Virtually Trying on New Clothing with Arbitrary Poses

Na Zheng
Shandong University
zhengnagrape@gmail.com

Xuemeng Song*
Shandong University
sxmustc@gmail.com

Zhaozheng Chen
Shandong University
zhaozhengcc@gmail.com

Linmei Hu
Beijing University of Posts and Telecommunications
hulinmei@bupt.edu.cn

Da Cao
Hunan University
caoda0721@gmail.com

Liqiang Nie*
Shandong University
nieliqiang@gmail.com

ABSTRACT

Thanks to the recent advance in the multimedia techniques, increasing research attention has been paid to the virtual try-on task, especially with the 2D image modeling. The traditional try-on task aims to align the target clothing item naturally to the given person's body and hence present a try-on look of the person. However, in practice, people may also be interested in their try-on looks with different poses. Therefore, in this work, we introduce a new try-on setting, which enables the changes of both the clothing item and the person's pose. Towards this end, we propose a pose-guided virtual try-on scheme based on the generative adversarial networks (GANs) with a bi-stage strategy. In particular, in the first stage, we propose a shape enhanced clothing deformation model for deforming the clothing item, where the user body shape is incorporated as the intermediate guidance. For the second stage, we present an attentive bidirectional GAN, which jointly models the attentive clothing-person alignment and bidirectional generation consistency. For evaluation, we create a large-scale dataset, FashionTryOn, comprising 28, 714 triplets with each consisting of a clothing item image and two model images in different poses. Extensive experiments on FashionTryOn validate the superiority of our model over the state-of-the-art methods.

CCS CONCEPTS

• **Computing methodologies → Artificial intelligence**; **Computer vision**; **Machine learning**.

KEYWORDS

Virtual Try-On System, Person Image Synthesis, Generative Adversarial Networks, Pose Transformation

ACM Reference Format:
Na Zheng, Xuemeng Song, Zhaozheng Chen, Linmei Hu, Da Cao, and Liqiang Nie. 2019. Virtually Trying on New Clothing with Arbitrary Poses. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19), October 21–25, 2019, Nice, France.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3343031.3350946

1 INTRODUCTION

It is reported that the retail sales of T-mall[1] in 2018 Double 11 shopping carnival have exceeded 46.9 billion US dollars, among which fashion apparel and clothing sales contribute 20.3%[2]. The huge economic value of online fashion market demonstrates people's great demand of online fashion shopping. Nevertheless, lacking the physical try-on, online fashion shopping is always criticized for its poor user experience. Owing to the recent advances in computer graphics, there emerge several practical services, such as TriMirror[3] and Fits Me[4], which work on synthesizing the try-on looks for users based on their 3D body shape measurements, desired poses and target clothing items. Despite that 3D-based methods have achieved promising success, the huge labor costs for 3D data annotation and potential economic costs for scanning equipment largely limit their real-world applications.

Fortunately, for more intuitive exhibition, fashion-oriented e-commerce websites, such as Zalando[5], usually display well-posed fashion model images wearing their products as well as the pure product image. In a sense, the tremendous try-on images online have opened the door to the possibility of fulfilling the virtual try-on task with the economic 2D modeling. Although several pioneer researches have achieved promising performance, most of existing efforts can only generate the single-view try-on result, that is, keeping the person's pose unchanged while simply changing the clothing item. However, in reality, people may want to check different views of themselves in the new clothing item before making the decision on whether to buy it or not. In the light of this, in this work, we define a new virtual try-on task, where given a person image, a desired pose, and a target clothing item, we aim to automatically generate the try-on look of the person with the target clothing item in his/her desired pose, as illustrated in Figure 1.

Indeed, advanced image generation models such as Generative Adversarial Networks (GANs) [8] and Variational Autoencoders (VAEs) [15] have demonstrated remarkable success in various image generation tasks. However, it is non-trivial to directly apply these methods to fulfil our proposed task due to the following challenges.

* Xuemeng Song (sxmustc@gmail.com) and Liqiang Nie (nieliqiang@gmail.com) are corresponding authors.

MM '19, October 21–25, 2019, Nice, France

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3350946

[1] https://www.tmall.com/.
[2] http://www.askci.com/news/chanye/20181116/1139281136843.shtml.
[3] https://www.trimirror.com/.
[4] https://fits.me/.
[5] https://zalando.com/.

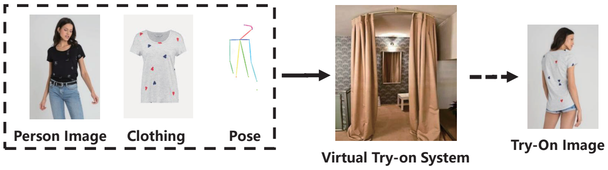

Figure 1: Illustration of our try-on task.

1) In the context of virtual try-on, the body shape and the desired pose of the person highly affect the final look of the target clothing item on the person. Accordingly, how to properly deform the new clothing item and seamlessly align with the target person is a major challenge. 2) How to generate the try-on image that maintains not only the detailed visual features of the clothing item, like the texture and color, but also the other body parts of the person, while changing the person pose is another tough challenge. And 3) there is no large-scale benchmark dataset that can support the research of our new virtual try-on task. Therefore, how to create a large-scale dataset constitutes a crucial challenge.

To address the aforementioned challenges and guarantee the try-on quality, we present a pose-guided virtual try-on scheme with two stages, similar to several state-of-the-art coarse-to-fine pipelines [21, 41]. In particular, in the first stage, we propose a shape enhanced clothing deformation approach working on deforming the given clothing item naturally to match the target body shape of the person, which can be internally predicted from the person's desired pose. In the second stage, our scheme focuses on generating the try-on image based on the deformed clothing item above, the conditional person image and the desired pose. In particular, we present an attentive bidirectional generative adversarial network to synthesize the realistic try-on image, named AB-GAN, which jointly models the attentive clothing-person alignment and bidirectional generation consistency. Pertaining to the evaluation, we create a new large-scale FashionTryOn dataset from the fashion-oriented e-commerce website Zalando[6], consisting of 28,714 triplets.

The main contributions are summarized as follows:

- We present a novel pose-guided virtual try-on scheme in a bi-stage manner. To our best knowledge, we are the first to address the new task of generating realistic try-on images with any desired pose, which has both great theoretical and practical significance.
- We propose a shape enhanced clothing deformation model, which aims to generate the warped clothing item based on both the target body shape and the desired pose. In addition, we present an attentive bidirectional generative adversarial network to synthesize the final try-on images, which simultaneously regularizes the attentive clothing-person alignment and the bidirectional generation consistency.
- We create a large-scale benchmark dataset, FashionTryOn, and extensive experiments conducted on that demonstrate the superiority of our proposed scheme over the state-of-the-art methods. Moreover, we have released the FashionTryOn dataset and codes to benefit other researchers[7].

