Learning to Wear: Details-Preserved Virtual Try-on via Disentangling Clothes and Wearer

Sangho Lee[1]
sangho.lee@snu.ac.kr
Seoyoung Lee[1]
seoyoung1215@snu.ac.kr
Joonseok Lee*[1,2]
joonseok@snu.ac.kr

[1] Graduate School of Data Science
Seoul National Univ.
Seoul, Korea

[2] Google Research
Mountain View, CA, USA

Abstract

Virtual try-on, fitting an image of a garment to an image of a person, has rapidly progressed recently. However, 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 helps to effectively learn disentangled representations of the clothing from the wearer. By disentangling these two major components for virtual try-on, we are able to preserve the wearer-agnostic structure and details of the clothing, and thus can 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.

Figure 1: Our proposed method better preserves the details of the reference person and the target clothes, without mixing the cue of original clothes in the reference.

*corresponding author

Figure 2: Examples of incorrect drawing of the target clothes by existing methods. Images in (a, b, c) are brought from Fig. 5, 10 in ACGPN [40], and (d) from Fig. 6 in PF-AFN [8].

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 [10], CP-VTON+ [26], ACGPN [40] and PF-AFN [8], 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 (GMM) 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 (TOM) to blend the warped clothing with the target person image.

Although previous methods can output images that look 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 [40] and PF-AFN [8]. 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. Other models [26, 38] also show similar limitations of misrepresenting important details of the target clothes, often struggling to generate a well-fitted image. This implies that previous 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, 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.

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 this problem, we propose a simple but effective way to disentangle the learning of clothes from that of the wearer. Specifically, we propose DP-VTON, a three-step model where an additional step called the Clothes Fitting Module (CFM) is inserted between the GMM and TOM, aimed at learning 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 to seamlessly connect the GMM and TOM while significantly improving the results.

Our contributions can be summarized as follows. First, we propose a three-step model called DP-VTON with 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

Image synthesis. Generative Adversarial Networks (GANs) have steered the progress in the fields of image synthesis and manipulation [12, 15, 17, 18]. To generate data with certain properties, additional information (text [31], class labels [28], or attributes [35]) has been incorporated to condition the generation procedure. ClothNet [20] generated a person inpainted with different clothing styles, by learning to condition on the pose, shape, and color. Convolutional neural networks (CNNs) [19, 21, 37] also have been widely utilized in image synthesis. U-net [33], originally developed for image segmentation, has been applied to image synthesis for high performance, *e.g.*, Generative Adversarial U-Net [1].

Virtual Try-on. Research on virtual try-on is rooted in studies on fashion editing [11, 22, 25, 36, 43]. Deep-learning-based virtual try-on models are roughly classified into 3D-based models [9, 29, 32, 34] and 2D models [4, 10, 14, 38]. 3D-based models tend to result in higher accuracy in the simulated clothes, while they require additional 3D measurements and more computing resources, making 2D-based methods to be more broadly adopted. 2D models can be further categorized into whether they emphasize the use of pose and person representations (*e.g.*, [7, 10, 24, 26, 38]) or segmentation maps (*e.g.*, [6, 8, 13, 40, 41]). Models generally follow two sequential stages proposed by CP-VTON [38], where clothes are first geometrically warped, then dressed to the target person. CP-VTON+ [26] improved the geometric warping process with regularization to prevent extreme distortion of the clothes. A few recent models [23, 25, 27] have attempted to refine these models to learn disentangled representations for the target clothes and reference person. However, due to the limitation in paired datasets of in-shop clothes and human models, these models were unable to learn fully disentangled representations. Moreover, recent works have expanded virtual try-on research into generating high-resolution images [2, 5], dressing multiple garments sequentially [3, 27, 30], and transferring garments between two people [23, 25, 27].

3 Preliminary

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. (Note that it is not necessarily $h = h'$ and $w = w'$, respectively.) The

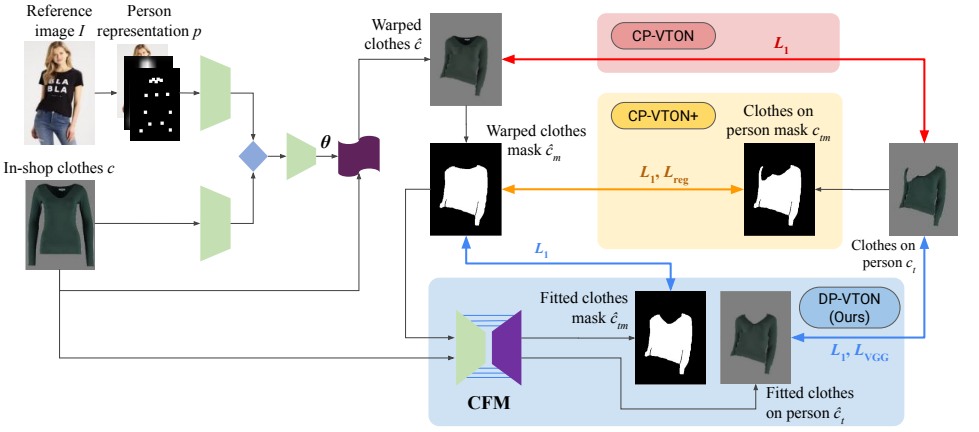

Figure 3: Our model (blue box), compared with CP-VTON [38] and CP-VTON+ [26].

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, existing 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 $c^{(i)}$.

CP-VTONs. Our work is inspired by the evolutionary achievements of VITON methods. Advancing from VITON [10], CP-VTON [38] proposed a two-stage approach, first warping the clothes using the Geometric Matching Module (GMM) and then dressing them to the target person using the Try-on Module (TOM).

Fig. 3 overviews the GMM proposed by CP-VTON. The reference image $I \in \mathbb{R}^{h\times w\times 3}$ is first pre-processed to a head image $H \in \mathbb{R}^{h\times w\times 3}$, the person mask $M \in \{0,1\}^{h\times w}$, and a pose map $P \in \mathbb{R}^{h\times w\times 18}$, where each layer of the pose map is a one-hot encoding indicating each pre-defined key point, *e.g.*, shoulder, elbow, etc. These preprocessed features are stacked to the person representation $p \in \mathbb{R}^{h\times w\times 22}$. The GMM geometrically transforms the target clothes c to a warped clothes $\hat{c}$ such that it is roughly aligned with the person in I, via a Thin-Plate Spline (TPS) transformation module T that warps c into $\hat{c} = T_\theta(c)$. (See CP-VTON [38] for more details.) GMM can be trained end-to-end by comparing the warped clothes ($\hat{c}$) and the actual clothes on person (c_t). CP-VTON uses the pixel-wise L_1 loss between them; that is, $\mathcal{L}_{\text{GMM}} = \|\hat{c} - c_t\|_1$. CP-VTON+ [26] improves CP-VTON by comparing the mask of the warped clothes ($\hat{c}_m$) and the clothes mask of the reference image (c_{tm}) instead of the RGB images ($\hat{c}$ and c_t), as shown in the yellow box of Fig. 3, and by applying regularization on the TPS parameters. After geometric warping, the TOM takes as input the warped clothes $\hat{c}$, which is roughly aligned with the body shape and the pose of the target person, to synthesize the final result by fusing $\hat{c}$ with the target person.

