Monocular-to-3D Virtual Try-On using Deep Residual U-Net

Hasib Zunair, ID: 40126681
COMP 6381 Digitial Geometric Modeling Project Paper, Fall 2021
Concordia University
hasibzunair@gmail.com

Abstract

3D virtual try-on aims to synthetically fit a target clothing image onto a 3D human shape while preserving realistic details such as pose, identity of the person. Existing methods heavily depend on annotated 3D shapes and garment templates which limits their practical use. While 2D virtual try-on is another alternative, it ignores the 3D body information and cannot fully represent the human body. Recently, M3D-VTON was proposed to generate textured 3D try-on meshes only from 2D images of person and clothing by formulating the 3D try-on problem as 2D try-on and depth estimation. However, we find that the synthesis model in the M3D-VTON pipeline uses a simple U-Net architecture. We hypothesize that this is insufficient to synthesize body parts and model complex relation between front and back parts of clothing only from the 2D clothing image, ultimately leading to unrealistic 3D try-on results. We improve this by implementing residual units in the existing synthesis model. Studying it's effect demonstrates that it improves 2D try-on outputs, mainly by differentiating between front and back part of clothing, preserving logo of clothing and reducing artifacts. This ultimately results in better textured 3D try-on mesh. Benchmarking our method on the MPV3D dataset shows that it performs better than previous works significantly. Code is available at https://hasibzunair.github.io/resm3dvton/.

1. Introduction

3D virtual try-on aims to synthetically fit a target clothing image onto a 3D human shape while preserving realistic details such as pose and identity of the person. Existing methods heavily depend on annotated 3D shapes and garment templates which limits their practical use. While 2D virtual try-on is another alternative, it is highly challenging because it involves several tasks such as cloth warping, image segmentation, image compositing, and image synthesis. It ignores the 3D body information and cannot fully represent the human body. M3D-VTON [5] was recently proposed to generate textured 3D try-on meshes only from 2D images of person and clothing by formulating the 3D try-on problems as 2D try-on and depth estimation.

However, we find that the synthesis model in the M3D-VTON pipeline uses a simple U-Net architecture. We hypothesize that this is insufficient to synthesize body parts and differentiate between front and back parts of clothing only from the 2D image. And this would ultimately lead to unrealistic outputs affecting the final 3D try-on result. We aim to improve this by implementing residual units in the existing synthesis model. Residual learning is known to ease training of these networks by reducing parameters and reducing compute cost. Further, the rich skip connections within the network could facilitate information propagation and effectively learn better representations to output better 2D try-on images, and finally better 3D try-on meshes.

2. Methodology & Experimental Results

2.1. Methodology

M3D-VTON. Figure 1 (left) is an overview of the 3D virtual try-on pipeline that we build on. We can see that there are many components involved. The major components are monocular prediction, depth refinement and texture fusion.

The monocular prediction module produces warped clothing, person segmentation and double depth maps which give a base 3D shape. The depth refinement module produces the refined depth maps which capture the warped clothing details as well as the high frequency details which the previous module oversmooths. The texture fusion module merges the warped clothing with unchanged person part to output 2D try-on results. After getting the 2D try-on and depth map, we unproject the front-view and back-view depth maps to get 3D point clouds and triangulate them with screened poisson reconstruction. Since the try-on image and depth maps are spatially aligned, the try-on image can be used to color the front side of the mesh. As for the back texture, the image is inpainted using fast marching method where the face area is filled with surrounding hair color and is then mirrored to finally texture the back side of the mesh.

