SieveNet: A Unified Framework for Robust Image-Based Virtual Try-On Supplementary Material

1. Duelling Triplet Loss Strategy

Figure 1 presents an illustration of the Duelling Triplet Loss Strategy proposed for training the Segmentation Assisted Texture Translation (SATT) module of our SieveNet framework. Training of the SATT module is done in multiple phases. The first K steps of training is a conditioning phase that minimizes the L_{tt} (see Section 3.4.2) to produce reasonable results. The subsequent phases (each lasting T steps) employ the L_{tt} loss augmented with a triplet loss (see Section 3.4.3) to fine-tune the results further. This strategy further improves the output significantly (see additional results in Figure 2).

As discussed earlier, a triplet loss is characterized by an anchor, a positive and a negative (w.r.t the anchor), with the objective being to simultaneously push the anchor result towards the positive and away from the negative. In the duelling triplet loss strategy, we pit the output obtained from the network with the current weights (anchor) against that from the network with weights from a previous phase (negative), and push it towards the ground-truth (positive). As training progresses, this online hard negative mining strategy helps push the results closer to the ground-truth by updating the negative at discrete step intervals (T steps).

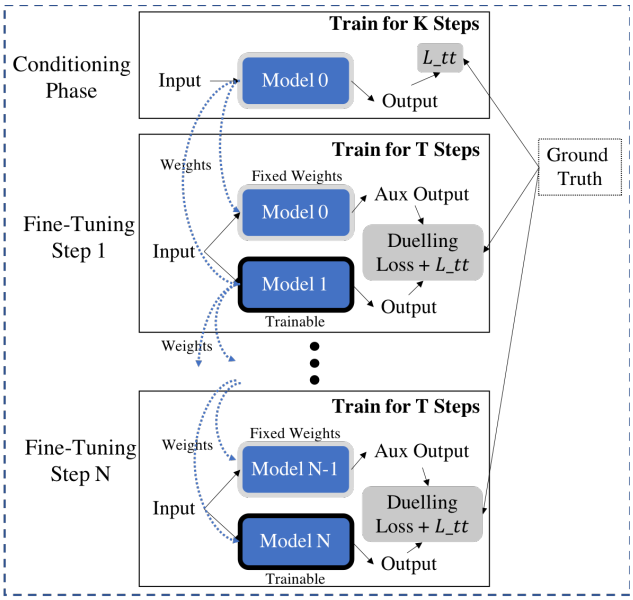


Figure 1: Visualization of the Duelling Triplet Loss strategy for training the SATT module. The model represented by thick black boundary is trainable unlike the one in the thick grey boundary. At the beginning of any fine-tuning step (say Fine Tuning Step 1 in above figure), the weights from the previous step are transferred to both the models (grey boundary and black boundary). Subsequently, during training in that fine-tuning step, the weights of the model in the black boundary are only modified and that of the grey boundary one are kept fixed. The thick grey boundary model is used to generate hard negatives which is used in the triplet loss for training the thick black boundary model. Please see Section 3.4.2 and 3.4.3 for mathematical formulation of L_{tt} and duelling loss.

