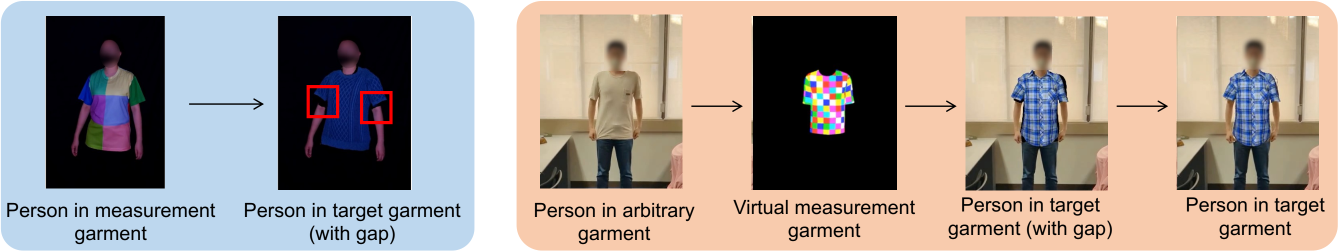
Virtual Measurement Garment for Per-Garment Virtual Try-On

Zaiqiang Wu* Jingyuan Liu* Toby Chong† I-Chao Shen‡ Takeo Igarashi§

The University of Tokyo


Person in measurement garment
Person in target garment (with gap)
Person in arbitrary garment
Virtual measurement garment
Person in target garment (with gap)
Person in target garment

(a) Prior per-garment virtual try-on method

(b) Our proposed method

Figure 1: (a) To use the prior per-garment virtual try-on method [5], the user must wear a physical measurement garment and stand in front of a black background. Furthermore, this method often synthesizes virtual try-on image with black gaps between the synthesized garment and body parts, as highlighted in the red box. (b) Our proposed method eliminates the need for wearing a physical measurement garment by introducing a virtual measurement garment. Additionally, we add a gap-filling module to improve the realism of the synthesized try-on image. By combining these features with a robust segmentation network, our proposed pipeline enables users to perform high-quality virtual try-on in more diverse environments.

Abstract

The popularity of virtual try-on methods has increased in recent years as they allow users to preview the appearance of garments on themselves without physically wearing them. However, existing image-based methods for general virtual try-on provide limited support to synthesize realistic and consistent garment images under different poses, due to two main difficulties: 1) the dataset used to train these methods contains a vast collection of garments, but they lack fine details of each garment; 2) they synthesize results by warping the front-view image of the target garment in a rest pose, which results in poor quality and detail for other viewpoints and poses. To overcome these drawbacks, per-garment virtual try-on methods train garment-specific networks that can produce high-quality results with fine-grained details for a particular target garment. However, existing per-garment virtual try-on methods require the use of a physical measurement garment, which limits their applicability. In this paper, we propose a novel per-garment virtual try-on method that leverages a virtual measurement garment, which eliminates the need for the physical measurement garment, to guide the synthesis of high-quality and temporally consistent garment images under various poses. Furthermore, we introduce a gap-filling module that effectively fills the gap between the synthesized garment and body parts. We conduct qualitative and quantitative evaluations against a state-of-the-art image-based virtual try-on method and ablation studies to demonstrate that our method achieves superior performance in terms of realism and consistency of the generated garment images.

Index Terms: Computing methodologies—Computer graphics—Image manipulation—Image processing;

1 Introduction

In recent years, the demand for online shopping of fashion items has significantly increased. However, many customers face challenges in making appropriate purchase decisions, as they are unable to determine whether a particular fashion item fits them or not. Various virtual try-on methods have been developed to address this issue. These methods are broadly categorized into 3D model-based and 2D image-based methods. 3D model-based methods [10, 22, 23, 25, 30] are effective in presenting the target garment from multiple viewpoints and under various poses. However, these methods entail a laborious 3D modeling process by expert designers, as well as additional 3D measurements and high computational costs, which limit their usability. On the other hand, image-based methods [4, 13, 15, 19] do not have such limitations as they utilize real-world images. Nonetheless, most image-based methods are trained on a general dataset, and thus cannot generate fine details of a particular garment. Moreover, they are constrained by their use of 2D static images and fixed camera positions, which restrict their ability to fully convey the 3D fit and style of a garment.

To overcome the aforementioned limitations of existing image-based virtual try-on methods, Chong *et al.* [5] proposed a per-garment virtual try-on method that focuses on garment-specific virtual try-on rather than general virtual try-on. Their key idea is to train the garment synthesis network on a garment-specific dataset captured using a robotic mannequin. The dataset contains images of the same garment across different body shapes, poses, and viewpoints, which enables the network to learn the fine-grained details of the target garment. Moreover, they leverage a physical measurement garment as a proxy to capture the human pose and body shape, which provides a strong prior for the synthesis of the target garment images. However, this also limits the applicability of their method, as it necessitates the person to wear a specially designed garment for the input. In addition, their method suffers from the problem of producing results with gaps due to the misalignment between the measurement garment and the target garment, which compromises the naturalness and realism of the results, as shown in Figure 1 (a).

In this paper, we present a novel per-garment virtual try-on framework that leverages a virtual measurement garment which eliminates

*e-mail: {wuzaiqiang,liujingyuan}@g.ecc.u-tokyo.ac.jp. Both authors contributed equally to the paper.

†e-mail:tobyclh@gmail.com

‡e-mail:ichaoshen@g.ecc.u-tokyo.ac.jp

§e-mail:takeo@acm.org

the need for a physical measurement garment, as illustrated in Figure 1 (b). The virtual measurement garment is a skinned mesh that can deform according to the estimated 3D human poses, thus serving as a guide for synthesizing the target garment images. Firstly, we estimate the 3D human pose from 2D images. Next, we deform the 3D virtual measurement garment according to the estimated human pose and render it into a 2D image. The rendered image is fed into a per-garment network to synthesize the target garment image whose pose conforms to the person in the input image. Finally, we compose the synthesized garment image with the input image and employ a gap-filling network to fill the gap to produce the final result.

