FashionOn: Semantic-guided Image-based Virtual Try-on with Detailed Human and Clothing Information

Chia-Wei Hsieh*
National Chiao Tung University
Hsinchu, Taiwan
maggie1209.tem04@nctu.edu.tw

Chieh-Yun Chen*
National Chiao Tung University
Hsinchu, Taiwan
sky4568520.ep05@nctu.edu.tw

Chien-Lung Chou*
National Chiao Tung University
Hsinchu, Taiwan
chienlung.eed04@nctu.edu.tw

Hong-Han Shuai
National Chiao Tung University
Hsinchu, Taiwan
hhshuai@nctu.edu.tw

Jiaying Liu
Peking University
Beijing, China
liujiaying@pku.edu.cn

Wen-Huang Cheng
National Chiao Tung University
Hsinchu, Taiwan
whcheng@nctu.edu.tw

ABSTRACT

The image-based virtual try-on system has attracted a lot of research attention. The virtual try-on task is challenging since synthesizing try-on images involves the estimation of 3D transformation from 2D images, which is an ill-posed problem. Therefore, most of the previous virtual try-on systems cannot solve difficult cases, e.g., body occlusions, wrinkles of clothes, and details of the hair. Moreover, the existing systems require the users to upload the image for the target pose, which is not user-friendly. In this paper, we aim to resolve the above challenges by proposing a novel FashionOn network to synthesize user images fitting different clothes in arbitrary poses to provide comprehensive information about how suitable the clothes are. Specifically, given a user image, an in-shop clothing image, and a target pose (can be arbitrarily manipulated by joint points), FashionOn learns to synthesize the try-on images by three important stages: pose-guided parsing translation, segmentation region coloring, and salient region refinement. Extensive experiments demonstrate that FashionOn maintains the details of clothing information (e.g., logo, pleat, lace), as well as resolves the body occlusion problem, and thus achieves the state-of-the-art virtual try-on performance both qualitatively and quantitatively.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; **Computer vision tasks**; **Image representations**; *Texturing*; Image segmentation; Shape inference.

KEYWORDS

Virtual try-on; image synthesis; pose transformation; semantic-guided learning

*Authors contributed equally to this research.

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

ACM ISBN 978-1-4503-6889-6/19/10…$15.00
https://doi.org/10.1145/3343031.3351075

ACM Reference Format:
Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. 2019. FashionOn: Semantic-guided Image-based Virtual Try-on with Detailed Human and Clothing Information. 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/10.1145/3343031.3351075

Figure 1: An example of virtual try-on with arbitrary poses.

1 INTRODUCTION

"Style is something each of us already has, all we need to do is find it."
— Diane von Furstenberg *(1946-)*

Everyone has their own style, but the process of finding it is challenging and time-consuming. For example, when users purchase fashion items online, they can quickly explore as many items as they want but may worry about the inconsistency of the looks between them and models. As such, users are more conservative in buying new fashion items online. On the other hand, users can try on various fashion items in brick-and-mortar stores, which helps them gain multi-aspect information of suitability, but it may require a lot of time to get to the stores and try on fashion items.

Among many promising approaches [6, 16, 17] bridging the gap between online and offline shopping, virtual try-on service is the most concerning task. Therefore, a recent line of studies attempts to realize the virtual try-on by utilizing clothing warping [15, 42], which successfully preserves the details of garments including logos, patterns, and decorative designs. However, when the occlusion (e.g., the users' arms cross over the chest and occlude the clothing in source images) or dramatic posture transformation (e.g., limbs from non-overlapping to overlapping) happen, the quality of the results decreases significantly. On the other hand, when users attempt to find their try-on images from different view angles or postures,

current approaches require users to upload their pictures with corresponding view angles or poses. To provide a user-friendly virtual try-on, a virtual try-on that can transform arbitrary poses by a single image is desirable.

Based on these observations, we propose a novel virtual try-on network, namely, FashionOn network, for synthesizing high-quality try-on images with arbitrary poses. As illustrated in Fig. 1, given a source user image, an in-shop garment for try-on, and a target pose only with keypoints[1], the goal is to synthesize a realistic try-on image for the user. To solve the issues of the occlusion or dramatic posture transformation based on clothing warping, one basic approach is to provide the body parsing information to the clothing warping frameworks. Nevertheless, this intuitive approach cannot solve the occlusion problems because of directly pasting the warped clothing on the target person. Therefore, instead of adopting clothing warping-based methods, the proposed FashionOn network first utilizes human segmentation information to synthesize the try-on segmentation. Afterward, a conditional GAN and a refinement network are consecutively exploited to generate the try-on images based on try-on segmentation and preserve the details.

Specifically, FashionOn network is comprised of three stages: pose-guided parsing translator, segmentation region coloring, and salient region refinement. In pose-guided parsing translator, inspired by [34], we first construct a deep neural network to transform the posture into the semantic segmentation form based on the source human parsing and the target pose keypoints for specifying the information of limbs such as location and size to guide the learning of the next stage. Afterward, in segmentation region coloring, we adopt cGAN to render the appearance information of the human and clothing to fill with the semantic segmentation result from the previous stage to generate a realistic human image. Finally, in the last stage, salient region refinement focuses on ameliorating two regions, i.e., face and clothing, with detailed information for achieving more realistic virtual try-on results.

To demonstrate the efficacy of the proposed model, we measure the performance of FashionOn in terms of Inception Score (IS) and Structural Similarity (SSIM) on the datasets collected by ourselves and DeepFashion [26], which are both composed of various types of clothes with a diversity of poses. Moreover, we conduct a user study to compare FashionOn with the state-of-the-art [42] by the A/B test. The experiments show that the proposed FashionOn outperforms the state-of-the-art method both quantitatively and qualitatively. The contributions are summarized as follows.

