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

Ayush Chopra*[1], Surgan Jandial†*[2], Kumar Ayush‡*[3], Mayur Hemani[1], and Balaji Krishnamurthy[1]

[1]Media and Data Science Research, Adobe Experience Cloud
[2]IIT Hyderabad
[3]Stanford University

Abstract

Image-based virtual try-on for fashion has gained considerable attention recently. The task requires trying on a clothing item on a target model image. An efficient framework for this is composed of two stages: (1) warping (transforming) the try-on cloth to align with the pose and shape of the target model, and (2) a texture transfer module to seamlessly integrate the warped try-on cloth onto the target model image. Existing methods suffer from artifacts and distortions in their try-on output. In this work, we present SieveNet, a framework for robust image-based virtual try-on. Firstly, we introduce a multi-stage coarse-to-fine warping network to better model fine grained intricacies (while transforming the try-on cloth) and train it with a novel perceptual geometric matching loss. Next, we introduce a try-on cloth conditioned segmentation mask prior to improve the texture transfer network. Finally, we also introduce a duelling triplet loss strategy for training the texture translation network which further improves quality of generated try-on result. We present extensive qualitative and quantitative evaluations of each component of the proposed pipeline and show significant performance improvements against the current state-of-the-art method.

1. Introduction

Providing interactive shopping experiences is an important problem for online fashion commerce. Consequently, several recent efforts have been directed towards delivering smart, intuitive online experiences including clothing retrieval [10, 2], compatibility prediction [3, 19] and virtual try-on [20, 5]. Virtual try-on is the visualization of fashion products in a personalized setting. The problem consists of trying on a specific garment on the image of a person. It is especially important for online fashion commerce because it compensates for the lack of a direct physical experience of in-store shopping.

Figure 1: The Task of Image Based Virtual Try-on

Recent methods based on deep neural networks [20, 5], formulate the problem as that of conditional image generation. As depicted in Figure 1, the objective is to synthesize a new image (henceforth referred to as the try-on output) from two images - a try-on cloth and a target model image, such that in the try-on output the target model is wearing the try-on cloth while the original body shape, pose and other model details (eg. bottom, face) are preserved.

Successful virtual try-on experience depends upon synthesizing images free from artifacts arising from improper positioning or shaping of the try-on garment, and inefficient composition resulting in blurry or bleeding garment textures in the final try-on output. Current solutions [5, 20] suffer from these problems especially when the try-on cloth is subject to extreme deformations or when characteristics of the try-on cloth and original clothing item in target model differ. For example, transferring a half-sleeves shirt image to a target model originally in a full-sleeves shirt often results in texture bleeding and incorrect warping. For alleviating these problems, we propose:

1. A multi-stage coarse-to-fine warping module trained with a novel perceptual geometric matching loss to bet-

*equal contribution
†work done as part of Adobe MDSR internship program
‡work done while at Adobe

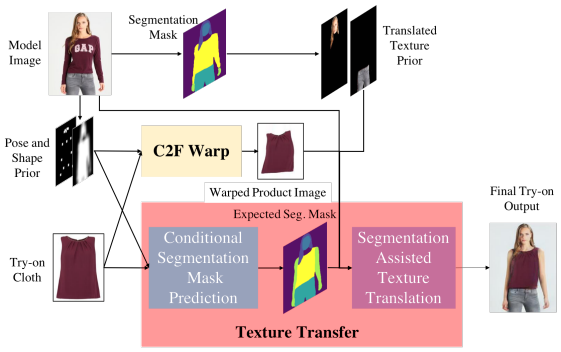


Figure 2: Inference Pipeline of the SieveNet framework

ter model fine intricacies while transforming the try-on cloth image to align with the body shape of the target model.

2. A conditional segmentation mask generation module to assist in handling complexities arising from variable pose, occlusion and bleeding during the texture transfer process, and
3. A duelling triplet loss strategy for training the texture translation network to further improve quality of the final try-on result.

We show significant qualitative and quantitative improvement over the current state-of-the-art method for image-based virtual try-on. An overview of our SieveNet framework is presented in Figure 2 and the training pipeline is detailed in Figure 3.

2. Related Work

Our work is related to existing methods for conditional person image synthesis that use pose and shape information to produce images of humans, and to existing virtual try-on methods - most notably [20].

2.1. Conditional Image Synthesis

Ma et al. [11] proposed a framework for generating human images with pose guidance along with a refinement network trained using an adversarial loss. Deformable GANs [18] attempted to alleviate the misalignment problem between different poses by using an affine-transformation on the coarse rectangle region, and warped the parts on pixel-level. In [4], Esser et al. introduced a variational U-Net [15] to synthesize the person image by restructuring the shape with stickman pose guidance. [13] applied CycleGAN directly to manipulate pose. However, all of these methods fail to preserve the texture details of the clothes in the output. Therefore, they cannot directly be applied to the virtual try-on problem.

2.2. Virtual Try-On

Initial works on virtual try-on were based on 3D modeling techniques and computer graphics. Sekine et al. [17] introduced a virtual fitting system that captures 3D measurements of body shape via depth images for adjusting 2D clothing images. Pons-Moll et al. [12] used a 3D scanner to automatically capture real clothing and estimate body shape and pose. Compared to graphics models, image-based generative models provide a more economical and computationally efficient solution. Jetchev et al. [8] proposed a conditional analogy GAN to swap fashion articles between models without using person representations. They do not take pose variant into consideration, and during inference, they required the paired images of in-shop clothes and a wearer, which limits their applicability in practical scenarios. In [5], Han et al. introduce a virtual try-on network to transfer a desired garment on a person image. It uses an encoder-decoder with skip connections to produce a coarse clothing mask and a coarse rendered person image. It then uses a Thin-plate spline (TPS) based spatial transformation (from [7]) to align the garment image with the pose of the person, and finally a refinement stage to overlay the warped garment image on to the coarse person image to produce the final try-on image. Most recently Wang et al. [20] present an improvement over [5] by directly predicting the TPS parameters from the pose and shape information and the try-on cloth image. Both of these methods suffer from geometric misalignment, blurry and bleeding textures in cases where the target model is characterized by occlusion and where pose variation or garment shape variation is high. Our method, aligned to the approach in [20], improves upon all of these methods. SieveNet learns the TPS parameters in multiple stages to handle fine-grained shape intricacies and uses a conditional segmentation mask generation step to aid in handling of pose variation and occlusions, and improve textures. In Section 5.2, we compare our results with those of [20].