Our method achieves superior temporal consistency when applied to videos, despite being an image-based approach that processes each frame independently without any temporal smoothing. Compared to general virtual try-on methods, our per-garment virtual try-on method entails more complex dataset capturing and network training. However, we argue that our target users are garment retailers who aim to promote their products and have the resources to conduct such processes. Moreover, our garment-specific method is less labor-intensive than 3D model-based virtual try-on techniques, which require professional 3D modeling, and can achieve superior results than general image-based methods.

To evaluate the effectiveness of our method, we collected videos of people doing try-on motions with arbitrary backgrounds for qualitative and quantitative evaluations. The experimental results show that our method outperforms existing methods in terms of visual quality and temporal consistency.

Our contributions can be summarized as below:

- A novel method for per-garment virtual try-on that utilizes a virtual measurement garment as an intermediate representation to synthesize high-quality and temporally consistent images of the target garment, without needing additional 3D measurements. (Section 3.1)
- A novel gap-filling module that is designed to fill the gaps that are created by the composition process of virtual try-on. To the best of our knowledge, this is the first work that introduces a dedicated gap-filling network for virtual try-on. (Section 3.4)
- We evaluate the effectiveness of several intermediate representations for garment image synthesis and demonstrate that our proposed virtual measurement garment with a grid pattern is the most effective. (Section 4)

2 Related Work

3D model-based virtual try-on. 3D model-based methods [10, 22, 23, 25, 30] capture 3D measurement data (e.g., 3D human pose and shape) and generate draped garments that conform to the measurement data. Some early works [6, 17, 21, 31] apply physically-based simulations to generate animations of garments in contact with the body. While such approaches are capable of generating visually plausible results with detailed wrinkles, they suffer from high computational expenses, and thus cannot be widely applicable. To reduce the computational cost of simulation, recent works [2, 9, 11, 18, 22, 23, 25–27, 36] utilize data-driven methods to produce garment animation and deformation. However, these methods are not guaranteed to produce accurate results for out-of-distribution inputs. Moreover, 3D model-based approaches necessitate capturing the physical and material properties of the garment in order to generate images that exhibit high realism and fidelity, which poses a significant challenge and requires a lot of effort.

Image-based virtual try-on. Image-based approaches [4, 13, 15, 19] do not require any 3D measurement data, thus are more widely applicable than 3D model-based methods. The pioneering work of CAGAN [15] first tackled the task of swapping fashion items on human images using a learning-based method. VITON [13] introduced a coarse-to-fine synthesis strategy to enhance the quality of the output and applied thin-plate spline (TPS) transformation to align the target garment with the corresponding body region. To preserve details of the target garment, CP-VTON [33] introduced a geometric matching module to learn the TPS transformation parameters explicitly. VTNFP [39] and ACGPN [38] utilized a body segmentation map prediction module to predict semantic layout, which provides critical information for image synthesis and is very beneficial for preserving clothing and body part details. VITON-HD [4] proposed alignment-aware segmentation normalization and generator to handle misaligned regions and preserve details of the target garment to synthesize high-resolution virtual try-on images. HR-VITON [19] further eliminates the pixel-squeezing artifacts that occur in VITON-HD by unifying the garment warping and segmentation generation stages. However, existing image-based virtual try-on methods often fail to synthesize target garments with fine-grained details, as they train the garment synthesis networks on a general dataset rather than a garment-specific dataset. Furthermore, these methods tend to produce unsatisfactory results for target garments that are not seen during training.

Video virtual try-on. The goal of video virtual try-on is to synthesize realistic and temporally consistent videos of a person wearing the target garment, based on image-based virtual try-on methods Previous methods, such as FW-GAN [7] and FashionMirror [3], utilize optical flow as a post-processing method to smooth the flicker between adjacent frames. However, the flicker artifacts still exist after smoothing. To further improve the temporal consistency, ClothFormer [16] proposed to use a two-stage warping module that predicts dense flow mapping to synthesize the target garment sequence with improved spatio-temporal consistency. However, these methods rely on a large dataset of video virtual try-on for training, which is challenging to collect. To date, only one video virtual try-on dataset, VVT [7], is publicly accessible. The lack of video dataset hinders the advancement of video virtual try-on techniques.

3 Method

In this section, we present our per-garment virtual try-on method with a virtual measurement garment. Figure 2 shows an overview of our proposed method. Our method consists of two main components: a Virtual-to-Target (V2T) network and a Gap-Filling (GF) network. The V2T network learns to translate the appearance of a virtual measurement garment, which is a 3D model of a generic garment that can be deformed and rendered according to the 3D human pose, to the appearance of the target garment. The GF network learns to inpaint the missing regions in the synthesized image caused by the removal of the original garment.

To synthesize the appearance of a target garment on a person from an input image, we first employ a pre-trained pose estimator to infer the 3D human pose (Figure 2(b)) from the input image. Second, we use the estimated pose to deform and render a virtual measurement garment (Figure 2(c)) that matches the body shape and pose of the person. Third, we apply the V2T network (Figure 2(d)) to generate a realistic image of the target garment (Figure 2(e)) based on the rendered image. Fourth, we combine the generated image with the input image, where the original garment has been segmented and removed using a garment segmentation tool, FashionFormer [37]. Finally, we fill in the black gap in the combined image using the GF network (Figure 2(h)), which performs gap-filling inpainting. In the training stage, we train the V2T network using paired images of the virtual measurement garment and the target garment. To train the GF network, we use images with black gaps and their corresponding masks, which are generated by erasing specific regions of the arms and areas near the periphery of the upper body garment in the images.

