M3D-VTON: A Monocular-to-3D Virtual Try-On Network

Fuwei Zhao[1], Zhenyu Xie[1], Michael Kampffmeyer[2], Haoye Dong[1]
Songfang Han[3], Tianxiang Zheng[4], Tao Zhang[4], Xiaodan Liang[1*]
[1]Shenzhen Campus of SYSU, [2]UiT The Arctic University of Norway, [3]UC San Diego, [4]Momo
{zhaofw@mail2,xiezhy6@mail2,donghy7@mail2,xdliang328@mail}.sysu.edu.cn
michael.c.kampffmeyer@uit.no,s5han@eng.ucsd.edu,{zhengtianxiang1128,allenxuejian}@gmail.com

Figure 1: Results of the proposed Monocular-to-3D virtual try-on network. Given the target clothing image and the reference person image, our M3D-VTON can reconstruct the 3D try-on mesh with the clothing changed and person identity retained.

Abstract

Virtual 3D try-on can provide an intuitive and realistic view for online shopping and has a huge potential commercial value. However, existing 3D virtual try-on methods mainly rely on annotated 3D human shapes and garment templates, which hinders their applications in practical scenarios. 2D virtual try-on approaches provide a faster alternative to manipulate clothed humans, but lack the rich and realistic 3D representation. In this paper, we propose a novel Monocular-to-3D Virtual Try-On Network (M3D-VTON) that builds on the merits of both 2D and 3D approaches. By integrating 2D information efficiently and learning a mapping that lifts the 2D representation to 3D, we make the first attempt to reconstruct a 3D try-on mesh only taking the target clothing and a person image as inputs. The proposed M3D-VTON includes three modules: 1) The Monocular Prediction Module (MPM) that estimates an initial full-body depth map and accomplishes 2D clothes-person alignment through a novel two-stage warping procedure; 2) The Depth Refinement Module (DRM) that refines the initial body depth to produce more detailed pleat and face characteristics; 3) The Texture Fusion Module (TFM) that fuses the warped clothing with the non-target body part to refine the results. We also construct a high-quality synthesized Monocular-to-3D virtual try-on dataset, in which each person image is associated with a front and a back depth map. Extensive experiments demonstrate that the proposed M3D-VTON can manipulate and reconstruct the 3D human body wearing the given clothing with compelling details and is more efficient than other 3D approaches. [1]

1. Introduction

3D virtual try-on, the process of fitting a specific clothing item onto a 3D human shape, has attracted increasing attention due to its promising research and commercial value. Recently, researchers' interest has moved from physics-

[1]code will be available at https://github.com/fyviezhao/M3D-VTON

based [2, 5, 6, 42, 13, 15] or scan-based approaches [37, 27, 44] to learning-based 3D try-on methods [3, 35, 31, 55, 8], dressing a 3D person directly from 2D images and getting rid of costly physics simulation or 3D sensors. However, most of these learning methods [3, 35, 31] build on the parametric SMPL [29] model and depend on some predefined digital wardrobe [3], limiting their real-world applicability. Moreover, the inference speed of these existing 3D approaches is still insufficient, largely due to the optimization cost introduced by the parametric 3D representation.

Related to this, research on image-based virtual try-on aims to fit an in-shop clothing onto the target person and has been explored intensively [17, 48, 52, 16, 51, 22, 9]. Most of these works utilize the Thin Plate Spline (TPS) transformation [4] to achieve the clothes-person alignment and fusion, obtaining photo-realistic try-on results. These 2D methods are attractive due to their small computation cost and extensive amount of available training data on shopping websites. Nevertheless, their try-on results are in 2D image space and ignore the underlying 3D body information, leading to inferior capability of representing the human body.

To address the above limitation of 2D/3D approaches, we propose a light-weight yet effective Monocular-to-3D Virtual Try-On Network (M3D-VTON), which integrates both 2D image-based virtual try-on and 3D depth estimation to reconstruct the final 3D try-on mesh. M3D-VTON consists of three modules as shown in Fig. 2. The first part is the Monocular Prediction Module (MPM), which utilizes a single network to serve the following three purposes: 1) regressing the parameters for the TPS [4] transformation; 2) predicting the conditional person segmentation that is compatible with the in-shop clothing; 3) estimating the full-body depth map. Different from the warping operation in existing 2D try-on methods, MPM first utilizes a novel self-adaptive affine transformation to transform the in-shop clothing to the appropriate size and location before the non-rigid TPS deformation. The second part is the Depth Refinement Module (DRM), which jointly uses the estimated depth map, the warped clothing, the non-target body part and the image gradient information to enhance the geometric details in the depth map. In particular, DRM introduces a depth gradient loss to better exploit the high-frequency details in the inputs. Finally, the Texture Fusion Module (TFM) leverages the 2D information (e.g., warped clothing) and the 3D information (e.g., estimated full-body depth) to synthesize the try-on texture. The collaborative use of the 2D information and the body depth map provides instructive information for the synthesizing process. Given the estimated 2D try-on texture and the refined body depth map, M3D-VTON obtains a colored point cloud and reconstructs the final textured 3D virtual try-on mesh.

We conduct extensive experiments on the new MPV-3D dataset, which is constructed by running PIFuHD [41] on the existing MPV dataset [9]. Compared with other 3D try-on methods, M3D-VTON recovers detailed body shapes and realistic texture color while being more computationally efficient. Our main contributions are:

- We are the first to exploit the merits of both 2D and 3D approaches to solve the monocular-to-3D try-on problem. Our approach reconstructs realistic 3D clothed humans while being faster than pure 3D methods.
- To facilitate more accurate geometric matching between the clothes and the reference person image, we introduce a self-adaptive pre-alignment strategy.
- We utilize the available shadow information in the images and incorporate a novel depth gradient constraint to guide the network to capture and recover intricate geometric changes.
- We construct a new synthesized 3D virtual try-on dataset, MPV-3D, which may stimulate the development of the Monocular-to-3D virtual try-on field. Extensive experiments show the surprising shape recovery and texture generation ability of our M3D-VTON.