2. Additional Qualitative Results

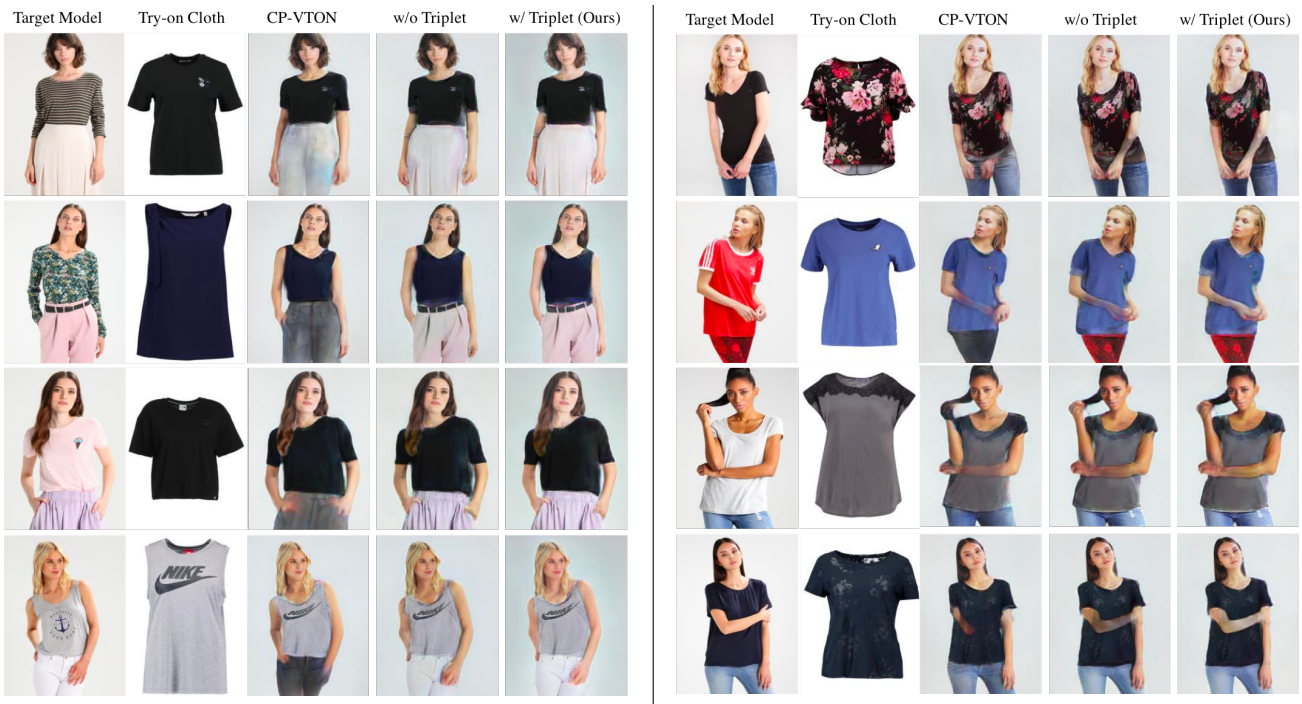
Figure 4 and 5 present additional results of comparison of the proposed SieveNet with those of CP-VTON.

2.1. Impact of Conditional Segmentation Mask Prediction

Figure 3 presents additional results obtained by training the texture transfer module of CP-VTON (TOM) with an additional prior of the try-on cloth conditioned segmentation mask and unaffected regions from the model image (I_m) to produce the final try-on image. It can be observed that this improves handling of skin generation, bleeding and variability of poses. Providing the expected segmentation mask of the try-on output image as prior equips the generation process to better handle these issues. Also, since the unaffected parts of the model image are also provided as prior, the proposed framework is also able to better translate texture of auxiliary products such as bottoms onto the final generated try-on image (unlike in CP-VTON).

