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

Anonymous ECCV submission

Paper ID ...

Abstract. Image-based virtual try-on (VTON) has drawn increasing attraction for online apparel shopping, mainly because of not requiring 3D information of try-on clothes and target humans. However, the existing 2D algorithms, even utilizing the advanced non-rigid deformation algorithm, can not handle the 3D shape changes for the postures of target humans. In this study, we propose the 3D cloth reconstruction method using 3D human body model. The 3D model of try-on cloth can be more easily deformed when applied to the rest posed standards human model. Thereafter the pose and shape of cloth can be transferred to the ones of the target humans estimated from an 2D image. Finally the deformed cloth model can be rendered and blended together with unchanged cloth and human parts. The experimental results with a open dataset shows the reconstructed cloth shapes are significantly more natural compared to the 2D imaged based deformation results, when the human pose and shape are estimated accurately.

Keywords: 3D Model, SMPL, 3D Reconstruction, Cloth deformation, Virtual Try-on

1 Introduction

Online fashion market has been growing rapidly every year. Unlike electronics, which makes it easy to standardize functions and performances, fashion apparels have infinite variations in style, forms, colors, texture, and materials. Also the difference between personal preferences is huge. As a result, clothing purchasing decisions are very difficult to make with current non-customized information, like the cloth and models' try fit images. Therefore, virtual try-on (VTON) is a highly demanding technology for the online shopping.

The early VTON technologies were based on 3D computer graphics technology that uses 3D models for a target human and clothing, which are usually expensive and difficult to obtain. Therefore, recently 2D image-based VTON technologies are being studied in academia and industry, fueled by the recent advances in computer vision technologies based on deep learning (DL).

There have been many assumptions in problem settings from the general conditional human image generation related to VTON application. We consider

the one with a try-on cloth and target human image is a practical condition which is assumed in many papers VITON[5], CP-VTON[10] , and the the following [9, 12]. Therefore we also consider the VTON problem that use the try-on cloth and human images and generated a new virtual image that the target human replaced the current top or bottom cloth with the try-on cloth. In this paper we limit our application to top cloth only due to the restricted data set but consider the bottom, e.g. pants cases would be easier than top cloth cases.

The existing image based algorithms seemingly generate high quality VTON images, but our classified analysis on the cloth styles, and human poses and shapes reveals significant problems[ ]. One reason of the seemingly high quality in the existing algorithms are mainly due to the low complexity of the dataset, i.e., most clothes are short-sleeved, and mono-colored, and the poses of humans are mild. Specifically the results with the long-sleeved cloth arm and body posed shows far low quality of the presented result in their papers. We identified 5 issues in CP-VTON[10] algorithms, some of which are tackled in the following papers. Firstly, the target try-on area is dependent upon current cloth shape. Especially, the neck area pixels are labeled as background and some body areas are occluded by hairs or accessories (Fig. 3 (a) left), which affects in cloth warping and blending. Secondly, all the unintended part, faces, bottom-clothes and legs have to be preserved in blending stage. But other parts except face and hair are missing in CP-VTON[10] human representation and generated at blending stage, which is all right for general synthesis application but not desirable in VTON application (Fig. 3 (b) left). Thirdly, the texture is often not vivid, which is due to the composition. Examining the original loss function of TON network, the term for the composition alpha mask are poorly formulated as simple regularization loss.

$$L = c_1|I_0 - I_{GT}| + c_2 L_{VGG} + c_3|1 - M_0| \tag{1}$$

Fourthly, since no label in the area of warped cloth is the same color as background, white colored clothes are confused and improperly processed in the blending stage (Fig. 3 (c)) Finally, GMM module using Spatial Transform Network[6] with TPS (Thin Plate Spline)[3] deformation cannot handle strong 3D deformation due to the target pose and also generates artifacts because of the person representation inputs. For example, hands-up and folded arms. Note that many errors in the warping stage are often hidden in the blending stage when the target clothes are single-colored, which can be expected in practical conditions (Fig. 3 (d)).