2. Related Work

2D Virtual Try-on. 2D virtual try-on aims to transfer a target clothing onto a reference person. A series of works [17, 48, 52, 51, 9, 34, 22, 19] have utilized the non-rigid TPS transformation [4] to obtain appealing virtual try-on results. Most of these works build upon VITON [17], which proposes a coarse-to-fine architecture that first warps the in-shop clothing by TPS and then renders the final try-on result. CP-VTON [48] further trains a geometric matching module and uses a composition mask to better fuse the clothes and person. VTNFP [52] utilizes body segmentation as the synthesis guidance, producing clearer skin texture. ACGPN [51] proposes a second-order constraint on TPS parameters to stabilize the warping process. Our method not only inherits the benefits of the aforementioned methods but also generates realistic 3D clothed human, providing an economic solution for monocular-to-3D virtual try-on.

3D Virtual Try-on. Compared to the tasks of 3D human reconstruction and performance capturing [54, 14, 11, 39, 53, 24, 36, 26, 21, 1], 3D virtual try-on is more challenging due to the complex deformation of clothes. PIFuHD [41] provides a high-fidelity single-view textureless 3D human reconstruction pipeline that produces realistic clothing details, however, it can not perform garment transfer. MGN [3] can predict parametric garment geometry and layer it on top of the SMPL [29] model. Thanks to the layered representation, MGN can dress varying body shapes and poses but is limited to garments from their predefined digital wardrobe. DeepFashion3D [55] provides more 3D clothes data to achieve more challenging clothing reconstruction. Pix2Surf [32] also aims to transfer more

Methods	CC	3D	FBT	SG	ED	FI
VITON [17]	Y	N	N	N	Y	Y
CP-VTON [48]	Y	N	N	N	Y	Y
ACGPN [51]	Y	N	N	Y	Y	Y
PIFuHD [41]	N	Y	N	N	N	Y
MGN [3]	Y	Y	Y	Y	N	N
DeepFashion3D [55]	N	Y	N	Y	N	-
Pix2Surf [31]	Y	Y	N	N	Y	Y
Deephuman [45]	N	Y	N	Y	N	Y
FACSMILE [43]	N	N	Y	N	Y	N
NormalGAN [49]	N	Y	Y	N	N	Y
M3D-VTON(ours)	Y	Y	Y	Y	Y	Y

Table 1: Comparison of M3D-VTON to related work in terms of their properties with Yes (Y) or No (N). The first three rows are 2D try-on methods, the middle rows are 3D try-on/reconstruction methods, and the bottom rows except ours are human depth estimation methods. Categorized based on: Changeable Clothes (CC); Clothed 3D Body (3D); Full Body Texture (FBT); Semantic Guidance (SG); Easy-to-get Dataset (ED); Fast Inference (FI).

in-the-wild clothes images onto the SMPL model by learning dense correspondences between 2D garment silhouettes and UV maps of 3D garment surfaces. However, both DeepFashion3D and Pix2Surf can not show the body texture. Besides, almost all these methods require a scanned 3D dataset for training, which is expensive to collect compared with our proposed high-quality synthesized dataset. Our method can recover both clothed body shape and texture, providing a more practical solution for 3D try-on.

Human Depth Estimation. Recently, non-parametric 3D human reconstruction has been proposed to better capture shape details by predicting depth maps. Moulding Humans [10] estimates the front and back depth map from a single RGB image to generate a textureless 3D human. FACSMILE [43] is similar and adds a normal constraint to carve local depth details but manipulates naked bodies and is not cloth-aware. DeepHuman [46] also utilizes a normal map to refine the estimated depth but only generates the frontal body part, limiting its practical application. NormalGAN [49] further uses an adversarial learning framework conditioned on normal maps to recover the textured 3D human body. However, NormalGAN requires the ground-truth depth map as input, which needs to be collected using expensive depth sensors. Compared with the above methods, our M3D-VTON is trained on high-quality synthesised data and allows for cloth-aware human manipulation. For ease of comparison, Table 1 presents an overview of the properties of M3D-VTON and the most related approaches.

3. M3D-VTON

To facilitate 3D virtual try-on, we propose a novel Monocular-to-3D Virtual Try-On Network (M3D-VTON) that takes a clothing image C and a person image I as inputs, and reconstructs a 3D try-on mesh O with clothes changed and person identity preserved. As illustrated in Fig. 2, M3D-VTON is composed of the Monocular Prediction Module (MPM), the Depth Refinement Module (DFM), and the Texture Fusion Module (TFM).

3.1. Monocular Prediction Module

This module plays a preparatory role in the proposed M3D-VTON. It provides constructive guidance for the other two modules by warping the in-shop clothing, predicting a conditional person segmentation, and by estimating a base 3D shape using a multi-target network. All these tasks can be accomplished by utilizing the features extracted from the target clothing C and the clothing-agnostic person representation A. A consists of a 25-channel pose map (obtained by applying OpenPose [7] on person image I), a 3-channel unchanged person part (I^p) (obtained by applying [28] on I), and a 1-channel coarse person mask that have been concatenated. We explain the three sub-branches of MPM in the following sections.