3. Proposed Methodology

The overall process (Figure 2) comprises of two main stages - warping the try-on cloth to align with pose and shape of the target model, and transferring the texture from the warped output onto the target model to generate the final try-on image. We introduce three major refinements into this process. To capture fine details in the geometric warping stage, we use a two-stage spatial-transformer based warp module (Section 3.2). To prevent the garment textures from bleeding onto skin and other areas, we introduce a conditional segmentation mask generation module (Section 3.4.1) that computes an expected semantic segmentation mask to reflect the bounds of the target garment on the model, which in turn assists the texture translation network

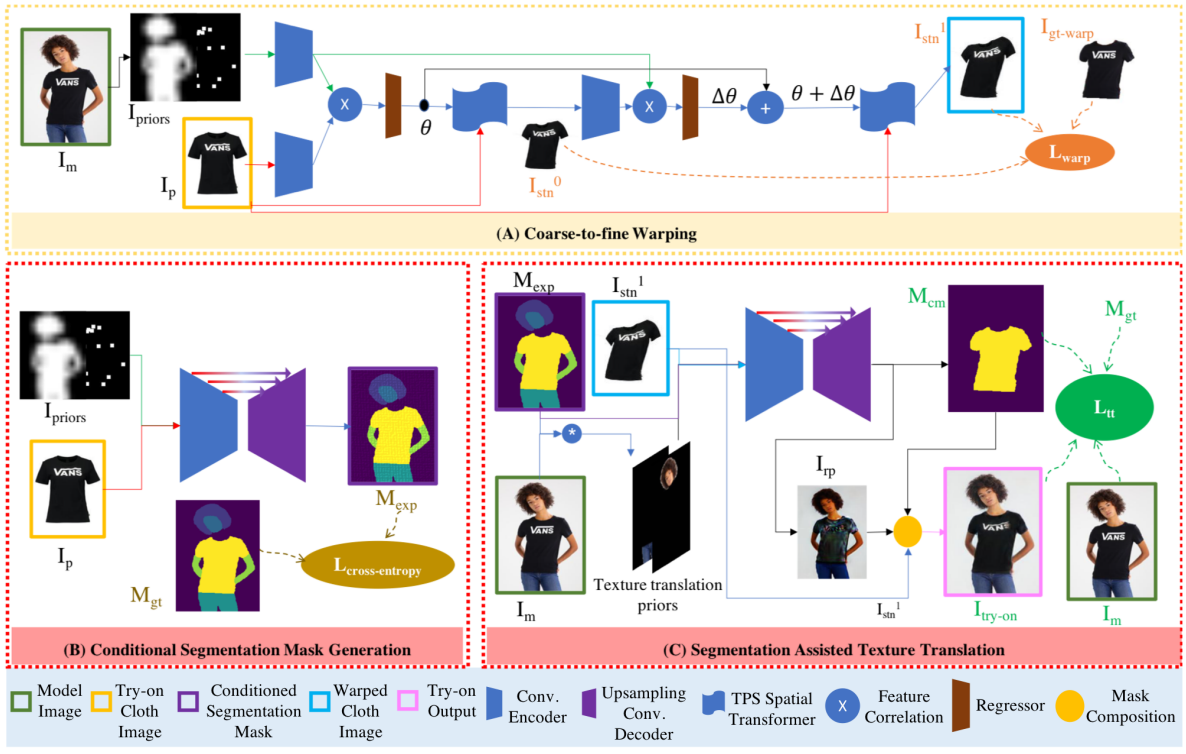


Figure 3: An overview of the training pipeline of SieveNet, containing (A) Coarse-to-Fine Warping Module, (B) Conditional Segmentation Mask Generation Module, and (C) Segmentation Assisted Texture Translation Module.

to produce realistic try-on results. We also propose two new loss computations - a perceptual geometric matching loss (Section 3.3) to improve the warping output, and a duelling triplet loss strategy (Section 3.4.3) to improve the output from the texture translation network.

3.1. Inputs

The framework uses the try-on cloth image (I_p), a 19-channel pose and body-shape map (I_{priors}) generated as described in [20] as input to the various networks in our framework. I_{priors} is a cloth-agnostic person representation created using the model image (I_m) to overcome the unavailability of ideal training triplets as discussed in [20]. A human parsing semantic segmentation mask (M_{gt}) is also used as ground-truth during training of the conditional segmentation mask generation module (described in Sections 3.3 and 3.4.1). For training, the task is set such that data consists of paired examples where the model in I_m is wearing the clothing product I_p.

3.2. Coarse-to-Fine Warping

The first stage of the framework warps the try-on product image (I_p) to align with the pose and shape of the target model (I_m). It uses the priors I_{priors} as guidance for achieving this alignment. Warping is achieved using thin-plate spline (TPS) based spatial transformers [7], as introduced in [20] with a key difference that we learn the transformation parameters in a two-stage cascaded structure and use a novel perceptual geometric matching loss for training.

3.2.1 Tackling Occlusion and Pose-variation

We posit that accurate warping requires accounting for intricate modifications resulting from two major factors:

1. Large variations in shape or pose between the try-on cloth image and the corresponding regions in the model image.

2. Occlusions in the model image. For example, the long hair of a person may occlude part of the garment near the top.

The warping module is formulated as a two-stage network to overcome these problems of occlusion and pose-variation. The first stage predicts a coarse-level transformation, and the second stage predicts the fine-level corrections on top of the coarse transformation. The transformation parameters from the coarse-level regression network (θ) is used to warp the product image to produce an approximate

warp output (I^0_{stn}). This output is then used to compute the fine-level transformation parameters ($\Delta\theta$) and the corresponding warp output (I^1_{stn}) is computed using ($\theta + \Delta\theta$) to warp the initial try-on cloth I_p and not I^0_{stn}. This is done to avoid the artifacts from applying the interpolation in the spatial transformer twice. To facilitate the expected hierarchical behaviour, residual connections are introduced to offset the parameters of the fine-transformation with the coarse-transformation. The network structure is schematized in Figure 3 (A). Ablation study to support the design of the network and losses is in Section 5.3.1.