In this paper, we focus on the last but most difficult problems that cannot be solved in pure 2D image-based algorithm. The 3D cloth deformation is inherently difficult for 2D warping method, including non-rigid one, like TPS algorithm, we propose to first reconstruct 3D model of try-on cloth, then apply the pose and shape transfer for the target human, and finally blending with unchanged image contents like the face, bottom cloth, and background. Therefore, one of main the tasks now is to reconstruct 3D cloth model from 2D try-on cloth image. The 3D cloth model reconstruction have been studied in previous studies [ ] but still

needs significant improvement for general condition. Our key idea in this step is that once we can control the human pose and shape to become similar to the try-on cloth's, the 3D reconstruction process can be made much easier and the reconstruction quality would be much higher than general pose and shape condition.

So in the Section 3, we describe the 3D cloth reconstruction algorithm. we divide the reconstruction step into 2D matching of cloth to the standard body silhouette and 3D reconstruction of cloth. The later 3D reconstruction step is done through the SMPLify[2] algorithm for the SMPL 3D body model[8]. In Section 4, the blending method described, where the 3D cloth model are transferred to the target human images, through SMPL body parameters of shapes and poses. Then the transferred 3D is rendered and blended to the target human image. In this step we reused the 2D VTON blending algorithm with the modification for the condition. The sampled results from dataset are presented in Section 5 and the paper is concluded in Section 6. In addition to our main study, we added the classified quality evaluation of the previous 2D image based VTON algorithms for the completeness of the paper.

2 Classified Image Based Performance Evaluation

2.1 Image-based VTON

In this Section, we started with evaluating the 2D image based VTON algorithms. We considered CP-VTON[10] published in 2018 as the benchmark algorithm. The previous and following in 2019 share same input image and information conditions with CP-VTON and compare the results with it. Here we include the SCM based-VTON, VITON[5], and CP-VTON[10], however, we believe the performance strength and weakness are similar in the other algorithms too.

The Image based VTON algorithms are mostly composed of two stages: (1) cloth warping step that warps the try-on cloth to align with the pose and shape of the target model (called GMM in CP-VTON: geometric Manipulation Module)[10], and (2) blending step that blends the warped cloth onto the target human image (called TON in CP-VTON: Try-On Network)[10]. CP-VTON assumes the target human image is pre-processed for a cloth agnostic human representation by a human pose estimation like OpenPose[4] and human parsing like LIP[7]. The human representation is composed of 1) heat maps for each joints 2) silhouette of human body, and 3) face and skin pixels patches (non-cloth and human identity area). We use the same dataset collected by Han et al. used in VITON[5] and CP-VTON[10] papers.

2.2 Classified quality

Even though the success and failure cases are presented and compared with other algorithms' results, the failure case analysis is not enough for understanding the origin of failure cases and therefore difficult to find the solution for them. A

classified evaluation would be better for this understanding. Here we summarized the classified results from our another study. We classify input try-on cloth and target human images according to the posture and body type of the person, the degree of occlusion of the clothes, and the characteristics of the clothes. Quality is compared in IoU for the warping step and in SSIM for the final blending step for same cloth re-try-on cases. We also tested for the new cloth try-on cases but did not include here for limitation of spaces, and the same cloth cases are enough to explain the tendencies of the performances. Though in general CP-VTON generates the best quality image, the relative comparison is not the main purpose of the analysis. Please refer to Figure 1 for detailed results.

////

Here I will describe the evaluation discussion, finally commenting that the GMM has serious problem.

////

Especially note that the warped cloth are often too much different for desired shape. It is originated two facts. First the 3D deformation that any 2D deformation including non-rigid transform such as TPS is quite limited, especially any 2D deformation cannot handle when the two area in the original image are overlapped in the destination images. There for when the arms of long sleeved cloth occlude the main body, 2D warping cannot approximate the 3D deformation properly. Second, the deformation needs corresponding points between the source nd target image. The cloth are extremely difficult object to find the corresponding points. The STN (spatial transform network)[6] and SCM (shape context matching)[1] cannot find the corresponding points when the target cloth and original cloth has different shapes. In conclusion, the 2D image based algorithm has serious limitation in the range of applications. It can apply to the mild posed target human only and simple short sleeved cloth, mainly because the inherent limitation of 2D deformation method including non-rigid ones, and the poor performance of matching algorithm. To overcome this limitation, we consider to model the try-on cloth into 3D model and apply the 3D deformation

