Size Does Matter: Size-aware Virtual Try-on via Clothing-oriented Transformation Try-on Network

Chieh-Yun Chen[1,2*] Yi-Chung Chen[1,3*] Hong-Han Shuai[2] Wen-Huang Cheng[3]
[1]Stylins.ai [2]National Yang Ming Chiao Tung University [3]National Taiwan University

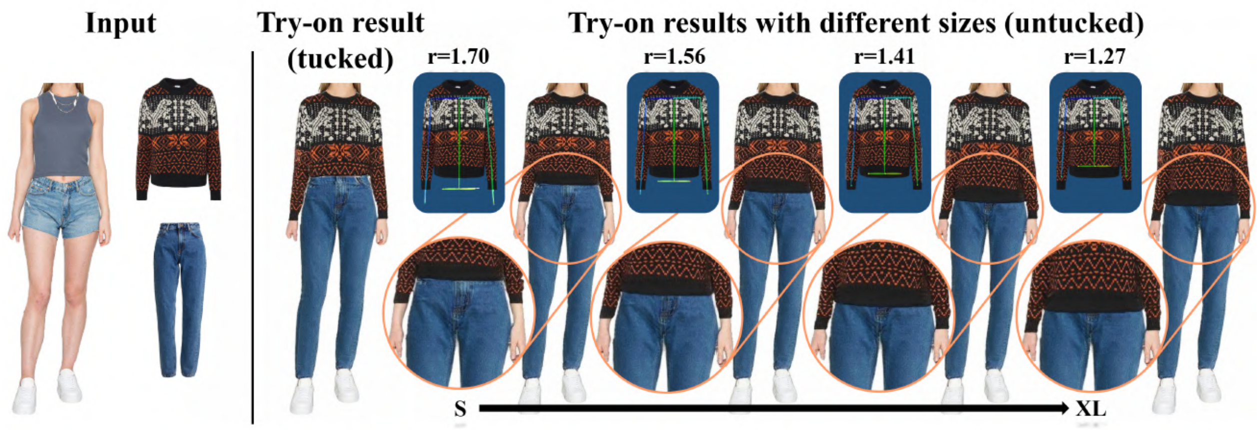

Figure 1. **Multi-garment virtual try-on results with different sizes (tucked or untucked).** r is the ratio of torso length over shoulder width, which can be decided by customers in online shopping scenarios. *Padding blue backgrounds for better visualizing clothing skeletons.*

Abstract

Virtual try-on tasks aim at synthesizing realistic try-on results by trying target clothes on humans. Most previous works relied on the Thin Plate Spline or appearance flows to warp clothes to fit human body shapes. However, both approaches cannot handle complex warping, leading to over distortion or misalignment. Furthermore, there is a critical unaddressed challenge of adjusting clothing sizes for try-on. To tackle these issues, we propose a Clothing-Oriented Transformation Try-On Network (COTTON). COTTON leverages clothing structure with landmarks and segmentation to design a novel landmark-guided transformation for precisely deforming clothes, allowing for size adjustment during try-on. Additionally, to properly remove the clothing region from the human image without losing significant human characteristics, we propose a clothing elimination policy based on both transformed clothes and human segmentation. This method enables users to try on clothes tucked-in or untucked while retaining more human characteristics. Both qualitative and quantitative results show that COTTON outperforms the state-of-the-art high-resolution virtual try-on approaches. All the code is available at https://github.com/cotton6/COTTON-size-does-matter.

1. Introduction

Image-based virtual try-on replaces clothing items on a person with the desired ones, creating realistic try-on results and lowering costs associated with on-model photos for the e-commerce industry[1]. Additionally, it enables customers to use virtual dressing rooms when shopping online, potentially enhancing the e-commerce experience and increasing conversion rates[2]. As online shopping becomes popular, the virtual try-on tasks have received more and more attention [2,4,5,9,13,17,23,28,29,33,35]. For example, to extend the virtual try-on resolution from low (256 × 192) to high (1024 × 768), the most obvious difficulty is the misalignment issue. [5] proposed the alignment-aware segment normalization to remove the misleading information in the misaligned area. Additionally, [23] further performs appearance flow-based warping and segmentation map generation simultaneously to tackle misalignment.

Despite the considerable progress made in previous works, there are still some challenges that have not been adequately addressed, as outlined below. **i) Handling complex warping without misalignment:** General image-

*These authors contributed equally to this work.
[1] https://www.zmo.ai/aimodels/
[2] https://www.revery.ai

based virtual try-on frameworks [5, 16, 21, 28, 31] typically employ a clothing deformation module, such as the Thin Plate Spline (TPS) method, to align clothing images with the target human pose. However, research [6, 15, 23] has found that TPS may not effectively handle complex warping when different garment regions require different deformations. In attempts to improve warping results, they have replaced TPS with dense appearance flows. However, when the transformation between the garment and corresponding body parts is significant (as illustrated in case II in Fig. 4), the performance of both TPS (VITON-HD [5]) and flow-based (HR-VITON [23]) methods deteriorates drastically. Hence, existing methods still can't properly address the challenge of clothing deformation. **ii) Adjusting clothing sizes:** Previous works predicted the try-on segmentation conditioned on the clothing image by only considering the shape of the clothes without the scale information. Thus, given a clothing image, it is impossible for previous works to change the clothing size. This limitation severely restricts the practical application of virtual try-on because people often try on different sizes in the fitting room. **iii) Appropriate clothing elimination policy:** Previous works [5, 23, 28] rely on segmentation maps to eliminate the original clothing region and utilize the remaining area as a guide for trying on target clothes. However, this method can lead to either excessive or insufficient removal of the original image. For instance, when prior works try on upper clothes, the lower part of the upper clothes remains intact, which limits the length of the upper clothes, as shown in Fig. 6. The lower part of the upper clothing is forced to be tucked in, which reduces the model's ability to generate complete upper clothes, thereby reducing the practicality of virtual try-on. Additionally, previous works remove the entire arm of input human images, which eliminates essential human characteristics such as tattoos and arm width that should be preserved in the final output, as illustrated in case I of Fig. 5. Therefore, developing an appropriate clothing elimination strategy that preserves crucial information while removing the clothing region is a significant challenge.