Clothing Warping Branch. Inspired by [38], the first branch of the MPM utilizes an end-to-end trainable geometric matching network to achieve the texture-preserving clothing-person alignment. Specifically, as part of the geometric matching network, the features extracted by the encoders $\mathcal{E}_C$ and $\mathcal{E}_A$ are fed into the feature correlation layer to calculate the matching score, which is used by the regressor $\mathcal{R}$ to predict the TPS transformation [4] parameters θ (see Fig. 2). However, directly estimating θ is non-trivial since there is a huge gap in size between the in-shop clothing C and the arm-torso region of the reference person I^{at}. We therefore extract I^{at} from I by applying person segmentation [28] and design a **self-adaptive pre-alignment** procedure to transform C to the proper position and size before conducting the TPS transformation. We formulate the procedure as an affine transformation:

$$C^{aff} = \begin{bmatrix} R & 0 \\ 0 & R \end{bmatrix} C + \begin{bmatrix} x^c_{I^{at}} - x^c_C \\ y^c_{I^{at}} - y^c_C \end{bmatrix}, \quad (1)$$

where C^{aff} denotes the transformed clothing item (see Fig. 2), $(x^c_{I^{at}}, y^c_{I^{at}})$ and (x^c_C, y^c_C) represent the center of I^{at} and C, respectively. R is a rescaling factor computed by comparing the aspect ratio to ensure that the aligned clothing is larger than or at least equal to the arm-torso region:

$$R = \begin{cases} \frac{h^{at}_I}{h_C}, & \frac{w_C}{h_C} \geqslant \frac{w^{at}_I}{h^{at}_I} \\ \frac{w^{at}_I}{w_C}, & \frac{w_C}{h_C} < \frac{w^{at}_I}{h^{at}_I}. \end{cases} \quad (2)$$

An intuitive understanding of Eq. 1 is that it first center aligns C and I^{at} and scales C to roughly the same size as I^{at} to simplify the TPS warping step. The effectiveness of the alignment procedure is illustrated in Fig. 3.

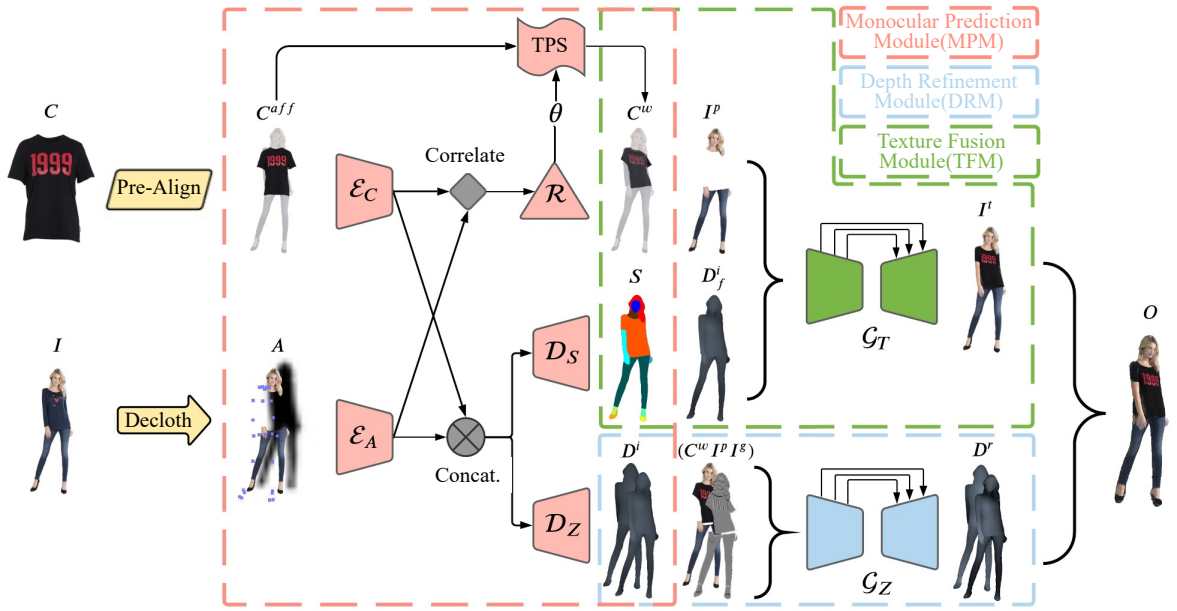

Figure 2: Overview of the proposed M3D-VTON. The pipeline contains three modules with the following tasks. a) **Monocular Prediction Module (MPM)**: obtaining the cloth-agnostic person representation A through the decloth process, deforming the in-shop clothing C to the warped clothing C^w via a self-adaptive pre-alignment followed by a TPS transformation, predicting a person segmentation S, and estimating an initial double-depth map D^i. b) **Depth Refinement Module (DRM)**: given the double-depth map D^i, the warped clothing C^w, the preserved person part I^p, and their shadow information I^g as inputs, this module refines the initial depth map and produces more local details (like cloth folds and face structure) by incorporating a novel depth gradient constraint. c) **Texture Fusion Module (TFM)**: rendering the results I^t under the guidance of the semantic layout from MPM by fusing the warped clothes and the preserved texture information. Once I^t and the refined depth map D^r are spatially aligned, forming an RGB-D representation, we can directly get colored point clouds and triangulate them to obtain the 3D clothed human O wearing the target clothes and with its identity preserved.

Figure 3: Verification of the pre-alignment strategy. The proposed self-adaptive transformations increase the quality of the warping (as shown in the third column).

Given C^{aff}, we pass both C^{aff} and the clothing-agnostic person representation A to the geometric matching network to regress the TPS parameters θ, which are then used to warp C^{aff} to the warped clothing C^w. During training, the difference between C^w and the ground-truth I^c (clothing-on-person) is used to define the warping loss:

$$\mathcal{L}_w = \|C^w - I^c\|_1. \tag{3}$$

Conditional Segmentation Estimation Branch. The goal of this branch is to estimate the person segmentation supposing now wearing the desired clothing, which delineates different parts of the reference person (e.g., the sleeve-arm boundary). The segmentation mask provides inpainting guidance for the following texture fusion module to mitigate skin texture degradation or clothing-skin penetration especially for the case of self-occlusion or large clothing variation. As shown in Fig. 2, the feature maps from $\mathcal{E}_C$ and $\mathcal{E}_A$ are concatenated together and sent to the segmentation decoder $\mathcal{D}_S$ to generate the conditional person segmentation S. Although only paired (C, I) images[2] are fed to the model during training, the network can generalize to unpaired data at inference time due to the benefit of its clothes-agnostic representation. During training, we use the pixel-level cross-entropy [12] $\mathcal{L}_s$ to optimize this branch.