3 3D model reconstruction of cloth

3.1 Overview

For 3D human body model, we use Skinned Multi-Person Linear model (SMPL)[8], because SMPL has well defined control variable for shape and pose and also well defined parameter estimation algorithms. For similar reasons, SMPL have been utilized in many research works. Furthermore because it is based on blend skinning, SMPL is compatible with existing rendering engines and we make it available for research purposes. SMPL is a skinned vertex-based model that accurately represents a wide variety of body shapes in natural human poses. The parameters of the model are learned from data including the rest pose template, blend weights, pose-dependent blend shapes, identity-dependent blend shapes, and a regressor from vertices to joint locations. Unlike previous models, the pose-dependent blend shapes are a linear function of the elements of the pose rotation

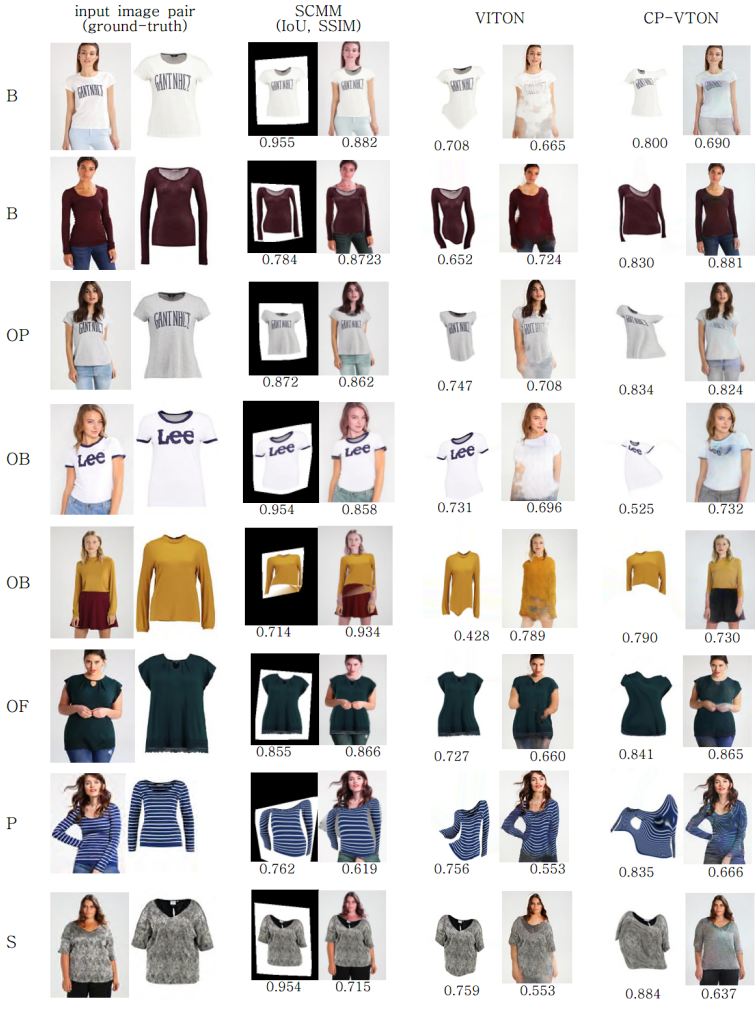

Fig. 1. Classified VTON result: same clothes

matrices. This simple formulation enables training the entire model from a relatively large number of aligned 3D meshes of different people in different poses. [8]

For estimating the SMPL parameters, we use SMPLify[2] method in this study. However any other methods can be used because we assume nothing on the procedure and use estimated parameters only. SMPLify uses 2D human body joint information often obtained from deep learning based method like DeepCut or OpenPose, and minimize the projected joint locations and the given (considered true) 2D joint locations. The cost function can include other priors and silhouette information. We made minor optimization for half body dataset, such as joint location mapping between the joints of used fashion data set and SMPLify joint definition, and conditional inclusion of invisible joints and initialization step. From our experiments with all 2032 test images, we found that the SMPLify quality should be much improved for fully automatic application to VTON application. So the result included in this paper excluded the bad matching cases which is around 30% of all test images.