To address these challenges, we propose a simple yet powerful approach called Clothing-Oriented Transformation Try-On Network (COTTON). Specifically, COTTON first exploits geometric information with *Clothing Landmark Predictor* and *Clothing Segmentation Network* to predict clothing landmarks and segmentation masks respectively. To overcome the first challenge posed by complex warping, we propose the *Landmark-guided Transformation* that first separates clothes into sub-parts via the segmentation mask and then uses clothing landmarks to estimate homography matrices that fit these sub-parts to the target human pose. To solve the second challenge, we introduce a clothing landmarks adjustment approach that allows users to change clothing sizes. Adjustment of landmarks will alter the homography matrices and then cause clothing size changes, as demonstrated in Fig. 1, which remarkably enhances the practicality of virtual try-on. To address the third challenge, we employ the transformed clothing images to identify and remove the areas that would be covered by the clothes. This enables our proposed *Clothing Elimination Policy* to effectively eliminate the clothing region while preserving important details and offer flexibility in tucking the clothes or not, as shown in Fig. 6. Extensive experiments show that COTTON achieves superior results than state-of-the-arts both quantitatively and qualitatively. We summarize our contributions as follows:

- We propose a Clothing-Oriented Transformation Try-On Network (COTTON), which improves clothing transformation quality and addresses complex warping misalignments by leveraging clothing geometry.
- We introduce adjustable clothing landmarks to provide clothing size information based only on images. To the best of our knowledge, this is the first work enabling 2D-based virtual try-on with different clothing sizes for approaching real-world try-on.
- To preserve critical information, we propose a *Clothing Elimination Policy* to properly remove clothing information while retaining valuable human characteristics, and offer flexibility in tucking the clothes or not.
- Extensive experiments on the Dress Code dataset [28] show that our model significantly outperforms SOTAs, *e.g.*, at least 41.1% improvement in terms of FID.

2. Related Works

2.1. Clothing Deformation

Virtual try-on requires clothing deformation to align patterns and preserve clothing details. Thin Plate Spline (TPS) warping is widely used in virtual try-on tasks [5, 16, 21, 28, 31], but it has limited ability to model geometric changes and result in unnatural deformations. Thus, [2, 6, 11, 13, 15, 17, 23] conducted appearance flows as an alternative to TPS, but misalignment still has not be well addressed. To address misalignment, [5] improved TPS by eliminating irrelevant clothing texture information in misaligned regions. Meanwhile, [23] improved the deformation by designing pathways to jointly predict appearance flows and segmentation features. However, neither TPS nor appearance flow can deform partial regions of clothing without affecting other parts, *e.g.*, warping the sleeves to present arm akimbo without affecting the torso part of the clothes, since they do not have the structural information of the clothes. To address this issue, [33] introduced normalized patches to learn spatial-agnostic clothing features but this approach may destroy the completeness of clothes

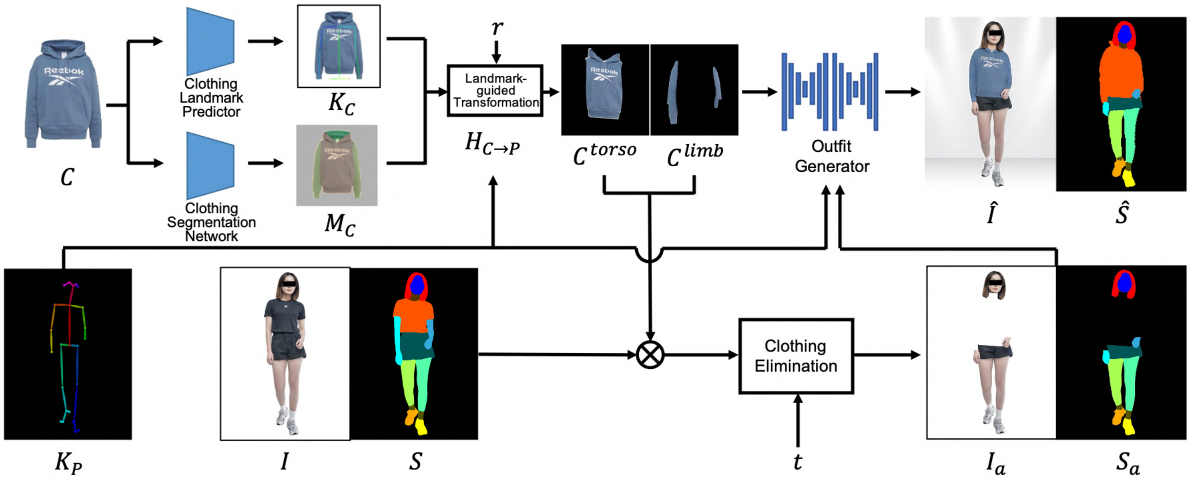

Figure 2. Overview of COTTON. Given a clothing image C, COTTON uses *Clothing Landmark Predictor* and *Clothing Segmentation Network* to predict clothing landmarks K_C and a clothing segmentation mask M_C. C is then warped by *Landmark-guided Transformation* with homography derived from human keypoints K_P and K_C to obtain the transformed images C^{torso} and C^{limb}. r is used to adjust K_C at this step for changing clothing size. The clothing elimination step uses C^{torso} and C^{limb} to mask human image I and segmentation mask S, resulting in clothing-agnostic image I_a and segmentation S_a. The variable t controls whether clothing is tucked in or not. Finally, *Outfit Generator* aggregates the previous outputs to predict segmentation mask $\hat{S}$ and synthesize the final try-on result $\hat{I}$.

due to lack of semantic information. In contrast, our proposed COTTON separates clothes based on landmarks and segmentation masks to maintain clothing completeness and aligns clothes based on specific landmarks to address misalignment. Besides, [8] is a concurrent work. Both of our works improve garment warping by human-garment correspondence. While [8] relies on the computationally intensive DensePose method, our proposed *LT* method is more efficient and enables clothing size adjustment.