- We propose a novel semantic-guided image-based virtual try-on, namely, FashionOn network, that is able to generate high-quality try-on images in arbitrary poses and solve the body occlusion and dramatic posture transformation. To the best of our knowledge, FashionOn is the first virtual try-on network to precisely address wearing details (pleats and shadows) and accurate facial characteristics.
- We collected a new dataset containing 11283 pairs of the same person in different poses and the corresponding in-shop clothes images.[2] Experimental results manifest that FashionOn network outperforms the state-of-the-art method and solves the body occlusion problem. Moreover, the user study shows that 78.22% of users are more willing to use try-on services with multiple view angles and in different poses than that with only one result, which can be a value-added service to fashion e-commerce websites.

[1]The target postures are specified by keypoints, which can be 1) obtained by classic posture images in the databases with keypoints prediction model [4] or 2) manually adjusted by users.

[2]Please find the examples in https://github.com/fashion-on/FashionOn.github.io.

2 RELATED WORK

2.1 Virtual try-on

Virtual try-on networks can be categorized into two kinds of approaches: 1) clothing warping-based and 2) 3D-model-based approaches. In the following, we briefly introduce these two kinds of approaches and compare FashionOn with them.

2.1.1 Clothing warping-based try-on. Thin Plate Spline (TPS) [2] is a spline-based technique that prevails in the non-rigid transformation of images without going through any generator [22, 46]. Therefore, warping clothes directly by TPS is widely-used in many try-on research [15, 42] since it warps clothes and preserves patterns, texture, and logos. For instance, Han et al. [15] presented a coarse-to-fine image-based virtual try-on network called VITON to warp in-shop clothes through TPS and an additional refinement network for synthesizing the warped clothing details with the coarse person. Although [15] successfully implemented virtual try-on and preserved most patterns and logos, some details are still missed since the warped clothes still require to be processed by the refinement generator. Therefore, Wang et al. [42] further improved [15] via constructing a novel network CP-VTON which combined the warped clothes with the generated person through a generated composition mask instead of using the refinement generator. Nevertheless, clothing warping-based methods relying on TPS warp the clothes smoothly but fail to change the detailed surface appearance (e.g., pleats and shadows) on the clothing to follow the human poses. In contrast, FashionOn is constructed as a semantic segmentation-based method that avoids the issue. As such, FashionOn not only preserves complete details of the in-shop clothing (patterns, logos, and texture) but also generates realistic appearance (pleats and shadows) according to body shapes and human poses.

2.1.2 3D model-based try-on. Although many studies based on 2D images work on the virtual try-on task, numerous research aimed to utilize 3D body shape and 4D sequence to make the results more realistic [14, 24, 31, 43]. For example, Pons-Moll et al. [31] used a high-resolution video to capture the garment geometry in motion on a body for getting a rough and low-resolution meshes body model and aligned cloth templates to the garments of the input scans again to generate more realistic and body-fitting clothes. Gundogdu et al. utilized a Point-Net [41] like architecture to extract the information about the person and encoded body features with the garment mesh to compute the point-wise, patch-wise, and global features to predict the fitted garment. To enhance the realism of the garment on the person, Lähner et al. [24] introduced a novel framework composed of two complementary modules: a learnable statistical model based on the non-rigidly aligned clothing templates and cGAN generating the high-resolution garment map. While the above 3D-model-based methods are capable of producing try-on videos, it requires plenty of manual labors or expensive equipment

to collect the 3D annotated data for constructing 3D models. In contrast, FashionOn network only requires two images to generate try-on images with arbitrary poses.

2.2 Pose transformation

Adding the pose transformation to the virtual try-on network helps consumers have more information for the clothing in multi-aspect. A variety of researches [1, 5, 7, 8, 18, 27, 28, 32, 37, 38, 47, 48] had been proposed to transform the poses in an image. Generally, human pose transformation operates in two stages, pose estimation and image generation. The first stage can be categorized into two classes, including human keypoints estimation [4, 13, 44] and human parsing segmentation [11, 21, 25]. In the second stage, most of the works [8, 32, 37] adopted the variant architecture of GAN [12] to generate realistic images. Moreover, in [5, 27, 48], they applied a coarse-to-fine framework with the refinement network to improve the final results effectively. However, previous works of pose transformation do not apply to virtual try-on, while FashionOn seamlessly integrates both pose transformation and virtual try-on.

3 FASHIONON NETWORK

To achieve the goal of virtual try-on with arbitrary poses, we propose a novel network as shown in Fig. 2, which is comprised of three stages: (I) pose-guided parsing translator, (II) segmentation region coloring, and (III) salient region refinement. Stage I is designed to fully exploit the human body parsing information of the source user image to try on the in-shop clothing mask and transform into the target posture. Then, after deriving the transformed segmentation image, we design a coloring generator in stage II to render human appearance on the transformed segmentation image from stage I. Additionally, we utilize refinement networks in stage III to generate more details and correct errors from the results of stage II.

3.1 Pose-guided parsing translator

The information of human body segmentation is useful to generate a realistic human image because the information explicitly shows the corresponding area of each body part and is also able to guide the learning to generate a clear texture and details of each body part. Therefore, a pose-guided parsing translator is introduced to translate the parsing masks of the source image M_s to the target parsing masks M_t according to the target pose p_t. We use the pre-trained CIHP [10] to generate the human parsing labels for representing body parts, which contains neck parsing distinguish from other virtual try-on network [15, 42], most of which use the common parsing method [11]. The synthesized human parsing labels contain 20 classes, which include clothing items and body parts, e.g., upper clothes, face, right-arm. To capture the shape and learn the mapping of each class item, we use one-hot encoding for the human parsing label to 20 channels tensor $M \in R^{20\times W\times H}$, where each channel is a binary mask representing one class item such as face or hair. To eliminate the redundant information of the source user image, we replace the corresponding parsing channel of clothing with the mask of original in-shop clothing M_c, which can simultaneously offer the information of in-shop clothing shape.