2 RELATED WORK

This work is related to image synthesis, pose-guided person synthesis and virtual try-on.

Image Synthesis. In the image synthesis domain, GANs have achieved compelling success in various tasks, ranging from the general image generation [27, 35, 36, 38, 41], to the pose-guided person synthesis [1, 5, 6, 21, 22, 26, 29, 32, 39]. In particular, one family of its derivatives, conditional GANs, have been extensively studied recently, especially in image-to-image translation tasks, like the style transfer [3, 4, 17, 23] and virtual try-on [12, 16], where certain conditional images should be given. For example, Isola et al. [11] employed the conditional GAN to fulfil the edge-to-image task, where a UNet-based architecture with skip connections is utilized. In addition, Zhu et al. [40] presented a CycleGAN in the context of style transfer, which focuses on regularizing the cycle consistency and can be trained in an unsupervised manner. Although both methods above have shown remarkable performance, the common assumption they share that the structure of the input is roughly aligned with that of the output is too rigorous and greatly hinders their practical applications in tasks with large misalignments, such as the virtual try-on and pose-guided person synthesis. Towards this end, Siarohit et al. [29] embedded the deformable skip connections to the decoder with affine transformations, while Dong et al. [5] introduced a Soft-Gated Warping-GAN with the geometric matching to alleviate the misalignment. Meanwhile, instead of one-stage models, some researchers explored the coarse-to-fine manner [21, 39], where a coarse intermediate image would be synthesized to help cope with the misalignment. As in our try-on task that involves large input-output misalignment, we follow the coarse-to-fine strategy and decompose our pipeline into two stages: 1) a shape enhanced clothing deformation module is introduced to deform the clothing item naturally, and 2) an attentive bidirectional GAN is proposed to synthesize the final try-on image.

Pose-Guided Person Synthesis. Recently, due to its great value in person re-identification and movie/game making, many efforts have been made on pose-guided person synthesis [1, 5, 6, 21, 22, 26, 29, 31, 32, 39], namely, the person image synthesis with pose transformation. For example, Ma et al. [21] introduced a PG2 network, which works on first synthesizing an intermediate coarse image that captures the global structure and then generating the final image by rendering the appearance details with the adversarial training. Later, to boost the performance, several researchers resorted to separating the human appearance from the image layout directly. For example, Ma et al. [22] presented a BodyROI7 model that particularly disentangles the three essential aspects of the input image, namely the foreground, background, and pose, and then generates the target person image based on the embedding features of all three aspects. In addition, Guha et al. [1] presented a modular generative neural network that decomposes the conditional image into different body parts and based on that synthesizes the target image. Different from the above methods that fulfil the task in a supervised manner, several unsupervised efforts have been made to alleviate the burden of data annotation. In particular, Albert et al. [26] proposed an unsupervised conditional bidirectional generator, where the generated image is able to rendered back

[6] https://zalando.com/.
[7] https://fashiontryon.wixsite.com/fashiontryon.

to the original one. In addition, Song et al. [31] introduced an unsupervised person image synthesis framework, comprising two key components: semantic parsing transformation and appearance generation. Beyond existing efforts, we work on simultaneously change the clothing item and pose of the person rather than simply performing the pose transformation.

Virtual Try-On. Existing virtual try-on systems can be roughly divided into two categories: 3D measurement based models and 2D image based models. As for the 3D based modeling, virtual try-on systems mainly depend on users' 3D body shape measurements. For example, Hauswiesner et al. [10] proposed a 3D based try-on method that allows different viewpoints as 3D measurements are captured by a multi-camera. Besides, Moll et al. [25] introduced a multi-part 3D model, ClothCap, for virtual try-on by jointly capturing the garment geometry on a body and the body shape under the new clothing item. Due to the fact that 3D measurements can be collected costly, some scholars have resorted to the rich 2D images to fulfil the virtual try-on task. For example, Jetchev et al. [12] presented a conditional analogy GAN (CAGAN), which casts the try-on task as an image analogy problem. As a matter of fact, CAGAN overlooks the clothing item deformation according to the user's body shape, and hence suffers from the unsatisfactory try-on performance. To address this issue, several efforts have been dedicated to synthesizing the virtual try-on images with the geometric alignment, such as VITON [9] and CP-VTON [33]. Although existing studies have achieved huge success in virtual try-on domain, they focused on simply changing the clothing item and presenting a single try-on viewpoint, but ignoring the fact that people may want to have different try-on viewpoints with different poses. Differently, in this work, we aim to devise a novel try-on system that can simultaneously change the clothing item and person pose to present a comprehensive try-on effect.

3 METHODOLOGY

In this work, we aim to fulfil the virtual try-on task comprehensively by not only naturally transferring the given clothing item to the corresponding part of the person, but also accurately transforming the person pose to get a novel view of the person. Formally, given a person image I_A, a desired pose P_B[8], and a new clothing item c, our goal is to learn a generator to synthesize the person image I_B in the new clothing item c with the pose P_B. To guarantee the try-on effect and alleviate the burden of the generator, we propose a bi-stage pose-guided virtual try-on GAN network, as illustrated in Figure 2. The first stage focuses on deforming the given clothing item c according to the target pose and the body shape with the *Shape Enhanced Clothing Deformation* (Sec. 3.1), where the target body shape is predicted as an auxiliary guidance. The second stage works on generating the ultimate try-on image based on the warped clothing item $\mathcal{T}_\theta(c)$, the given person image I_A, the predicted body shape $\hat{S}$, and the target pose P_B with the *Attentive Bidirectional GAN* (Sec. 3.2).

3.1 Shape Enhanced Clothing Deformation

In fact, the target clothing item deformation plays a pivotal role in generating the natural try-on image. In a sense, the given clothing item c should be warped according to not only the target pose but also the target human body shape. Due to the fact that the target body shape is not directly available in our context, we first focus on the prediction of the target body shape mask, which acts as an auxiliary guidance on the clothing item deformation.

Body Shape Mask Prediction. In fact, owing to the recent advance of deep neural networks [24], several efforts have been dedicated to the pose-guided parsing, which aims to learn the interplay between the human parsing and human pose. Inspired by [5], we propose to predict the target body shape mask based on the given target pose and the conditional body shape of the given person image.