2.2. Virtual Try-on

Virtual try-on methods can be categorized into 3D-based methods [1,14,22,24,26,27,37] and 2D-based methods [5,6,7,8,9,10,11,12,15,16,17,19,23,28,29,31,33,34,35]. Since 3D-based methods are time-consuming and real-world scenarios require fast inference speeds, we focus on 2D-based methods. [16] proposed a coarse-to-fine network to synthesize image-based virtual try-on results based on TPS warping. However, try-on results synthesized by [16, 31] do not provide a clear silhouette of human body parts since they consider the human to be a general mask instead of considering the semantic information, *e.g*., legs, head, limbs, etc. To tackle this problem, [10,20] proposed the first semantic-guided virtual try-on network, which predicted the target semantic segmentation to provide human structural information for the network to synthesize photo-realistic body parts. Afterward, many try-on methods [5, 6, 7, 9, 23, 29, 35] are designed with an individual stage for predicting the target semantic segmentation before synthesizing virtual try-on results. However, the individual stage for predicting target semantic segmentation would lead to error accumulation as pointed out by [21]. In this paper, we propose an end-to-end *Outfit Generator* that updates semantic segmentation and try-on results simultaneously, and introduce a *Clothing Elimination Policy* that retains valuable human characteristics while removing original clothing information.

3. Methodology

Given a human image I and a clothing image C, the goal of virtual try-on is to obtain a synthetic image $\hat{I}$, which shows the human of I wearing the clothing of C. We design a Clothing-Oriented Transformation Try-On Network (COTTON) for generating high-resolution virtual try-on results. Fig. 2 provides an overview of the proposed COTTON. We use the *Clothing Landmark Predictor* and the *Clothing Segmentation Network* to respectively capture clothing landmarks and a clothing segmentation map of C. Afterward, we separate the clothes into torso part C^{torse} and limbs part C^{limb} according to the clothing segmentation map. The separation of clothes enables COTTON to treat each part differently, and thus makes it adapt to handling complex warping well. To align the clothes with the target human pose, the clothing landmarks are adjusted with a ratio r to fit human body proportions and used to calculate homography matrices. COTTON then warps the clothes with affine transformation. The warped clothes shed light on how to properly remove the clothing region of image I without losing necessary information. We use a user-determined variable t to control whether to eliminate the overlap region of clothes. In this case, we can generate tucked-in or untucked try-on results by choosing the value of t. Finally, the *Outfit Generator* takes both human and clothing representation to generate the try-on result $\hat{I}$.

3.1. Clothing Landmark Predictor (CLP)

One of the most challenging problems of virtual try-on is to transform a clothing image C to fit the target human pose. To solve this challenge, several approaches have been proposed in previous works, including hyperparameter learning of TPS warping [5, 16, 28, 31], and flow-based deep deformation [6, 15, 23]. However, these approaches implicitly learn the underlying geometric information from the training dataset and only achieve mediocre results when handling complex warping. We argue that explicit geometric information plays the most critical role in virtual try-on and can be leveraged to further improve performance. For example, by thoroughly understanding the clothes' geometry, humans can easily capture the relation between clothes and humans, even if the clothes are with different textures and styles. Therefore, we propose the *Clothing Landmark Predictor (CLP)* to extract explicit geometric information.

Specifically, to exploit the geometric information of clothes, our proposed *CLP* is trained to predict a set of landmarks K_C, which correlate to different parts of human pose K_P, where K_P can be obtained from off-the-shelf human pose estimation methods, *e.g.*, [3]. The ground truth of K_C is labeled manually, and the number of landmarks varies with regard to the clothing types. For upper clothing, $|K_C| = 10$, including neck, shoulders, elbows, wrists, and hips. These landmarks explicitly indicate the relation between clothes and human pose, and therefore can be used as a reference to generate transformation matrices. Given clothing image C, the *CLP* is trained to output clothing joint heatmaps $\hat{M}_{JH}$ and Part Affinity Field $\hat{M}_{PAF}$. Afterward, the predicted landmarks K_C can be obtained from $\hat{M}_{JH}$ and $\hat{M}_{PAF}$ by using the Hungarian algorithm.

3.2. Clothing Segmentation Network

Each part of clothes has distinct properties that should be taken into consideration during transformation. For example, the sleeves of clothes usually suffer severe distortion during transformation while the torso part only undergoes a mild distortion. Another common noise in clothing try-on is the region around the neckline that must be concealed when the garment is worn. This region always introduces undesired variation into try-on results, and as such, should be eliminated. Though the *Clothing Landmark Predictor* provides the connection between the clothes and human poses, it is difficult to identify different regions with only a few landmarks. To find the explicit boundary of each part, pixel-level semantic segmentation is necessary. Therefore, we propose a *Clothing Segmentation Network* to predict the clothing segmentation mask M_C, which separates the clothing into three sub-region, including torso part, sleeve part, and invisible part around the neckline. As such, each sub-region can be treated properly based on its characteristics.