The architecture of the pose-guided parsing translator is adapted from pix2pix [19], of which the generator contains 2 downsampling layers, 9 residual blocks, and 2 upsampling layers. Each residual block is composed of convolution layers and skip connection combining the input and the output of the corresponding block. The objective of our pose-guided parsing translator G_t adopt conditional GAN as following:

$$\begin{aligned}\mathcal{L}_{cGAN}(G_t, D_t) =& \mathbb{E}_{M_s,p_t,M_c,M_t}[logD((M_s,p_t,M_c),M_t)]+\\ & \mathbb{E}_{M_s,p_t,M_c}[log(1-D((M_s,p_t,M_c),G_t(M_s,p_t,M_c)))],\end{aligned}$$

where G_t tries to minimize the objective against D_t that tries to maximize it, i.e. $\arg\min_{G_t}\max_{D_t}\mathcal{L}_{cGAN}(G_t, D_t)$.

To precisely classify each pixel as the corresponding body part or the clothing item, we combine a pixel-wise binary-cross entropy loss of the G_t, denoted as $\mathcal{L}^{G_t}_{BCE}$, with our cGAN objective and the discriminator stay the same:

$$\mathcal{L}^{G_t}_{BCE}(G_t) = -\sum_{n_c} M_t \log(G_t(M_s,p_t,M_c))+(1-M_t)\log(1-G_t(M_s,p_t,M_c)),$$

where n_c represents the total number of channels of human parsing mask. The final objective is

$$\arg\min_{G_t}\max_{D_t}\mathcal{L}_{cGAN}(G_t, D_t) + \lambda_{bce}\mathcal{L}^{G_t}_{BCE}(G_t).$$

3.2 Segmentation region coloring

After deriving the target segmentation, in the second stage, the goal is to synthesize the rough results on the segmentation regions, denoted as $T = G_t(M_s, p_t, M_c)$. Therefore, we adopt the architecture of conditional GAN (cGAN) [30] for synthesizing the results with a pair of generator and discriminator. For the generator, we propose a coloring generator G_c to fill the human detailed information into the parsing regions according to the appearance of the source person image I_s and the texture of the in-shop clothing C_t. Due to the limited dataset, our model learns to change the same garment for the source person the training procedure. Therefore, we remove the garment information of I_s to avoid providing G_c any clothing information of source person.

Specifically, the in-shop clothing C_t in $R^{3\times W\times H}$, the source person image without clothing information after masking clothing $I'_s \in R^{3\times W\times H}$, and the target segmentation $T \in R^{Ch\times W\times H}$ are offered as the input of G_c. As illustrated in Fig. 2, we adopt the architecture of convolutional auto-encoder and utilize skip connections between encoder and decoder to help transmit input information to output directly. For the encoder of G_c, we design six residual blocks. Each of them is stacked with two convolution layers and ReLU to integrate T, I'_s and C_t from small local regions to broader ones so that appearance information of I'_s and C_t is able to be extracted. Additionally, local skip connections are employed to avoid vanishing gradient [3, 9] and improve our network performance. The decoder is similar and symmetric to the encoder for generating the result image I_g of G_c which fills appearance information into corresponding body parts of T.

For forcing G_c to concentrate on generating correct human part instead of the whole image, we filter out background information of the generation result $I_g = G_c(C_t, I'_s, T)$ with $1 - T_{bg}$ and so do as the ground truth I_t with $1 - T_{bg}$, where T_{bg} and $M_{t_{bg}}$ respectively represent the background channel of T and M_t. After that, we compare them with L1 distance loss function to capture the global

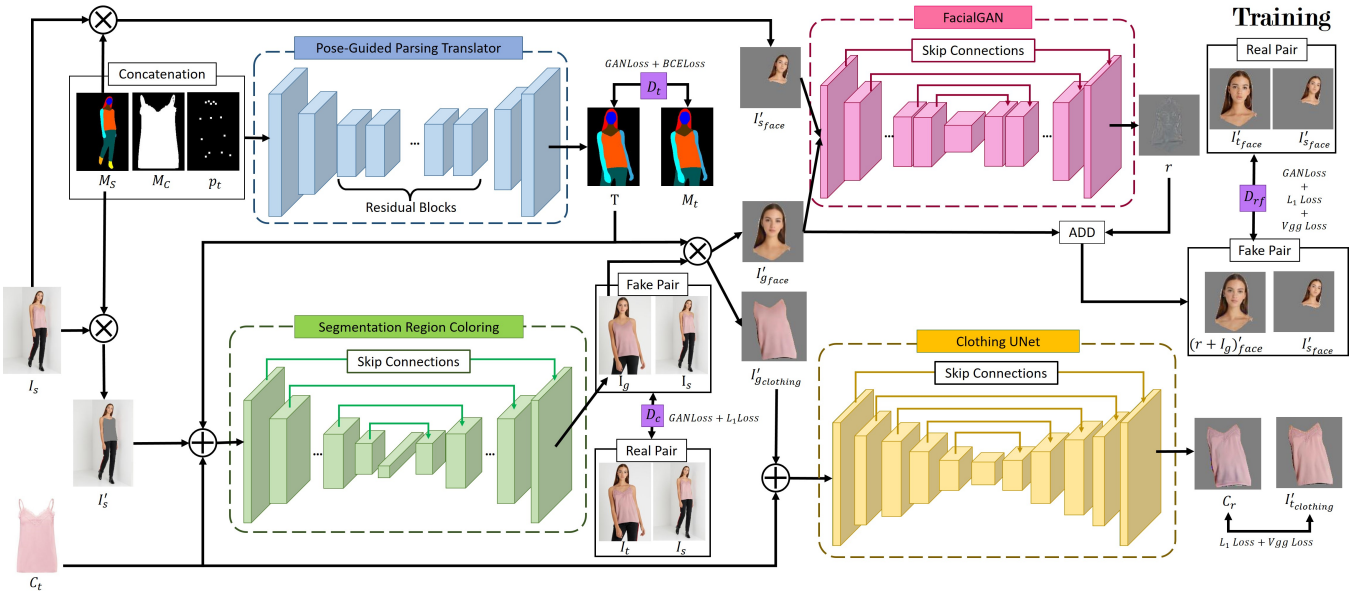

Figure 2: Training overview, which consists of three stages. Stage I (*Pose-guided parsing translator*) transforms the source human pose and generates the target parsing image in target pose T according to M_s, M_c, and P_t by a DCGAN [33]. Stage II (*Segmentation region coloring*) renders the information of clothing and human by cGAN [30] to generate a realistic source person image I_g with the target posture. Stage III (*Salient region refinement*) elaborately extracts two critical regions, face, and clothes, to additionally generate more detailed information separately by FacialGAN and Clothing UNet.

