Supplementary Material
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]

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

In this supplementary material, we provide 1) mathematical details for pose and shape transfer of a standard 3D clothing model to the target person, 2) additional comparison results on the VITON [1] dataset in comparison with the state-of-the-art methods, and 3) extra experiment results with in-the-wild images.

1 Transfer of 3D Clothing Model to the Target Human

After reconstructing the 3D clothing models from the standard human model, we get the 3D deformed clothes, by transferring the vertices' displacements of the clothing model to the corresponding target human body model (See Figure 1).

The displacement of a cloth vertex from a human body vertex is expressed in the local coordinate frame at the corresponding human body vertex as shown in Figure 2. The local coordinate frame at a human body vertex (both source and target body models) is defined as follows: the surface normal vector as z-axis u_z, the vector to the smallest indexed neighborhood vertex as x-axis u_x, and their cross product vector as y-axis u_y.

$$\vec{u}_z = \mathbf{n}(\tilde{\mathbf{v}}_{\mathbf{body}}) \tag{1}$$

$$\vec{u}_x = \vec{u}_x^{''}/|\vec{u}_x^{''}|,\ \vec{u}_x^{''} = \vec{u}_x^{'} - \vec{u}_x^{'} \cdot \vec{u}_z,\ \vec{u}_x^{'} = (\vec{v}_{\min(N_v)} - \vec{v}_{body}) \tag{2}$$

$$\vec{u}_y = \vec{u}_z \times \vec{u}_x, \tag{3}$$

where N_v is the set of neighboring vertices of $\vec{v}$.

The displacement of cloth vertex form the corresponding source body vertex is calculated in the local coordinates and then transferred to the target body surfaces to get the transferred position of the clothing vertex.

$$\vec{d} = (d_x, d_y, d_z) = \vec{v}_{clothed} - \vec{v}_{body}\ in\ (u_x, u_y, u_z|\vec{v}_{body}) \tag{4}$$

$$\vec{v}^t_{clothed} = \vec{v}^t_{body} + \vec{d}\ in\ (u_x, u_y, u_z|\vec{v}^t_{body}) \tag{5}$$

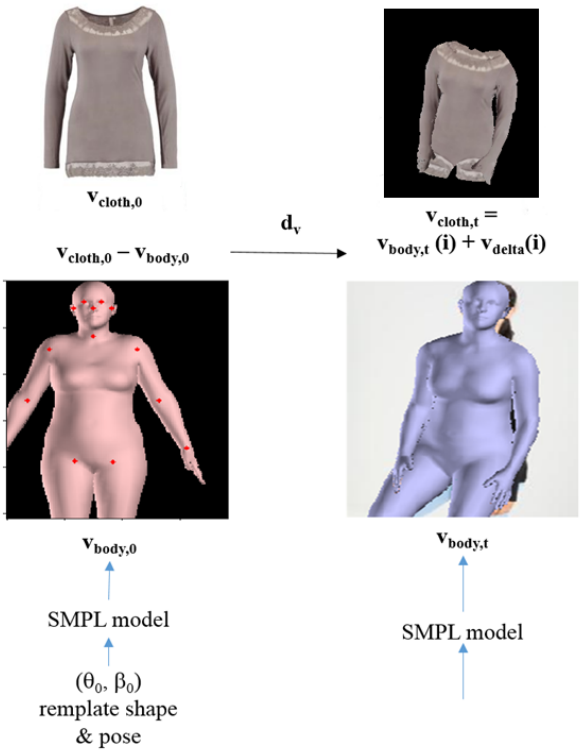

Fig. 1. Clothing shape and pose transfer method: the differences between the corresponding clothing and body in the reference model are added to the target body vertex positions.

2 More Try on Results with VITON Dataset

We present additional experimental results for comparison with state-of-the-art (SOTA) methods and detailed qualitative analyses.

2.1 2D Cloth Matching for 3D Cloth Reconstruction

The 2D clothing silhouette matching to the cloth mask on the standard SMPL [2] body model is critical to the quality of 3D clothing reconstruction. We apply the Shape-Context Matching (SCM) [3] between the in-shop input cloth masks and the reference matching masks according to the standard body model. In the categorical approach [4], one matching mask is specified manually per clothing category. In comparison with the category-based semi-automatic approach from [4], this paper proposed an automatic approach that generates the matching masks using the Mask Generation Network (MGN) automatically and for a more wide range of clothing categories. Examining the visual results from these

