PG-3DVTON: Pose-Guided 3D Virtual Try-on Network

Sanaz Sabzevari[1][a], Ali Ghadirzadeh[2][b], Mårten Björkman[1][c] and Danica Kragic[1][d]
[1]*Division of Robotics, Perception and Learning, KTH Royal Institute of Technology, Stockholm, Sweden*
[2]*Department of Computer Science, Stanford University, California, U.S.A.*

Keywords: 3D Virtual Try-on, Multi-Pose, Spatial Alignment, Fine-Grained Details.

Abstract: Virtual try-on (VTON) eliminates the need for in-store trying of garments by enabling shoppers to wear clothes digitally. For successful VTON, shoppers must encounter a try-on experience on par with in-store trying. We can improve the VTON experience by providing a complete picture of the garment using a 3D visual presentation in a variety of body postures. Prior VTON solutions show promising results in generating such 3D presentations but have never been evaluated in multi-pose settings. Multi-pose 3D VTON is particularly challenging as it often involves tedious 3D data collection to cover a wide variety of body postures. In this paper, we aim to develop a multi-pose 3D VTON that can be trained without the need to construct such a dataset. Our framework aligns in-shop clothes to the desired garment on the target pose by optimizing a consistency loss. We address the problem of generating fine details of clothes in different postures by incorporating multi-scale feature maps. Besides, we propose a coarse-to-fine architecture to remove artifacts inherent in 3D visual presentation. Our empirical results show that the proposed method is capable of generating 3D presentations in different body postures while outperforming existing methods in fitting fine details of the garment.

1 INTRODUCTION

3D virtual try-on (3DVTON) platforms apply realistic image synthesis for online marketing, enabling the process of fitting target clothes on human bodies in the 3D world. The use of 3DVTON holds the promise of eliminating a fair amount of online shopping returns that are due to a mismatch in style, size and body shape. Despite significant advances in prior works (Han et al., 2018; Wang et al., 2018; Dong et al., 2019; Issenhuth et al., 2019; Zheng et al., 2019; Yang et al., 2020; Chou et al., 2021; Ge et al., 2021; Xie et al., 2021) on virtual cloth try-on, the 3D aspect of the solution has not yet been well explored.

Multi-pose virtual fitting requires generating intuitive and realistic views for users in line with real try-on experience. Most of the existing works (Pons-Moll et al., 2017; Bhatnagar et al., 2019; Mir et al., 2020; Patel et al., 2020) focus on dressing a 3D person directly from 2D images built on the parametric Skinned Multi-Person Linear (SMPL) (Loper et al., 2015) model. Furthermore, typical manipulations are carried out by image-based virtual try-on systems to fit target in-shop clothes onto a reference person (Han et al., 2018; Yu et al., 2019; Han et al., 2019; Issenhuth et al., 2020). Most of these works adopt geometric warping by utilizing Thin Plate Spline (TPS) (Bookstein, 1989) transformations to deal with cloth-person misalignment. However, they cannot flexibly be applied to arbitrary poses and neglect the underlying 3D human body information. Besides, some fine-grained 2D details are not preserved well in synthesized images without using a reference model like SMPL, even in 3D approaches, e.g., Monocular-to-3D Virtual Try-On (M3D-VTON) (Zhao et al., 2021). One stream of work proposed to reconstruct a rigged 3D human model to address the artefact problem occurring at the boundaries of clothing (Kubo et al., 2019; Tuan et al., 2021). Nevertheless, it requires huge computational costs and efforts, which limits its practical application.

In this work, we address the above by multi-pose image manipulation that is neither restricted to coarse output results nor needs excessive manual effort to reconstruct a 3D human model. The proposed Pose-Guided 3D Virtual Try-On Network (PG-3DVTON) manipulates the target in-shop outfit beforehand and spatially aligns it to a 3D target human pose. The framework integrates 2D image-based virtual try-on

[a] https://orcid.org/0000-0003-0355-8977
[b] https://orcid.org/0000-0001-6738-9872
[c] https://orcid.org/0000-0003-0579-3372
[d] https://orcid.org/0000-0003-2965-2953

Sabzevari, S., Ghadirzadeh, A., Björkman, M. and Kragic, D.
PG-3DVTON: Pose-Guided 3D Virtual Try-on Network.
DOI: 10.5220/0011658100003417
In *Proceedings of the 18th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2023) - Volume 4: VISAPP*, pages 819-829
ISBN: 978-989-758-634-7; ISSN: 2184-4321

Figure 1: **Pose-Guided 3D virtual try-on network.** We present a 3D single-image human body guided by arbitrary poses and garments. Our method first generates the multi-pose cloth virtual try-on. We then synthesize the double depth map based on the target posture to construct 3D mesh photo-realistic results. The first three columns represent inputs, columns 4 to 8 are generated warped clothes, target semantic maps, and try-on meshes with double depth maps of our proposed approach, respectively.

and 3D depth estimation to generate 3D try-on of a dressed person with the identity of the person preserved, as shown in Figure 1. The main contributions of the paper are:

- We extend M3D-VTON to a multi-pose scenario in a multi-stage network conditioning on arbitrary poses and target garments through coarse-to-fine generation.
- We utilize dual geometric matching modules to reduce the artefacts generated at boundaries of outfits, especially around neckline, which is crucial for achieving more realistic results.
- We incorporate tree dilated fusion blocks to capture more spatial information with dilated convolution. We also aggregate multi-scale features to generate an initial double depth map for 3D virtual try-on.
- We present a training strategy for end-to-end training of our proposed approach which preserves high-quality details, specifically for the texture of garments.

2 RELATED WORK

2.1 Virtual Try-on Network

VTONs commonly consist of several multi-module pipelines and data preprocessing steps. Below, we overview approaches closely related to our work.