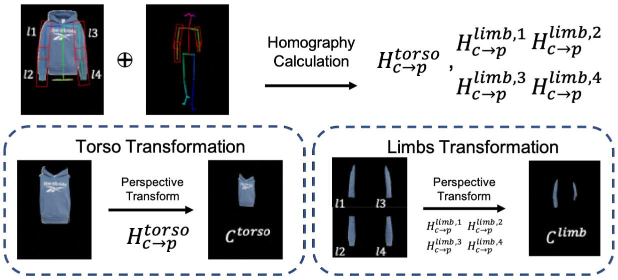
l1
l3
l2
l4
⊕
Homography Calculation
$H_{c\to p}^{torso}$, $H_{c\to p}^{limb,1}$ $H_{c\to p}^{limb,2}$ $H_{c\to p}^{limb,3}$ $H_{c\to p}^{limb,4}$
Torso Transformation
Perspective Transform
$H_{c\to p}^{torso}$
C^{torso}
Limbs Transformation
l1
l3
l2
l4
Perspective Transform
$H_{c\to p}^{limb,1}$ $H_{c\to p}^{limb,2}$ $H_{c\to p}^{limb,3}$ $H_{c\to p}^{limb,4}$
C^{limb}

Figure 3. Clothing transformation steps. For the torso part, we use keypoints including shoulders and hips to calculate matrix $H_{c\to p}^{torso}$. For the limbs part, we first create bounding boxes around limbs and use them to calculate matrices $H_{c\to p}^{limb,i}, i \in \{1, 2, 3, 4\}$.

3.3. Landmark-guided Transformation

Here, we introduce how to leverage the landmarks K_C and segmentation mask M_C to achieve high-fidelity clothing transformation. We first remove the undesired part around the neckline such as the green region of M_C in Fig. 2, which is invisible in the final try-on result, via the clothing segmentation map. We then separate the remaining part into two parts, including the torso part and the limbs part. Let K_C^{torso} and K_P^{torso} denote the clothing and human keypoints belonging to the torso part, such as shoulders and hips. For aligning the torso part, we compute a homography matrix $H_{c\to p}^{torso} \in R^{3\times 3}$ that can project clothing landmarks to human keypoints, as follows:

$$k_p = H_{c\to p}^{torso} k_c, \quad \forall k_c \in K_C^{torso}, \, \forall k_p \in K_P^{torso}. \tag{1}$$

Afterward, this perspective transformation is applied to get the transformed torso part C^{torso}. On the other hand, as the limbs part usually suffers from severe distortion, it requires a more sophisticated procedure. Therefore, instead of directly warping the whole limbs part, we break the limbs part into four sub-limbs based on the clothing landmarks. Take a hoodie as an example in Fig. 3. The limbs part is divided into two upper arms and two forearms. We create bounding boxes for each sub-limb based on clothing landmarks and human keypoints to obtain the area of each sub-limb. The corners of bounding boxes are then used to compute 4 homography matrices, $H_{c\to p}^{limb,i}, i \in \{1, 2, 3, 4\}$, which project each sub-limb to the target human. Finally, the four sub-limbs are collected as the warped limbs part C^{limb}.

It is worth noting that the predicted clothing landmarks K_C only provide a clothes-human relation in the general case. However, since body proportion varies from person to person, the relative size of the same clothing on different persons may change. To generate a more realistic try-on result, we can adjust the clothing landmark K_C to fit the body proportion of the target human. For instance, if two people have the same shoulder width but different torso lengths, the portion of their bodies covered by clothing should be different. In that case, the predicted landmarks are adjusted accordingly to fit their body proportions. Let r be the ratio of torso length over shoulder width, which can be derived from extra body information provided by customers in online shopping scenarios. Fig. 1 visualizes the results of clothing landmarks adjustment with different r, which shows that the region covered by the clothes is varied with regards to different r, and can therefore generate try-on results with different clothing sizes.

3.4. Clothing Elimination Policy

Since the goal of the virtual try-on task is to replace the clothes of a given person image I with a target one, the information about the original clothes should be eliminated. One common approach is eliminating the clothing according to the human segmentation S. However, [5] pointed out that simply removing the clothing from image I leads to a performance drop for paired data training due to information leakage about the clothing shape, and therefore proposed a clothing-agnostic person representation to address this problem. Nevertheless, this approach may lead to valuable information loss, *e.g.*, arm width in case I of Fig. 5, when whole arms are directly masked during inference.

Hence, we propose a *Clothing Elimination Policy* for the inference phase to preserve the important information while generating clothing-agnostic person representation. The regions that ought to be eliminated in image I are i) the regions of original clothes and ii) the regions covered by target clothes C when worn. Note that the latter is hard to measure with previous methods. Fortunately, since we have already obtained the reliable warped clothing image at the previous step, we can use it as a reference to find the region that would be covered after try-on. Specifically, we create a clothing-elimination mask M_{elm} by union the original clothing region and the region covered by the transformed target clothes, which is then expanded using morphological dilation to cover the entire body. Moreover, when wearing tops, we make the clothing tucked or not in the try-on result by controlling $t \in \{0, 1\}$ to determine whether to eliminate the overlap region of bottoms. As such, we can effectively obtain clothing-agnostic person representation without overly masking the irrelevant regions, *i.e.*,

$$M_{elm} = \mathcal{M}(S^t \cup (C^{torse} \cup C^{limb}) - t \cdot (C^{torse} \cap S^b)), \quad (2)$$

where $\mathcal{M}$ represents the morphological dilation; S^t and S^b denote top and bottom clothing segmentation respectively. Fig. 1 shows the results of different t. After eliminating clothing information, we obtain clothing-agnostic images I_a and segmentation maps S_a. To bridge the performance gap between the training and inference phases, we randomly mask the whole arms of human images with a 50% probability during the training phase and apply our *Clothing Elimination Policy* only in the inference phase.

3.5. Outfit Generator

In the final stage, the *Outfit Generator* takes personal (I_a, S_a and K_P) and clothing (C^{torse} and C^{limb}) information as input and synthesizes the final try-on result. Since learning the shape and texture of clothing simultaneously is challenging for the generator, a segmentation network is commonly used to predict a segmentation mask as guidance to lead the try-on generator network [5, 20, 23]. However, they regarded segmentation prediction as an independent task and trained the segmentation network and try-on generator separately. In that case, the optimization of the two networks may be suboptimal. Moreover, the accumulated noise from the segmentation network may further deteriorate the performance of the final synthesis result [21].

