Powering Virtual Try-On via Auxiliary Human Segmentation Learning

Kumar Ayush*
Stanford University
kayush@stanford.edu

Surgan Jandial*
IIT Hyderabad
jandialsurgan@gmail.com

Ayush Chopra*
Adobe Inc.
ayuchopr@adobe.com

Balaji Krishnamurthy
Adobe Inc.
kbalaji@adobe.com

Abstract

Image-based virtual try-on for fashion has gained considerable attention recently. This task requires to fit an in-shop cloth image on a target model image. An efficient framework for this is composed of two stages: (1) warping the try-on cloth to align with the body shape and pose of the target model, and (2) an image composition 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 propose to use auxiliary learning to power an existing state-of-the-art virtual try-on network. We leverage prediction of human semantic segmentation (of the target model wearing the try-on cloth) as an auxiliary task and show that it allows the network to better model the bounds of the clothing item and human skin, thereby producing a better fit. Using exhaustive qualitative and quantitative evaluation we show that there is a significant improvement in the preservation of characteristics of the cloth and person in the final try-on result, thereby outperforming the existing state-of-the-art virtual try-on framework.

1. Introduction

Various applications require solving several tasks from the computer vision domain. While each task is traditionally tackled individually, close connections between them exist. Exploiting those by solving them jointly can increase the performance of each individual task. This concept of learning several outputs from a single input simultaneously is called multi-task learning [7] and is applied to numerous tasks and techniques, including artificial neural network architectures with parameter sharing.

Building upon multi-task learning, auxiliary tasks were introduced to multi-task setups. Unlike the main tasks, which are the primary required output for an application, auxiliary tasks serve solely for learning a rich and robust common representation of an image. In the context of deep learning, the standard approach is to use a single neural network for both tasks, with shared layers followed by task-specific layers, and to apply a gradient descent-based method to minimize the weighted sum of the two losses. This leads to more meaningful representations in the shared layers and that these representations will be leveraged by the layers specific to the primary task. Auxiliary learning has often been successfully applied in other settings [8, 12, 6].

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 to use 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. Recent image-based virtual try-on systems [1, 9] 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 [1, 9]. CP-VTON [9] 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 (fusing

*Equal Contribution

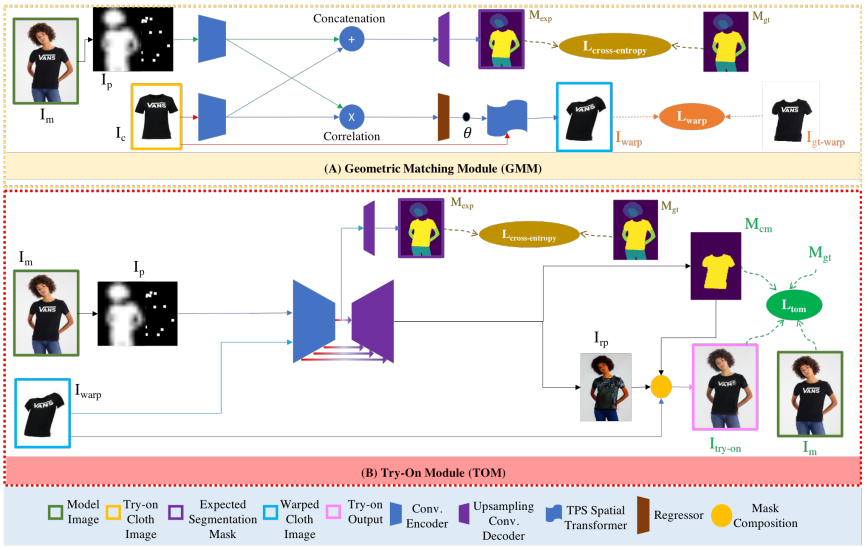


Figure 1: An overview of the proposed pipeline.

the warped cloth with the target model) with a U-net generator.