3.3. Perceptual Geometric Matching Loss

The interim (I^0_{stn}) and final (I^1_{stn}) output from the warping stage are subject to a matching loss L_{warp} against the $I_{gt-warp}$ (segmented out from the model image) during training. L_{warp} is defined below which includes a novel perceptual geometric matching loss L_{pgm} component. The intuition behind this loss component L_{pgm} is to have the second stage warping incrementally improve upon that from the first stage.

$$\begin{aligned} L_{warp} &= \lambda_1 L^0_s + \lambda_2 L^1_s + \lambda_3 L_{pgm} \\ L^0_s &= |I_{gt-warp} - I^0_{stn}| \\ L^1_s &= |I_{gt-warp} - I^1_{stn}| \end{aligned} \tag{1}$$

Here, $I_{gt-warp} = I_m * M^{cloth}_{gt}$, and L_{pgm} is the perceptual geometric matching loss which comprises of two components. $I_{gt-warp}$ is the cloth worn on the target model in I_m and M^{cloth}_{gt} is the binary mask representing the cloth worn on the target model.

$$L_{pgm} = \lambda_4 L_{push} + \lambda_5 L_{align} \tag{2}$$

Minimizing L_{push} pushes the second stage output I^1_{stn} closer to the ground-truth $I_{gt-warp}$ compared to the first stage output.

$$L_{push} = k * L^1_s - |I^1_{stn} - I^0_{stn}| \tag{3}$$

The scalar k is a multiplicative margin used to ensure stricter bound for the difference ($k = 3$ is used for our experiments). For L_{push}, I^0_{stn}, I^1_{stn} and $I_{gt-warp}$ are first mapped to the VGG-19 activation space, and then the loss attempts to align the difference vectors between I^0_{stn} and $I_{gt-warp}$, and I^0_{stn} and $I_{gt-warp}$ in the feature space.

$$\begin{aligned} V^0 &= VGG(I^0_{stn}) - VGG(I_{gt-warp}) \\ V^1 &= VGG(I^1_{stn}) - VGG(I_{gt-warp}) \\ L_{align} &= (CosineSimilarity(V^0, V^1) - 1)^2 \end{aligned} \tag{4}$$

Minimizing L_{align} loss facilitates the goal of minimizing L_{push}.

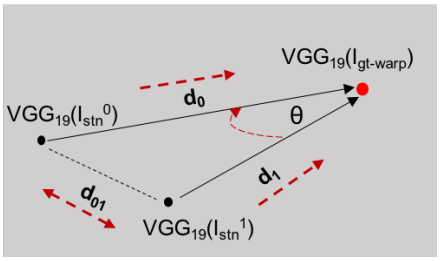


Figure 4: Visualization of the Perceptual Geometric Matching Loss in VGG-19 Feature Space.

Figure 5: Illustrating work of Conditional Segmentation Mask Prediction Network

3.4. Texture Transfer

Once the product image is warped to align with the pose and shape of the target model, the next stage transfers the warped product to the model image. This stage computes a rendered model image, and a fractional composition mask to compose the warped product image onto the rendered model image. We break down this stage into two steps - conditional segmentation mask prediction and segmentation assisted texture translation.

3.4.1 Conditional Segmentation Mask Prediction

A key problem with existing methods is their inability to accurately honor the bounds of the clothing product and human skin. The product pixels often bleeds into the skin pixels (or vice-versa), and in the case of self-occlusion (such as with the case of folded arms), the skins pixels may get replaced entirely. This problem is exacerbated for cases where the try-on clothing item has a significantly different shape than the clothing in the model image. Yet another scenario that aggravates this problem is when the target model is in a complex pose. To help mitigate these problems of bleeding and self-occlusion as well as to handle variable and complex poses, we introduce a conditional segmentation mask prediction network.

Figure 3 (B) illustrates the schematics of the network.

It takes the pose and shape priors (I_{priors}) and the product image (I_p) as input, to generate an "expected" segmentation mask (M_{exp}). This try-on clothing conditioned seg. mask represents the expected segmentation of the generated try-on output where the target model is now wearing the try-on cloth. Since we are constrained to train with coupled data (I_p and I_m), this expected (generated) segmentation mask (M_{exp}) is matched against the ground-truth segmentation mask (M_{gt}) itself. We intend to highlight that the network is able to generalize to unseen models at inference since it learns from a sparse clothing agnostic input (I_{priors}) that does not include any effects of worn cloth in target model image or segmentation mask (to avoid learning identity). At inference time, the generated M_{exp} is directly used downstream. Figure 5 demonstrates some examples of the corrected segmentation masks generated with our network, and ablation studies to support the use of the conditional segmentation mask is in Section 5.3.3.

The network (a 12-layer U-Net [15] like architecture) is trained with a weighted cross-entropy loss, which is the standard cross-entropy loss for semantic segmentation with increased weights for skin and background classes. The weight of the skin is increased to better handle occlusion cases, and the background weight is increased to stem bleeding of the skin pixels into the background.

3.4.2 Segmentation Assisted Texture Translation

The last stage of the framework uses the expected segmentation mask (M_{exp}), the warped product image (I^1_{stn}), and unaffected regions from the model image (I_m) to produce the final try-on image. The network is a 12-layer U-Net [15] that takes the following inputs:

- The warped product image I^1_{stn}
- The expected seg. mask M_{exp}, and
- Pixels of I_m for the unaffected regions, (Texture Translation Priors in Figure 3). E.g. face and bottom cloth, if a top garment is being tried-on.

The network produces two output images - an RGB rendered person image (I_{rp}) and a composition mask M_{cm}, which are combined with the warped product image I^1_{stn} using the following equation to produce the final try-on image:

$$I_{try-on} = M_{cm} * I^1_{stn} + (1 - M_{cm}) * I_{rp} \tag{5}$$

Because the unaffected parts of the model image are 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 [20] and [5]).

The output of the network is subject to the following matching losses based on L_1 distance and a perceptual distance based on VGG-19 activations:

$$\begin{aligned} L_{tt} &= L_{l1} + L_{percep} + L_{mask} \\ L_{l1} &= |I_{try-on} - I_m| \\ L_{percep} &= |VGG(I_{try-on}) - VGG(I_m)| \\ L_{mask} &= |M_{cm} - M^{cloth}_{gt}| \end{aligned} \tag{6}$$

The training happens in multiple phases. The first K steps of training is a conditioning phase that minimizes the L_{tt} to produce reasonable results. The subsequent phases (each lasting T steps) employ the L_{tt} loss augmented with a triplet loss (Section 3.4.3) to fine-tune the results further. This strategy further improves the output significantly (see ablation study in Section 5.3.2).