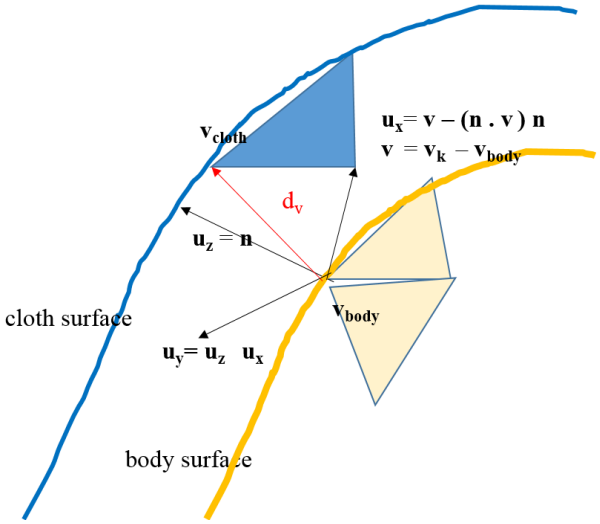

Fig. 2. The local coordinate frame definition for vertex displacement representation.

two approaches in Figure 3, categorical cloth matching & reconstruction produces similar or slightly better results due to the silhouette masks made directly from the standard body model. Also, the back neck parts of input clothes are manually segmented and removed for better visualization in categorical 2D clothing matching. However, automatic mask generation for cloth matching provides cloth-specific silhouettes, which is more applicable and can cover diverse fashion clothing. Also, the automatic approach will be much better if the exclusive SGN and MGN networks for generating matching masks are trained with standard posed data.

2.2 Try-on Results

Try-on with Step-wise Results In Figure 4, We present the step-wise results for generating the final try-on output from our method. First, we do 2D clothing matching of the input in-shop clothes with the standard SMPL [2] body model. Then, we reconstruct the 3D cloth models from the 2D matched clothes, and get the 3D warped clothes by transferring the cloth models to the corresponding 3D body models. We also produce the target segmentation maps of the target humans from the Segmentation Generation Network (SGN), according to the target clothes. Based on the generated segmentation, we generate the target skin body parts, using the Parts Generation Network (PGN). Finally, we produce the final try-on results by fusion of the target human representation, rendered 3D warped clothes, and the target skin body parts, according to the generated target human segmentation.

We show the step-by-step output in Figure 4. Input to our method is a pair of images, i.e., target in-shop cloth image and target human image. 2D matched

and 3D warped clothes are produced from the in-shop cloth, according to the target human. And the target human body semantic segmentation and the skin parts are generated from the input human, according to the target cloth. Finally, the try-on output is produced from the results of the consecutive steps.

Try-on Results Comparison We provide an additional qualitative comparison in Figure 5, among state-of-the-art (SOTA) image-based virtual try-on (VTON) approaches and our proposed CloTH-VTON. Our method produces the try-on output with the highest details and quality.

In addition to the results presented in the paper, we show more samples from VITON [1], CP-VTON [5], CP-VTON+ [6], ACGPN [7], and our CloTH-VTON, reproduced on the VITON [1] test dataset. From the results in Figure 5, VITON [1] loses many details from the human representations and the non-target areas. CP-VTON [5] fails to preserve clothing shape and texture details, and generates blurry results. CP-VTON+ [6] solves the limitations of CP-VTON, but it also generates blurry output. ACGPN [7] generates photo-realistic output with high resolution and details but fails to preserve clothing textures accurately, especially when the target cloth has detailed textures. Lastly, our proposed CloTH-VTON produces the try-on output with the highest resolution and the natural details of the target cloth. Since our method does not use deep generative networks to produce the final output directly, it does not suffer from any blurry effects. Hence, our method can generate photo-realistic results with the highest quality possible.

3 Results on In-the-wild Images

To prove the generalization ability of the proposed CloTH-VTON method, we provide extra results on in-the-wild images collected from the Internet for testing. We test our 3D cloth reconstruction method with in-the-wild in-shop cloth images, as shown in Figure 6. We collected the images mostly with detailed textures from the popular TV/comic series. Reconstructed clothes are displayed as an overlay on top of standard SMPL silhouette, which proves that our 3D reconstruction method works for any in-shop clothes and can be applied to target human images for try-on.

We also apply the proposed try-on method to celebrity images, collected from the Internet with the 3D reconstructed clothes from Figure 6. Figure 7 shows that our method is capable of generating realistic virtual try-on output while retaining the details with any random background.