Previously, [3, 5] have successfully used multi-task learning for clothing landmark detection and fashion classification. In this work, we seek to leverage auxiliary learning to boost image-based virtual try-on. We develop on the work of CP-VTON [9] by using human semantic segmentation prediction as an auxiliary task to improve the virtual try-on performance. We propose a branched architecture to simultaneously predict the try-on result and the expected segmentation mask of the generated try-on output where the target model is now wearing the inshop cloth. A key problem with existing methods is their inability to accurately honor the bounds of the clothing item and human skin, thereby failing to produce a better fit. The cloth 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 expected segmentation mask prediction as an auxiliary task in both the Geometric Matching Module (GMM) and Try-On Module (TOM) of CP-VTON [9]. To the best of our knowledge this is the first work leveraging the paradigm of auxiliary learning in the domain of virtual try-on networks, and by significant qualitative and quantitative improvement we show that such auxiliary learning not only handles well the transformations of the inshop cloth in the warping stage while preserving the texture details and characteristics of the inshop cloth, but also improves the quality of the final try-on result via image composition in the Try-On Module (TOM). An overview of our proposed approach is schematized in Figure 1.

2. Proposed Approach

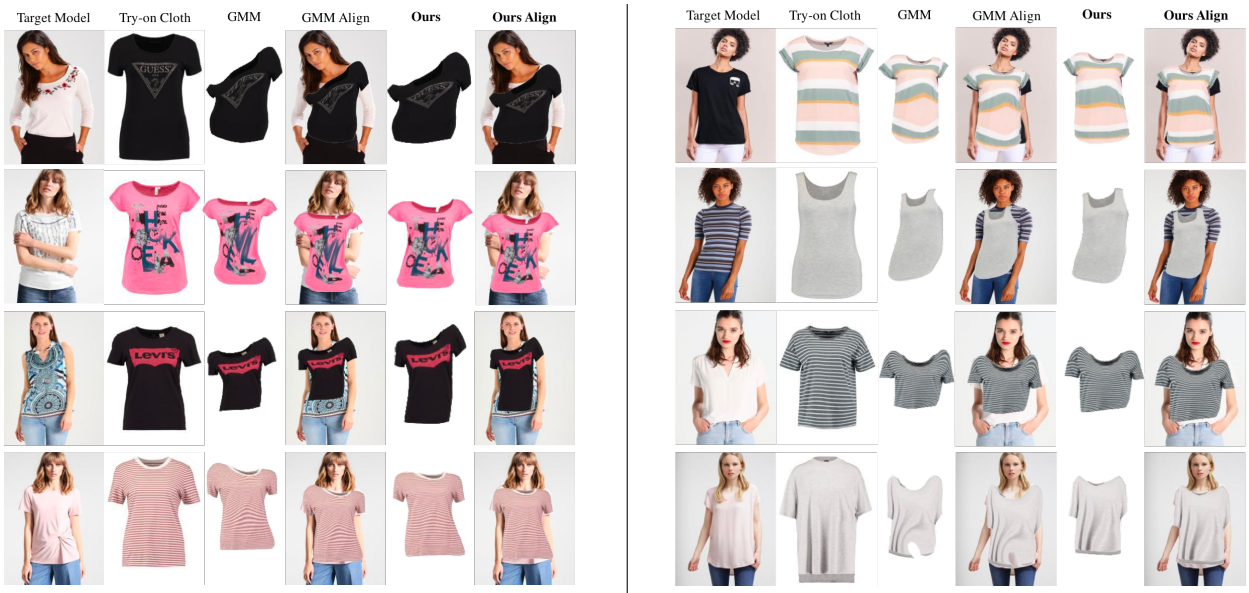
The original GMM of CP-VTON [9] consists of (1) two networks for extracting high-level features of I_p and I_c respectively, (2) a correlation layer to combine two features into a single tensor as input to the regressor network, (3) the regression network for predicting the spatial transformation parameters θ, (4) a Thin-Plate Spline (TPS) transformation module for warping an image into the output. The pipeline is end-to-end learnable and trained under the pixel-wise L1 loss between the warped result I_{warp} and ground truth $I_{gt-warp}$.

In TOM, given a concatenated input of person representation I_p and the warped cloth I_{warp}, a UNet simultaneously renders a person image I_{rp} and predicts a composition mask M_{cm}. The rendered person and the warped cloth are then fused together using the composition mask to synthesize the final try-on result I_{try-on}:

$$I_{try-on} = M_{cm} * I_{warp} + (1 - M_{cm}) * I_{rp} \quad (1)$$

L_{tom} shown in Figure 1 (B) is the same loss used in CP-VTON to train the TOM.

