CloTH-VTON: Clothing Three-dimensional reconstruction for Hybrid image-based Virtual Try-ON

Matiur Rahman Minar[0000−0002−3128−2915] and Heejune Ahn[0000−0003−1271−9998]

Seoul National University of Science and Technology, Seoul, South Korea.
{minar,heejune}@seoultech.ac.kr

Abstract. Virtual clothing try-on, transferring a clothing image onto a target person image, is drawing industrial and research attention. Both 2D image-based and 3D model-based methods proposed recently have their benefits and limitations. Whereas 3D model-based methods provide realistic deformations of the clothing, it needs a difficult 3D model construction process and cannot handle the non-clothing areas well. Image-based deep neural network methods are good at generating disclosed human parts, retaining the unchanged area, and blending image parts, but cannot handle large deformation of clothing. In this paper, we propose CloTH-VTON that utilizes the high-quality image synthesis of 2D image-based methods and the 3D model-based deformation to the target human pose. For this 2D and 3D combination, we propose a novel 3D cloth reconstruction method from a single 2D cloth image, leveraging a 3D human body model, and transfer to the shape and pose of the target person. Our cloth reconstruction method can be easily applied to diverse cloth categories. Our method produces final try-on output with naturally deformed clothing and preserving details in high resolution.

Keywords: Virtual try-on, 3D cloth reconstruction, Generative model.

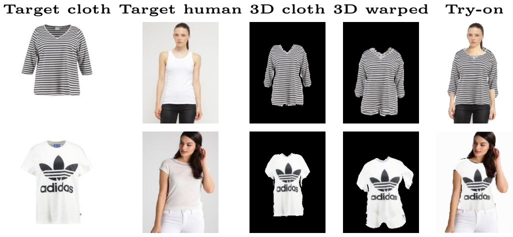

Fig. 1. Results of our CloTH-VTON. From left to right: input clothes, reference humans, reconstructed 3D clothes (shape and pose transferred respectively), and final fusion results. CloTH-VTON produces realistic output with high details.

1 Introduction

Virtual try-on (VTON) technologies can help customers in making their clothing purchase decisions for shopping online. Although 3D model-based VTON approaches could produce realistic 3D VTON results, 3D modeling and scanning for real clothing and human body are time-consuming and expensive [1,2]. Since the emergence of deep learning, 2D image-based approaches have gained more attention, mainly due to the lower costs of 2D data collection and less computational time than 3D [3,4]. However, manipulating the human and clothing shape and texture is an extremely challenging task, due to the huge variety of poses and shapes. Thus, deep neural networks, i.e., image-based and 2D VTON technologies suffer from variations of clothing and human styles. Figure 2 shows a comparison of state-of-the-art (SOTA) methods and their limitations. Especially, 2D image-based methods cannot deform the input clothing to the 3D pose of the target person [5]. From the statistical 3D human model [6] and 3D reconstruction studies, many research works are ongoing on 3D human or clothed-human digitization. Recently, works have been done in 3D garments or clothing reconstruction [7,8,2,1]. However, since the separate reconstruction of clothing and humans is necessary for VTON, prior works on 3D garment reconstruction [1,2,7] methods work only with very limited clothing categories. Also, full 3D reconstruction of human parts like the face with hair for VTON is a more difficult problem [9,10,2,7].

In this paper, our idea is to leverage the advantages from both virtual try-on domains, i.e., 3D model-based and image-based approaches, and make a hybrid pipeline for the image-based virtual try-on task, which is simple and fully-automatic. Hence, we propose *Clothing Three-dimensional reconstruction for Hybrid image-based Virtual Try-ON* (CloTH-VTON). Since 2D non-rigid deformations suffer due to complex 3D poses [5] and 3D model-based techniques are good at realistic deformation of any poses/styles [8], we propose a novel 3D cloth reconstruction method, from a single in-shop cloth image, using 3D SMPL human body mesh model [6]. Using the SMPL model for reconstruction and deformation of clothes can handle any complex human poses. We also exploit the latest deep networks-based VTON techniques for generating the final try-on results. We use a fusion mechanism for blending 3D warped clothes to 2D human images, which generates the photo-realistic outputs with preserving the original pixel quality (Fig. 1).

Our contributions are as follows:

- We propose a hybrid image-based VTON approach, utilizing the benefits of 2D GAN [13] based methods for synthesizing the dis-occluded parts, and 3D model for the 3D posing of the clothing.
- We introduce a novel 3D clothing reconstruction method from a single in-shop cloth image of any style or category, through 2D image matching and 3D depth reconstruction using body depth.
- We provide a highly effective fine alignment and fusion mechanism for combining the rendered 3D warped clothing with the generated and original 2D human parts.

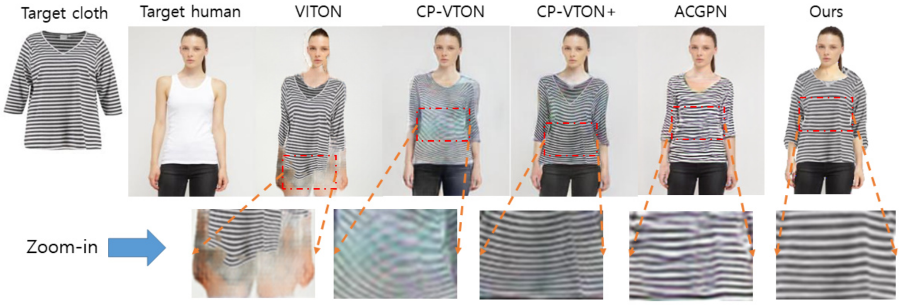

Fig. 2. Limitations from the existing image-based VTON methods. VITON [11] has fusion problems, CP-VTON [4] and CP-VTON+[5] has texture & blur issue, ACGPN [12] produces the best result among SOTA having texture alteration issue, while CloTH-VTON generates output with the highest possible quality including full details.

2 Related Works

2.1 Image-Based Virtual Try-On (VTON)

(Fixed-Pose) Virtual Try-On This task is to transfer in-shop cloth to humans, keeping the pose fixed, same as ours. VITON [11] and CP-VTON [4] propose VTON pipelines with two main stages - clothing warping and try-on image synthesis. Sun et al. [14], VTNFP [15], SieveNet [16] proposed an extra stage for full target human segmentation generation including target clothing. ACGPN [12] proposed a two-stage based target segmentation generation for target human body parts and target clothing mask respectively. Some other works are [8,5,17,18,19,3,20,21]. However, challenges remain, such as self-occlusions, heavy misalignment among different poses, and complex clothes shape or textures (Fig 2).