3.4.3 Duelling Triplet Loss Strategy

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 the 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). In the fine-tuning phase, at step i ($i > K$) the triplet loss is computed as:

$$\begin{aligned} i_{prev} &= K + T * (\lfloor \frac{i-K}{T} \rfloor - 1) \\ D^i_{neg} &= |I^i_{try-on} - I^{i_{prev}}_{try-on}| \\ D^i_{pos} &= |I^i_{try-on} - I_m| \\ L^i_d &= max(D^i_{pos} - D^i_{neg}, 0) \end{aligned} \tag{7}$$

Here I^i_{try-on} is the try-on image output obtained from the network with weights at the i^{th} iteration. The overall loss with the duelling triplet strategy in use is then computed for a training step i as:

$$L^i_{tryon} = \begin{cases} L_{tt} & i \leq K \\ L_{tt} + L^i_d & i > K \end{cases} \tag{8}$$

4. Experiments

4.1. Datasets

We use the dataset collected by Han et al. [5] for training and testing. It contains around 19,000 images of front-facing female models and the corresponding upper-clothing

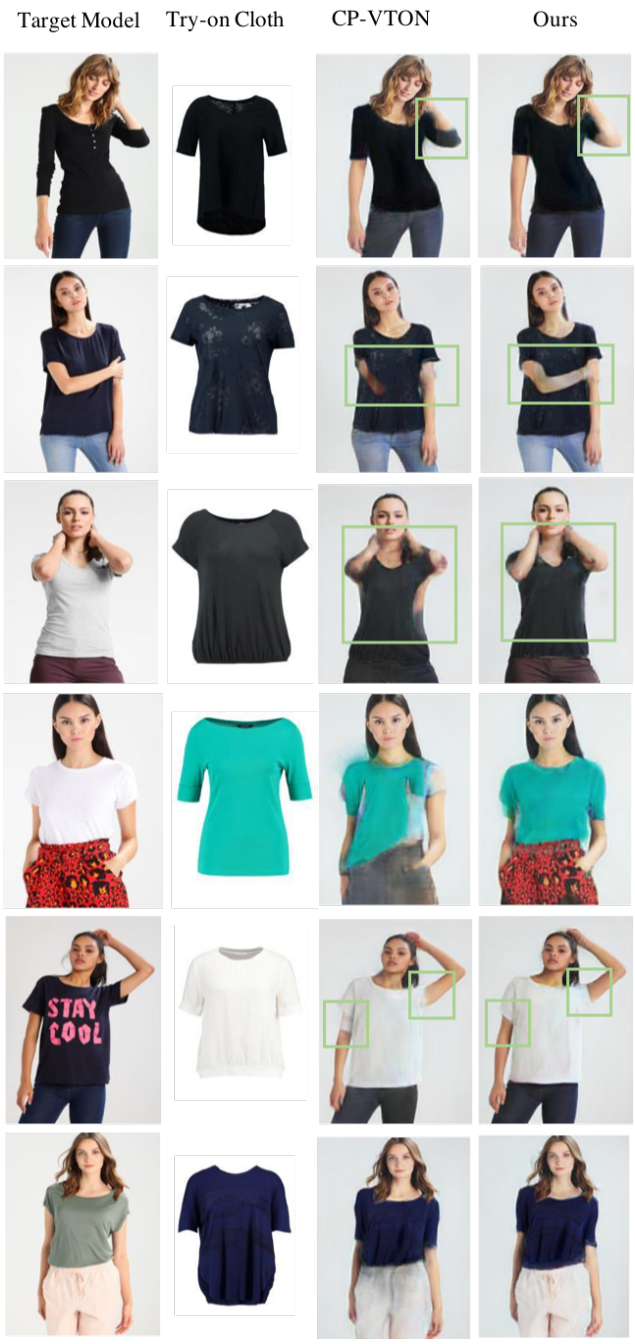


Figure 6: SieveNet can generate more realistic try-on results compared to the state-of-the-art CP-VTON.

isolated product images. There are 16253 cleaned pairs, which are split into a training set and a testing set with 14221 and 2032 pairs, respectively. The images in the testing set are rearranged into unpaired sets for qualitative evaluation and kept paired for quantitative evaluation otherwise.

4.2. Implementation Details

All experiments are conducted on 4 NVIDIA 1080Ti on a machine with 16 GB RAM. The hyper-parameter configurations were as follows: batch size=16, epochs=15, optimizer=Adam[9], lr=0.002, λ_1=λ_2=λ_3=1, λ_4=λ_5=0.5.

4.3. Quantitative Metrics

To effectively compare the proposed approach against the current state-of-the-art, we report our performance using various metrics including Structural Similarity (SSIM) [21], Multiscale-SSIM (MS-SSIM) [22], Fréchet Inception Distance (FID) [6], Peak Signal to Noise Ratio (PSNR), and Inception Score (IS) [16]. We adapt the Inception Score metric in our case as a measure of generated image quality by estimating similarity of generated image distribution to the ground truth distribution. For computing pairwise MS-SSIM and SSIM metrics, we use the paired test data.

4.4. Baselines

CP-VTON[20] and VITON [5] are the latest image based virtual try-on methods, with CP-VTON being the current state-of-the-art. In particular, [5] directly applied shape context [1] matching to compute the transformation mapping. By contrast, [20] estimates the transformation mapping using a convolutional network and has superior performance than [5]. We therefore use results from CP-VTON [20] as our baselines.

5. Results

The task of virtual try-on can be broadly broken down into two stages, *warping* of the product image and *texture transfer* of the warped product image onto the target model image. We conduct extensive quantitative and qualitative evaluations for both stages to validate the effectiveness of our contributions (coarse-to-fine warping trained with perceptual geometric matching loss, try-on cloth conditioned segmentation mask prior, and the duelling triplet loss strategy for training the texture translation network) over the existing baseline CP-VTON [20].

5.1. Quantitative Results

Table 1 summarizes the performance of our proposed framework against CP-VTON on benchmark metrics for image quality (IS, FID and PSNR) and pair-wise structural similarity (SSIM and MS-SSIM). To highlight the benefit of our contributions in warp and texture transfer, we experiment with different warping and texture transfer configurations (combining modules from CP-VTON with our modules). All scores progressively improve as we swap-in our modules. Using our final configuration of coarse-to-fine warp (C2F) and segmentation assisted texture translation with duelling triplet strategy (SATT-D) improved FID from

20.331 (for CP-VTON) to 14.65. Also, PSNR increased by around 17% from 14.554 to 16.98. While a higher Inception score (IS) is not necessarily representative of output quality for virtual try-on, we argue that the proposed approach is able to better model the ground truth distribution as it produces an IS (2.82 ± 0.09) which is closer to the IS for ground-truth images in the test set (2.83 ± 0.07) than CP-VTON (2.66 ± 0.14). These quantitative claims are further substantiated in subsequent sections where we qualitatively highlight the benefit from each of the components.

5.2. Qualitative Results

Figure 6 presents a comparison of results of the proposed framework with those of CP-VTON. The results are presented to compare the impact on different aspects of quality - skin generation (row 1), handling occlusion (row 2), variation in poses (row 3), avoiding bleeding (row 5), preserving unaffected regions (row 4), better geometric warping (row 4) and overall image quality (row 6). For all aspects, our method produces better results than CP-VTON for most of the test images. These observations are corroborated by the quantitative results reported in Table 1. More results from our framework are presented in the supplementary material.