Let S_A $(S_B) \in \mathbb{R}^{W \times H}$ denote the binary body shape mask of the person image I_A (I_B), where $S_A^{ij} = 1$ $(S_B^{ij} = 1)$ if and only if the (i, j)-th pixel of the corresponding image refers to the person's body part, and $S_A^{ij} = 0$ $(S_B^{ij} = 0)$ otherwise. In addition, we define the network $\mathcal{P}$ for the target body shape mask prediction as follows:

$$\hat{S}_B = \mathcal{P}(S_A, P_B | \Theta_P), \tag{1}$$

where $\hat{S}_B$ represents the predicted body shape mask in line with the target pose P_B and Θ_P refers to the network parameters.

In particular, we devise the body shape mask prediction network $\mathcal{P}$ with the encoder-decoder architecture, where the concatenation of the given body shape mask S_A and desired pose P_B are fed as the input. As the body shape mask prediction can be deemed as a set of binary classification problems, on the top of the decoder, we adopt the cross-entropy loss for each entry as follows:

$$\mathcal{L}_C = \sum_{i=1}^{H} \sum_{j=1}^{W} [-S_B^{ij} \log \hat{S}_B^{ij} - (1 - S_B^{ij}) \log(1 - \hat{S}_B^{ij})]. \tag{2}$$

In addition, to supervise the body shape mask prediction, we incorporate the L1 loss for minimizing the difference between the predicted one and the ground truth:

$$\mathcal{L}_1 = \left\| \hat{S}_B - S_B \right\|_1. \tag{3}$$

Ultimately, we have the following objective function for the target body shape prediction:

$$\mathcal{L}_{BodyShape} = \mathcal{L}_C + \mathcal{L}_1. \tag{4}$$

Clothing Item Deformation. With the guidance of the above predicted target body shape mask, we can perform the target clothing item deformation. Towards this end, we follow the state-of-the-art Geometric Matching Module (GMM) in [33], a variant of CNNGeometric [28], which is able to deform the clothing item naturally by matching it with a cloth-agnostic person representation. In particular, we define the person information p as the concatenation of the predicted target body shape mask $\hat{S}_B$ and the target pose P_B. We feed the person information p and the target clothing item c to two convolutional networks to learn their latent representations, respectively, based on which we can measure the clothing-person alignment scores with a matching layer. Thereafter, according to [33], given the alignment scores, we can obtain the spatial transformation parameters θ by a regression network, and hence can perform the geometric transformation $\mathcal{T}_\theta$ with the thin-plate spline (TPS) layer to derive the warped clothing item $\mathcal{T}_\theta(c)$. Essentially, to minimize the discrepancy between the

[8] We denote the human pose with 18 heatmaps, each of which corresponds to a keypoint of the human body.

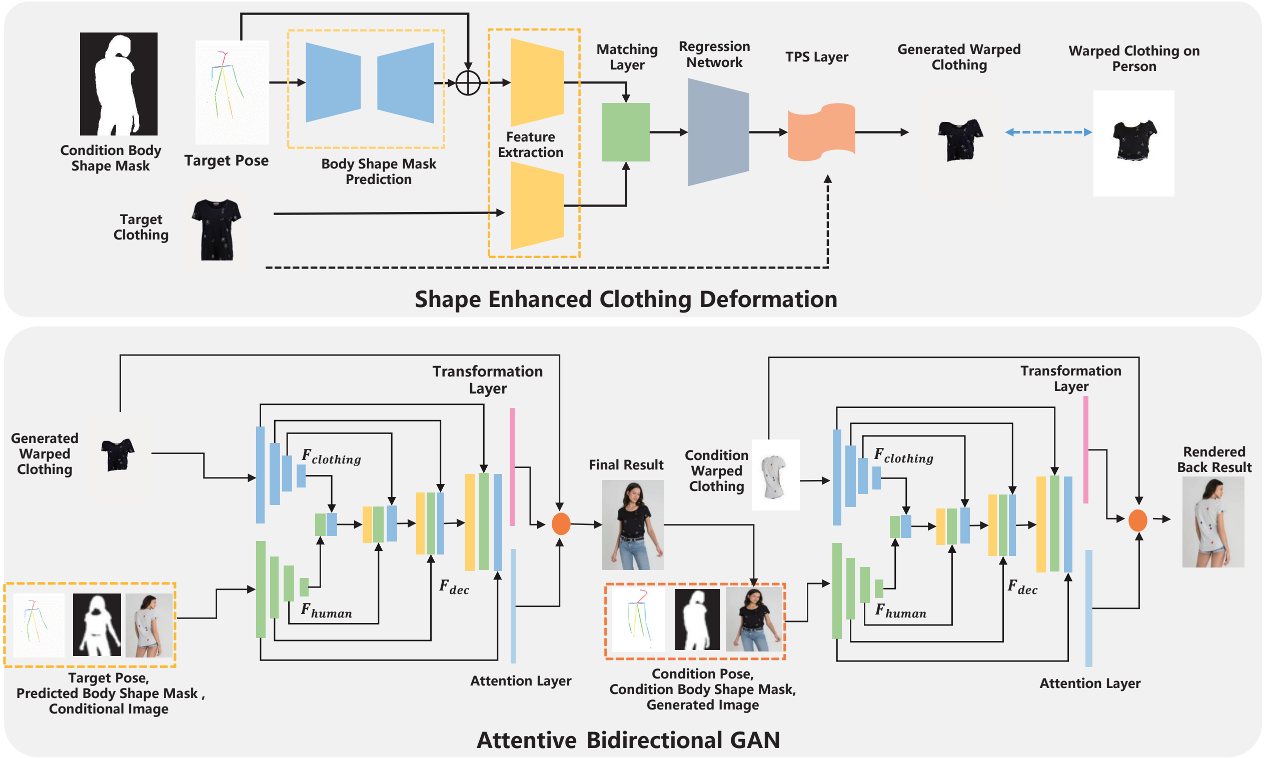
Condition Body Shape Mask
Target Pose
Body Shape Mask Prediction
Feature Extraction
Matching Layer
Regression Network
TPS Layer
Generated Warped Clothing
Warped Clothing on Person
Target Clothing
Shape Enhanced Clothing Deformation
Generated Warped Clothing
$F_{clothing}$
F_{human}
F_{dec}
Transformation Layer
Attention Layer
Target Pose, Predicted Body Shape Mask , Conditional Image
Final Result
Condition Warped Clothing
Condition Pose, Condition Body Shape Mask, Generated Image
Rendered Back Result
Attentive Bidirectional GAN

Figure 2: Illustration of the proposed scheme. Our pipeline is composed of two stages. The Shape Enhanced Clothing Deformation works towards warping the new clothing item naturally to align it with the target body shape while the Attentive Bidirectional GAN aims to synthesize the final try-on image.

warped clothing item $\mathcal{T}_\theta(c)$ and the ground truth $\hat{c}$ that can be effortlessly segmented from the target person image I_B, we adopt the L1 loss at the pixel level:

$$\mathcal{L}_{Clothing} = \|\mathcal{T}_\theta\,(c) - \hat{c}\|_1. \tag{5}$$

Ultimately, the final objective function for shape enhanced clothing deformation is defined as follows:

$$\mathcal{L}_{CD} = \lambda_{BS}\mathcal{L}_{BodyShape} + \lambda_C\mathcal{L}_{Clothing}, \tag{6}$$

where λ_{BS} and λ_C are the trade-off parameters.