Multi-Pose Guided Virtual Try-On This task applies a new pose along with the target cloth to the target human, e.g., MG-VTON [22], FIT-ME [23], Zheng et al. [24], FashionOn [25]. They use multi-stage architectures due to the high complexity and large information of features to transfer. Human pose transfer is related to this task except for the target clothing, e.g., ClothFlow [17], Ma et al. [26], Balakrishnan et al. [27] VU-Net [28], and others [29,30,31,32,33,34].

Person to Person Transfer Another popular application of image-based VTON is the person to person clothing transfer. SwapNet [35] proposed a garment exchange method between two human images. Outfit-VITON [36] transfers multiple clothing from different human images to another person. Zanfir et al. [37] proposed appearance transfer between human images using 3D SMPL models [6].

However, most works use conditional GANs where results show the limitations in blurring in dis-occluded area and misalignment of transferred clothing to the target human when there is a big difference between two persons' poses and shapes.

2.2 3D Garment/Human Reconstruction and Dressing

3D Pose and Shape Estimation 3D human pose and shape estimation is one of the most active research areas in 3D human reconstruction. Statistical and parametric human body models, such as SMPL [6] and SMPL-X [38] are accelerating this area rapidly. Frank and Adam [10] models capture markerless motions. To estimate single 3D human pose and shape in an image, SMPLify [39] and SMPLify-X [38] uses optimization techniques, HMR [40] uses learning with 3D supervision, SPIN [41] makes a combination of neural network regression and optimization. OOH [42] estimates 3D humans from object-occluded images, Jiang et al. [43] detect multiple 3D humans from single images, and VIBE [44] estimates multiple 3D humans from videos.

3D Clothed Human Reconstruction Fully-clothed reconstruction of human texture/depth/geometry from image/video/point-cloud is popular due to AR/VR potentials, although not for VTON, e.g., PIFuHD [45], PIFusion [46], IF-Nets [47], PIFU [48], Tex2Shape [49], Photo Wake-Up [50], SiClope [51], 360° textures [52], human depth [53].

3D Garment Reconstruction One major sub-task in our method is to reconstruct 3D cloth models from images. Due to the enormous variety of clothing and fashion, it's highly difficult to reconstruct 3D garment models covering all categories. ClothCap [1] captures cloth models of shirts, pants, jerseys, and skirts from 4D scans of people. Multi-Garment Net [2] makes 3D garment models from 3D scans of people for 3D VTON. They use 3D garment templates for 5 categories: shirt, t-shirt, coat, short-pants, long-pants [2]. Pix2Surf [7] learns to reconstruct 3D clothing from images for 3D VTON, leveraging garment meshes from MGN [2]. Our work is most similar to Minar et al. (2020) [8], where they reconstruct and deform 3D cloth models for image-based VTON. However, they consider 5 clothing categories based on sleeve lengths only, and the final try-on result suffers badly from blurring effects. Some other related works includes Tailornet [54] for predicting realistic 3D clothing wrinkle details, 3D garments from sketches [55], garment animation [56], DeepWrinkles [57].

Despite the high details of 3D clothing models, they mostly require 3D scanning data/templates, fixed categories, and the modeling techniques outside of clothing are still in early stages, which is difficult to apply in VTON task.

3 CloTH-VTON

Figure 3 shows the overall architecture of our proposed CloTH-VTON, which takes a pair of an in-shop cloth image C_{in} and a human image I_{in}, and generates

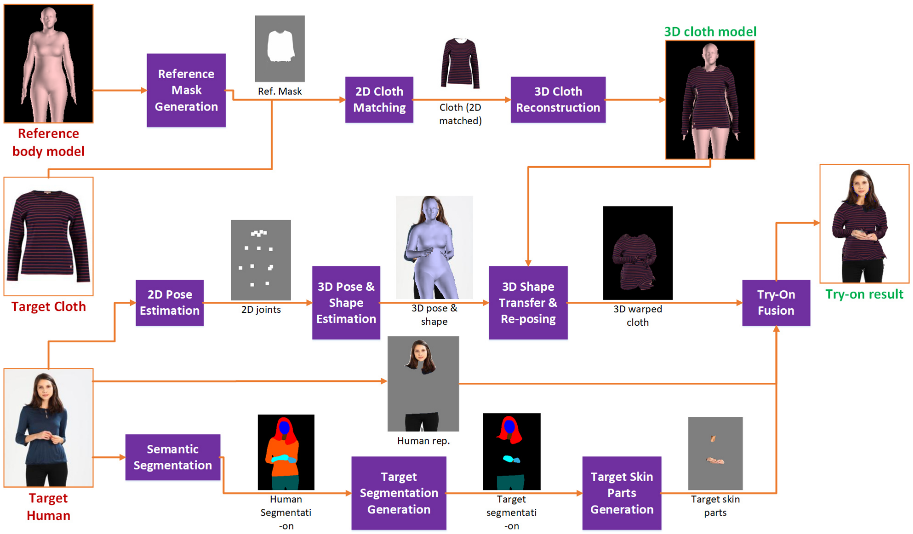

Fig. 3. Overview of our proposed CloTH-VTON. We reconstruct the 3D cloth model of the target cloth by matching silhouettes to the standard body model. Then, we transfer the 3D cloth model to the estimated 3D target human model, to produce the 3D warped cloth. We generate target human skin parts and blend it to the rendered warped cloth along with human representations.

a virtual try-on output I_{out} in the same pose J of I_{in}. First, we generate the target segmentation map S_{body} which guides the following processes for synthesizing the dis-occluded human parts P_{out} and 2D matching mask $M_{out,ref}$ for cloth deformation. Our method reconstructs 3D target cloth model $V_{clothed}$, first through 2D matching between the cloth C_{in} and the matching mask $M_{out,ref}$, then reconstructing $V_{clothed}$ using the standard SMPL [6] model V_{body}. Then, the vertices displacements of $V_{clothed}$ are applied to the estimated 3D target human model $\vec{V}^{t}_{body}$. The non-cloth areas are retained from the original images or synthesized if invisible in the input human image. The final try-on image is blended by a fusion mechanism, retaining the target human properties and high original details of the target cloth. Since the final output is fused by blending masks, not generating using GAN [13] networks, it does not suffer from blurring effects or texture alterations (Fig. 2) which are very common in deep neural network-based image synthesis. Figure 1 shows sample results from our approach.