5.3. Ablation Studies

In this section, we present a series of ablation studies to qualitatively highlight the particular impact of each our contributions: the coarse-to-fine warp, try-on product conditioned segmentation prediction and the duelling triplet loss strategy for training the texture translation module. Additional results corresponding to each of the ablation studies are included in the supplementary material.

5.3.1 Impact of Coarse-to-Fine Warp

Figure 7 presents sample results comparing outputs of the proposed coarse-to-fine warp approach against the geometric matching module used in [20]. Learning warp parameters in a multi-stage framework helps in better handling of large variations in model pose and body-shape in comparison to the single stage warp in [20]. The coarse-to-fine (C2F) warp module trained with our proposed perceptual geometric matching loss does a better job at preserving textures and patterns on warping. This is further corroborated through the quantitative results in Table 1 (row 2 vs row 3).

5.3.2 Impact of Duelling Triplet Loss

In Figure 9, we present sample results depicting the particular benefit of training the texture translation network with the duelling triplet strategy. As highlighted by the results, using the triplet loss for online hard negative mining in the fine-tuning stage refines the quality of the generated results.

Figure 7: Comparison of our C2F warp results with GMM warp results. Warped clothes are directly overlaid onto target persons for visual checking. C2F produces robust warp results which can be seen from the preservation of text (row 1) and horizontal stripes (row 2 and row 3) along with better fitting. GMM of CP-VTON produces highly unnatural results.

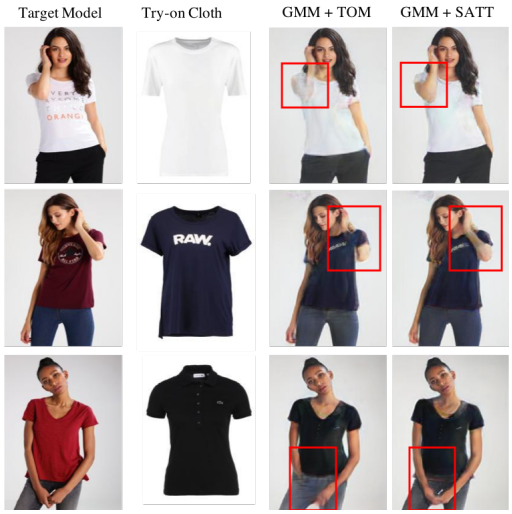


Figure 8: Using the conditional segmentation mask as prior to texture transfer aids in better handling of variable pose, occlusion and helps avoid bleeding.

This arises from better handling of occlusion, bleeding and skin generation. These observations are corroborated by results in Table 1 (row 3 vs 4).

5.3.3 Impact of Conditional Segmentation Mask Prediction

Figure 8 presents results obtained by training the texture transfer module of CP-VTON (TOM) [20] with an additional prior of the try-on cloth conditioned segmentation

Configuration	SSIM	MS-SSIM	FID	PSNR	IS
GMM + TOM (CP-VTON)	0.698	0.746	20.331	14.544	2.66 ± 0.14
GMM + SATT	0.751	0.787	15.89	16.05	2.84 ± 0.13
C2F + SATT	0.755	0.794	14.79	16.39	2.80 ± 0.08
C2F + SATT-D (SieveNet)	**0.766**	**0.809**	**14.65**	**16.98**	$\mathbf{2.82 \pm 0.09}$

Table 1: Quantitative comparison of Proposed vs CP-VTON. GMM, TOM are the warping and texture transfer modules from CP-VTON. C2F is the coarse-to-fine warp network and SATT is the segmentation assisted texture translation network we introduce in this framework. SATT-D is SATT trained with the duelling triplet loss strategy.

Figure 9: Finetuning texture translation with the duelling triplet strategy refines quality of generated images by handling occlusion and avoiding bleeding.

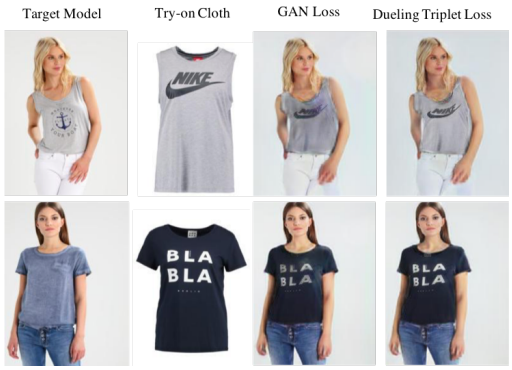


Figure 10: Proposed Duelling Triplet Loss helps in better handling of texture and avoiding blurring in generated results than the GAN Loss.

mask. 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. These observations are corroborated through results in Table 1 (row 1 vs 2).

5.3.4 Impact of Adversarial Loss on Texture Transfer

Many recent works on conditional image generation [20] [14, 11] employ a discriminator network to help improve quality of generated results. However, we observe that finetuning with the duelling triplet strategy instead results in better handling of texture and blurring in generated images without the need for any additional trainable parameters. Sample results are presented in Figure 10 to corroborate the claim.

5.3.5 Failure Cases

While SieveNet performs significantly better than existing methods, it has certain limitations too. Figure 11 highlights some specific failure cases. In some cases, generated result is unnatural due to presence of certain artifacts (as the gray neckline of the t-shirt in the example in row 1) that appear in

(a) Failure in correctly occluding the back portion of the t-shirt.

(b) Failure in predicting the correct segmentation mask owing to errors in key-point prediction.

Figure 11: Failure Cases

the output despite the correct fit and texture being achieved. This problem can be alleviated if localized fine-grained keypoints are available. Further, texture quality in try-on output may be affected by errors in the predicted conditional segmentation mask. This happens due to errors in predicting

pose key-points. For instance, this may happen in model images with low-contrast regions (example in row 2). Using dense pose information or introducing a dedicated pose prediction network can help alleviate this problem.

6. Conclusion

In this work, we propose SieveNet - a fully learnable image-based virtual try-on framework. We introduce a coarse-to-fine warp network trained with a novel perceptual geometric matching loss to better handle occlusion, variable pose whilst preserving texture. Next, we achieve accurate texture transfer using a try-on cloth conditioned segmentation mask prior and training the texture translation network with a novel duelling triplet loss strategy. We report qualitatively and quantitatively superior results over the state-of-the-art [20] for the dataset collected by Han et al. [5].