4 The Proposed Method: DP-VTON

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 undesirable phenomenon 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.

What is happening when we train the GMM? Simply speaking, GMM learns to map the frontal view of clothes c to their distorted shape according to the person's body and pose in the reference image I, assuming the clothes will be worn by the person. CP-VTON and CP-VTON+ assume that this is mainly geometric conversion, but in fact, this conversion includes more than that. For instance, some area in the frontal view actually belongs to the backside of the clothes (*e.g.*, above the neckline, as in Fig. 2(b, d)), so this area should not appear in the warped image. Since the GMM is designed to solely learn geometric transformation, however, the warped clothes image $\hat{c}$ often fails to preserve these kinds of fine details required when wearing clothes, and sometimes even the general characteristics of the target clothes.

To resolve this issue, we introduce DP-VTON, where the Clothes Fitting Module (CFM) is inserted between the GMM and 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. 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.

At a glance, this might look redundant, since the GMM is supposed to produce this directly from c. However, from the existing models, we realize that the GMM is not sufficient to model the natural details of the clothes when they are worn. As the input $\hat{c}_m$ provides the geometrically transformed mask this time, however, the CFM concentrates purely on "how to wear". In other words, the CFM is now completely independent of the source clothes in I, using the mask of the warped clothes $\hat{c}_m$ rather than the reference image I directly.

Specifically, we first get the warped clothes mask $\hat{c}_m \in \{0,1\}^{h\times w}$ by applying the same learned θ to the mask of c, which is provided in the training data, instead of $\hat{c}$. The CFM consists of an encoder-decoder structure (we use a U-Net [33], but other encoder-decoder networks can be used as well), mapping the warped clothes mask $\hat{c}_m$ and the in-shop clothes image c to the fitted clothes image $\hat{c}_t \in \mathbb{R}^{h\times w\times 3}$ and its mask $\hat{c}_{tm} \in \{0,1\}^{h\times w}$. 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 $\mathcal{L}_{\text{VGG}}$ [16] between $\hat{c}_t$ and c_t. Overall, our loss function is composed of three terms, with λ_{mask}, λ_{L1}, and λ_{VGG} to control the relative importance of each term:

$$\mathcal{L}_{\text{ours}} = \lambda_{\text{mask}} \cdot \|\hat{c}_{tm} - \hat{c}_m\|_1 + \lambda_{L1} \cdot \|\hat{c}_t - c_t\|_1 + \lambda_{\text{VGG}} \cdot \mathcal{L}_{\text{VGG}}(\hat{c}_t, c_t). \quad (1)$$

Discussion. How does the CFM help to disentangle the source clothes from the person? In the existing models without CFM, the GMM is fully in charge of generating the warped clothes. The GMM, however, is in-nature imperfect, in that it maps a 2D image to another 2D image, projecting 3D clothes from different angles. As the 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 since it refers to the source clothes mask of I, information about the source clothes is not completely ignored. This might 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 CFM no longer directly uses the incompletely warped clothes $\hat{c}$ made by the GMM, but rather generates the clothes on a person $\hat{c}_t$, relying only on the *binary mask* $\hat{c}_m$ of the warped clothes attained from $\hat{c}$, completely independent of the *2D RGB* input reference image I. By explicitly separating the learning process of geometric transformation and inpainting of the clothes, our approach disentangles information from the source clothes more robustly.

In other words, the GMM in our model learns only about the person (pose and body shape) from I, by using I **only as a source** for the person's identifiable traits. By ignoring the warped clothes image $\hat{c}$ produced by the GMM but keeping only its mask $\hat{c}_m$, our method drops undesirable information coming from the source clothes. The CFM, on the other hand, uses I **only as the ground truth** image. As the body shape and pose is provided with $\hat{c}_m$, it concentrates only on inpainting c within the mask, guided by c_t extracted from I as the ground truth. After learning the geometric warping of the clothes in GMM and RGB visualization of the warped clothes in CFM, our model continues to its third step of TOM to integrate the output from CFM with the target person.

The GMM was initially proposed by CP-VTON to learn the geometric gap between the clothes in c and I. As the person in I already wears the target clothes in c at training, however, what the GMM actually learns is a combination of how the clothes look when a person wears them as well as the geometric difference between c and I. Without the CFM, CP-VTONs use the reference image I as both the source and ground truth at the same time, thereby confusing the model to entangle information from source and target clothes (again, which are the same at training). We conduct extensive experiments in Sec. 5 to verify this claim.

5 Experiments

Dataset. We conduct experiments on the VITON dataset [10], containing 14,221 samples for training and 2,032 for testing. Each sample is a pair of a frontal image of a top clothing (c) and an image of a front-view person wearing the clothes (I). Image resolution is 256×192 both for c and 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.

Baselines. We compare our proposed method against three state-of-the-art baselines, including CP-VTON+ [26], ACGPN [40], and PF-AFN [8]. We expect the proposed approach to improve VITON-HD [2], another state-of-the-art virtual try-on with high-resolution images, but we do not compare against this model as its training dataset is not publicly available. Applying the proposed idea to high-resolution images will be a promising future work.

Quantitative Metrics. We use Structural SIMilarity (`SSIM`) [39], Learned Perceptual Image Patch Similarity (`LPIPS`) [42], and pixel-wise Mean-Squared Error (`MSE`) to measure the similarity (or distance) between generated images and ground truth. `LPIPS` measures a semantic distance between two images based on embeddings extracted from a pre-trained network (we use VGG [37]). `LPIPS` scores based on AlexNet [19] show a similar pattern, available in the Supplementary Materials.

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

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

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

$$\texttt{Metric}^p_\varepsilon(I) = \frac{1}{k}\sum_{i=1}^{k} \texttt{Metric}\left(I\left[x_i - \frac{\varepsilon}{2} : x_i + \frac{\varepsilon}{2}, y_i - \frac{\varepsilon}{2} : y_i + \frac{\varepsilon}{2}\right]\right), \tag{2}$$

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, ε is the number of pixels to be included in each axis around the key point. `Metric` can be any existing metric above. The traditional way of using the entire image is a special case, where $\varepsilon = \infty$.