structural information and other low-frequency information:

$$\mathcal{L}_{L1} = \sum_{W}\sum_{H}\left\|G_c(C_t, I_s^{'}, T)\otimes(1-T_{bg}) - I_t\otimes(1-M_{t_{bg}})\right\|_1.$$

To complete GAN architecture, we construct the coloring discriminator D_c to differentiate two pairs: one including I_t and I_s, and the other including I_g and I_s. With inserting additional real image I_s, D_c impels G_c to generate more realistic image. Furthermore, the GAN loss is calculated based on the binary cross-entropy loss since it is a binary classification problem, i.e., true or fake images.

$$\mathcal{L}_{GAN}^{G_c} = \mathcal{L}_{BCE}(D_c(G_c(C_t, I_s^{'}, T), I_s), 1) + \lambda_{L_1}L_1$$

$$\begin{aligned}\mathcal{L}_{GAN}^{D_c} = {} & \mathcal{L}_{BCE}(D_c(G_c(C_t, I_s^{'}, T), I_s), 0) \\ & + \mathcal{L}_{BCE}(D_c(I_t, I_s), 1)\end{aligned}$$

The goal of G_c is to deceive D_c discerning its result as real image, thus, the target of $\mathcal{L}_{BCE}$ in $\mathcal{L}_{GAN}^{G_c}$ equals to 1. In contrast, the targets of $\mathcal{L}_{BCE}$ in $\mathcal{L}_{GAN}^{D_c}$ equal to 0 and 1 separately, since D_c is built for distinguishing generated and real images correctly.

3.3 Salient region refinement

The quality of virtual try-on highly depends on the characteristics (face details or body shape) of users, the clothing information (stripe, logo, bow tie), and 3D physics (pleat, shadow). Therefore, in the third stage, two refinement generators are exploited on the facial and clothing region separately to synthesize the realistic details.

3.3.1 FacialGAN. Human face and hair are complicated but play an important role in synthesizing user try-on images. Therefore, to generate the residual face details, we modify the architecture of segmentation region coloring (G_c) for the face refinement network G_{rf} by removing the fully-connected layer to preclude the loss of input details when compressing. Let T_{face} denote the human parsing mask of head regions (including the face, neck, and hair) from stage I. We use T_{face} to extract facial region from I_g and I_s for forcing G_{rf} to focus on facial details. As such, G_{rf} generates the details as a residual output $r = G_{rf}(I_g \otimes T_{face}, I_s \otimes T_{face})$. After processing images through G_{rf}, we derive the fine-tuned results by adding r to I_g.

Moreover, inspired by [20, 35], we employ the VGG perceptual loss to sharpen the fine-tuned images. When calculating the VGG perceptual distance, only the regions within T_{face} of $(r + I_g)$ and I_t are considered, which are denoted as $(r + I_g)^{'}_{face}$ and $I^{'}_{t_{face}}$, respectively. Both $(r + I_g)^{'}_{face}$ and $I^{'}_{t_{face}}$ are mapped into the feature space through the differentiable function ϕ instead of calculating the distance in the image space directly. This additional loss encourages $(r + I_g)^{'}_{face}$ and $I^{'}_{t_{face}}$ to have similar feature representation which allows the model reconstructing the details and edges better. Specifically, the VGG perceptual loss is defined as:

$$\begin{aligned}&\mathcal{L}_{vgg}^{G_{rf}}((r + I_g)^{'}_{face}, I^{'}_{t_{face}}) \\ &\qquad = \sum_{i=1}^{n}\lambda_i\left\|\phi_i((r + I_g)^{'}_{face}) - \phi_i(I^{'}_{t_{face}})\right\|_1,\end{aligned}$$

where ϕ_i represents the feature map obtained from the i^{th} layer in VGG19 model [39]. We further add L1 loss to reduce the artifacts from VGG perceptual loss and integrate GAN loss with binary cross-entropy as follows:

$$\begin{aligned}\mathcal{L}_{GAN}^{G_{rf}} &= \mathcal{L}_{vgg}^{G_{rf}}((r+I_g)'_{face}, I'_{t_{face}}) \\ &+ \sum_W \sum_H \left\| (r+I_g)'_{face} - I'_{t_{face}} \right\|_1 \\ &+ \lambda_{f1} \mathcal{L}_{BCE}(D_{rf}(I_s \otimes T_{face}, (r+I_g)'_{face}), 1) \\ &+ \lambda_{f2} \sum_W \sum_H \left\| I_{res} - I_t \otimes (1 - M_{t_{bg}}) \right\|_1,\end{aligned}$$

$$\begin{aligned}\mathcal{L}_{GAN}^{D_{rf}} &= \mathcal{L}_{BCE}(D_{rf}(I_s \otimes T_{face}, (r+I_g)'_{face}), 0) \\ &+ \mathcal{L}_{BCE}(D_{rf}(I_s \otimes T_{face}, I'_{t_{face}}), 1).\end{aligned}$$

3.3.2 Clothing UNet. Most state-of-the-art virtual try-on networks [15, 42, 45] preserved clothes detailed information by directly fusing the pre-warped clothes and users. However, this kind of approach faced the problem of veiling the limbs in front of the clothing. Therefore, to solve this problem, we re-design the try-on network by first transforming the human pose into the semantic segmentation form via pose-guided parsing translator and then coloring the clothing textures, instead of using clothing-warping. Nevertheless, after the second stage, FashionOn fills most clothing information (color and shape) back but there still exists obvious missing information (e.g., texture, logo, pleat, shadow). Therefore, we design a clothing refinement generator G_{rc}, named Clothing UNet, to extract clothing features directly from the in-shop clothing C_t and fill into the clothing part of I_g. The encoder of G_{rc} contains five downsampling convolutional layers and each layer is followed by one instance normalization layer [40] and one Leaky ReLU [29]. The decoder of G_{rc} is symmetric to the encoder.

