Low-Barrier Dataset Collection with Real Human Body for Interactive Per-Garment Virtual Try-On

Zaiqiang Wu
The University of Tokyo
Japan

Yechen Li
The University of Tokyo
Japan

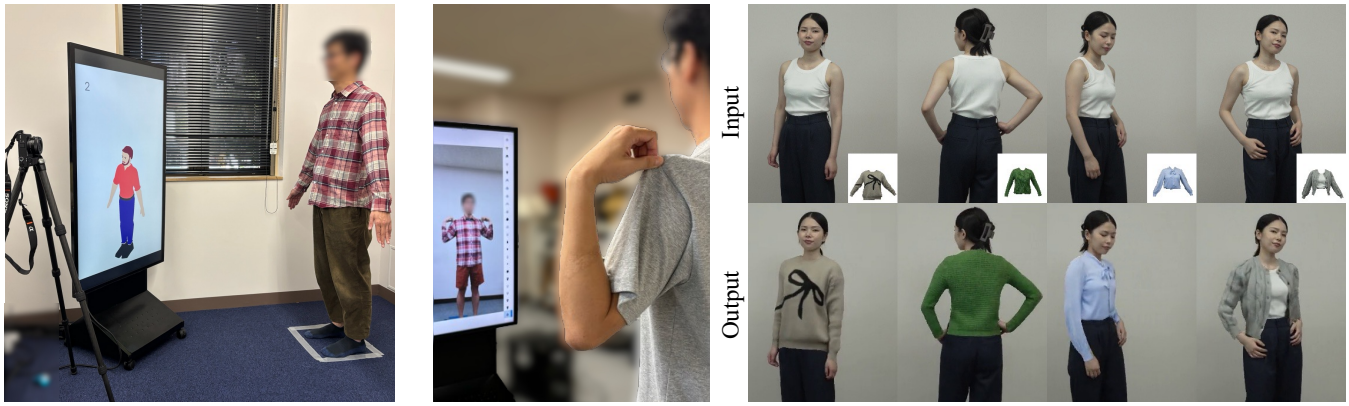
Jingyuan Liu
The University of Tokyo
Japan

Yuki Shibata
SoftBank Corp
Japan

Takayuki Hori
SoftBank Corp
Japan

I-Chao Shen
The University of Tokyo
Japan

Takeo Igarashi
The University of Tokyo
Japan



(a) Dataset collection with human body (b) Interactive virtual try-on (c) Virtual try-on results

Figure 1. (a) Our low-barrier dataset collection method requires only a camera, a monitor, and a human body wearing the target garment. No customized devices are required. (b) Our system is efficient, achieving a frame rate of approximately 8 frames per second on a moderate PC, thereby supporting interactive virtual try-on experiences. (c) Our system can generate highly plausible virtual try-on results for various target garments from different viewpoints, including side and back.

Abstract

Existing image-based virtual try-on methods are often limited to the front view and lack real-time performance. While per-garment virtual try-on methods have tackled these issues by capturing per-garment datasets and training per-garment neural networks, they still encounter practical limitations: (1) the robotic mannequin used to capture per-garment datasets is prohibitively expensive for widespread adoption and fails to accurately replicate natural human body deformation; (2) the synthesized garments often misalign with the human body. To address these challenges, we propose a low-barrier approach for collecting per-garment datasets using real human bodies, eliminating the necessity for a customized robotic mannequin. We also introduce a hybrid person representation that enhances the existing intermediate representation with a simplified DensePose map. This ensures accurate alignment of synthesized garment images with the human body and enables human-garment interaction without the need for customized wearable devices. We performed qualitative and quantitative evaluations against other state-of-the-art image-based virtual try-on methods and conducted ablation studies to demonstrate the superiority of our method regarding image quality and temporal consistency. Finally, our user study results indicated that most participants found our virtual try-on system helpful for making garment purchasing decisions.

CCS Concepts: • **Computing methodologies** → **Image processing**; • **Human-centered computing** → **Mixed / augmented reality**.

Keywords: Virtual Try-On, Image Synthesis, Human-in-the-loop Machine Learning

1 Introduction

Virtual try-on technology, which enables users to try on garments virtually without needing physical access to them,

has recently garnered significant attention from researchers. Existing image-based virtual try-on methods [5, 8, 9, 17, 20, 26] use a 2D image of a target garment to create realistic virtual try-on results. However, these methods face several limitations that hinder their practical application in real-world scenarios. Firstly, they often produce poor results for general users because they are typically trained on biased datasets that include only fashion models and a limited range of poses. Secondly, their complexity prevents them from running in real-time, significantly impacting the virtual try-on experience.

A subfield of image-based methods, per-garment virtual try-on methods [6, 43], addresses these challenges by collecting garment-specific datasets and training garment-specific neural networks. These methods leverage per-garment datasets collected using a robotic mannequin, enabling the capture of massive paired images of the target and measurement garments in identical poses. Despite the advantages of using a robotic mannequin, several limitations persist:

- The robotic mannequin fails to replicate the natural human body dynamics and the garment-skin interactions. This results in unnatural wrinkles appearing on the target garment, as shown in Figure 2.
- The high cost of the robotic mannequin limits its widespread use and poses a major barrier to dataset collection.
- Significant differences in appearance between the robotic mannequin and real humans hinder the application of existing human-centric vision models for generating annotations to guide alignment, resulting in misalignment between the synthesized garment and the human body.

To address these limitations, we propose a low-barrier dataset collection method that uses real human bodies instead of a customized robotic mannequin. Our approach requires only a camera, a monitor, and a human model wearing the target garment, as shown in Figure 1(a). Non-experts can create their own virtual try-on model for a specific garment by physically wearing it and demonstrating how it looks under various poses to a camera, which can be regarded as a form of human-in-the-loop machine learning. Our experiments indicate that it takes approximately two minutes for a real human, without prior professional training, to complete the task, whereas using a customized robotic mannequin requires around two hours [6].

Additionally, using real humans allows us to enhance the prior intermediate representation [43] with a simplified DensePose [14] map, ensuring precise alignment between the synthesized garment and the human body. Furthermore, DensePose's ability to capture rough garment deformation allows users to interact with the synthesized garment by pulling on the garment they are wearing, eliminating the need for a customized measurement garment as required by [6]. Moreover, the flexibility of real humans facilitates the collection of datasets for a wider variety of garment types and poses, enabling our methods to generate plausible results for various garments across diverse poses, as illustrated in Figure 1(c).

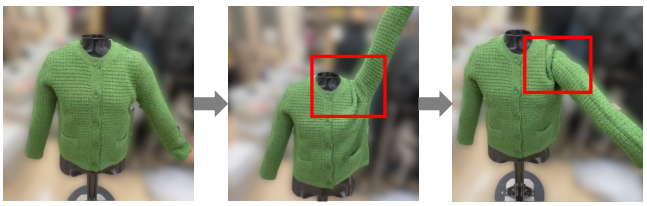


Figure 2. Unnatural wrinkles of the target garment occur around the shoulder region when collecting per-garment datasets using a customized robotic mannequin.

Our method achieves a frame rate of approximately 8 frames per second on a moderate PC, enabling an interactive virtual try-on experience, as shown in Figure 1(b). Both qualitative and quantitative evaluations indicate that our method surpasses existing approaches in terms of visual quality and temporal consistency.