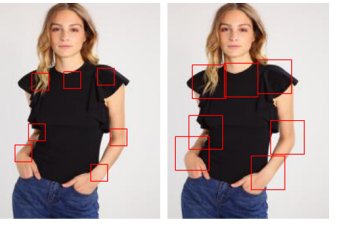

Figure 4: Key points used in patch-based `Metrics`.

We choose as key points 7 important joints (the neck, both sides of the shoulders, elbows, and wrists) illustrated in Fig. 4. 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 $\varepsilon = \{10, 20, 40, 60\}$ for `SSIM` and `MSE`, while we drop $\varepsilon = 10$ for `LPIPS` since a 10×10 image patch is not sufficiently large to perform inference on VGG or AlexNet.

Implementation Details. Our GMM and TOM are built on top of CP-VTON+ [26]. For training GMM, a similar setting in the original paper is used, *i.e.*, λ_{L1}, λ_{VGG}, $\lambda_{\text{mask}} = 1$ and $\lambda_{\text{reg}} = 0.5$. We use U-Net [33] for CFM, whose full architecture is available in the Supplementary Materials. We use Adam optimizer with $\beta_1 = 0.5$ and $\lambda_{\text{VGG}} = 0.999$. 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.

5.1 Quantitative Comparisons

Overall Performance. Tab. 1 compares the scores of `SSIM`, `LPIPS`, and `MSE` of CP-VTON+, ACGPN, PF-AFN, and our method with various window sizes (ε) around the key points. Under the traditional metrics taken over the entire output image ($\varepsilon = \infty$), the proposed method outperforms baselines only in `MSE`, while PF-AFN and ACGPN perform better in

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 warped clothes ($\hat{c}$) and warped clothes mask ($\hat{c}_m$)	0.449	0.244	36.7

Table 2: Comparison on various CFM input configurations.

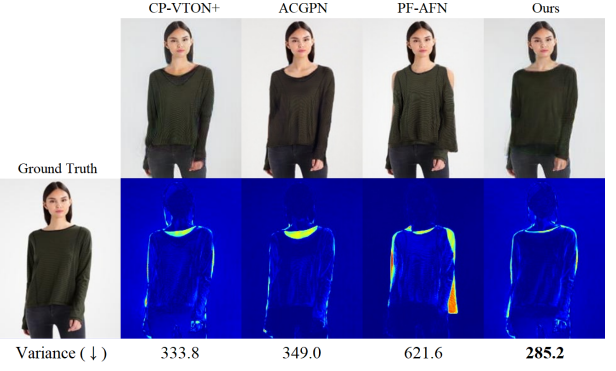

Figure 5: Pixelwise difference with ground truth.

SSIM and LPIPS, respectively. However, when we consider only around the key points, representing the major joints in the torso area, DP-VTON outperforms all other baselines in all three metrics, 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, such as the background or lower garment, which is not the main target of virtual try-on.

To further demonstrate why evaluating the selected areas is important, we visualize the pixelwise difference between the generated images and ground truth in Fig. 5. Discordant pixels are concentrated more on the target clothes area in the baselines (3 in the middle), while for our model they are more *evenly* scattered across the entire image, including the background. This verifies that the traditional scores using *all* pixels ($\varepsilon = \infty$) for baselines may look better thanks to better accordance on less important non-clothes areas, even though their try-on results are not visually superior. We also report the variance of difference across all pixels in Fig. 5, averaged over 2,000+ test images. Our model clearly shows lower variance, demonstrating its robustness and consistency.

Ablation on CFM Inputs. We present an ablation study on the configuration of the CFM. After the geometric warping, we have two warped images, the warped clothes $\hat{c}$ in RGB and the warped clothes mask $\hat{c}_m$. The CFM may take as input either or both of these, together with the in-shop clothes image c.

Table 2 compares the performance of each setting. We observe that feeding only the mask $\hat{c}_m$ outperforms the other two. In other words, directly using the warped clothes $\hat{c}$ deteriorates the overall performance. 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.