2.2. Impact of Duelling Triplet Loss

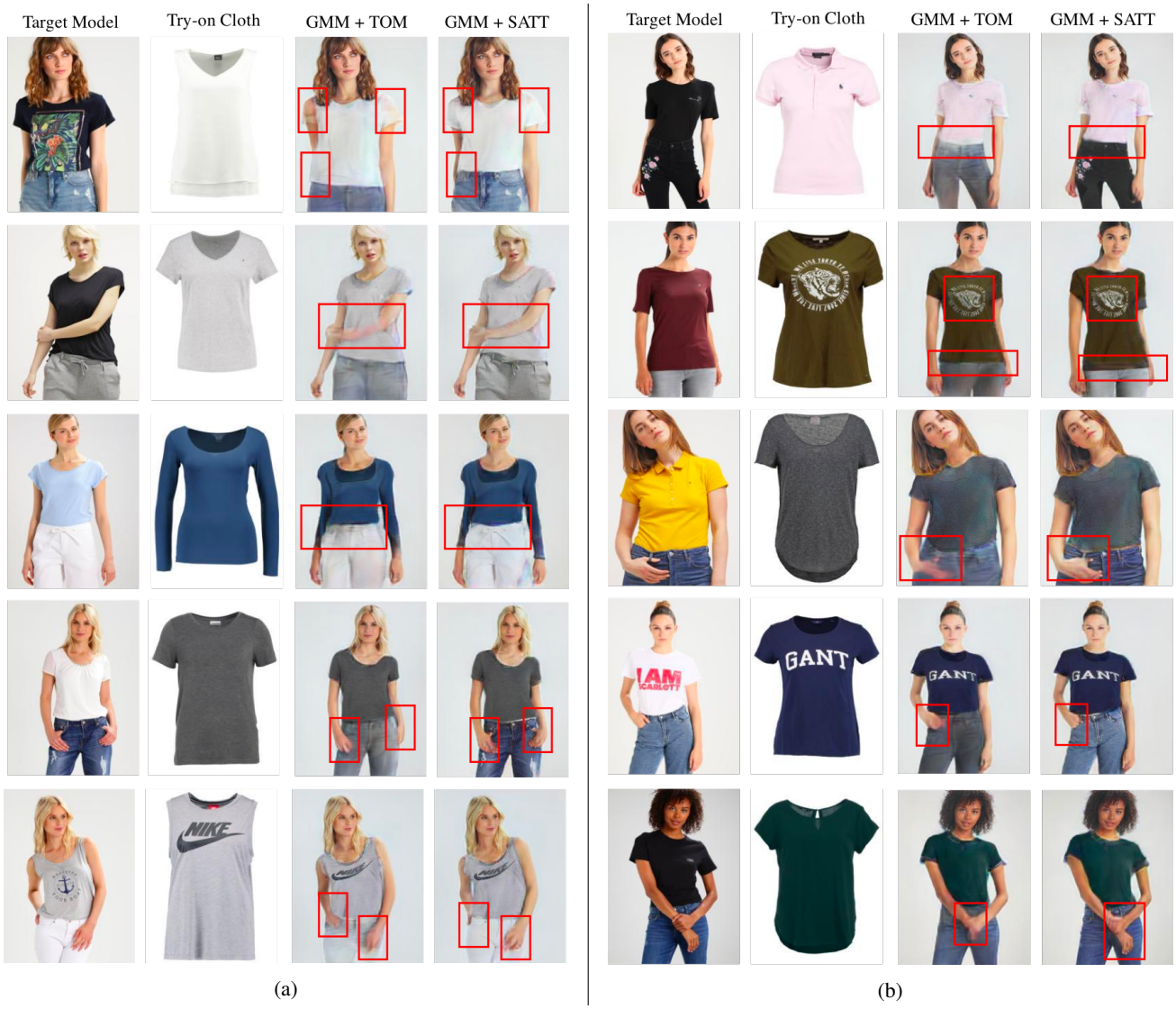
In Figure 2, we present additional results depicting the particular benefit of training the texture translation network



(a) (b)

