VTNFP: An Image-based Virtual Try-on Network with Body and Clothing Feature Preservation

Ruiyun Yu[1] Xiaoqi Wang[1*] Xiaohui Xie[2]
[1]Software College, Northeastern University, China
[2]Department of computer science, University of California, Irvine, CA 92617
yury@mail.neu.edu.cn, 1701290@stu.neu.edu.cn, xhx@uci.edu

Abstract

Image-based virtual try-on systems with the goal of transferring a desired clothing item onto the corresponding region of a person have made great strides recently, but challenges remain in generating realistic looking images that preserve both body and clothing details. Here we present a new virtual try-on network, called VTNFP, to synthesize photo-realistic images given the images of a clothed person and a target clothing item. In order to better preserve clothing and body features, VTNFP follows a three-stage design strategy. First, it transforms the target clothing into a warped form compatible with the pose of the given person. Next, it predicts a body segmentation map of the person wearing the target clothing, delineating body parts as well as clothing regions. Finally, the warped clothing, body segmentation map and given person image are fused together for fine-scale image synthesis. A key innovation of VTNFP is the body segmentation map prediction module, which provides critical information to guide image synthesis in regions where body parts and clothing intersects, and is very beneficial for preventing blurry pictures and preserving clothing and body part details. Experiments on a fashion dataset demonstrate that VTNFP generates substantially better results than state-of-the-art methods.

1. Introduction

As more and more consumers are shopping apparel and accessories online, technologies that allow consumers to virtually try on clothes can not only enhance consumers' shopping experience, but also help transform the way how people shop for fashion items. Motivated by this, a number of methods have been proposed to solve the virtual try-on problem, which can be broadly classified into two categories: methods based on 3D modeling [10, 43, 28, 35, 4], and methods based on 2D images [13, 30, 11, 39].

*Corresponding author.

Classical virtual try-on methods are primarily 3D based. Applications in this category include SenseMi, triMirror, etc. 3D-based methods rely on computer graphics to build 3D models and render the resulting images, which can well control clothing deformation, material performance and other issues. However, they are computationally intensive and require additional information to build 3D models [35], which has constrained their adoption in online e-commerce or real-time AR applications.

Recently virtual try-on methods based solely on RGB images have also been proposed [11, 39, 13, 30]. These methods formulate virtual try-on as a conditional image generation problem, which are much less resource intensive and having the potential for widespread applications, if proven effective.

On the other hand, generating perceptually convincing virtual try-on images without 3D information is challenging. For a synthetic image to be realistic and effective, it has to meet the following criteria: 1) the posture and body shape of the person should be preserved, and body parts should be clearly rendered; 2) clothing items not intended to be replaced, such as trousers, should be preserved; 3) the target clothing item should well fit to the intended body part of the person; and 4) the texture and embroidery details of the target clothing should be retained as much as possible.

Recent methods have taken a two-stage approach by first aligning the target clothing to the body shape of a given person and then fusing warped clothing and person images together. VITON [11] implemented a coarse-to-fine framework, generating warped clothing using thin-plate spline (TPS) transformation. CP-VTON [39] proposed a geometric matching module to directly learn the parameters of TPS for clothing warping, and a single-step synthetic network to merge rendered person and warped clothing images. CP-VTON improved the preservation of clothing details, but it has drawbacks on preserving body parts and clothing items that should not be changed.

Figure 1 shows example synthetic images generated by VITON and CP-VTON. A few issues are worth noting: 1)

Figure 1. Visual comparison of three different methods.

both models didn't preserve trousers in the reference image, 2) the left forearm is either deformed (VITON) or incorrectly clothed (CP-VTON), and 3) CP-VTON is better than VITON in preserving clothing details, but the regions at the interaction of clothing and body are blurry. We believe there are two main reasons behind these shortcomings. First, the clothing-agnostic representations used by both VITON and CP-VTON don't retain enough body part information. Second, important body parts information such as arms and trousers is not fully represented in the final synthesis.

To address the challenges mentioned above, we propose a new image-based virtual try-on method, called VTNFP. Figure 2 gives an overview of VTNFP, consisting of three modules: 1) Clothing Deformation Module for aligning the target clothing to the posture of a given person. Different from CP-VTON, we incorporate a self-attention mechanism to make the correlation matching component more robust; 2) Segmentation Map Generation Module, the goal of which is to generate a body segmentation map of the person wearing the target clothing. This module is a key contribution of our method and is primarily responsible for its improved performance; and 3) Try-on synthesis Module, which fuses the warped clothing, the predicted body segmentation map and additional auxiliary information together for final image synthesis. Experiments show that VTNFP significantly improved the state-of-the-art methods for virtual try-on image synthesis, generating images with better preservation of both clothing details and body parts (Figure 1).

The main contributions of our work are summarized as follows:

- We propose a new segmentation map generation module to predict the body parts of a person wearing the target clothing. We show that such a module can be efficiently trained, and is instrumental for improving the performance of image synthesis.
- We present a new image synthesis network to fuse information from the predicted body part segmentation map, warped clothing and other auxiliary body information to preserve clothing and body part details.
- We demonstrate that our new method performs substantially better than the state-of-the-art methods both qualitatively and quantitatively.