Clothed human reconstruction using SMPL have been studied in several previous works [11, 13]. Even we are successful in modeling human body, there are further difficulty to recover the clothed human model from body model. It is because the cloth vertices are not directly corresponds to the human body's, and even though it has it is still difficult to estimate the difference between two. Also the texture of cloth can be occluded by other part of cloth and human body parts. The previous work try to solve the problem in the given image condition. Therefore the results are strongly dependent upon the input image. In this paper, we make this step easy using simple standard human pose, where all frontal part of cloth is well separated and visible. This setup cannot handle all problem in the clothed human model reconstruction but can greatly make it easy. The following subsection describe the procedure in details.

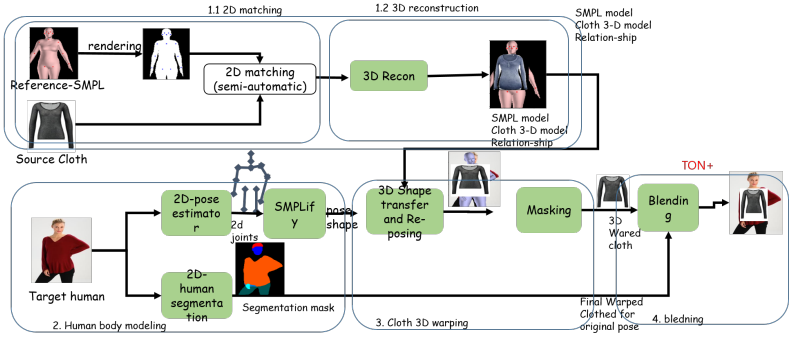

Fig. 2. Pipeline

3.2 2D Standard Cloth matching

To align the try-on cloth image with 3D SMPL body model[8], first their dimension spaces should be matched. Natural way would be first rendering the SMPL body model into 2D image space. However, again the matching with cloth image and body silhouette is not a simple task, for simplicity we assume we can segment the silhouette so that the the remaining area can be easily matched by SCM algorithms. We argue that this step can be monitored by service provider which is practically acceptable; the manual operation from the customer in the try-on step would be not acceptable in general service environment.

$$(I_{c,warped}, M_{c,warped}) = T_{SMPL}((I_c, M_c)) \tag{2}$$

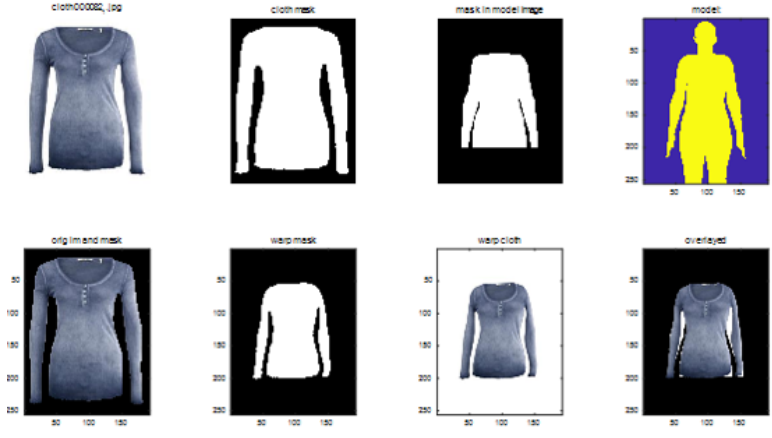

Fig. 3. 2D Matching

3.3 3D cloth model reconstruction

The 3D reconstruction process from aligned cloth image and projected silhouette consists of 2 steps. First, the vertices of 3D body mesh are projected into 2D image space, the boundary vertices in 2D spaces and the cloth boundaries are used for corresponding points. The corresponding points in the cloth boundary i defined the closest points from the projected vertices. This step works well in our cases differently from PhotoWakeup study, because the part of body and cloth are not self-overlapped. This is a implementation benefits of our approach. From the corresponding point pairs, a TPS paramter are estimated and applied to the mesh points. The new mesh points are considered the vertices projected from 3D mesh of cloth.

