3D Reconstruction of Clothes using a Human Body Model and its Application to Image-based Virtual Try-On

Matiur Rahman Minar[1], Thai Thanh Tuan[1], Heejune Ahn[1], Paul L. Rosin[2], and Yu-Kun Lai[2]

[1]Department of Electrical and Information Engineering, Seoul National University of Science and Technology, South Korea
[2]School of Computer Science and Informatics, Cardiff University, UK

Abstract

Image-based virtual try-on (VTON) approaches are getting attention since they do not require 3D modeling. However, 2D cloth warping algorithms cannot cover 3D spatial transformations for diverse target human poses. To solve this problem, we propose a 2D and 3D hybrid method. First, a 3D clothing mesh is reconstructed leveraging a 3D human body model in a rest pose. Due to the correspondence, resulting 3D clothing models can be easily transferred to the target human models with different poses and shapes estimated from 2D images. Finally, the deformed clothing models can be rendered and blended with target human representations. Experimental results with an open dataset show that shapes of reconstructed clothing are more natural, compared to the 2D image-based deformation results, when human poses and shapes are estimated accurately.

1. Introduction

Due to the difficulty and high costs in obtaining 3D models of humans and clothing for 3D virtual try-on (VTON), 2D image-based VTON is getting more popular [4, 9]. However, a close look at the results of image-based VTON reveals that their seemingly high qualities are thanks to 1) simple datasets, i.e., mostly short-sleeved, single-colored clothes and simple human poses, and 2) blending algorithms that hide and mitigate the misalignment and warping distortion of the try-on clothing for simple style clothes. Figure 2 shows several significant problems in image-based VTON. Among them, especially the misalignment of the warped clothing with the target human is considered the critical and inherent limitation of 2D transforms, even for non-rigid deformations like the thin plate spline (TPS) algorithm. However, reconstructing a 3D clothing model from a 2D image is not a trivial problem either. 3D clothed human model reconstruction studies [10, 7, 8] confirm the limitations and challenges.

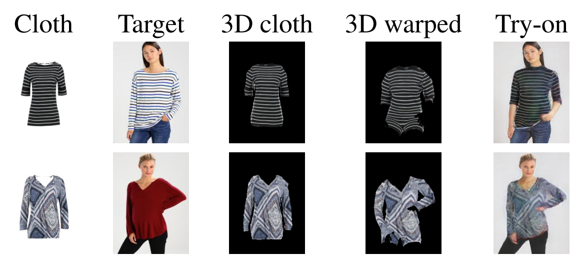

Figure 1. Sample results of our proposed pipeline. Left to right: input try-on clothes, target human images, reconstructed 3D clothes (shape and pose transferred respectively) and final blended results.

In this paper, we leverage the rest-posed 3D body model to facilitate the reconstruction of the 3D clothing model and apply this to an image-based VTON application. We consider the setting of using input pairs of try-on clothing and target humans. We propose an approach for reconstructing the 3D model of try-on clothing, using a 3D template body model. This template-based approach makes the 2D and 3D clothing reconstruction process easier, also making clothing deformation more natural. The clothing image is aligned with the rendered silhouette of the 3D body model and reconstructed using the depth and mesh information of the 3D body model. The warped clothing images are obtained through the pose and shape transfer to the target human parameters and rendering. The final VTON image is obtained by blending the rendered warped clothing image with target human representation.

2. Related Work

2.1. 3D Human model and reconstruction

We use the Skinned Multi-Person Linear (SMPL) [5] model for 3D reconstruction, since SMPL has well-defined

Figure 2. Failure cases of image-based Virtual Try-On algorithms. Left to right: try-on clothing and target human images, VITON (warped clothes and VTON results), CP-VTON (warped clothes and VTON results)

control variables for shape and pose, and parameter estimation algorithms. SMPL is a skinned vertex-based statistical model that accurately represents a wide variety of body shapes in natural human poses [5]. SMPLify [2] estimates SMPL parameters and performs 3D body shape reconstruction from a single image, by optimizing 2D human body joint information. [7, 8, 10] are examples of 3D clothed human model reconstruction.