References

[1] S. Belongie, J. Malik, and J. Puzicha. Shape context: A new descriptor for shape matching and object recognition. In *Advances in neural information processing systems*, pages 831–837, 2001.

[2] A. Chopra, A. Sinha, H. Gupta, M. Sarkar, K. Ayush, and B. Krishnamurthy. Powering robust fashion retrieval with information rich feature embeddings. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[3] G. Cucurull, P. Taslakian, and D. Vazquez. Context-aware visual compatibility prediction. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[4] P. Esser, E. Sutter, and B. Ommer. A variational u-net for conditional appearance and shape generation. 2018.

[5] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. VITON: an image-based virtual try-on network. *CoRR*, abs/1711.08447, 2017.

[6] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 6626–6637. Curran Associates, Inc., 2017.

[7] M. Jaderberg, K. Simonyan, A. Zisserman, and k. kavukcuoglu. Spatial transformer networks. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 2017–2025. Curran Associates, Inc., 2015.

[8] N. Jetchev and U. Bergmann. The conditional analogy gan: Swapping fashion articles on people images. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[9] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[11] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool. Pose guided person image generation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 406–416. Curran Associates, Inc., 2017.

[12] G. Pons-Moll, S. Pujades, S. Hu, and M. Black. Clothcap: Seamless 4d clothing capture and retargeting. *ACM Transactions on Graphics, (Proc. SIGGRAPH)*, 36(4), 2017. Two first authors contributed equally.

[13] A. Pumarola, A. Agudo, A. Sanfeliu, and F. Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8620–8628, 2018.

[14] A. Raj, P. Sangkloy, H. Chang, J. Hays, D. Ceylan, and J. Lu. Swapnet: Image based garment transfer. In *ECCV*, 2018.

[15] O. Ronneberger, P. Fischer, and T. 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.

[16] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, X. Chen, and X. Chen. Improved techniques for training gans. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 2234–2242. Curran Associates, Inc., 2016.

[17] M. Sekine, K. Sugita, F. Perbet, B. Stenger, and M. Nishiyama. Virtual fitting by single-shot body shape estimation. In *Int. Conf. on 3D Body Scanning Technologies*, pages 406–413. Citeseer, 2014.

[18] A. Siarohin, E. Sangineto, S. Lathuilière, and N. Sebe. Deformable gans for pose-based human image generation. *CoRR*, abs/1801.00055, 2018.

[19] M. I. Vasileva, B. A. Plummer, K. Dusad, S. Rajpal, R. Kumar, and D. Forsyth. Learning type-aware embeddings for fashion compatibility. In *ECCV*, 2018.

[20] B. Wang, H. Zhang, X. Liang, Y. Chen, L. Lin, and M. Yang. Toward characteristic-preserving image-based virtual try-on network. *CoRR*, abs/1807.07688, 2018.

[21] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE TRANSACTIONS ON IMAGE PROCESSING*, 13(4):600–612, 2004.

[22] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multi-scale structural similarity for image quality assessment. In *in Proc. IEEE Asilomar Conf. on Signals, Systems, and Computers, (Asilomar*, pages 1398–1402, 2003.

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).

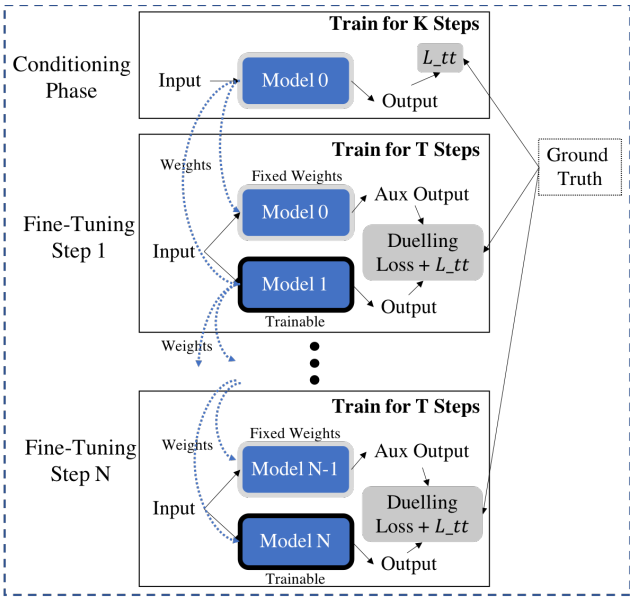


Figure 1: Visualization of the Duelling Triplet Loss strategy for training the SATT module. The model represented by thick black boundary is trainable unlike the one in the thick grey boundary. At the beginning of any fine-tuning step (say Fine Tuning Step 1 in above figure), the weights from the previous 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

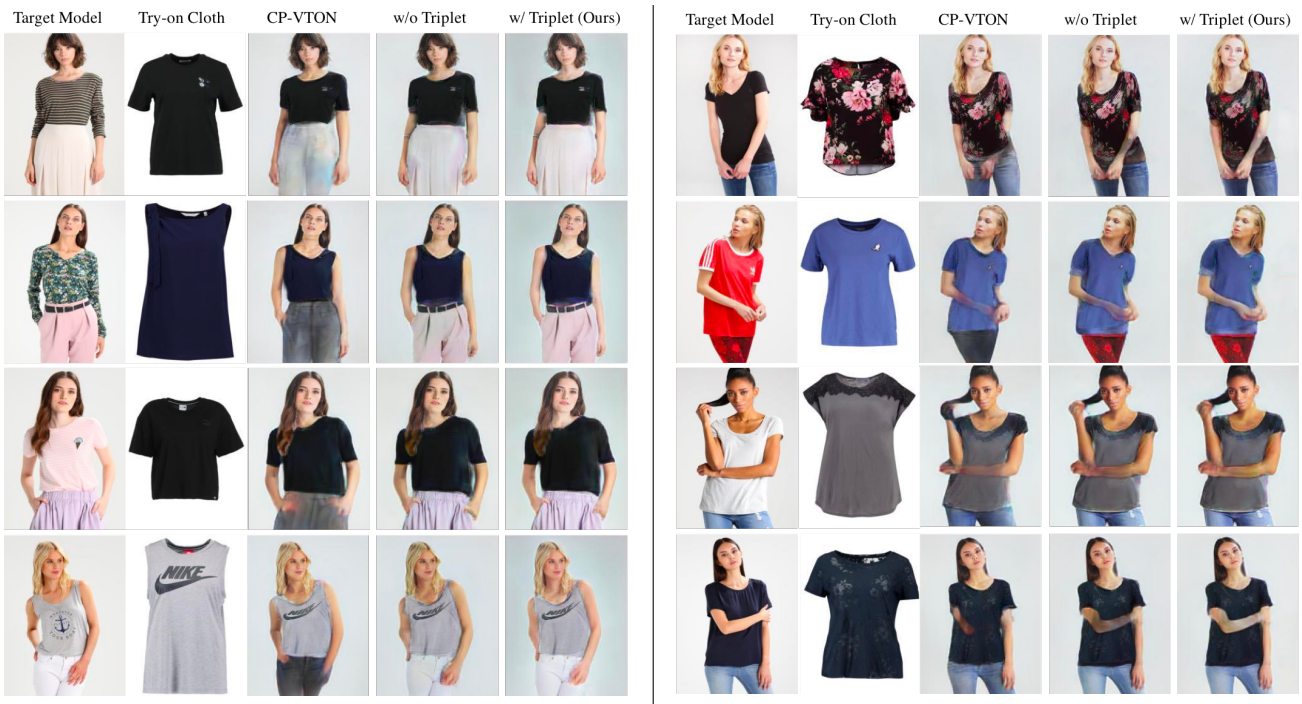
Figure 4 and 5 present additional results of comparison of the proposed SieveNet with those of CP-VTON.