Our contributions can be summarized as follows:

- We propose a low-barrier dataset collection approach that eliminates the need for expensive customized robotic mannequins, significantly reducing the cost of per-garment dataset collection.
- We enhance the virtual measurement garment [43] by incorporating a simplified DensePose map, ensuring precise human-garment alignment and facilitating human-garment interaction without requiring customized wearable devices.
- We perform extensive qualitative and quantitative evaluations, along with a user study, to assess the usability and utility of our virtual try-on system.

2 Related Work

2.1 3D model-based Virtual Try-on

3D model-based virtual try-on methods [13, 29, 30, 32, 35] involve creating a digital representation of the human body and developing 3D models of clothing items. The try-on results are generated by simulating the draping, fitting, and movement of the clothing items on the 3D body model, utilizing either physically-based simulations [7, 19, 28, 36] or data-driven methods [3, 12, 15, 24, 29, 30, 32–34, 44]. Despite their ability to realistically capture garment deformations across diverse poses and viewpoints, these methods face two critical limitations: (1) They require time-consuming manual 3D modeling for each garment, which significantly hinders scalability; (2) They primarily generate try-on results for digital avatars rather than real users, leading to substantial misalignment when overlaying the 3D garment on human bodies, thereby reducing their effectiveness for personalized

purchase decisions. In contrast, our method can automatically generate the necessary data for each garment item with minimal human intervention and ensure precise alignment between the synthesized garment and the human body without the need for depth sensors or wearable devices.

2.2 Image-based Virtual Try-on

Image-based virtual try-on methods aim to synthesize a realistic image of a person wearing a specific garment using a combination of the person's image and an in-shop garment image [38]. Unlike 3D model-based methods, 2D image-based methods do not need manual 3D modeling, making them more accessible while still achieving a realistic appearance of the target garment. Some early works [5, 17, 26] are based on Generative Adversarial Networks (GAN) [11] and its variations [21, 22, 48]. More recently, many recent works [23, 40, 45] have leveraged diffusion models to improve image quality in virtual try-on methods.

While these methods generate realistic static try-on images, users often prefer to see try-on videos that showcase a person dressing and moving naturally. Initial approaches, such as FW-GAN [8], FashionMirror [4], and ClothFormer [18], incorporate optical flow for garment image warping to maintain temporal consistency over time. More recent methods, including ViViD [9], Tunnel Try-on [46], and Fashion-VDM [20], utilize image and video diffusion models to enhance try-on quality.

Typically, these methods train a universal network to accommodate all types of garments using human identity-aware representations and target garment images. As a result, they need a large dataset that includes a diverse range of human identities and garment styles to ensure generalizability. Moreover, these methods face challenges such as slow processing speeds, data bias towards front-view fashion models, and issues with altering human identity. In contrast, our method focuses on generating high-fidelity and temporal consistent garments from multiple viewpoints while preserving the user's human identity.

2.3 Per-garment Virtual Try-on

Per-garment virtual try-on methods achieve superior try-on results for specific target garments by training garment-specific networks. The pioneer work by Chong et al. [6] employed a robotic mannequin to collect per-garment datasets and trained image-to-image translation networks. These networks map the images of a physical measurement garment to the images of the target garment. While this method has demonstrated superior try-on results for specific garments, it requires users to wear a customized measurement garment, limiting its application. Wu et al. [43] proposed an alternative method that estimates a 3D human pose and deforms a virtual measurement garment mesh accordingly, serving as an intermediate representation and eliminating the need for a physical measurement garment. However, this approach suffers from noticeable misalignment between the synthesized garment and the human body. Most importantly, both approaches rely on using a customized robotic mannequin for dataset collection, which hinders their reproducibility. Our method overcomes this limitation by capturing per-garment datasets using real human bodies, significantly reducing the barrier to data collection.

2.4 Human-in-the-loop Machine Learning

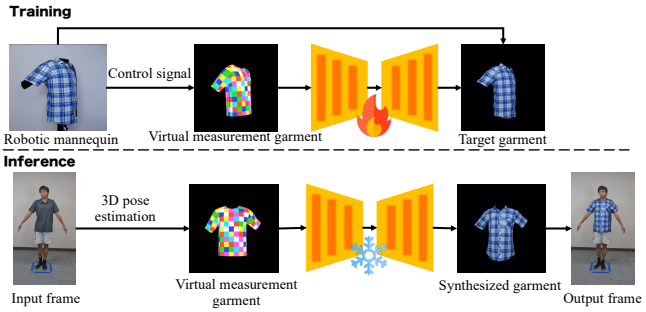
Human-in-the-loop machine learning aims to enhance the performance of machine learning models by incorporating human expertise [42]. Interactive Machine Teaching [31, 37], a subfield of human-in-the-loop machine learning, aims to help non-experts easily create their own machine learning model without the need for technical skills. It aims to bridge the gap between domain knowledge holders and machine learning technology, making the model creation process more accessible. For instance, Teachable Machine [1] allows novice users, without any machine learning skills, to transfer their knowledge of object classification to a machine learning model by presenting various views of each object to a camera. To reduce the burden on users and make it easier to highlight the object of interest, LookHere [47] employs users' deictic gestures to annotate the region of interest.

Our method can be considered as a variant of human-in-the-loop machine learning. It enables novice users, without machine learning skills or specialized devices, to create a virtual try-on model by physically wearing the target garment and demonstrating how it looks under various poses to a camera. Unlike typical systems that rely on human knowledge for annotation or instruction for training, our system leverages the users' physical bodies.

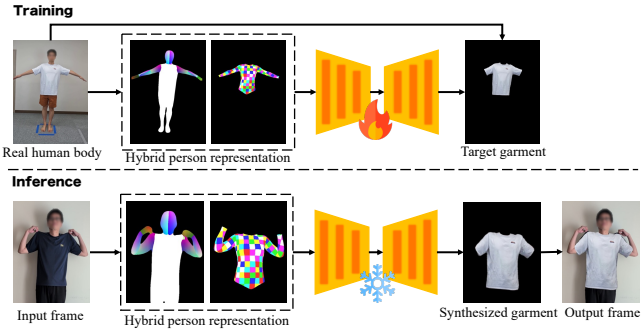
3 Method

Our method builds upon the previous per-garment virtual try-on approach by Wu et al. [43] and follows a similar image-to-image translation workflow. As illustrated in Figure 3(a), during training, Wu et al. [43] utilized a robotic mannequin and its control signals to generate paired images of the virtual measurement garment, which only contained pose information, and the target garment for training the per-garment network. During inference, they extracted the virtual measurement garment from the input frame using 3D pose estimation and fed it to the network to generate the try-on result. However, the absence of alignment information led to misalignment issues in their results (Figure 11(a)).

In contrast, our method employs real human bodies to generate per-garment datasets for training the per-garment networks, as depicted in Figure 3(b). We propose to use hybrid person representation $\mathbf{I}_{hybrid} \in \mathbb{R}^{6\times H\times W}$ as the intermediate representation:



(a) Training/inference pipeline of Wu et al. [43].



(b) Training/inference pipeline of our method.

Figure 3. Training/inference pipeline comparison: (a) prior per-garment method [43] vs. (b) our method.