Depth Estimation Branch. The last branch in MPM aims at estimating a base 3D shape of the reference person. We represent the 3D shape in a double-depth form similar to [10], i.e. a front and a back depth map corresponding to the respective sides of the 3D human representation. In this branch, the concatenated feature map is upsampled by the depth decoder $\mathcal{D}_Z$ to generate the front and the back depth. During training, the loss function can be formulated as:

$$L_z = \|D_f^i - D_f^{gt}\|_1 + \|D_b^i - D_b^{gt}\|_1, \tag{4}$$

[2]Reference person I is wearing clothing C.

where D_f^i and D_b^i represent the estimated front and back depth, and the superscript i means "initial". D_f^{gt} and D_b^{gt} are the corresponding ground-truth depth maps.

We refer to the estimated depth maps as "initial" depth since there are not enough clues for $\mathcal{D}_Z$ to infer the complete details of the warped clothing, such as the pleat details. To obtain more precise 3D information, the initial depth map will be refined in the depth refinement module, which will be explained in Section 3.2.

We train the three branches together within a multi-target network and combine the three aforementioned losses to yield the full loss of MPM:

$$\mathcal{L}_{\text{MPM}} = \mathcal{L}_w + \mathcal{L}_s + \mathcal{L}_z. \tag{5}$$

3.2. Depth Refinement Module

The reasons that the initially estimated depth map from MPM fails to capture geometric details (e.g., clothing details, face characteristics) are twofold: (1) the inputs of the MPM lack the warped clothing, which is crucial to carve clothing pleats; (2) the L1 depth loss used in MPM tends to penalize low-frequency differences between the estimated and the ground truth depth map, resulting in an over-smoothed depth result. To add high-frequency depth details, we propose the Depth Refinement Module (DRM), which further exploits the brightness changes in the warped clothing C^w and the preserved person part I^p to refine the initial depth map. Specifically, we apply the Sobel operator on C^w and I^p and concatenate the gradient images to obtain the image gradient I^g, representing the changes in brightness. Then, I^g, C^w, I^p and the initial depth map D^i are sent to an UNet-like generator $\mathcal{G}_Z$ to produce the refined depth map D^r. During training, we propose two special losses to enable the network to capture the high-frequency details. Firstly, inspired by [20], we replace the vanilla L1 depth loss with a Log-L1 version, which penalizes close points more heavily and therefore guides the estimation to focus on intricate local details, which is formulated as:

$$\mathcal{L}_{\text{depth}} = \frac{1}{n}\sum_{i=1}^{n} \ln\left(\epsilon_i + 1\right), \tag{6}$$

where ϵ_i is the L1 loss of the i-th depth point, and n is the total number of the front/back depth map points.

Secondly, to further strengthen the depth estimation and capture geometric details especially at the boundary of adjacent body parts, we incorporate a **depth gradient loss** :

$$\mathcal{L}_{\text{grad}} = \frac{1}{n}\sum_{i=1}^{n} \left(\ln\left(\nabla_x\left(\epsilon_i\right) + 1\right) + \ln\left(\nabla_y\left(\epsilon_i\right) + 1\right)\right), \tag{7}$$

where ∇ denotes the Sobel operator.

Note that normal maps can be generated from depth gradient maps [33] and that Eq. 7 thus also penalizes the difference in normal maps. It is shown in [49] that normal maps tend to contain more detailed geometric information than depth maps, therefore constraints along the normal direction can help recover geometric details and delineate the boundary of adjacent body parts, where the depth gradient is generally large.

The above two losses work in a complementary manner to constrain different types of errors: a) $\mathcal{L}_{\text{depth}}$ ensures consistency along the z-direction, b) $\mathcal{L}_{\text{grad}}$ does the same for the x-, y- and thus normal direction. We therefore utilize a weighted sum of the aforementioned losses to train DRM:

$$\mathcal{L}_{\text{DRM}} = \lambda_{\text{depth}}\mathcal{L}_{\text{depth}} + \lambda_{\text{grad}}\mathcal{L}_{\text{grad}}, \tag{8}$$

where λ_{depth}, λ_{grad} are set to 1.0, 0.5 respectively.

3.3. Texture Fusion Module

To synthesize photo-realistic body texture for the final 3D human mesh, we propose the Texture Fusion Module (TFM) which fuses the warped clothing with the unchanged person part to render seamless try-on results. TFM takes the preserved person part I^p, the warped clothing C^w, the predicted segmentation S, and the estimated initial front depth D_f^i as inputs, and generates a coarse try-on result $\tilde{I^c}$ as well as a fusion mask $\tilde{M}$. The 2D clues of I^p, C^w, and S provide the person appearance, clothing texture, and semantic guidance for the network. Further, TFM also considers the body depth map D_f^i, which contains the spatial information of different body parts along the z-axis. Under the extra guidance of D_f^i, TFM is capable of synthesizing the try-on result more precisely even in challenging self-occlusion cases. Finally, the fusion mask $\tilde{M}$ is used to fuse C^w and $\tilde{I^c}$ into the refined try-on result I^t, which can be formulated as:

$$I^t = C^w \odot \tilde{M} + \tilde{I^c} \odot (1 - \tilde{M}). \tag{9}$$

TFM is trained using the perceptual loss $\mathcal{L}_{\text{perc}}$ [23] between the refined try-on result I^t and the real person image I, the L1 loss $\mathcal{L}_{\text{try-on}}$ between I^t and I, as well as the L1 loss $\mathcal{L}_{\text{mask}}$ between the estimated fusion mask $\tilde{M}$ and the real clothing-on-person mask M. The combined loss for TFM can thus be formulated as:

$$\mathcal{L}_{\text{TFM}} = \mathcal{L}_{\text{perc}} + \mathcal{L}_{\text{try-on}} + \mathcal{L}_{\text{mask}}. \tag{10}$$

In the end, we can unproject the front-view and the back-view depth maps from DRM to get the 3D point clouds and triangulate them with screened Poisson reconstruction [25]. Since the try-on result from TFM is spatially aligned with the depth map, it can directly be used to color the front side of the mesh. As for the back texture, we first inpaint the try-on image using the fast matching method proposed in [47], filling the face area with the surrounding hair color, and then mirror the inpainted "back" view image to texture the back-side of the mesh. This allows us to successfully achieve the monocular-to-3D conversion, producing the reconstructed 3D clothed human with retained identity.