We concatenate $I'_{g_{clothing}} = I_g \otimes T_c$ (the clothing part of I_g) and in-shop clothing C_t as the input to generate a refined clothing $C_r = G_{rc}(I'_{g_{clothing}}, C_t)$, where $T_c \in R^{1\times W\times H}$ represents the clothing channel of T. To minimize the discrepancy between the refined clothing C_r and the target clothing region $I'_{t_{clothing}} = I_t \otimes M_{t_c}$, where M_{t_c} represents the clothing channel of M_t, we introduce the L1 loss ($\mathcal{L}_{L_1}^{G_{rc}}$) and the VGG perceptual loss ($\mathcal{L}_{vgg}^{G_{rc}}$) to refine the clothing as follows:

$$\mathcal{L}_{L1}^{G_{rc}}(C_r, I'_{t_{clothing}}) = \sum_W \sum_H \left\| C_r - I'_{t_{clothing}} \right\|_1,$$

$$\mathcal{L}_{vgg}^{G_{rc}}(C_r, I'_{t_{clothing}}) = \sum_{i=1}^{5} \lambda_i \left\| \phi_i(C_r) - \phi_i(I'_{t_{clothing}}) \right\|_1,$$

where $\phi_i(C)$ represents the feature map of the clothing C of the i^{th} layer in VGG19 model [39].

Notice that L1 loss is exploited instead of L2 loss since the final stage aims to generate sharp try-on images and avoid the blurry results. Besides, to avoid spatial misalignment, we fuse the refined clothing C_r into I_g, where the clothing region is removed, to synthesize a refined human $I_{rg} = C_r \otimes T_c + I_g \otimes (1 - T_c)$. Here, we use the parsing mask T_c to select the clothing regions which helps to exclude limbs in front of the clothing when fusing the clothing. To avoid the refined clothing C_r being misaligned to I_g, we introduce L1 loss function to help C_r locate in the right position on the human body. The loss for the refined clothing try-on is defined as:

$$\mathcal{L}_{L1}^{G_{rc}}(I_{rg}, I_t) = \sum_W \sum_H \left\| I_{rg} - I_t \right\|_1.$$

The overall loss function of clothing UNet is defined by

$$\mathcal{L}^{G_{rc}} = \lambda_{c1}\mathcal{L}_{vgg}^{G_{rc}} + \lambda_{c2}\mathcal{L}_{L1}^{G_{rc}}(C_r, I'_{t_{clothing}}) + \lambda_{c3}\mathcal{L}_{L1}^{G_{rc}}(I_{rg}, I_t).$$

4 EXPERIMENTS

In this section, we first present the details of the datasets and implementation. Afterward, qualitative and quantitative analyses are conducted with the state-of-the-art method. Finally, the limitations of FashionOn are discussed for future research.

4.1 Dataset

Since most of the existing datasets [15, 42] contain only one pose for each person, we collected a new large-scale dataset comprising 10,895 in-shop clothes and 10,895 pairs of human images of the same person in 2 different poses.[3] Moreover, we also use the DeepFashion dataset [26]. There are 11,283 in-shop clothes and 11,283 pairs of human images of the same person in 2 different poses in total with the resolution of 288×192. We further wrap one in-shop clothing and two human images in different poses into a triplet for training. The dataset is randomly split into the training set and the testing set with 9,590 and 1,693 triplets, respectively.

4.2 Implementation Details

The architecture of the pose-guided parsing translator is based on ResNet, where the generator contains 2 downsampling layers, 9 residual blocks, and 2 upsampling layers. Each residual block is composed of 2 convolution layers and 1 skip connection combining the input and the output of the corresponding block. All the convolution layers in the residual blocks are with stride=1, kernel=3, and followed by ReLU.

For the segmentation region coloring, the encoder and decoder of G_c are symmetric and consists of 6 residual blocks. Each block contains 2 convolutional layers with stride=1, 1 sub-sampling convolutional layer with stride=2 except the last block, and 1 fully-connected layer. All convolutional layers contain 3×3 filters and the number of the filters linearly increases and decreases respectively for encoder and decoder.

The face refinement network (G_{rf}) is similar to segmentation region coloring but without the fully-connected layer. G_{rf} uses 4 residual blocks containing 2 convolutional layers with stride=1 and 1 sub-sampling convolutional layer with stride=2 for both encoder and decoder. Moreover, for Clothing UNet, G_{rc} uses 5 convolutional layers for both encoder and decoder to construct a UNet with stride=2. All convolutional layers contain 4×4 filters and the number of kernels linearly increases and decreases respectively for encoder and decoder. Also, it contains an Instance Normalization layer[40] and Leaky ReLU[29] following each convolutional layer.

[3] Please refer to the images in https://github.com/fashion-on/FashionOn.github.io.

Figure 3: Qualitative results sampled from our results through manipulating both numerous clothes and various poses. The input source person images are shown in the first row; further, the joint points of target pose and in-shop clothes are shown in the first and second columns respectively. The results generated based on the aforementioned clothes and poses are shown in the other columns.