3.1 Segmentation Generation Network (SGN)

The segmentation layout in the try-on output becomes different from the input human image because of the different cloth shape and occluded/dis-occluded parts of the human. Early works like VITON [11] and CP-VTON [4] do not generate an explicit target segmentation. We use an explicit target segmentation

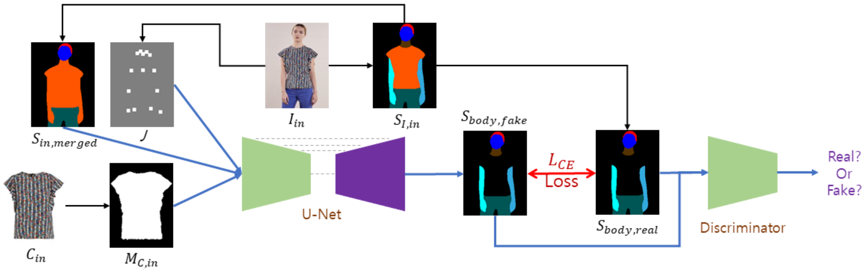

Fig. 4. Segmentation Generation Network (SGN) architecture; to generate the target human body segmentation according to the target cloth.

generation as ACGPN [12], for utilizing in the cloth matching and parts synthesis. We refer to this stage as the Segmentation Generation Network (SGN), presented in Figure 4. SGN has U-Net [58] as the generator and the discriminator from Pix2PixHD [59]. SGN learns to generate the target human segmentation, except for the target clothing area.

SGN takes merged human segmentation $S_{in,merged}$, 2D pose as joints J from the input human I_{in}, and cloth mask $M_{C,in}$ from the input cloth C_{in} as inputs. $S_{in,merged}$ comes from the 2D human segmentation $S_{I,in}$ of I_{in}, where clothes on person and affected human body parts, i.e., top-clothes, torso-skin, right and left arms, are merged to a single upper-cloth label. SGN produces $S_{body,fake}$ as output. We calculate Cross-Entropy loss L_{CE} between $S_{body,fake}$ and $S_{body,real}$, along with loss L_{GAN}, which is the sum of GAN losses and GAN feature matching loss from Pix2PixHD [59]. $S_{body,real}$ is parsed from $S_{I,in}$ as the ground-truth for SGN network.

$$L_{SGN} = \lambda_1 L_{CE} + \lambda_2 L_{GAN} \tag{1}$$

3.2 3D Cloth Reconstruction from In-shop Image

For reconstructing the 3D shape of clothing from in-shop images, we extend the 2D to 3D approach in [8], where they need manual category selection (of 5). We present a fully automatic approach, not restricted to any cloth categories.

Mask Generation Network (MGN) Prior works on 3D garment reconstruction and modeling work with fixed clothing categories [1,2,7,54,8,55] works. They use reference garment templates [1,2,7], trains separate network for predicting 3D models [2,7], or standard body model silhouette [8]. Since our cloth modeling is similar to Minar et al. [8] without requiring 3D garment templates, we also use silhouette masks of a standard A-posed SMPL [6] body model V_{body}, as the reference for 2D silhouette matching between the target cloth and SMPL [6]. Figure 5 shows 5 reference masks generated for 2D matching of clothes.

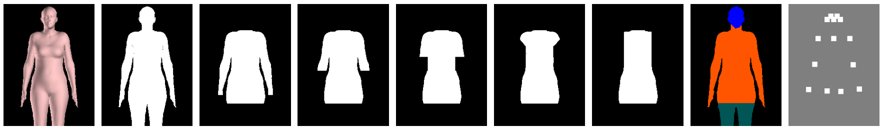

Fig. 5. 2D cloth mapping process to the reference SMPL silhouette. From left to right: standard body model and its silhouette, matching masks for long-sleeve, half-sleeve, short-sleeve (half-elbow and quarter elbow respectively), sleeveless clothing categories, and the standard model inputs for generating matching masks with SGN & MGN.

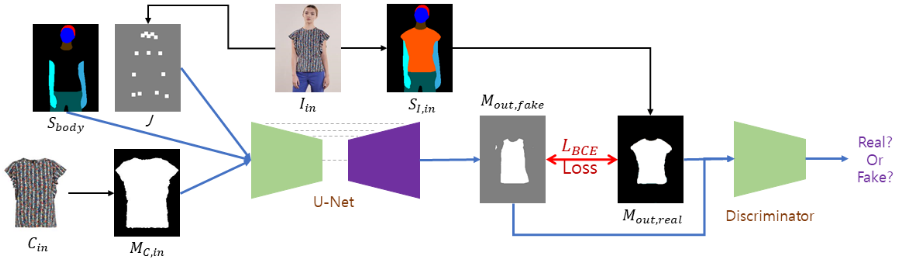

Fig. 6. Mask Generation Network (MGN) architecture; to generate the cloth mask for 2D matching.

To apply cloth matching to other categories, manual matching masks are needed for corresponding clothing categories. While this procedure is easier and generates a tight-fit clothing texture for 3D reconstruction, it is difficult to cover all the clothing categories manually due to diverse clothing styles. Also, making cloth matching masks from SMPL [6] body silhouette does not provide loose-fitting clothing textures. To alleviate this problem and make the process fully-automatic, we use the Mask Generation Network (MGN) for 2D matching, following the semantic generation module of ACGPN [12], since they separated the target cloth mask generation network from segmentation generation. MGN has similar architecture as SGN, as illustrated in Figure 6.

MGN takes the generated body parts segmentation S_{body} from SGN output, human joints J, and in-shop input cloth mask $M_{C,in}$, as the inputs, and produces target cloth mask on person $M_{out,fake}$ as the output. We calculate binary cross-entropy loss L_{BCE} between $M_{out,fake}$ and real clothes mask on person $M_{out,real}$ from $S_{I,in}$, along with the GAN losses L_{GAN} (Eq. 2).

$$L_{MGN} = \lambda_1 L_{BCE} + \lambda_2 L_{GAN} \tag{2}$$

We made a fused segmentation map of our standard body model, and generated 2D joints as in A-pose (see Figure 5), to provide input to SGN, which generates the target body segmentation S_{body} in standard pose. MGN takes S_{body} as an input, and infers the matching masks $M_{out,ref}$ for silhouette matching of target cloth to standard SMPL [6] model V_{body}. Figure 7 shows an example of SGN and MGN inference for V_{body}.