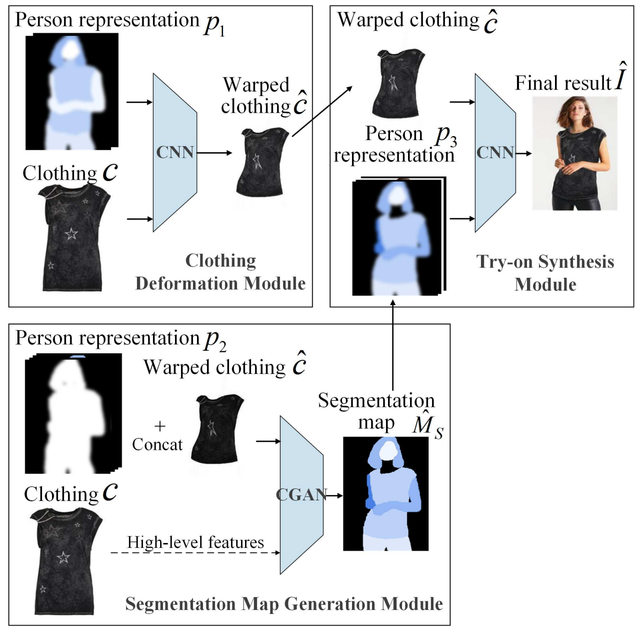

Figure 2. Overview of VTNFP, consisting of three modules - clothing Deformation Module, Segmentation Map Generation Module and Try-on Synthesis Module.

2. Related Work

2.1. Image Synthesis

Generative adversarial network (GAN) [9, 29, 6, 48] is one of the most popular deep generative models for image generation, and has shown impressive results in many applications [44, 1, 15, 49, 27]. Conditional GAN (cGAN) generates images conditional on certain input signals such as attributes [36], class information [25], sketch [33, 20, 41], text [31, 46], and pose [21]. Image-to-image translation networks [12] synthesize new images conditional on an input image, but tend to generate blurry images when the conditional image is not well aligned with the target image.

In the domain of apparel image synthesis, [47] generates multi-view clothing images from only a single view. [45] generates isolated clothing images from the image of a clothed person. [17] and [21] synthesize images of clothed people with different poses. FashionGAN [50] generates clothed images based on text descriptions of fashion items.

2.2. Human Parsing and Understanding

Human parsing and understanding have been used in many tasks, such as traffic supervision [2], behavior recognition [23], and so on. Current algorithms can be generally divided into three categories: 1) clothing parsing [18, 42, 8], 2) body parts parsing [38, 7], and 3) body posture parsing, including 2D pose [3], 3D pose [32] or body shape [34] parsing.

2.3. Virtual Try-on

Virtual try-on methods can be broadly classified into two categories: methods based on 3D body modeling [10, 43, 28, 35, 4], and methods based solely on 2D images [13, 30, 11, 39]. 3D methods can generate great results for virtual try-on, but require additional 3D measurements and more computing power.

2D image-based methods are more broadly applicable. Jetchev and Bergmann [13] proposed a conditional analogy GAN to swap clothing on people images, but requires paired clothing images to train the model. SwapNet [30] proposed a method to interchange garment appearance between two single views of people. VITON [11] and CP-VTON [39] generate new images given a target clothing item and a clothed person image, and are most relevant to the problem we are trying to solve.

3. VTNFP

Given a target clothing image c and a reference image I containing a clothed person (wearing different clothing), the goal of VTNFP is to generate a new image $\hat{I}$ of the person wearing clothing c such that the body shape and pose of the person are retained. Ideally, the training data for our model should be in the form of triplets (I, c, $\hat{I}$). However, such data are uncommon; instead, we are provided with more readily available training data in pairs of (I, c) only. In order to train an image generation model, we create clothing-agnostic person representations of I, and train a model to generate a synthetic image $\hat{I}$ based on the clothing-agnostic person representations and c.

VTNFP consists of three modules (Figure 2): a) a clothing deformation module $\hat{c} = M_1(p_1, c)$, which transforms c to a warped version $\hat{c}$ that aligns with the posture of the person, given person representation p_1; b) a segmentation map generation module $\hat{M}_s = M_2(p_2, \hat{c})$, which generates a new segmentation of body parts as well as body regions covered by the target clothing, given person representation p_2 and $\hat{c}$; and c) a try-on synthesis module $\hat{I} = M_3(p_3, \hat{c})$, which synthesizes the final target image. Key to our model are three person representations, among which p_1 and p_2 are derived directly from I, whereas p_3 is predicted based on both I and $\hat{c}$. p_3 contains information on segmentation of body parts and clothing of the intended target image, and is critical for preserving clothing details and body parts in the synthesized image $\hat{I}$.

3.1. Person Representation

To retain human body and clothing features, we propose a hybrid clothing-agnostic person representation (HCPR) method to derive three levels of person representations, p_1, p_2 and p_3 (Figure 3).

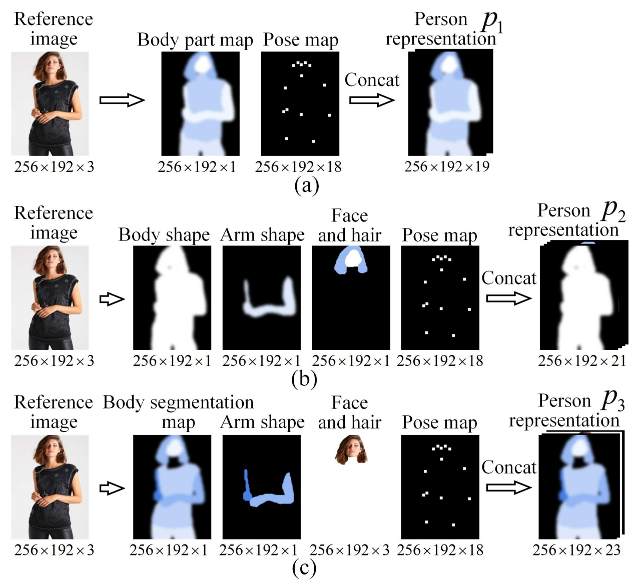

Figure 3. Hybrid clothing-agnostic person representation.

