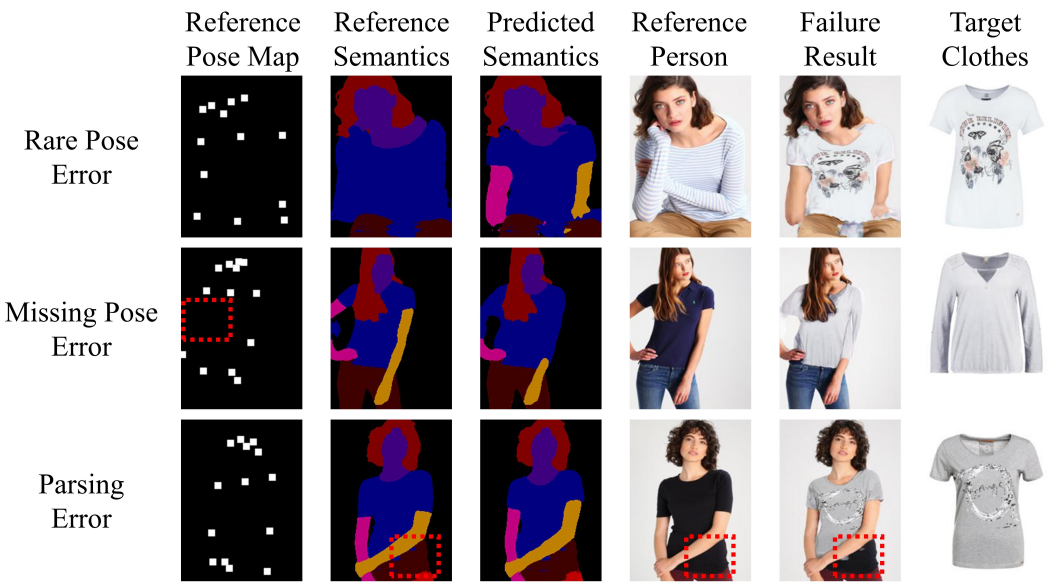

Figure 1. Three main failure cases are presented including rare pose error, missing pose error, and parsing error. Typical failure causes are red-boxed.

A. More Try-on Results

The extensive results of RT-VTON are given in our homepage https://lzqhardworker.github.io/RT-VTON/.

B. Limitations

In Fig. 1, we present the failure cases of our method. The first row shows a person with huge self-occlusions and complex body intersections. Depth information can be utilized to enhance the overall understanding of the reference person. In the second row, the elbow of the reference person is out of the image, which makes the pose map incomplete. Simply applying other dense pose representations can ameliorate this problem. In the last row, the initial semantics of the reference person are incorrect, where a part of the clothing region is mis-classified as bottom clothes. Better semantic parser or distillation trick can be used to improve this case. **Notably, RT-VTON is designed for predicting more accurate semantic layout given the correct initial semantics. Error-handling of the pretrained semantic parser or the pose estimator is not our main focus.**

C. Moving Least Squares

To make the problem of Moving Least Squares clear, we largely follow the presentation as addressed in [5] and extract the essence for better understanding.

Let p be a set of control points in the image and q the target control points in the image, where p moves to q after the transformation. Given a certain point v, we look for the optimal affine transformation $l_v(x)$ which minimizes

$$\sum_i w_i \left| l_v\left(p_i\right) - q_i \right|^2, \tag{1}$$

where p_i and q_i are vectors and the weights w_i are defined as

$$w_i = \frac{1}{\left|p_i - v\right|^{2\alpha}}. \tag{2}$$

The weights w_i in this least squares problem depend on the evaluation point v. Therefore, we get a different transformation $l_v(x)$ for each v.