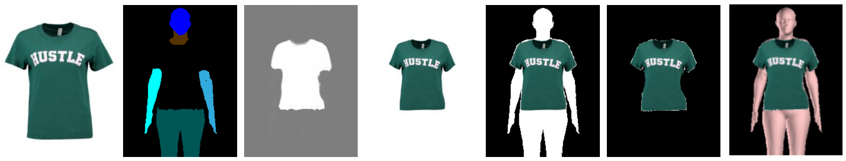

Fig. 7. Example of 2D cloth matching for 3D reconstruction. From left to right: Input cloth image, target body segmentation generated by SGN for standard body model, cloth matching mask generated by MGN, cloth texture after 2D matching and overlapped on SMPL silhouette respectively, 3D reconstructed cloth and overlaid on SMPL model respectively.

2D Clothing Matching We apply Shape-Context Matching (SCM) [60] between the target clothes and their corresponding categorical masks (Fig. 5) or matching masks generated by MGN (Sec. 3.2). Then, we apply Thin-Plate Spline (TPS) [61] transformation on the target clothes, based on the shape correspondences from SCM, to generate the clothing textures to be aligned to the standard SMPL [6] body model V_{body}. Pix2Surf [7] argues that a combination of SCM-TPS may generate holes and artifacts at the clothing boundary. However, since we only use the front images of clothes for image-based VTON, SCM-TPS provides better matching for specific clothes. Figure 7 shows an example of 2D cloth matching and texture extraction for 3D reconstruction.

3D Clothing Model Reconstruction For 3D reconstruction from the aligned clothing image and projected silhouette, first, vertices of the 3D body mesh V_{body} are projected into 2D image space. When boundary vertices are in 2D space, clothing boundaries are used to find the corresponding points. To make the clothing transfer, i.e., change of its pose and shape easily, a 3D clothing model's vertices are mapped to an SMPL [6] body vertices. We assume that the relation between the clothing and human vertices is isotropic, i.e., the difference in the projection space is also retained in the 3D model. Although this is not strictly true, we make this assumption for practical applications. We define the corresponding points in the clothing boundary as the closest points from the projected vertices. We estimate Thin-Plate Spline (TPS) [61] parameters and apply them to the mesh points. New mesh points are considered as the vertices projected from the 3D mesh of clothing $\mathbf{V_{clothed}}$. From 2D points to 3D points are done with inverse projection with depth obtained from the body with a small constant gap. In reality, the gap between the clothing and body cannot be constant but it works with tight or simple clothes.

$$\mathbf{V_{clothed}} = \mathbf{P^{-1}} \cdot \mathbf{T_{TPS}}(\mathbf{P} \cdot \mathbf{V_{body}}), \mathbf{depth}(\mathbf{V_{body}})), \tag{3}$$

Here, $\mathbf{P}$ is the projection matrix with the camera parameters $\mathbf{K} \cdot [\mathbf{R}|\mathbf{t}]$, $\mathbf{P^{-1}}$is the inverse projection matrix of the same camera, and $depth(V_{body})$ is the distance from the camera to the vertices. Target clothing images are used as the textures for the 3D clothing mesh. Finally, we obtain the clothing 3D model

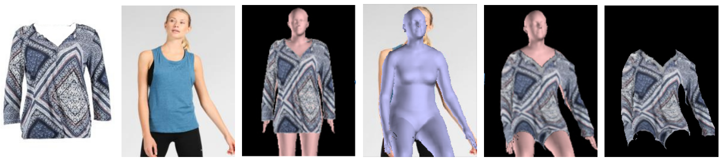

Fig. 8. Sample 3D clothing deformation with 3D human body estimation. From left to right: input cloth, target human, 3D reconstructed cloth model, 3D estimated human body model, 3D deformed cloth model, and the rendered image of 3D warped cloth.

$V_{clothed}$ by selecting the vertices that are projected onto the clothing image area.

Target Human Model Parameter Estimation To estimate $\vec{V}^{t}_{body}$, the SMPL [6] parameters (β, θ) for a human image, we use the SMPLify [39] method. However, any newer and better method can be used since we use estimated parameters only. SMPLify [39] is for full-body images, so we made a few minor optimizations, mainly for our half-body dataset, such as - head pose correction, joints location mapping between the joints of the dataset used in this paper and SMPL's, the SMPLify joints definition, conditional inclusion of invisible joints and initialization step.

3D Clothing Model Deformation 3D clothing model $V_{clothed}$ and texture information obtained from 3D reconstruction is for the standard shaped and posed person (β_0, θ_0). For the VTON application, we have to apply the shape and pose parameters (β, θ) of the target human image, $\vec{V}^{t}_{body}$, estimated from the previous step (Sec. 3.2). Instead of applying the shape and pose parameters to the obtained clothed 3D model, we transfer the displacements of clothing vertices to $\vec{V}^{t}_{body}$, since the application of new parameters to the body model provides much better natural results. Several options can be considered for the transfer, e.g., transferring the physical size of clothing or keep the fit, i.e., keep the displacements from the body to clothing vertices as before. We simply decide the fit-preserving option for showing more natural results for final fitting.

$$\vec{V}^{t}_{clothed} = \vec{V}^{t}_{body} + \vec{d}\, in\, (u_x, u_y, u_z | \vec{V}^{t}_{body}) \tag{4}$$

Hence, we get the 3D deformed model $\vec{V}^{t}_{clothed}$ of the target cloth. Then, we render $\vec{V}^{t}_{clothed}$ to get the warped cloth image C_{warped}, to apply to the final try-on, following Minar et al. [8]. Figure 8 shows an example of applying 3D cloth deformation. Additional details are provided in the supp. mat.