Figure 4: Qualitative comparison for the 2D try-on task. The first columns represent the inputs, columns 3 to 6 are prior approaches, and column 7 illustrates our proposed approach.

4. Experiments

4.1. Dataset Generation

We construct the first monocular-to-3D try-on dataset MPV-3D based on the MPV dataset [9], which contains person images covering a wide range of poses and upper-body garments[3]. MPV-3D contains 6566 clothes-person image pairs (C, I) of size 512×320, in which each person image is associated with a front and a back depth map, D_f and D_b, respectively. We obtain the depth maps and set them as the pseudo ground truth of our M3D-VTON by applying PIFuHD [41] on the full-body front-faced person images from the MPV dataset and then orthographically projecting the generated human mesh to the double-depth maps. The dataset is further divided into a train set and a test set with 5632 and 934 four-tuples (C, I, D_f, D_b) respectively, and the test set is shuffled to form the unpaired (C, I) list for quality evaluation.

4.2. Implementation Details

The MPM is trained separately from the DRM and the TFM as it provides the inputs to these modules, while DRM and TFM are trained together[4]. Each module is trained for 100 epochs using the Adam optimizer, with $\beta_1 = 0.5$, $\beta_2 = 0.999$, and the learning rate is initialized as 0.0002 with a linearly decay to 0 in the last 50 epochs. The batch size is 8. The model is implemented in Pytorch and trained on a single NVIDIA 2080ti GPU. During training, the reference person wears the same clothing as the target in-shop clothes as the try-on result for unpaired clothes and persons are not available as supervision. However, during testing, the target clothing is different from the clothing on the person and inference is performed in an end-to-end manner.

[3]Examples are shown in the supplementary.

[4]We provide the complete architecture details in the supplementary.

Method	SSIM ↑	FID ↓	HE ↑
VITON [17]	0.8807	28.43	21.35%
CP-VTON [48]	0.8503	20.05	10.65%
CP-VTON+ [30]	0.8782	23.18	12.57%
ACGPN [51]	**0.8924**	20.19	13.50%
M3D-VTON	0.8804	**20.04**	**41.92%**

Table 2: Quantitative comparisons to other 2D try-on methods. For fair comparison, we crop and resize our full-body try-on results to half-body like those in Fig. 4, as other methods originally perform half-body try-on.

4.3. 2D Try-on Comparison with SOTA methods

We compare our 2D try-on results with the existing state-of-the-art 2D try-on methods: VITON [17], CP-VTON [48], CP-VTON+ [30], and ACGPN [51].

A qualitative comparison is shown in Fig. 4. VITON lacks texture details of the clothing and fails to synthesize arms in self-occlusion cases. Although CP-VTON and CP-VTON+ can better preserve clothing texture, they perform poorly when the clothing is occluded by body parts. ACGPN fails to synthesize complete arms and may synthesize artifacts in the clothes region due to the stochasticity introduced by its segmentation estimation network. Due to our two-stage warping strategy, M3D-VTON more accurately preserves the clothing texture, and synthesizes body parts precisely through the collaborative guidance of the conditional segmentation and the body depth map.

For the quantitative comparison, we adopt the Structural SIMilarity index measure (SSIM) [50] and the Fréchet Inception Distance (FID)[18] to measure the similarity between the synthesized and the real images. Further, we conduct a human evaluation (HE) to assess the 2D try-on results from M3D-VTON and the other four baselines. Specifically, we invited 26 volunteers to complete a questionnaire that contains 40 assignments. In each assignment, given a