$l_v(x)$ consists of two parts: a linear transformation matrix M and a translation T, for $l_v(x)$ is an affine transformation:

$$l_v(x) = xM + T. \tag{3}$$

We solve for T to get that

$$T = q_* - p_* M, \tag{4}$$

where p and q are weighted centers.

$$\begin{aligned} p_* &= \frac{\sum_i w_i p_i}{\sum_i w_i}, \\ q_* &= \frac{\sum_i w_i q_i}{\sum_i w_i}. \end{aligned} \tag{5}$$

Based on the above observation, we can replace T with Equation 3. $l_v(x)$ can be rewritten as a linear matrix M,

$$l_v(x) = \left(x - p_*\right) M + q_*. \tag{6}$$

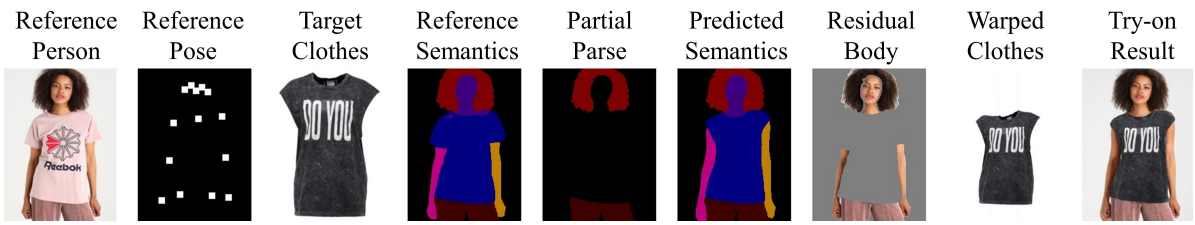

Figure 2. The intermediate results of our method.