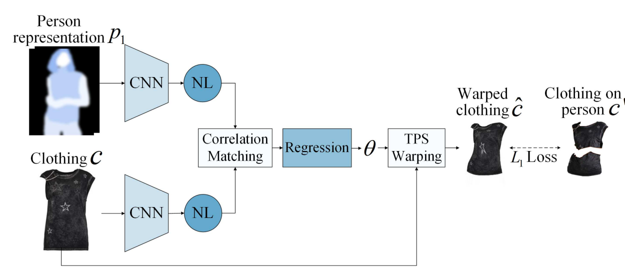

Figure 4. Clothing Deformation Module.

Person representation p_1 consists of two components - 1-channel body part map and 18-channel pose map. The body part map contains the class labels of 6 body parts, derived from the reference image I using the method described in [7]. The pose map contains predicted positions of 18 keypoints in I [3] with every keypoint represented by an 11×11 rectangle centered at the predicted position.

Person representation p_2 consists of four components - 1-channel body shape map, 1-channel arm shape map, 1-channel face and hair map, and 18-channel pose map. The pose map is the same as the one in p_1, while the other maps are generated by combining the body part map in p_1 and additional semantic part labels extracted by the method described in [8].

Person representation p_3 consists of four components - 1-channel body segmentation map, 1-channel arm shape map, 3-channel face and hair map, and 18-channel pose map. The body segmentation map contains the class labels of 13 semantic regions of the person wearing the target clothing (not the original clothing), including upper and

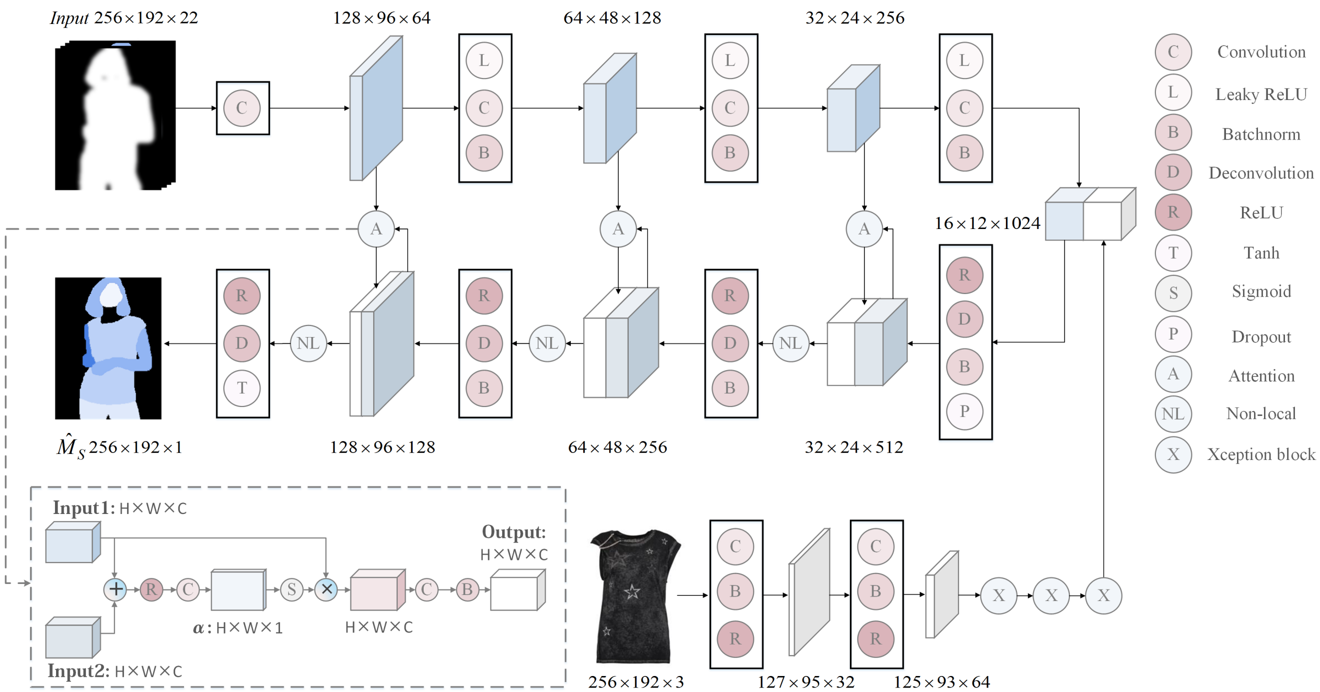

Figure 5. Segmentation Map Generation Module.

lower body clothing, hat, hair, glove, sunglasses, scarf, face, left and right arms, left and right legs, and shoes, predicted by the segmentation map generation module (sec 3.3) based on p_2 and $\hat{c}$. The arm shape map, derived from the predicted body segmentation map, indicates regions of the exposed arms in the target image. (Note its difference from the arm shape map in p_2, which is derived from I and is clothing-agnostic.) The face and hair map contains the RGB image of face and hair. The pose map is the same as before.