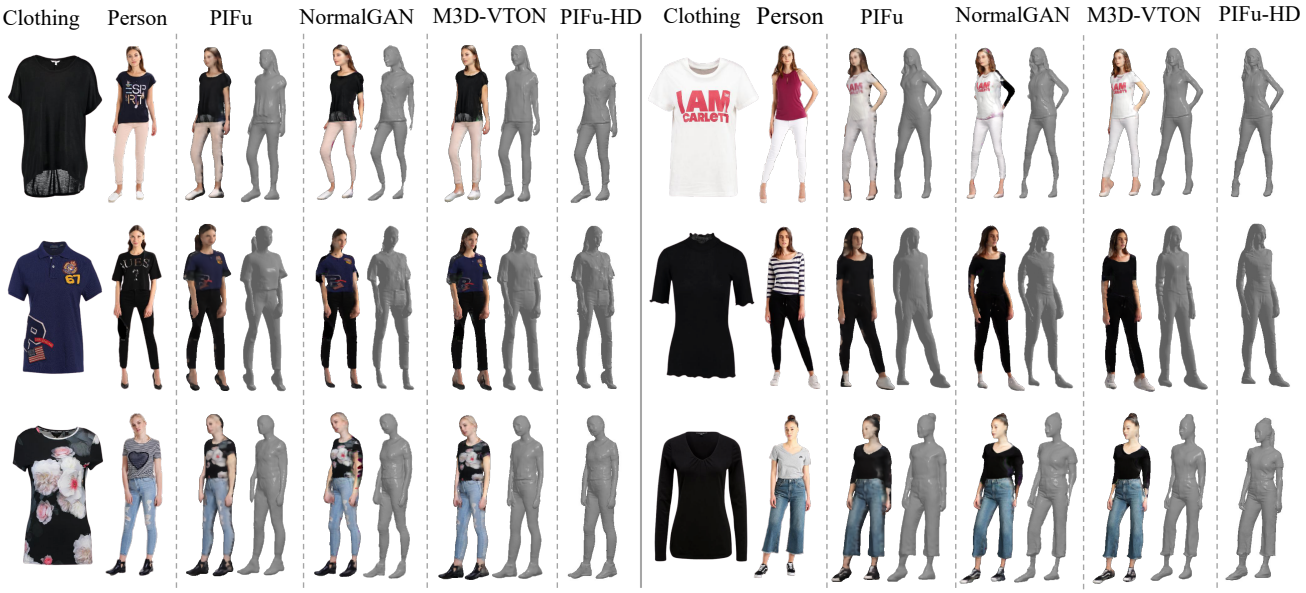

Figure 5: Qualitative comparisons of 3D try-on results. The first two and the last columns respectively represent the inputs and the (pseudo) ground truth PIFu-HD mesh, while the others are 3D try-on results (w/ and w/o texture) from the different methods. The human mesh generated by our M3D-VTON contains more texture details and a more accurate shape compared to PIFu [40] and NormalGAN [49] (note that NormalGAN uses GT front depth map as its input).

person image and a clothing image, the volunteers are required to select the most realistic try-on image out of the ones produced by the five methods.

As shown in Table 2, M3D-VTON obtains the lowest FID and highest human evaluation score, outperforming other baseline methods. Its SSIM score is on-par with the best performing model. To fairly compare with baseline methods which trained on the half-body VITON dataset [17], the baseline methods take the cropped half-body images from MVP-3D as inputs and synthesize half-body results during testing. The full-body results of M3D-VTON are cropped to half-body images (as shown in Fig. 4) following the same cropping procedure.

4.4. 3D Try-on Comparison with SOTA methods

Since this is the first work that explores the monocular-to-3D virtual try-on setting, we design three hybrid models to conduct 3D try-on comparisons. Specifically, we first obtain the 2D virtual try-on result using CP-VTON and then generate the 3D try-on mesh using the state-of-the-art 3D human reconstruction approaches PIFu [40], NormalGAN [49], and Deephuman [45]. The qualitative and quantitative comparisons are shown in Fig. 5 and Table 3, respectively. Since Deephuman does not recover the backside of the 3D shape, we compare with it only quantitatively.

In Fig. 5, the hybrid CP-VTON+PIFu model produces plausible 3D shape results but fails to recover detailed texture due to its unreliable implicit texture color inference.

Method	Abs. ↓	Sq. ↓	RMSE ↓	HE ↑
Deephuman [45]	17.35	1.271	22.44	-
NormalGAN [49]	15.41	0.778	18.94	21.3%
PIFu [40]	8.376	1.813	27.57	11.3%
M3D-VTON (ours)	**7.880**	**0.385**	**11.27**	**67.4%**

Table 3: Quantitative comparisons to other 3d human reconstruction methods. All values have been divided by 10^{-3} for readability. Note for PIFu and M3D-VTON, we average their double-depth score. The scores for Deephuman and NormalGAN are computed from single depth since they only recover front and back depth, respectively.

Unlike PIFu, NormalGAN uses the double-depth representation and directly sets the 2D image as mesh texture. However, NormalGAN requires a noisy ground truth depth map as input to infer the back shape and although we simulate the depth generation process in NormalGAN, it still tends to produce over-slim 3D persons. Compared with these hybrid methods, our M3D-VTON generates more realistic 3D persons and preserves detailed texture within a single model.

The results of the quantitative comparison are shown in Table 3. We use three common depth estimation metrics: Absolute Relative error (Abs.), Squared Relative error (Sq.) and Root Mean Squared Error (RMSE). Our method outperforms the benchmark models on all of the four measurements including human evaluation (HE), illustrating the superior shape generation ability of M3D-VTON. Finally, our method takes about 4 seconds to run for a given MPV-

MTM	Pre-align	IoU	Pre-align	IoU
	✗	0.708	✓	**0.737**
	Segmt. S	Depth D_f^i	SSIM ↑	FID ↓
TFM	✗	✗	0.9348	16.52
	✓	✗	0.9434	16.01
	✗	✓	0.9418	15.96
	✓	✓	**0.9435**	**15.74**
	Grad. I^g	Loss $\mathcal{L}_{\text{grad}}$	Sq. ↓	RMSE ↓
DRM	✗	✗	0.0824	5.8369
	✓	✗	0.0896	5.7650
	✓	✓	**0.0801**	**5.7420**

Table 4: The ablation study on the three modules. Note values of Sq. and RMSE are divided by 10^{-3} for readability. Here the scores of DRM are for front depth, as we do not have the back-side gradient image. And the TFM scores are computed from full-body try-on results.

Figure 6: Visual comparisons to verify the effectiveness of the segmentation guidance and depth guidance in TFM.

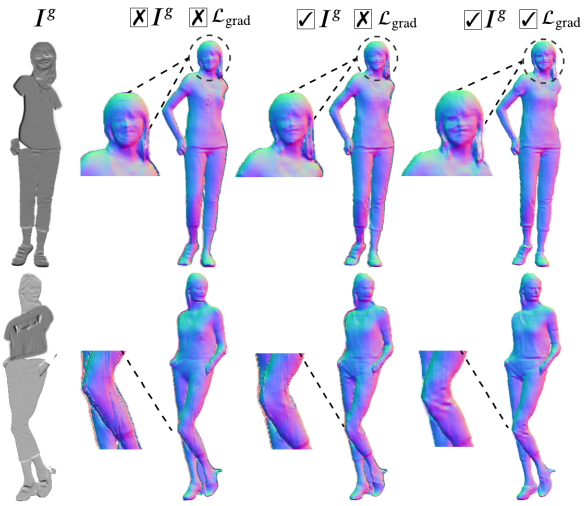

Figure 7: Visualized ablation study on DRM. $\mathcal{L}_{\text{grad}}$ is beneficial for producing geometric details (see last column).