References

1. 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) 1, 4, 8

2. 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
3. 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 2
4. 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
5. 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) 4, 8
6. 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) 4, 8
7. 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) 4, 8

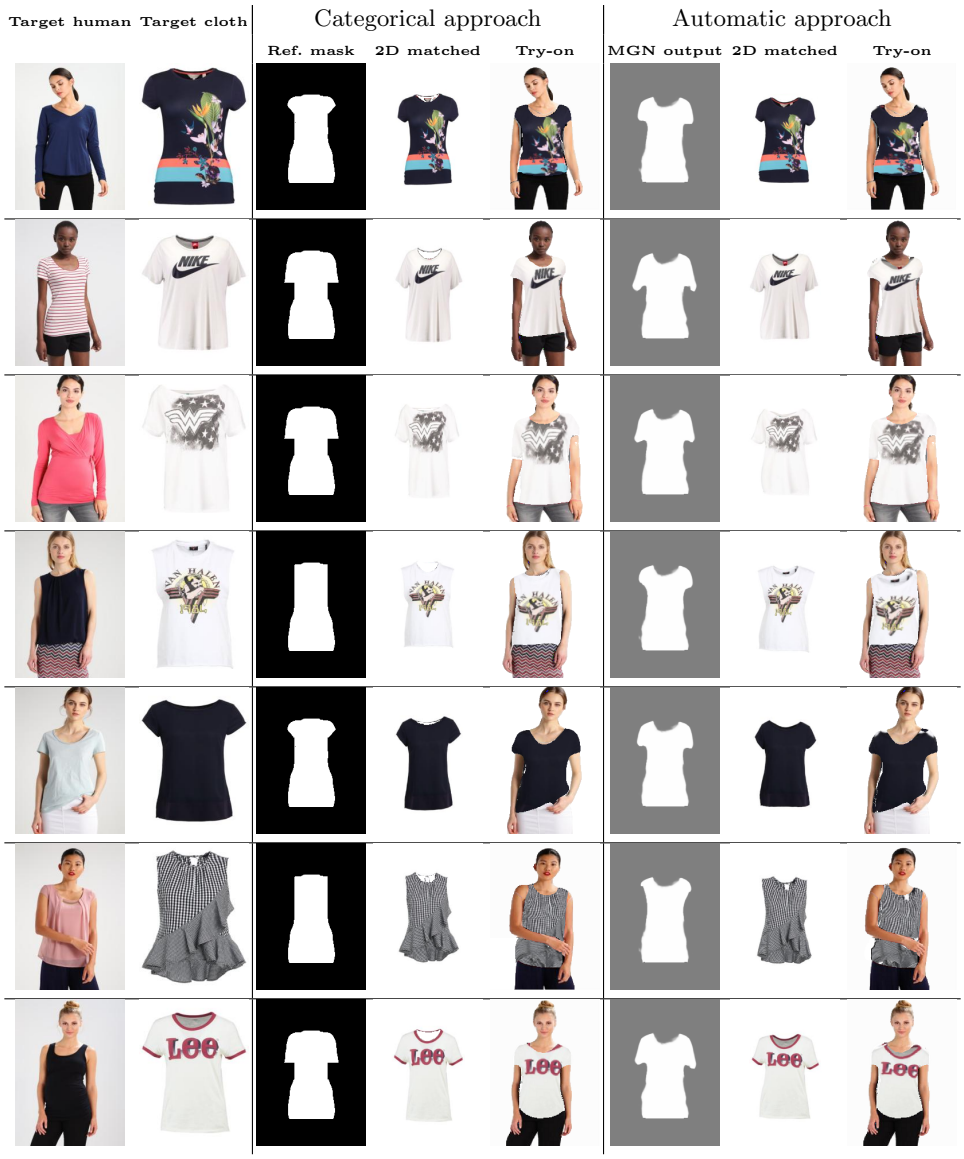

Fig. 3. Examples of 2D cloth matching by generating masks from the categorical and the automatic approaches for comparison. Cloth matching and reconstruction using the categorical approach produces better output since the reference masks are taken from the standard model silhouette. The automatic approach provides the freedom of clothing variety and auto-generation of the matching masks.

Fig. 4. Sample results of our CloTH-VTON, showing step-wise results from each stage (See Section 2.2). The first two columns are the inputs, then the corresponding results from the consecutive steps, i.e., 2D cloth matching, 3D reconstruction and deformation, Segmentation Generation Network (SGN), Parts Generation Network (PGN), and final fusion.

Fig. 5. Additional visual comparison among VITON [1], CP-VTON [5], CP-VTON+ [6], ACGPN [7] and CloTH-VTON. VITON [1] has blending issues. CP-VTON [5] loses texture details and gets blurry. CP-VTON+ [6] also generates blurry results. ACGPN [7] produces the best results among SOTA methods with better resolution and blending, with clothing texture alterations. Our method preserves the full texture details and generates the photo-realistic images with the highest quality.

Fig. 6. 3D cloth reconstruction testing on in-the-wild images. The top row shows collected random in-shop clothing images from the internet. The middle row shows the overlaid 2D clothing after silhouette matching with the standard A-posed SMPL model. And the bottom row shows the 3D reconstructed clothing models.

Fig. 7. Try-on Results with in-the-wild images, with the clothes reconstructed from Figure 6. From left to right: celebrity images as the target humans, two consecutive examples of the target clothes, and their corresponding try-on results.