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

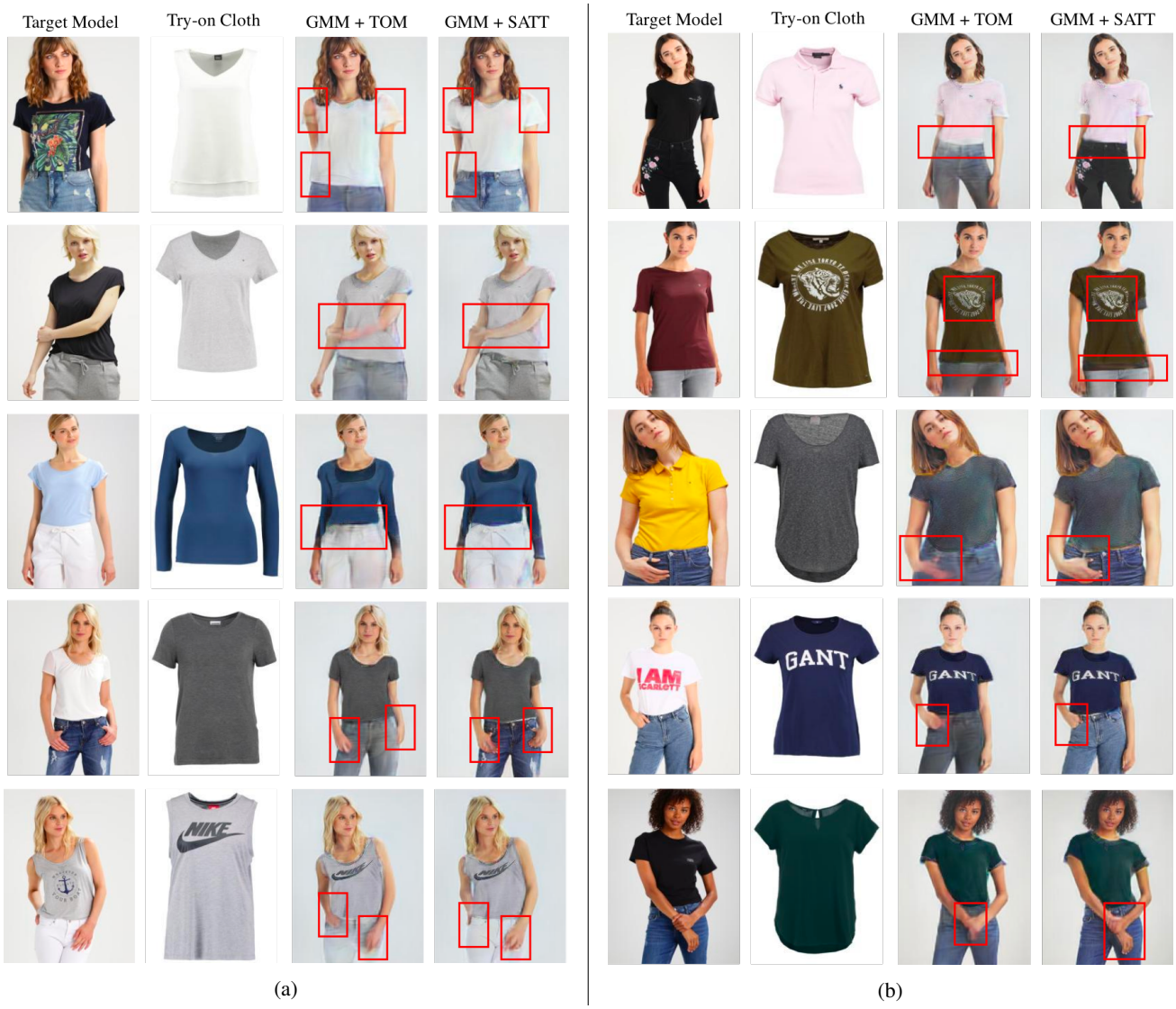
In Figure 2, we present additional results depicting the particular benefit of training the texture translation network



(a) (b)

