
\]

where

\[
\begin{align*}
(1 - r)^k & = \begin{cases} (1 - r)^v & 0 \leq r < 1 \\ 0 & r \geq 1 \end{cases}, \\
I_\psi(r) & := \int_r^\infty t \psi(t) dt \quad r \geq 0.
\end{align*}
\]

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

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

where \( \alpha \) is a learnable parameter as same as the spatial transformation parameters \( \theta \). Inspired by CP-VTON \([14] \), we use the same structure to learn these parameters. As shown in Figure 2, our IGMM firstly extract high-level features of the target clothes \( c \) and the refined human body information \( \hat{H} = d \oplus l \oplus h \) (\( h \) 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 \( \theta \) and \( \alpha \). Finally, a transformation \( \Psi_{\theta, \alpha} \) based on Equation (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 \( \alpha \), which shows a significant positive correlation between spatial warping range and \( \alpha \), and we concluded that the best \( \alpha \) should meet the condition: \( \alpha \geq D \), where \( D \) is the maximum distance among control points in Delaunay triangles \([62] \). 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 \( \alpha = \lambda_a \hat{\alpha} + D \), where \( \hat{\alpha} \) is the sigmoid’s output of the regression network and \( \lambda_a \) 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 \( I_r \):

\[
L_{\text{IGMM}} = \| \hat{c} - c_r \|_1 = \| \Psi_{\theta, \alpha}(c) - c_r \|_1. \tag{5}
\]

### 3.3 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 them to make images realistic. One of the common methods \([33] \) is to produce a composition mask for fusing UNet \([45] \) rendered person with warped clothes.
and finally to produce a refined result. Although it can refine the coarse try-on image, it lacks preserving characteristics of the warped clothes. Another common method [14] is to utilize a UNet to render a person image and predict a composition mask simultaneously, and then synthesize 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 [34, 35] have some problems with trading off characteristics between the warped clothes and the reference person (e.g., artifacts near the neck regions).

In this article, we adopt a GAN [32] 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 (Figure 2 residual blocks) for handling the above problems. Specifically, in the proposed module, the warped clothes’ features \( \mathcal{F}_c \) and the refined human body features are firstly summed element-wise, so that the fused features can be obtained, i.e., \( \mathcal{F} (\mathcal{F}_c + \hat{F} \rightarrow \mathcal{F}) \). Then the fused features \( \mathcal{F} \) are concatenated with \( \mathcal{F}_c \) and \( \hat{F} \), and decoded respectively to get predicted mask \( \hat{m} \) and predicted rendered person \( r \). Finally, similarly to CP-VTON [14], \( \hat{c} \) and \( r \) are fused together using \( \hat{m} \) to synthesize try-on image \( I_t \):

\[
I_t = \hat{m} \odot \hat{c} + (1 - \hat{m}) \odot r,
\]

where \( \odot \) represents element-wise matrix multiplication. Additionally, we use the multi-scale discriminators \( D \) that is similar to pix2pixHD [63].

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 (Equation (2)) and LSGAN loss:

\[
L_G = \lambda_{L1} \| I_t - I_g \|_1 + \lambda_{vgg} L_{VGG}(I_t, I_g) + \\
\lambda_{mask} \| \hat{m} - m_r \|_1 + \lambda_{lsgan}(D(\hat{c}, \hat{H}, I_t) - 1)^2,
\]

while the discriminator’s loss is

\[
L_D = ((D(\hat{c}, \hat{H}, I_t))^2 + (D(\hat{c}, \hat{H}, I_g) - 1)^2)/2,
\]

where \( I_g \) is the ground truth image, and \( I_g = I_t \) in training stage, \( m_r \) is the mask of \( c_r \). In our experiments, we set \( \lambda_{L1}, \lambda_{vgg}, \) and \( \lambda_{mask} \) to 10, while set \( \lambda_{lsgan} \) to 1.

4 EXPERIMENTS

4.1 Dataset Description

In this work, we evaluate the performance of our proposed work based on two datasets: the VITON dataset [33] that is used in VITON [33], CP-VTON [14] and ACGPN [35] et al., and the newly-collected Zalando dataset. In this article, we call them VITON-Dataset and Zalando-Dataset respectively.