3.2 Attentive Bidirectional GAN (AB-GAN)

Having obtained the warped clothing item $\mathcal{T}_\theta(c)$, we can proceed to the presentation of our pose-guided try-on network, AB-GAN, which aims to synthesize the ultimate try-on image with the desired pose and the naturally deformed clothing item. Due to the huge success of GANs in various image generation tasks, we adopt the GAN as the backbone of our pose-guided try-on network. A typical GAN consists of a generator $\mathcal{G}$ and a discriminator $\mathcal{D}$, between whom a min-max strategy game would be performed. The generator $\mathcal{G}$ attempts to fool the discriminator $\mathcal{D}$ by generating realistic images, while the discriminator $\mathcal{D}$ tries to distinguish the synthesized fake images from the real ones.

In our context, we aim to devise a virtual try-on GAN, whose generator $\mathcal{G}$ is able to generate the realistic try-on image $\hat{I}_B$, given the conditional person image $I_A \in \mathbb{R}^{3\times H\times W}$, a desired pose $P_B \in \mathbb{R}^{18\times H\times W}$ the warped new clothing item $\mathcal{T}_\theta(c) \in \mathbb{R}^{3\times H\times W}$ as well as the auxiliary predicted target body shape mask $\hat{S}_B \in \mathbb{R}^{1\times H\times W}$. Formally, we have:

$$\hat{I}_B = \mathcal{G}(I_A, \mathcal{T}_\theta(c), P_B, \hat{S}_B). \tag{7}$$

Try-On Image Generator. To accomplish the above generator, we introduce a human feature encoder F_{human}, a clothing feature encoder $F_{clothing}$ and a unified try-on image decoder F_{dec} to our generator $\mathcal{G}$. In particular, the human feature encoder F_{human} works on embedding the conditional person image I_A and the desired pose P_B and the predicted body shape mask $\hat{S}_B$, while the clothing feature encoder $F_{clothing}$ is designed to extract the key features of the warped clothing item $\mathcal{T}_\theta(c)$. Then the network seamlessly fuses the human features and clothing features by the dilation-based bottleneck, which has been proved to be effective in image inpainting [37]. Thereafter, the fused features would be decoded to the target person image $\hat{I}_B$ by the try-on image decoder F_{dec}. For each encoder, we adopt the UNet network [11] with skip connections. In a sense, the skip connections between F_{human} and F_{dec} serve to propagate the human appearance and the desired pose, while those between $F_{clothing}$ and F_{dec} work on transferring the features of desired clothing item.

In order to push the try-on network to pay more attention to the (target) try-on area and achieve the natural alignment of the warped clothing item, we adopt the attention mechanism, which has shown great success in various computer vision tasks [4, 18, 19]. In a sense, the attention mechanism serves to help the generator to focus more on the region that the warped clothing item should be aligned.

Towards this end, we introduce a transformation layer L_I that aims to synthesize a rough target person image $\tilde{I}_B \in \mathbb{R}^{3\times H\times W}$ as the template, and an attention layer L_A for generating the attention mask $\mathbf{A} \in \mathbb{R}^{1\times H\times W}$ with the same shape of the person image. Accordingly, we can generate the target person image $\hat{I}_B$ as follows:

$$\hat{I}_B = \tilde{I}_B \odot (\mathbf{1} - \mathbf{A}) + \mathcal{T}_\theta(c) \odot \mathbf{A}, \tag{8}$$

where $\odot$ denotes element-wise matrix multiplication. In a sense, the attention mask $\mathbf{A}$ weighs the relative importance of the rough person image $\tilde{I}_B$ and the warped clothing item $\mathcal{T}_\theta(c)$ in the final try-on image generation.

Try-on Network Optimization. As aforementioned, we expect higher attention scores can be given to the try-on region while lower scores to elsewhere. Accordingly, we adopt the L1 loss that is widely used in image generation tasks as follows:

$$\mathcal{L}_{Atten} = \mathbb{E}_{I_A}\left[\|\mathbf{M} - \mathbf{A}\|_1\right] + \lambda_{TV}\mathbb{E}_{I_A}\left[\|\nabla\mathbf{A}\|_2\right], \tag{9}$$

where $\mathbf{M}$ denotes the binary ground truth try-on region mask that can be easily segmented from the target person image I_B. Notably, the second term refers to the total variation (TV) regularization [9], which is introduced to penalize the gradients of the attention mask $\mathbf{A}$ and ensure the spatial smoothness. λ_{TV} is the trade-off non-negative hyperparameter.

According to the standard GAN loss that regularizes the generated $\hat{I}_B$ and the ground truth person image I_B to come from the same domain, we have:

$$L_{GAN_B} = \mathbb{E}_{I_B\sim\mathbb{P}}\left[\log \mathcal{D}\left(I_B\right)\right] + \mathbb{E}_{\hat{I}_B\sim\mathbb{P}}\left[\log\left(1 - \mathcal{D}(\hat{I}_B)\right)\right], \tag{10}$$

where $\mathbb{P}$ denotes the data distribution. Apparently, the simple GAN loss is insufficient to fulfil our task, as a generated person image with either undesired pose or clothing item cannot meet our initial requirement no matter how realistic it is. Towards this end, we adopt the L1 regularization to penalize the loss from the content perspective. To better characterize the content feature of each image, we resort to the perceptual features, like the edge, color and texture features extracted by the pre-trained VGG19 model [30]. In particular, we comprehensively measure the content loss from both the pixel and perceptual levels as follows:

$$\mathcal{L}_{Con_B} = \sum_{i=1}^{5} \lambda_i \left\|\varphi_i\left(\hat{I}_B\right) - \varphi_i\left(I_B\right)\right\|_1 + \left\|\hat{I}_B - I_B\right\|_1, \tag{11}$$

where $\varphi_i\left(I_B\right)$ stands for the feature map of image I regarding the i-th layer of the pre-trained VGG19. Here, the i-th layer refers to the layer "conv(i_2)" of VGG19, which is in line with [9, 33].