Given a reference person image $I \in \mathbb{R}^{H \times W \times 3}$, H and W denote the height and width of the image. Our method aims to synthesize an

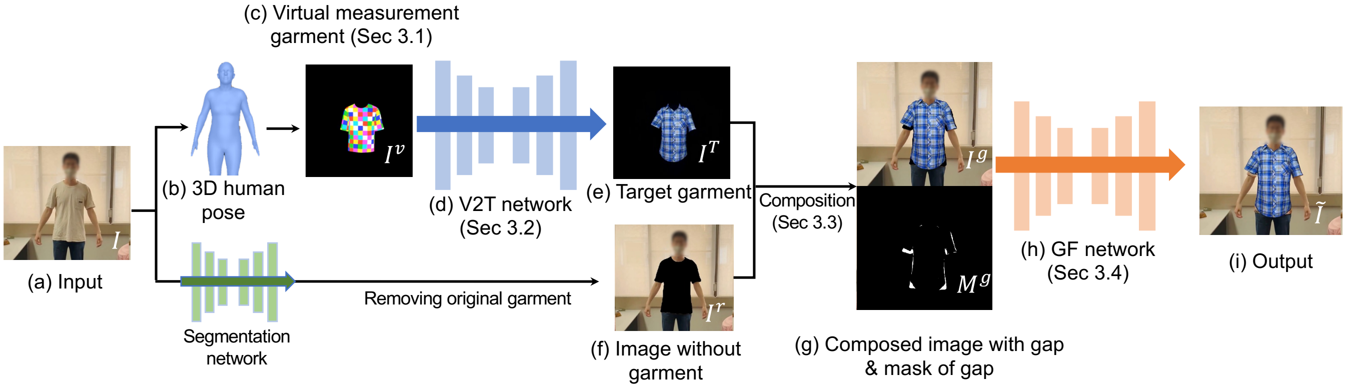


Figure 2: Overview of our method. During the inference stage, we first estimate (b) the 3D human pose of (a) the input image using a pre-trained pose estimator. Then, we deform and render (c) the virtual measurement garment according to the estimated pose. Next, we feed the rendered image to (d) the Virtual-to-Target (V2T) network to obtain the synthesized image of the target garment. We then compose the synthesized image with (f) the input image from which the original garment has been removed using a pre-trained segmentation model. Finally, we inpaint the black gap in the composed image using (h) the Gap-Filling (GF) network.

image $I^T \in \mathbb{R}^{H\times W\times 3}$ that depicts the same person wearing the target garment. Let θ be the estimated 3D human pose from the input frame I, the virtual measurement garment model $\bar{\mathbf{V}}$ is then deformed to conform to the estimated pose $\mathbf{V}$. Following this deformation, the deformed virtual measurement garment model is rendered to produce a 2D image I^v. Subsequently, the rendered virtual measurement garment image I^v is converted to the target garment image I^T via the Virtual-to-Target (V2T) network, expressed formally as:

$$I^T = V2T(I^v) \tag{1}$$

Subsequent to the synthesis of the target garment image I^T, it is composed with the input frame I^r, from which the original garment has been removed, yielding a composite image featuring a discernible gap. The resultant composite image and the corresponding gap mask are denoted as I^g and M^g respectively, which are then input into the Gap-Filling (GF) network to generate the final output image with the gap effectively and aesthetically filled:

$$\tilde{I} = GF(I^g,\ M^g) \tag{2}$$

In Section 3.1, we provide a concise overview of the construction of the virtual measurement garment model. Section 3.2 offers an in-depth account of the target garment synthesis. Section 3.3 presents specific aspects of the composition procedure. Lastly, Section 3.4 presents the details of the gap-filling module designed for gap filling.

3.1 Virtual measurement garment

To build the virtual measurement garment, we first build a 3D template garment mesh that can represent various types of short-sleeved garments, whose vertices in A-pose can be represented as $\bar{\mathbf{V}} \in \mathbb{R}^{N\times 3}$, with N representing the number of vertices. The skinning matrix $\mathscr{W}$ is obtained by deforming an SMPL model with an average body shape to the A-pose, and then projecting each vertex of the virtual measurement garment onto the nearest triangle of the SMPL model. The skinning weight of each vertex is then computed by barycentric interpolation. In addition, we apply Laplacian smoothing to the skinning weights to eliminate artifacts caused by skinning. Following the approach of [5], we texture the virtual measurement garment with a grid pattern from [12] to enhance the quality of synthesized garment images.

Our proposed virtual measurement garment shares the same skeleton as the SMPL model, allowing it to be easily animated by the SMPL pose parameters. Let θ be the estimated 3D human pose from the input frame, the virtual measurement garment model is then deformed to conform to the estimated pose, resulting in a set of new vertices represented as:

$$\mathbf{V} = LBS(\bar{\mathbf{V}}, \theta, \mathscr{W}) \tag{3}$$

where $LBS(\cdot)$ denotes the Linear Blend Skinning deformation.

3.2 Target garment synthesis

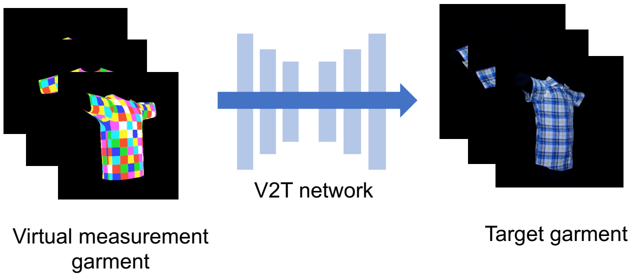


Figure 3: The Virtual-to-Target (V2T) network is trained on the paired images of the virtual measurement garment and the target garment.

Training stage. The Virtual-to-Target (V2T) network performs the translation from rendered virtual measurement garment images to the corresponding target garment images. This network adopts the same architecture as pix2pixHD [34], an image-to-image translation method. The V2T network is trained for a specific target garment captured using a robotic mannequin, as illustrated in Figure 3, allowing it to learn the detailed appearance of the target garment under various poses. This implies that our method cannot handle unseen target garments, which limits its generality. However, we argue that this trade-off improves the quality of the synthesized images, as we demonstrate in our experiments.