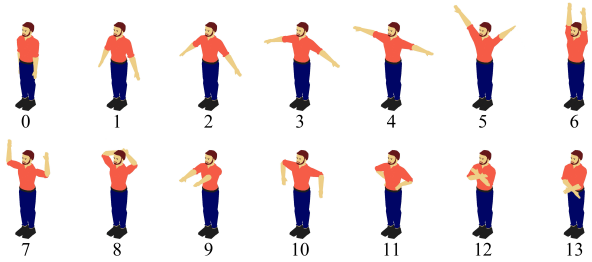


Figure 4. The predefined pose set we used during the data capturing procedure.

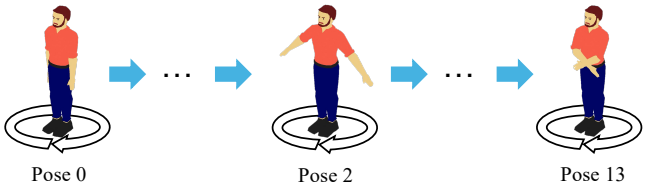


Figure 5. The data capturing procedure: The person replicates each predefined pose, rotating 360 degrees slowly in place for each pose.

$$\mathbf{I}_{hybrid} = \mathbf{I}_{vm} \oplus \mathbf{I}_{sdp} \tag{1}$$

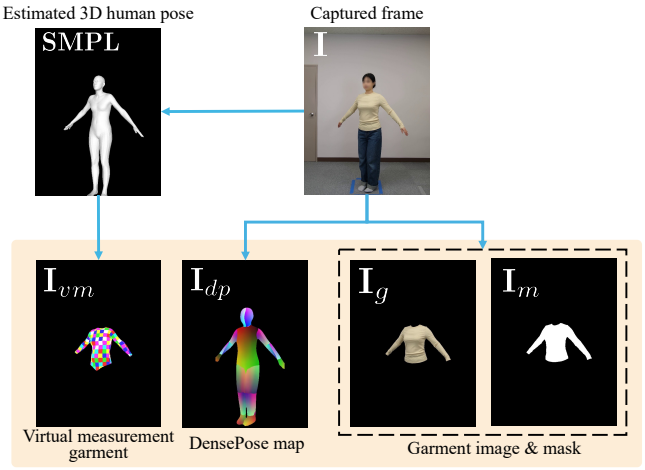


Figure 6. Per-garment dataset generation from recorded video of a person performing predefined poses.

where $\oplus$ denotes channel-wise concatenate. In this representation, $\mathbf{I}_{vm}$ denotes the image of the virtual measurement garment, which provides the annotation for the 3D human pose, while $\mathbf{I}_{sdp}$ denotes the simplified DensePose [14] map, which provides the annotation for 2D image space alignment.

To obtain $\mathbf{I}_{vm}$ from a person image $\mathbf{I}$, we first estimate its 3D pose represented by a SMPL [27] mesh. We then remove the unnecessary body parts (head, hands, and lower body) and texture the remaining parts with a grid pattern, and render it into a virtual measurement garment image $\mathbf{I}_{vm} \in \mathbb{R}^{3\times H\times W}$. Unlike [43], our method employs the SMPL mesh with full arms, which allows our method to handle both short-sleeved and long-sleeved garments.

To obtain $\mathbf{I}_{sdp}$ from a person image $\mathbf{I}$, we estimate the DensePose map $\mathbf{I}_{dp}$ from the input image. We then simplify the DensePose map by painting the torso and lower body parts white, resulting in the simplified DensePose map $\mathbf{I}_{sdp} \in \mathbb{R}^{3\times H\times W}$:

$$\mathbf{I}_{sdp} = SIM(\mathbf{I}_{dp}) \tag{2}$$

where $SIM(\cdot)$ denotes the simplification process. This simplification is necessary because the boundary between the upper and lower body in the original DensePose map is not robust, causing flickering artifacts. This simplification improves temporal consistency and reduces artifacts, as demonstrated in Section 4.4.

During inference, we extract the hybrid person representation $\mathbf{I}_{hybrid}$ and feed it to the per-garment network to generate the synthesized target garment image and garment mask. Then, we composite the synthesized target garment image onto the input frame using the predicted garment mask. With the simplified DensePose map, the synthesized garment image aligns accurately with the human body. Therefore, our

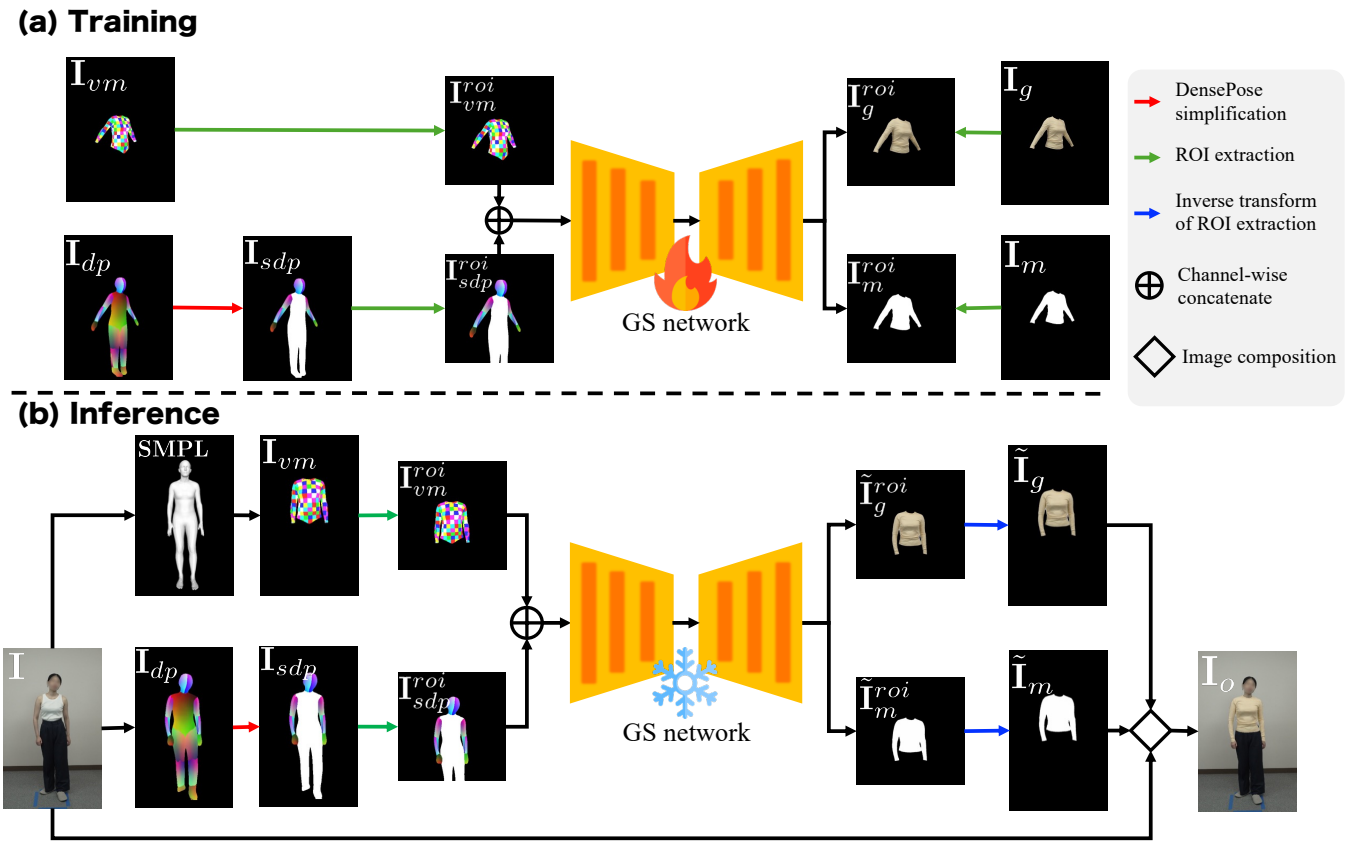


Figure 7. The details of the (a) training and (b) inference stages of our proposed per-garment virtual try-on method.