Inspired by [40], we expect the generator $\mathcal{G}$ to be capable of not only synthesizing a realistic try-on image conditioned on the given person image, the new clothing item, the desired pose and the predicted body shape mask, but also rendering back to the original person image conditioned on original constraints (i.e., the original pose, original clothing item and original body shape mask). Accordingly, given the generated try-on image $\hat{I}_B$, the original pose P_A and the original clothing item $\hat{c}$, which can be directly extracted from I_A, the generator $\mathcal{G}$ should be able to synthesize a original person image $\hat{I}_A$. It is worth noting that the rendered back person image $\hat{I}_A$ and the original person image I_A should keep the same data distribution. As a result, to regularize the bidirectional generation consistency, we have:

$$\mathcal{L}_{GAN_A} = \mathbb{E}_{I_A\sim\mathbb{P}}\left[\log D\left(I_A\right)\right] + \mathbb{E}_{\hat{I}_A\sim\mathbb{P}}\left[\log\left(1 - D(\hat{I}_A)\right)\right], \tag{12}$$

and

$$\mathcal{L}_{Con_A} = \sum_{i=1}^{5} \lambda_i \left\|\varphi_i\left(\hat{I}_A\right) - \varphi_i\left(I_A\right)\right\|_1 + \left\|\hat{I}_A - I_A\right\|_1. \tag{13}$$

Ultimately, our pose-guided try-on loss can be written as:

$$\begin{aligned}\mathcal{L} =& \lambda_{GAN_B}\mathcal{L}_{GAN_B} + \lambda_{GAN_A}\mathcal{L}_{GAN_A} + \lambda_{Atten}\mathcal{L}_{Atten} + \\ & \lambda_{Con_B}\mathcal{L}_{Con_B} + \lambda_{Con_A}\mathcal{L}_{Con_A},\end{aligned} \tag{14}$$

where λ_{GAN_B}, λ_{GAN_A}, λ_{Atten}, λ_{Con_B} and λ_{Con_A} are the hyperparameters that control the relative importance of each loss item.

4 DATASET

Although several fashion datasets have been constructed for the virtual try-on related tasks, such as the DeepFashion [20], and Zalando Dataset [9], they can only support the conventional try-on tasks. For example, DeepFashion contains numerous person images with different poses, which can only facilitate the pose transformation research; whereas in our scenario, we need the ground truth image with both the target clothing item and the target pose. In fact, it is intractable to build an ideal training dataset of triplets, each of which comprises a conditional person image, a new clothing item image and the target person image wearing the new clothing item with the target pose. Fortunately, inspired by VITON [9], we can train our model with the slightly modified triplet, where the target person image share the same clothing item with the conditional person image (i.e., the new clothing item is exactly the original one). Accordingly, to create our virtual try-on dataset, we turn to the fashion-oriented e-commercial website Zalando[9], where each clothing item is associated with multiple fashion model images with different poses wearing it. In total, we crawled 4, 327 clothing items with their corresponding model images.

To ensure the quality of our dataset, we removed the noisy model images that only show limited part of, even without, the human body. In particular, we extracted the keypoints of each model image by the state-of-the-art pose estimator [2]. Based on the presence of these keypoints, we removed the images with less than 10 keypoints, as they indicate the limited human body parts. Ultimately, we obtained our FashionTryOn dataset, consisting of 28, 714 triplets with each comprising one clothing item image and two model images in different poses, corresponding to the original (conditional) person image and the target person image. Notably, all images are resized to 256 × 192.

5 EXPERIMENTS

To evaluate the proposed method, we conducted extensive experiments on the real-world dataset FashionTryOn by answering the following research questions:

- Does AB-GAN outperform the state-of-the-art methods?
- What is the performance of our shape enhanced clothing deformation?

[9] www.zalando.com/.

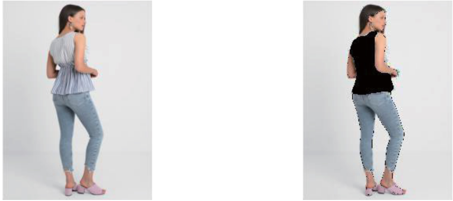

Figure 3: The left is the original person image, while the right is the same image overlaid with a clothing mask.

- How do the key components of our pipeline affect the final try-on image synthesis?

5.1 Implementation Details

Here we give the implementation details of our work.

Pose Embedding. Similar to [9], we first extracted the pose of each person image with the state-of-the-art pose estimator [2], where coordinates of 18 human body keypoints can be obtained. Based on these coordinates, we derived 18 binary heatmaps for the keypoints, respectively. Each heatmap is filled with all zeros, except for the ones in the neighborhood area of 11 × 11 around each keypoint. Afterwards, all heatmaps are stacked into 18 channels, embedding the pose information of the person image.

Network Architecture. For the body shape mask prediction module in the shape enhanced clothing deformation, we adopted the ResNet-like architecture, as it has shown remarkable performance in various generation tasks [5, 40]. In particular, we designed the encoder with one 1-strided convolutional layer and three 2-strided ones followed by nine residual blocks, while the decoder with three 2-strided deconvolutional layers as well as two 1-strided ones. All convolutional layers are followed by the Instance Normalization and Relu activation functions, except for the last layer which takes the softmax activation function. As for the clothing deformation module, following the network structure of the GMM in [33], we devised both the person and clothing item feature encoder networks with four 2-strided convolutional layers and two 1-strided ones. Then we deployed the regression network with two 2-strided convolutional layers, two 1-strided ones and one fully-connected layer. The structure of the matching layer and TPS layer are similar to those in CNNGeometric [28].

Pertaining to the pose-guided try-on image generation, both the human feature encoder and the clothing feature encoder have four down-sampling convolutional layers, whose numbers of 3 × 3 filters are 64, 128, 256 and 512, respectively. The dilation-based bottleneck consists of three dilation blocks with the dilation rates of 1, 2 and 4, respectively. All three blocks have 512 filters of the size 3 × 3. The architecture of the decoder is symmetric to that of the human/clothing feature encoder. Regarding the discriminator, we adopted the same architecture as pixel2pixel [11].