Inference stage. To synthesize the target garment image that conforms to the user's body orientation and pose, we first employ BEV [32] to estimate the 3D human pose from the input frame, which is denoted as $\theta \in \mathbb{R}^{24\times 3}$. Next, we deform the virtual measurement garment according to the estimated pose parameters. Following the deformation, the virtual measurement garment model is rendered to produce a 2D image:

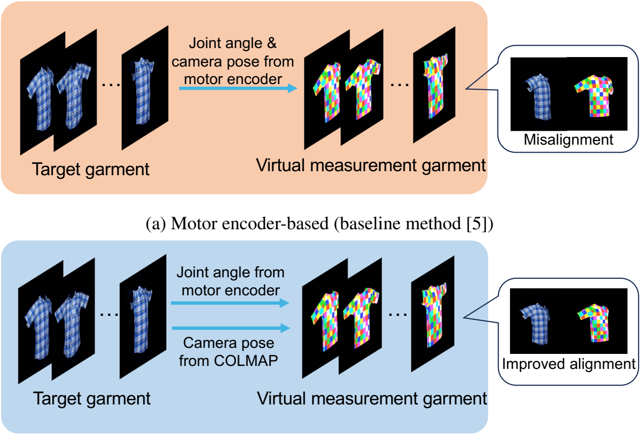
$$I^v = R(\mathbf{V}) \tag{4}$$

where $R(\cdot)$ denotes the rendering function, $\mathbf{V}$ represents the vertices of the deformed virtual measurement garment.

Subsequently, the rendered virtual measurement garment image I^v is transformed to the target garment image I^T via the pre-trained V2T network, expressed formally as:

$$I^T = V2T(I^v) \tag{5}$$

Data preparation. The original per-garment try-on method [5] used a motor encoder for generating paired images of the target garment and the virtual measurement garment is illustrated in Figure 4 (a). This method estimates the joint angles and camera pose from the motor encoder. However, we observe that the camera pose inferred from the motor encoder is inaccurate and leads to significant misalignment between the target garment and the rendered virtual measurement garment. To address this issue, we propose a reconstruction-based method that estimates the camera pose by COLMAP [28, 29] and relies on the motor encoder for joint angles, as shown in Figure 4b. Our proposed approach achieves a significant improvement in the alignment quality.



(a) Motor encoder-based (baseline method [5])

(b) Reconstruction-based (our proposed method)

Figure 4: This figure illustrates (a) the baseline method [5] and (b) our proposed method of generating the paired images of the virtual measurement garment and target garment for training the V2T network. The baseline method estimates the joint angles and camera pose from the motor encoder annotation, whereas our proposed method leverages both the motor encoder and the reconstruction results obtained by COLMAP [28, 29] to estimate the joint angles and camera pose respectively.

3.3 Composition

To seamlessly blend the synthesized garment image with the input frame, we adopt a three-step procedure. First, we leverage FashionFormer [37] to extract and erase the upper body garment from the input frame. Second, we apply Self-Correction-Human-Parsing [20] to segment the human body parts in the frame, which allows us to preserve the correct occlusion relationship between the synthesized garment and the human body. Third, we compose the synthesized garment image, denoted by I^T, with the input frame, denoted by I^r, from which the original garment has been removed.

To ensure the correct occlusion relationship between synthesized garment and body parts in the composite image, we utilize the masks represented as boolean matrices with size $H \times W$. The mask of the synthesized garment is denoted by $\mathbf{M}^s$, the mask of the original garment by $\mathbf{M}^o$, and the mask of arm skin by $\mathbf{M}^a$. We define M^c as:

$$\mathbf{M}^c = ((\neg \mathbf{M}^s) \vee \mathbf{M}^a) \wedge \mathbf{M}^o \tag{6}$$

Then we have the composite image with gaps I^g:

$$I^g_{i,j} = \begin{cases} I^r_{i,j}, & \text{if } \mathbf{M}^c_{i,j} \text{ is True} \\ I^T_{i,j}, & \text{if } \mathbf{M}^c_{i,j} \text{ is False} \end{cases} \tag{7}$$

3.4 Gap filling

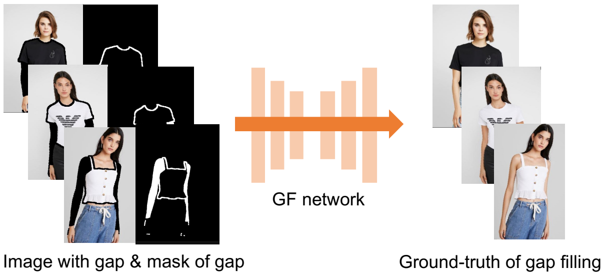


Figure 5: The Gap-Filling (GF) network is trained on the paired images of try-on images with and without gaps and their corresponding masks.

Training stage. The Gap-Filling (GF) network, which follows the architecture design of [14], is responsible for inpainting the gaps produced in the composition step of the synthesized target garment I^T and the input frame I. As illustrated in Figure 5, the network receives as input the try-on images with gaps and the corresponding gap masks, and produces as output the try-on images without gaps. This training procedure enables the GF network to learn how to generate realistic pixels for the missing regions.

Inference stage. After the composition, we obtain the fused frame and the gap mask. To remove the black gap in the fused frame, we feed both the fused frame and the gap mask to the GF network to fill in the black gap.

Data preparation. To train the GF network that can fill the gap in composite images of virtual try-on, we need to synthesize realistic images with missing regions that resemble the composite images of virtual try-on. We leverage existing virtual try-on datasets, DeepFashion2 [8] and VITON-HD [4], as the ground-truth for the gap-filling task. We employ FashionFormer [37] and Self-Correction-Human-Parsing [20] to extract the mask of the upper body garment and arms from the input image, and then we generate the gaps by applying morphological operations on the mask.

4 Experiments

4.1 Experiment setup

Dataset for evaluation. To demonstrate that our method generates results with temporal consistency, we require a video dataset for the evaluation process. Existing video virtual try-on methods use the VVT dataset [7] for evaluation. However, this dataset is not suitable for our method, because our method requires multi-view images of the target garment for training, while the VVT dataset only provides front-view images of the target garment. Therefore, we collected a new dataset that meets our requirements. Our dataset consists of several videos of people wearing target garments that have been used for training our method.

Training. The V2T and GF networks are trained separately and then combined together to generate the virtual try-on images. The V2T network is trained with a learning rate of $l_r = 0.0001$ for 10 epochs, using the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We also apply random affine transformation for data augmentation. The GF network adopts the same training settings as the V2T network. Both networks take approximately one day to complete the training process on an NVIDIA A100 GPU.