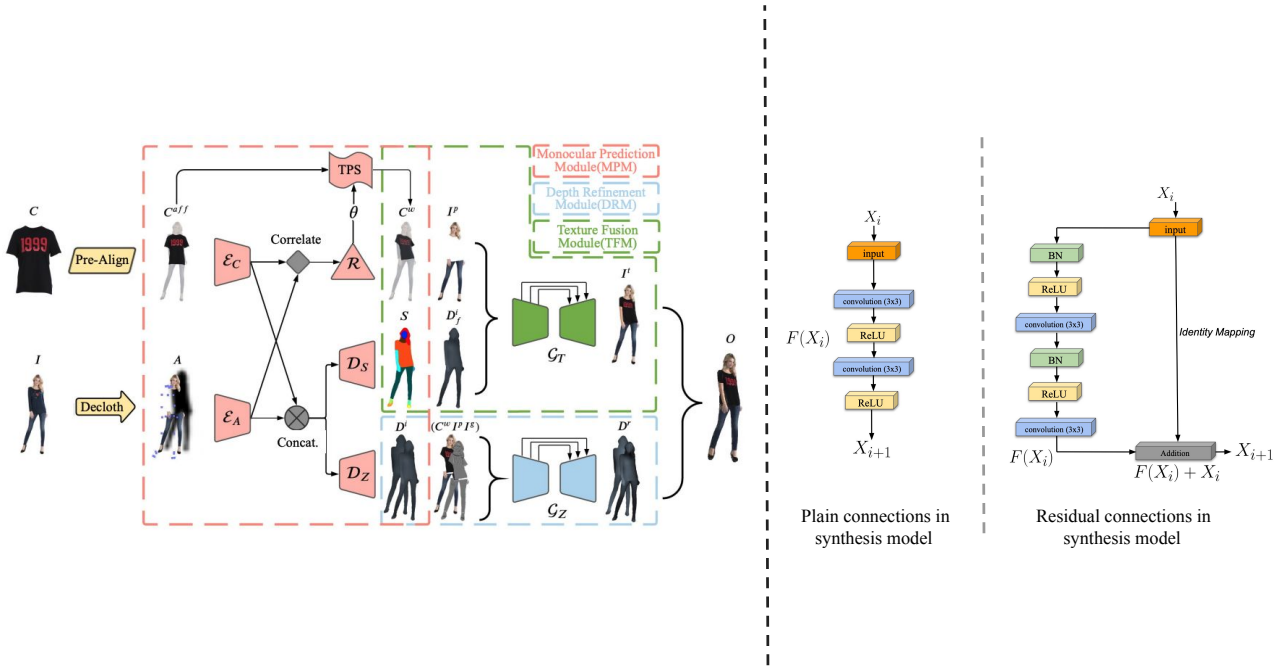

Figure 1. Overview of the proposed framework (left) with an illustration of a plain unit and it's residual counterpart (right). We can see that there are many components involved. The major components are monocular prediction, depth refinement and texture fusion. **Left image taken from M3D-VTON [5]**.

This allows us to achieve the monocular-to-3D conversion, producing the reconstructed 3D try-on mesh with the clothing changed and person identity retained.

Residual connections. The existing synthesis model in texture fusion module which combines all previous outputs comprises of an 5-layer encoder and a 5-layer decoder architecture, similar to that of a U-Net [3]. We argue that the plain connections in this encoder-decoder network in the texture fusion module is not enough to synthesize body parts and differentiate between front and back parts of clothing only from the 2D image. And errors in this step would ultimately lead to unrealistic outputs affecting the final 3D try-on result.

To address the above problem, we propose to use residual connections [1]. The main idea is that residual connections are proven to have better information propagation and effectively learn better representations of the input data, especially known for image recognition tasks [1]. Each connection can be mathematically defined as:

where x_i and x_{i+1} are the input and output of the i-th residual unit, $F(.)$ is the residual function, $f(.)$ is activation function and $h(x_i)$ is an identity mapping function, for instance $h(x_i) = x_i$. Figure 1 (right) shows an illustration of a plain unit and it's residual counterpart. The residual block also consists of batch normalization (BN), ReLU activation and convolutional layers. This approach uses identity mapping [2] that facilitates training and addresses the degradation problem mainly due to vanishing gradients. We refer readers to [1, 2] for more details.

We augment the existing U-Net [3] model in texture fusion module by replacing the plain connections with residual connections. This results in a new synthesis architecture where the encoder and decoder layers consists of residual blocks, similar to that of Deep Residual U-Net [4]. We think that residual connection in the synthesis model is capable on handling to problem of front and back part of clothing, preserve logo as well as reduce artifacts to output better 2D try-on results and eventually better looking textured 3D try-on mesh. Since our work builds on M3D-VTON [5] directly, we follow the same architecture design in the other modules as well as follow the same training and testing protocols. All experiments are performed on a Linux workstation running 4.8Hz and 64GB RAM with and RTX 3080 GPU. Experiments are conducted using Python programming language and PyTorch deep learning framework.