Fixed-Pose VTON. Image-based VTON systems involve a two-stage process. First, the in-shop clothing is warped to align to the target area in the human body. The second stage consists of texture fusion of the warped garments and target reference image while synthesizing the disclosed parts. There are extensive works that rely on this process. Specifically, the pioneering one is VITON (Han et al., 2018) which uses Shape Context Matching (SCM) (Belongie et al., 2002) as a matching method for warping the in-shop clothing. Some other related works like CP-VTON (Wang et al., 2018), CP-VTON+ (Minar et al., 2020), and ACGPN (Yang et al., 2020) apply TPS for geometry matching using convolution neural network (Rocco et al., 2017). For the fusion

stage, encoder-decoder networks like U-Net (Ronneberger et al., 2015) are used to synthesize try-on images to preserve the try-on cloth texture.

A feature warping module is present in (Han et al., 2019), known as ClothFlow, enhance the prediction of an appearance flow for aligning the source and target clothing areas in a cascaded manner. To improve textural integrity of try-on clothing and handle large deformations, ZFlow (Chopra et al., 2021) uses gated appearance flow. Further, ZFlow integrates UV projection maps with dense body-part segmentation (Güler et al., 2018) to mitigate undesirable artefacts, particularly around necklines. (Raffiee and Sollami, 2021), (Minar et al., 2021), and (Xie et al., 2021) use garment transfer as a more concrete synthesizing method for the try-on images. GarmentGAN deals with complex body pose and occlusion by employing a Generative Adversarial Network (GAN), alleviating the loss of target clothing details (Raffiee and Sollami, 2021). To further improve the warping network for different clothing categories, (Xie et al., 2021) propose a dynamic warping exploration strategy called Warping Architecture Search for Virtual Try-ON (WAS-VTON), searching a fusion network for various kinds of clothes. Cloth-VTON+ reconstructs a 3D garment model through the SMPL model to generate realistic try-on images using conditional generative networks (Minar et al., 2021). However, estimating the 3D parameters for the input person and leverage the standard SMPL body model for 3D cloth reconstruction is rather time consuming.

Another aspect of integrating pose information is stated in the following. TryOnGAN (Lewis et al., 2021) incorporates a pose-conditioned StyleGAN2 interpolation to create a try-on experience. This work typically results in high-resolution to visualize fashion on any person, but it also fails to extreme poses and underrepresented garments. To further improve photo-realistic try-on images without human segmentation, Parser Free Appearance Flow Network (PF-AFN) employs a teacher-tutor-student approach. It is initially designed to train a parser-based teacher model as a tutor network. Then it treats tutor knowledge as inputs of the parser-free student model in a distillation scheme. A similar counterpart parser-free method is the StyleGAN-based warping module to overcome significant misalignment between a person and a garment image (He et al., 2022). Despite the recent advances, the results are constrained to poses similar to the input image.

Multi-Pose VTON. The work towards a multi-pose guided virtual try-on network is initially presented in (Dong et al., 2019) by proposing Multi-pose Guided Virtual Try-On Network (MG-VTON). This work aims to transfer a garment onto a person with diverse poses and consists of three stages: a conditional human parsing network, a deep Warping Generative Adversarial Network (Warp-GAN), and a refinement render network. Attentive Bidirectional Generation Adversarial Network (AB-GAN) is another similar approach to refine the quality of the try-on image through a bi-stage strategy, including a shape-enhanced clothing deformation model and an attentive bidirectional GAN (Zheng et al., 2019). FashionOn (Hsieh et al., 2019) introduces FacialGAN and clothing U-Net to extract salient regions like faces and clothes for refining the try-on images. However, some fine-grained details, particularly around necklines, were still missing, likewise in the earlier works (Dong et al., 2019; Zheng et al., 2019; Wang et al., 2020a). Reposing of humans based on a single source image is proposed through a pose-conditioned StyleGAN network (Albahar et al., 2021). While this approach provides high-quality human pose transfer, it remains challenging to transfer both poses and garments concurrently. Another recent study relies on swapping both pose and garments. 2D multiple-pose virtual try-on based 3D clothing reconstruction called 3D-MPVTON (Tuan et al., 2021) renders natural clothing deformations while imposing limitations due to the rigged reconstructed 3D garment model. Dressing in Order (DiOr) introduces a flexible person generation for several fashion editing tasks, including layering multiple garments of the same kind (Cui et al., 2021). It is, however, a limitation that could not overcome both reposing and garment transfer simultaneously. Semantic Prediction Guidance for Multi-pose Virtual Try-on Network (SPG-VTON) (Hu et al., 2022) includes three sub-modules by conducting a global and a local discriminator to control the generated results using DeepFashion (Liu et al., 2016) and Multi-Pose Virtual try on (MPV) (Dong et al., 2019). Despite achieving a photo-realistic try-on, the method cannot be used on a 3D virtual try-on and disregards the underlying 3D body information.

2.2 3D Virtual Try-on

3D virtual try-on without a scanned 3D dataset is an intriguing and challenging problem due to the complex deformation of a garment. Prior work has demonstrated successful 3D human reconstructions as well as generating fine-detail clothes, but still, these methods cannot transfer clothes from one domain to another (Saito et al., 2019; Saito et al., 2020; Li et al., 2020). The most popular of these methods is PIFuHD (Saito et al., 2020). It renders a high-quality

3D human mesh based on 2D images through a multi-level Pixel-Aligned Implicit Function and circumvents premature decisions regarding explicit geometry. One way to predict the underlying human shape and clothing is MGN (Bhatnagar et al., 2019) that is associated with the body represented by an SMPL model while limiting to the predefined garments from a digital wardrobe. Pix2Surf (Mir et al., 2020) also proposed to translate texture from 2D garment images to a 3D virtually dressed SMPL. It uses silhouette shape instead of clothing texture to make it robust to highly varying garment textures. However, the body texture is ignored in this method, and it requires considerable costs to collect scanned 3D datasets for training. Our work involves translating both a cloth image and poses of a human body into a target one for a 3D try-on task.