Optimization. We split our dataset into two trucks: the training set with 21, 197 triples and the testing set with 7, 517 triples. In all experiments, we used the Adam [14] optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and a fixed learning rate of 0.0001. For training the shape enhanced clothing deformation, to boost the performance, we pre-trained our body shape mask prediction module for $100K$ steps and clothing deformation module for $200K$ steps, respectively. Then we fine-tuned the shape enhanced clothing deformation with $\lambda_{BS} = 1$ and $\lambda_C = 1$. To avoid the artifacts caused by the conflict of the body shape and the clothing item, similar to [9], we first down-sampled the predicted body shape mask to 16 × 12 and then reconstructed it back to 256 × 192. For the try-on image generation, we trained our attentive bidirectional GAN for $200K$ steps, where $\lambda_{GAN_B} = 0.1$, $\lambda_{GAN_A} = 0.1$, $\lambda_{Atten} = 1$, $\lambda_{Con_B} = 5$, $\lambda_{Con_A} = 5$, and $\lambda_{TV} = 0.0001$. The batch size of all experiments are set as 4.

Table 1: Performance of different methods.

Method	SSIM ↑	L1 Error ↓	VGG Error↓
UNet	0.7032	0.2064	0.6588
CP-VTON	0.7040	0.1937	0.6193
Vari-UNet+CAGAN	0.6055	0.2473	0.8506
CA-GAN+Vari-UNet	0.7580	0.1580	**0.5119**
Vari-UNet+CP-VTON	0.7335	0.1963	0.5719
CP-VTON+Vari-UNet	**0.7582**	0.1574	0.5203
Ours	0.7541	**0.1230**	0.5272

5.2 Baselines

Due to the fact of lacking proper baselines for our novel task, we introduced six baselines for the performance comparison: two pipelines to simultaneously transform the pose of the person and overlay the desired clothing item, two pipelines that first transform the pose then change the clothing item, and two pipelines that first transfer the clothing item onto the person and then synthesize the final try-on image with the desired pose.

UNet. We employed the UNet in [11] as one baseline. In particular, we adapted to take the concatenation of the predicted body shape mask, the desired pose, the warped clothing item and the 3-channel head segmentation[10] as the input and employed L1 loss and VGG loss for optimization.

CP-VTON. As CP-VTON [33] is originally designed to tackle the simple try-on task of overlaying a new clothing item on the person, for fairness, we adapted it by feeding the same concatenation as UNet for this baseline.

Vari-UNet+CP-VTON. In this baseline, we first utilized the Vari-UNet [6] for the pose transformation, and then applied the CAGAN [12] to change the clothing item.

Vari-UNet+CAGAN. Similarly, here we first performed the pose transformation by Vari-UNet, followed by the clothing item replacement by CAGAN.

CP-VTON+Vari-UNet. In the similar manner, we changed the order of Vari-UNet and CP-VTON in Vari-UNet+CP-VTON to replace the clothing item before changing the person's pose.

CAGAN+Vari-UNet. Similarly, we changed the order of Vari-UNet and CAGAN in Vari-UNet+CAGAN.

5.3 On Model Comparison (RQ1)

To get a thorough understanding of our model, we performed both the quantitative and qualitative evaluations.

Quantitative Evaluation. Following [1], we adopted the Structure Similarly Index Measure (SSIM) [34], L1 error and VGG error [13] to measure the quality of synthesized images by different models. In particular, SSIM measures the similarity between the synthesized image and the ground truth image based on the three

[10] The head segmentation is obtained with the help of the state-of-the-art LIP human parser [7].

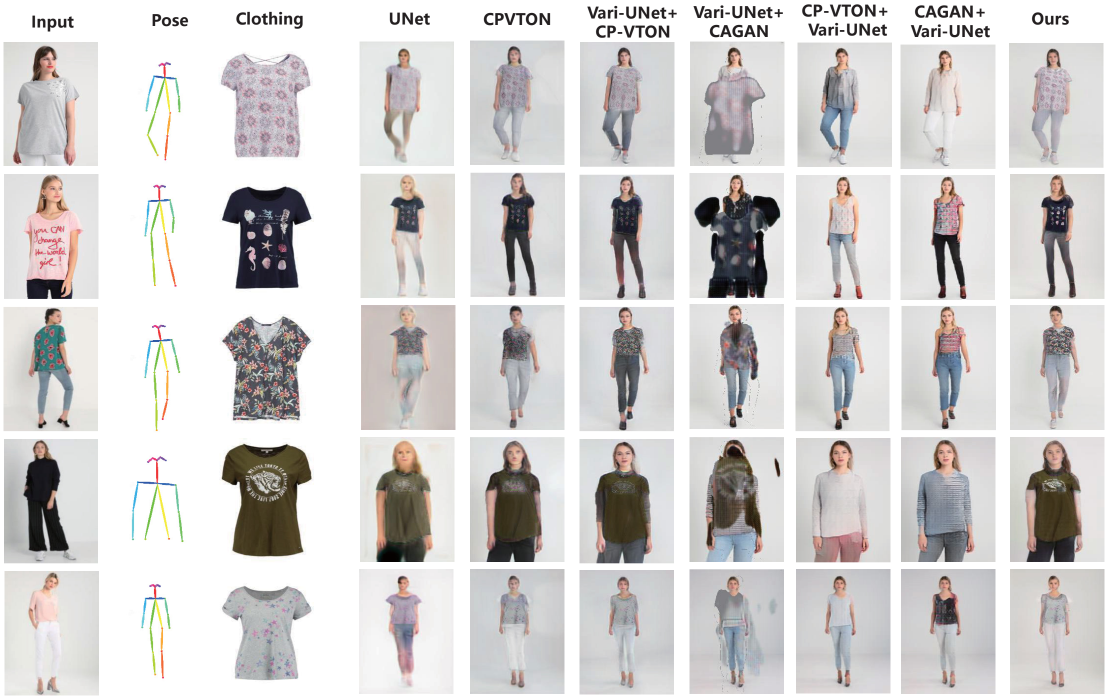

Figure 4: Try-on results of different methods.

aspects of brightness, structure and contrast, ranging from 0 to 1. Both the L1 error and VGG error assess the content difference between the synthesized image and the ground truth image.