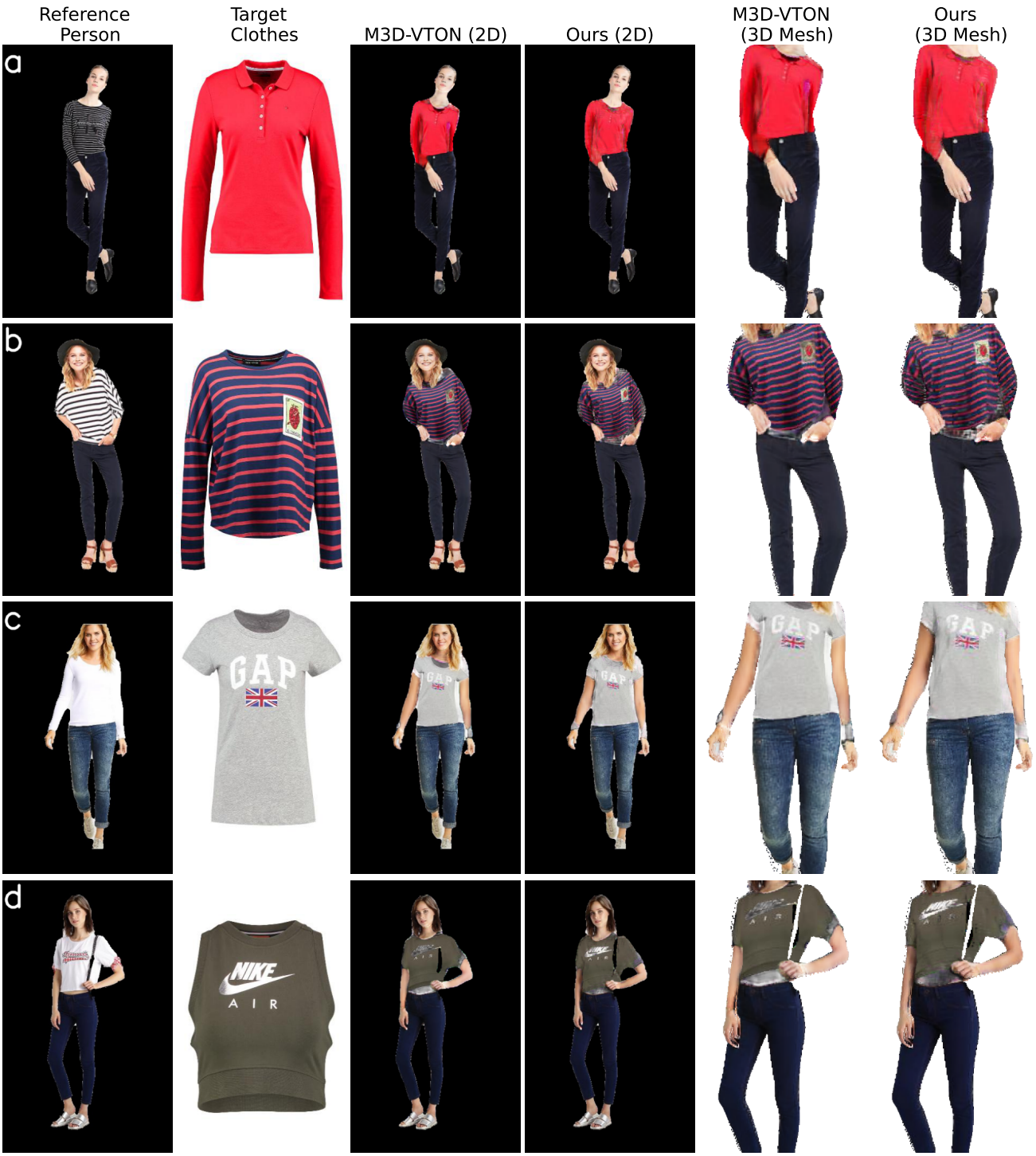

Figure 2. Comparison of 2D and 3D try-on mesh outputs with recent state-of-the-art M3D-VTON.

2.2. Experimental Results

We find that better 2D try-on results lead to better textured 3D meshes. In particular the 3D try-on meshes do not have back part of clothing in the front, preserves logo of clothing and reduces artifacts shown in Figure 2.

