Towards Detailed Characteristic-Preserving Virtual Try-On

Sangho Lee*
Seoul National Univ.
sangho.lee@snu.ac.kr

Seoyoung Lee*
Seoul National Univ.
seoyoung1215@snu.ac.kr

Joonseok Lee
Seoul National Univ.
joonseok@snu.ac.kr

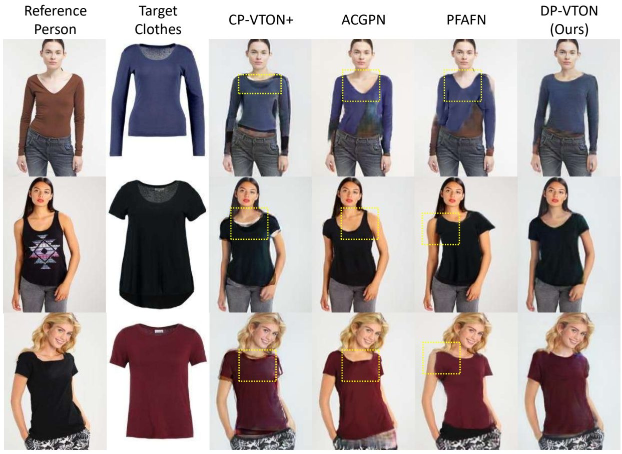

Figure 1. DP-VTON, our proposed method, better preserves the details of the reference person and the target clothes.

Abstract

While virtual try-on has rapidly progressed recently, existing virtual try-on methods still struggle to faithfully represent various details of the clothes when worn. In this paper, we propose a simple yet effective method to better preserve details of the clothing and person by introducing an additional fitting step after geometric warping. This minimal modification enables disentangling representations of the clothing from the wearer, hence we are able to preserve the wearer-agnostic structure and details of the clothing, to fit a garment naturally to a variety of poses and body shapes. Moreover, we propose a novel evaluation framework applicable to any metric, to better reflect the semantics of clothes fitting. From extensive experiments, we empirically verify that the proposed method not only learns to disentangle clothing from the wearer, but also preserves details of the clothing on the try-on results.

1. Introduction

The objective of the virtual try-on task is to fit an image of a garment to an image of a person wearing another garment. Most existing methods, such as VITON [7], CP-VTON+ [13], ACGPN [20] and PF-AFN [5], approach virtual-try-on as an image inpainting problem. Specifically, these models attempt to fit in an image of a new garment onto the torso region of a person wearing another set of clothing. The models generally involve two major steps: 1) a Geometric Warping Module to learn how clothes should be geometrically warped to fit in the pose and body shape of the target person, and 2) a Try-on Module to blend the warped clothing with the target person image.

Although previous methods can output images that looks natural, we observe that they often fail to reflect how the input clothes should be worn naturally considering all the fine details of clothed garments, without fully understanding the semantics of wearing them. Fig. 2 shows four examples from current state-of-the-art models, ACGPN [20] (Fig. 5, 10) and PF-AFN [5] (Fig. 6). We observe that some parts that are invisible when worn (*e.g.*, inner side of the shirt neckline) are still shown in (b, d), while some other parts that should be represented in the outputs (*e.g.*, spaghetti straps in (a), high neck in (c)) are not retained. Results of other models [13, 18] also show similar limitations of misrepresenting important details of the target clothes, and often struggle to generate a well-fitted image. This implies that previous try-on models might simply be fitting the target garment on top of the target person's torso, without fully understanding how the garment is actually worn tridimensionally. In other words, the learned features of the clothing and the wearer are not fully disentangled, and thus those models frequently fail to adequately select and preserve details of the target clothes, especially when they are significantly different from the source clothes. Even when such details are retained, models face difficulty in accurately fitting the clothing on the person [5].

An ideal virtual try-on model should be able to separate signals from each independent factor involved in try-on by fully understanding their semantics and transformations, so that it can generate an image that preserves details of wearing behavior. To address the problems mentioned above, we propose a simple but effective way to disentangle the learning of clothes from that of the wearer. Specifically, we propose to insert after the geometric warping, an additional step called the Clothes Fitting Module (CFM), which learns

*These authors contributed equally.