4.2 Alternative methods

We conduct comparisons of our proposed method with HR-VITON [19], a state-of-the-art image-based method for virtual try-on, and two alternative methods that we use for ablation study:

- **DP**: This method adopts the part-specific UV coordinates estimated by DensePose [24] as an intermediate representation for garment synthesis, instead of using a virtual measurement garment. The garment synthesis pipeline is presented in Figure 6 (a). The purpose of this method is to evaluate the role of the virtual measurement garment in the proposed framework.

- **TF**: This method employs a virtual measurement garment without any grid pattern texture as an intermediate representation for garment synthesis, as shown in Figure 6 (b). The aim of this method is to assess the impact of the grid pattern texture on the quality of the synthesized images.

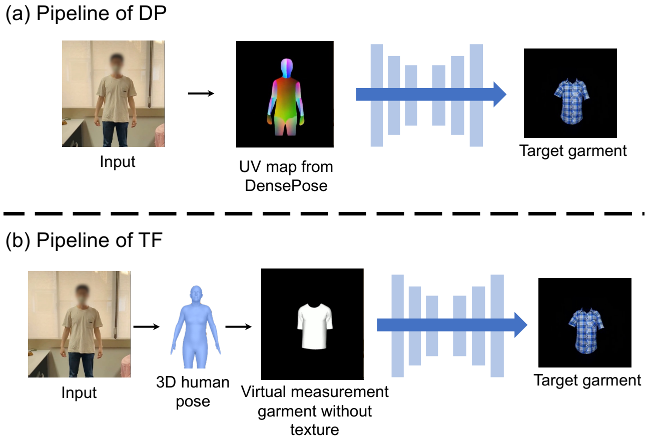


Figure 6: Two alternative methods for ablation study. (a) DP method uses UV map estimated by DensPose [24] as an intermediate representation for the garment synthesis. (b) TF method uses a virtual measurement garment without any grid pattern texture as an intermediate representation for the garment synthesis.

4.3 Results

Qualitative comparison. Figure 7 presents a qualitative comparison of the image quality among HR-VITON, DP, TF, and our proposed method. It can be observed from the figure that our method effectively preserves the fine details and the overall shape of the target garment, whereas the other methods introduce noticeable distortions.

In addition, we compare the temporal consistency of our method with HR-VITON, DP, and TF in Figure 8 (also see supplementary video). As can be seen from this figure, our method preserves the shape and appearance of the target garment across different frames, while the other methods suffer from noticeable distortions and flickering. Therefore, our method achieves superior temporal consistency compared to other methods.

Quantitative comparison. Figure 9 illustrates our quantitative evaluation procedure. We capture videos of people wearing the physical target garment and use different virtual try-on methods to replace it with a synthesized target garment. We then compute the similarity between the input and output frames using quantitative metrics. We apply quantitative metrics for both image quality and temporal consistency. For image results, we employ the structural similarity (SSIM) [35] and the learned perceptual image patch similarity (LPIPS) [40] to evaluate image quality. For video results, we use the Video Frechet Inception Distance (VFID) to measure visual quality and temporal consistency in the unpaired setting, and both temporal and spatial features are extracted by a pre-trained video recognition CNN backbones: I3D [1].

The quantitative results of our method and other methods are presented in Table 1, where we compare the image quality and temporal consistency metrics. It can be seen that our method achieves the best performance in all aspects, indicating that our method can benefit video virtual try-on research by effectively maintaining temporal consistency.

Table 1: Quantitative comparison on our collected video dataset. For SSIM, a higher value indicates better image quality. For LPIPS and VFID, a lower value implies better image quality and temporal consistency.

Method	SSIM ↑	LPIPS ↓	VFID ↓
DP	0.874	0.065	5.982
TF	0.881	0.057	4.156
HR-VITON [19]	0.855	0.081	10.091
Ours	**0.886**	**0.056**	**3.441**

Efficiency comparison. To evaluate the efficiency of our proposed method, we compare it with HR-VITON [19]. Both methods are implemented using Pytorch and run on a PC with one NVIDIA A100 GPU and AMD EPYC 73F3 16-Core Processor CPU. We use the official implementation [1] of HR-VITON for the comparison. Since HR-VITON only supports an output resolution of 1024×768, we test our method at 3 different resolutions: 512 × 512, 1024 × 1024, and 2048 × 2048. This way, we can demonstrate the scalability and robustness of our method in terms of efficiency.

Table 2: Comparison of efficiency. The data presented in this table are the average elapsed times for processing one frame, measured in seconds.

Method	Resolution	Elapsed time (s)
HR-VITON [19]	1024 × 768	37.06
Ours	512 × 512	0.67
Ours	1024 × 1024	0.71
Ours	2048 × 2048	0.81

The efficiency comparison between our method and HR-VITON is presented in Table 2. It can be observed that our method achieves a significant improvement over HR-VITON in terms of efficiency.

4.4 Dataset requirement analysis

Unlike general image-based virtual try-on methods, our method requires a per-garment dataset for each particular target garment. According to [5], a per-garment dataset contains 83,750 images, and the capturing process using the robotic mannequin takes around four hours for each target garment.

To investigate the minimum number of images required for a per-garment dataset, we train V2T networks on various downsampled datasets and compare the results qualitatively and quantitatively. We downsampled the original dataset uniformly by ratios of $d = 1, 1/2, 1/4, 1/8, \ldots, 1/16384$, and then replicating the remaining