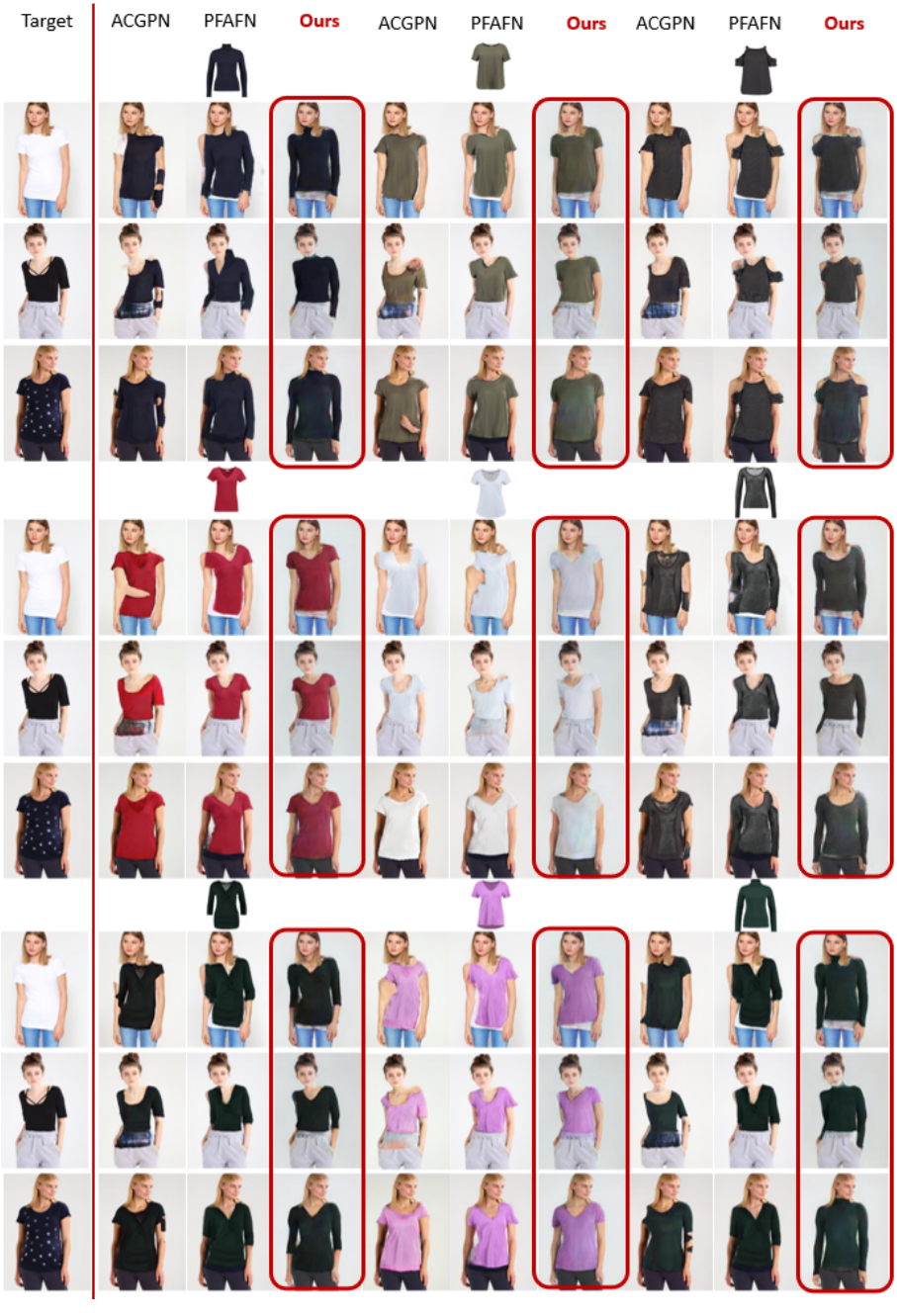

Figure 6: Qualitative results with various poses

5.2 Qualitative Analysis and User Study

Qualitative Comparison. Fig. 6 visually compares DP-VTON against baselines, CP-VTON+, ACGPN, and PF-AFN on four examples. 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 shows better results but the shape of clothes looks similar to the reference images, especially around the neckline and arm parts. PF-AFN produces more vivid images, but it also faces difficulty in handling a variety of body shapes, as shown in Fig. 6 (top-right and bottom-left). In contrast, our method better preserves the characteristics of the clothes, regardless of the source clothes that the reference person wears. In the top-left case, for example, DP-VTON dresses the blue round-neck clothes naturally without being mixed with the brown V-neck long sleeve t-shirt in the source. These examples empirically verify that DP-VTON better disentangles the characteristics of the person and those of source clothes.

Fig. 6 show additional examples with various poses. Our method dresses the target clothes human-agnostically, regardless of pose or body shape. We again observe that our method faithfully expresses the detailed characteristics of the target clothes and fits well on diverse poses and body shapes, while others show limited preservation of such details.

User Study. We additionally conducted a user study to compare the models, to reflect general human perception. We randomly sampled 200 examples from the VITON test set and divided them into two sets, 100 in each. We invited 60 volunteers who are unfamiliar with virtual try-on techniques, and randomly assigned them to either set A or B. Each question shows the reference person (I), the target clothes (c), and 4 randomly-ordered virtual try-on images generated by CP-VTON+, ACGPN, PF-AFN, and our DP-VTON. The participant was asked to choose the best one among them. (Optionally, they could leave a comment if it is hard to choose only one or if none of them is dressed properly.) More details about the user study are provided in the Supplementary Materials with examples.

Table 3 summarizes the results. The first line (participant-centric) shows the ratio of participants who select each method most frequently. That is, 65.3% of the participants answer that our method produces the best result most often. The second line shows a question-centric aggregation. For each question, one method is chosen as the best by majority vote, and the table lists the ratio of questions that each method is chosen as the best for. Our approach is chosen as the best for 76.5% (153 questions out of 200), significantly outperforming others. This result confirms that our method actually produces better quality of try-on images than existing methods in general.

Aggregation	CP-VTON+	ACGPN	PF-AFN	Ours
Participant-centric	7.2%	16.1%	11.4%	**65.3%**
Question-centric	2.5%	11.0%	10.0%	**76.5%**

Table 3: User study results on VITON

6 Summary

We propose a simple yet effective method to better preserve details of the clothing and reference person for virtual try-on. With an additional module called Clothes Fitting Module (CFM) after geometric warping, our DP-VTON model learns representations of the clothing disentangled from the human figure or identity. By disentangling these two major components of virtual try-on, the proposed method preserves the wearer-agnostic structure and details of the clothing, and thus can fit a garment naturally to a variety of poses and body shapes of the target person. Our model learns the behavior of "wearing clothes" in general, just as a person would dress up in real life. This is confirmed by quantitative and qualitative evaluations, with our novel patch-based metrics that reflect the semantics of clothes fitting.

Acknowledgement

This work was supported by the New Faculty Startup Fund from Seoul National University and by National Research Foundation (NRF) grant (No. 2021H1D3A2A03038607/25%, 2022R1C1C1010627/25%) and Institute of Information & communications Technology Planning & Evaluation (IITP) grant (No. 2022-0-00264/20%, 2021-0-01778/30%) funded by the government of Korea.

References

[1] Xiaocong Chen, Yun Li, Lina Yao, Ehsan Adeli, and Yu Zhang. Generative adversarial U-Net for domain-free medical image augmentation. *arXiv:2101.04793*, 2021.

[2] 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.

[3] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[4] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *Proc. of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2019.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

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

[10] 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.

[11] 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.

[12] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[13] 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)*, 2020.

[14] Nikolay Jetchev and Urs Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *Proc. of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV) Workshop*, 2017.

[15] Youngjoo Jo and Jongyoul Park. SC-FEGAN: Face editing generative adversarial network with user's sketch and color. In *Proc. of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2019.

[16] 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.

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

[18] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[20] Christoph Lassner, Gerard Pons-Moll, and Peter V Gehler. A generative model of people in clothing. In *Proc. of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2017.

[21] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[22] Sangho Lee, Seoyoung Lee, and Joonseok Lee. Towards detailed characteristic-preserving virtual try-on. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2022.

[23] Kathleen M Lewis, Srivatsan Varadharajan, and Ira Kemelmacher-Shlizerman. TryOnGAN: Body-aware try-on via layered interpolation. *arXiv:2101.02285*, 2021.

[24] 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.

[25] 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.

[26] 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 *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2020.

[27] 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.

[28] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier GANs. In *Proc. of the International Conference on Machine Learning (ICML)*, 2017.

[29] 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):1–15, 2017.

[30] Amit Raj, Patsorn Sangkloy, Huiwen Chang, James Hays, Duygu Ceylan, and Jingwan Lu. SwapNet: Image based garment transfer. In *Proc. of the European Conference on Computer Vision (ECCV)*. Springer, 2018.

[31] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proc. of the International Conference on Machine Learning (ICML)*, 2016.

[32] Damien Rohmer, Tiberiu Popa, Marie-Paule Cani, Stefanie Hahmann, and Alla Sheffer. Animation wrinkling: augmenting coarse cloth simulations with realistic-looking wrinkles. *ACM Transactions on Graphics (ToG)*, 29(6):1–8, 2010.

[33] 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)*, 2015.

[34] Igor Santesteban, Miguel A Otaduy, and Dan Casas. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, 2019.

[35] Wei Shen and Rujie Liu. Learning residual images for face attribute manipulation. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

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

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

[38] 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.