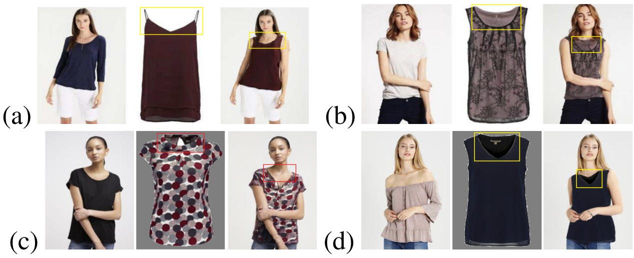

Figure 2. Examples of incorrect drawing of the target clothes by existing methods. Each image shows a set of reference image, target clothes, and try-on output. All of these examples are brought from ACGPN [20] (a, b, c) and from PF-AFN [5] (d).

how the clothes should be naturally worn completely independent of the input reference image. As opposed to previous models where the reference image (wearing the source clothes) is directly referred to perform warping, CFM fills the target clothes within the mask of the already warped target clothes, learning how they should appear when worn by the given person. As long as the backbone model follows the common two-step approach of warping and try-on, the CFM can be easily incorporated to fit the warped clothes image after the first step with minimal extra overhead.

Our contributions can be summarized as follows. First, we propose a novel 'Clothes Fitting Module (CFM)', which imitates the human behavior of wearing clothes. By clearly separating the geometric warping and inpainting of clothes before blending with the person, the proposed method successfully disentangles representation of the clothes and that of the wearer in the reference image. Second, we propose a novel way of applying evaluation metrics more suitable for the virtual try-on task, focusing on a few critical body points instead of equally weighting all pixels. Lastly, we empirically verify that the proposed approach produces try-on images of higher quality, outperforming several recent state-of-the-art methods both qualitatively and quantitatively.

2. Related Work

Research on virtual try-on is rooted in studies on fashion editing [6, 12, 16, 23]. 2D deep-learning based virtual try-on models can be categorized into whether they emphasize the use of pose and person representations [4, 7, 11, 13, 18] or segmentation maps [3, 5, 8, 20, 21]. Models generally follow two sequential stages proposed by CP-VTON [18], where clothes are first warped using the Geometric Matching Module (GMM), then dressed to the target person using the Try-on Module (TOM). CP-VTON+ [13] improved the geometric warping process with regularization to prevent extreme distortion of the clothes. However, with the limitation in paired datasets of in-shop clothes and human models, previous models do not learn fully disentangled representations for the target clothes and reference person, despite recent efforts to tackle the issue [10, 12, 14]. Generating high-resolution images is another active area of research; *e.g.*, VITON-HD [1] and FE-GAN [2] tackled this problem.

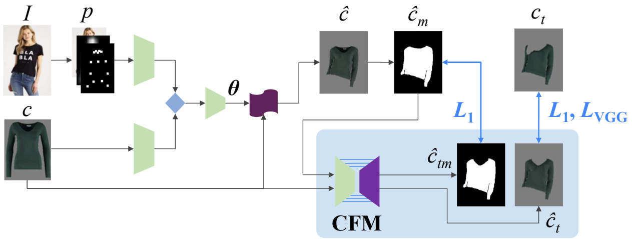

Figure 3. Overview of the Geometric Matching Module (GMM) and Clothes Fitting Module (CFM). Reference image I is preprocessed into person representation p. In-shop cloth c and p are fed into GMM. Then, the CFM takes warped clothes mask $\hat{c}_m$ and c as input and produces fitted clothes mask $\hat{c}_{tm}$ and fitted clothes on person $\hat{c}_t$, comparing to ground truth clothes on person c_t.

3. Problem Formulation

Virtual try-on task takes two inputs, an image $c \in \mathbb{R}^{h' \times w' \times 3}$ of an in-shop clothes and a reference image $I \in \mathbb{R}^{h \times w \times 3}$ of the target person, wearing another garment called source clothes. The goal of this task is generating an image $I_t \in \mathbb{R}^{h \times w \times 3}$, where the person in I wears the target clothes in c. Qualitatively, an ideal virtual try-on model should output a natural photo-like image, preserving the identity of the target person (*e.g.*, appearance, body shape, and pose), properties of the target clothes (*e.g.*, shape and texture), and interactions between them (*e.g.*, how specific parts of clothes or body should appear when clothed).