[1] https://github.com/lastdefiance20/TryYours-Virtual-Try-On

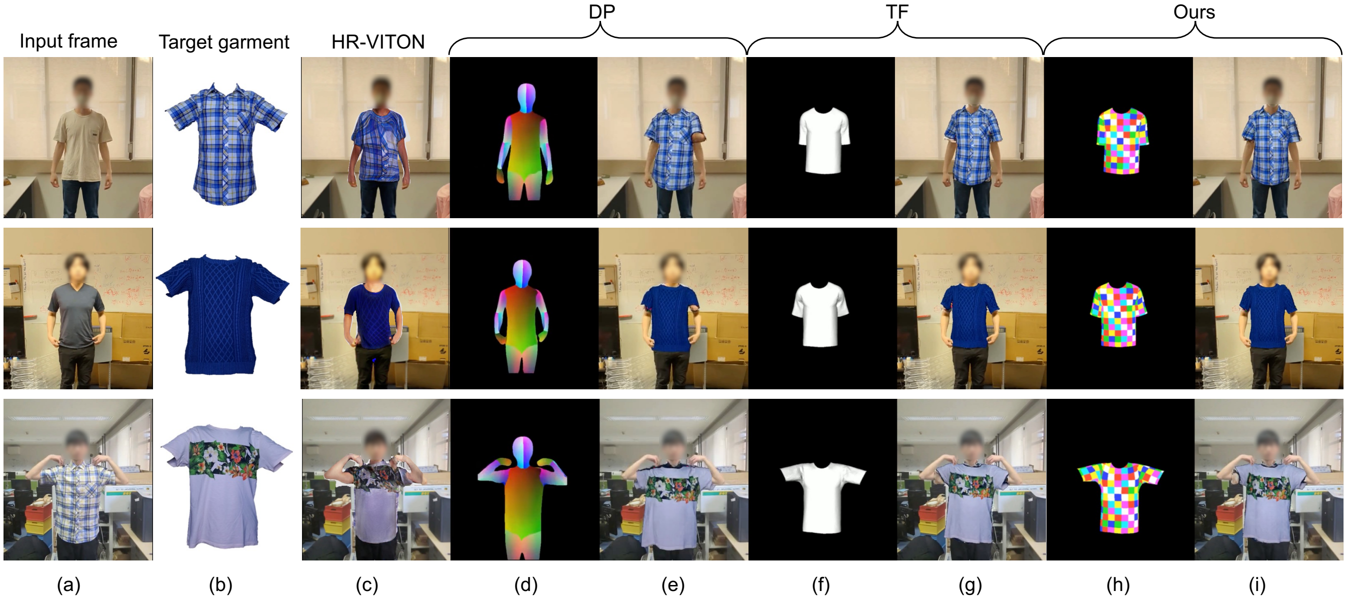


Figure 7: Image quality comparison with HR-VITON [19] and two alternative methods, DP and TF, for ablation study. DP is the method that uses the UV coordinate map estimated by DensePose as an intermediate representation, as shown in column (e). TF is the method that uses a virtual measurement garment without texture as an intermediate representation, as shown in column (f). Our method preserves the details of the target garment and exhibits less distortion than other methods.

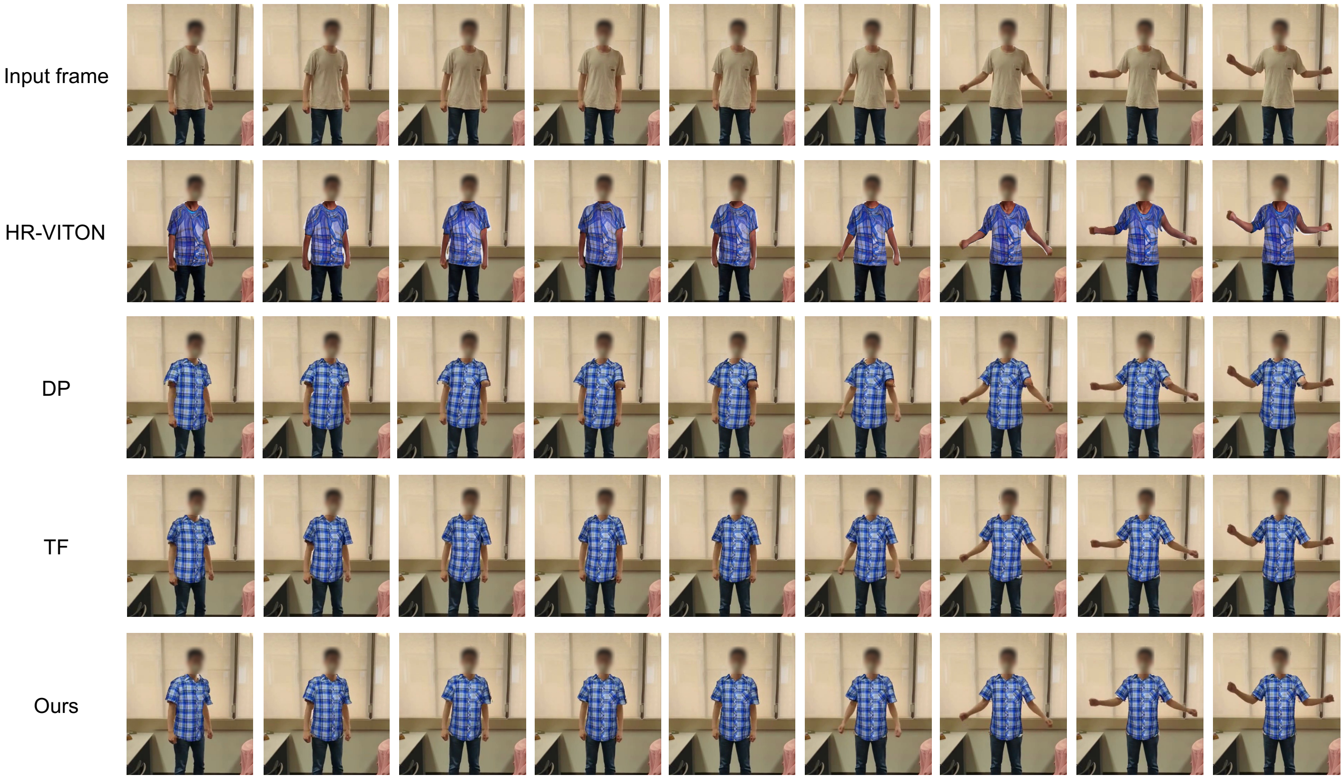


Figure 8: Temporal consistency comparison with HR-VITON [19] and two alternative methods, DP and TF, for ablation study. The first row shows the input frames that are adjacent frames extracted from a video. Our method achieves superior temporal consistency compared to other methods.