The VITON-Dataset contains 16253 frontal-view woman and top clothing image pairs, which is split into a training set and a testing set with 14,221 and 2,032 pairs respectively. We also use the strategy in ACGPN [35] to score the complexities of images in dataset and divide them into Easy, Medium, and Hard levels, which is used to further evaluate the proposed method and other state-of-the-art methods for handling different levels of try-on task.
Though VITON-Dataset is currently the most widely-used benchmark dataset for virtual try-on, it has a key problem that it only contains top clothes while lacks of bottom and whole clothes. In other words, a model trained only on top clothes will definitely not be able to handle try-on task for other types of clothes such as bottom clothes (skirt, pants et al.) and whole clothes. The reason is obvious since the bottom and whole clothes are not involved in training the model. In order to solve the above problem, we in this work collect a new dataset, namely Zalando-Dataset, which is utilized to handle Arbitrary Virtual Try-On task.

4.2 The Newly-collected Try-on Dataset

We in this subsection introduce the newly-collected Zalando-Dataset. The Zalando-Dataset is crawled from https://www.zalando.co.uk/. It contains 34928 frontal-view human (include man and woman) and clothing (include top, bottom, and whole) image pairs. In our study, we split it into training set and testing set with 32,746 and 2,182 image pairs, respectively. In detail, the training set contains 19,185 tops, 10,587 bottoms, and 2,974 whole clothes, while the testing set contains 1,310 tops, 692 bottoms, and 180 whole clothes. We also score the complexity of each image following the work in [35].

In order to show the superiority of the collected Zalando-Dataset, we first illustrate the categories of the dataset, which can be shown in Figure 3. In detail, the division of dataset based on the categories and genders is shown in Figure 3 (a). From Figure 3(a), we can see that the collected try-on dataset has involved almost all categories of clothes including both top and bottom clothes for man as well as top, bottom and whole clothes for woman. Figure 3(b) further show the statistics and distribution of all categories of clothes according to gender and types, where the images of top clothes take up 58.68% while those of bottom and whole clothes take up the remaining 41.32%. Since the conventional VITON-Dataset only involves images with top clothes while lacks of those of bottom and whole clothes, the collected Zalando-Dataset is an extension to VITON-Dataset by involving more clothing categories, which is good for handling real-world arbitrary try-on task.

In order to further show the superiority, we then analyze the characteristics of dataset and compare with those of VITON-Dataset. Here, we choose two measurements for comparisons: (1) the size of the dressed clothes taking up the whole referenced image, and (2) the complexity scores for the referenced image (only for top clothes), which are firstly defined in ACGPN [35] and are
Fig. 4. On the VITON-Dataset. Qualitative comparisons of VITON [33], CP-VTON [14], VTNFP [34], ACGPN [35] 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 target clothes with the IGMM. What’s 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.

Fig. 5. (a) the distribution of the size of the dressed clothes taking up the whole referenced image; (b) the distribution of the complexity scores for the referenced image (only for top clothes).

divided into easy, medium and hard cases. The characteristics analysis of proposed dataset and VITON-Dataset based on two measures are shown in Figure 5(a).

From Figure 5(a), we can see that the size of the dressed clothes taking up the whole referenced image has a wide range from 0.2 to 0.6 in the collected dataset. While for VITON-Dataset, such range mainly focus on about 0.45. This means that the collected dataset has different sizes of the dressed clothes and is more extensive in data selection. This can also be more realistic and
closer to the real-world cases. As a result, a try-on method that can handling different size of the dressed clothes is more useful and practical, which is good to satisfy the requirement of user. From Figure 5(b), we can see that the distribution of the score curve of the collected dataset is on the left of that of VITON-Dataset. This indicates that the complexity of the proposed dataset is higher than that of VITON-Dataset. Therefore, a method trained on such dataset can handle more complex try-on task such as limb intersections and torso occlusions, which is good to enhance the robustness and adaptiveness of the model.

It should be noted that a key technique is how to get the human body mask (i.e., head, limbs and other non-target human body parts), which is already integrated in the VITON-Dataset. A feasible solution for our collected Zalando-Dataset is to involve additional annotations from human laboring for utilization. But labeling a large number of images is time-consuming and costly. In order to solve the problem, we in this work choose a special human parsing method, namely, Self-Correction-Human-Parsing (SCHP) [10], to get the correct head, hand detail, and non-target human body parts in our collected Zalando-Dataset. In detail, we utilize the inference model of SCHP to directly infer the mask of the human body parts we want. The reason we choose SCHP for human parsing is that it is a noise-tolerant method, which can well handle incorrect labels in ground-truth masks. This can usually happen when the annotators cannot distinguish the boundary between different semantic parts that have similar appearances. In addition, we use the Dense-Pose [59] to extract the pose and shape information into IUV values.