The three person representations serve different purposes. p_1 and p_2 provide a clothing-agnostic representation of the input image I, with p_1 used to generate warped clothing $\hat{c}$ and p_2 used to predict the body segmentation map. p_3 provides a blueprint of the target image to be generated, derived from p_2 and $\hat{c}$.

3.2. Clothing Deformation Module M_1

This module transforms the target clothing c from a positive perspective into a warped form $\hat{c}$, aligned with the posture and body shape represented by p_1. Similar to CP-VTON, M_1 also utilizes a geometric matching module (GMM) to generate parameters of thin-plate spline (TPS) for warping c, but incorporates a non-local (NL) mechanism [40] to improve feature learning and matching.

Figure 4 shows the basic architecture of M_1. It starts with convolutional neural networks, followed by non-local layers to extract features of p_1 and c respectively, which are then combined to produce a tentative correspondence map. At last, the correspondence map is fed to a regression layer to predict the parameters of TPS for warping c to $\hat{c}$. The entire module was trained by minimizing the ℓ_c loss between the warped clothing $\hat{c}$ and the ground truth c' segmented from the original image I,

$$L(\hat{c}, c') = \|\hat{c} - c'\|_1. \tag{1}$$

3.3. Segmentation Map Generation Module M_2

The goal of M_2 is to generate a semantic segmentation map $\hat{M}_s$ of the person wearing the target clothing. Note that from the input image we can only obtain the semantic body map of the person in the origin clothes. However, we show in this module that $\hat{M}_s$ can be predicted based on clothing-agnostic person representation p_2 and warped clothing $\hat{c}$.

Figure 5 shows the overall architecture of M_2, which consists of an encoder to encode the information from p_2 and $\hat{c}$, and a decoder to generate the semantic segmentation map $\hat{M}_s$. An attention sub-network is incorporated to the model to gate the lateral information flow from the encoder to the decoder [26]. Non-local operations are used throughout the network to capture long-range dependencies.

Although warped clothing $\hat{c}$ is an input to the above encoder-decoder framework, some texture and embroidery information in the target clothing c might get lost after warping. To better preserve the original information, we further add a branch to extract clothing features directly from c and concatenate them to the encoder features.

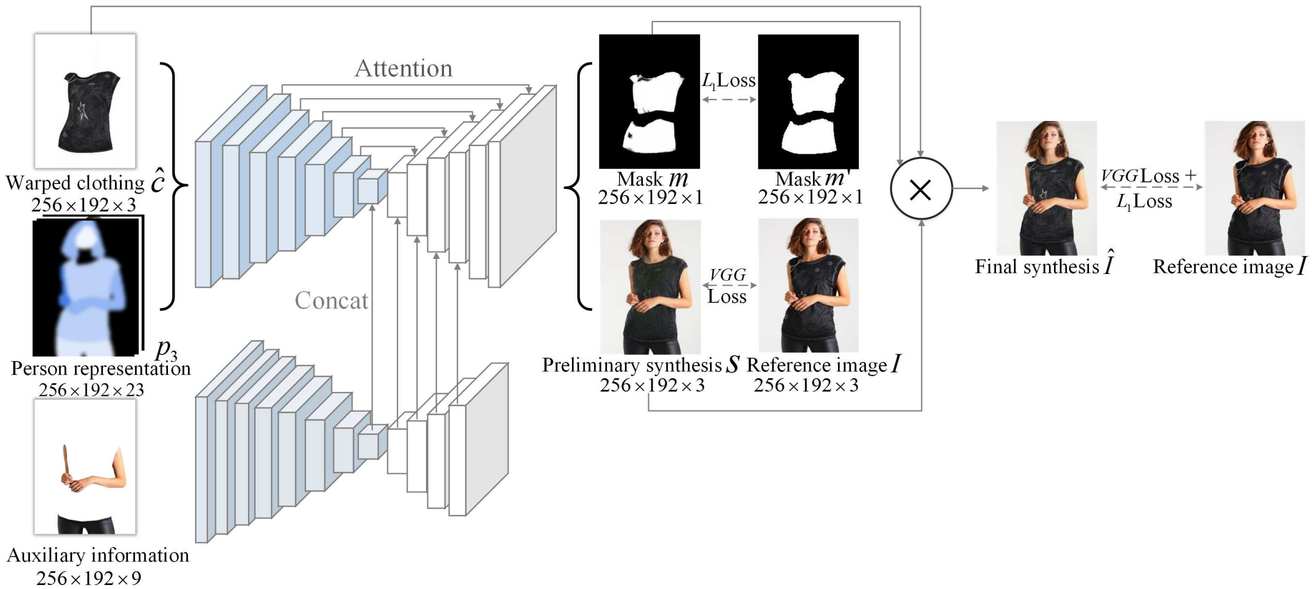

Figure 6. Try-on Synthesis Module.