(a) Chong et al. [6] **(b)** Ours

Figure 8. Human-garment interaction: (a) prior per-garment method (Chong et al. [6]) vs. (b) our method. Chong et al. [6] requires a customized measurement garment, whereas our method does not require such a garment, thanks to the incorporation of the simplified DensePose map.

method does not require any post-processing to achieve accurate human-garment alignment.

3.1 Per-garment dataset collection with a human body

We employ a Garment Synthesis (GS) network to synthesize target garment images. The input is hybrid person representation $\mathbf{I}_{hybrid}$ and the outputs are a target garment image $\mathbf{I}_g \in \mathbb{R}^{3\times H\times W}$ and a mask $\mathbf{I}_m \in \mathbb{R}^{1\times H\times W}$:

$$\mathbf{I}_g, \mathbf{I}_m = GS(\mathbf{I}_{hybrid}) \tag{3}$$

$$= GS(\mathbf{I}_{vm} \oplus SIM(\mathbf{I}_{dp})) \tag{4}$$

Therefore, to train the GS network, we need to collect a dataset that includes target garment image $\mathbf{I}_g$, target garment mask $\mathbf{I}_m$, virtual measurement garment image $\mathbf{I}_{vm}$, and the DensePose map $\mathbf{I}_{dp}$.

In contrast to previous research [6, 43] that utilized a robotic mannequin and took about two hours to collect a per-garment dataset, our method employs real humans for the following reasons:

- Try-on movements exhibit limited diversity, allowing them to be effectively represented by a small set of predefined poses, which can be performed within a short, tolerable duration for real humans.

- Developing a robotic mannequin that can faithfully replicate human body deformation caused by the interactions between bone, flesh, and skin is challenging and costly.
- Significant differences in appearance between robotic mannequins and real humans make it difficult to use existing human-centric vision models for generating alignment annotations. By using real humans, we can utilize the simplified DensePose map $\mathbf{I}_{sdp}$ as the annotation for alignment.

Data capturing environment setup. As shown in Figure 1(a), we positioned a camera at chest height and at a sufficient distance to ensure the entire body is captured within the frame. Additionally, we placed a large monitor diagonally in front of the person to display the predefined poses, making it easier for them to follow along without needing to remember all the poses. Natural indoor lighting was used for the video recording. The recording of the participant performing the predefined movements was conducted in 4K resolution at 30 fps, and the entire process took approximately two minutes.

Data capturing procedure. We designed a set of predefined poses for the per-garment dataset collection, as illustrated in Figure 4. There are 14 predefined symmetrical poses in total. Despite the limited number and the absence of asymmetrical poses, experimental results demonstrate that our trained model generalizes well to a wide range of try-on movements. To collect a dataset for a target garment, we instruct a person to wear the garment and mimic each predefined pose in order. If the garment is occluded by hair, we ask the person to tie his/her hair up. For each pose, the person slowly turns 360° in place, allowing us to capture the garment's appearance from multiple perspectives (Figure 5). Notably, the users only need to perform each pose roughly because we will utilize 3D human pose estimation to obtain the actual pose during dataset generation.

After capturing the video, we generate a per-garment dataset for the target garment by processing the video frame-by-frame, as shown in Figure 6. First, we employ BEV [39] to estimate a 3D human pose represented by an SMPL mesh and then transform it into an image of a virtual measurement garment. Then, we apply DensePose [14] to obtain the DensePose map for each frame. Finally, we obtain the target garment image and mask using Graphonomy [10], an off-the-shelf human parsing tool.

3.2 Training

We consider the garment synthesis from the hybrid person representations as an image-to-image translation task. We employ the GAN-based image-to-image translation method, pix2pixHD [41], due to its efficiency, which aligns with our requirement for real-time virtual try-on. The training pipeline is illustrated in Figure 7(a), where the GS network and discriminator architectures are derived from pix2pixHD [41].

Since we only focus on upper-body garment virtual try-on, utilizing the entire original image for training is inefficient. Therefore, we perform ROI (Region of Interest) extraction to crop the upper-body region into a square image of size s. This extraction is straightforward due to the availability of body part segmentation from DensePose maps. We perform ROI extraction on each image in the per-garment dataset:

$$\begin{cases} \mathbf{I}_{vm}^{roi} = ROI_s(\mathbf{I}_{vm}) \\ \mathbf{I}_{sdp}^{roi} = ROI_s(\mathbf{I}_{sdp}) \\ \mathbf{I}_{g}^{roi} = ROI_s(\mathbf{I}_{g}) \\ \mathbf{I}_{m}^{roi} = ROI_s(\mathbf{I}_{m}) \end{cases} \tag{5}$$

where $ROI_s(\cdot)$ denotes the image transformation that crops the original image into a square region of interest with size s. During training, we apply random affine transformations following the $ROI_s(\cdot)$ transformation as a method of data augmentation.

Consequently, the input and output pair data $(\mathbf{I}_{hybrid}^{roi}, \mathbf{I}_{gm}^{roi})$ for training are defined as:

$$\begin{cases} \mathbf{I}_{hybrid}^{roi} = \mathbf{I}_{vm}^{roi} \oplus \mathbf{I}_{sdp}^{roi} \\ \mathbf{I}_{gm}^{roi} = \mathbf{I}_{g}^{roi} \oplus \mathbf{I}_{m}^{roi} \end{cases} \tag{6}$$

For simplicity, we denote $x = \mathbf{I}_{hybrid}^{roi}$ and $y = \mathbf{I}_{gm}^{roi}$. The object function of the GAN can be expressed as:

$$\mathcal{L}_{GAN} = \mathbb{E}[\log D(x, y)] + \mathbb{E}[\log(1 - D(x, GS(x)))] \tag{7}$$

where $GS(\cdot)$ denotes the Garment Synthesis network, $D(\cdot)$ denotes the discriminator.

Following pix2pixHD [41], we incorporate feature matching loss to stabilize the training process and VGG loss [25] to enhance the quality of the synthesized images. The feature matching loss is defined as:

$$\mathcal{L}_{FM} = FM(GS(x), y) \tag{8}$$

where $FM(\cdot)$ denotes the feature matching loss function.

The VGG loss can then be defined as:

$$\mathcal{L}_{VGG} = VGG(\tilde{\mathbf{I}}_{g}^{roi} \odot \tilde{\mathbf{I}}_{m}^{roi}, \mathbf{I}_{g}^{roi}) \tag{9}$$

where $\tilde{\mathbf{I}}_{g}^{roi}$ represents the synthesized garment image, $\tilde{\mathbf{I}}_{m}^{roi}$ denotes the synthesized garment mask, and $VGG(\cdot)$ denotes VGG loss function.