4.3 Implementation Details

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

**Training.** Followed by steps in Figure 2, we first train the LPM and then use the LPM’s trained results to train the IGMM, followed by training the TOFM with the trained results from 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 the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$, and the learning rate is fixed at 0.0001. Additionally, the resolution for all input and output images is $256 \times 192$, and we use a single NVIDIA 2080Ti GPU in our experiments.

**Testing.** We use the same steps as the training stage to test the modules. During the VITON-Dataset testing stage, we test our modules in easy, medium, and hard cases respectively, and evaluate the results qualitatively and quantitatively. During the Zalando-Dataset testing stage, we evaluate the qualitative and quantitative results for the top, bottom, and whole clothing cases respectively. We then perform ablation experiments both on the VITON-Dataset and Zalndo-Dataset testing stages. Note that we also analyze the qualitative results in the easy, medium, and hard cases respectively in the Zalando-Dataset testing stage (there are complexity labels in Figure 6 and Figure 7).

4.4 Qualitative Results

**On the VITON-Dataset.** Figure 4 shows visual comparisons of our proposed method with VITON [33], CP-VTON [14], VTNFP [34], and ACGPN [35]. To save a lot of work on reproducing them (VTNFP has no official code), we refer to the results from the article of ACGPN. Note that we validate the official code of ACGPN qualitatively and quantitatively and receive the same results as with ACGPN.

Among these results, VITON is the worst method of preserving characteristics, such as the blurred hands (i.e., the red box on the first row), the disappeared pattern (i.e., the red box on the second row), and so on. CP-VTON has a slight improvement on characteristics-preserving, but it still
fails to keep the non-target parts (i.e., the green box on the second row) and deal with large occlusions (i.e., the green box on the fourth row). In short, VITON and CP-VTON are bad at preserving characteristics of the target clothes and reference person (mentioned in Section 3.1 and Section 3.2).

In comparison to VITON and CP-VTON, VTNFP preserves more characteristics by 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 on the first row) and the hands are unclear (i.e., the red box on the third row). This happens because of an unawareness of the semantic layout and the relationship within the layout. ACGPN performs better than VTNFP, in that it can preserve hand details, but it also fails to generate try-on details. This is because ACGPN uses TPS to warp clothes (mentioned in Section 3.2), and it uses a simple UNet to fuse the features, which makes it difficult to tradeoff characteristics of the target clothes and the reference person (mentioned in Section 3.3). These include the sleeves of error lengths (i.e., the green box on the third row) that are caused by TPS [44], and the artifacts near the neck (i.e., the red box on the fourth row, as the inner collar is not ignored) that resulted from preserving too many characteristics of the warped clothes.

However, AVTON does better both in preserving characteristics and trading off characteristics. Benefiting from the LPM, it predicts limbs first and then provides limb information to the TOFM, which helps solve occlusion problems (e.g., the blue box on the fourth row, the arm is clearer than others). What’s more, the IGMM warps clothes more reasonably to preserve styles and patterns (e.g., the blue box on the third row, sleeves are the same length and patterns are clear), and the TOFM makes the try-on image more realistic due to tradeoff (e.g., the yellow box on the fourth row, the inner collar should be ignored). In a nutshell, AVTON can generate more realistic try-on images than VITON, CP-VTON, VTNFP, and ACGPN.
Table 1. On the VITON-Dataset

| Methods          | SSIM↑ / LPIPS↓ | IS↑ |
|------------------|----------------|-----|
| VITON            | 0.783 / 0.183  | 0.779 / 0.199 | 2.650 |
| CP-VTON          | 0.745 / 0.238  | 0.742 / 0.243 | 2.757 |
| VTNFP            | 0.803 / 0.155  | 0.801 / 0.158 | 2.784 |
| ACGPN            | 0.845 / 0.107  | 0.841 / 0.110 | 2.829 |
| DCTON            | 0.847/0.105    | 0.847/0.090  | 2.881 |
| AVTON (Vanilla)  | 0.813 / 0.135  | 0.810 / 0.137 | 2.880 |
| AVTON (w/o Ψ)   | 0.819 / 0.123  | 0.816 / 0.126 | 2.983 |
| AVTON (w/o LPM)  | 0.856 / 0.090  | 0.852 / 0.093 | 2.859 |
| AVTON (Full)     | **0.859 / 0.077** | **0.855 / 0.080** | **3.024** |

SSIM and LPIPS are measured at different complexity levels. AVTON (Vanilla), AVTON (w/o Ψ), and AVTON (w/o LPM) are used for ablation studies.