We trained the module $\hat{M}_s = M_2(p_2, \hat{c})$ using training data with target clothing and reference image pairs (c, I), where I shows the image of a person wearing c. We first derive clothing-agnostic representation p_2 from I. The predicted semantic segmentation map $\hat{M}_s$ is then compared to the ground-truth segmentation map M_s, extracted directly from I based on the method in [8]. This module can also be viewed as a conditional GAN model. The final loss L_{SMGM} for training module M_2 consists of a focal loss [19] on pixel-wise segmentation performance and an adversarial loss to distinguish true semantic segmentation maps from fake ones:

$$L_{fl} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{C}(1-\hat{y}_{ik})^{\gamma} y_{ik}\log(\hat{y}_{ik}) \quad (2)$$

$$L_{cGAN} = \mathbb{E}_{x,y}[\log D(x,y)] + \mathbb{E}_{x,z}[\log(1-D(x,G(x,z)))] \quad (3)$$

$$L_{SMGM} = \alpha L_{fl} + (1-\alpha)L_{cGAN}, \quad (4)$$

where i and k denote the indices of pixels and semantic body parts, respectively. y_{ik} denotes semantic segmentation ground-truth, while $\hat{y}_{ik}$ denotes the predicated probability. Eq. (3) indicates the conditional GAN loss, where x is the input data (combination of p_2 and $\hat{c}$), y is a ground-truth segmentation map, and z is the noise in the form of dropout [12].

3.4. Try-on Synthesis Module M_3

The aim of M_3 is to synthesize the final virtual try-on image $\hat{I}$ based on the outputs from the first two modules. Overall, we use three sources of information: warped clothing $\hat{c}$ from M_1, p_3 from M_2, auxiliary information on pants and arms extracted from the original image I.

Figure 6 shows the overall architecture of M_3, consisting of two parts. The upper branch uses an attention-gated U-Net to extract features from p_3 and $\hat{c}$. The lower branch consists of 7 encoding layers, designed based on the idea of Xception [5], and 4 decoding layers to extract features from the auxiliary information, which are then concatenated to the features extracted from the upper branch. The main motivation for including the lower branch is to retain the original pants and arms feature in the synthesized images.

The synthesis module outputs a mask m, denoting the clothing regions in the target image, and a preliminary synthesis s. The final synthesis $\hat{I}$ is obtained by fusing s and $\hat{c}$, guided by m as follows,

$$\hat{I} = m \odot \hat{c} + (1-m) \odot s, \quad (5)$$

where $\odot$ denotes element-wise matrix multiplication.

The loss function L_{TSM} in M_3 includes four components shown in Eq. (10). $L(m, m')$ is an ℓ_1 loss between the predicted clothing mask and the ground truth m'. The ground-truth mask is derived from the warped clothing segmentation map $\hat{c}$ by removing the arm part, as shown in Figure 6. This loss encourages the network to retain as many clothing details as possible. $L(\hat{I}, I)$ measures the ℓ_1 loss between the synthesized image $\hat{I}$ and the ground-truth I. In addition to pixel-wise intensity differences, we also consider a perceptual loss between two images, measured by features extracted from the VGG model [14]. $L_{VGG}(s, I)$ measures the perceptual loss between the preliminary synthesis s and I, and $L_{VGG}(\hat{I}, I)$ the perceptual loss between $\hat{I}$ and I. The perceptual losses help make the synthesized images more photo-realistic. The overall loss is a weighted sum of the four losses described above:

$$L(m, m') = \|m - m'\|_1 \tag{6}$$

$$L(\hat{I}, I) = \|\hat{I} - I\|_1 \tag{7}$$

$$L_{VGG}(s, I) = \sum_{i=1}^{5} \lambda_i ||f_i(s) - f_i(I)||_1 \tag{8}$$

$$L_{VGG}(\hat{I}, I) = \sum_{i=1}^{5} \lambda_i ||f_i(\hat{I}) - f_i(I)||_1 \tag{9}$$

$$\begin{aligned} L_{TSM} &= \lambda_1 L(m, m') + \lambda_2 L_{VGG}(s, I) \\ &+ \lambda_3 L(\hat{I}, I) + \lambda_4 L_{VGG}(\hat{I}, I) \end{aligned} \tag{10}$$

4. Experiments and Analysis

4.1. Dataset

The dataset used for experiments is the same as the one in VITON and CP-VTON, consisting of 19,000 pairs of top-clothing images and positive perspective images of female models. Some incomplete image pairs are removed, leaving behind 14,006 pairs for training and 2,002 pairs for testing. In the training set, the target clothing and the clothing worn by the model is the same. However, in the test set, the two are different. All of our evaluations and visualizations are performed on images from the test set.

4.2. Implementation Details

The size of all the input images and the output images is fixed to 256×192.

Clothing Deformation Module. We trained this module for 200K epochs with batch size 4. The Adam [16] optimizer is used with $\beta_1 = 0.5$, $\beta_2 = 0.999$. The learning rate is first fixed at 0.0001 for 100K epochs and then linearly reduced to zero in the remaining 100K epochs. The structure of two CNN networks for feature extraction is similar. Each has six convolutional layers, including four 2-strided layers and two 1-strided layers, followed by a non-local [40] layer. The numbers of filters are 64, 128, 256, 512, 512. The regression convolutional network for parameter estimation consists of two 2-strided convolutional layers, one 1-strided convolutional layer and one fully-connected output layer. The numbers of filters are 512, 256, 128, 64, respectively.

Segmentation Map Generation Module. In this module, parameters in Eq. (4) are set as $\alpha = 0.5$. We trained this module for 15 epochs with batch size 5. The generator contains four encoding layers and four decoding layers, where the 2-strided filter size is 4×4. The numbers of filters in encoding layers are 64, 128, 256, 512, respectively. For decoding layers, the numbers of channels are 512, 256, 128, 1, respectively. The non-local layers are added after the concatenated layers. The convolutional neural network for extracting high-level features of undeformed clothing contains two convolutional layers with 3×3 spatial filters, and three Xception blocks [5], where the numbers of filters are 32, 64, 128, 256, 512, respectively. The discriminator is designed as in [12].