3 BACKGROUND

3.1 Problem Formulation

PG-3DVTON focuses on generating a 3D clothed human body wearing a target garment under arbitrary poses, see Figure 2. It takes as input a target pose P^t, an image of the in-shop cloth C, and an input image of a person I, and outputs a synthesized 3D try-on mesh I^O which represents the same person wearing the in-shop cloth at the target posture.

3.2 Geometric Matching Module

Each module uses a Geometric Matching Module (GMM) to preserve the details of the person's image and the texture of the clothes due to huge pixel-to-pixel misalignment. In CGM, GMM_1 warps the in-shop garment under the target pose, and in FGM, GMM_2 converts the warped garment back to the target garment. We denote ε_i^A and ε_i^B as feature extractors for each $GMM_i, i \in \{1,2\}$. Features extracted are correlated in a single tensor as the input of a regressor network. The output of the correlation map contains all pairwise similarities between the corresponding features. The regression network consists of two 2-strided convolutional layers, two 1-strided ones and one fully-connected output layer to predict the spatial transformation parameter θ_i. The architecture of a GMM is shown in Figure 3.

The thin-plate spline (TPS) is an algebraic tool to interpolate surfaces over a set of known corresponding control points in the plane (Bookstein, 1989). The TPS transformation in the GMM performs this interpolation based on control points of two images. These control points are defined as a fixed uniform grid over the second image and their corresponding points in the first image. Thus, the control point position of the first image plays an important role in TPS transformation parametrization because the control points in the second image are fixed.

4 PG-3DVTON

PG-3DVTON is based on two modules: a Coarse Generation Module (CGM) and a Fine Generation Module (FGM). CGM estimates the region of the desired garment and a base 3D shape of the input person. The FGM is then applied to refine the final 3D try-on mesh results. The purpose of this module is to preserve rich details of the garment on the reference person and the details of the face. These modules are described below.

4.1 Coarse Generation Module

The CGM module is responsible to generate a coarse representation of the final output. It consists of four sub-modules including semantic parsing prediction, spatial cloth warping, double-depth map estimation, and coarse appearance generation which are described below.

Semantic Parsing. This module predicts the semantics of the generated image at the new pose to better fit the garment. It receives as input the semantic parsing at the initial pose S^I and a garment mask M^C and outputs a semantic parsing $\hat{S}^t$ for the target pose P^t; $(S^I, P^t, M^C) \rightarrow \hat{S}^t$. The network is implemented as a U-Net (Ronneberger et al., 2015), and is trained by optimizing $\mathcal{L}_S = \mathcal{L}_{ad} + \lambda_{ce}\mathcal{L}_{ce}$, where $\mathcal{L}_{ce}$ denotes the cross-entropy loss, $\mathcal{L}_{ad}$ is an adversarial loss, and λ is a hyper-parameter to balance the two losses. The cross-entropy loss is defined as:

$$\mathcal{L}_{ce} = -\left\| S^{gt} \odot log(\hat{S}^t) \odot (1 + M^C) \right\|_1, \quad (1)$$

where, S^{gt} is the ground truth data, $\odot$ is the element-wise multiplication, and $\|.\|_1$ denotes the L1 norm. The adversarial loss is defined as:

$$\mathcal{L}_{ad} = \mathbb{E}_X[log(D(X))] + \mathbb{E}_Z[log(1 - D(G(Z)))], \quad (2)$$

where $G(Z)$ is a generator that generates a target semantic parsing $\hat{S}^t$ from a random sample in a latent space Z, and $D(X)$ is a discriminator trained to tell $\hat{S}^t$ apart from the ground truth in $X = \{S^{gt}\}$. Both $G(Z)$ and $D(X)$ are conditioned on the inputs $[S^I, P^t, M^C]$.