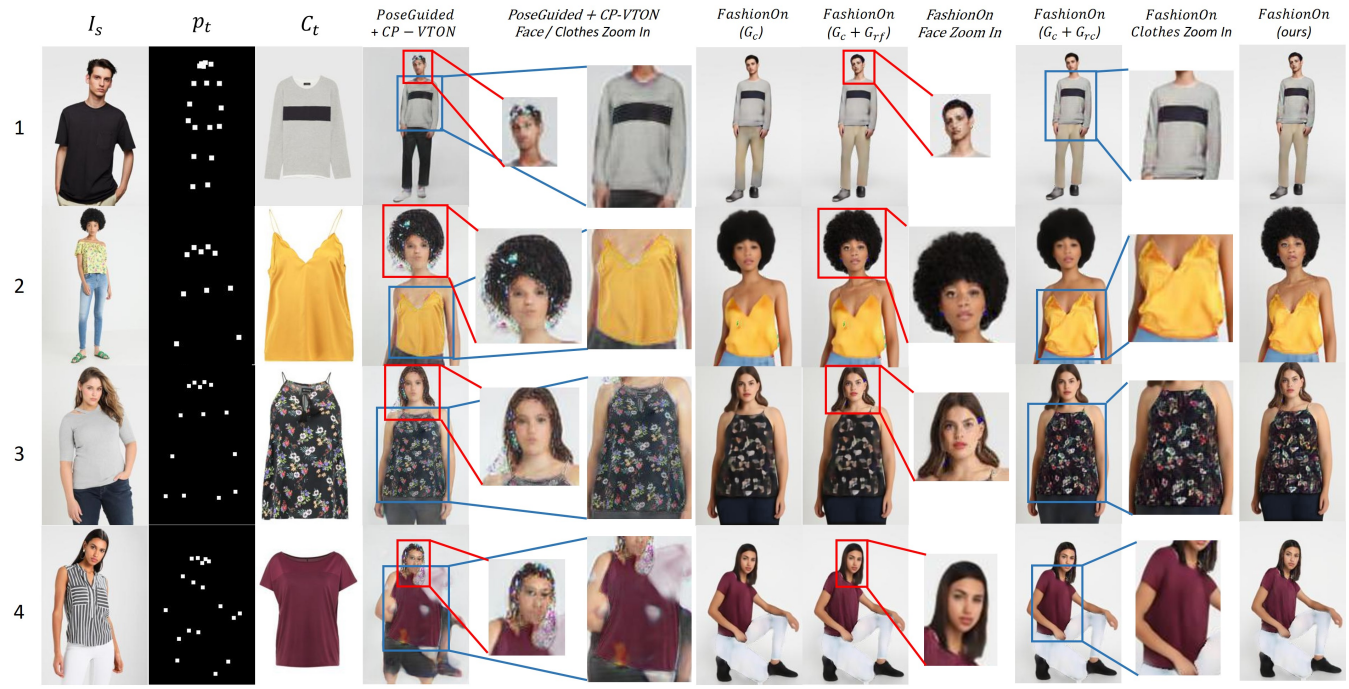

Figure 4: Visual Comparison: The most left three columns are inputs, and we show the comparison of the same conditions within different models. FashionOn network prevails the state-of-the-art work shown in the fourth column a lot.

We use Adam [23] as the optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for all stages. The learning rates of pose-guided parsing translator and other stages are 2e-4 and 2e-5, respectively.

Runtime. We randomly select 2000 image sets and report the average running time for each module with NVIDIA 1080-Ti GPU as: pose-guided parsing translator (2.6ms), segmentation region coloring (3.1ms), and salient region refinement (1.9ms).

4.3 Qualitative results

Since none of the previous researches can synthesize image-based virtual try-on with arbitrary poses, we build the baseline by first applying the state-of-the-art pose transformation (PoseGuided [27]) to transform the source user pose, and then use the state-of-the-art clothing warping (CP-VTON [42]) to warp the in-shop clothing and paste on the warped source user image. Fig. 4 shows the visual comparison of different approaches and Fig. 3 displays our sample results. The results manifest that both methods accomplish the task of virtual try-on with arbitrary poses, but the results of PoseGuided+CP-VTON contain some artifacts or lack of details, especially for the following cases.

Human limbs occlusion Row 4 in Fig. 4 shows that the proposed FashionOn successfully resolves the human limbs occlusion problems of CP-VTON since we preserve the clothing details by simultaneously warping the clothing mask and body parsing masks, and then rendering human appearance sequentially, instead of simply warping it through TPS [2] and synthesizing directly. Since G_c renders the appearance based on all masks at the same time, our model FashionOn resolves the problems of limb occlusion.

Duplicating erroneous pleats and shadows Another failure case of virtual try-on via TPS warping is the erroneous pleats and shadows as shown in row 2 in Fig. 4. The pleats and shadows on the clothing by PoseGuided+CP-VTON are exactly the same as in-shop clothing with only slight distortion. Nevertheless, these details seldom remain the same as in-shop clothes when people try on them. In contrast, FashionOn generates the pleats and shadows based on the body shape and the posture of the source person which achieves far more realistic and reasonable results.

Limitation of dramatic posture transformation PoseGuided [27] often presents mutilated limbs when transferring the posture dramatically. Compared to [27] that warps the person image based on human joints, FashionOn uses human parsing segmentation and divides the pose transform generation task into two sub-tasks to simplify warping tasks for both posture and clothing. In the Fig. 4, the case in row 4 shows that FashionOn surmounts the mutilation problem even in extreme posture transformation.

In addition, as shown in Fig. 4, FashionOn generates realistic results for a variety of clothes, races and body shapes. The demonstration proves that our model makes a great achievement in virtual try-on with arbitrary postures and preserves the details (Row 1 and 3) well. Further, FashionOn is much more effective than PoseGuided+CP-VTON since it is inevitable to lose some details when producing images through a two-phase conditional generator.

Failure case As Fig. 5 demonstrates, our FashionOn network fails to synthesize symmetric eyes in the case of transforming the sideways photo into front one since the face of the sideways photo often contains either a hidden eye or asymmetric size of eyes, which

Figure 5: Failure cases of our FashionOn network.

may require a more sophisticated model to deal with facial geometry. Besides, there is another failure case shown in Fig. 5. The fingers of generated person images are blurry because current human parsing regards the fingers as a region, regardless of each finger. Therefore, it is envisaged to develop a fine-grained human parsing for generating better results.

4.4 Quantitative results

We evaluated the quantitative performance of our FashionOn network with the other virtual try-on system, CP-VTON [42], in terms of two widely-used metrics, i.e., SSIM and IS. Moreover, a user study is conducted with 206 users to evaluate visual quality.