[39] 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.

[40] 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.

[41] 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.

[42] 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.

[43] 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.

Supplementary Material for Learning to Wear: Details-Preserved Virtual Try-on via Disentangling Clothes and Wearer

A Detailed Model Architectures

Architecture of the CFM (Clothes Fitting Module) follows the same structure of U-Net described in [11]. As described in Table i, it consists of an encoder (the left columns; from the Input to conv5b layers) and a decoder (the right columns; from upsample1 to Output layers). The encoder consists of 5 repeated blocks of two 3×3 conv-layers with stride and padding of size 1, respectively. After two convolution layers, a 2×2 max-pooling layer is applied for downsampling. For each convolution layer, we apply an instance normalization, and we use a Rectified Linear Unit (ReLU) for activation. Every layer in the decoder consists of an upsampling of the feature map with halved number of channels, a concatenation with the corresponding feature map from the encoding path, and one 3×3 convolution, each followed by a ReLU. At the final layer, a 3×3 convolution with stride and padding of size 1 is used, slightly modified from the original U-Net [11]. In total, the network contains 22 convolutional layers.

B Baselines

There are additional current state-of-the-art works that we do not compare against for the following reasons. VITON-HD [1] shows impressive quality in high-resolution images, but we do not compare it qualitatively or quantitatively because their training dataset in high-resolution is not publicly available. Dress in Order [2] is another recent work, focusing on dressing clothes in a suggested order. Since the task is different from our standard virtual try-on, this model is not comparable to our work. DCTON [3] tries to address the lack of paired data by adopting a double-cyclic architecture. Since this model is similar in architecture and shows similar performance with PF-AFN [4], we choose to compare with PF-AFN instead of this work. LM-VTON [9] tries to express the characteristics of the clothes using landmarks of clothes. As this model utilizes an additional information that is not a part of the VITON dataset, we do not choose this as a baseline for fair comparison.

Layer	Details	Output Size	Layer	Details	Output Size
Input	Clothes image ($3 \times H \times W$) Warped mask ($1 \times H \times W$)	$4 \times H \times W$	upsample1	2×2 upsampling	$1024 \times H/8 \times W/8$
conv1a	$3 \times 3 \times 64$ Instance Normalization ReLU	$64 \times H \times W$	conv6a	$3 \times 3 \times 512$ Instance Normalization ReLU	$512 \times H/8 \times W/8$
conv1b	$3 \times 3 \times 64$ Instance Normalization ReLU	$64 \times H \times W$	conv6b	$3 \times 3 \times 512$ Instance Normalization ReLU	$512 \times H/8 \times W/8$
pool1	2×2 max pool	$64 \times H/2 \times W/2$	conv6c	$3 \times 3 \times 512$ Instance Normalization ReLU	$512 \times H/8 \times W/8$
conv2a	$3 \times 3 \times 128$ Instance Normalization ReLU	$128 \times H/2 \times W/2$	upsample2	2×2 upsampling	$512 \times H/4 \times W/4$
conv2b	$3 \times 3 \times 128$ Instance Normalization ReLU	$128 \times H/2 \times W/2$	conv7a	$3 \times 3 \times 256$ Instance Normalization ReLU	$256 \times H/4 \times W/4$
pool2	2×2 max pool	$128 \times H/4 \times W/4$	conv7b	$3 \times 3 \times 256$ Instance Normalization ReLU	$256 \times H/4 \times W/4$
conv3a	$3 \times 3 \times 256$ Instance Normalization ReLU	$256 \times H/4 \times W/4$	conv7c	$3 \times 3 \times 256$ Instance Normalization ReLU	$256 \times H/4 \times W/4$
conv3b	$3 \times 3 \times 256$ Instance Normalization ReLU	$256 \times H/4 \times W/4$	upsample3	2×2 upsampling	$256 \times H/2 \times W/2$
pool3	2×2 max pool	$256 \times H/8 \times W/8$	conv8a	$3 \times 3 \times 128$ Instance Normalization ReLU	$128 \times H/2 \times W/2$
conv4a	$3 \times 3 \times 512$ Instance Normalization ReLU	$512 \times H/8 \times W/8$	conv8b	$3 \times 3 \times 128$ Instance Normalization ReLU	$128 \times H/2 \times W/2$
conv4b	$3 \times 3 \times 512$ Instance Normalization ReLU Dropout (0.5)	$512 \times H/8 \times W/8$	conv8c	$3 \times 3 \times 128$ Instance Normalization ReLU	$128 \times H/2 \times W/2$
pool4	2×2 max pool	$512 \times H/16 \times W/16$	upsample4	2×2 upsampling	$128 \times H \times W$
conv5a	$3 \times 3 \times 1024$ Instance Normalization Relu	$1024 \times H/16 \times W/16$	conv9a	$3 \times 3 \times 64$ Instance Normalization ReLU	$64 \times H \times W$
conv5b	$3 \times 3 \times 1024$ Instance Normalization ReLU Dropout(0.5)	$1024 \times H/16 \times W/16$	conv9b	$3 \times 3 \times 64$ Instance Normalization ReLU	$64 \times H \times W$
			conv9c	$3 \times 3 \times 64$ Instance Normalization ReLU	$64 \times H \times W$
			Output	$3 \times 3 \times 64$ Instance Normalization ReLU	$4 \times H \times W$

Table i: Full architectural details of the CFM.

(↓)	LPIPS$^{\text{all}}$	LPIPS$^{p}_{20}$	LPIPS$^{p}_{25}$	LPIPS$^{p}_{30}$
CP-VTON+	0.133	0.067	0.114	0.115
ACGPN	0.406	0.237	0.349	0.358
PF-AFN	0.123	0.060	0.102	0.102
Ours	**0.116**	**0.054**	**0.095**	**0.095**

Table ii: Quantitative comparisons to state-of-the-art models with LPIPS using a pretrained AlexNet [8].

CFM inputs	LPIPS$^{p}_{20}$(↓)
Warped clothes mask ($\hat{c}_m$)	**0.054**
Warped clothes ($\hat{c}$)	0.147
Both warped clothes ($\hat{c}$) and warped clothes mask ($\hat{c}_m$)	0.124

Table iii: Comparison on various CFM input configurations with LPIPS using a pretrained AlexNet [8].

C Additional Evaluation Results

Table ii compares LPIPS$_{\text{alex}}$ with other baselines, and Table iii compares LPIPS$_{\text{alex}}$ with various CFM input configurations, namely the warped clothes $\hat{c}$ in RGB and/or the warped clothes mask $\hat{c}_m$. The CFM may take as input either or both of these, together with the in-shop clothes image c.