Figure 4 shows some examples of the final 2D try-on outputs compared to previous work. In many cases, we see

Method	FID	SSIM
VITON (*CVPR, 2018*)	28.43	0.8807
CP-VTON (*ECCV, 2018*)	20.05	0.8503
CP-VTON+ (*2020*)	23.18	0.8782
ACGPN (*CVPR, 2020*)	20.19	0.8924
M3D-VTON (*ICCV, 2021*)	19.87	0.9725
Ours	**15.16**	**0.9814**

Table 1. 2D try-on SSIM and FID scores on the MPV3D test set. Bolface numbers indicate better performance. Our method consistently outperforms baseline methods.

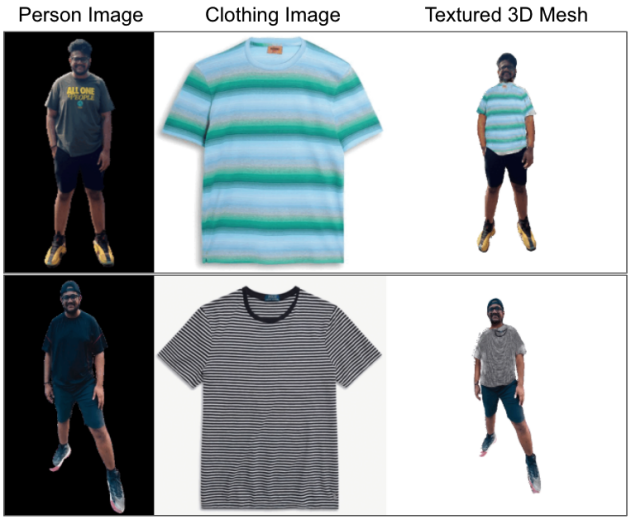

Figure 3. Results of our method on out-of-distribution images. Given the reference person image (left) and target clothing image (middle), our method can reconstruct the 3D try-on mesh (right) with the clothing changed and person identity retained.

that the baseline model is unable to differentiate between the front and back of clothing. It also tends to change the skin color of persons. The baseline model also fails to preserve the logo of clothing image. This is due to the limited capability of the U-Net architecture employed in the baseline model. In comparison, the proposed method generates realistic try-on results which differentiates front and back part of clothing as well as preserve logo of clothing. It also reduces artifacts in non-target body parts such as skin. We show some more examples, where the baseline tends to output blurry logo, synthesize back part of clothing in the front. In comparison, our method mitigates these problems.

We also show two examples in Figure 3 of outputs from our model on out-of-distribution images. Out-of-distribution in the sense that the model is trained on MPV3D dataset which consists of only women images and women top clothing, while the images here are of men. We can see that the model is able to reconstruct the 3D try-on mesh with the clothing changed and person identity retained.

Finally, we show some quantitative results in Table 1 on two metrics which are currently used to benchmark try-on methods. Our method consistently outperforms baseline methods with an improvement of almost 5% over the previous best method on FID score.

3. Conclusions

To summarize, we integrate residual connections into the synthesis model of a recent 3D virtual try-on pipeline. Studying it's effect demonstrates that it improves 2D try-on outputs, mainly by differentiating between front and back part of clothing, preserving logo of clothing and reducing artifacts. This ultimately results in better textured 3D try-on mesh. Benchmarking our method on the MPV3D dataset shows that it performs better than previous works significantly.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European conference on computer vision*, pages 630–645. Springer, 2016. 2

[3] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 2

[4] Zhengxin Zhang, Qingjie Liu, and Yunhong Wang. Road extraction by deep residual U-Net. *IEEE Geoscience and Remote Sensing Letters*, 15(5):749–753, 2018. 2

[5] Fuwei Zhao, Zhenyu Xie, Michael Kampffmeyer, Haoye Dong, Songfang Han, Tianxiang Zheng, Tao Zhang, and Xiaodan Liang. M3d-vton: A monocular-to-3d virtual try-on network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13239–13249, 2021. 1, 2

Appendix: Extra Results

Figure 4 shows more examples of the final try-on outputs compared to previous work. In many cases, we see that the baseline model is unable to differentiate between the front and back of clothing. It also tends to change the skin color of the person. The baseline model also fails to preserve the logo of clothing image. This is due to the limited capability of the U-Net architecture employed in the baseline model. In comparison, the proposed method generates realistic try-on results which differentiates front and back part of clothing, preserve logo of clothing. It also reduces artifacts in non-target body parts such as skin.