To address the problem, we propose a novel *Outfit Generator* by combining the segmentation network and try-on generator as one end-to-end training multi-task network.[3] Specifically, S_a, K_P, C^{torse}, C^{limb} are first concatenated and down-sampled to a lower resolution as the input of the segmentation network to predict segmentation mask $\hat{S}$. Afterward, $\hat{S}$ is up-sampled back to the original resolution and combined with I_a as well as C^{torse} and C^{limb} as the input of the synthesis step. Finally, the generator integrates all the information and renders the synthesis result $\hat{I}$.

$$\hat{I}, \hat{S} = G(I_a, S_a, K_P, C^{torse}, C^{limb}). \quad (3)$$

The objective function of the *Outfit Generator* includes the supervision of both the predicted segmentation mask and the try-on synthesis result. We apply focal loss $\mathcal{L}_{focal}$ [25] for segmentation. For try-on results, we use losses including reconstruction loss $\mathcal{L}_{rec}$, perceptual loss $\mathcal{L}_{VGG}$, and adversarial loss $\mathcal{L}_{adv}$. The overall loss functions are as follows:

$$\begin{aligned} \mathcal{L}_{overall} &= \lambda_{rec}\mathcal{L}_{rec} + \lambda_{VGG}\mathcal{L}_{VGG} \\ &\quad + \lambda_{focal}\mathcal{L}_{focal} + \lambda_{adv}\mathcal{L}_{adv}, \end{aligned} \quad (4)$$

$$\mathcal{L}_{rec} = \|\hat{I} - I\|_1, \quad (5)$$

$$\mathcal{L}_{VGG} = \sum_{k=1}^{5} \lambda_k \|\phi_k(\hat{I}) - \phi_k(I)\|_1, \quad (6)$$

$$\mathcal{L}_{focal} = \sum_i -\alpha_i (1 - p_i)^\gamma \log(p_i), \ p_i \in \hat{S}, \quad (7)$$

$$\mathcal{L}_{adv} = \|\mathcal{D}(\hat{I}) - \mathbf{1}\|_2, \quad (8)$$

where ϕ_k represents the k^{th} layer output of VGG-19 network. α_i and p_i are respectively the importance weighting for the i^{th} class and the probabilities of pixels belonging to the i^{th} class. γ is the focusing parameter. $\lambda_{rec}, \lambda_{VGG}, \lambda_{focal}$ and λ_{adv} denote the balance weights. $\mathcal{D}$ is the multi-layer discriminators [32].

[3] The comparison of two-stage and end-to-end training is in supplement.

4. Experiments

4.1. Experimental Setup

Dataset. We created the Pure Cotton, a high-resolution (1024×768) outfit dataset, to mitigate the impact of noisy data on model performance. We strictly employed several rules to filter out unsuitable data, which are clearly outlined in the supplements, resulting in 16,428 frontal-view human models and in-shop clothing pairs, including 10,636 upper clothes and 5,792 bottom clothes. For training and testing, we divided the upper clothing set into 8,451 and 2,185 pairs, respectively, and the bottom clothing set into 4,626 and 1,166 pairs. To evaluate the model's robustness, we also experimented with the public Dress Code dataset [28].

Implementation Details. To train *Clothing Landmark Predictor (CLP)* and *Clothing Segmentation Network (CSN)*, we collected a small clothing dataset with 50 images for each clothing type, including shirts, turtlenecks, hoodies, etc, and manually labeled landmarks and segmentation masks as ground truth. The architecture of *CLP* is based on the one in [3], and that of *CSN* is based on FCN-resnet50 [30]. Moreover, *Outfit Generator* is designed using an Unet and an encoder-decoder network. To ensure a fair comparison with previous studies, all figures are generated in the tucked-in setting unless stated otherwise. This setting is consistent with prior arts.

Baselines. We compare our proposed method with 3 relevant state-of-the-art baselines, including HR-VITON [23], VITON-HD [5], and PASTA-GAN [33], by training all the models with our self-collected Pure Cotton dataset and the public Dress Code dataset [28]. Specifically, HR-VITON and VITON-HD both represent SOTA high-resolution virtual try-on methods with different warping modules. HR-VITON warps clothes with appearance flows and designs a feature fusion network for eliminating the mismatch issue, while VITON-HD warps clothes with TPS and proposes an alignment-aware segmentation normalization method to deal with the misalignment problem. Meanwhile, PASTA-GAN represents another warping method, a patch-routed disentanglement module, which is more related to our proposed clothing deformation method.

Evaluation metrics. We evaluate the model performance by three widely-used metrics for virtual try-on, including i) Structural Similarity (SSIM) [38], ii) Learned Perceptual Image Patch Similarity (LPIPS) [36], and iii) Fréchet Inception Distance (FID) [18]. SSIM measures the quality of three key features: luminance, contrast, and structure between the reconstruction results and the ground truth. LPIPS evaluates the perceptual similarity and FID is computed for evaluating the GAN performance.

4.2. Qualitative Results

Fig. 4 and 5 show the visual comparisons in Pure Cotton and Dress Code datasets. Overall, our model achieves the most visually convincing high-resolution try-on results. The comparison analysis is discussed as follows.