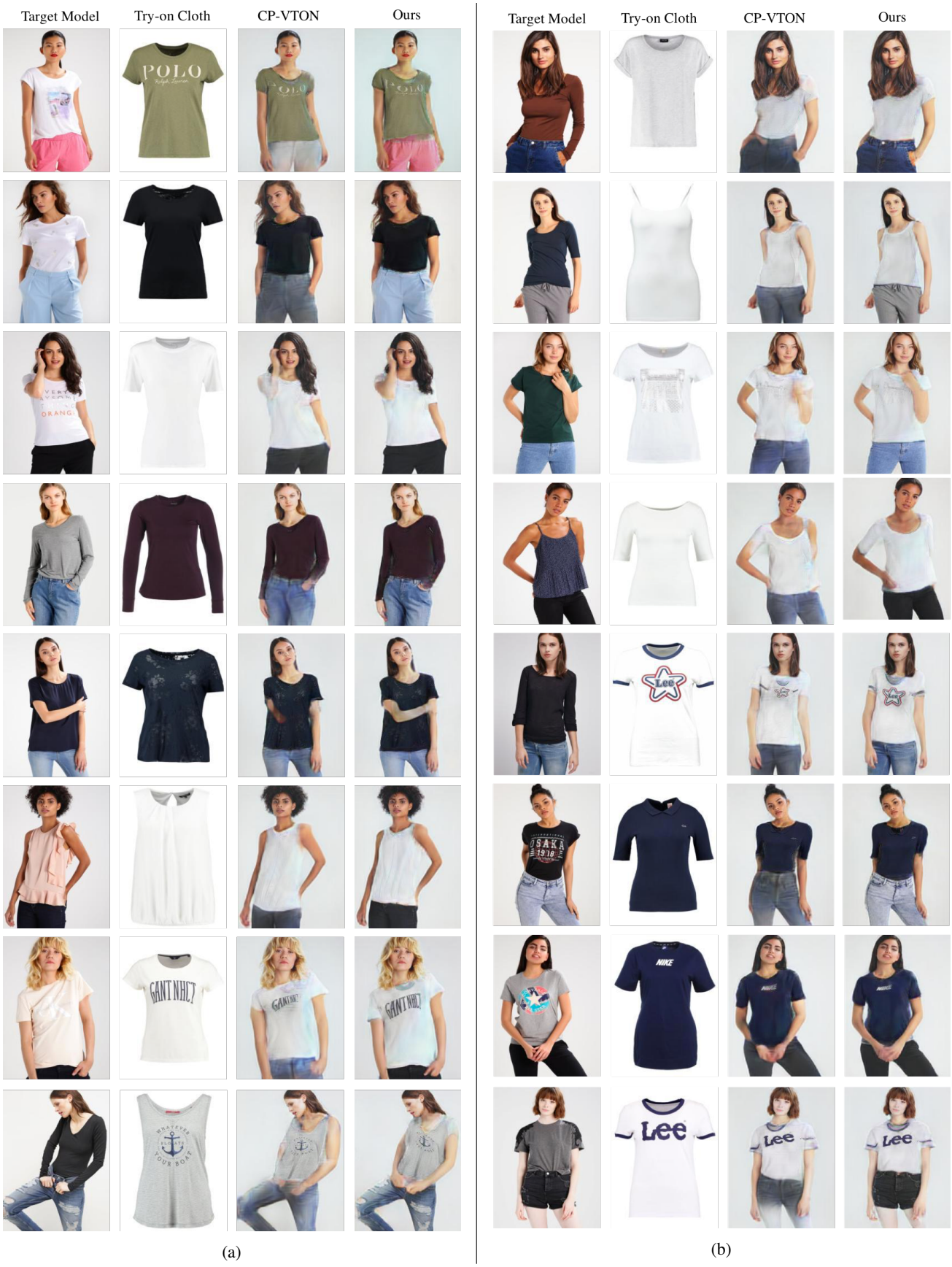
Figure 2: Fine-tuning texture translation with the duelling triplet strategy refines quality of generated images by handling occlusion and avoiding bleeding.

with the duelling triplet loss strategy. As highlighted by the results, this duelling triplet loss strategy behaves as an online hard negative mining strategy in the fine-tuning stage and subsequently refines the quality of the generated results. This arises from better handling of occlusion, bleeding and skin generation.



(a) (b)

Figure 3: Using the conditional segmentation mask as prior to texture transfer aids in better handling of variable pose, occlusion and helps avoid bleeding. Also, since the unaffected parts of the model image are also provided as prior, the proposed framework is also able to better translate texture of auxiliary products such as bottoms onto the final generated try-on image



(a) (b)

Figure 4: Comparison of SieveNet with CP-VTON. SieveNet can generate more realistic try-on results compared to the current state-of-the-art CP-VTON.