Overall, we observe a similar pattern in both tables to the results in Table 1 in the main manuscript, reporting LPIPS$_{\text{VGG}}$, except for the LPIPS$^{\text{all}}$. Interestingly, the ACGPN turns out to perform the best in terms of LPIPS based on VGG (in Table 1), while it is the worst according to the metric based on AlexNet. This indicates that the metric based on the entire pixels is highly unstable. On the other hand, we observe that the proposed metrics LPIPS$^{p}_{\varepsilon}$ reserve the same ordering of all four methods across all εs we report, for Table 1 (in the main manuscript) and Table ii.

D More on the User Study

Experimental Settings. We conducted a user study to compare the quality of generation results. We randomly sampled 200 pairs from the VITON test set without cherry-picking and divided them into 2 sets, 100 pairs in each with 60 volunteers who are unfamiliar with virtual try-on techniques, and randomly assigned them to either set A or B. As illustrated in Fig. i, each question shows the reference person (I), the target clothes (c), and 4 randomly-ordered virtual try-on images generated by CP-VTON+, ACGPN, PF-AFN, and our method. The participant was asked to choose the best try-on result among them. Fig. ii shows more examples we used in the user study.

Results and Discussions. As mentioned in Sec. 6.4, in total 60 users participated in the survey, and 65.3% of them chose our method most frequently as the best performing one. Fig. iii shows the percentage of participants sorted by the number of questions they chose ours as the best. We see that more than half of the participants chose ours as the best on ≥ 60 questions (out of 100). All participants chose ours as the best for at least 35 questions (out of 100), which is significantly higher than 25, the expected number if randomly chosen from 4 choices. From the question-centric view, as described in Table iv, participants answered that our model generated the best results compared to the other 3 baselines in 153 (76.5%)

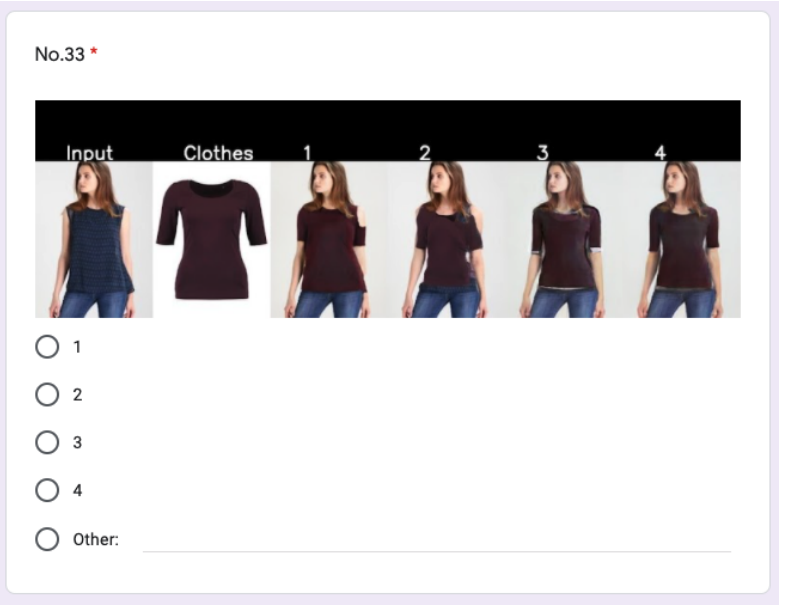

Figure i: A screenshot of a survey question.

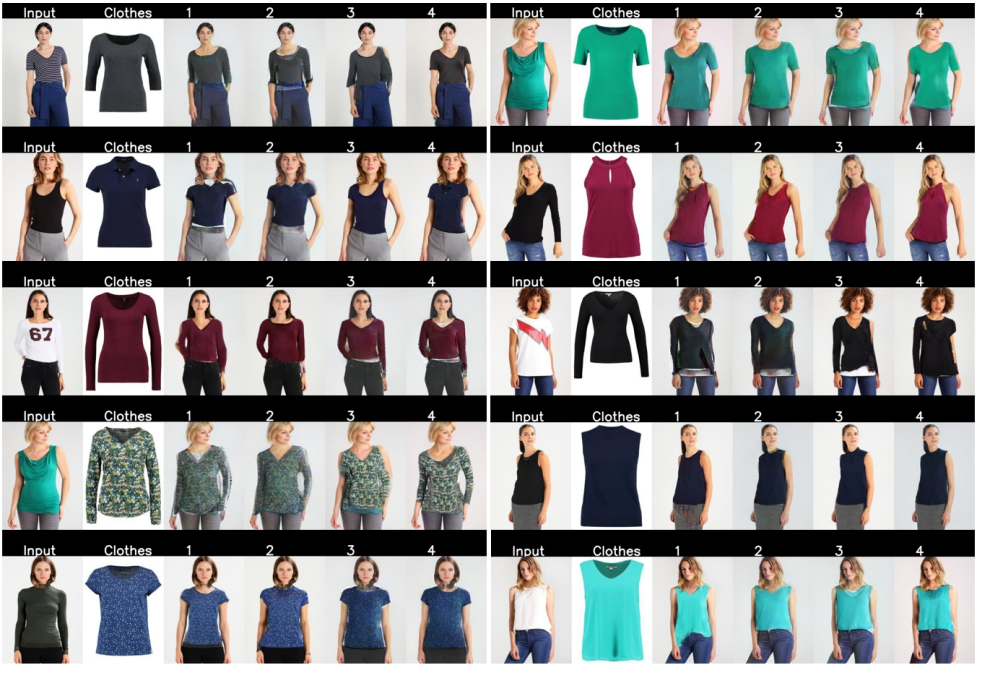

Figure ii: Examples of user study questions.

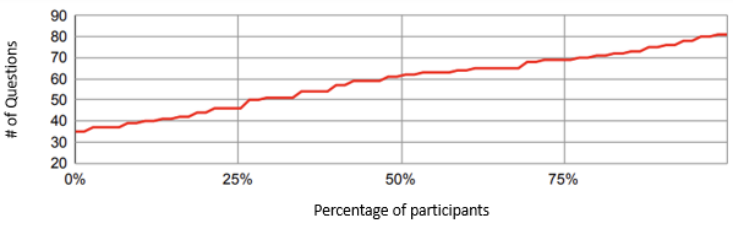

Figure iii: Number of questions in which participants chose our result as the best.

	#Questions	Ratio
CPVTON+	5	2.5%
ACGPN	22	11.0%
PF-AFN	20	10.0%
Ours	**153**	76.5%

Table iv: Number of questions that each method was chosen as the best.

questions among 200 questions.