To encourage better propagation of texture and body shape and accurately honor the bounds of the clothing item



(a) (b)

Figure 2: Comparison of our warp results with GMM warp results. Warped clothes are directly pasted onto target persons for visual checking. Our approach produces robust warp results which can be seen from the preservation of text, patterns like horizontal stripes, etc., along with better fitting.

and human skin, we introduce expected human segmentation prediction as an auxiliary task in both GMM and TOM. In GMM, we pass the concatenated features from the two encoders through a decoder to generate the M_{exp}. Similarly, in TOM, we pass the features from the encoder of the UNet through another decoder to predict M_{exp}. We use weighted cross-entropy loss $L_{cross-entropy}$ as the auxiliary loss in both the modules, 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.

These auxiliary losses allow the encoders in GMM and TOM (shared with their respective auxiliary modules) to learn meaningful and robust features that are leveraged by the warp and try-on specific layers to produce high quality warp and try-on results respectively.

3. Dataset

We conduct our experiments on the dataset used in [1, 9]. It contains around 16253 front-view woman and top-clothing image pairs. We use a train/test split of 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. We use [4] to get human semantic segmentation mask as pseudo ground truth for our auxiliary task.

4. Results

Table 1 summarizes the performance of our proposed framework against CP-VTON on benchmark metrics for image quality (FID and PSNR) and pair-wise structural similarity (SSIM and MS-SSIM). Using human segmentation prediction as an auxiliary task in the warping module (GMM) allows the module to effectively transform the in-shop cloth into fitting the body shape of the target person and preserve texture details when facing large geometric deformations (Figure 2). Such accurate transformations in the warp stage is important for image composition with the person image in TOM. The same human segmentation mask prediction in the TOM stage helps in improving the quality of try-on results according to both objective (Table 1) and perceptual qualities (Figure 3). Additionally, to highlight the benefit of our approach, we perform an ablation analysis and experiment with auxiliary learning only in the TOM module. Scores and image-synthesis quality progressively improve as we swap-in our auxiliary module in the GMM as well.

5. Conclusion and Future Work

In this paper, we propose to exploit human semantic segmentation prediction as an auxiliary task to facilitate the performance of an existing virtual try-on framework. Our proposed approach produces robust cloth transformations to fit the body shape and pose whilst preserving the tex-

(a) (b)

Figure 3: The proposed approach results in robust cloth transformations and generates more realistic image-based virtual try-on results that preserve well key characteristics of the in-shop clothes and fit the body well. From the example results, it can be seen that CP-VTON fails at fitting the cloth properly, handling bleeding, occlusion and preserving texture details.

Metric	CP-VTON	only TOM	**Both (Ours)**
FID [2]	20.31	19.87	**18.93**
SSIM [10]	0.698	0.701	**0.712**
MS-SSIM [11]	0.744	0.750	**0.767**
PSNR	14.54	14.64	**15.19**

Table 1: Quantitative comparison of CP-VTON vs. Auxiliary learning only in TOM vs Auxiliary learning in both GMM and TOM (Ours).

ture details and improves image composition to produce good quality try-on results. We demonstrate the effectiveness of an assistive auxiliary task in the domain of virtual try-on through significant quantitative and qualitative improvement in the performance of CP-VTON.

References

[1] 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, 3

[2] 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. 4

[3] P. Li, Y. Li, X. Jiang, and X. Zhen. Two-stream multi-task network for fashion recognition. *arXiv preprint arXiv:1901.10172*, 2019. 2

[4] X. Liang, K. Gong, X. Shen, and L. Lin. Look into person: Joint body parsing & pose estimation network and a new benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018. 3

[5] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 2

[6] T. Mordan, N. Thome, G. Henaff, and M. Cord. Revisiting multi-task learning with rock: a deep residual auxiliary block for visual detection. In *Advances in Neural Information Processing Systems*, pages 1310–1322, 2018. 1

[7] S. Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017. 1

[8] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning*, pages 1096–1103. ACM, 2008. 1

[9] 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, 3

[10] 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. 4

[11] 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. 4

[12] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014. 1