2.2. Virtual Try-on

VITON [4] and CP-VTON [9] both presented image-based virtual try-on approaches, where they transfer try-on clothes to a target person by using a warping module, and later a blending module. VITON directly computes the TPS [3] transformation mapping using the shape context descriptor [1]. CP-VTON introduces a learning method to estimate the transformation parameters. However, a TPS transform cannot warp the try-on clothing to a target human with a 3D pose, as in the examples of hands-up and folded arms demonstrated in Figure 2.

3. 3D Model Reconstruction of Clothing from an Image

Figure 3 illustrates our proposed end-to-end VTON processing steps. Our pipeline is composed of 3D clothing model reconstruction and clothing model transfer and blending stages. The upper path runs when the new clothing images are entered into the online system and the bottom path runs when the customer uploads her/his photo and chooses a try-on clothing. The 3D clothing model reconstruction stage is composed of (1) 2D matching and alignment between a try-on clothing image and a SMPL [5] silhouette, and (2) 3D clothing mesh model reconstruction from the 2D input clothing image though an SMPL template body model.

The transfer and blending path is done through (3) estimating the 3D body model (SMPL pose and shape parameters) from a 2D target human image, (4) transferring the 3D clothing model's information, and (5) blending the rendered warped clothing image into the target human image. Figure 1 shows sample results of our approach, and Figure 5 shows the visual flow of 3D SMPL model reconstruction and deformation.

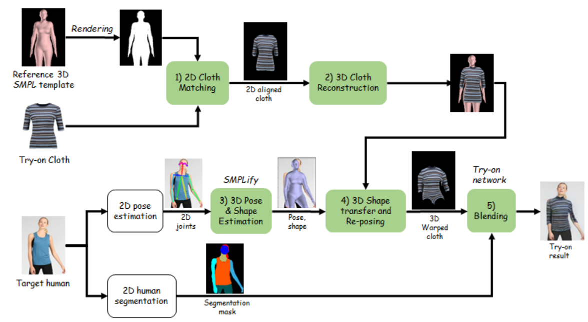

Figure 3. The proposed pipeline: We reconstruct a 3D clothing model from a single 2D clothing image. Then, given a target human image, we transfer the reconstructed 3D clothing model properties (shape and pose) to the target human image and render the transferred 3D clothing onto a 2D image as the warped clothing. Finally, we blend the rendered clothing with the target human image, generating a virtual try-on image.

3.1. Clothing matching to a 2D standard silhouette

First, we render a standard (i.e. a fixed pose and shape parameter (β_0, θ_0)) SMPL [5] model into a silhouette image. The pose and shape are chosen so that no self-occlusion of the silhouette can occur. (For simplicity's sake, we chose a single shape and pose parameter (A-pose), but it can be varied for better or easier matching. Please see Figure 4 for example). For high accuracy matching, we categorize the clothes into long sleeve, sleeveless, short sleeve elbow, short sleeve half elbow, and short sleeve quarter elbow (Please see details in the supplementary material). According to each category, the rendered silhouette is clipped so that the matching algorithm, SCM (Shape Context Matching) [1], can extract the matching information easily. The clipping parameters are decided manually in the current version. We can consider an automatic parameter estimation method in the future. Then, we apply SCM between clothing input and the processed template silhouette as in Figure 4. Finally, we apply a thin-plate-spline (TPS) [3] transformation $T_{SMPL}(\cdot)$ on the input clothing image I_c and the corresponding mask image M_c and generate the 2D matched clothing $I_{c,warped}$ and mask $M_{c,warped}$ images.

$$(I_{c,warped}, M_{c,warped}) = T_{SMPL(\beta_0,\theta_0)}(I_c, M_c) \quad (1)$$

3.2. 3D clothing model reconstruction

The 3D reconstruction process from the aligned clothing image and projected silhouette consists of 2 steps. First, vertices of the 3D body mesh are projected into 2D image

Figure 4. 2D matching between a clothing image and the 3D body model silhouette mask.

space, boundary vertices are in 2D space, and the clothing boundaries are used to find corresponding points. To facilitate the clothing transfer, i.e., change of its pose and shape, a 3D clothing model's vertex is mapped to an SMPL [5] body vertex. We assume that the relation between the clothing and human vertex is isotropic, i.e., the difference in the projection space is also retained in the 3D model.