The previous warping methods, *e.g.*, TPS (VITON-HD) and appearance flows (HR-VITON), did not consider clothing skeleton and sleeves segmentation, rendering them incapable of accurately synthesizing clothing patterns and features in respective positions, such as hit color mosaic (case I in Fig. 4) and logo in special positions (case II in Fig. 5). Additionally, these methods could not tackle regional clothing warping, such as arm akimbo in complex postures (case II in Fig. 4). In contrast, our model conducts the *Clothing Landmark Predictor* and *Clothing Segmentation Network* to obtain clothing structural information. Then, our proposed *Landmark-guided Transformation* helps to warp the clothes regionally. As such, our model could synthesize try-on results with clothing patterns and features located in the right position that closely follow the target clothing image, even in complex postures. On the other hand, PASTA-GAN conducts a regional clothing deformation method similar to our proposed method. However, there are three differences between PASTA-GAN and our model. Firstly, we split the clothes into 5 regions, *i.e.*, torso, right/left upper arms, and right/left forearms, and kept the torso undivided. Secondly, we further consider clothing segmentation, including neckline and limb segmentation. Thirdly, PASTA-GAN deforms clothes from clothes on another human image, as demonstrated on the right-bottom side of the target clothes columns in Fig. 4. Due to these differences, PASTA-GAN would produce strange try-on results that highly follow the original clothing tied up and limb postures (the blue circle and the red arrow on Fig 4). In contrast, our model leverages clothing structural information to synthesize natural clothing wrinkles.

Additional Clothing Control. The elimination policy in our proposed approach allows for greater flexibility in virtual try-on, with options for both tucked-in and untucked clothing. As shown in Fig. 6, previous methods would limit the location of the upper clothes' lower boundary based on the original lower clothes' upper boundary. However, our approach allows for the clothing to be untucked based on the clothes' body length, providing additional cues for customers to judge the clothing style.

Partial Clothing Resizing: Apart from demonstrating COTTON's clothing size adjustment in Fig. 1, we also present its ability to adjust partial clothing sizes in Fig. 7. In the real world, clothing designers would adjust different sizes only for specific parts, e.g., body or sleeve length, instead of the entire garment. COTTON can smoothly manipulate partial clothing sizes by adjusting the corresponding landmarks without impacting the rest of the clothing.

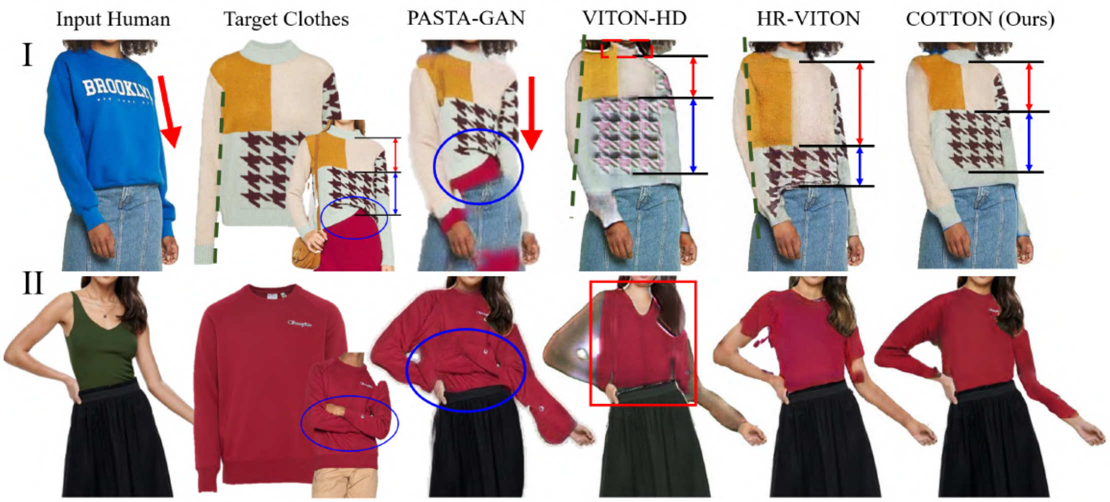

Figure 4. The visual comparison in the Pure Cotton dataset. The left two columns are the inputs and the rest present the results of baselines and our model. VITON-HD and HR-VITON did not consider sleeve segmentation, resulting in misaligned clothes with wrong pattern shifts (green dash line in case I) and broken warped clothes (case II). Moreover, without clothing skeletons, VITON-HD and HR-VITON failed to maintain pattern proportions (red and blue double-headed arrows in case I). PASTA-GAN also had issues in case I, where the partially tucked-in clothing shape from the original human image was directly copied to the try-on result, and in case II, where the try-on result had two additional limbs in front of the abdomen due to the human, representing target clothes, crossing their arms over the chest.

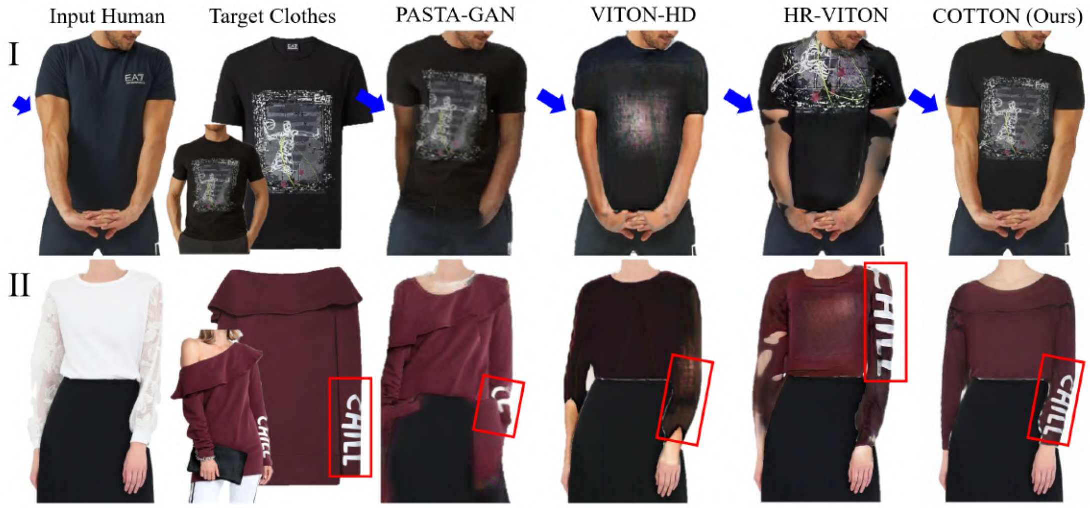

Figure 5. The visual comparison in the Dress Code dataset. In case I, our *Clothing Elimination Policy* well retains valuable human features, e.g., bulging muscles. For case II, our *Landmark-guided Transformation* aligns textures better based on the obtained clothing structure.