Figure 7. The effect of the high-level feature extraction branch and the non-local layer. (a) is the reference image; (b) is the target clothing image; (c) is the result of removing the high-level feature extraction branch; (d) is the result of removing the non-local layer; (e) is the results of our VTNFP.

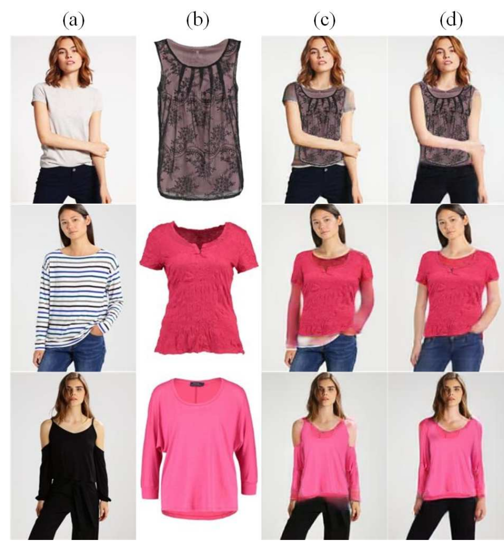

Figure 8. The effect of lower branch in the synthesis module. (a) is the reference image; (b) is the target clothing image; (c) shows the result of removing the lower branch of the try-on synthesis module; (d) is the result of our VTNFP.

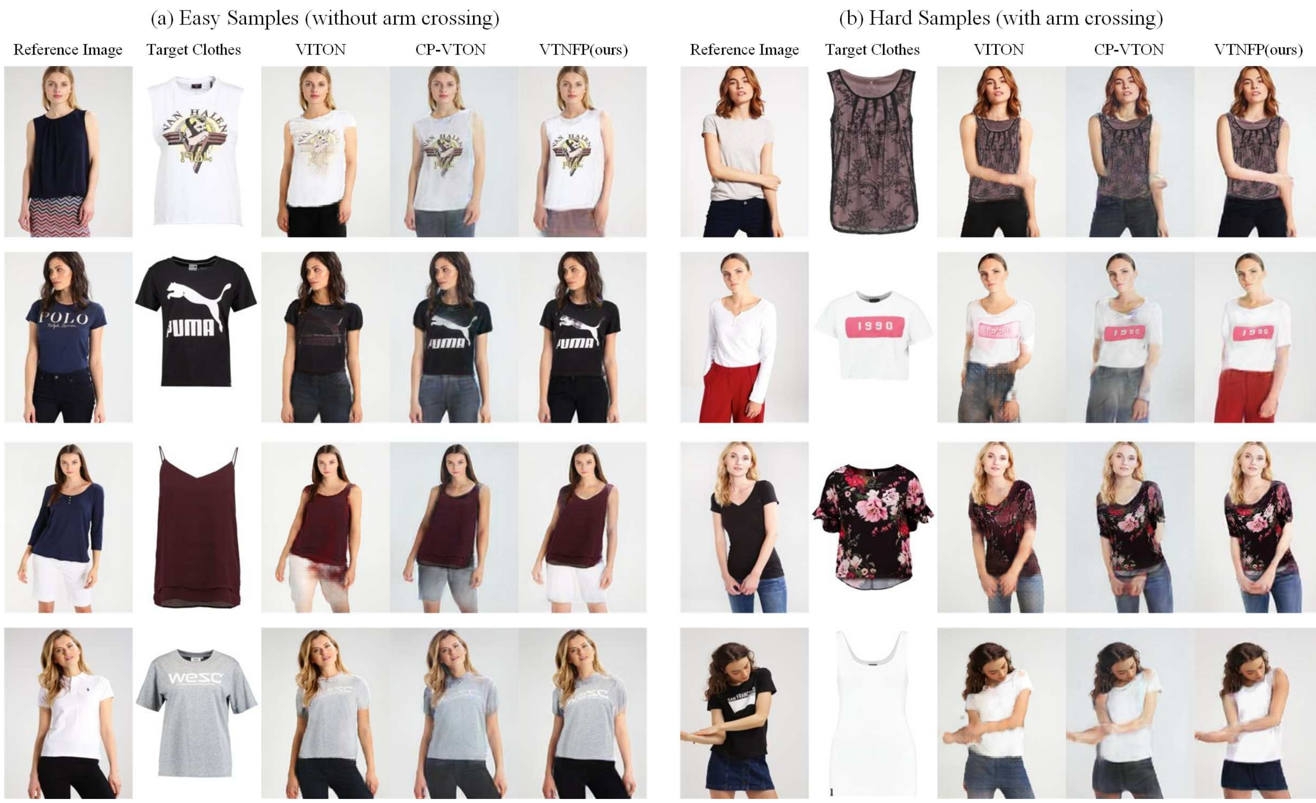

Figure 9. Visual comparison of three different methods. Our method VTNFP generates more realistic try-on results, which preserves both the clothing texture and person body features.

Try-on Synthesis Module. In this module, we set $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 1$ in Eq. (10). The settings of training steps, optimizer and learning rate are the same as those in the clothing deformation module.