The corresponding points in the clothing boundary are defined as the closest points from the projected vertices. Thin Plate Spline (TPS) [3] parameters are estimated and applied to the mesh points. The new mesh points are considered as the vertices projected from the 3D mesh of clothing $\mathbf{v_{clothed}}$. Mapping 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 for tight or simple clothes. Further research works are needed for accurate depth estimation.

$$\mathbf{v_{clothed}} = \mathbf{P^{-1}} \cdot \mathbf{T_{TPS}}(\mathbf{P} \cdot \mathbf{v_{body}}), \mathbf{depth}(\mathbf{v_{body}})), \tag{2}$$

where $\mathbf{P}$ 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 distance from camera to the vertex. Try-on clothing images are used as the texture for the 3D clothing mesh. Finally, the clothing 3D model is obtained by selecting the vertices that are projected onto the clothing image area.

3.3. Target human model parameter estimation

To estimate the SMPL [5] parameters (β, θ) for a human image, we use the SMPLify [2] method. However, any newer and better method can be used since we assume nothing on the procedure and use estimated parameters only. Since the original work is for full-body images, we made a few minor optimizations for the half body dataset, such as joint location mapping between the joints of the VITON [4] data set used and the SMPLify joint definition, and conditional inclusion of invisible joints and the initialization step.

3.4. Transfer of 3D clothing model to the target human

3D model and texture information obtained from 3D reconstruction is for the standard shape and posed person (β_0, θ_0). To apply this information for the virtual try-on application, we have to apply the shape and pose parameters (β, θ) of the target human image estimated with the SMPLify [2] step. Instead of applying the shape and pose parameters to the obtained clothed 3D model, we transfer the displacement of clothing vertices to the target human body model, since the application of new parameters to the body model provides much better natural results.

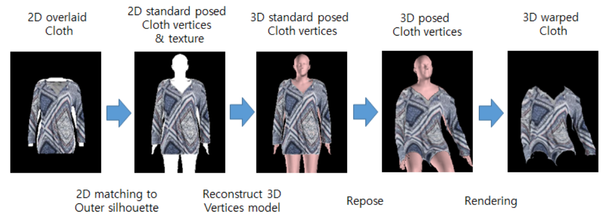

Figure 5. Visual flow of our method: from 2D matching of clothing to 3D reconstruction and clothing transfer.

3.5. Blending of warped clothing with target human image

For our experiment, we use an extended version of the try-on module (TOM) from CP-VTON. We update three things from the original try-on module (TOM) of CP-VTON for our implementation. Firstly, we include the un-intended body and clothing areas into the person representation input to the try-on module (TOM) network. Secondly, we include the warped clothing mask into the network inputs, so that the network can differentiate the white clothing area from the background. Thirdly, we update the composite mask loss function. In the mask loss term in the try-on module loss function, we replace the composition mask with supervised ground truth mask for a strong alpha mask.

4. Experiments and Results

4.1. Implementation Details

The VITON clothing-human pair dataset [4] is used as the training and test dataset, which provides the 2D joint estimation, human parsing maps of target human images, and binary masks for clothing images. However, examining the test dataset, we chose 1789 clothes out of 2032 for 3D reconstruction (see supplementary material) for clothing categories and number of their images). We filtered out the rest due to being side views or falling in other categories.

For implementation, we used the publicly available SMPL [5] models[1] and SMPLify SMPLify [2] implementation[2], which is based on OpenDR [6] that in turn uses Chumpy[3]. We used the MATLAB implementation of SCM

[1]https://smpl.is.tue.mpg.de/
[2]http://smplify.is.tue.mpg.de/
[3]https://github.com/mattloper/chumpy