The full objective function, which combines the GAN loss, feature matching loss, and VGG loss, can be expressed as:

$$\arg\min_{GS} \left(\left(\max_{D} \mathcal{L}_{GAN} \right) + \lambda_0 \mathcal{L}_{FM} + \lambda_1 \mathcal{L}_{VGG} \right) \tag{10}$$

where λ_0 and λ_1 are coefficients that control the importance of the respective loss functions.

3.3 Inference

Figure 7(b) illustrates the inference pipeline. Given an input image $\mathbf{I}$, we generate the corresponding virtual measurement garment image $\mathbf{I}_{vm}$ and simplified DensePose map $\mathbf{I}_{sdp}$. We then extract the upper-body regions of $(\mathbf{I}_{vm}, \mathbf{I}_{sdp})$ using the body parts segmentation from the DensePose map and obtain the cropped images $(\mathbf{I}^{roi}_{vm}, \mathbf{I}^{roi}_{sdp})$:

$$\begin{cases} \mathbf{I}^{roi}_{vm} = ROI_s(\mathbf{I}_{vm}) \\ \mathbf{I}^{roi}_{sdp} = ROI_s(\mathbf{I}_{sdp}) \end{cases} \tag{11}$$

Next, we concatenate $\mathbf{I}^{roi}_{vm}$ and $\mathbf{I}^{roi}_{sdp}$ to obtain the hybrid person representation $\mathbf{I}^{roi}_{hybrid}$. This hybrid person representation is then fed into the GS network to synthesize the target garment image and garment mask:

$$\tilde{\mathbf{I}}^{roi}_g, \tilde{\mathbf{I}}^{roi}_m = GS(\mathbf{I}^{roi}_{hybrid}) \tag{12}$$

We then transform the synthesized target garment image and mask into the size of the input image using the inverse transform of the aforementioned ROI extraction:

$$\begin{cases} \tilde{\mathbf{I}}_g = ROI_s^{-1}(\tilde{\mathbf{I}}^{roi}_g) \\ \tilde{\mathbf{I}}_m = ROI_s^{-1}(\tilde{\mathbf{I}}^{roi}_m) \end{cases} \tag{13}$$

Finally, we compose the synthesized target garment image with the input image using the synthesized target garment mask to obtain the try-on result $\mathbf{I}_o$:

$$\mathbf{I}_o = \mathbf{I} \odot (\mathbf{1} - \tilde{\mathbf{I}}_m) + \tilde{\mathbf{I}}_g \odot \tilde{\mathbf{I}}_m \tag{14}$$

where $\odot$ denotes element-wise multiplication. Unlike [43], which relies solely on 3D pose estimation for coarse human-garment alignment, our approach integrates DensePose to achieve pixel-level alignment, thereby eliminating the need for post-processing.

3.4 Human-garment interaction

As shown in Figure 8(a), a previous method by Chong et al. [6] requires the user to wear a customized measurement garment to facilitate human-garment interaction. In contrast, our method enables human-garment interaction without the need for such measurement garments, as demonstrated in Figure 8(b). The user can wear any arbitrary garment and manipulate it to interact with the synthesized target garment.

Our method supports this interaction by utilizing the simplified DensePose map, which captures general garment deformation during human-garment interaction, as shown in Figure 8(b). Although DensePose is designed for dense human pose estimation and aims to be garment-invariant, it does not achieve complete garment invariance. Nevertheless, by leveraging this characteristic, our method enables human-garment interaction without the need for any wearable devices.

4 Technical Evaluation

4.1 Experiment setting

Per-garment dataset. We collected 25 per-garment datasets, including both long-sleeved and short-sleeved garments for males and females. Each dataset contains approximately 3,000 images. In addition to capturing predefined movements, we recorded a video of the same person wearing the same garment while performing free-style try-on movements lasting about 20 seconds (approximately 600 frames in total). These recordings are intended for both qualitative and quantitative evaluation purposes.

Implementation details. We set the resolution of the region of interest to 512×512 and set both λ_0 and λ_1 to 1. For each garment, we trained a per-garment model using the corresponding per-garment dataset. We trained each model on an NVIDIA RTX A6000 GPU for 100 epochs, with a batch size of 8 and a learning rate of 2×10^{-4}. We use the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The training duration for a garment-specific model was approximately 15 hours.

Comparison setting. We conduct both qualitative and quantitative comparisons of our method against OOTDiffusion [45] and ViViD [9], which represent the state-of-the-art in image-based and video virtual try-on techniques. Additionally, we perform a qualitative comparison with [43] regarding human-garment alignment, utilizing their publicly available results due to the non-reproducibility of their method.

4.2 Qualitative evaluation

In Figure 9, we provide a qualitative comparison of image quality among OOTDiffusion [45], ViViD [9], and our method. Our method produces more realistic try-on results, whereas the other methods tend to alter the garment characteristics and human identity. Furthermore, we present a qualitative comparison of temporal consistency in Figure 10. The results demonstrate that our method consistently generates plausible results across all frames, while the other methods produce varying artifacts in different frames and struggle to maintain temporal consistency.

Additionally, we compare the human-garment alignment of Wu et al. [43] with our method. Since the results of [43] are non-reproducible, we cannot compare them with the exact same target garment and person. Therefore, we selected a similar garment with collars and similar human body orientations from their supplementary video for comparison. As shown in Figure 11, the results generated by Wu et al. [43] exhibit noticeable human-garment misalignment, particularly around the collar region. In contrast, our method ensures precise human-garment alignment. Please check our supplementary video for more detailed video results.

4.3 Quantitative evaluation

To quantitatively evaluate the try-on results generated by our method and previous methods, we leverage the recorded

Figure 9. Qualitative comparison with OOTDiffusion [45] and ViViD [9]. Our method effectively preserves the details of the target garment and produces more realistic try-on results compared to the other two methods. Note that we use the ground-truth image (captured from free-style movements and not included in the training data) shown in the first row as input to demonstrate our method can faithfully recover the real-world wearing result.

video with free-style try-on movements as both input and ground truth. Although our method supports high-resolution output, for fairness in comparison, we resize the output of all methods to 512×384, the highest resolution supported by ViViD. We evaluate the image quality, video quality, and frame rate. To measure image quality, we calculated the Structural Similarity Index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) between the generated frames and the ground-truth frames. To measure video quality, we calculated Video Frechet Inception Distance (VFID) using two backbones: I3D [2] and 3DResNeXt101 [16].

As shown in Table 1(a), our method generates try-on results that outperform all other methods across all metrics. This result indicates that our method generates high-quality try-on results in real-time and maintains strong temporal consistency.

4.4 Ablation study

To verify the effectiveness of different components in our hybrid person representation, we conduct ablation studies with the following alternative intermediate representations:

- **VM**: Using only the virtual measurement garment as the intermediate representation to demonstrate the benefit of the DensePose map.
- **VMDP**: Using the virtual measurement garment and the original DensePose map as the intermediate representation to show the benefit of simplification of DensePose.
- **SDP**: Using only the simplified DensePose map as the intermediate representation to illustrate the benefit of the virtual measurement garment.

