VITON-GAN: Virtual Try-on Image Generator Trained with Adversarial Loss

Shion Honda[1,2]

[1]The University of Tokyo
[2]The International Research Center for Neurointelligence

Figure 1: *Samples from generated images. The models in the left column virtually wear the clothes from the top row.*

Abstract

Generating a virtual try-on image from in-shop clothing images and a model person's snapshot is a challenging task because the human body and clothes have high flexibility in their shapes. In this paper, we develop a Virtual Try-on Generative Adversarial Network (VITON-GAN), that generates virtual try-on images using images of in-shop clothing and a model person. This method enhances the quality of the generated image when occlusion is present in a model person's image (e.g., arms crossed in front of the clothes) by adding an adversarial mechanism in the training pipeline.

CCS Concepts

• ***Computing methodologies*** → *Image representations;* • ***Applied computing*** → *Online shopping;*

1. Introduction

Despite the recent growth of online apparel shopping, there is a tremendous demand by consumers for buying clothes after trying them on in real shops. If e-commerce sites can offer virtual try-on images from a snapshot of the customer, they can improve their user experience.

To solve this task, previous approaches combined a U-net generator and thin plate spline (TPS) transform [HWW*18] [WZL*18]. The TPS transform keeps the patterns and letters of the clothes accurate when mapped on the human body. The latest work (CP-VTON) [WZL*18] used a human parser [GLZ*17] and pose estimator [CSWS17] in its pipeline to extract the person's representation (explanatory variable) independent of wearing clothes (objective variable). As shown in Figure 3, however, we report that these methods fail when arms are crossed in front of the clothes (occlusion), generating blurred arms due to reconstruction loss.

For generating realistic images, generative adversarial networks (GANs) have been successfully used [BDS19] [KLA18]. Unlike other generative models using reconstruction loss such as VAE, GANs are able to generate fine, high-resolution, and realistic images because adversarial loss can incorporate perceptual features that are difficult to define mathematically.

In this paper, we propose an image generator that alleviates the occlusion problem, called Virtual Try-On GAN (VITON-GAN). This generator consists of two modules, the geometry matching module (GMM) and the try-on module (TOM) as was implemented

DOI: 10.2312/egp.20191043

in CP-VTON, except adversarial loss is additionally included in the TOM to address the occlusion problem.

2. Methods

The whole training pipeline of VITON-GAN is presented in Figure 2. There are three major updates from CP-VTON. First, TOM is trained adversarially against the discriminator that uses the TOM result image, in-shop clothing image, and person representation as inputs and judges whether the result is real or fake. Second, the loss function of GMM includes the L1 distance between the generated and real images of clothes layered on the body. Finally, random horizontal flipping is used for data augmentation. The source codes and the trained model are available at https://github.com/shionhonda/viton-gan.

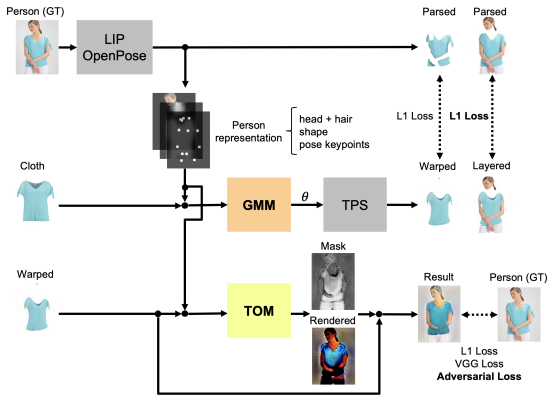

Figure 2: *Overview of the VITON-GAN training pipeline.*

3. Experiments and Results

To show the effect on the occlusion problem of VITON-GAN, a virtual try-on experiment was conducted using the same dataset as CP-VTON. The dataset contains 16,253 female model's snapshots and top clothing image pairs, which were split into 13,221, 1,000, and 2,032 pairs for training, validation, and test sets, respectively. All result images presented in this paper are from the test set.

As shown in Figure 3, VITON-GAN generated hands and arms more clearly than CP-VTON in occlusion cases. However, arm generation failed when the model's original clothing was half-sleeve and the tried-on clothing was long-sleeve (see Figure 4: upper row). This was because the TPS transform was not able to deal with topological changes that often occurred in the case of occlusion with long-sleeve shirts. Also, although in most cases VITON-GAN generated images as fine as CP-VTON (see Figure 1), it occasionally generated blurred images as shown in the lower row of Figure 4.

4. Conclusions

Here, we propose a virtual try-on image generator from 2D images of a person and top clothing that alleviates the occlusion problem. Future work will include improving the quality of generated parts of the human body and addressing topological changes in the clothes.

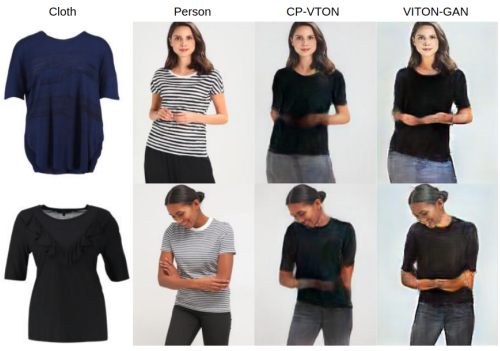

Figure 3: *Successful cases of the proposed method.*

Figure 4: *Failed cases of the proposed method.*

Acknowledgements

This work was supported by the Chair for Frontier AI Education, the University of Tokyo. Also, we would like to thank I. Sato, T. Harada, T. Mano, and C. Yokoyama for correcting this paper.

References

[BDS19] Brock A., Donahue J., Simonyan K.: Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations* (2019). 1

[CSWS17] Cao Z., Simon T., Wei S.-E., Sheikh Y.: Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017). 1

[GLZ*17] Gong K., Liang X., Zhang D., Shen X., Lin L.: Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017). 1

[HWW*18] Han X., Wu Z., Wu Z., Yu R., Davis L. S.: Viton: An image-based virtual try-on network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018). 1

[KLA18] Karras T., Laine S., Aila T.: A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948* (2018). 1

[WZL*18] Wang B., Zheng H., Liang X., Chen Y., Lin L., Yang M.: Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision* (2018). 1