All encoding layers of upper branch use 4×4 spatial filters with a stride of 2, and the numbers of filters are 64, 128, 256, 512, 512, 512, respectively. As recommended by [39, 24], we use the combination of nearest-neighbor interpolation layer and 1-strided convolutional layer instead of 2-strided deconvolutional layer for the decoding layers. So all the decoding layers consist of up-sampling layer with scale_factor of 2 and convolutional layer of 3×3 spatial filters with 1 stride, and the numbers of filters are 512, 512, 256, 128, 64, 4, respectively. We use LeakyReLU [22] for encoding layers and ReLU for decoding layers, and each convolutional layer is followed by an instance normalization layer [37].

The lower branch is a different encoding and decoding network. In the encoding part, the numbers of filters are 32, 64, 128, 256, 512, 512, 512, respectively. The first and second convolutional layers contain 3×3 spatial filters with a stride of 2 and 1, respectively. The last five convolutional layers are Xception blocks. In the decoding part, we use the same structures as the first four layers of the upper branch.

4.3. Qualitative Results

In this section, we provide some qualitative results of our model. Through visualization, we demonstrate the contributions to model performance from various network components we incorporated to our model. We also show that VTNFP produces more realistic looking virtual try-on images than two state-of-the-art models, VITON and CP-VTON.

The effect of non-local layers and features from the undeformed clothing on body segmentation map generation. Figure 7 illustrates the effects of these two components on predicting the body segmentation map from module M_2. Shown on the column (a) is the reference image, column (b) is the target clothing image, and the last three columns of images represent the segmentation map of the current person wearing target clothing. Column (c) is the result of removing the features from undeformed clothing, column (d) is the result of removing non-local layers, and column (e) is the result of VTNFP. It shows that without the features from the undeformed clothing or non-local layers, the results are less stable.

The effect of lower branch in the synthesis module. In Figure 8, column (a) is the reference image, column (b)

is the target clothing image, column (c) shows the result of removing the lower branch of the try-on synthesis module by putting the arm and pants information, p_3 and $\hat{c}$ in one upper branch. As we can see, the results are not as good as the results of VTNFP (column (d)), because, without the lower branch, the network learns hybrid features of up-clothing, pants, and arms, and can't recover the pants and arm information well in the testing phase.

Comparison of try-on results. Figure 9 presents a visual comparison of three different methods. Compared with CP-VTON, VITON performs better on preserving persons' posture, but does not preserve the clothing details as well. On the other hand, CP-VTON performs better on retaining clothing details, but worse on body posture preservation. In both models, pants is often not well retained after replacing tops.

By contrast, VTNFP is able to retain both the body posture and clothing details at the same time. In Figure 9, most of the pants in the original images are well preserved, unlike the other two models. We can observe that VTNFP is able to retain more clothing details in all cases compared with VITON and CP-VTON. Most importantly, when a person's posture is complex, e.g. when the arms are crossing, VTNFP performs substantially better than other two models on retaining the person's body information, as shown in column (b) of Figure 9.

The main reason behind VITON's under performance is that the mask used in VITON tends to preserve coarse person image information, such as body information, while ignoring the details of the warped clothing. As shown in Figure 9, VITON loses the texture of clothing. To get better results, CP-VTON generates a rendered coarse person image and a mask at the same time, replacing the coarse-to-fine strategy in VITON. However, the mask tends to preserve more clothing details and ignores persons' body information. As we can see in Figure 9, CP-VTON sometimes generates images with severe arm deformation.

In order to preserve the features of both human body and clothing, we propose to generate a new segmentation map of the person wearing the target clothing before the final image is synthesized. Hence, the final image is guided by the generated segmentation map, rather than relying solely on pose map. The ground truth of the mask is the warped clothing segmentation map after removing the arm parts. As a result, VTNFP can not only preserve a person's complete body information, but also retain the details of clothing, leading to a significant performance gain against VITON and CP-VTON.

4.4. Quantitative Results

To further evaluate the performance of our model, we conducted a user perception study. In this study, we designed an A/B test to compare the quality of the images synthesized by VTNFP over the images synthesized by either VITON or CP-VTON.

Method	Human	Method	Human
VITON	32.13%	CP-VTON	22.62%
VTNFP	67.87%	VTNFP	77.38%

Table 1. Quantitative evaluation of different methods.

We recruited 80 volunteers, and presented them with 500 groups of testing data, with each group consisting of four images - inference image, target clothing, VTNFP result, and VITON result (or CP-VTON result). Each volunteer was randomly assigned 50 groups of testing data, and was asked to choose the synthetic image in each group that he/she thinks have better quality.

In the A/B test conducted between VTNFP and VITON, 67.87% of the images generated by VTNFP were chosen by the volunteers to have a better quality. In the A/B test conducted between VTNFP and CP-VTON, 77.38% of the images generated by VTNFP were chosen by the volunteers (Table 1). These randomized tests confirm the qualitative results shown the previous section, demonstrating that VTNFP performs significantly better than previous models.

5. Conclusion

We have presented a new method for image-based virtual try-on applications. Our model follows a three-stage design strategy by first generating warped clothing, followed by generating a body segmentation map of the person wearing the target clothing, and ending with a try-on synthesis module to fuse together all information for a final image synthesis. We introduced several methodological innovations to improve the quality of image synthesis, and demonstrated that our method is able to generate substantially better realistic looking virtual try-on images than the state-of-the-art methods.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (61672148), the Program for Liaoning Innovative Research Team in University (LT2016007), and the Fundamental Research Funds for the Central Universities (N182608004, N171702001, N171604016).

References

[1] David Berthelot, Thomas Schumm, and Luke Metz. Began: boundary equilibrium generative adversarial networks. *arXiv preprint arXiv:1703.10717*, 2017.