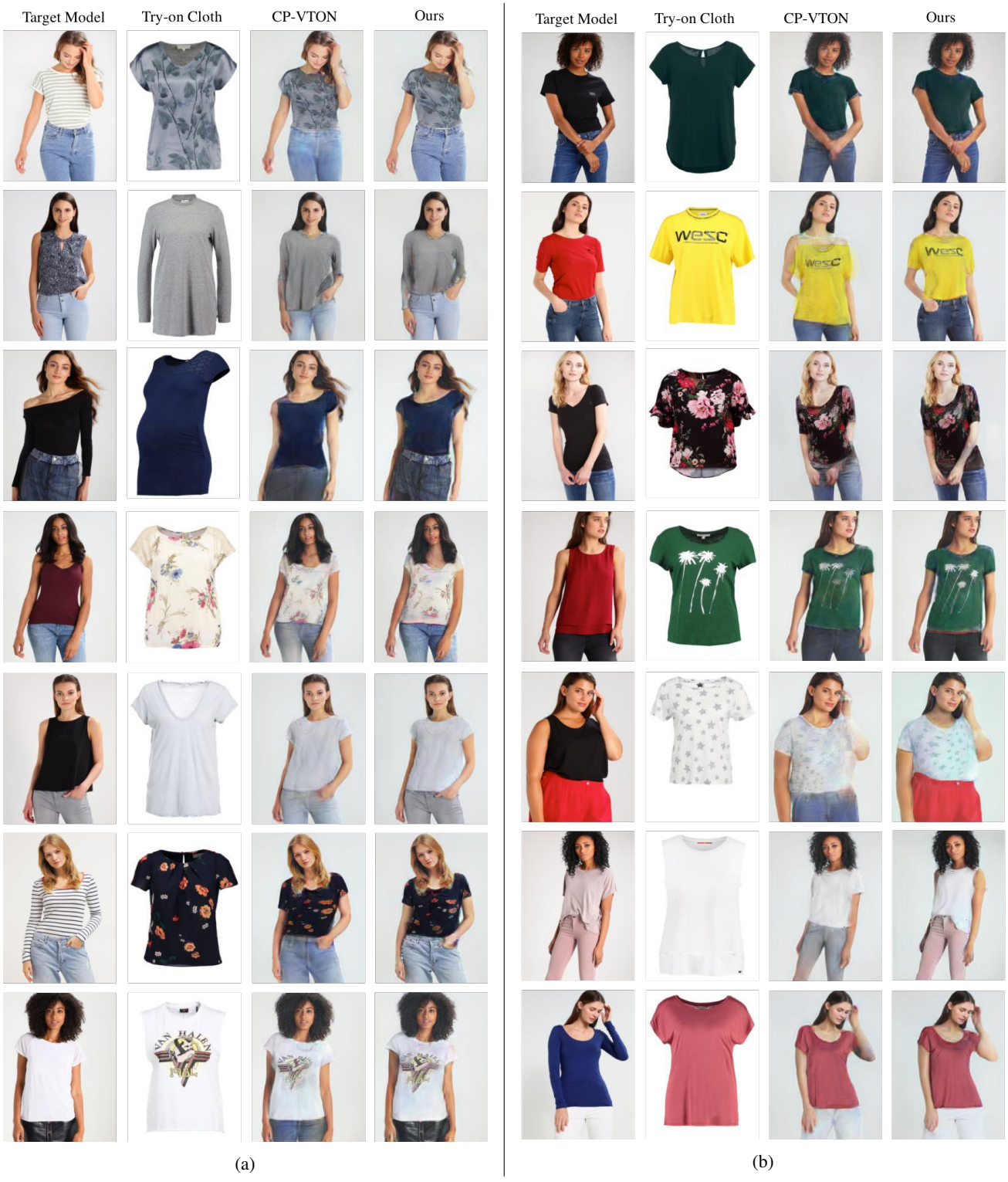


Figure 5: Comparison of SieveNet with CP-VTON. SieveNet can generate more realistic try-on results compared to the current state-of-the-art CP-VTON.