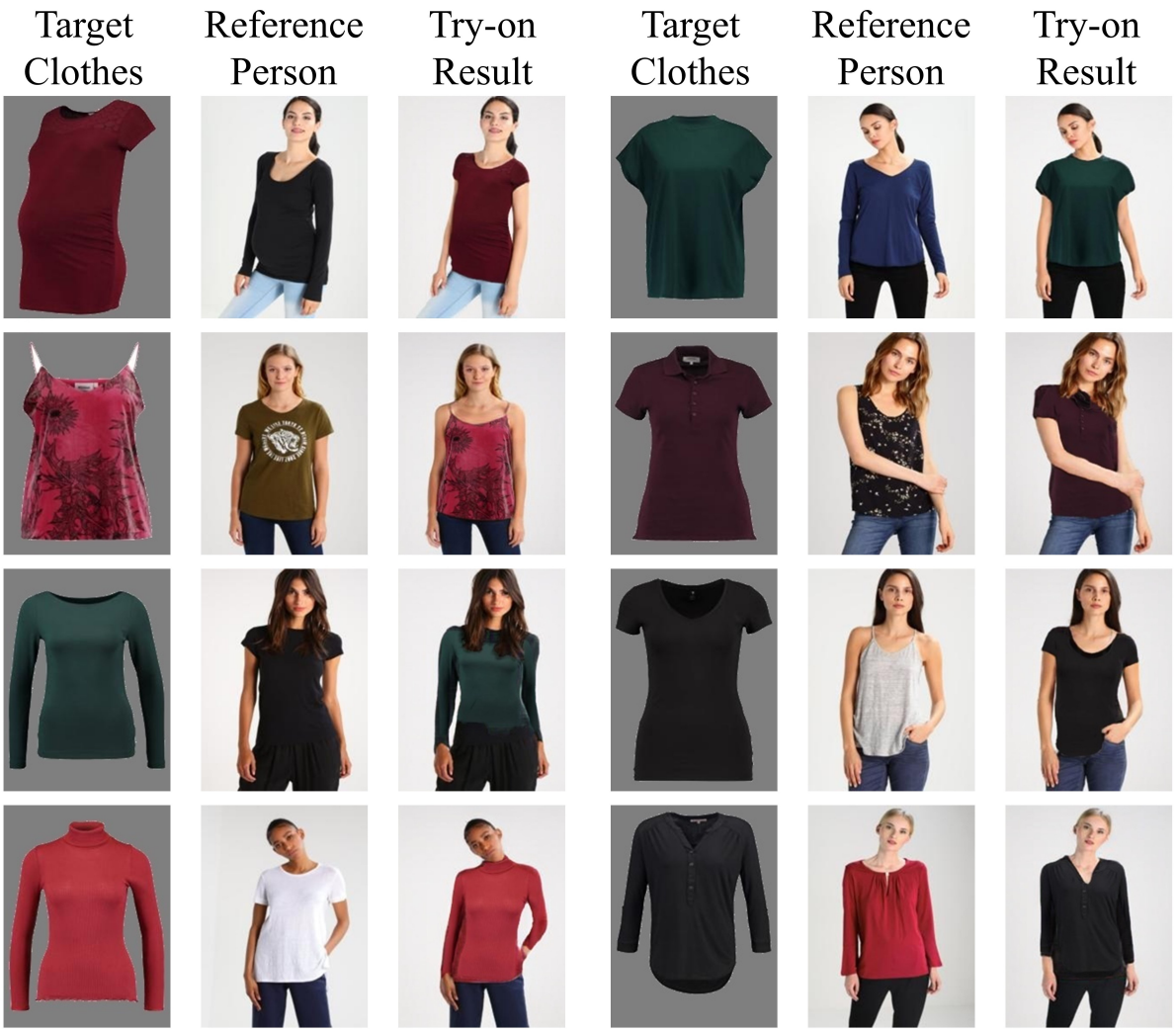

Figure 3. Extensive results regarding intricate neckline and collar structures.

Based on the above formula, the least squares problem can be redefined as

$$\sum_i w_i \left| \hat{p}_i M - \hat{q}_i \right|^2, \tag{7}$$

where $\hat{p}_i = p_i - p_*$ and $\hat{q}_i = q_i - q_*$. Then we find an affine deformation that minimizes Equation 7 by least squares.

$$M = \left(\sum_i \hat{p}_i^T w_i \hat{p}_i \right)^{-1} \sum_j w_j \hat{p}_j^T \hat{q}_j. \tag{8}$$

D. Intermediate Results

We show the intermediate results of RT-VTON in Fig. 2 for better understanding of our overall pipeline. We firstly remove the face, upper clothes, and arm labels from the reference semantics to derive partial parse; the original clothing shape is thus agnostic to the network. Then SGM predicts the "after-try-on" semantic layout *i.e*. predicted semantics, given the target clothes as well as the reference pose map. With accurate semantic segmentation, we can adaptively generate and preserve the image contents by computing the intersection of skin regions, *i.e*. residual body. Eventually we combine the predicted semantics, residual body with the warped clothes to produce the final try-on results.

E. Results on Collar Types

In order to validate the effectiveness of RT-VTON on handling intricate collar structures, we show an extensive visual results in Fig. 3. RT-VTON performs well in capturing the detailed collar shapes.

F. Extra Quantitative Results

Here we give an extra quantitative experiment in Tab. 1 with some recent works without official implementations. The numbers are directly copied from their papers, which are thus only for reference.

Table 1. Without the official implementations, we compare the FID score with their reported values.

Method	FID
SieveNet [4]	26.67
ClothFlow [3]	23.68
ZFlow [1]	15.17
RT-VTON	**11.66**

Structural Similarity (SSIM) [6] and Peak Signal to Noise Ratio (PSNR) are not used in our experiment for the following reasons. **1)** Since we do not have the ground-truth images (i.e. reference person wearing the target clothes) as discussed in [2], FID can best depict the unpaired try-on quality. **2)** SSIM and PSNR are reconstruction metrics, which are computed by putting on the same clothes of the reference person. A severe problem for reconstruction metrics is that the method which generates trivial identical result of reference person can score the highest in SSIM and PSNR.

References

[1] Ayush Chopra, Rishabh Jain, Mayur Hemani, and Balaji Krishnamurthy. Zflow: Gated appearance flow-based virtual try-on with 3d priors. *CoRR*, abs/2109.07001, 2021. 3

[2] Ge et al. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, 2021. 3

[3] Xintong Han, Weilin Huang, Xiaojun Hu, and Matthew R. Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, pages 10470–10479. IEEE, 2019. 3

[4] Surgan Jandial, Ayush Chopra, Kumar Ayush, Mayur Hemani, Abhijeet Kumar, and Balaji Krishnamurthy. Sievenet: A unified framework for robust image-based virtual try-on. In *WACV*, pages 2171–2179. IEEE, 2020. 3

[5] Scott Schaefer, Travis McPhail, and Joe D. Warren. Image deformation using moving least squares. *ACM Trans. Graph.*, 25(3):533–540, 2006. 1

[6] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4):600–612, 2004. 3