3D image pair (most computational cost occurs during the poisson reconstruction process), which is clearly faster and more efficient than pure 3D virtual try-on (such as Multi-Garment Net [3], roughly 17s to run) or 3D human reconstruction (such as PIFu [40], roughly 10s to run) methods.

4.5. Ablation Study

We conduct ablation experiments on the three modules of M3D-VTON to verify their effectiveness.

Effectiveness of the Self-Adaptive Pre-Alignment in MPM. Fig. 3 shows that directly applying TPS results in excessive deformation and fails to warp clothes properly. Our two-stage warping with pre-alignment, instead, can generate gentle deformation and obtains precisely warped clothes. Quantitative results (Table 4, 1st row), verify this as the IoU between the warped clothes and the clothes region from the reference person increases with pre-alignment.

Effectiveness of Depth and Segmentation Guidance in TFM. Fig. 6 illustrates the need that using these two guidance independently can help alleviate the self-occlusion issue. Furthermore, under their collaborative guidance, TFM can further improve the fidelity of arms in the synthesized results. Table 4 corroborates that they both contribute positively to the M3D-VTON. Note that the SSIM and FID scores here are calculated on the full-body results, while the scores in Table 2 are reported for the cropped half-body results to fit the setting of ACGPN for fair comparison.

Effectiveness of depth gradient constraint in DRM. Table 4 and Fig. 7 show that the image gradient inputs and the proposed depth gradient constraint can improve the depth prediction and guide the DRM to carve more intricate details onto the 3D shape. The black dotted circles in Fig. 7 highlight the improvements brought by these terms.

5. Conclusion

In this work, we propose a computational efficient Monocular-to-3D Virtual Try-On Network (M3D-VTON) that builds on the merits of both 2D and 3D approaches to produce the 3D try-on mesh from 2D information. Our M3D-VTON decomposes the 3D try-on task into a 2D try-on and a body depth estimation problem. In future work, we will investigate if the two can further promote each other in a cyclic-manner. To get more realistic texture fusion results, M3D-VTON utilizes a two-stage warping strategy as well as segmentation and depth guidance . We also introduce a novel depth gradient constraint to generate more detailed depth maps. Our method provides a faster and more economic solution for the monocular-to-3D virtual try-on task.

6. Acknowledgements

This work was supported in part by National Key R&D Program of China under Grant No. 2020AAA0109700, National Natural Science Foundation of China (NSFC) under Grant No.U19A2073 and No.61976233, Guangdong Province Basic and Applied Basic Research (Regional Joint Fund-Key) Grant No.2019B1515120039, Guangdong Outstanding Youth Fund (Grant No. 2021B1515020061), Shenzhen Fundamental Research Program (Project No. RCYX20200714114642083, No. JCYJ20190807154211365), Zhejiang Lab's Open Fund (No. 2020AA3AB14).

References

[1] Thiemo Alldieck, Gerard Pons-Moll, Christian Theobalt, and Marcus Magnor. Tex2shape: Detailed full human body geometry from a single image. In IEEE International Conference on Computer Vision (ICCV). IEEE, 2019. 2

[2] David Baraff and Andrew Witkin. Large steps in cloth simulation, 1998. 2

[3] Bharat Lal Bhatnagar, Garvita Tiwari, Christian Theobalt, and Gerard Pons-Moll. Multi-garment net: Learning to dress 3d people from images. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 5420–5430, 2019. 2, 3, 8

[4] F. L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 11(6):567–585, 1989. 2, 3

[5] Robert Bridson, Ronald Fedkiw, and John Anderson. Robust treatment of collisions, contact and friction for cloth animation. ACM Trans. Graph., 21(3):594–603, July 2002. 2

[6] R. Bridson, S. Marino, and R. Fedkiw. Simulation of clothing with folds and wrinkles. In Proceedings of the 2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA '03, page 28–36, Goslar, DEU, 2003. Eurographics Association. 2

[7] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019. 3

[8] Enric Corona, Albert Pumarola, Guillem Alenyà, Gerard Pons-Moll, and Francesc Moreno-Noguer. Smplicit: Topology-aware generative model for clothed people. In CVPR, 2021. 2

[9] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 9026–9035, 2019. 2, 6

[10] Valentin Gabeur, Jean-Sebastien Franco, Xavier Martin, Cordelia Schmid, and Gregory Rogez. Moulding humans: Non-parametric 3d human shape estimation from single images. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019. 3, 4

[11] Andrew Gilbert, Marco Volino, John Collomosse, and Adrian Hilton. Volumetric performance capture from minimal camera viewpoints. In Proceedings of the European Conference on Computer Vision (ECCV), September 2018. 2

[12] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 932–940, 2017. 4

[13] Peng Guan, Loretta Reiss, David A Hirshberg, Alexander Weiss, and Michael J Black. Drape: Dressing any person. ACM Transactions on Graphics, 31(4), 2012. 2

[14] Marc Habermann, Weipeng Xu, Michael Zollhofer, Gerard Pons-Moll, and Christian Theobalt. Deepcap: Monocular human performance capture using weak supervision. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. 2

[15] Fabian Hahn, Bernhard Thomaszewski, Stelian Coros, Robert WSumner, Forrester Cole, Mark Meyer, Tony DeRose, and Markus Gross. Subspace clothing simulation using adaptive bases. ACM Transactions on Graphics, 33(4), 2014. 2

[16] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R. Scott. Clothflow: A flow-based model for clothed person generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 10471–10480, 2019. 2

[17] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. Viton: An image-based virtual try-on network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7543–7552, 2018. 2, 3, 6, 7

[18] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In Advances in Neural Information Processing Systems, pages 6626–6637, 2017. 6

[19] Chia-Wei Hsieh, Chieh-Yun Chen, Chien-Lung Chou, Hong-Han Shuai, Jiaying Liu, and Wen-Huang Cheng. Fashionon: Semantic-guided image-based virtual try-on with detailed human and clothing information. In Proceedings of the 27th ACM International Conference on Multimedia, pages 275–283, 2019. 2