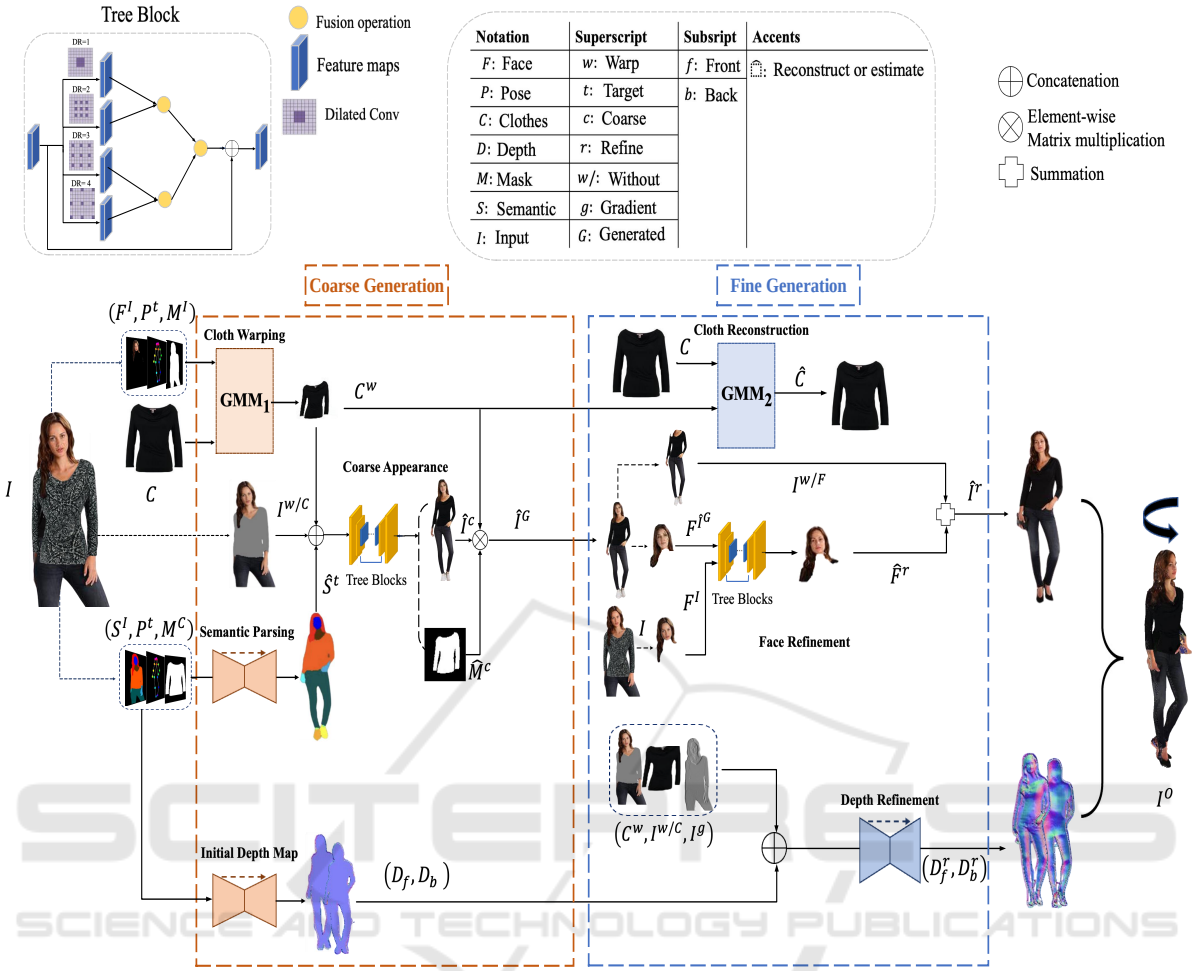

Figure 2: Overview of the proposed PG-3DVTON. The pipeline consists of two parts: a) Coarse Alignment: The Coarse Generation Module (CGM) is introduced for prediction of the human segmentation and depth map, producing the initial rigged try-on image via aligning the warped clothing and the composition mask. b) Refinement: To generate fine-grained details of a 3D reference image wearing the target cloth, the Fine Generation Module (FGM) is adopted. Once RGB-D representation is achieved, the 3D clothed human with the target pose and garment converted to get colored point clouds and finally remeshing the predicted point cloud.

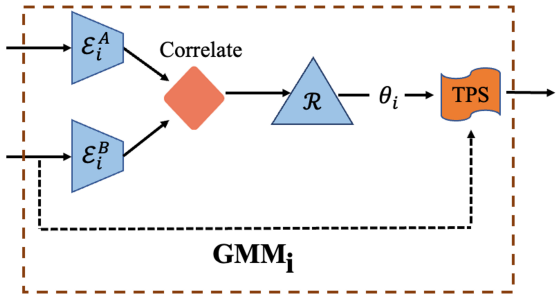

Figure 3: Diagram of the Geometric Matching Module.

Cloth Warping. As shown in Figure 2, this module warps the in-shop cloth to fit the target pose using a GMM. It receives the face region of the input image F^I, target pose P^t, the binary mask of the input image M^I, and clothes C and outputs the warped image of the clothes C^w; $(F^I, P^t, M^I, C) \rightarrow C^w$.

Initial Depth Map. We model a 3D representation of the resulting try-on using a double-depth map to model the front D_f and the back D_b depth images; $(S^I, P^t, M^C) \rightarrow (D_f, D_b)$. This module is implemented as a U-Net and takes the same inputs as the semantic parsing prediction module. The model is pre-trained using the following loss function:

$$\mathcal{L}_d = \left\| D_f - D_f^{gt} \right\|_1 + \left\| D_b - D_b^{gt} \right\|_1, \qquad (3)$$

where, D_f^{gt} and D_b^{gt} denotes the ground-truth front and back depth maps, respectively. The ground truth

maps are created by first generating 3D meshes using Pixel-Aligned Implicit Function for High-Resolution 3D Human Digitization (PIFuHD) (Saito et al., 2020), which are then projected orthographically by Pyrender (Matl et al., 2019) to create the pseudo ground-truth depth maps.

Coarse Appearance. To generate the appearance, it is required to colour the translated semantic parsing map according to the input image and warped in-shop cloth image. Inspired by (Wang et al., 2020a), a tree block generator (Fu et al., 2018) utilizes dilated convolutions to retain some specific parts of the try-on image by aggregating multi-scale features and retrieving more spatial information. The goal of this module is to predict the initial coarse try-on result $\hat{I}^c$ and the binary mask of the target garment $\hat{M}^c$ conditioned on the warped cloth image C^w, the person image without the garment $I^{w/C}$, and the target semantic parsing $\hat{S}^t$; $\left(C^w, I^{w/C}, \hat{S}^t\right) \rightarrow \left(\hat{I}^c, \hat{M}^c\right)$. Then, the output is fed into the FGM for further refinement. The generated coarse result $\hat{I}^G$ is captured by:

$$\hat{I}^G = \underbrace{C^w \odot \hat{M}^c}_{\hat{C}^w} + \underbrace{\hat{I}^c \odot \left(1 - \hat{M}^c\right)}_{\hat{I}^{w/C}}, \tag{4}$$

where $\hat{C}^w$ is the cloth region in the coarse result, and $\hat{I}^{w/C}$ is the coarse generated result without clothes. The corresponding loss function $\mathcal{L}_C$ used in this paper is:

$$\begin{aligned}\mathcal{L}_C = {} & \lambda_{atten}\mathcal{L}_{atten}\left(\hat{M}^c, M^w\right) + \lambda_{smooth}\mathcal{L}_{smooth}\left(\hat{I}^G, I^{gt}\right) \\ & + \lambda_{percept}\mathcal{L}_{percept}\left(\hat{I}^G, I^{gt}\right) + \mathcal{L}_{adv}\left(\hat{I}^G, I^{gt}\right).\end{aligned} \tag{5}$$

where the attention loss $\mathcal{L}_{atten}$ is

$$\mathcal{L}_{atten} = \left\|\hat{M}^c - M^w\right\|_1 + \lambda_{TV}\left\|\nabla\hat{M}^c\right\|_2, \tag{6}$$

and λ_{TV} is the total variation regularization parameter to preserve edges (Liang et al., 2011), $\|.\|_2$ is the L2 or Euclidean norm, M^w is the ground-truth warped cloth mask as well as $\nabla\hat{M}^c$ is the gradient of the composition mask. Moreover, the smooth loss $\mathcal{L}_{smooth}$ (Girshick, 2015) is employed to be more robust to the outliers compared to L2 loss which is sum of all the squared differences in between the ground-truth data and the generated output:

$$\mathcal{L}_{smooth}(\hat{I}^G, I^{gt}) = \begin{cases} 0.5(\hat{I}^G - I^{gt})^2 & \text{if } |\hat{I}^G - I^{gt}| < 1 \\ |\hat{I}^G - I^{gt}| - 0.5 & otherwise \end{cases} \tag{7}$$

where I^{gt} is the ground-truth input image under target posture.