3.3 Try-on Generation

To generate the final try-on output image, its common to utilize the generative neural networks [4,12]. However, due to not having enough training data, we

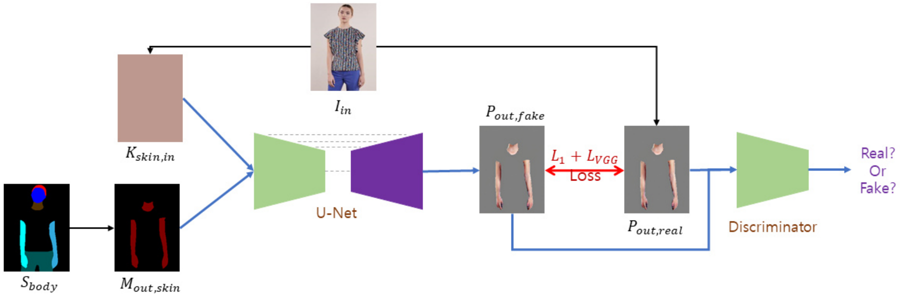

Fig. 9. Parts Generation Network (PGN) architecture; to generate the target skin parts of the input human according to the target cloth.

chose not to train a network to generate the final results directly. Also, results from the generative networks suffer from blurry output issues due to up-sampling e.g. and texture alterations [4,8]. Therefore, we chose to simply merge the warped clothes into the target human, leveraging the target segmentation from SGN. One problem is getting the accurate target clothing affected human skin parts, i.e., torso-skin, right and left arms/hands. For transition cases like long-sleeve cloth to short-sleeve, hidden skin parts cause artifacts [2]. So, we train a separate network for generating the target skin parts, using the S_{body} output from SGN.

Parts Generation Network (PGN) We refer to this stage as the Parts Generation Network (PGN), which uses similar networks as SGN and MGN. PGN pipeline is drawn in Figure 9.

PGN takes 2 inputs: mask of the target skin parts, $M_{out,skin}$ from the target segmentation S_{body} generated by SGN, and average skin color $K_{skin,in}$ of body skin parts i.e., torso-skin, left-arm and right-arm from the target human image I_{in}. PGN generator produces the target body skin parts, $P_{out,fake}$. We calculate L_1 loss, VGG loss L_{VGG}, and GAN losses L_{GAN} from Pix2PixHD [59], between the generator output $P_{out,fake}$ and the real human body skin parts $P_{out,real}$ from $S_{I,in}$ (Eq. 5).

$$L_{PGN} = \lambda_1 L_{L1} + \lambda_2 L_{VGG} + \lambda_3 L_{GAN} \tag{5}$$

Try-on Fusion The final step is to merge all the parts (Eq. 6), i.e., human representation $I_{in,fixed}$ from the target person I_{in}, rendered 3D deformed clothing image C_{warped}, and the generated skin parts P_{out} from PGN, to get the try-on output I_{out}. To make a successful fusion of all these segments, accurate target body segmentation S_{body} and 3D body estimation plays a critical role.

$$I_{out} = I_{in,fixed} + C_{warped} + P_{out} \tag{6}$$

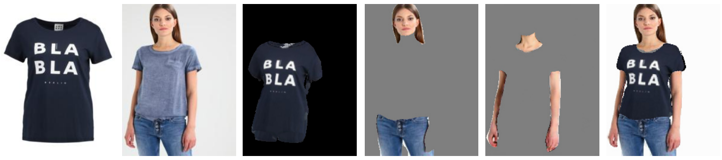

Fig. 10. Sample final fusion for try-on. From left to right: Input cloth image, target human image, rendered 3D warped cloth, human representation from target person, target skin parts generated by PGN, and the final fusion output for virtual try-on.

4 Experiments

4.1 Dataset Preparation

For training and testing, We used the VITON [11] dataset, which contains 14221 training and 2032 testing pairs of in-shop clothes and human images (half-body). VITON resized dataset [4] contains in-shop cloth masks, human poses predicted by OpenPose [62], and human segmentation generated by LIP-SSL [63]. However, as noted by [5], VITON dataset has several problems. LIP [63] segmentation does not have labels for skin in the torso area, i.e., neck, chest, or belly, and labeled those as background. These areas are crucial to estimate the clothing area, so we generated new segmentation with pre-trained CIHP-PGN [64]. Also, many cloth masks are wrong when cloth colors are similar to the background (white). We re-implemented a new mask generator considering the dataset characteristics.

4.2 Implementation Details

Neural Network Training and Testing All three neural networks in our approach, SGN (Sec. 3.1), MGN (Sec. 3.2) and PGN (Sec. 3.3), shares the common network architecture: U-Net [58] as the generators, and the discriminators from Pix2PixHD [59] network. GAN losses include the generator loss, discriminator losses for real and fake outputs, and the feature-matching loss [59]. All networks are implemented in PyTorch [65], based on the public implementation [66] of ACGPN [12], and each network is trained for 20 epochs with a batch size of 8. It takes 17-20 hours of training for each network with 4 TITAN Xp GPUs. For testing, we used two kinds of VITON test input pairs - same-clothes and different-clothes. Same-clothes input pairs are used for evaluating with ground-truth, and different clothes pairs for visual comparison.

Mask Generation for Cloth Matching We use 5 silhouette masks from the reference SMPL [6] model for 5 clothing categories based on sleeve lengths [8], i.e., long-sleeve, half-sleeve, short-sleeve half-elbow, short-sleeve quarter-elbow, and sleeveless. These categories contains a total of 465 + 130 + 780 + 162 + 252 = 1789 in-shop clothing images, out of 2032 VITON [11] test dataset clothes (See Figure 5). For the rest of the clothes, we use the fully-automatic process from

Section 3.2, to generate a specific matching mask for each cloth. Direct inference with SGN and MGN networks gives several unexpected results when we test with the standard A-posed model input. We assume that, since these networks are trained with the full training set, which is full of various poses and different from A-pose, cause the artifacts. It would be best to train simple networks with fixed A-pose data. However, due to the lack of such data and annotations, we choose to go with a closer path. We collected 1095 human images from the VITON dataset, having very simple poses, e.g., straight hands and standing. These are selected based on the Easy criterion from ACGPN [12]. We train a very simple version of SGN and MGN networks with the easy pose pairs, exclusively for generating reference masks for 2D clothing matching. We follow the same training procedures for these networks as discussed in Section 4.2. Then, we generate the cloth-specific silhouette matching mask, using our standard SMPL model inputs, as shown in Figures 5, 6 & 7.