On the Zalando-Dataset. For fair comparisons, we retrain CP-VTON, Outfit-VITON [37] and ACGPN using the Zalando-Dataset and select the best-trained models. We put on different types of clothes for a person (Figure 6) and put on one type of clothes for persons of different complexity levels (Figure 7). 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 3.2 and uses a single UNet to generate limbs in the try-on step that we motioned in Section 3.1. Outfit-VITON performs better than CP-VTON, but the content of the generated images are incomplete (the first row in Figure 6). This can be reasonable because the Outfit-VITON first decouples the features of the clothes and human body, and then decodes them to generate the try-on image. This procedure does not consider the preservation of spatial information, making the image incomplete. In comparison to CP-VTON and Outfit-VITON, ACGPN generates more realistic try-on results, but there are still some issues mentioned in the VITON-Dataset. However, AVTON can deal with these issues by allowing the LPM to accurately predict the color of skin, the IGMM to warp clothes accurately, and the TOFM to make results more realistic.

4.5 Quantitative Results

We employ Structure SIMilarity (SSIM) [64] and Learned Perceptual Image Patch Similarity (LPIPS) [65] to measure the similarity between try-on images and groundtruths, and Inception Score (IS) [66] to measure the visual quality of try-on images. Specifically, we measure SSIM and LPIPS at different complexity levels on the VITON-Dataset, and measure them with different types of clothes on the Zalando-Dataset.

On the VITON-Dataset. Table 1 shows quantitative comparisons of our AVTON with VITON, CP-VTON, VTNFP, and ACGPN. In our experiments, AVTON obtains a significant lead in all these metrics over baseline methods. Specifically, the SSIM of our method improves by 0.010, 0.014, and 0.023 respectively over that of the best baseline method (i.e., ACGPN) at each complexity level. For LPIPS, our method beats the best baseline method (i.e., ACGPN) by 0.028, 0.030, and 0.036 respectively at each complexity level. And our method surpasses the best baseline method (i.e., ACGPN) by 0.195 in terms of IS.

On the Zalando-Dataset. As shown in Table 2, we present quantitative comparisons of our AVTON with CP-VTON, Outfit-VITON, and ACGPN, in which CP-VTON, Outfit-VITON, and ACGPN are retrained as mentioned in Section 4.4. The SSIM of our method improves by 0.007, 0.020, and 0.001 respectively over that of the best baseline method (i.e., ACGPN) for each type of clothing. For LPIPS, our method beats the best baseline method (i.e., ACGPN) by 0.014, 0.030, and 0.015 respectively for each type of clothing. And our method surpasses the best baseline method (i.e., ACGPN) by 0.099 in terms of IS.
4.6 Ablation Study

Similarly to the quantitative comparisons, we evaluate the effectiveness of both the LPM and $\Psi$ (Wendland’s $\Psi$-function) using SSIM, LPIPS, and IS. As shown in Table 1 and Table 2, Wendland’s $\Psi$-function plays an important role, where AVTON (Full) surpasses the AVTON (w/o $\Psi$) by 0.046 and 0.008 in terms of the mean of SSIM, respectively. Here we show the visual comparison in Figure 8 and Figure 9, where Wendland’s $\Psi$-function can warp clothes locally and smoothly, while TPS warps clothes globally. Additionally, the LPM has a significant impact on LPIPS and IS.

The LPIPS of AVTON (Full) is reduced by 0.013 (Table 1) and the IS of AVTON (Full) is increased by 0.165 (Table 1). Note that in Table 2, the IS of AVTON (Full) is lower than that of AVTON (w/o LPM). It can be explained that the testing set contains Bottom and Whole cases, most of which 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. In summary, the LPM is necessary for the cross-category try-on task in our experiments.

Here, in order to further show why LPM is good for handling cross-category try-on task (e.g., long sleeves/pants $\leftrightarrow$ short sleeves/pants), we give another ablation study for analysis, where we aim to show the full LPM can well predict some exposed arms or legs when long sleeves/pants $\leftrightarrow$ short sleeves/pants, etc.) compared with LPM w/o correlation layer, LPM w/o U-Net, or even w/o LPM. The simulation results are shown in Figure 10 and Figure 11. From simulation results, we can see that the performances of w/o LPM is the worst, since the features of human body and clothes are not well fused. LPM w/o correlation layer and LPM w/o U-Net can achieve relatively better performance. By merit from the advantages of both correlation layer and U-Net structure, the full LPM can achieve the best performance of limb prediction, since the correlation layer is good to keep the correlated information of human body and clothes, while U-Net structure can well preserve the detailed information of original human body by utilizing shallow features.