Evaluation Metrics Inception Score (IS) [36] is usually used to quantitatively evaluate the synthesis quality of images [15, 27, 42]. The IS score will be higher if the models can produce visually diverse and semantically meaningful images. On the other hand, Structural Similarity (SSIM) is utilized to measure the similarity between the reference image and the generated image ranging from zero (dissimilar) to one (similar).

Quantitative comparisons of the above metrics are summarized in Table 1. The results manifest that FashionOn outperforms CP-VTON in terms of both IS and SSIM by 9.1% and 12.7% respectively. Note that although FashionOn (G_r) without salient region refinement achieves the best SSIM score, it obtains the worst IS score among all our FashionOn networks since the results are blurry as illustrated in Fig. 4. On the other hand, we can also observe that FashionOn ($G_r + G_{rf}$), FashionOn ($G_r + G_{rc}$), and FashionOn (Ours) achieve higher IS score than FashionOn (G_r), which means that our salient region refinement network successfully generates the details to improve the visual quality of synthesis images.

User Study There are 206 volunteers participating in our user study. All the 1,693 humans and clothes in the test dataset are randomly composed into 15 problem sets. We compare our results with the results generated by CP-VTON [42] using the same conditioned humans and clothes. We use an A/B test for the evaluation, i.e., asking users to vote for the better try-on result. The result of the user study is summarized in Table 2. We get 2,417 votes and CP-VTON gets 673 votes, which shows the superiority of the proposed FashionOn. Moreover, after filtering 50% of the data in the middle and choosing the best results from both methods, the survey also shows

Table 1: Comparison on the test part of our virtual try-on data.

Method	IS	SSIM
CP-VTON [42]	2.9243 ± 0.0057	0.7930
FashionOn (G_r)	3.0397 ± 0.0593	**0.8992**
FashionOn ($G_r + G_{rf}$)	3.0460 ± 0.0550	0.8974
FashionOn ($G_r + G_{rc}$)	3.1655 ± 0.0894	0.8954
FashionOn (Ours)	**3.1914 ± 0.0935**	0.8935
Real Data	3.3354 ± 0.0631	1

Table 2: The user study results that FashionOn prevails CP-VTON [42].

Method	CP-VTON[42]	FashionOn (Ours)
Mean	21.78%	**78.22%**
Max (25%~75%)	26.7%	**86.7%**

FashionOn is state-of-the-art. Besides, we make a questionnaire to survey the user's preference for virtual try-on. The questions and the corresponding results are reported as follows.

Q1. How many poses should virtual try-on present? The average of answers is 5.6 and the median is 5, which means users want the virtual try-on systems could give them multi-aspect photos to buy clothes. Besides, this also proves our virtual try-on work with arbitrary poses has a larger potential than the traditional virtual try-on systems.

Q2. What is the most important consideration when using virtual try-on work? The most important consideration is the authenticity of the virtual try-on results and the second is multi-aspect photos. This shows FashionOn network totally hit the spot.

5 CONCLUSIONS AND FUTURE WORK

In this paper, we present a unique virtual try-on FashionOn network based on part-level learning to precisely generate try-on human images with arbitrary poses. FashionOn is devoted to preserving the critical human information, e.g., face, hairstyle, body shape, and clothing characteristic (pleat, shadow, and logo) to make virtual try-on results the most lifelike. Besides, compared with recent virtual try-on networks, FashionOn surpasses the recent methods on diverse clothing types and also supply multi-perspective try-on results via presenting on various poses for users to make the right choice for the most satisfying clothing. In the future, we plan to study the complicated multi-layer outfits problem by providing additional information such as categories of clothes.

6 ACKNOWLEDGEMENT

This work was supported in part by the Ministry of Science and Technology of Taiwan under Grants MOST-108-2634-F-009-006, MOST-108-2218-E-009-050, MOST-108-2823-8-002-004, MOST-108-2745-8-009-002, MOST-108-2634-F-007-009, MOST-107-2218-E-009-062, and MOST-107-2218-E-002-010.

REFERENCES

[1] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frédo Durand, and John V. Guttag. 2018. Synthesizing Images of Humans in Unseen Poses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8340–8348.

[2] Serge J. Belongie, Jitendra Malik, and Jan Puzicha. 2002. Shape matching and object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, 4 (2002), 509–522.

[3] Yoshua Bengio, Patrice Y. Simard, and Paolo Frasconi. 1994. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks* 5, 2 (1994), 157–166.

[4] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. 2017. Realtime Multi-person 2D Pose Estimation Using Part Affinity Fields. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1302–1310.

[5] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. 2018. Everybody Dance Now. In *European Conference on Computer Vision (ECCV) Workshop*.

[6] SHung-Jen Chen, Ka Ming Hui, Szu Yu Wang, Li-Wu Tsao, Hong-Han Shuai, , and Wen-Huang Cheng. 2019. BeautyGlow: On-Demand Makeup Transfer Framework with Reversible Generative Network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] 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 (NIPS) 31*. Curran Associates, Inc., 474–484.

[8] Yixiao Ge, Zhuowan Li, Haiyu Zhao, Guojun Yin, Shuai Yi, Xiaogang Wang, and hongsheng Li. 2018. FD-GAN: Pose-guided Feature Distilling GAN for Robust Person Re-identification. In *Advances in Neural Information Processing Systems (NIPS) 31*. Curran Associates, Inc., 1222–1233.

[9] Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, Vol. 9. PMLR, 249–256.

[10] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. 2018. Instance-Level Human Parsing via Part Grouping Network. In *European Conference on Computer Vision (ECCV)*. 770–785.

[11] Ke Gong, Xiaodan Liang, Xiaohui Shen, and Liang Lin. 2017. Look into Person: Self-Supervised Structure-Sensitive Learning and a New Benchmark for Human Parsing. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6757–6765.

[12] 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 (NIPS) 27*. Curran Associates, Inc., 2672–2680.