2D Clothing Matching We implemented this step in MATLAB, utilizing the original script of SCM [60]. We chose 10 * 10 control points for describing shape contexts between the silhouette masks, and then apply TPS [61] transformation on the input clothes.

3D Clothing Reconstruction and Re-posing We use the available public models from SMPL [67] and SMPLify [68], and their python implementations for 3D reconstruction and model transfer. Based on the SMPLify [39] implementation, we also make our implementation for this step using Chumpy [69] and OpenDR [70]. First, the standard SMPL [6] model is reconstructed. Then, we transform the model from 2D space to 3D space, according to the cloth texture from 2D matching, to get the shape information of cloth. Pose and shape parameters are estimated from the human image using SMPLify [39] optimization. Finally, the cloth model is transferred to the 3D body model to get the warped cloth.

Clothing and Human Image Fusion For final try-on output, we utilize the generated target body segmentation to fuse the human representations, warped clothes, and the generated skin parts into output images.

4.3 Results

We provide both qualitative and quantitative analyses of our results, comparing with existing image-based VTON methods. For qualitative comparisons, we retrained the networks from the available public implementations of the SOTA approaches and reproduced the results.

Qualitative Analysis We present the qualitative comparisons in Figure 11 among VITON [11], CP-VTON [4], CP-VTON+ [5], ACGPN [12] and CloTH-VTON, for different clothes input pairs. Newer methods such as CP-VTON+ [5]

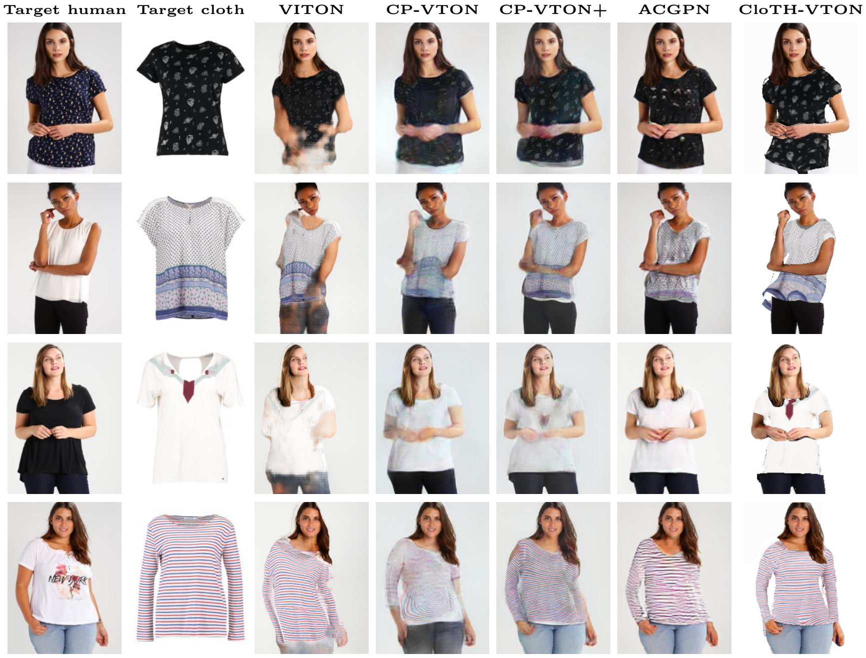

Fig. 11. Visual comparisons between SOTA image-based VTON methods and our CloTH-VTON. For fair comparisons, we present samples from the dataset with complex human poses e.g. cross arms, and target clothes with detailed textures. Our method produces try-on results with the highest details and quality possible.

and ACGPN [12] generates competitive results. However, we can see the clear improvements in our results than the existing methods, especially when the target clothes have detailed textures or the target humans have complex poses. Figure 2 shows the differences in the details of the methods' results. More comparisons & results are provided in the supplementary material.

Quantitative Analysis We present quantitative comparisons in Table 1. We use the Structural Similarity Index (SSIM) [71] and Inception Score (IS) [72], for comparing with and without ground truths respectively. The values of VT-NFP [15] and ACGPN [12] are added from the reported scores in ACGPN [12] paper. The values of CP-VTON+ [5] is added from their paper.

Our SSIM score is slightly lower than CP-VTON+ and ACGPN. SSIM is originally developed for video compression quality measures without geometric distortion, according to a recent study on the limitation of SSIM [73]. We argue that, even though we are comparing with the image-based methods, our method is a hybrid approach from the 3D reconstruction. Therefore, SSIM scores are lower since rendered 3D warped clothes are structurally different from the clothes

Table 1. Quantitative comparison with the SOTA image-based VTON methods: VITON [11], CP-VTON [4], VTNFP [15], ACGPN [12], and CP-VTON+ [5].

Metric	VITON	CP-VTON	VTNFP	CP-VTON+	ACGPN	CloTH-VTON
SSIM	0.783	0.745	0.803	0.816	**0.845**	0.813
IS	2.650	2.757	2.784	3.105	2.829	**3.111**

in images. However, human synthesizing metric like inception score provides the highest score for our results, which proves the capability of our approach.

4.4 Discussion

From the results, it is clear that our method is highly competitive against the SOTA image-based VTON approaches. However, there are many rooms for improvement. Such as - target human body segmentation generation, matching mask generation for 2D clothing matching, 3D shape and pose estimation of the human body, realistic clothing deformation, and final fusion. Target human segmentation plays a crucial role in almost all stages, making it one of the most important areas for improvement. The next performance bottleneck is the 2D silhouette matching for transferring the clothing textures to the standard 3D model. Hence, it is important to generate accurate silhouette matching masks for input clothes. 3D body estimation from human images can be done with any SOTA 3D human pose and shape estimation methods, e.g. SPIN [41] or OOH [42]. Also, our current approach is mostly applicable to close-fitting clothes. For reconstruction and deformation of loose-fitting clothes, e.g. dress and skirt, separate clothing deformation techniques like TailorNet [54] can be applied. Final fusion output will be far better based on the improvements in previous stages.

5 Conclusion

We propose a hybrid approach for image-based virtual try-on tasks, combining the benefits of 2D image-based GAN[13] and 3D SMPL [6] model-based cloth manipulation. We present a 2D to 3D cloth model reconstruction method using only a 3D body model, applicable to diverse clothing without requiring 3D garment templates. To integrate from two different domains, we develop target semantic segmentation and clothing-affected body parts generation networks. Our final try-on output provides the photo-realistic results which come with great details, high resolution, and quality.