4.7 User Study

As shown in Table 3 and Table 4, we conduct two user studies on the VITON-Dataset and Zalando-Dataset. Within these two studies, we compare the ACGPN [35] and our proposed method AVTON on the VITON-Dataset in easy, medium, and hard cases, respectively. And we compare the retrained ACGPN and our proposed method AVTON on the Zalando-Dataset in 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 of all the VITON-Dataset based methods.

Specifically, we invite 40 volunteers to complete the experiment. In each study, each volunteer is assigned 50 image pairs and is asked to select the most realistic image out of two virtual try-on results. Both studies show that the AVTON performs better than other methods in all-type clothing and cross-category try-on tasks.
| Target Person and Clothes | TPS | Wendland’s $\Psi$-function |
|---------------------------|-----|---------------------------|
| ![Target Person and Clothes](image1) | ![TPS](image2) | ![Wendland’s $\Psi$-function](image3) |
| ![Target Person and Clothes](image4) | ![TPS](image5) | ![Wendland’s $\Psi$-function](image6) |
| ![Target Person and Clothes](image7) | ![TPS](image8) | ![Wendland’s $\Psi$-function](image9) |
| ![Target Person and Clothes](image10) | ![TPS](image11) | ![Wendland’s $\Psi$-function](image12) |
| ![Target Person and Clothes](image13) | ![TPS](image14) | ![Wendland’s $\Psi$-function](image15) |

Fig. 8. Comparison of image warping between TPS and Wendland’s $\Psi$-function on the VITON dataset.
| Target Person and Clothes | TPS          | Wendland’s $\Psi$-function |
|---------------------------|-------------|----------------------------|
| ![Target Person and Clothes](image1) | ![TPS](image2) | ![Wendland’s $\Psi$-function](image3) |
| ![Target Person and Clothes](image4) | ![TPS](image5) | ![Wendland’s $\Psi$-function](image6) |
| ![Target Person and Clothes](image7) | ![TPS](image8) | ![Wendland’s $\Psi$-function](image9) |
| ![Target Person and Clothes](image10) | ![TPS](image11) | ![Wendland’s $\Psi$-function](image12) |
| ![Target Person and Clothes](image13) | ![TPS](image14) | ![Wendland’s $\Psi$-function](image15) |

Fig. 9. Comparison of image warping between TPS and Wendland’s $\Psi$-function on the Zalando dataset.
| Target Clothes and Person | w/o LPM | LPM w/o Correlation | LPM w/o U-Net | Full LPM |
|---------------------------|---------|---------------------|---------------|---------|
| ![Jacket](image)          | ![Jacket](image) | ![Jacket](image) | ![Jacket](image) | ![Jacket](image) |
| ![Tank Top](image)        | ![Tank Top](image) | ![Tank Top](image) | ![Tank Top](image) | ![Tank Top](image) |
| ![T-shirt](image)         | ![T-shirt](image) | ![T-shirt](image) | ![T-shirt](image) | ![T-shirt](image) |
| ![Sleeveless Top](image)  | ![Sleeveless Top](image) | ![Sleeveless Top](image) | ![Sleeveless Top](image) | ![Sleeveless Top](image) |
| ![Sweater](image)         | ![Sweater](image) | ![Sweater](image) | ![Sweater](image) | ![Sweater](image) |

Fig. 10. Comparison of the limb prediction results with full LPM, w/o LPM, LPM w/o correlation layer, LPM w/o U-Net structure on the VITON-Dataset
| Target Clothes and Person | w/o LPM | LPM w/o Correlation | LPM w/o U-Net | Full LPM |
|--------------------------|---------|---------------------|---------------|---------|
| ![image](image1.png)     | ![image1.png] | ![image1.png]      | ![image1.png] | ![image1.png] |
| ![image2.png]            | ![image2.png] | ![image2.png]      | ![image2.png] | ![image2.png] |
| ![image3.png]            | ![image3.png] | ![image3.png]      | ![image3.png] | ![image3.png] |
| ![image4.png]            | ![image4.png] | ![image4.png]      | ![image4.png] | ![image4.png] |
| ![image5.png]            | ![image5.png] | ![image5.png]      | ![image5.png] | ![image5.png] |
| ![image6.png]            | ![image6.png] | ![image6.png]      | ![image6.png] | ![image6.png] |

Fig. 11. Comparison of the limb prediction results with full LPM, w/o LPM, LPM w/o correlation layer, LPM w/o U-Net structure on the Zalando-Dataset
Table 3. On the VITON-Dataset

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

User study compares ACGPN and our proposed method AVTON at different complexity levels.