[13] Riza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. DensePose: Dense Human Pose Estimation in the Wild. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 7297–7306.

[14] Erhan Gundogdu, Victor Constantin, Amrollah Seifoddini, Minh Dang, Mathieu Salzmann, and Pascal Fua. 2018. GarNet: A Two-stream Network for Fast and Accurate 3D Cloth Draping. *CoRR* (2018). arXiv:1811.10983

[15] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. 2018. VITON: An Image-Based Virtual Try-on Network. In *2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Shintami Chusnul Hidayati, Yu-Ting Chang Cheng-Chun Hsu, Kai-Lung Hua, Jianlong Fu, and Wen-Huang Cheng. 2018. What Dress Fits Me Best? Fashion Recommendation on the Clothing Style for Personal Body Shape. In *ACM International Conference on Multimedia*.

[17] Shintami Chusnul Hidayati, Chuang-Wen You, Wen-Huang Cheng, and Kai-Lung Hua. 2018. Learning and Recognition of Clothing Genres from Full-body Images. *IEEE Transactions on Cybernetics* (2018), 1647–1659.

[18] Zhongyue Huang, Jingwei Xu, and Bingbing Ni. 2018. Human Motion Generation via Cross-Space Constrained Sampling. In *International Joint Conferences on Artificial Intelligence Organization (IJCAI)*. 757–763.

[19] 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 (CVPR)*. IEEE, 5967–5976.

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. 2016. Perceptual Losses for Real-Time Style Transfer and Super-Resolution. In *European Conference on Computer Vision (ECCV)*, Vol. 9906. 694–711.

[21] Mahdi M. Kalayeh, Emrah Basaran, Muhittin Gokmen, Mustafa E. Kamasak, and Mubarak Shah. 2018. Human Semantic Parsing for Person Re-identification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1062–1071.

[22] Angjoo Kanazawa, David W. Jacobs, and Manmohan Krishna Chandraker. 2016. WarpNet: Weakly Supervised Matching for Single-View Reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3253–3261.

[23] Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

[24] Zorah Lähner, Daniel Cremers, and Tony Tung. 2018. DeepWrinkles: Accurate and Realistic Clothing Modeling. In *European Conference on Computer Vision (ECCV)*, Vol. 4. 698–715.

[25] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. 2018. Look into Person: Joint Body Parsing and Pose Estimation Network and a New Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 4 (2018), 871–885.

[26] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. 2016. DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1096–1104.

[27] 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 (NIPS) 30*. Curran Associates, Inc., 406–416.

[28] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. 2018. Disentangled Person Image Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 99–108.

[29] Andrew L. Maas, Awni Y. Hannun, and Andrew Y. Ng. 2013. Rectifier nonlinearities improve neural network acoustic models. In *International Conference on Machine Learning (ICML) Workshop on Deep Learning for Audio, Speech and Language Processing*.

[30] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. *arXiv preprint* (2014). arXiv:1411.1784

[31] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael Black. 2017. ClothCap: Seamless 4D Clothing Capture and Retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)* 36, 4 (2017).

[32] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. 2018. Unsupervised Person Image Synthesis in Arbitrary Poses. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8620–8628.

[33] Alec Radford, Luke Metz, and Soumith Chintala. 2016. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. In *International Conference on Learning Representations (ICLR)*.

[34] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. 2018. SwapNet: Image Based Garment Transfer. In *European Conference on Computer Vision (ECCV)*.

[35] Mehdi S. M. Sajjadi, Bernhard Schölkopf, and Michael Hirsch. 2017. EnhanceNet: Single Image Super-Resolution Through Automated Texture Synthesis. In *IEEE International Conference on Computer Vision (ICCV)*. 4501–4510.

[36] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. 2016. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems (NIPS) 29*. Curran Associates, Inc., 2234–2242.

[37] Chenyang Si, Wei Wang, Liang Wang, and Tieniu Tan. 2018. Multistage Adversarial Losses for Pose-Based Human Image Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 118–126.

[38] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. 2018. Deformable GANs for Pose-Based Human Image Generation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3408–3416.

[39] Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *International Conference on Learning Representations (ICLR)*.

[40] Dmitry Ulyanov, Andrea Vedaldi, and Victor S. Lempitsky. 2017. Improved Texture Networks: Maximizing Quality and Diversity in Feed-forward Stylization and Texture Synthesis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4105–4113.

[41] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer Networks. In *Advances in Neural Information Processing Systems (NIPS) 28*. Curran Associates, Inc., 2692–2700.

[42] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Meng Yang. 2018. Toward Characteristic-Preserving Image-Based Virtual Try-On Network. In *European Conference on Computer Vision (ECCV)*. 589–604.

[43] Tuanfeng Y. Wang, Duygu Ceylan, Jovan Popovic, and Niloy J. Mitra. 2018. Learning a Shared Shape Space for Multimodal Garment Design. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 203:1–203:13.

[44] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. 2016. Convolutional Pose Machines. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4724–4732.

[45] Zhonghua Wu, Guosheng Lin, Qingyi Tao, and Jianfei Cai. 2018. M2E-Try On Net: Fashion from Model to Everyone. *arXiv preprint* (2018). arXiv:1811.08599

[46] Xuemiao Xu, Liang Wan, Xiaopei Liu, Tien-Tsin Wong, Liansheng Wang, and Andrew Chi-Sing Leung. 2008. Animating animal motion from still. *ACM Transactions on Graphics (TOG)* 27, 5 (2008), 117:1–117:8.

[47] Yichao Yan, Jingwei Xu, Bingbing Ni, Wendong Zhang, and Xiaokang Yang. 2017. Skeleton-Aided Articulated Motion Generation. In *Proceedings of the 25th ACM International Conference on Multimedia*. ACM, 199–207.

[48] Bo Zhao, Xiao Wu, Zhi-Qi Cheng, Hao Liu, Zequn Jie, and Jiashi Feng. 2018. Multi-View Image Generation from a Single-View. In *Proceedings of the 26th ACM International Conference on Multimedia*. ACM, 383–391.