The perceptual loss $\mathcal{L}_{percept}$ (Johnson et al., 2016) is used to preserve the high-level content and the style of the garment. It includes two perceptual loss functions based on the network's loss Φ (pretrained network for image classification): (i) feature reconstruction loss is Euclidean distance between feature representations and (ii) style reconstruction loss is squared Frobenius norm of the difference between the Gram matrices of the generated and ground-truth images. The first one encourages the pixels of the output images to have similar feature representations, while the latter penalizes it in a case that deviates in content from the ground-truth data. More details of the this loss function can be found in (Johnson et al., 2016).

The objective function to guide distinguishing between real and fake labels is introduced by $\mathcal{L}_{adv}$ through Least Squares Generative Adversarial Networks (LSGANs) (Mao et al., 2017). Also, λ_{atten}, λ_{smooth}, $\lambda_{percept}$, and λ_{adv} are hyper parameters.

4.2 Fine Generation Module

After generating the coarse result, the following process adds more information to the try-on image by synthesizing photo-realistic body texture.

Face Refinement. Retaining the facial characteristics of the person is one of the challenges in VTON systems. We use a GAN-based tree block network to address this. The network uses the face regions of generated try-on image $F^{\hat{I}^g}$ and the original reference image F^I, resulting in a refined face region $\hat{F}^r$; $\left(F^I, F^{\hat{I}^g}\right) \rightarrow \hat{F}^r$, which is combined with the coarse try-on image without the face $I^{w/F}$ to produce a new refined try-on image $\hat{I}^r$. For training the following objective function is used:

$$\begin{aligned}\mathcal{L}_F^r = {} & \mathcal{L}_{VGG}\left(\hat{F}^r, F^{I_{gt}}\right) + \lambda_{\mathcal{L}_1}\left\|\hat{F}^r - F^{I_{gt}}\right\|_1 \\ & + \lambda_{adv}^F\mathcal{L}_{adv}\left(F^I, \hat{F}^r\right) + \lambda_{smooth}^F\mathcal{L}_{smooth}\left(\hat{I}^r, I^{gt}\right).\end{aligned} \tag{8}$$

where $\mathcal{L}_{VGG}$ is a perceptual loss defined as (Wang et al., 2018)

$$\mathcal{L}_{VGG}\left(\hat{F}^r, F^{I_{gt}}\right) = \sum_{i=1}^{5}\lambda_i\left\|\phi_i\left(\hat{F}^r\right) - \phi_i\left(F^{I_{gt}}\right)\right\|_1 \tag{9}$$

and $F^{I_{gt}}$ is the face region of the target person ground-truth image I_{gt}, ϕ_i is the feature map of i-th layer in the visual VGG19 network (Simonyan and Zisserman, 2014).

Cloth Reconstruction. To preserve some specific parts of the in-shop cloth image, such as the neckline, in the generated try-on images, a second GMM is

used to reconstruct the in-shop cloth image C from the warped cloth image C^w in CGM. The reconstructed cloth image is given by $\hat{C} = TPS_{\theta_2}(C^w)$, which captures rich details through the geometric cloth warping method in (Wang et al., 2018) and is trained with the loss $\mathcal{L}_w$; $(C, C^w) \rightarrow \hat{C}$.

$$\mathcal{L}_w = \mathcal{L}_{smooth}\left(C^w, C^I\right) + \left\|\hat{C} - C\right\|_1 \quad (10)$$

Here $\mathcal{L}_{smooth}$ is defined similarly to (7). The image C^I denotes the ground-truth region of the cloth image, C, extracted from the reference image under the target pose P^t.

Depth Refinement. To minimize the discrepancy between the ground-truth depth map and the reconstructed one, it is necessary to keep the high-frequency depth details. To achieve this, the image gradient I^g is acquired by concatenating the gradient images of C^w and $I^{w/C}$. We apply the Sobel operator to detect edges and capture the gradient images on the aforementioned images. Then, the triplets C^w, $I^{w/C}$, I^g are concatenated with initial depth maps to produce the refined double-depth map (D_f^r, D_b^r) through a U-Net; $\left(\left(C^w, I^{w/C}, I^g\right), (D_f, D_b)\right) \rightarrow (D_f^r, D_b^r)$. Inspired by (Hu et al., 2019), the weighted sum of two loss functions is considered during training as follows

$$\mathcal{L}_d^r = \lambda_{depth} \underbrace{\left(\frac{1}{n}\sum_{i=1}^{n} log\left(\left\|D_i^r - D_i^{gt}\right\|_1 + 1\right)\right)}_{\mathcal{L}_{depth}} + \lambda_{grad}\left(\frac{1}{n}\sum_{i=1}^{n} log\left(\nabla_x\left(\left\|D_i^r - D_i^{gt}\right\|_1\right) + 1\right)\right. \quad (11)$$
$$\left. \underbrace{+ log\left(\nabla_y\left(\|D_i^r - D_i^{gt}\|_1\right) + 1\right)}_{\mathcal{L}_{grad}}\right).$$

where D_i^r and D_i^{gt} are the i-th refined depth point, and the ground-truth one, respectively, n is the total number of front/back depth map points, and ∇ represents the Sobel operator.