Table 4. On the Zalando-Dataset

| Methods      | Mean | Top  | Bottom | Whole |
|--------------|------|------|--------|-------|
| ACGPN        | 36.1%| 32.1%| 40.5%  | 35.6% |
| AVTON (Full) | 63.9%| 67.9%| 59.5%  | 64.4% |

User study compares the retrained ACGPN and our proposed method AVTON with different types of clothes.

Fig. 12. The clothing collocation results in 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.
4.8 Arbitrary Clothing Collocation

In real life, a single clothing virtual try-on cannot meet people’s needs, and clothing collocation can improve people’s preference for virtual try-on. Hence, we conduct an additional experiment to show the results of arbitrary clothing collocation (Figure 12). During the experiment, we first try on tops with our AVTON and got 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 characteristics of the target clothes and the reference person can still be retained after two try-on steps. And benefit from the characteristic tradeoff function of TOFM, the final try-on images are natural and realistic.

5 CONCLUSION

In this article, we collect a new dataset to enhance the robustness and adaptiveness of the model. And we propose a novel virtual try-on network, named AVTON, which aims at handling all-type clothing try-on task (tops, bottoms, and whole clothes) and cross-category try-on task (e.g., long sleeves ↔ short sleeves or long pants ↔ skirts). 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 cross-category try-on task, 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 TOFM can fuse both the information from render person as well as warped clothes to generate a realistic try-on 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. While the proposed work can generally perform well in-shop try-on task, our future work can lie in in-the-wild case with more arbitrary clothing item such as handbag or special dress like frog or bear.

REFERENCES

[1] Xianlin Zhang, Mengling Shen, Xueming Li, and Xiaojie Wang. 2022. AABLSTM: A novel multi-task based CNN-RNN deep model for fashion analysis. ACM Transactions on Multimedia Computing, Communications and Applications 00, 0 (2022), 00 pages. DOI: http://dx.doi.org/10.1145/3519029

[2] Xin Yang, Xuemeng Song, Fuli Feng, Haokun Wen, Ling-Yu Duan, and Liqiang Nie. 2021. Attribute-wise explainable fashion compatibility modeling. ACM Transactions on Multimedia Computing, Communications and Applications 17, 1 (2021), 21 pages. DOI: http://dx.doi.org/10.1145/3425636

[3] Zunlei Feng, Zhenyun Yu, Yongcheng Jing, Sai Wu, Mingli Song, Yezhou Yang, and Junxiao Jiang. 2019. Interpretable partitioned embedding for intelligent multi-item fashion outfit composition. ACM Transactions on Multimedia Computing, Communications and Applications 15, 2s, Article 61 (Jul(2019), 20 pages. DOI: http://dx.doi.org/10.1145/3326332

[4] Feifei Zhang, Mingliang Xu, and Changsheng Xu. 2022. Tell, imagine, and search: End-to-end learning for composing text and image to image retrieval. ACM Transactions on Multimedia Computing, Communications and Applications 18, 2 (2022), 23 pages. DOI: http://dx.doi.org/10.1145/3478642

[5] Gerald Häubl and Valerie Trifts. 2000. Consumer decision making in online shopping environments: The effects of interactive decision aids. Marketing Science 19, 1 (2000), 4–21.

[6] Stefan Hauswiesner, Matthias Straka, and Gerhard Reitmayr. 2013. Virtual try-on through image-based rendering. IEEE Transactions on Visualization and Computer Graphics 19, 9 (2013), 1552–1565.

[7] Shizhan Zhu, Sanja Fidler, Raquel Urtasun, Dahua Lin, and Chen Change Loy. 2017. Be your own prada: Fashion synthesis with structural coherence. In Proceedings of the International Conference on Computer Vision.

[8] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering robust clothes recognition and retrieval with rich annotations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.
[9] Yuying Ge, Ruimao Zhang, Xiaogang Wang, Xiaou Tang, and Ping Luo. 2019. Deepfashion2: A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5337–5345.

[10] Peike Li, Yuning Xu, Yunhao Wei, and Yi Yang. 2020. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 6 (2020), 3260–3271.

[11] Xiaodan Liang, Si Liu, Xiaohui Shen, Jianchao Yang, Luqi Liu, Jian Dong, Liang Lin, and Shuicheng Yan. 2015. Deep human parsing with active template regression. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, 12 (2015), 2402–2414.