From the 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 cloth and body cannot be constant but it works with tight or simple clothes. Further research should be needed for accurate depth estimation.

$$V_{clothed} = Pjt^{-1}(T((Pjt(V_{body})), depth(V_{body})) \tag{3}$$

The try-on cloth images are used for the texture for the 3D cloth mesh. We can filter the vertices corresponding to cloth and get the cloth 3D mesh model. Figure xxxxx shows the reconstructed cloth examples.

Discussions needed

Fig. 4. 3D reconstructed cloth

4 Transfer of 3D cloth model to the target Human and Virtual Try-On

4.1 Transfer of 3D cloth model to the target Human

The 3D model and texture information obtained above are for the standard shape and posed person. To apply this information to the target human image, we have to apply the shape and pose parameters of estimated from SMPLify[2] step. In stead of apply the shape and pose parameters to the obtained clothed 3D model, we transferred the displacement of cloth vertices to the target human body model, because the application of new parameters to the Body model provide much natural results.

Multiple option can be considered for the transferring. We could transfer the physical size of cloth or keep the fit, i.e., keep the displacement from the body to cloth vertex as before. We simply decide the Fit-preserving option for showing more natural results for final fitting.

Technically the displacement should be calculated locally. First we calculated the local coordinate at each vertices. We defines the local coordinates: surface normal vector as z -axis, and the vector to smallest indexed edge as x-axis, and their cross product vector as y-axis as the following equations.

$$u_z = normal(V_{body}) \tag{4}$$

$$u_x = u_x^{''}/|u_x^{''}|, u_x^{''} = u_x^{'} - u_x^{'} \cdot u_z, u_x^{'} = (V_{argmin(N_V)} - V_{body}) \tag{5}$$

$$u_y = u_z \otimes u_x, \tag{6}$$

where N_V is the neighbor vertex of V.

The displacement is expressed in the local coordinates and then used the same way in the new target body surfaces for location transfer

$$\overrightarrow{d} = (d_x, d_y, d_z) = V_{clothed} - V_{body}\ in\ (u_x, u_y, u_z|V_{body}) \tag{7}$$

$$V'_{clothed} = V'_{body} + \overrightarrow{d}\ in\ (u_x, u_y, u_z|V'_{body}) \tag{8}$$

4.2 Blending of warped cloth with target human image

This part is under implementation.

We first tried to use TOM. But we found when the reconstruction is not perfect the blending is not natural

Other option is first reconstruct all the clothed information from the target user and overlay the transferred cloth.

5 Conclusions

In this paper, we proposed 3D cloth model reconstruction method using single cloth image. Leveraging the 3D body model, we can make it easy to reconstruct 3D shape information. The 3D cloth model is used for transferring the cloth to target human model. The transferred clothes can be integrated with the human image contents for realizing the pose and shape changes which can not be realizable by existing image based VTON methods.

However, the algorithms in each step of the pipeline are not perfect and have many things to improve at present. Especially the in-accuracy in estimating human pose and shape makes the integrated VTON results not natural enough. Therefore we can consider to improve the SMPLify[2] algorithm or use different blending step that suits for the 3D model input.