4.3 Joint Training and Final 3D Human Mesh

While the training process is handled separately for each network, the performance deteriorates when fine-grained details are desired. Therefore, we jointly train all the sub-modules of the proposed approach except depth modules to remedy the influence of coarse try-on results. We formulate the overall objective function as:

$$\mathcal{L}_{total} = \mathcal{L}_S + \mathcal{L}_w + \mathcal{L}_C + \mathcal{L}_F^r. \quad (12)$$

Consequently, we extract the front and back view depth maps to convert the 3D point clouds. The front depth map is incorporated in the try-on result. However, there is a need to inpaint the try-on image for the back texture inspired by (Telea, 2004). Finally, we remesh the predicted point cloud viewer for 3D presentation.

5 EXPERIMENTS

5.1 Dataset

The MPV dataset (Dong et al., 2019) consists of pairs of female models and top garment images performed for experiments used for both train and test sets. It should be noted that we need to construct the pseudo depth dataset for a monocular-to-3D virtual try-on dataset, in which each person image has the corresponding front and back depth maps (D_f and D_b), respectively. For this we use PIFuHD (Saito et al., 2020) to obtain the relative generated human mesh. Then it is orthographically projected to the depth maps. We divide the whole dataset into a 12997 image train set and a 2577 image test set. The images in this dataset have a resolution of 256×192.

5.2 Implementation Detail

The sub-modules of the CGM are trained to provide the inputs for the FGM. The Adam optimizer is adopted to train the combined network for 200 epochs with the initial learning rate set to 0.0002. We additionally set different batch sizes for each modules; 64 for semantic parsing and GMM modules, and 8 for the remaining modules, while using 2 GPUs. We implement the model in Pytorch and trained on NVIDIA RTX 2080Ti GPUs.

5.3 Qualitative Results

We compare the results of the proposed network with the following baseline methods: MG-VTON (Dong et al., 2019), Down to the last detail (Wang et al., 2020a), and M3D-VTON (Zhao et al., 2021).

MG-VTON: is an improved version of the 2D Virtual Try-ON (VTON) system, including the change of input posture. We define a 3D virtual try-on presentation and adapt it for diverse poses.

Down to the Last Detail: is the baseline to tackle the

VTON system for multiple poses. Despite its effort to preserve the carving details, some mismatches explode, especially around the neckline. We augment the cycle consistency loss to this network for better visual tracking of the clothing region in the generated results.

M3D-VTON: is the state-of-the-art method to present 3D VTON with the simple try-on task of fitting the desired garment on the reference person. We enhance this network through matching the arbitrary poses and taking care of carving details.

We present the comparison of our method with the existing state-of-the-art in Figure 4. It should be noted that we perform the comparison for a single pose since the benchmark model includes the estimation of depth maps applied to a single posture. Although the resolution of M3D-VTON is roughly two times higher than the rest, it performs poorly for different clothing regions. The face has not been generated in M3D-VTON since it does not allow changing the posture; accordingly, we could just compare qualitatively with the pseudo ground-truth PIFu-HD in Figure 5. This figure illustrates that our PG-3DVTON generates realistic monocular 3D VTON while preserving mesh texture.

5.4 Ablation Study

We perform an ablation study, including removing the end-to-end joint training strategy on held-out test data. It is verified that this strategy could enhance the generated results due to optimizing the entire framework and could help to reduce artifacts for various cloth synthesizes. It is also illustrated in Figure 6 that

Figure 4: The visualized 2D-virtual try-on result comparison. Our method has better performance at generating cloth rich details illustrated in red boxes.

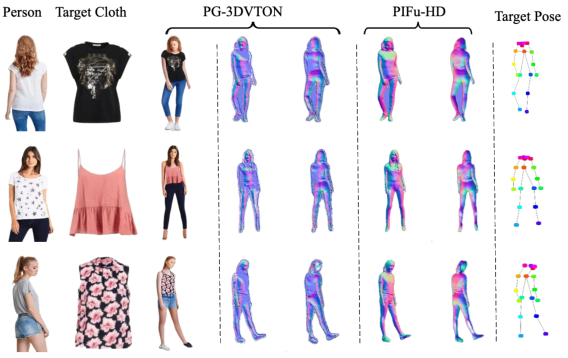

Figure 5: The visualized generated double-depth maps. The first two columns and the last one represent the inputs, while the others are generated 3D try-on results and PIFU-HD mesh, respectively.

training separately of the modules as semantic parsing leads to the unpleasing generation. Then it is required to fine-tune in the end-to-end training process.

5.5 Evaluation Metrics

There are two different metrics to evaluate the effectiveness of the proposed structure in a 2D presentation: Image-based and Feature-based metrics. We use the Structural Similarity Index Measure (SSIM) as a representative for image-based metrics and the Fréchet Inception Distance (FID) for feature-based ones. A higher score for SSIM and a lower value for FID indicate the higher accuracy of the generated images compared with the ground-truth images. We also use two common depth evaluation metrics: Root Mean Squared Error (RMSE) and Absolute Relative error (Abs.). Our approach outperforms the baselines in terms of geometric details of the depth estimation. It should be noted that the PIFU score is captured based on the average double-depth score, while NormalGAN is computed from either front or back depth.

In Table 1, we summarize evaluation results from 1500 generated try-on images cropped around the generated clothes, disregarding the face area. These show that our PG-3DVTON achieves the maximum SSIM scores on the MPV dataset. A greater score of SSIM and a lower score of FID demonstrate that the quality of the generated image is closer to the ground-truth image. Thus PG-3DVTON is better at fitting the in-shop clothes onto the input person under different postures. However, failure cases are also presented, primarily due to the stochasticity of the semantic segmentation, with examples shown in Figure 7, especially for the facial area or the area unrelated to clothes between the generated image and the original image for evaluation.