[12] Jianshu Li, Jian Zhao, Yunhao Wei, Congyan Lang, Yidong Li, Terence Sim, Shuicheng Yan, and Jiashi Feng. 2017. Multiple-human parsing in the wild. arXiv:1705.07206. Retrieved from https://arxiv.org/abs/1705.07206

[13] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. 2017. 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*. 932–940.

[14] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. 2018. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision*. 589–604.

[15] Tomoharu Iwata, Shinji Wanatabe, and Hiroshi Sawada. 2011. Fashion coordinates recommender system using photographs from fashion magazines. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. Citeseer, 2262.

[16] Mingbo Zhao, Yu Liu, Xianrui Li, Zhao Zhang, and Yue Zhang. 2020. An end-to-end framework for clothing collocation based on semantic feature fusion. *IEEE Multimedia* 27, 4 (2020), 122–132.

[17] Andreas Veit, Balazs Kovacs, Sean Bell, Julian McAuley, Kavita Bala, and Serge Belongie. 2015. Learning visual clothing style with heterogeneous dyadic co-occurrences. In *Proceedings of the IEEE International Conference on Computer Vision*. 4642–4650.

[18] Ruining He, Charles Packer, and Julian McAuley. 2016. Learning compatibility across categories for heterogeneous item recommendation. In *Proceedings of the 2016 IEEE 16th International Conference on Data Mining*. IEEE, 937–942.

[19] Yong-Siang Shih, Kai-Yueh Chang, Hsuan-Tien Lin, and Min Sun. 2018. Compatibility family learning for item recommendation and generation. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*.

[20] Yuncheng Li, Liangliang Cao, Jiang Zhu, and Jiebo Luo. 2017. Mining fashion outfit composition using an end-to-end deep learning approach on set data. *IEEE Transactions on Multimedia* 19, 8 (2017), 1946–1955.

[21] Xintong Han, Zuxuan Wu, Yu-Gang Jiang, and Larry S. Davis. 2017. Learning fashion compatibility with bidirectional lstms. In *Proceedings of the 25th ACM International Conference on Multimedia*. 1078–1086.

[22] Zeyu Cui, Zejun Li, Shu Wu, Xiao-Yu Zhang, and Liang Wang. 2019. Dressing as a whole: Outfit compatibility learning based on node-wise graph neural networks. In *Proceedings of the World Wide Web Conference*. 307–317.

[23] Wei Zeng, Mingbo Zhao, Yuan Gao, and Zhao Zhang. 2020. TileGAN: category-oriented attention-based high-quality tiled clothes generation from dressed person. *Neural Computing and Applications* 32 (2020), 17587–17600.

[24] Jingyuan Liu and Hong Lu. 2018. Deep fashion analysis with feature map upsampling and landmark-driven attention. In *Proceedings of the European Conference on Computer Vision*. 0–0.

[25] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaou Tang. 2016. Fashion landmark detection in the wild. In *Proceedings of the European Conference on Computer Vision*. 3596–3605.

[26] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11307–11317.

[27] Yen-Liang Lin, Son Tran, and Larry S. Davis. 2020. Fashion outfit complementary item retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3311–3319.

[28] Mehrdad Hosseinzadeh and Yang Wang. 2020. Composed query image retrieval using locally bounded features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3596–3605.

[29] Surgan Jandial, Ayush Chopra, Pinkesh Badjatiya, Pranit Chawla, Mausoom Sarkar, and Balaji Krishnamurthy. 2020. TRACE: Transform aggregate and compose visiolinguistic representations for image search with text feedback. arXiv:2009.01485. Retrieved from https://arxiv.org/abs/2009.01485

[30] Wei-Lin Hsiao, Isay Katsman, Chao-Yuan Wu, Devi Parikh, and Kristen Grauman. 2019. Fashion++: Minimal edits for outfit improvement. In *Proceedings of the IEEE International Conference on Computer Vision*. 5047–5056.

[31] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Zhenyu Xie, Bowen Wu, Ziqi Zhang, Xiaohui Shen, and Jian Yin. 2019. Fashion editing with multi-scale attention normalization. arXiv:1906.00884. Retrieved from https://arxiv.org/abs/1906.00884

[32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Communications of the ACM* 63, 11 (2020), 139–144.
[33] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. 2018. Viton: An image-based virtual try-on network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7543–7552.

[34] Ruiyuan Yu, Xiaoli Wang, and Xiaohui Xie. 2019. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In Proceedings of the IEEE International Conference on Computer Vision. 10511–10520.