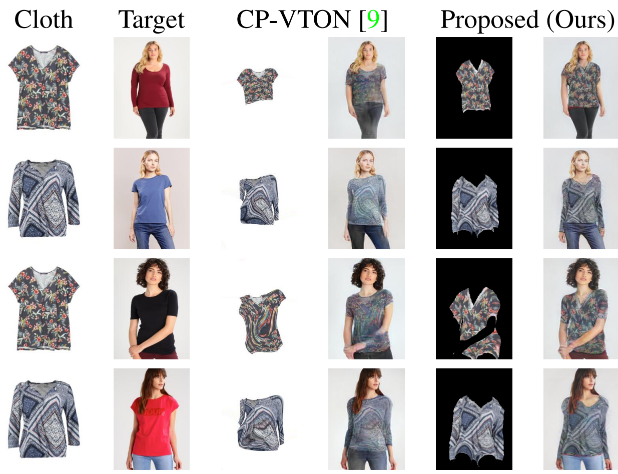

Figure 6. Qualitative comparison between the baseline CP-VTON [9] and our approach. For each row, the first pair of images are the inputs: try-on clothing and target human. The second pair of images are the warped clothing and final try-on results of CP-VTON [9]. And last pair includes the results of our approach: 3D reconstructed-deformed clothing, and the try-on results from the blending network.

matching and TPS deformation for 2D clothing and template matching stage.

For the blending stage, we use an extended TOM (Try-On Module) network of the CP-VTON implementation described above. Our 3D model based warped clothing processing corresponds to GMM (Geometric Matching Module) in CP-VTON. We train our updated try-on module with the dataset collected by Han et al. [4], and follow the same training procedures as CP-VTON. Please refer to their original work for details.

4.2. Results

We present 3D cloth reconstruction results in Figure 1 and 6. Figure 6 shows visual comparison results between the state-of-the-art image-based virtual try-on model CP-VTON [9] and our approach. As discussed in Section 1, CP-VTON generates better try-on outputs when there are small spatial deformations between try-on clothes and target human images, and try-on clothes are mono-colored. However, for warping try-on clothes with large spatial transformations and preserving clothing characteristics realistically, 3D model- based deformations are far better than 2D image-based non-rigid transformation algorithms. Hence, we present the examples of the results from our approach, where the try-on clothes have detailed textures, and/or the target humans have big poses. Our proposed approach can realistically deform try-on clothes while preserving better clothing characteristics than 2D image-based approaches, and also generates better final virtual try-on output.

5. Conclusion

We proposed 3D clothing model reconstruction method from a single clothing image, leveraging the correspondence between the clothing and human body shapes. Then, we transfer the 3D clothing model to the target human model, using SMPL [5] human body pose and shape parameters. Finally, we render the transferred 3D clothing model and integrate it with target human image contents. Our 3D reconstruction approach can provide clothing deformation of diverse human images where existing image-based VTON methods fail. However, the virtual try-on image quality is still not high enough. In particular, the inaccuracy in estimating human pose and shape makes the integrated VTON results appear less natural. Therefore, we can consider in the future either improving the 3D model estimation algorithm, or using a different blending step that is better suited to the 3D warped clothing input.

References

[1] Serge Belongie, Jitendra Malik, and Jan Puzicha. Shape matching and object recognition using shape contexts. *IEEE TPAMI*, 24(4):509–522, 2002. 2

[2] Federica Bogo, Angjoo Kanazawa, Christoph Lassner, Peter Gehler, Javier Romero, and Michael J Black. Keep it smpl: Automatic estimation of 3d human pose and shape from a single image. In *ECCV*, pages 561–578, 2016. 2, 3

[3] Fred L. Bookstein. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE TPAMI*, 11:567–585, 1989. 2, 3

[4] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *CVPR*, pages 7543–7552, 2018. 1, 2, 3, 4

[5] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. Smpl: a skinned multi-person linear model. *ACM TOG*, 34:248:1–248:16, 2015. 1, 2, 3, 4

[6] Matthew M Loper and Michael J Black. Opendr: An approximate differentiable renderer. In *ECCV*, pages 154–169, 2014. 3

[7] Ryota Natsume, Shunsuke Saito, Zeng Huang, Weikai Chen, Chongyang Ma, Hao Li, and Shigeo Morishima. Siclope: Silhouette-based clothed people. In *CVPR*, pages 4480–4490, 2019. 1, 2

[8] 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 *ICCV*, pages 2304–2314, 2019. 1, 2

[9] Bochao Wang, Hongwei Zhang, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018. 1, 2, 4

[10] Chung-Yi Weng, Brian Curless, and Ira Kemelmacher-Shlizerman. Photo wake-up: 3d character animation from a single photo. In *CVPR*, pages 5908–5917, 2019. 1, 2