Figure 12 illustrates the qualitative ablation results. The simplified DensePose map contributes to precise alignment between the synthesized garment and the human body (Figure 12(a)). Additionally, simplifying the DensePose map reduces artifacts and enhances the temporal consistency of

Figure 10. Temporal consistency comparison with OOTDiffusion [45] and ViViD [9]. Both OOTDiffusion and ViViD generate inconsistent garments, e.g., the target garment becomes long sleeve and no sleeve. Moreover, they generate body parts that are not consistent with the input frames.

Table 1. (a) Quantitative comparison with OOTDiffusion [45], ViViD [9], and (b) ablation study on hybrid person representation. Metrics marked with ↑ indicate higher values are better, while those marked with ↓ indicate lower values are better. Our method achieves the best performance among all methods on all metrics.

Experiment	Method	SSIM ↑	LPIPS ↓	$VFID_{I3D}$ ↓	$VFID_{ResNeXt}$ ↓	fps ↑
(a)	OOTDiffusion [45]	0.851	0.086	0.459	0.334	0.21
	ViViD [9]	0.831	0.096	0.478	0.365	0.89
	Ours (**VM**+**SDP**)	**0.920**	**0.036**	**0.273**	**0.184**	**7.93**
(b)	**VM** (Virtual measurement garment only)	0.902	0.043	0.304	0.216	-
	VMDP (**VM**+original DensePose map)	0.905	0.043	0.300	0.208	-
	SDP (Simplified DensePose map only)	0.904	0.044	0.302	0.207	-

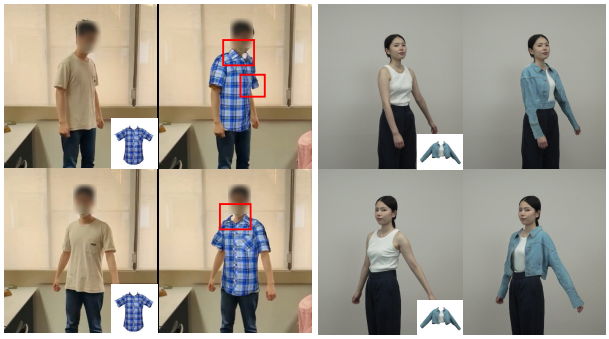
the synthesized garment by removing non-robust parts of the original DensePose map (Figure 12(b)). Furthermore, the virtual measurement garment ensures consistent orientation between the synthesized garment and the human body (Figure 12(c)). As demonstrated in Table 1(b), our proposed hybrid person representation outperforms other alternative intermediate representations by a large margin.

5 User Study

We conducted a user study to evaluate the usability of our dataset collection method and the utility of our interactive virtual try-on system. Participants were asked to try our virtual try-on system and then act as human bodies for dataset collection. We prepared two questionnaires: the first assessed participants' expectations regarding virtual try-on technology, while the second evaluated the performance of our system. We did not include comparisons with OOTDiffusion [45] or ViViD [9], as these methods are incapable of running in real-time. Additionally, comparisons with previous per-garment methods [6, 43] were omitted due to the non-reproducibility of their results, which rely on a customized robotic mannequin.

We recruited 18 female participants, aged 18 to 50, all of whom have experience purchasing garments online. These participants were non-experts with no professional training or prior knowledge of the procedure. We specifically chose female participants due to their typically stronger interest in virtual try-on technologies. Additionally, our virtual try-on prototype focuses primarily on female garments, which are generally more diverse and complex.

Each participant spent about 5 minutes trying our system. As shown in Figure 1(a), we place a large display and a web camera in front of participants, allowing them to see the



(a) Wu et al. [43] **(b)** Ours

Figure 11. Qualitative comparison of human-garment alignment: (a) prior per-garment method (Wu et al. [43]) vs. (b) our method. Wu et al. [43] produces noticeable human-garment misalignment, particularly around the collar region.

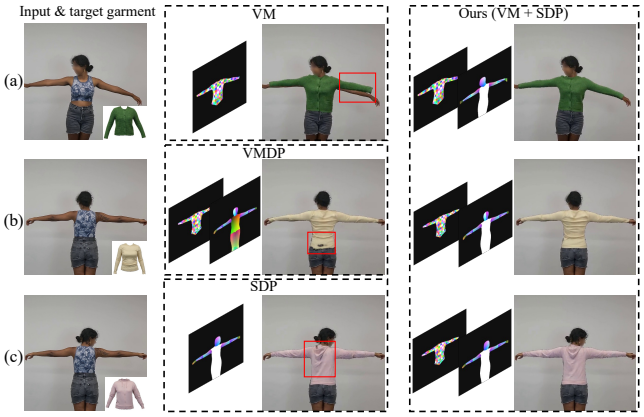


Figure 12. Qualitative ablation results for hybrid person representation. (a) Removing the simplified DensePose map leads to misalignment issues. (b) Using the original DensePose map without simplification introduces artifacts. (c) Omitting the virtual measurement garment results in inconsistent orientation of the synthesized garment with the human body.

virtual try-on results on the display in real-time. They could switch to other virtual garments by touching the screen. After experiencing our virtual try-on system, participants were asked to act as the human body to collect two per-garment datasets. Finally, we asked them to complete the questionnaires. The entire session lasted about 30 minutes, and we compensated the participants for their time.

5.1 Feedback for dataset collection

Following the instructions in Section 3.1, each participant helped collect two per-garment datasets, with each dataset taking about two minutes to collect. Overall, participants completed the dataset collection session without significant difficulty. Only one participant reported feeling "*a little bit dizzy when rotating in place*". Therefore, we believe that using real human bodies for dataset collection is feasible, even without prior professional training.

5.2 Feedback for virtual try-on system

Expectation for virtual try-on. Figure 13 presents an overview of the questions and the ratings provided by participants regarding their expectations for virtual try-on. In Q1, "For virtual try-on, static image results are insufficient; it is important to see video results" (Mean=4.28, SD=0.65), most participants agreed on the importance of video results. Additionally, they emphasized the significance of being able to see try-on results in 360° (Mean=4.61, SD=0.59) and in real-time (Mean=4.33, SD=0.67). In Q4, "Viewing virtual try-on results on a fashion model instead of yourself can lead to poor purchasing decisions" (Mean=4.33, SD=0.67), the majority of participants concurred that using fashion models for virtual try-ons can be misleading. These findings support our approach of prioritizing a customer-oriented virtual try-on by collecting per-garment datasets.

System utility. In Q1 of Figure 14, most participants agreed that our system is helpful for making decisions when purchasing garments (Mean=4.39, SD=0.49). Q2 and Q3 reveal that many participants also found our system useful for home use with portable devices (Mean=4.11, SD=0.94) and for in-store use with large displays (Mean=4.17, SD=0.50).

System performance. In Q4 and Q5 of Figure 14, most participants agreed that the garment deforms appropriately (Mean=4.28, SD=0.45) and that the wrinkles appear realistic (Mean=4.11, SD=0.74) across various poses. In Q6, while many participants agreed that the image quality is sufficient for virtual try-on (Mean=3.56, SD=0.90), four participants expressed disagreement. We believe this results from our decision to prioritize real-time performance over image quality, as most participants were satisfied with our system's frame rate (Mean=3.94, SD=0.52). Finally, in Q8, most participants agreed that the human-garment alignment is perfect (Mean=3.94, SD=0.97).