For reference, ACGPN [13] conducted a user study comparing with CP-VTON [12], VITON [6], and VTNFP [14]. Unlike our study, they asked questions in an A/B (binary) manner; that is, each question had two choices, one by their ACGPN method and the other by one of the baselines. 66.7%, 89.8%, and 76.6% among the participants chose the proposed method (ACGPN) against VTNFP, CP-VTON, and VITON, respectively. PFAFN [4] also conducted a similar user study with 50 volunteers, comparing against CP-VTON, ClothFlow [8], CP-VTON+ [10], ACGPN, and WUTON [7]. 84.3% of their participants chose the proposed method against a single baseline, just as in the user study conducted by [13]. We emphasize that these two user studies asked the participants to choose one out of the two candidates, one from the proposed method and the other from another baseline. Thus, the arithmetic expectation of their experiment was 50%. On the other hand, our user study showed all four images at the same time, making the expectation 25%, while we achieved a 65.3% (participant-centric) or 76.5% (question-centric) winning rate.

E More Qualitative Results

Fig. iv shows additional virtual try-on results using our proposed approach and three baselines [4, 10, 13]. Fig. v illustrates our and baselines' generation results focusing on various body shapes. Table v–vi list additional results on various poses using ACGPN, PF-AFN, and ours.

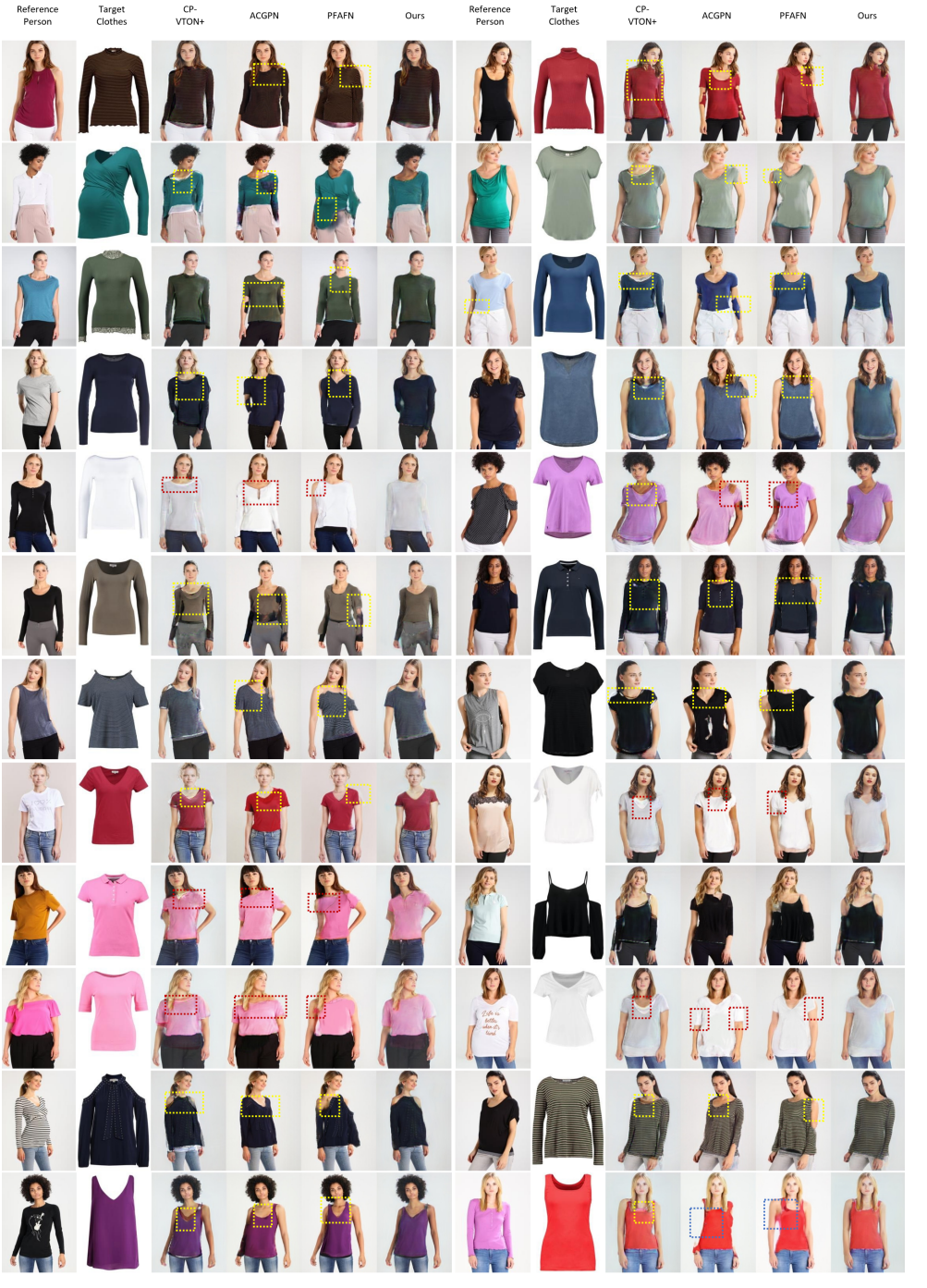

Figure iv: Additional qualitative virtual try-on results.