[35] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. 2020. 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. 7850–7859.

[36] Nilesch Pandey and Andreas Savakis. 2020. Poly-GAN: Multi-conditioned GAN for fashion synthesis. Neurocomputing 414 (2020), 356–364.

[37] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. 2020. Image based virtual try-on network from unpaired data. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 5184–5193.

[38] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. 2018. Swapnet: Image based garment transfer. In Proceedings of the European Conference on Computer Vision. Springer, 679–695.

[39] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J. Black. 2017. ClothCap: Seamless 4D clothing capture and retargeting. ACM Transactions on Graphics 36, 4 (2017), 1–15.

[40] Sangwoo Mo, Minsu Cho, and Jinwoo Shin. 2018. Instagan: Instance-aware image-to-image translation. arXiv:1812.10889. Retrieved from https://arxiv.org/abs/1812.10889

[41] Peng Guan, Loretta Reiss, David A. Hirschberg, Alexander Weiss, and Michael J. Black. 2012. Drape: Dressing any person. ACM Transactions on Graphics 31, 4 (2012), 1–10.

[42] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R. Scott, and Larry S. Davis. 2019. Finet: Compatibile and diverse fashion image inpainting. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 4481–4491.

[43] Holger Wendland. 1995. Piecewise polynomial, positive definite and compactly supported radial functions of minimal degree. Advances in Computational Mathematics 4, 1 (1995), 389–396.

[44] Jean Duchon. 1977. Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In Proceedings of the Constructive Theory of Functions of Several Variables. Springer, 85–100.

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In Proceedings of the International Conference on Medical Image Computing and Computer-assisted Intervention. Springer, 234–241.

[46] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-image translation with conditional adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 1125–1134.

[47] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. arXiv:1710.10196. Retrieved from https://arxiv.org/abs/1710.10196

[48] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4401–4410.

[49] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2357–2366.

[50] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision. 2223–2232.

[51] Wenguan Wang, Hailong Zhu, Jifeng Dai, Yanwei Pang, Jianbing Shen, and Ling Shao. 2020. Hierarchical human parsing with typed part-relation reasoning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 8929–8939.

[52] Wenguan Wang, Zhijie Zhang, Siyuan Qi, Jianbing Shen, Yanwei Pang, and Ling Shao. 2019. Learning compositional neural information fusion for human parsing. In Proceedings of the IEEE/CVF International Conference on Computer Vision. 5703–5713.

[53] Wenguan Wang, Yuanlu Xu, Jianbing Shen, and Song-Chun Zhu. 2018. Attentive fashion grammar network for fashion landmark detection and clothing category classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 4271–4280.

[54] Xianrui Li, Zhiling Ye, Zhao Zhang, and Mingbo Zhao. 2021. Clothes image caption generation with attribute detection and visual attention model. Pattern Recognition Letters 141 (2021), 68–74. DOI: http://dx.doi.org/10.1016/j.patrec.2020.12.001

[55] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 11307–11317.

[56] Thiemo Alldieck, Marcus Magnor, Weipeng Xu, Christian Theobalt, and Gerard Pons-Moll. 2018. Video based reconstruction of 3d people models. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 8387–8397.
[57] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. 2020. Learning to transfer texture from clothing images to 3d humans. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 7023–7034.

[58] Bingwen Hu, Ping Liu, Zhedong Zheng, and Mingwu Ren. 2021. SPG-VTON: Semantic prediction guidance for multi-pose virtual try-on. IEEE Transactions on Multimedia 24 (2021), 1233–1246. Retrieved from https://api.semanticscholar.org/CorpusID:236881063

[59] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 7297–7306.

[60] Mike Fornefett, Karl Rohr, and H. Siegfried Stiehl. 1999. 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). IEEE, 402–407.

[61] Mike Fornefett, Karl Rohr, and H. Siegfried Stiehl. 2001. Radial basis functions with compact support for elastic registration of medical images. Image and Vision Computing 19, 1–2 (2001), 87–96.

[62] B Delaunay, S Vide, A Lamémoire, and V De Georges. 1934. Bulletin de l’Académie des Sciences de l’URSS. Classe Des Sciences Mathématiques et Naturelles 6 (1934), 793–800.

[63] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 8798–8807.

[64] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing 13, 4 (2004), 600–612.

[65] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 586–595.

[66] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. Advances in Neural Information Processing Systems 29 (2016), 2234–2242.

Received 30 March 2022; revised 18 November 2023; accepted 24 November 2023