A training example of index i consists of a pair of images $(c^{(i)}, I^{(i)})$, and the model produces $\hat{I}_t^{(i)}$. We need the ground truth $I_t^{(i)}$ in a supervised setting, but in practice, it is tricky to have a pair of pictures of a model wearing two different garments with exactly the same pose. Thus, previous virtual try-on models have used $I^{(i)}$ wearing the same clothes in $c^{(i)}$, and we follow the same approach in this paper. At inference, a query $(c^{(i)}, I^{(i)})$ usually contains two different garments in $c^{(i)}$ and $I^{(i)}$, where $c^{(i)}$ is the target clothes and $I^{(i)}$ shows a person wearing the source clothes, different from $c^{(i)}$.

4. The Proposed Method: DP-VTON

Our work is highly inspired by the evolutionary achievements of VITON methods [13, 18]. We use the TOM of CP-VTON+ without modification. Ideally, the roughly warped clothes $\hat{c}$ by GMM should be synthesized with the person, keeping the wearer's attributes (*e.g.*, identity, body shape, and pose) only, independent of the garments she was wearing. However, we observe from Fig. 2 that the previous methods often retain some characteristics of the source clothes, worn by the person in I. This indicates that the characteristics of the person and those of source clothes are not completely disentangled. Our hypothesis is that this is because of the training scheme, where we use the same clothes in I and in c, due to the reason mentioned in Sec. 3.

To resolve this, we introduce the Clothes Fitting Module (CFM), inserted between the GMM and the TOM. As illustrated in the blue box of Fig. 3, we use another network that learns to fit, instead of directly using the imperfectly warped clothes $\hat{c}$ in the TOM. This CFM takes the warped clothes mask $\hat{c}_m$ and the initial target clothes image c as input, and learns to do two things: 1) estimate the mask of the target clothes $\hat{c}_{tm}$, and 2) generate the clothes image $\hat{c}_t$, both when they are actually worn by the target person.

Specifically, we first get the warped clothes mask $\hat{c}_m \in \mathbb{R}^{h \times w \times 1}$ by applying the same learned θ to the mask of c provided in the training data, instead of $\hat{c}$. The CFM consists of an encoder-decoder structure (we use U-Net [15], but other encoder-decoder networks can also be used), mapping the warped clothes mask $\hat{c}_m$ and in-shop clothes image c to the fitted clothes image $\hat{c}_t \in \mathbb{R}^{h \times w \times 3}$ and its mask $\hat{c}_{tm} \in \mathbb{R}^{h \times w \times 1}$. The generated $\hat{c}_t$ is trained to be close to the ground truth clothes image on the target person (c_t), and the fitted mask $\hat{c}_{tm}$ is trained to preserve the geometric warping in $\hat{c}_m$. We apply L_1 loss for both, and additionally we apply the VGG perceptual loss $\mathcal{L}_{\text{VGG}}$ [9] between $\hat{c}_t$ and c_t. Overall, our loss function is composed of three terms:

$$\mathcal{L} = \lambda_{\text{m}} \|\hat{c}_{tm} - \hat{c}_m\|_1 + \lambda_1 \|\hat{c}_t - c_t\|_1 + \lambda_{\text{v}} \mathcal{L}_{\text{VGG}}(\hat{c}_t, c_t),$$

where λ_{m}, λ_1, and λ_{v} are coefficients controlling relative importance of each term. We name our three-step model consisting of GMM, CFM, and TOM as **Details-Preserving Virtual Try-On (DP-VTON)**.

Discussion. How does the CFM help disentangle the source clothes from the person? In existing models without CFM, GMM is fully in charge of generating the warped clothes. The GMM, initially proposed by CP-VTON to learn the geometric gap between the clothes in c and I, is in-nature imperfect, as it maps a 2D image to another 2D image, projecting 3D clothes from different angles. As input c is already reduced to a 2D image, it is challenging for the GMM to estimate the 3D structure of the clothes. It does some level of inference on 3D structure, but as it refers to the source clothes mask of I, information about the source clothes is not completely ignored. This may look okay at training since each training example is a pair with the same clothes, but this entanglement results in lower quality of images at inference, which uses different clothes images on I and c.