Specifically, participant P3 commented, "*The wrinkles looked realistic. Additionally, the video played smoothly without any lag, providing a comfortable try-on experience.*" Participant P6 added, "*This system allows for virtual try-on from various angles, making it practical as you can check the appearance from the sides and back.*"

6 Limitations and Discussion

Inability to handle extremely loose garments. Our method relies on the DensePose map for human-garment alignment; however, DensePose map estimation often fails when individuals wear extremely loose garments that occlude the

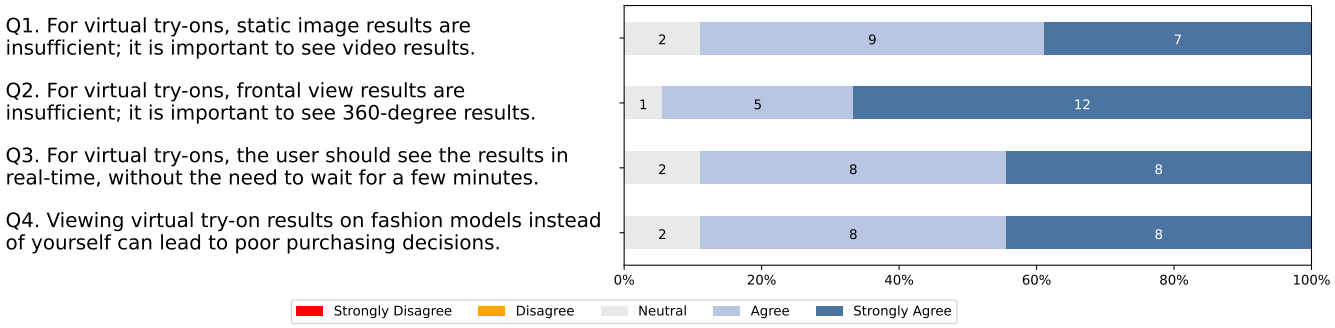


Figure 13. The results of the questionnaire asked about the participants' expectations of virtual try-on.

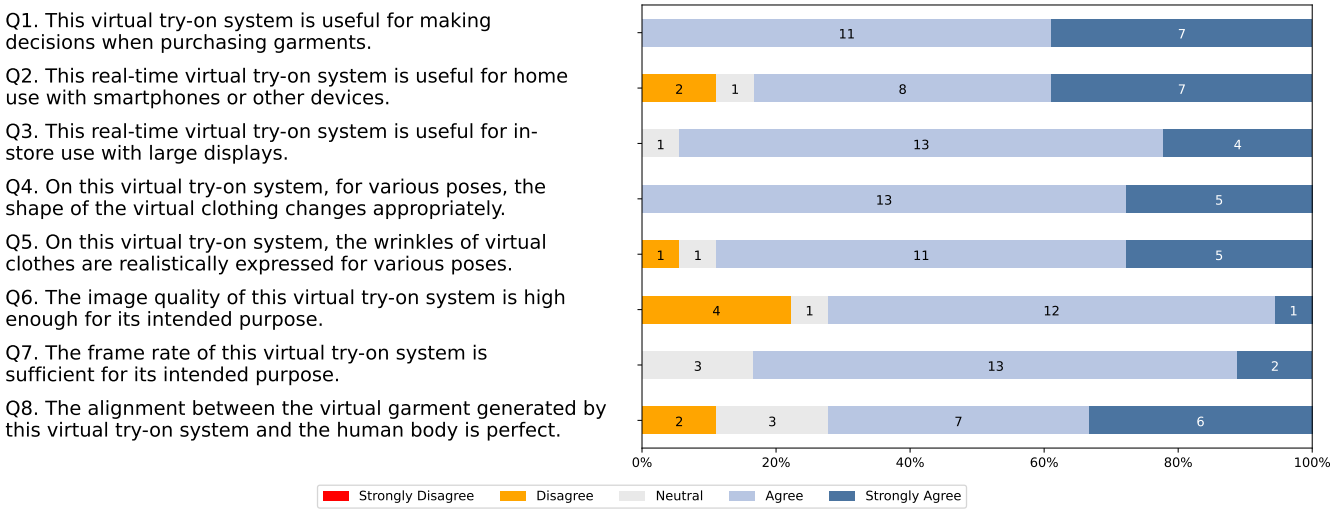


Figure 14. Results of the questionnaire asking the participants' rating on eight statements about the usefulness and user experience of experiencing our virtual try-on system.

Figure 15. Limitations of our method: (a) Our method is unable to process inputs with extremely loose garments or generate datasets for such garments due to the failure of DensePose map estimation for these inputs; (b) Inaccurate composition mask for hands; (c) Original garment is not removed.

body, as illustrated in Figure 15(a)[1]. Consequently, we cannot collect per-garment datasets for such garments, and users must avoid wearing them when using our system. In the

[1]The fashion model image is sourced from https://v.douyin.com/0hnko66OvwQ/

future, we plan to enhance our intermediate representation to robustly estimate human body shape.

Inaccurate composition mask. We utilize the predicted garment mask for composition, which generally performs well. However, as illustrated in Figure 15(b), our method struggles to generate an accurate composition mask for the hands when they are in front of the synthesized garment. This limitation arises from the DensePose map's inability to capture the silhouettes of fingers. Exploring additional representations to capture such fine-grained details is a promising direction for future work.

Presence of original garment in the input images. Our method overlays the synthesized garment directly onto the input frame without removing the original garment, as shown in Figure 15(c). In the future, we plan to develop a skin inpainting module to inpaint the gaps created by the removal of the original garment.

Necessity of data capture for each garment. Our method involves capturing a dataset for each garment using real humans. This is certainly an overhead. However, since retailers already hire fashion models for product photos, the additional effort required to collect per-garment datasets is minimal compared to the substantial benefits of personalized, interactive, and 360° virtual try-on experiences. Our user study supports this claim, showing that users find our system helpful in making purchase decisions.

7 Conclusion

We present a low-barrier dataset collection method for an interactive per-garment virtual try-on system that delivers high-fidelity, temporally smooth results. Our approach leverages real human bodies to collect per-garment datasets, eliminating the need for an expensive customized robotic mannequin and significantly improving the practical applicability of per-garment virtual try-on techniques. Additionally, we propose a hybrid person representation for garment synthesis that ensures the synthesized garment images are consistently oriented and precisely aligned with the human body without requiring additional wearable devices. This innovation also facilitates human-garment interaction without the need for extra wearable devices, further improving the practical usability of our method. We demonstrate the superiority of our approach through qualitative and quantitative evaluations against other state-of-the-art virtual try-on methods. We also validate the necessity of all components of the hybrid person representation through ablation studies. Finally, a user study revealed that most participants found our virtual try-on system helpful for making purchasing decisions.

References

[1] Michelle Carney, Barron Webster, Irene Alvarado, Kyle Phillips, Noura Howell, Jordan Griffith, Jonas Jongejan, Amit Pitaru, and Alexander Chen. 2020. Teachable machine: Approachable Web-based tool for exploring machine learning classification. In *Extended abstracts of the 2020 CHI conference on human factors in computing systems*. 1–8.

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

[3] Andrés Casado-Elvira, Marc Comino Trinidad, and Dan Casas. 2022. Pergamo: Personalized 3d garments from monocular video. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 293–304.