images by $1/d$ times to preserve the dataset size. We keep the other training parameters unchanged across different datasets. We showed the qualitative and quantitative results in Figure 10 and Figure 11. In Figure 11, we can observe that the visual quality and

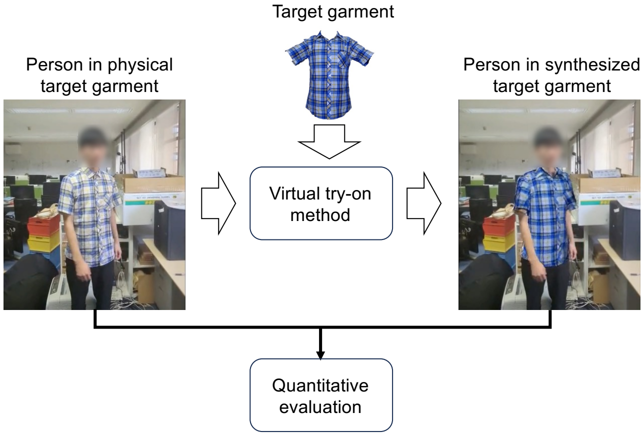

Target garment
Person in physical target garment
Person in synthesized target garment
Virtual try-on method
Quantitative evaluation

Figure 9: Illustration of how we conduct quantitative evaluation. We capture videos of people wearing the physical target garment and use virtual try-on methods to replace it with a synthesized target garment. We then measure the similarity between the input and output frames using quantitative metrics.

temporal consistency decrease significantly after $d < 1/512$. The experimental results demonstrate that our V2T network achieves satisfactory performance with a subset of the full dataset. However, the full dataset remains useful for more sophisticated networks that can exploit its rich information.

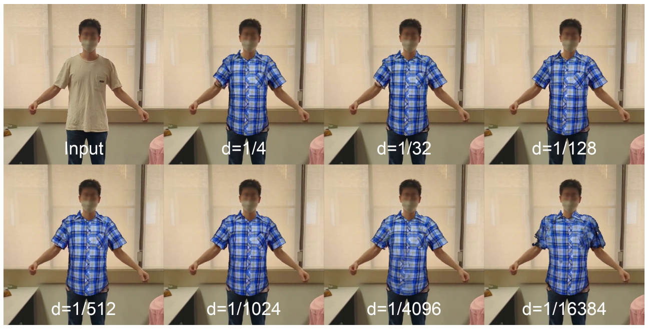


Figure 10: Some results obtained from V2T networks trained on diverse downsampled datasets, where the downsampling ratio d varies. The figure illustrates the impact of different downsampling levels on network performance.

5 Conclusion

In this paper, we present a novel per-garment virtual try-on method that utilizes a virtual measurement garment. Our method differs from existing image-based and per-garment virtual try-on methods in that it employs a virtual measurement garment model to guide the synthesis of the target garment under various poses, without requiring additional 3D measurements. Moreover, our method can generate high-quality and temporally consistent garment images, even when each frame is processed independently. Our main technical contributions are: (1) the introduction of the virtual measurement garment; (2) the introduction of the gap-filling module; and (3) reconstruction-based registration. We conduct qualitative and quantitative experiments to show that our method surpasses the state-of-the-art image-based virtual try-on method in terms of quality and temporal consistency. Furthermore, efficiency comparison reveals that our method is significantly faster even when setting the resolution higher than the existing method. We also perform an ablation study to validate the effectiveness of components of our method.

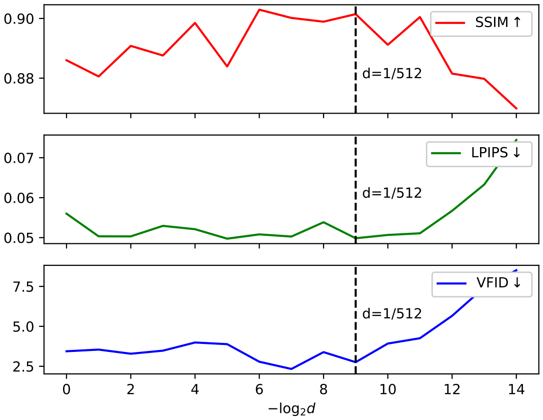


Figure 11: This figure illustrates the relationship between quantitative metrics and $-\log_2 d$, where d denotes the downsampling ratio. Significant image quality and temporal consistency degradation are observed when $d < 1/512$.

Our current method has the following limitations that we plan to address in the future:

Limited garment types support Although Our method can be applied to long-sleeve garments by constructing a virtual measurement garment model with full arms, there is no per-garment dataset available for long-sleeve garments. The reason is because the existing robotic mannequin from [5] has only upper arms and cannot capture images of long-sleeve garments. In the future, we plan to develop a new robotic mannequin with full arms to overcome this problem.

Inability to depict the fittingness. Our method cannot accurately depict how well a target garment fits the user. This is because we rely on a simple assumption that the user has an average body shape, which may not reflect the actual body shape of the user. Estimating precise body shapes from 2D images is inaccurate because garments draped on people conceal their body shape. A possible future direction would be to develop a learning-based method to estimate the body shape from user input, such as weight, height, and gender.

Lack of user-garment interaction. Our method does not support user-garment interaction, such as stretching or pulling the target garment. This is because we use a virtual measurement garment that is not physically attached to the user, unlike the physical measurement garment used in [5]. A possible future direction would be to run cloth simulation for the virtual measurement garment and to use off-the-shelf gesture recognition tools to identify if the user is interacting with the garment by grabbing.

Inability to capture time-dependent appearance. Our method uses a single frame of the virtual measurement garment as input to the garment synthesis network. Consequently, it cannot generate time-dependent behavior of the garment from a sequence of movement. To address this limitation, future work could involve collecting a new dataset with temporal annotations, and training a recurrent network to learn the time-dependent appearance of the garment.