With the CFM, however, the GMM is now only in charge of learning the *geometric* warping to generate a roughly warped clothes mask $\hat{c}_m$. That is, the incompletely warped clothes $\hat{c}$ is abandoned, and the CFM generates the clothes on a person $\hat{c}_t$, relying only on $\hat{c}_m$, completely independent of input reference image I. By explicitly separating geometric transformation and inpainting of the clothes, our approach disentangles information from the source clothes more robustly.

5. Experimental Settings

Dataset. We conduct experiments on the VITON dataset [7], the most commonly used one for virtual try-on which contains 14,221 pairs for training and 2,032 for testing. Each pair consists of a frontal image of a top clothing (c) and an image of a front-view person wearing the clothes (I). For quantitative evaluation, we use the same clothes for the clothes image (c) and the reference image (I), similarly to the training, as it requires ground truth.

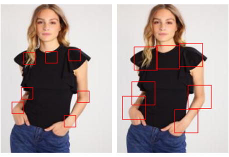

Figure 4. Key points used in patch-based `Metrics`.

Quantitative Metrics. Unlike general image synthesis, it is particularly crucial to naturally fit the clothes to each body part in virtual try-on. Existing metrics only consider how the generated images are similar to the original ones at pixel or feature level *overall*. However, we claim that these metrics cannot adequately measure the quality of how well clothes are fitted on a person, and propose a novel way to apply these metrics in a more suitable for virtual try-on. Specifically, we propose to measure the quality of the generated images only around k important body parts ('key points') of size $\epsilon \times \epsilon$ using an existing metric and average them to judge how well the clothes are fitted. Formally, we define a **patch-based** `Metric` of an image I with patch size ϵ, denoted by $\texttt{Metric}^p_\epsilon(I)$ as follows:

$$\texttt{Metric}^p_\epsilon(I) = \frac{1}{k} \sum_{i=1}^{k} \texttt{Metric}^{\text{all}} \Big(I \Big[x_i\text{-}\frac{\epsilon}{2} : x_i\text{+}\frac{\epsilon}{2}, y_i\text{-}\frac{\epsilon}{2} : y_i\text{+}\frac{\epsilon}{2} \Big] \Big),$$

where I is an image to be evaluated, (x_i, y_i) is the i-th key point, k is the number of pre-defined key points, and ϵ is the number of pixels included in each axis around a key point. We denote the regular `Metric` taken over the entire I as $\texttt{Metric}^{\text{all}}$ to distinguish it from the proposed $\texttt{Metric}^p_\epsilon$.

`Metric` can be any existing metric. We adopt `SSIM` [19], `LPIPS` [22], and pixel-wise `MSE` to measure the similarity (or distance) between generated images and ground truth. We report `LPIPS` scores based on VGG [17].

We choose 7 important joints in the torso region of the human body as key points, illustrated in Fig. 4: the neck, both sides of the shoulders, elbows, and wrists. This specific setting may be flexibly adjusted for a different task, *e.g.*, including knees, ankles, or feet for a full-body virtual try-on. We use $\epsilon = \{10, 20, 40, 60\}$ for `SSIM` and `MSE`, while we drop $\epsilon = 10$ for `LPIPS` since a 10×10 image patch is not sufficiently large to perform inference on VGG.