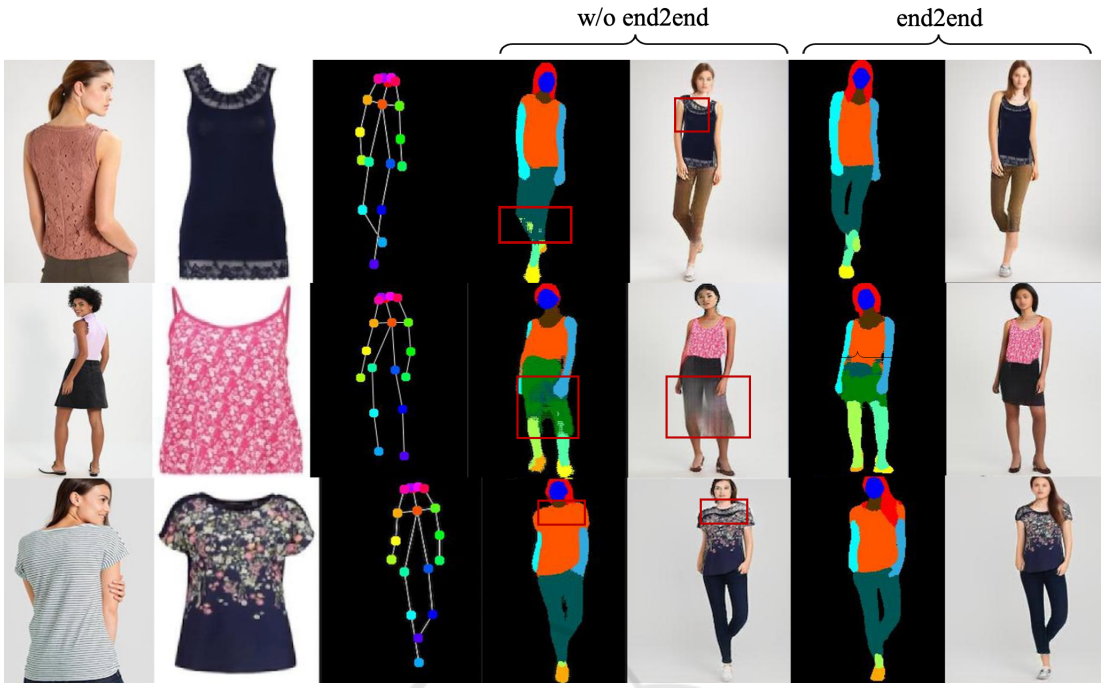

Figure 6: The effective results of our end-to-end training pipeline and ablation study.

Figure 7: Failure Cases for real-world applications due to stochasticity provided by the semenatic segmentation estimation network.

Since this is the first work that explores the multi-pose 3D-VTON, we compare the generated 3D try-on mesh with the 3D human reconstruction and baseline approaches, as shown in Table 2. To make a fair comparison, we only consider cases for which the input and target poses are the same, even if PG-3DVTON can handle also changes in pose. To compare with these benchmark models, we evaluate both the global metrics RMSE and Abs., which estimate the depth between the generated depth map and that of the ground-truth. The lower score of PG-3DVTON in Table 2 illustrates superior shape generation ability compared to the state-of-the-art methods.

Table 1: Quantitative comparison with the state-of-the-art methods on the MPV dataset.

Method	SSIM ↑	FID ↓
MG-VTON (Dong et al., 2019)	0.705	22.42
Down-to-the-Last-Detail (Wang et al., 2020a)	0.723	16.01
M3D-VTON (Zhao et al., 2021)	0.685	22.05
PG-3DVTON (Ours)	**0.797**	**14.64**

Table 2: Quantitative comparison for double-depth score (All values have been multiplied by 10^3 to improve readability in the table).

Method	RMSE ↓	Abs. ↓
PIFU (Saito et al., 2019)	27.07	8.12
NormalGAN (Wang et al., 2020b)	18.21	11.23
M3D-VTON (Zhao et al., 2021)	14.68	8.79
PG-3DVTON (Ours)	**14.16**	**6.87**

6 CONCLUSIONS

We have presented a 3D synthesis approach for a multi-pose virtual try-on. The core novelties lie in 1) producing the 3D try-on mesh through body depth estimation under arbitrary poses and 2) a geometric matching module augmentation in the end-to-end training process. Our experiment demonstrates that the proposed methodology could enhance transferring the in-shop garment to the person image in the target posture while synthesizing the corresponding depth maps. In addition, this framework outperforms the

benchmark models in estimating the front and back body depth maps. We have validated the VTON task by performing an ablation study and quantitative evaluation concerning the state-of-the-art. Our model provides an economical and alternative way to 3D scanning for the monocular 3D multi-pose virtual try-on. In future work, we will explore the application of the proposed method to the tailoring industry with sewing pattern datasets. Furthermore, the multi-stage network is dependent on the success of previous levels, such as the semantic parsing module in our pipeline. Subsequent work may include incorporating the distillation process to alleviate the human parsing for a multi-pose try-on.

REFERENCES

Albahar, B., Lu, J., Yang, J., Shu, Z., Shechtman, E., and Huang, J.-B. (2021). Pose with style: Detail-preserving pose-guided image synthesis with conditional stylegan. *ACM Transactions on Graphics (TOG)*, 40(6):1–11.

Belongie, S., Malik, J., and Puzicha, J. (2002). Shape matching and object recognition using shape contexts. *IEEE transactions on pattern analysis and machine intelligence*, 24(4):509–522.

Bhatnagar, B. L., Tiwari, G., Theobalt, C., and Pons-Moll, G. (2019). Multi-garment net: Learning to dress 3d people from images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5420–5430.

Bookstein, F. L. (1989). Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Transactions on pattern analysis and machine intelligence*, 11(6):567–585.

Chopra, A., Jain, R., Hemani, M., and Krishnamurthy, B. (2021). Zflow: Gated appearance flow-based virtual try-on with 3d priors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5433–5442.

Chou, C.-L., Chen, C.-Y., Hsieh, C.-W., Shuai, H.-H., Liu, J., and Cheng, W.-H. (2021). Template-free try-on image synthesis via semantic-guided optimization. *IEEE Transactions on Neural Networks and Learning Systems*.