As aforementioned, the conditional and target person images in each triplet of our dataset share the same clothing item. Therefore, for the quantitative evaluation, we overlaid a clothing mask of all zeros on the try-on region of the testing conditional person image, illustrated in Figure 3. To be specific, we utilized the LIP human parser [7] to obtain the try-on region. Table 1 shows the performance of different methods with respect to SSIM, L1 error and VGG error. As can be seen, our model consistently outperforms all the baselines in terms of the L1 error and achieves comparable results regarding the SSIM and the VGG error. The philosophy behind may be that our method adopts two encoders for extracting the human features and clothing features, respectively, rather than the single encoder that is used by baselines, like CP-VTON and Vari-UNet+CP-VTON. Moreover, we incorporated both the content loss and bidirectional generation consistency that can regularize the generator to render more sharp and realistic images with less content loss. In addition, we observed that CAGAN+Vari-UNet and CP-VTON+Vari-UNet surpass Vari-UNet+CAGAN and Vari-UNet+CP-VTON, respectively, suggesting that it is advisable to change the clothing item before transforming the human pose. The plausible explanation is that the prior pose transformation can make the following clothing try-on more difficult by introducing misalignments between the human pose and the clothing item. Meanwhile, the pose transformation pipeline Vari-UNet that disentangles the human shape and appearance can be insensitive to the appearance noises caused by the simple clothing try-on pipelines and contributes the better performance of CAGAN+Vari-UNet and CP-VTON+Vari-UNet.

Qualitative Evaluation. Towards the qualitative evaluation, we first shuffled the clothing item images in the testing dataset to ensure the clothing item of the target person image is different from that of the conditional person image. Figure 4 shows the try-on results of different methods. From this figure, we can draw the following observations: 1) our method achieves the best visual try-on effects as compared to all the baselines. The synthesized person image of our method not only meets the desired pose but also realizes the natural change of clothing item. 2) Although the stacked pipelines (e.g., CAGAN+Vari-UNet and CPVTON+Vari-UNet) obtain promising scores pertaining to the quantitative evaluation, the try-on results are unsatisfactory as the detailed visual features, like the pattern and color, of the target clothing item cannot be well preserved, which is improper for our try-on task. 3) CP-VTON+Vari-UNet can be seen as the best baseline while Vari-UNet+CAGAN the worst. The strange try-on results (see the 2-nd and 4-th cases) of Vari-UNet+CAGAN may be attributed to the lack of effective clothing deformation in CAGAN. 4) Interestingly, although Vari-UNet+CPVTON performs worse than CPVTON+Vari-UNet in the quantitative evaluation, Vari-UNet+CPVTON surpasses CPVTON+Vari-UNet in terms of the qualitative try-on effect assessment, as more clothing details are successfully transferred by Vari-UNet+CPVTON. This reconfirms the importance of the order in performing the clothing item try-on and the pose transformation.

5.4 On Clothing Deformation (RQ2)

To evaluate our shape enhanced clothing deformation (SECD) model, we compared it with the GMM in CP-VTON [33]. Figure 5 shows the qualitative evaluation of our SECD and GMM. We noticed that SECD outperforms GMM in terms of both the clothing

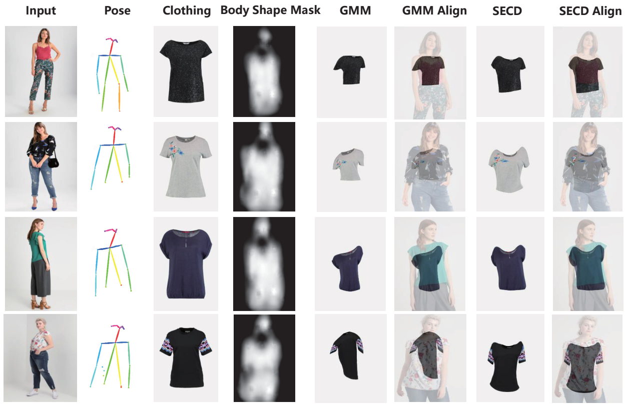

Figure 5: Matching results of SECD and GMM.

deformation quality and the clothing alignment with the target person. As can be seen, GMM can lead to the abnormal deformations (see the 4-th case) and the misalignments of the clothing item (see the 2-nd case). The underlying philosophy is that the GMM learns the clothing mapping directly based on the conditional body shape mask and the desired human pose, where the potential of the target human body shape in the guidance of the clothing deformation is overlooked. On the contrary, our method takes the body shape mask prediction as an intermediate step to enhance the performenace. As we can see, the predicted body shape masks can roughly capture the target body shapes, and hence promote the performance on clothing alignment. This suggests that it is advisable to incorporate the target body shape mask as an auxiliary guidance towards the clothing deformation in the complex virtual try-on task.

5.5 On Ablation Study (RQ3)

To get a better understand of our pipeline, we conducted the ablation study on three key components of our pipeline: target body shape prediction, the attention mechanism and the bidirectional generation consistency. Accordingly, we introduced three ablation baseline: **w/o body shape**, **w/o attention** and **w/o consistency**, which can be derived from our pipeline by disabling the corresponding module. Table 2 shows the quantitative experiment results of different ablation methods. As can be seen, our pipeline shows superiority over each w/o ablation method, verifying the importance of the predicted body shape mask in guiding the clothing deformation, and the attention mechanism as well as the bidirectional generation consistency in the final try-on image generation. Figure 6 shows the qualitative results of different pipelines. For the better illustration, we also listed the learned attention masks of our full pipeline. As can be seen, the qualitative results are consistent with the quantitative ones: 1) our full pipeline surpasses the w/o body shape pipeline, as the latter fails to align the clothing item accurately with the human's body, implying that the predicted body shape mask does provide the auxiliary guidance for the clothing alignment. 2) Our full pipeline synthesizes more natural person images than the w/o consistency pipeline. One plausible explanation is that the bidirectional generation consistency can propel the generator to synthesize realistic person images while ignoring noises caused by the abnormal clothing deformations, and ultimately boosts the robustness of the generator. And 3) our pipeline outperforms the w/o attention pipeline in terms of the detailed feature (e.g., the pattern) preservation of the clothing item. The underlying philosophy is that the generator without the guidance of the attention mask can fail to focus on the try-on region, and hence missing the texture details of the clothing item.

Table 2: Performance of different ablation methods.

Method	SSIM ↑	L1 Error ↓	VGG Error↓
w/o Attention	0.7343	0.1466	0.5652
w/o Consistency	0.7374	0.1457	0.5656
w/o Body Shape	0.7118	0.1642	0.5995
Ours	**0.7541**	**0.1230**	**0.5272**

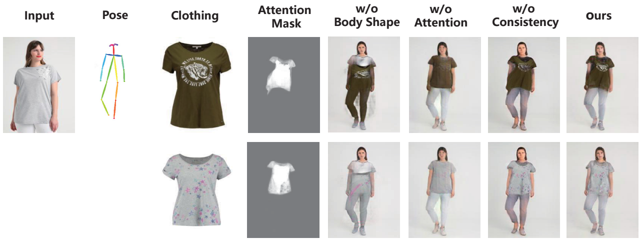

Figure 6: Results of our pipeline and the ablation methods.