Implementation Details. Our GMM and TOM are built on

Method	CP-VTON+	ACGPN	PF-AFN	DP-VTON (ours)
$\texttt{SSIM}^{\text{all}}$	0.368	0.387	**0.511**	0.392
$\texttt{SSIM}^{p}_{10}$	0.805	0.361	0.811	**0.847**
$\texttt{SSIM}^{p}_{20}$	0.531	0.231	0.582	**0.589**
$\texttt{SSIM}^{p}_{40}$	0.549	0.249	0.599	**0.604**
$\texttt{SSIM}^{p}_{60}$	0.577	0.279	0.627	**0.628**
$\texttt{LPIPS}^{\text{all}}$	0.082	**0.066**	0.077	0.075
$\texttt{LPIPS}^{p}_{40}$	0.231	0.485	0.202	**0.197**
$\texttt{LPIPS}^{p}_{50}$	0.230	0.478	0.200	**0.198**
$\texttt{LPIPS}^{p}_{60}$	0.230	0.475	0.199	**0.197**
$\texttt{MSE}^{\text{all}}$	1874.4	18703.5	2192.5	**1394.9**
$\texttt{MSE}^{p}_{10}$	7.0	53.4	9.3	**4.6**
$\texttt{MSE}^{p}_{20}$	27.6	211.7	36.8	**18.7**
$\texttt{MSE}^{p}_{40}$	103.5	819.9	136.4	**71.6**
$\texttt{MSE}^{p}_{60}$	214.3	1767.1	275.6	**149.8**

Table 1. Quantitative comparisons to state-of-the-art models.

CFM inputs	$\texttt{SSIM}^{p}_{20}(\uparrow)$	$\texttt{LPIPS}^{p}_{20}(\downarrow)$	$\texttt{MSE}^{p}_{20}(\downarrow)$
Warped clothes mask ($\hat{c}_m$)	**0.589**	**0.198**	**18.7**
Warped clothes ($\hat{c}$)	0.414	0.275	45.3
Both $\hat{c}$ and $\hat{c}_m$	0.449	0.244	36.7

Table 2. Comparison on various CFM input configurations.

top of CP-VTON+ [13]. For training GMM, a similar setting in the original paper is used, *i.e.*, λ_1, λ_v, λ_m = 1 and $\lambda_{\text{reg}} = 0.5$. We follow the U-Net [15] architecture for CFM, except for the final layer where we use a 3×3 convolution instead of the original 1×1 convolution. For the CFM, we use VGG loss ($\mathcal{L}_{\text{VGG}}$) and L1 loss ($\mathcal{L}_1$) for training to minimize the difference between the shape and style of the generated clothes with ground truth clothes. Then with the fitted clothes after CFM, we train the TOM module. We use Adam optimizer with $\beta_1 = 0.5$ and λ_{VGG} = 0.999. We train the model for 200K steps, with a constant learning rate of 0.0001 for the first 100K steps and linearly decay the rate to zero for the remaining 100K steps.

6. Results and Discussion

Quantitative Comparisons. Table 1 compares the scores of SSIM, LPIPS, and MSE of CP-VTON+ [13], ACGPN [20], PF-AFN [5], and our method with various patch sizes (ϵ) around the key points. Under the traditional metrics taken over the entire output image ($^{\text{all}}$), the proposed method outperforms baselines only in $\texttt{MSE}^{\text{all}}$, while PF-AFN and ACGPN perform better in SSIM and LPIPS, respectively. However, when we consider only the major joints in the torso area, we observe that our DP-VTON outperforms all other baselines in all three metrics, $\texttt{SSIM}^p$, $\texttt{LPIPS}^p$, and $\texttt{MSE}^p$ with all ϵs we tried. Putting these two facts together, we can conclude that the proposed method generates semantically and graphically more plausible try-on images near the key points that are critical to human perception (recall Fig. 4), while the baselines get better scores thanks to better matches to the ground truth outside of these critical regions.

We additionally perform an ablation study on the configuration of the CFM. After the geometric warping, CFM may take as input either or both of the warped clothes $\hat{c}$ in RGB and the warped clothes mask $\hat{c}_m$, together with the in-shop clothes image c. Table 2 compares the performance for each input setting. We observe that feeding only the mask $\hat{c}_m$ outperforms the other two. This confirms that it is indeed important to let the CFM solely learn to dress independently of the reference image I, instead of leaking information of the warped image from the GMM into the TOM.