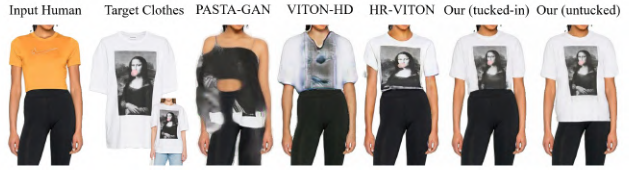

Figure 6. Additional clothing control for tucked-in or untucked.

4.3. Quantitative Results

Table 1 compares COTTON's performance with SOTAs in terms of SSIM, LPIPS, and FID. We evaluated SSIM and LPIPS on 2,000 random paired test sets from our Pure Cotton dataset and 1,800 test sets from the Dress Code dataset. FID was evaluated on 2,000 and 1,800 unpaired test sets from Pure Cotton and Dress Code datasets, respectively. Human images wearing original clothing served as ground truth for FID calculation. The results show that COTTON consistently outperforms the baselines across all evaluation metrics on both datasets, with at least 18.6% improvement in LPIPS on Pure Cotton dataset and 41.1% in FID on Dress Code dataset. The results manifest that our method excels in generating detailed contents on Pure Cotton dataset and robust results on Dress Code dataset. Notably, COTTON shows lower performance with untucked clothes, as it ob-

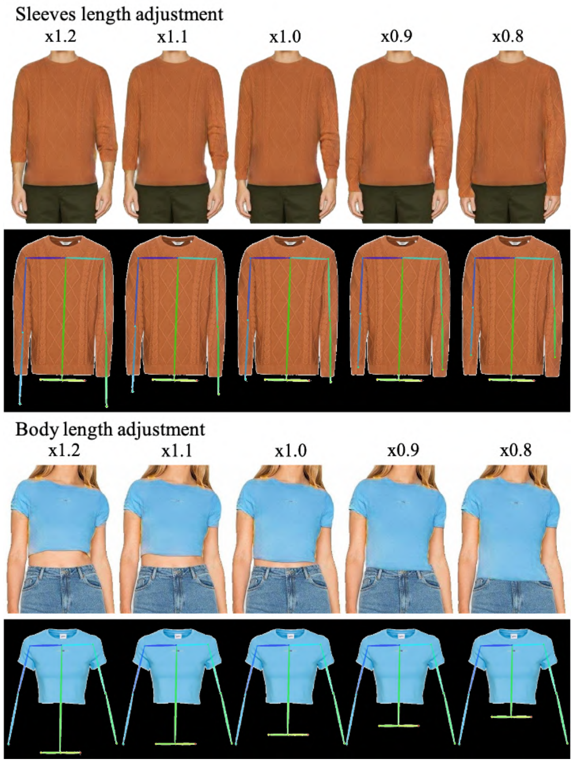

Figure 7. Visualization of partial clothing adjustment.

Table 1. Quantitative comparison on the test dataset.

Method	Dataset	SSIM↑	LPIPS↓	FID↓
PASTA-GAN	Self-collected Pure Cotton	0.918	0.1215	29.43
VITON-HD		0.944	0.0566	17.16
HR-VITON		0.948	0.0387	11.16
Ours (untucked)		0.956	0.0349	10.96
Ours (tucked-in)		**0.958**	**0.0315**	**10.17**
PASTA-GAN	Public Dress Code	0.800	0.3066	33.92
VITON-HD		0.927	0.0906	22.53
HR-VITON		0.927	0.0755	20.12
Ours (untucked)		0.949	0.0482	12.58
Ours (tucked-in)		**0.953**	**0.0438**	**11.86**

Table 2. The user study shows that COTTON is the most photo-realistic and remains the most human and clothing characteristics.

Method	PASTA-GAN	VITON-HD	HR-VITON	COTTON (Ours)
Photo-realistic	9.08%	7.56%	29.82%	**53.54%**
Try-on accuracy	8.66%	4.62%	27.40%	**59.32%**

scures the upper portion of lower clothing, causing the try-on results to differ from the ground truth. Nevertheless, Fig. 6 demonstrates that COTTON can generate more visually appealing results when clothing is untucked.

Moreover, we conduct a user study on the Pure Cotton dataset to further evaluate the visual quality by humans. The experiment involved 127 volunteers. We randomly sampled 30 image sets and divided them equally into two groups for evaluating: i) photo-realistic quality and ii) try-on accuracy. To evaluate the photo-realistic quality, volunteers would see four try-on results (without input information) synthesized by four different models, and choose the most photo-realistic one. Furthermore, for evaluating try-on accuracy, we provide a source human image, one target try-on clothing, and four try-on results synthesized by four different models. Based on the source human and the target clothes, volunteers should select the best try-on result. Table 2 shows that COTTON outperforms all other baselines in both photo-realistic quality and try-on accuracy and achieves even higher ratings for try-on accuracy. Volunteers reported that they mainly used realistic reproduced clothing patterns, including loose/fit torso and accurate sleeve length, with fewer artifacts as the criteria to determine the better model. Our proposed *Clothing Landmark Predictor* and *Clothing Segmentation Network* effectively retaine clothing characteristics such as patterns and sleeve length, contributing significantly to COTTON's superior performance over other state-of-the-art models.