Inability to handle illumination difference. The proposed method does not account for the difference in lighting conditions between the training and inference stages, which affects its ability to synthesize the correct appearance of the target garment according to the user's lighting condition. A possible solution to ensure the realism of the target garment's appearance is to capture its material parameters and measure the user's lighting condition. However, this solution would require costly equipment and restrict its applicability. An alternative and more affordable way to address the issue of lighting difference is to apply image harmonization techniques for image composition.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback. This work was supported by the Institute for AI and Beyond of the University of Tokyo. Zaiqiang Wu was also supported by JST SPRING, Grant Number JPMJSP2108.

References

[1] J. Carreira and A. Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

[2] A. Casado-Elvira, M. C. Trinidad, and D. Casas. Pergamo: Personalized 3d garments from monocular video. In *Computer Graphics Forum*, vol. 41, pp. 293–304. Wiley Online Library, 2022.

[3] C.-Y. Chen, L. Lo, P.-J. Huang, H.-H. Shuai, and W.-H. Cheng. Fashionmirror: Co-attention feature-remapping virtual try-on with sequential template poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13809–13818, 2021.

[4] S. Choi, S. Park, M. Lee, and J. Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14131–14140, 2021.

[5] T. Chong, I.-C. Shen, N. Umetani, and T. Igarashi. Per garment capture and synthesis for real-time virtual try-on. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pp. 457–469, 2021.

[6] G. Cirio, J. Lopez-Moreno, D. Miraut, and M. A. Otaduy. Yarn-level simulation of woven cloth. *ACM Transactions on Graphics (TOG)*, 33(6):1–11, 2014.

[7] H. Dong, X. Liang, X. Shen, B. Wu, B.-C. Chen, and J. Yin. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1161–1170, 2019.

[8] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo. A versatile benchmark for detection, pose estimation, segmentation and re-identification of clothing images. *CVPR*, 2019.

[9] A. Grigorev, M. J. Black, and O. Hilliges. Hood: Hierarchical graphs for generalized modelling of clothing dynamics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16965–16974, 2023.

[10] P. Guan, L. Reiss, D. A. Hirshberg, A. Weiss, and M. J. Black. Drape: Dressing any person. *ACM Transactions on Graphics (ToG)*, 31(4):1–10, 2012.

[11] O. Halimi, E. Larionov, Z. Barzelay, P. Herholz, and T. Stuyck. Physgraph: Physics-based integration using graph neural networks. *arXiv preprint arXiv:2301.11841*, 2023.

[12] O. Halimi, T. Stuyck, D. Xiang, T. Bagautdinov, H. Wen, R. Kimmel, T. Shiratori, C. Wu, Y. Sheikh, and F. Prada. Pattern-based cloth registration and sparse-view animation. *ACM Transactions on Graphics (TOG)*, 41(6):1–17, 2022.

[13] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 7543–7552, 2018.

[14] S. Iizuka, E. Simo-Serra, and H. Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017.

[15] N. Jetchev and U. Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *Proceedings of the IEEE international conference on computer vision workshops*, pp. 2287–2292, 2017.

[16] J. Jiang, T. Wang, H. Yan, and J. Liu. Clothformer: Taming video virtual try-on in all module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10799–10808, 2022.

[17] J. M. Kaldor, D. L. James, and S. Marschner. Simulating knitted cloth at the yarn level. In *ACM SIGGRAPH 2008 papers*, pp. 1–9. 2008.

[18] Z. Lahner, D. Cremers, and T. Tung. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 667–684, 2018.

[19] S. Lee, G. Gu, S. Park, S. Choi, and J. Choo. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*, pp. 204–219. Springer, 2022.

[20] P. Li, Y. Xu, Y. Wei, and Y. Yang. Self-correction for human parsing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. doi: 10.1109/TPAMI.2020.3048039

[21] R. Narain, A. Samii, and J. F. O'brien. Adaptive anisotropic remeshing for cloth simulation. *ACM transactions on graphics (TOG)*, 31(6):1–10, 2012.

[22] X. Pan, J. Mai, X. Jiang, D. Tang, J. Li, T. Shao, K. Zhou, X. Jin, and D. Manocha. Predicting loose-fitting garment deformations using bone-driven motion networks. In *ACM SIGGRAPH 2022 Conference Proceedings*, pp. 1–10, 2022.

[23] C. Patel, Z. Liao, and G. Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7365–7375, 2020.

[24] I. K. Rıza Alp Güler, Natalia Neverova. Densepose: Dense human pose estimation in the wild. 2018.

[25] I. Santesteban, M. A. Otaduy, and D. Casas. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, vol. 38, pp. 355–366. Wiley Online Library, 2019.

[26] I. Santesteban, M. A. Otaduy, and D. Casas. Snug: Self-supervised neural dynamic garments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8140–8150, 2022.

[27] I. Santesteban, N. Thuerey, M. A. Otaduy, and D. Casas. Self-supervised collision handling via generative 3d garment models for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11763–11773, 2021.

[28] J. L. Schönberger and J.-M. Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[29] J. L. Schönberger, E. Zheng, M. Pollefeys, and J.-M. Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.

[30] M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama. Virtual fitting by single-shot body shape estimation. In *Int. Conf. on 3D Body Scanning Technologies*, pp. 406–413. Citeseer, 2014.

[31] A. Selle, J. Su, G. Irving, and R. Fedkiw. Robust high-resolution cloth using parallelism, history-based collisions, and accurate friction. *IEEE transactions on visualization and computer graphics*, 15(2):339–350, 2008.

[32] Y. Sun, W. Liu, Q. Bao, Y. Fu, T. Mei, and M. J. Black. Putting People in their Place: Monocular Regression of 3D People in Depth. In *CVPR*, 2022.

[33] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 589–604, 2018.

[34] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8798–8807, 2018.

[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[36] D. Xiang, F. Prada, T. Bagautdinov, W. Xu, Y. Dong, H. Wen, J. Hodgins, and C. Wu. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics (TOG)*, 40(6):1–15, 2021.

[37] S. Xu, X. Li, J. Wang, G. Cheng, Y. Tong, and D. Tao. Fashionformer: A simple, effective and unified baseline for human fashion

segmentation and recognition. *ECCV*, 2022.

[38] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, and P. Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 7850–7859, 2020.

[39] R. Yu, X. Wang, and X. Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10511–10520, 2019.

[40] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.