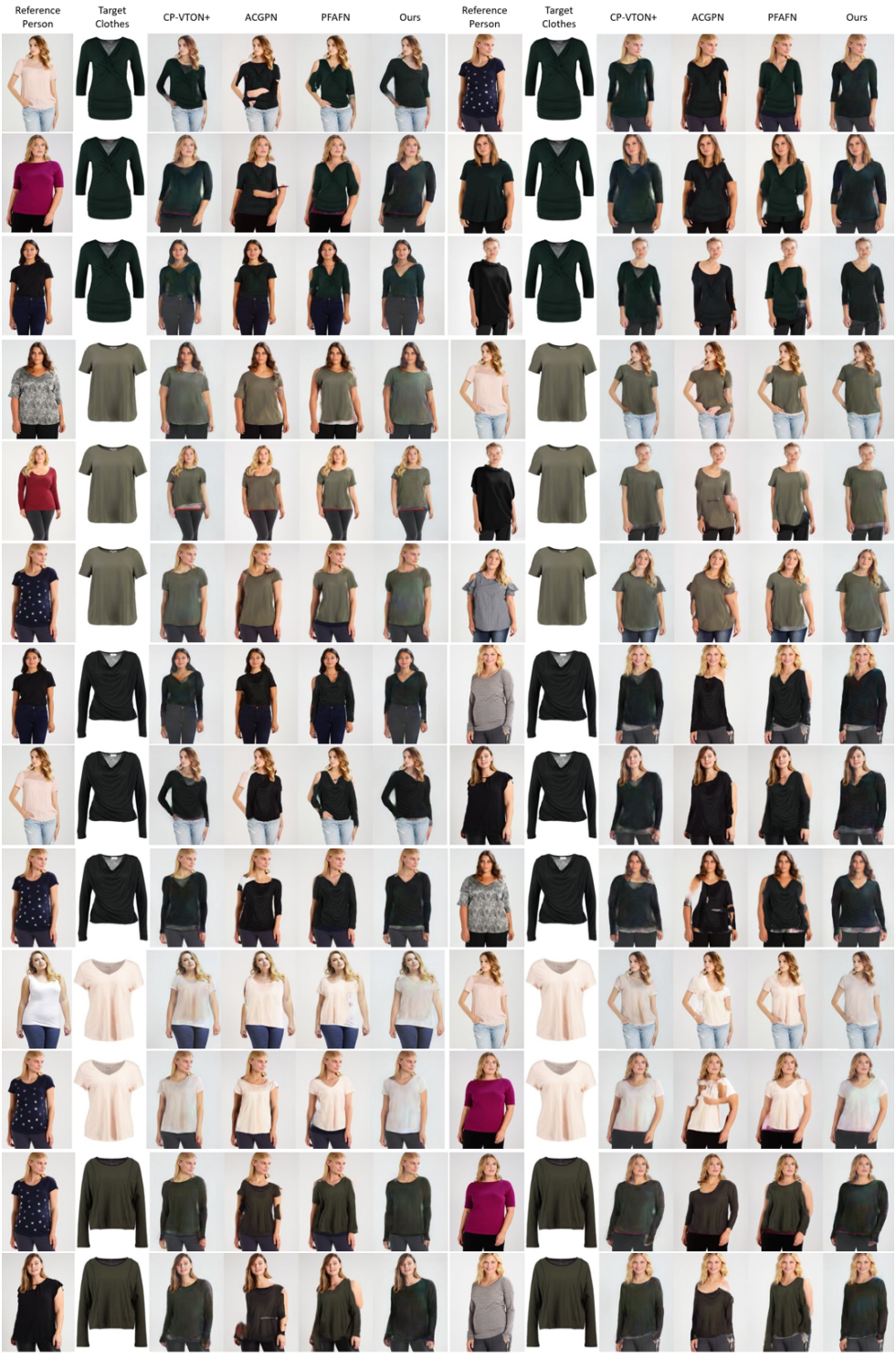

Figure v: Additional results on various body shapes.

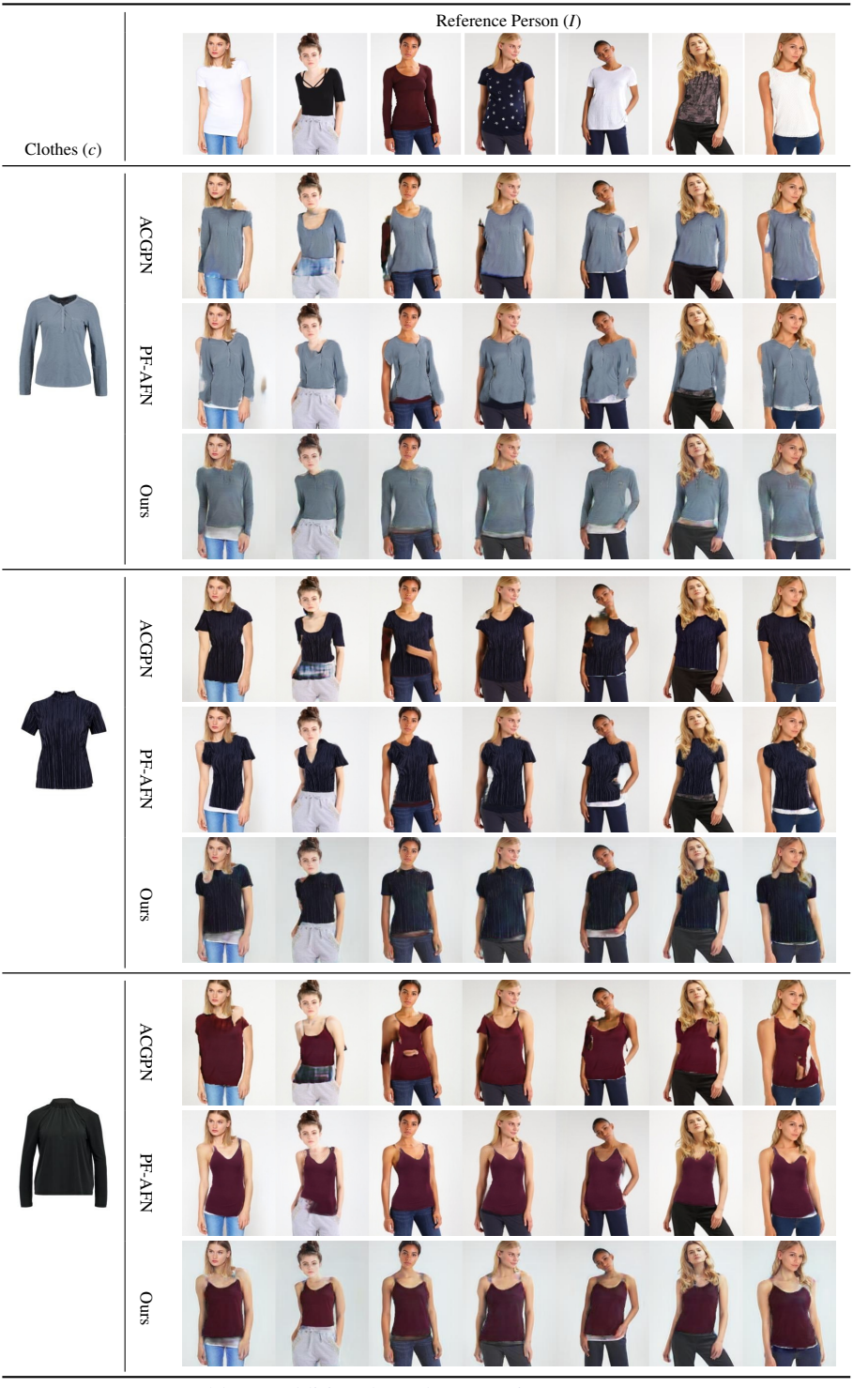

Table v: Additional results on various poses (1/2)

Table vi: Additional results on various poses (2/2)

F Failure cases

We observe the proposed method struggles when arms are folded (marked with red box in Fig. vi). Although it still shows better synthesized results than baselines, better handling such complicated poses is an interesting direction for future work.

Figure vi: Qualitative comparisons with various poses.

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] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[3] 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.

[4] 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.

[5] 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.

[6] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. ClothFlow: A flow-based model for clothed person generation. In *Proc. of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2019.

[7] 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)*, 2020.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[9] 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.

[10] 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 *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2020.

[11] 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)*, 2015.

[12] 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.

[13] 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.

[14] 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.