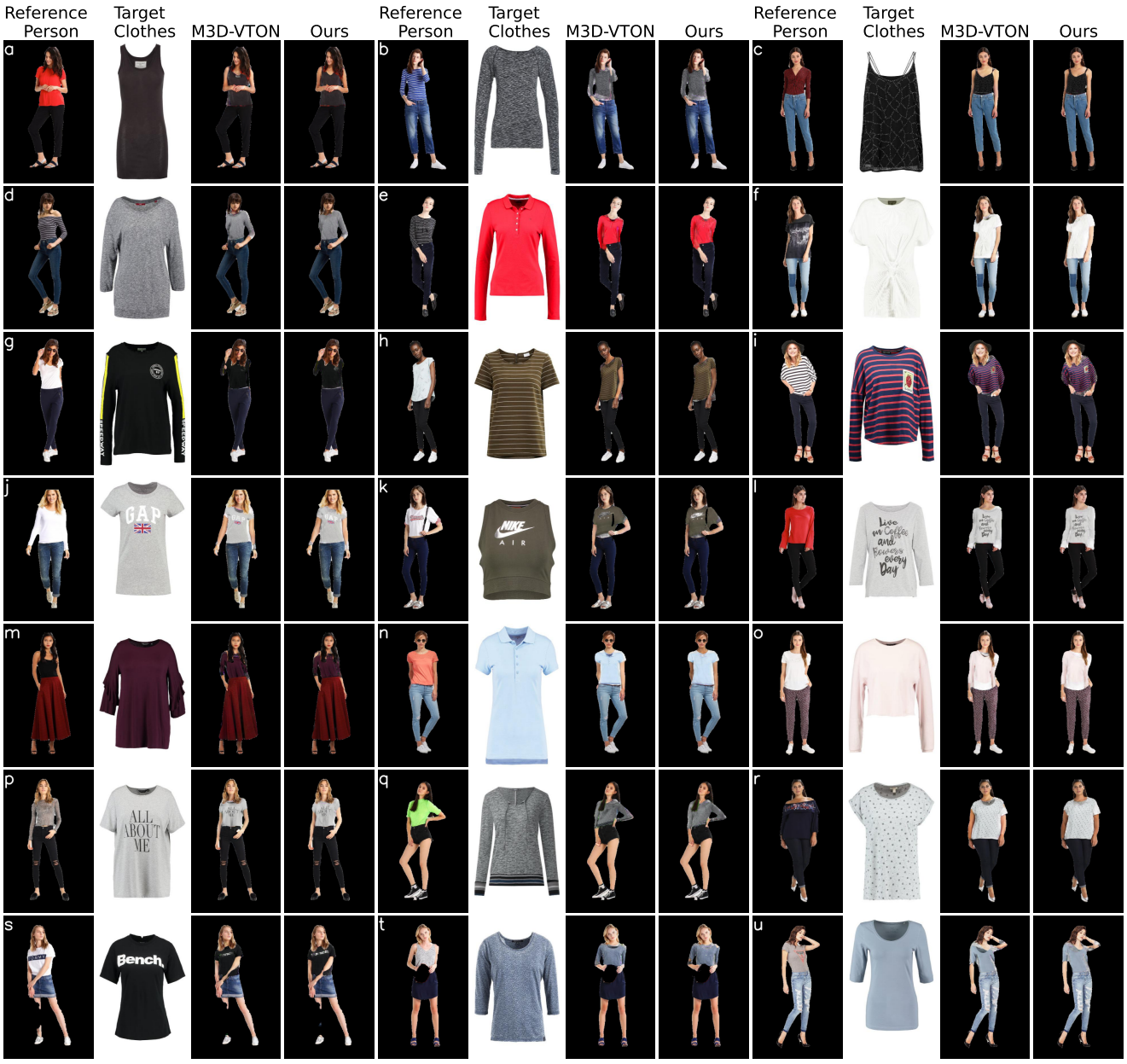

Figure 4. Extensive visual results of 2D try-on outputs with M3D-VTON. Our method generates realistic try-on results which differentiates front and back part of clothing, preserve logo of clothing. It also reduces artifacts in non-target body parts such as skin