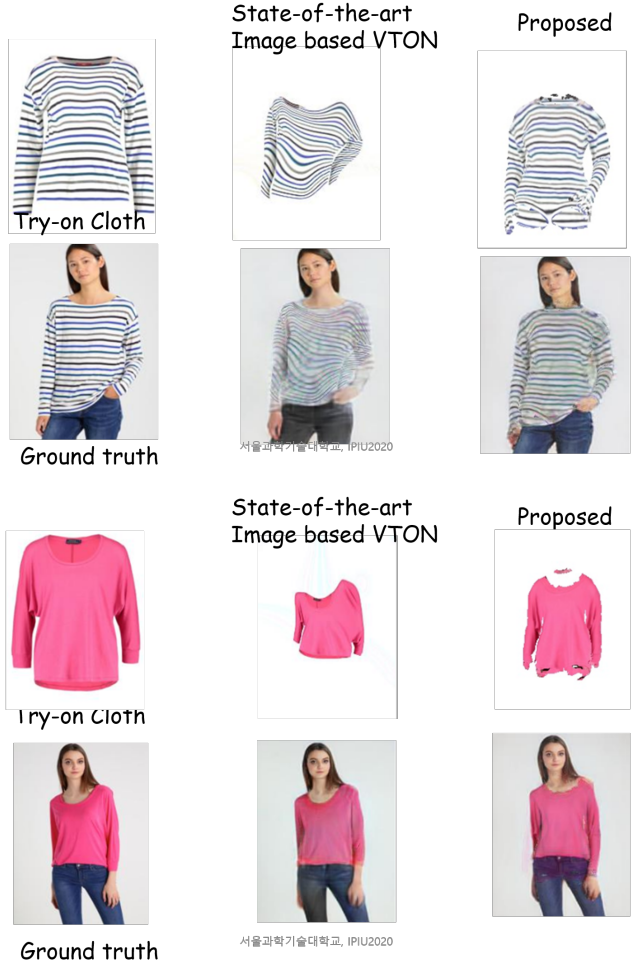

Fig. 5. VTON results

References

1. Belongie, S.J., Malik, J., Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell. **24**(4), 509–522 (2002). https://doi.org/10.1109/34.993558, https://doi.org/10.1109/34.993558
2. Bogo, F., Kanazawa, A., Lassner, C., Gehler, P.V., Romero, J., Black, M.J.: Keep it SMPL: automatic estimation of 3d human pose and shape from a single image. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V. Lecture Notes in Computer Science, vol. 9909, pp. 561–578. Springer (2016). https://doi.org/10.1007/978-3-319-46454-1_34, https://doi.org/10.1007/978-3-319-46454-1_34
3. Bookstein, F.L.: Principal warps: Thin-plate splines and the decomposition of deformations. IEEE Trans. Pattern Anal. Mach. Intell. **11**, 567–585 (1989)
4. Cao, Z., Martinez, G.H., Simon, T., Wei, S.E., Sheikh, Y.: Openpose: Realtime multi-person 2d pose estimation using part affinity fields. IEEE transactions on pattern analysis and machine intelligence (2018)
5. Han, X., Wu, Z., Wu, Z., Yu, R., Davis, L.S.: Viton: An image-based virtual try-on network. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 7543–7552 (2017)
6. Jaderberg, M., Simonyan, K., Zisserman, A., Kavukcuoglu, K.: Spatial transformer networks. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada. pp. 2017–2025 (2015), http://papers.nips.cc/paper/5854-spatial-transformer-networks
7. Liang, X., Gong, K., Shen, X., Lin, L.: Look into person: Joint body parsing and pose estimation network and a new benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**, 871–885 (2018)
8. Loper, M., Mahmood, N., Romero, J., Pons-Moll, G., Black, M.J.: Smpl: a skinned multi-person linear model. ACM Trans. Graph. **34**, 248:1–248:16 (2015)
9. Sun, F.D., Guo, J., Su, Z., Gao, C.Y.: Image-based virtual try-on network with structural coherence. 2019 IEEE International Conference on Image Processing (ICIP) pp. 519–523 (2019)
10. Wang, B., Zhang, H., Liang, X., Chen, Y., Lin, L., Yang, M.: Toward characteristic-preserving image-based virtual try-on network. In: ECCV (2018)
11. Weng, C.Y., Curless, B., Kemelmacher-Shlizerman, I.: Photo wake-up: 3d character animation from a single photo. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) pp. 5901–5910 (2018)
12. 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) (October 2019)
13. Zanfir, M., Popa, A.I., Zanfir, A., Sminchisescu, C.: Human appearance transfer. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 5391–5399 (2018)