Cui, A., McKee, D., and Lazebnik, S. (2021). Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14638–14647.

Dong, H., Liang, X., Shen, X., Wang, B., Lai, H., Zhu, J., Hu, Z., and Yin, J. (2019). Towards multi-pose guided virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9026–9035.

Fu, X., Qi, Q., Huang, Y., Ding, X., Wu, F., and Paisley, J. (2018). A deep tree-structured fusion model for single image deraining. *arXiv preprint arXiv:1811.08632*.

Ge, Y., Song, Y., Zhang, R., Ge, C., Liu, W., and Luo, P. (2021). Parser-free virtual try-on via distilling appearance flows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8485–8493.

Girshick, R. (2015). Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448.

Güler, R. A., Neverova, N., and Kokkinos, I. (2018). Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306.

Han, X., Hu, X., Huang, W., and Scott, M. R. (2019). Clothflow: A flow-based model for clothed person generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10471–10480.

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

He, S., Song, Y.-Z., and Xiang, T. (2022). Style-based global appearance flow for virtual try-on. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3470–3479.

Hsieh, C.-W., Chen, C.-Y., Chou, C.-L., Shuai, H.-H., Liu, J., and Cheng, W.-H. (2019). Fashionon: Semantic-guided image-based virtual try-on with detailed human and clothing information. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 275–283.

Hu, B., Liu, P., Zheng, Z., and Ren, M. (2022). Spg-vton: Semantic prediction guidance for multi-pose virtual try-on. *IEEE Transactions on Multimedia*.

Hu, J., Ozay, M., Zhang, Y., and Okatani, T. (2019). Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1043–1051. IEEE.

Issenhuth, T., Mary, J., and Calauzènes, C. (2019). End-to-end learning of geometric deformations of feature maps for virtual try-on. *arXiv preprint arXiv:1906.01347*.

Issenhuth, T., Mary, J., and Calauzènes, C. (2020). Do not mask what you do not need to mask: a parser-free virtual try-on. In *European Conference on Computer Vision*, pages 619–635. Springer.

Johnson, J., Alahi, A., and Fei-Fei, L. (2016). Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer.

Kubo, S., Iwasawa, Y., Suzuki, M., and Matsuo, Y. (2019). Uvton: Uv mapping to consider the 3d structure of a human in image-based virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0.

Lewis, K. M., Varadharajan, S., and Kemelmacher-Shlizerman, I. (2021). Tryongan: Body-aware try-on via layered interpolation. *ACM Transactions on Graphics (TOG)*, 40(4):1–10.

Li, Z., Yu, T., Pan, C., Zheng, Z., and Liu, Y. (2020). Robust 3d self-portraits in seconds. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 1344–1353.

Liang, D., Wang, H., Chang, Y., and Ying, L. (2011). Sensitivity encoding reconstruction with nonlocal total variation regularization. *Magnetic resonance in medicine*, 65(5):1384–1392.

Liu, Z., Luo, P., Qiu, S., Wang, X., and Tang, X. (2016). Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104.

Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., and Black, M. J. (2015). Smpl: A skinned multi-person linear model. *ACM transactions on graphics (TOG)*, 34(6):1–16.

Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., and Paul Smolley, S. (2017). Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2794–2802.

Matl, M. et al. (2019). Pyrender.

Minar, M. R., Tuan, T. T., and Ahn, H. (2021). Cloth-vton+: Clothing three-dimensional reconstruction for hybrid image-based virtual try-on. *IEEE Access*, 9:30960–30978.

Minar, M. R., Tuan, T. T., Ahn, H., Rosin, P., and Lai, Y.-K. (2020). Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *CVPR Workshops*.

Mir, A., Alldieck, T., and Pons-Moll, G. (2020). Learning to transfer texture from clothing images to 3d humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7023–7034.

Patel, C., Liao, Z., and Pons-Moll, G. (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*, pages 7365–7375.

Pons-Moll, G., Pujades, S., Hu, S., and Black, M. J. (2017). Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics (ToG)*, 36(4):1–15.

Raffiee, A. H. and Sollami, M. (2021). Garmentgan: Photo-realistic adversarial fashion transfer. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 3923–3930. IEEE.

Rocco, I., Arandjelovic, R., and Sivic, J. (2017). Convolutional neural network architecture for geometric matching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6148–6157.

Ronneberger, O., Fischer, P., and Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer.

Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., and Li, H. (2019). Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2304–2314.

Saito, S., Simon, T., Saragih, J., and Joo, H. (2020). Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 84–93.

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Telea, A. (2004). An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34.

Tuan, T. T., Minar, M. R., Ahn, H., and Wainwright, J. (2021). Multiple pose virtual try-on based on 3d cloth reconstruction. *IEEE Access*, 9:114367–114380.

Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., and Yang, M. (2018). Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604.

Wang, J., Sha, T., Zhang, W., Li, Z., and Mei, T. (2020a). Down to the last detail: Virtual try-on with fine-grained details. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 466–474.

Wang, L., Zhao, X., Yu, T., Wang, S., and Liu, Y. (2020b). Normalgan: Learning detailed 3d human from a single rgb-d image. In *European Conference on Computer Vision*, pages 430–446. Springer.

Xie, Z., Zhang, X., Zhao, F., Dong, H., Kampffmeyer, M. C., Yan, H., and Liang, X. (2021). Was-vton: Warping architecture search for virtual try-on network. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 3350–3359.

Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., and Luo, P. (2020). 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*, pages 7850–7859.

Yu, R., Wang, X., and Xie, X. (2019). 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*, pages 10511–10520.

Zhao, F., Xie, Z., Kampffmeyer, M., Dong, H., Han, S., Zheng, T., Zhang, T., and Liang, X. (2021). M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13239–13249.

Zheng, N., Song, X., Chen, Z., Hu, L., Cao, D., and Nie, L. (2019). Virtually trying on new clothing with arbitrary poses. In *Proceedings of the 27th ACM International Conference on Multimedia*, pages 266–274.