5. Conclusion

This paper presents COTTON, a clothing-oriented transformation try-on network that synthesizes high-resolution virtual try-on results. By considering the clothing structure and leveraging it, COTTON improves the clothing transformation quality and allows for size adjustment during clothing try-on. Additionally, the proposed elimination policy enables COTTON to try on outfits untucked or not while preserving valuable human characteristics. The experiments demonstrate that COTTON outperforms the state-of-the-art virtual try-on works at 1024 × 768 resolution, both qualitatively and quantitatively. In future work, we plan to improve the accessories segmentation and extend the system to try on more clothing categories, such as shoes, hats, and accessories. Overall, the proposed approach shows great promise for improving the online shopping experience and reducing the need for physical try-on.

Acknowledgement

We would like to express our gratitude to Josephine, Wendy, Hans, Joyce, Leslie and May from the Epoch Foundation for their valuable support. This work was supported in part by the National Science and Technology Council of Taiwan under Grants NSTC-109-2221-E-009-114-MY3, NSTC-112-2221-E-A49-094-MY3, NSTC-109-2223-E-009-002-MY3, NSTC-112-2634-F-002-002-MBK and NSTC-112-2628-E-002-033-MY4.

References

[1] Alakh Aggarwal, Jikai Wang, Steven Hogue, Saifeng Ni, Madhukar Budagavi, and Xiaohu Guo. Layered-garment net: Generating multiple implicit garment layers from a single image. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2022. 3

[2] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. Single stage virtual try-on via deformable attention flows. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2

[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 (CVPR)*, 2017. 4, 6

[4] Chieh-Yun Chen, Ling Lo, Pin-Jui Huang, Hong-Han Shuai, and Wen-Huang Cheng. Fashionmirror: Co-attention feature-remapping virtual try-on with sequential template poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1

[5] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. VITON-HD: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1, 2, 3, 4, 5, 6

[6] Ayush Chopra, Rishabh Jain, Mayur Hemani, and Balaji Krishnamurthy. Zflow: Gated appearance flow-based virtual try-on with 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 2, 3, 4

[7] Chien-Lung Chou, Chieh-Yun Chen, Chia-Wei Hsieh, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. Template-free try-on image synthesis via semantic-guided optimization. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*, 2021. 3

[8] Aiyu Cui, Sen He, Tao Xiang, and Antoine Toisoul. Learning garment densepose for robust warping in virtual try-on. In *arXiv*, 2023. 3

[9] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 1, 3

[10] 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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 3

[11] Xin Dong, Fuwei Zhao, Zhenyu Xie, Xijin Zhang, Kang Du, Min Zheng, Xiang Long, Xiaodan Liang, and Jianchao Yang1. Dressing in the wild by watching dance videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3

[12] Ruili Feng, Cheng Ma, Chengji Shen, Xin Gao, Zhenjiang Liu, Xiaobo Li, Kairi Ou, Deli Zhao, and Zheng-Jun Zha. Weakly supervised high-fidelity clothing model generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

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

[14] Oshri Halimi, Tuur Stuyck, Donglai Xiang, Timur Bagautdinov, He Wen, Ron Kimmel, Takaaki Shiratori, Chenglei Wu, Yaser Sheikh, and Fabian Prada. Pattern-based cloth registration and sparse-view animation. *ACM Trans. Graph.*, 41(6), 2022. 3

[15] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R. Scott. ClothFlow: a flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019. 2, 3, 4

[16] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. VITON: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3, 4

[17] Sen He, Yi-Zhe Song, and Tao Xiang. Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 2, 3

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 2017. 6

[19] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, and Wen-Huang Cheng. Fit-me: Image-based virtual try-on with arbitrary poses. In *IEEE International Conference on Image Processing (ICIP)*, 2019. 3

[20] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. FashionOn: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proceedings of the 27^{th} ACM International Conference on Multimedia (ACM MM)*, 2019. 3, 5

[21] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *European Conference on Computer Vision (ECCV)*, 2020. 2, 3, 5

[22] Maria Korosteleva and Sung-Hee Lee. Neuraltailor: Reconstructing sewing pattern structures from 3d point clouds of garments. *ACM Trans. Graph.*, 41(4), 2022. 3

[23] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 4, 5, 6

[24] Ren Li, Benoit Guillard, Edoardo Remelli, and Pascal Fua. Dig: Draping implicit garment over the human body. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2022. 3

[25] Tsung-Yi Lin, Priya Goyal, Ross B. Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020. 5

[26] Xinqi Liu, Jituo Li, Guodong Lu, Dongliang Zhang, and Shihai Xing. Robust and automatic clothing reconstruction based on a single rgb image. *Computers & Graphics*, 2023. 3

[27] S. Majithia, S. N. Parameswaran, S. Babar, V. Garg, A. Srivastava, and A. Sharma. Robust 3d garment digitization from monocular 2d images for 3d virtual try-on systems. In *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2022. 3

[28] Davide Morelli, Matteo Fincato, Marcella Cornia, Federico Landi, Fabio Cesari, and Rita Cucchiara. Dress Code: High-Resolution Multi-Category Virtual Try-On. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022. 1, 2, 3, 4, 6

[29] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. Image based virtual try-on network from unpaired data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3

[30] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 6

[31] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 3, 4

[32] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 5

[33] Zhenyu Xie, Zaiyu Huang, Fuwei Zhao, Haoye Dong, Michael Kampffmeyer, and Xiaodan Liang. Towards scalable unpaired virtual try-on via patch-routed spatially-adaptive GAN. In *Advances in Neural Information Processing Systems (NIPS)*, 2021. 1, 2, 3, 6

[34] Han Yang, Xinrui Yu, and Ziwei Liu. Full-range virtual try-on with recurrent tri-level transform. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[35] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating $\leftrightarrow$ preserving image content. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1, 3

[36] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[37] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[38] 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 (TIP)*, 2004. 6