6 CONCLUSION AND FUTURE WORK

In this paper, we presented a novel bi-stage pose-guided virtual try-on pipeline to address a new task of virtually trying on a new clothing item with the desired pose. In particular, in the first stage, we proposed a shape enhanced clothing deformation scheme to accurately warp the given new clothing item, where a target human body shape mask prediction module is introduced to provide the intermediate guidance for the clothing deformation. In the second stage, we proposed an attentive bidirectional generation adversarial network, AB-GAN, to generate the final try-on person image, which jointly regularizes the attentive clothing-person alignment and bidirectional generation consistency. Moreover, we constructed a large-scale dataset, FashionTryOn, comprising 28, 714 clothing-person-person triplets. Extensive experiments have been conducted on FashionTryOn, and the qualitative and quantitative results show the superiority of our pipeline over the state-of-the-art methods. Interestingly, we found that introducing the target body shape mask as an guidance in clothing deformation dose help to boost the performance. In the future, we plan to enhance our try-on scheme to further cope with the clothing matching problem, where a try-on image in a complete and compatible outfit would be synthesized.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China, No.: 61772310, No.:61702300, No.:61702302, No.: 61802231, and No. U1836216; the Project of Thousand Youth Talents 2016; the Shandong Provincial Natural Science and Foundation, No.: ZR2019JQ23, No.:ZR2019QF001; the Future Talents Research Funds of Shandong University, No.: 2018WLJH 63.

REFERENCES

[1] Guha Balakrishnan, Amy Zhao, Adrian V Dalca, Fredo Durand, and John Guttag. 2018. Synthesizing images of humans in unseen poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 8340–8348.

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 7291–7299.

[3] Huiwen Chang, Jingwan Lu, Fisher Yu, and Adam Finkelstein. 2018. Pairedcyclegan: Asymmetric style transfer for applying and removing makeup. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 40–48.

[4] Xinyuan Chen, Chang Xu, Xiaokang Yang, and Dacheng Tao. 2018. Attention-GAN for object transfiguration in wild images. In *Proceedings of the European Conference on Computer Vision*. Springer, 164–180.

[5] Haoye Dong, Xiaodan Liang, Ke Gong, Hanjiang Lai, Jia Zhu, and Jian Yin. 2018. Soft-gated warping-GAN for pose-guided person image synthesis. In *Advances in Neural Information Processing Systems*. MIT Press, 472–482.

[6] Patrick Esser, Ekaterina Sutter, and Björn Ommer. 2018. A variational u-net for conditional appearance and shape generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 8857–8866.

[7] 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*. IEEE, 932–940.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in Neural Information Processing Systems*. MIT Press, 2672–2680.

[9] 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*. IEEE, 7543–7552.

[10] Stefan Hauswiesner, Matthias Straka, and Gerhard Reitmayr. 2011. Free viewpoint virtual try-on with commodity depth cameras. In *Proceedings of the International Conference on Virtual Reality Continuum and Its Applications in Industry*. ACM, 23–30.

[11] 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*. IEEE, 1125–1134.

[12] Nikolay Jetchev and Urs Bergmann. 2017. The conditional analogy gan: Swapping fashion articles on people images. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 2287–2292.

[13] Justin Johnson, Alexandre Alahi, and Fei Fei Li. 2016. Perceptual losses for real-time style transfer and super-resolution. In *Proceedings of the European Conference on Computer Vision*. Springer, 694–711.

[14] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[15] Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (2013).

[16] Zorah Lahner, Daniel Cremers, and Tony Tung. 2018. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European Conference on Computer Vision*. Springer, 667–684.

[17] Mingyu Liu, Thomas Breuel, and Jan Kautz. 2017. Unsupervised image-to-image translation networks. In *Advances in Neural Information Processing Systems*. MIT Press, 700–708.

[18] Meng Liu, Xiang Wang, Liqiang Nie, Xiangnan He, Baoquan Chen, and Tat-Seng Chua. 2018. Attentive Moment Retrieval in Videos. In *In Proceedings of the International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. ACM, 15–24.

[19] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-modal Moment Localization in Videos. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 843–851.

[20] 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 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1096–1104.

[21] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. 2017. Pose guided person image generation. In *Advances in Neural Information Processing Systems*. MIT Press, 406–416.

[22] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Disentangled person image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 99–108.

[23] Youssef Alami Mejjati, Christian Richardt, James Tompkin, Darren Cosker, and Kwang In Kim. 2018. Unsupervised attention-guided image-to-image translation. In *Advances in Neural Information Processing Systems*. MIT Press, 3697–3707.

[24] Liqiang Nie, Xuemeng Song, and Tat-Seng Chua. 2016. Learning from multiple social networks. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 8, 2 (2016), 1–118.

[25] 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), 73.

[26] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Unsupervised person image synthesis in arbitrary poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 8620–8628.

[27] Scott E Reed, Zeynep Akata, Santosh Mohan, Samuel Tenka, Bernt Schiele, and Honglak Lee. 2016. Learning what and where to draw. In *Advances in Neural Information Processing Systems*. MIT Press, 217–225.

[28] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. 2017. Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 6148–6157.

[29] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. 2018. Deformable gans for pose-based human image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 3408–3416.

[30] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[31] Sijie Song, Wei Zhang, Jiaying Liu, and Tao Mei. 2019. Unsupervised person image generation with semantic parsing transformation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2357–2366.

[32] Hao Tang, Wei Wang, Dan Xu, Yan Yan, and Nicu Sebe. 2018. Gesturegan for hand gesture-to-gesture translation in the wild. *arXiv preprint arXiv:1808.04859* (2018).

[33] 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*. Springer, 589–604.

[34] Z Wang, Alan Bovik, H.R. Sheikh, and Eero Simoncelli. 2004. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612.

[35] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 1316–1324.

[36] Guojun Yin, Bin Liu, Lu Sheng, Nenghai Yu, Xiaogang Wang, and Jing Shao. 2019. Semantics disentangling for text-to-image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2327–2336.

[37] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S. Huang. 2018. generative image inpainting with contextual attention. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 5505–5514.

[38] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. 2017. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 5907–5915.

[39] Bo Zhao, Xiao Wu, Zhiqi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. 2018. Multi-view image generation from a single-view. In *ACM International Conference on Multimedia*. ACM, 383–391.

[40] Junyan 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*. IEEE, 2223–2232.

[41] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. 2017. Be your own prada: Fashion synthesis with structural coherence. In *Proceedings of the IEEE International Conference on Computer Vision*. IEEE, 1680–1688.