[4] Chieh-Yun Chen, Ling Lo, Pin-Jui Huang, Hong-Han Shuai, and Wen-Huang Cheng. 2021. Fashionmirror: Co-attention feature-remapping virtual try-on with sequential template poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13809–13818.

[5] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14131–14140.

[6] Toby Chong, I-Chao Shen, Nobuyuki Umetani, and Takeo Igarashi. 2021. Per garment capture and synthesis for real-time virtual try-on. In *The 34th Annual ACM Symposium on User Interface Software and Technology*. 457–469.

[7] Gabriel Cirio, Jorge Lopez-Moreno, David Miraut, and Miguel A Otaduy. 2014. Yarn-level simulation of woven cloth. *ACM Transactions on Graphics (TOG)* 33, 6 (2014), 1–11.

[8] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bowen Wu, Bing-Cheng Chen, and Jian Yin. 2019. Fw-gan: Flow-navigated warping gan for video virtual try-on. In *Proceedings of the IEEE/CVF international conference on computer vision*. 1161–1170.

[9] Zixun Fang, Wei Zhai, Aimin Su, Hongliang Song, Kai Zhu, Mao Wang, Yu Chen, Zhiheng Liu, Yang Cao, and Zheng-Jun Zha. 2024. ViViD: Video Virtual Try-on using Diffusion Models. *arXiv preprint arXiv:2405.11794* (2024).

[10] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. 2019. Graphonomy: Universal human parsing via graph transfer learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7450–7459.

[11] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in neural information processing systems* 27 (2014).

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

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

[14] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. 2018. Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 7297–7306.

[15] Oshri Halimi, Egor Larionov, Zohar Barzelay, Philipp Herholz, and Tuur Stuyck. 2023. Physgraph: Physics-based integration using graph neural networks. *arXiv preprint arXiv:2301.11841* (2023).

[16] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*. 6546–6555.

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

[18] Jianbin Jiang, Tan Wang, He Yan, and Junhui Liu. 2022. Clothformer: Taming video virtual try-on in all module. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10799–10808.

[19] Jonathan M Kaldor, Doug L James, and Steve Marschner. 2008. Simulating knitted cloth at the yarn level. In *ACM SIGGRAPH 2008 papers*. 1–9.

[20] Johanna Karras, Yingwei Li, Nan Liu, Luyang Zhu, Innfarn Yoo, Andreas Lugmayr, Chris Lee, and Ira Kemelmacher-Shlizerman. 2024. Fashion-VDM: Video Diffusion Model for Virtual Try-On. *arXiv preprint arXiv:2411.00225* (2024).

[21] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.

[22] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 8110–8119.

[23] Jeongho Kim, Guojung Gu, Minho Park, Sunghyun Park, and Jaegul Choo. 2024. Stableviton: Learning semantic correspondence with latent diffusion model for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8176–8185.

[24] Zorah Lahner, Daniel Cremers, and Tony Tung. 2018. Deepwrinkles: Accurate and realistic clothing modeling. In *Proceedings of the European conference on computer vision (ECCV)*. 667–684.

[25] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4681–4690.

[26] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. 2022. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *European Conference on Computer Vision*. Springer, 204–219.

[27] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. 2015. SMPL: A Skinned Multi-Person Linear Model. *Acm Transactions on Graphics* 34, Article 248 (2015).

[28] Rahul Narain, Armin Samii, and James F O'brien. 2012. Adaptive anisotropic remeshing for cloth simulation. *ACM transactions on graphics (TOG)* 31, 6 (2012), 1–10.

[29] Xiaoyu Pan, Jiaming Mai, Xinwei Jiang, Dongxue Tang, Jingxiang Li, Tianjia Shao, Kun Zhou, Xiaogang Jin, and Dinesh Manocha. 2022. Predicting loose-fitting garment deformations using bone-driven motion networks. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–10.

[30] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. 2020. 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*. 7365–7375.

[31] Gonzalo Ramos, Christopher Meek, Patrice Simard, Jina Suh, and Soroush Ghorashi. 2020. Interactive machine teaching: a human-centered approach to building machine-learned models. *Human–Computer Interaction* 35, 5-6 (2020), 413–451.

[32] Igor Santesteban, Miguel A Otaduy, and Dan Casas. 2019. Learning-based animation of clothing for virtual try-on. In *Computer Graphics Forum*, Vol. 38. Wiley Online Library, 355–366.

[33] Igor Santesteban, Miguel A Otaduy, and Dan Casas. 2022. Snug: Self-supervised neural dynamic garments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8140–8150.

[34] Igor Santesteban, Nils Thuerey, Miguel A Otaduy, and Dan Casas. 2021. 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*. 11763–11773.

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

[36] Andrew Selle, Jonathan Su, Geoffrey Irving, and Ronald Fedkiw. 2008. Robust high-resolution cloth using parallelism, history-based collisions, and accurate friction. *IEEE transactions on visualization and computer graphics* 15, 2 (2008), 339–350.

[37] Patrice Y Simard, Saleema Amershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, et al. 2017. Machine teaching: A new paradigm for building machine learning systems. *arXiv preprint arXiv:1707.06742* (2017).

[38] Dan Song, Xuanpu Zhang, Juan Zhou, Weizhi Nie, Ruofeng Tong, and An-An Liu. 2023. Image-Based Virtual Try-On: A Survey. *arXiv preprint arXiv:2311.04811* (2023).

[39] Yu Sun, Wu Liu, Qian Bao, Yili Fu, Tao Mei, and Michael J Black. 2022. Putting people in their place: Monocular regression of 3d people in depth. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13243–13252.

[40] Haoyu Wang, Zhilu Zhang, Donglin Di, Shiliang Zhang, and Wangmeng Zuo. 2024. MV-VTON: Multi-View Virtual Try-On with Diffusion Models. *arXiv preprint arXiv:2404.17364* (2024).

[41] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. 2018. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8798–8807.

[42] Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems* 135 (2022), 364–381.

[43] Zaiqiang Wu, Jingyuan Liu, Long Hin Toby Chong, I-Chao Shen, and Takeo Igarashi. 2024. Virtual Measurement Garment for Per-Garment Virtual Try-On. In *Graphics Interface 2024*.

[44] Donglai Xiang, Fabian Prada, Timur Bagautdinov, Weipeng Xu, Yuan Dong, He Wen, Jessica Hodgins, and Chenglei Wu. 2021. Modeling clothing as a separate layer for an animatable human avatar. *ACM Transactions on Graphics (TOG)* 40, 6 (2021), 1–15.

[45] Yuhao Xu, Tao Gu, Weifeng Chen, and Chengcai Chen. 2024. Ootdiffusion: Outfitting fusion based latent diffusion for controllable virtual try-on. *arXiv preprint arXiv:2403.01779* (2024).

[46] Zhengze Xu, Mengting Chen, Zhao Wang, Linyu Xing, Zhonghua Zhai, Nong Sang, Jinsong Lan, Shuai Xiao, and Changxin Gao. 2024. Tunnel Try-on: Excavating Spatial-temporal Tunnels for High-quality Virtual Try-on in Videos. *arXiv preprint arXiv:2404.17571* (2024).

[47] Zhongyi Zhou and Koji Yatani. 2022. Gesture-aware interactive machine teaching with in-situ object annotations. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–14.

[48] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*. 2223–2232.