Robust Cloth Warping via Multi-Scale Patch Adversarial Loss for Virtual Try-On Framework

Kumar Ayush*
Stanford University
kayush@stanford.edu

Surgan Jandial*
IIT Hyderabad
jandialsurgan@gmail.com

Ayush Chopra*
Adobe Inc.
ayuchopr@adobe.com

Mayur Hemani
Adobe Inc.
mayur@adobe.com

Balaji Krishnamurthy
Adobe Inc.
kbalaji@adobe.com

Abstract

With the rapid growth of online commerce, image-based virtual try-on systems for fitting new in-shop garments onto a person image presents an exciting opportunity to deliver interactive customer experience. Current state-of-the-art methods achieve this in a two-stage pipeline, where the first stage transforms the in-shop cloth into fitting the body shape of the target person and the second stage employs an image composition module to seamlessly integrate the transformed in-shop cloth onto the target person image. In the present work, we introduce a multi-scale patch adversarial loss for training the warping module of a state-of-the-art virtual try-on network. We show that the proposed loss produces robust transformation of clothes to fit the body shape while preserving texture details, which in turn improves image composition in the second stage. We perform extensive evaluations of the proposed loss on the try-on performance and show significant performance improvement over the existing state-of-the-art method.

1. Introduction

Online apparel shopping has huge commercial advantages compared to traditional shopping but lacks physical apprehension. To create an interactive and real shopping environment, virtual try-on models have garnered a lot of attention recently. The traditional approach is use to computer graphics to build 3D models and render the output images since graphics methods provide precise control of geometric transformations and physical constraints. But these approaches require plenty of manual labor or expensive hardware to collect necessary information for building 3D models along with huge computations.

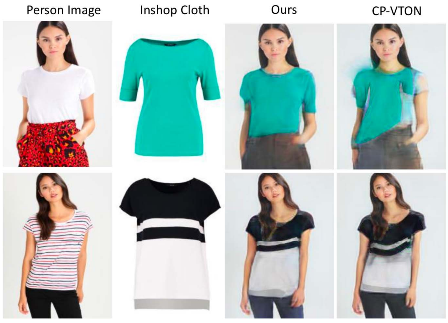

Figure 1. The proposed loss for training the GMM results in robust cloth transformations thus generating more realistic image-based virtual try-on results that preserve well key characteristics of the in-shop clothes. From the example results, it can be seen that CP-VTON fails at handling bleeding and preserving texture details. In the first row example, CP-VTON is not able to fit the cloth properly as well.

Recent image-based virtual try-on systems [3, 6] provide a more economical solution without resorting to 3D information and show promising results by reformulating it as a conditional image generation problem. Given two images, a person and an in-shop cloth, such systems aim to fit the cloth image on the person image while preserving cloth patterns and characteristics, along with realistic composition and retainment of original body shape and pose.

The best practice in image-based virtual try-on is a two-stage pipeline [3, 6]. CP-VTON [6] uses a convolutional geometric matcher (geometric matching module) which learns the deformations (i.e. thin-plate spline transform) to align the cloth with the target body shape and learns an image composition with a U-net generator.

In this work, we introduce a multi-scale patch adversarial loss to train the geometric matching module of CP-VTON. Extensive experiments show that the proposed loss han-

*Equal Contribution

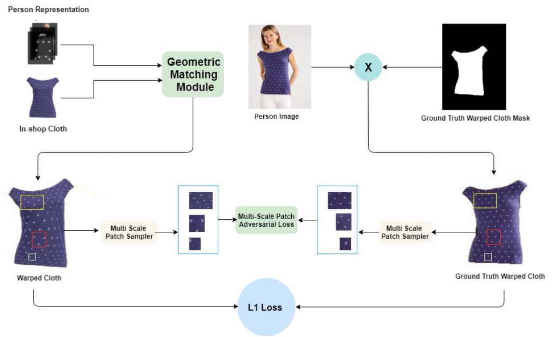
Person Representation
Geometric Matching Module
In-shop Cloth
Person Image
X
Ground Truth Warped Cloth Mask
Warped Cloth
Multi Scale Patch Sampler
Multi-Scale Patch Adversarial Loss
Multi Scale Patch Sampler
Ground Truth Warped Cloth
L1 Loss

Figure 2. An overview of our Multi-Scale Patch Adversarial Loss for training the Geometric Matching Module of CP-VTON [6]

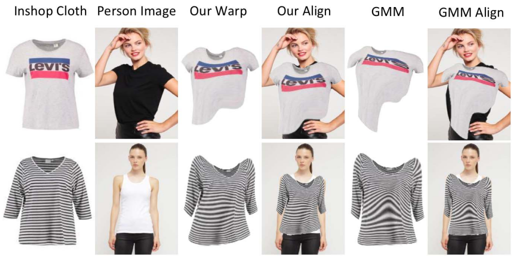

Figure 3. Warp results of GMM with and without our loss. Warped clothes are directly pasted onto target persons for visual checking. Our proposed loss produces robust warp results which can be seen from the preservation of text (first row) and horizontal stripes (second row) along with better fitting. Original GMM in CP-VTON produces highly un-natural results.

dles well the shape and pose transformations in the warping stage while preserving the texture details which in turn significantly improves the final try-on output via image composition in the second stage, achieving state-of-the-art performance on the image-based virtual try-on task.