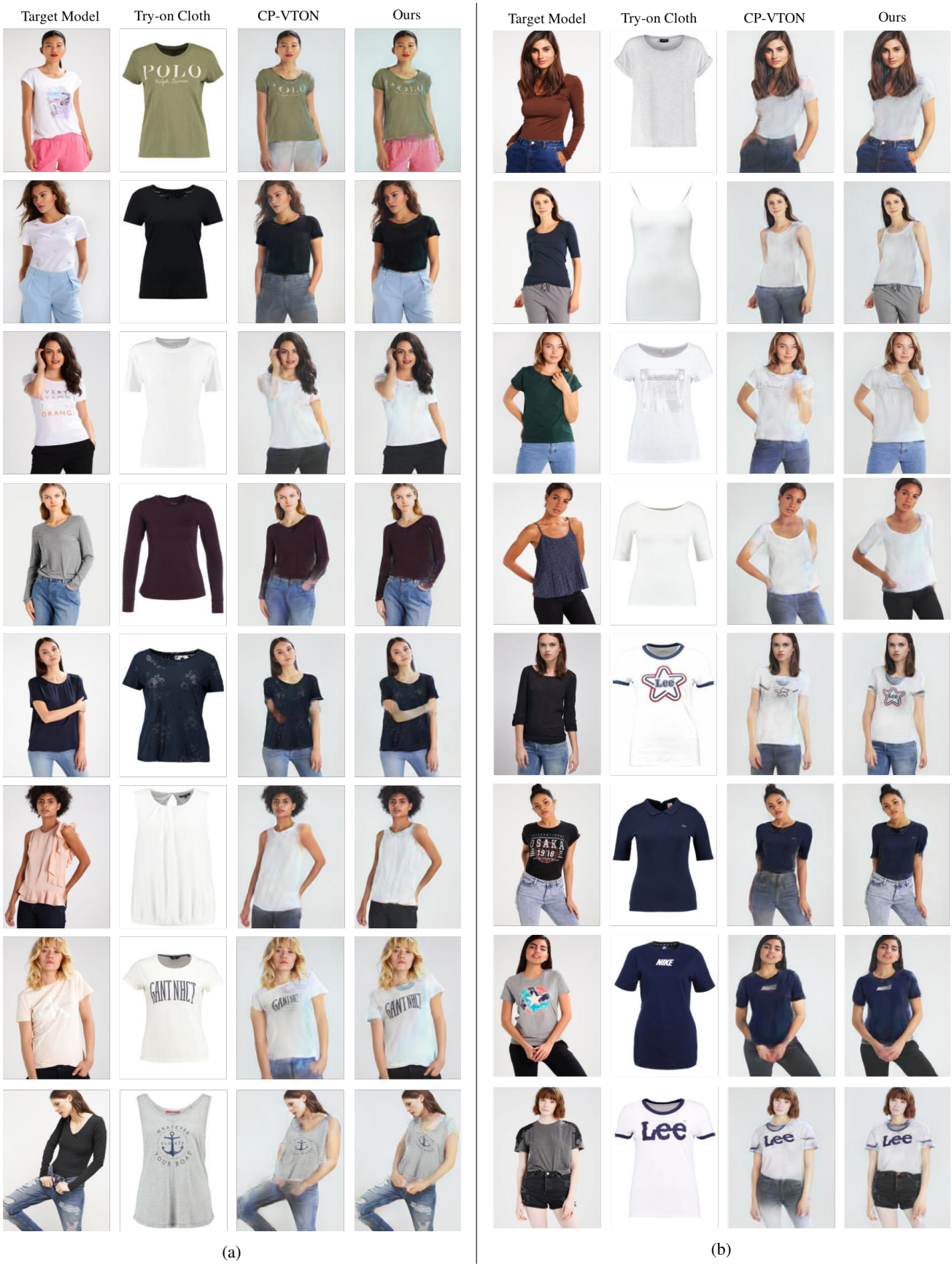
Figure 2: Fine-tuning texture translation with the duelling triplet strategy refines quality of generated images by handling occlusion and avoiding bleeding.

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.

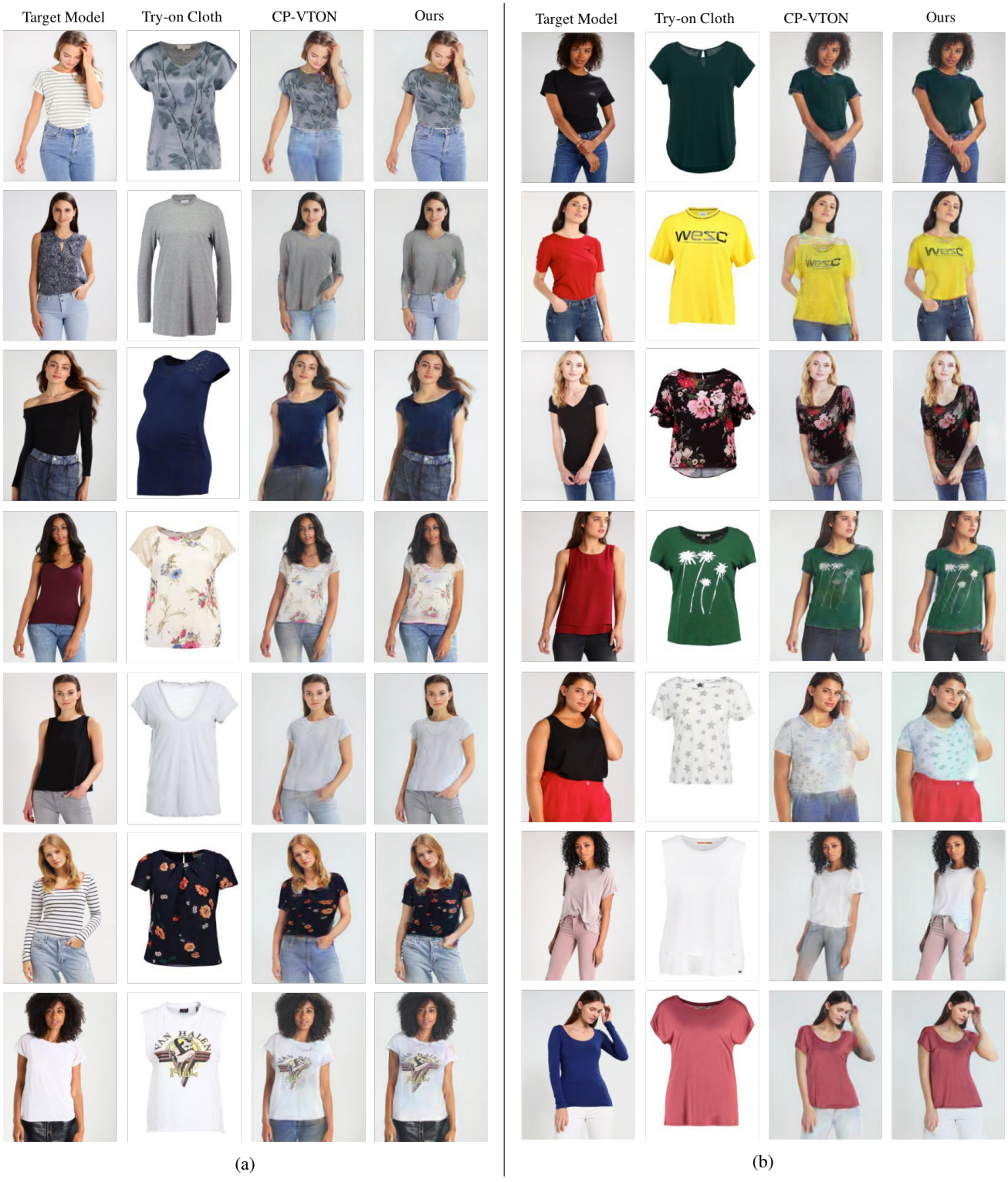


Figure 5: Comparison of SieveNet with CP-VTON. SieveNet can generate more realistic try-on results compared to the current state-of-the-art CP-VTON.