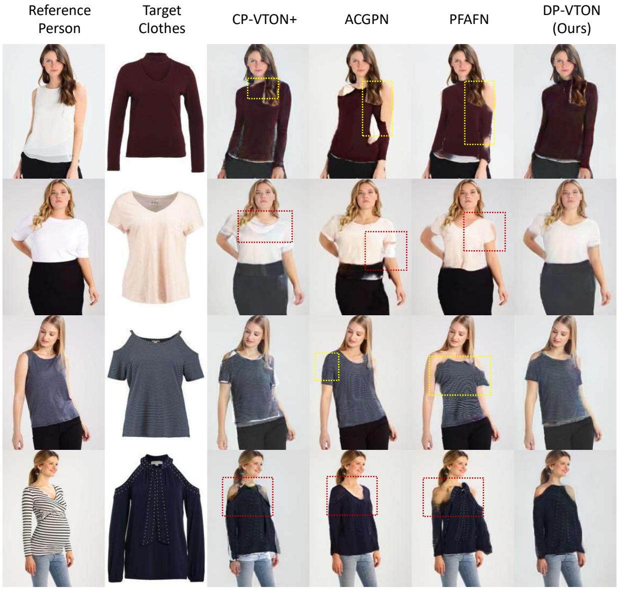

Figure 5. Qualitative comparisons.

Qualitative Analysis. We perform visual comparisons with recent state-of-the-art methods, including CP-VTON+, ACGPN, and PF-AFN. As shown in Fig. 1, the images generated by CP-VTON+ show the backside of a shirt around the neckline, and the overall color of the clothes is blended and blurred. ACGPN makes the shape of clothes look similar to the reference images, especially for the neckline and arm parts, and PF-AFN faces difficulty in handling various body shapes. In contrast, our method better preserves the characteristics of the clothes, regardless of the source clothes that the reference person wears. In the top case, for example, ACGPN and PF-AFN keep the V-neck trait mixed with the brown color, while our method dresses the blue round-neck clothes naturally without being mixed with the source clothes. For the second example, ACGPN and PF-AFN make a similar mistake, leaving the shape of the tank-top in one shoulder. CP-VTON+ preserves the characteristics of the target clothes better in this example, but there are some undesirable artifacts, such as white regions around the neckline and at the edge of a sleeve. These examples empirically verify that our proposed method better disentangles the characteristics of the person and those of source clothes.

Figure 5 illustrates additional examples with various poses and clothes. We again observe that DP-VTON faithfully expresses the detailed characteristics of the target clothes and fits well on a variety of poses and body shapes, while others show limited preservation of such details.

References

[1] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-HD: High-resolution virtual try-on via misalignment-aware normalization. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[2] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. Fashion editing with adversarial parsing learning. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[3] Benjamin Fele, Ajda Lampe, Peter Peer, and Vitomir Struc. C-VTON: Context-driven image-based virtual try-on network. In *Pro. of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 2

[4] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[5] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 4

[6] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott, and Larry S Davis. FiNet: Compatible and diverse fashion image inpainting. In *Proc. of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2019. 2

[7] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. VITON: An image-based virtual try-on network. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3

[8] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *Proc. of the European Conference on Computer Vision (ECCV)*. Springer, 2020. 2

[9] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2016. 3

[10] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. Vogue: Try-on by stylegan interpolation optimization. *arXiv e-prints*, 2021. 2

[11] Guoqiang Liu, Dan Song, Ruofeng Tong, and Min Tang. Toward realistic virtual try-on through landmark-guided shape matching. In *Proc. of the AAAI Conference on Artificial Intelligence*, 2021. 2

[12] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[13] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. CP-VTON+: Clothing shape and texture preserving image-based virtual try-on. In *CVPR Workshops*, 2020. 1, 2, 4

[14] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. Image based virtual try-on network from unpaired data. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[15] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. of the International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI)*. Springer, 2015. 3, 4

[16] Wu Shi, Tak-Wai Hui, Ziwei Liu, Dahua Lin, and Chen Change Loy. Learning to synthesize fashion textures. *arXiv:1911.07472*, 2019. 2

[17] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556*, 2014. 3

[18] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2018. 1, 2

[19] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 3

[20] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 2, 4

[21] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proc. of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2019. 2

[22] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[23] Shizhan Zhu, Raquel Urtasun, Sanja Fidler, Dahua Lin, and Chen Change Loy. Be your own prada: Fashion synthesis with structural coherence. In *Proc. of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2017. 2