Acknowledgement

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2018R1D1A1B07043879).

References

1. Pons-Moll, G., Pujades, S., Hu, S., Black, M.: Clothcap: Seamless 4d clothing capture and retargeting. ACM Transactions on Graphics, (Proc. SIGGRAPH) **36** (2017) 2, 4, 6
2. Bhatnagar, B.L., Tiwari, G., Theobalt, C., Pons-Moll, G.: Multi-garment net: Learning to dress 3d people from images. In: IEEE International Conference on Computer Vision (ICCV), IEEE (2019) 2, 4, 6, 10
3. Song, D., Li, T., Mao, Z., Liu, A.A.: Sp-viton: shape-preserving image-based virtual try-on network. Multimedia Tools and Applications (2019) 1–13 2, 3
4. Wang, B., Zheng, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: The European Conference on Computer Vision (ECCV). (2018) 2, 3, 5, 9, 10, 11, 12, 14
5. Minar, M.R., Tuan, T.T., Ahn, H., Rosin, P., Lai, Y.K.: Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (2020) 2, 3, 11, 12, 13, 14
6. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: A skinned multi-person linear model. ACM transactions on graphics (TOG) **34** (2015) 1–16 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 14
7. Mir, A., Alldieck, T., Pons-Moll, G.: Learning to transfer texture from clothing images to 3d humans. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2020) 2, 4, 6, 8
8. Minar, M.R., Tuan, T.T., Ahn, H., Rosin, P., Lai, Y.K.: 3d reconstruction of clothes using a human body model and its application to image-based virtual try-on. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. (2020) 2, 3, 4, 6, 9, 10, 11
9. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36** (2017) 2
10. Joo, H., Simon, T., Sheikh, Y.: Total capture: A 3d deformation model for tracking faces, hands, and bodies. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2018) 8320–8329 2, 4
11. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 3, 5, 11, 12, 14
12. Yang, H., Zhang, R., Guo, X., Liu, W., Zuo, W., Luo, P.: Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2020) 3, 6, 7, 9, 11, 12, 13, 14
13. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A.C., Bengio, Y.: Generative adversarial nets. In: NIPS. (2014) 2, 5, 14
14. Sun, F., Guo, J., Su, Z., Gao, C.: Image-based virtual try-on network with structural coherence. In: 2019 IEEE International Conference on Image Processing (ICIP). (2019) 519–523 3
15. Yu, R., Wang, X., Xie, X.: Vtnfp: An image-based virtual try-on network with body and clothing feature preservation. In: The IEEE International Conference on Computer Vision (ICCV). (2019) 3, 13, 14

16. Jandial, S., Chopra, A., Ayush, K., Hemani, M., Krishnamurthy, B., Halwai, A.: Sievenet: A unified framework for robust image-based virtual try-on. In: The IEEE Winter Conference on Applications of Computer Vision (WACV). (2020) 3
17. Han, X., Hu, X., Huang, W., Scott, M.R.: Clothflow: A flow-based model for clothed person generation. In: The IEEE International Conference on Computer Vision (ICCV). (2019) 3
18. Jae Lee, H., Lee, R., Kang, M., Cho, M., Park, G.: La-viton: A network for looking-attractive virtual try-on. In: The IEEE International Conference on Computer Vision (ICCV) Workshops. (2019) 3
19. Kubo, S., Iwasawa, Y., Suzuki, M., Matsuo, Y.: Uvton: Uv mapping to consider the 3d structure of a human in image-based virtual try-on network. In: The IEEE International Conference on Computer Vision (ICCV) Workshops. (2019) 3
20. Ayush, K., Jandial, S., Chopra, A., Krishnamurthy, B.: Powering virtual try-on via auxiliary human segmentation learning. In: The IEEE International Conference on Computer Vision (ICCV) Workshops. (2019) 3
21. Yildirim, G., Jetchev, N., Vollgraf, R., Bergmann, U.: Generating high-resolution fashion model images wearing custom outfits. In: The IEEE International Conference on Computer Vision (ICCV) Workshops. (2019) 3
22. Dong, H., Liang, X., Shen, X., Wang, B., Lai, H., Zhu, J., Hu, Z., Yin, J.: Towards multi-pose guided virtual try-on network. In: The IEEE International Conference on Computer Vision (ICCV). (2019) 3
23. Hsieh, C.W., Chen, C.Y., Chou, C.L., Shuai, H.H., Cheng, W.H.: Fit-me: Image-based virtual try-on with arbitrary poses. In: 2019 IEEE International Conference on Image Processing (ICIP), IEEE (2019) 4694–4698 3
24. Zheng, N., Song, X., Chen, Z., Hu, L., Cao, D., Nie, L.: Virtually trying on new clothing with arbitrary poses. In: Proceedings of the 27th ACM International Conference on Multimedia. (2019) 266–274 3
25. Hsieh, C.W., Chen, C.Y., Chou, C.L., Shuai, H.H., Liu, J., Cheng, W.H.: Fashionon: Semantic-guided image-based virtual try-on with detailed human and clothing information. In: Proceedings of the 27th ACM International Conference on Multimedia. (2019) 275–283 3
26. Ma, L., Jia, X., Sun, Q., Schiele, B., Tuytelaars, T., Van Gool, L.: Pose guided person image generation. In: Advances in neural information processing systems. (2017) 406–416 3
27. Balakrishnan, G., Zhao, A., Dalca, A.V., Durand, F., Guttag, J.: Synthesizing images of humans in unseen poses. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 3
28. Esser, P., Sutter, E., Ommer, B.: A variational u-net for conditional appearance and shape generation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 3
29. Ma, L., Sun, Q., Georgoulis, S., Van Gool, L., Schiele, B., Fritz, M.: Disentangled person image generation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 3
30. Siarohin, A., Sangineto, E., Lathuilière, S., Sebe, N.: Deformable gans for pose-based human image generation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2018) 3
31. Qian, X., Fu, Y., Xiang, T., Wang, W., Qiu, J., Wu, Y., Jiang, Y.G., Xue, X.: Pose-normalized image generation for person re-identification. In: The European Conference on Computer Vision (ECCV). (2018) 3