[20] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries, 2018. 5

[21] Zeng Huang, Yuanlu Xu, Christoph Lassner, Hao Li, and Tony Tung. Arch: Animatable reconstruction of clothed humans. In IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. 2

[22] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on. In Proceedings of the European Conference on Computer Vision, 2020. 2

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In Proceedings of the European Conference on Computer Vision, pages 694–711, 2016. 5

[24] Angjoo Kanazawa, Michael J. Black, David W. Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In Computer Vision and Pattern Recognition (CVPR), 2018. 2

[25] Michael Kazhdan and Hugues Hoppe. Screened poisson surface reconstruction. ACM Trans. Graph., 32(3), July 2013. 5

[26] Muhammed Kocabas, Nikos Athanasiou, and Michael J. Black. Vibe: Video inference for human body pose and shape estimation. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2020. 2

[27] Zorah Lahner, Daniel Cremers, and Tony Tung. Deepwrinkles: Accurate and realistic clothing modeling. In

Proceedings of the European Conference on Computer Vision (ECCV), pages 667–684, 2018. 2

[28] Xiaodan Liang, Ke Gong, Xiaohui Shen, and Liang Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018. 3

[29] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. ACM Transactions on Graphics, 34(6), 2015. 2

[30] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In CVPRW, 2020. 6

[31] Aymen Mir, Thiemo Alldieck, and Gerard Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7023–7034, 2020. 2, 3

[32] A. Mir, T. Alldieck, and G. Pons-Moll. Learning to transfer texture from clothing images to 3d humans. In 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 7021–7032, 2020. 2

[33] Y. Nakagawa, H. Uchiyama, H. Nagahara, and R. Taniguchi. Estimating surface normals with depth image gradients for fast and accurate registration. In 2015 International Conference on 3D Vision, pages 640–647, 2015. 5

[34] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. Image based virtual try-on network from unpaired data. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5184–5193, 2020. 2

[35] Chaitanya Patel, Zhouyingcheng Liao, and Gerard Pons-Moll. Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7365–7375, 2020. 2

[36] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2019. 2

[37] Gerard Pons-Moll, Sergi Pujades, Sonny Hu, and Michael J Black. Clothcap: seamless 4d clothing capture and retargeting. ACM Transactions on Graphics, 36(4), 2017. 2

[38] I. Rocco, R. Arandjelović, and J. Sivic. Convolutional neural network architecture for geometric matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017. 3

[39] Nitin Saini, Eric Price, Rahul Tallamraju, Raffi Enficiaud, Roman Ludwig, Igor Martinovic, Aamir Ahmad, and Michael J. Black. Markerless outdoor human motion capture using multiple autonomous micro aerial vehicles. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019. 2

[40] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In The IEEE International Conference on Computer Vision (ICCV), pages 2304–2314, 2019. 7, 8

[41] Shunsuke Saito, Tomas Simon, Jason Saragih, and Hanbyul Joo. Pifuhd: Multi-level pixel-aligned implicit function for high-resolution 3d human digitization. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 84–93, 2020. 2, 3, 6

[42] Andrew Selle, Jonathan Su, Geoffrey Irving, and Ronald Fedkiw. Robust high-resolution cloth using parallelism, history-based collisions, and accurate friction. IEEE Transactions on Visualization and Computer Graphics, 15(2):339–350, Mar. 2009. 2

[43] David Smith, Matthew Loper, Xiaochen Hu, Paris Mavroidis, and Javier Romero. Facsimile: Fast and accurate scans from an image in less than a second. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019. 3

[44] Carsten Stoll, Juergen Gall, Edilson De Aguiar, Sebastian Thrun, and Christian Theobalt. Video-based reconstruction of animatable human characters. ACM Transactions on Graphics (TOG), 29(6):1–10, 2010. 2

[45] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. A neural network for detailed human depth estimation from a single image. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 7750–7759, 2019. 3, 7

[46] Sicong Tang, Feitong Tan, Kelvin Cheng, Zhaoyang Li, Siyu Zhu, and Ping Tan. A neural network for detailed human depth estimation from a single image. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), October 2019. 3

[47] Alexandru Telea. An image inpainting technique based on the fast marching method. J. Graphics, GPU, & Game Tools, 9(1):23–34, 2004. 5

[48] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In Proceedings of the European Conference on Computer Vision, pages 589–604, 2018. 2, 3, 6

[49] Lizhen Wang, Xiaochen Zhao, Tao Yu, Songtao Wang, and Yebin Liu. Normalgan: Learning detailed 3d human from a single rgb-d image. In Proceedings of the European Conference on Computer Vision, 2020. 3, 5, 7

[50] Zhou Wang, Alan Conrad Bovik, Hamid Rahim Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing, 13(4):600–612, 2004. 6

[51] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7850–7859, 2020. 2, 3, 6

[52] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In The IEEE International Conference on Computer Vision, pages 10511–10520, 2019. 2

[53] Tao Yu, Zerong Zheng, Yuan Zhong, Jianhui Zhao, Qionghai Dai, Gerard Pons-Moll, and Yebin Liu. Simulcap : Single-view human performance capture with cloth simulation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2019. 2

[54] Zerong Zheng, Tao Yu, Hao Li, Kaiwen Guo, Qionghai Dai, Lu Fang, and Yebin Liu. Hybridfusion: Real-time performance capture using a single depth sensor and sparse imus. In Proceedings of the European Conference on Computer Vision (ECCV), September 2018. 2

[55] Heming Zhu, Yu Cao, Hang Jin, Weikai Chen, Dong Du, Zhangye Wang, Shuguang Cui, and Xiaoguang Han. Deep fashion3d: A dataset and benchmark for 3d garment reconstruction from single images. In Proceedings of the European Conference on Computer Vision, pages 512–530, 2020. 2, 3