[2] Guido Borghi, Riccardo Gasparini, Roberto Vezzani, and Rita Cucchiara. Embedded recurrent network for head pose estimation in car. In *Proceedings of the 28th IEEE Intelligent Vehicles Symposium*, 2017.

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

[4] Wenzheng Chen, Huan Wang, Yangyan Li, Hao Su, Zhenhua Wang, Changhe Tu, Dani Lischinski, Daniel Cohen-Or, and Baoquan Chen. Synthesizing training images for boosting human 3d pose estimation. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 479–488. IEEE, 2016.

[5] François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[6] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015.

[7] Hao-Shu Fang, Guansong Lu, Xiaolin Fang, Jianwen Xie, Yu-Wing Tai, and Cewu Lu. Weakly and semi supervised human body part parsing via pose-guided knowledge transfer. *arXiv preprint arXiv:1805.04310*, 2018.

[8] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 932–940, 2017.

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[10] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. *ACM Trans. Graph.*, 31(4):35–1, 2012.

[11] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7543–7552. IEEE, 2018.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

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

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

[15] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

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

[17] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 6, 2017.

[18] Xiaodan Liang, Liang Lin, Wei Yang, Ping Luo, Junshi Huang, and Shuicheng Yan. Clothes co-parsing via joint image segmentation and labeling with application to clothing retrieval. *IEEE Transactions on Multimedia*, 18(6):1175–1186, 2016.

[19] Tsung-Yi Lin, Priyal Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE transactions on pattern analysis and machine intelligence*, 2018.

[20] Yifan Liu, Zengchang Qin, Zhenbo Luo, and Hua Wang. Auto-painter: Cartoon image generation from sketch by using conditional generative adversarial networks. *arXiv preprint arXiv:1705.01908*, 2017.

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

[22] Andrew L Maas, Awni Y Hannun, and Andrew Y Ng. Rectifier nonlinearities improve neural network acoustic models. In *Proc. icml*, volume 30, page 3, 2013.

[23] Amira Ben Mabrouk and Ezzeddine Zagrouba. Abnormal behavior recognition for intelligent video surveillance systems: A review. *Expert Systems with Applications*, 91:480–491, 2018.

[24] Augustus Odena, Vincent Dumoulin, and Chris Olah. Deconvolution and checkerboard artifacts. *Distill*, 1(10):e3, 2016.

[25] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org, 2017.

[26] Ozan Oktay, Jo Schlemper, Loic Le Folgoc, Matthew Lee, Mattias Heinrich, Kazunari Misawa, Kensaku Mori, Steven McDonagh, Nils Y Hammerla, Bernhard Kainz, et al. Attention u-net: Learning where to look for the pancreas. *arXiv preprint arXiv:1804.03999*, 2018.

[27] Guim Perarnau, Joost van de Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

[28] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (TOG)*, 36(4):73, 2017.

[29] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

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

[31] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Genera-

tive adversarial text to image synthesis. *arXiv preprint arXiv:1605.05396*, 2016.

[32] Gregory Rogez, Philippe Weinzaepfel, and Cordelia Schmid. Lcr-net: Localization-classification-regression for human pose. In *CVPR 2017-IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[33] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling deep image synthesis with sketch and color. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2017.

[34] Hosnieh Sattar, Gerard Pons-Moll, and Mario Fritz. Fashion is taking shape: Understanding clothing preference based on body shape from online sources. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 968–977. IEEE, 2019.

[35] Masahiro Sekine, Kaoru Sugita, Frank Perbet, Björn Stenger, and Masashi Nishiyama. Virtual fitting by single-shot body shape estimation. In *Int. Conf. on 3D Body Scanning Technologies*, pages 406–413. Citeseer, 2014.

[36] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1225–1233. IEEE, 2017.

[37] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6924–6932, 2017.

[38] Gül Varol, Javier Romero, Xavier Martin, Naureen Mahmood, Michael J Black, Ivan Laptev, and Cordelia Schmid. Learning from synthetic humans. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, pages 4627–4635. IEEE, 2017.

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

[40] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 1, page 4, 2018.

[41] Wenqi Xian, Patsorn Sangkloy, Varun Agrawal, Amit Raj, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Texturegan: Controlling deep image synthesis with texture patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2018.

[42] Kota Yamaguchi, M Hadi Kiapour, and Tamara L Berg. Paper doll parsing: Retrieving similar styles to parse clothing items. In *Proceedings of the IEEE international conference on computer vision*, pages 3519–3526, 2013.

[43] Shan Yang, Tanya Ambert, Zherong Pan, Ke Wang, Licheng Yu, Tamara Berg, and Ming C Lin. Detailed garment recovery from a single-view image. *arXiv preprint arXiv:1608.01250*, 2016.

[44] Xi Yin, Xiang Yu, Kihyuk Sohn, Xiaoming Liu, and Manmohan Chandraker. Towards large-pose face frontalization in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3990–3999, 2017.

[45] Donggeun Yoo, Namil Kim, Sunggyun Park, Anthony S Paek, and In So Kweon. Pixel-level domain transfer. In *European Conference on Computer Vision*, pages 517–532. Springer, 2016.

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

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

[48] Junbo Zhao, Michael Mathieu, and Yann LeCun. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126*, 2016.

[49] Jun-Yan Zhu, Philipp Krähenbühl, Eli Shechtman, and Alexei A Efros. Generative visual manipulation on the natural image manifold. In *European Conference on Computer Vision*, pages 597–613. Springer, 2016.

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