32. Dong, H., Liang, X., Gong, K., Lai, H., Zhu, J., Yin, J.: Soft-gated warping-gan for pose-guided person image synthesis. In: Advances in Neural Information Processing Systems 31. Curran Associates, Inc. (2018) 474–484 3
33. Zhu, Z., Huang, T., Shi, B., Yu, M., Wang, B., Bai, X.: Progressive pose attention transfer for person image generation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 3
34. Song, S., Zhang, W., Liu, J., Mei, T.: Unsupervised person image generation with semantic parsing transformation. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2019) 3
35. Raj, A., Sangkloy, P., Chang, H., Lu, J., Ceylan, D., Hays, J.: Swapnet: Garment transfer in single view images. In: The European Conference on Computer Vision (ECCV). (2018) 3
36. Neuberger, A., Borenstein, E., Hilleli, B., Oks, E., Alpert, S.: Image based virtual try-on network from unpaired data. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2020) 3
37. Zanfir, M., Popa, A.I., Zanfir, A., Sminchisescu, C.: Human appearance transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 5391–5399 3
38. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3d hands, face, and body from a single image. In: Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR). (2019) 4
39. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P., Romero, J., Black, M.J.: Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In: European Conference on Computer Vision, Springer (2016) 561–578 4, 9, 12
40. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 7122–7131 4
41. Kolotouros, N., Pavlakos, G., Black, M.J., Daniilidis, K.: Learning to reconstruct 3d human pose and shape via model-fitting in the loop. In: Proceedings of the IEEE International Conference on Computer Vision. (2019) 2252–2261 4, 14
42. Zhang, T., Huang, B., Wang, Y.: Object-occluded human shape and pose estimation from a single color image. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). (2020) 4, 14
43. Jiang, W., Kolotouros, N., Pavlakos, G., Zhou, X., Daniilidis, K.: Coherent reconstruction of multiple humans from a single image. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 5579–5588 4
44. Kocabas, M., Athanasiou, N., Black, M.J.: Vibe: Video inference for human body pose and shape estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 5253–5263 4
45. Saito, S., Simon, T., Saragih, J., Joo, H.: 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. (2020) 84–93 4
46. Li, Z., Yu, T., Pan, C., Zheng, Z., Liu, Y.: Robust 3d self-portraits in seconds. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 1344–1353 4
47. Chibane, J., Alldieck, T., Pons-Moll, G.: Implicit functions in feature space for 3d shape reconstruction and completion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. (2020) 6970–6981 4

48. Saito, S., Huang, Z., Natsume, R., Morishima, S., Kanazawa, A., Li, H.: Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. 2019 IEEE/CVF International Conference on Computer Vision (ICCV) (2019) 2304–2314 4
49. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2shape: Detailed full human body geometry from a single image. In: IEEE International Conference on Computer Vision (ICCV). (2019) 4
50. Weng, C.Y., Curless, B., Kemelmacher-Shlizerman, I.: Photo wake-up: 3d character animation from a single photo. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 5908–5917 4
51. Natsume, R., Saito, S., Huang, Z., Chen, W., Ma, C., Li, H., Morishima, S.: Siclope: Silhouette-based clothed people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2019) 4480–4490 4
52. Lazova, V., Insafutdinov, E., Pons-Moll, G.: 360-degree textures of people in clothing from a single image. In: 2019 International Conference on 3D Vision (3DV), IEEE (2019) 643–653 4
53. Tang, S., Tan, F., Cheng, K., Li, Z., Zhu, S., Tan, P.: A neural network for detailed human depth estimation from a single image. In: The IEEE International Conference on Computer Vision (ICCV). (2019) 4
54. Patel, C., Liao, Z., Pons-Moll, G.: Tailornet: Predicting clothing in 3d as a function of human pose, shape and garment style. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE (2020) 4, 6, 14
55. Wang, T.Y., Ceylan, D., Popovic, J., Mitra, N.J.: Learning a shared shape space for multimodal garment design. ACM Trans. Graph. **37** (2018) 1:1–1:14 4, 6
56. Wang, Y., Shao, T., Fu, K., Mitra, N.: Learning an intrinsic garment space for interactive authoring of garment animation. ACM Trans. Graph. **38** (2019) 4
57. Lahner, Z., Cremers, D., Tung, T.: Deepwrinkles: Accurate and realistic clothing modeling. In: The European Conference on Computer Vision (ECCV). (2018) 4
58. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In Navab, N., Hornegger, J., III, W.M.W., Frangi, A.F., eds.: Medical Image Computing and Computer-Assisted Intervention - MICCAI 2015 - 18th International Conference Munich, Germany, October 5 - 9, 2015, Proceedings, Part III. Volume 9351., Springer (2015) 234–241 6, 11
59. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2018) 6, 10, 11
60. Belongie, S., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Transactions on Pattern Analysis and Machine Intelligence **24** (2002) 509–522 8, 12
61. Bookstein, F.L.: Principal warps: thin-plate splines and the decomposition of deformations. IEEE Transactions on Pattern Analysis and Machine Intelligence **11** (1989) 567–585 8, 12
62. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: CVPR. (2017) 11
63. Gong, K., Liang, X., Zhang, D., Shen, X., Lin, L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR). (2017) 11
64. Gong, K., Liang, X., Li, Y., Chen, Y., Yang, M., Lin, L.: Instance-level human parsing via part grouping network. In: The European Conference on Computer Vision (ECCV). (2018) 11

65. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library. In: Advances in Neural Information Processing Systems 32. Curran Associates, Inc. (2019) 8026–8037 11
66. : Acgpn. (https://github.com/switchablenorms/DeepFashion_Try_On) 11
67. : Smpl. (https://smpl.is.tue.mpg.de/) 12
68. : Smplify. (http://smplify.is.tue.mpg.de/) 12
69. : Chumpy. (https://github.com/mattloper/chumpy) 12
70. Loper, M.M., Black, M.J.: Opendr: An approximate differentiable renderer. In Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T., eds.: Computer Vision – ECCV 2014, Cham, Springer International Publishing (2014) 154–169 12
71. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13** (2004) 600–612 13
72. Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., Chen, X.: Improved techniques for training gans. In: Advances in Neural Information Processing Systems 29. Curran Associates, Inc. (2016) 2234–2242 13
73. Nilsson, J., .A.M.T.: Understanding ssim. In: arXiv 2006.13846. (2020) 13