2. Multi-Scale Patch Adversarial Loss

Inspired by the local texture loss in [9], to encourage better propagation of texture and body shape, we introduce a multi-scale patch level adversarial loss (Figure 2) to train the GMM. We randomly sample pairs of patches of multiple scales from same corresponding locations in the generated warped cloth and the ground truth warped cloth. We only sample patches which completely fall inside the cloth region. This local loss decides whether a pair of patches represent the same local region and have the same local texture in the predicted warped cloth and the ground truth warped cloth. We train a local patch discriminator to recognize a pair of cropped patches from the same corresponding regions as a positive example ($D_{patch}(.) = 1$), and a pair of patches from different regions as a negative example ($D_{patch}(.) = 0$).

Metric	orig. GMM	**GMM w/ our loss**
IS[1, 5] - val	2.705 ± 0.0958	**2.730 ± 0.0988**
IS[1, 5] - test	2.546 ± 0.084	**2.558 ± 0.127**
FID[4]	17.8396	**15.7261**
SSIM[7]	0.7151	**0.731**
MS-SSIM[8]	0.7731	**0.7901**
PSNR[2]	15.344	**16.02**

Table 1. Comparison of image synthesis performance on validation set except the second row which shows IS on test set. Additionally, corresponding to the first row, IS of ground-truth data in validation set is 2.832 ± 0.0908. Ours is more closer to the ground-truth IS.

Let $p_{1,s}(\hat{wc}, s)$ and $p_{2,s}(wc, s)$ be cropped patches of size $s \times s$ from the GMM output $\hat{wc}$ and ground truth warped cloth wc respectively. Given pairs of cropped patches of multiple scales, we define L_{ms-adv} as

$$L_{ms-adv} = -\sum_{s}\sum_{j}(D_{patch}(p_{1j,s}, p_{2j,s}) - 1)^2 \qquad (1)$$

Here, j corresponds to number of cropped patches of size $s \times s$. We use L_{ms-adv} as an additional loss along with L_1 loss for training the GMM of CP-VTON [6].

3. Dataset

We conduct our experiments on the dataset used in [3, 6]. It contains around 16253 front-view woman and top clothing image pairs. We use a train/val/test split of 14221, 2845 and 2032 pairs, respectively. The images in the test set are rearranged into unpaired pairs.

4. Results

Table 1 summarizes the performance of our proposed approach against CP-VTON on benchmark metrics for image quality [1, 5, 4, 2] and pair-wise structural similarity [7, 8]. The proposed loss function is not only able to effectively transform the in-shop cloth into fitting the body shape of the target person but is also able to preserve texture details when facing large geometric deformations (see Figure 3). Such accurate transformations while preserving texture details in the warp stage significantly improves image composition in the second stage according to both objective (Table 1) and perceptual qualities (Figure 1).

5. Conclusion

In this paper, we propose a multi-scale patch adversarial loss to train the Geometric matching module of CP-VTON [6]. Our proposed loss function produces robust cloth transformations whilst preserving the texture details which is important for image composition in the second stage of CP-VTON [6]. We demonstrate the effectiveness of our loss via a comprehensive comparison with state-of-the-art virtual try-on framework of CP-VTON [6].

References

[1] S. Barratt and R. Sharma. A note on the inception score. *arXiv preprint arXiv:1801.01973*, 2018. 2

[2] P. Gupta, P. Srivastava, S. Bhardwaj, and V. Bhateja. A modified psnr metric based on hvs for quality assessment of color images. In *2011 International Conference on Communication and Industrial Application*, pages 1–4. IEEE, 2011. 2

[3] X. Han, Z. Wu, Z. Wu, R. Yu, and L. S. Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7543–7552, 2018. 1, 2

[4] 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 *Advances in Neural Information Processing Systems*, pages 6626–6637, 2017. 2

[5] T. Salimans, I. Goodfellow, W. Zaremba, V. Cheung, A. Radford, and X. Chen. Improved techniques for training gans. In *Advances in neural information processing systems*, pages 2234–2242, 2016. 2

[6] B. Wang, H. Zheng, X. Liang, Y. Chen, L. Lin, and M. Yang. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 589–604, 2018. 1, 2

[7] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 2

[8] Z. Wang, E. P. Simoncelli, and A. C. Bovik. Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. Ieee, 2003. 2

[9] W. Xian, P. Sangkloy, V. Agrawal, A. Raj, J. Lu, C. Fang, F. Yu, and J. Hays. Texturegan: Controlling deep image synthesis with texture patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2018. 2