# CP-VTON+: Clothing Shape and Texture Preserving Image-Based Virtual Try-On

Matiur Rahman Minar[1], Thai Thanh Tuan[1], Heejune Ahn[1], Paul L. Rosin[2], and Yu-Kun Lai[2]

[1]Seoul National University of Science and Technology, South Korea
[2]Cardiff University, UK

## Abstract

*Recently proposed Image-based virtual try-on (VTON) approaches have several challenges regarding diverse human poses and clothing styles. First, clothing warping networks often generate highly distorted and misaligned warped clothes, due to the erroneous clothing-agnostic human representations, mismatches in input images for clothing-human matching, and improper regularization transform parameters. Second, blending networks can fail to retain the remaining clothes due to the wrong representation of humans and improper training loss for the composition-mask generation. We propose CP-VTON+ (Clothing shape and texture Preserving VTON) to overcome these issues, which significantly outperforms the state-of-the-art methods, both quantitatively and qualitatively.*

## 1. Introduction

Due to the difficulty and high costs in 3D model based approaches, 2D image-based VTON technologies are getting popular nowadays. Among different system settings in the previous works, one with an image of try-on clothing and a target human image has been considered practical in many works [1, 4]. A common processing pipeline for this setting has two stages: first, the try-on clothing is warped to align with the target human (called the Geometric Matching Module (GMM)), and then the warped clothing is blended with the target human image (called the Try-On Module (TOM). We also use this setting.

Previous works demonstrated the feasibility of image-based VTON technologies. However, as Figure 1 shows, they work fairly well for the cases of mono-colored short-sleeved clothes and up-front poses, but not for cases with rich-textured or long-sleeved clothing, or a diversely posed human. We reveal the origins of these problems and redesign the pipeline by employing better input representations and using improved training cost functions. First, we correct the erroneous clothing-agnostic human representation: wrong labeling of the chest area in human parsing maps, and omission of clothing from reserved areas in the human representation. Second, we observe the problems in clothing warping networks: the unbalanced geometric matching inputs and training loss function. Finally, we improve the composition mask using the input clothing mask and a concrete loss function. Our proposed system, named CP-VTON+ after the baseline CP-VTON [4], outperforms CP-VTON by large margins, in both perceptible and subjective evaluations.

## 2. Related Works

VITON [1] first proposed the system setting and dataset of an in-shop clothing and a target human image. VITON also first used the two stage architecture (a warping and a blending module) and CP-VTON [4] refined VITON for improving the clothing texture transfer, where the clothing area is blended with the warped clothing generated from the original clothing image, not reconstructing through a decoder network. We include [1, 4] in our comparisons, since their implementations are publicly available.

## 3. CP-VTON+

### 3.1. Overview

Our new VTON pipeline is designed based on the pipeline structure of CP-VTON [4], hence named CP-VTON+. Figure 2 illustrates the architecture.

#### 3.1.1 Clothing Warping Stage

The improvement of the GMM stage is in three aspects. First, it is crucial to obtain the complete target body silhouette area from the target human image. However, in the VITON dataset, the neck and bare chest area is wrongly la-


Figure 1. Qualitative evaluation of image-based VTONs (new clothing try-on): left to right, input pairs, VITON results, CP-VTON results, CP-VTON+ (ours) results). We present examples of different type of clothes here: from sleeveless to short, half and long sleeve, with different human body shapes/poses. Our method correctly preserves original clothing shapes, textures, and target human reserved regions.

beled as background and the body shape is often distorted by hair occlusion. We correct this as follows: A new label 'skin' is added to the label set, and then the label of the corresponding area is restored from the wrong label, 'background', considering the original human image and joint locations. The skin-labeled area is now included in the silhouette in the human representation. To recover the hair occlusion over the body, first the hair occlusion areas are identified as the intersection of the convex contour of the upper clothing and the hair-labeled area, and the intersections are re-labeled as upper cloth.

Second, the CP-VTON GMM network is built on CNN geometric matching [2]. Whereas the CNN geometric matching uses a pair of color images, CP-VTON GMM inputs are binary mask information, silhouette, and joint heatmap and the colored try-on clothing. Since the colored texture from try-on clothing does not help in the matching process, our GMM uses a clothing mask $M_{C_i}$ instead of colored $C_i$, i.e.,

$$\theta = f_\theta(f_H(H_t), f_C(M_{C_i})) \tag{1}$$

Finally, the experiments with existing methods reveal that warped clothing is often severely distorted. We could not clearly determine the reason, but can conclude that the estimation of the TPS parameters needs regularization, taking into account the restriction of clothing textures. Our grid warping regularization is defined on the grid deformation and not directly on the TPS parameters for easy visualization and understanding, so as to not have too much different warping from the before and next grid-gap in equation 3.

$$L_{GMM}^{CP-VTON+} = \lambda_1 \cdot L1(C_{warped}, I_{C_t}) + \lambda_{reg} \cdot L_{reg} \tag{2}$$

$$\begin{aligned} L_{reg}(G_x, G_y) = &\sum_{i=-1,1} \sum_x \sum_y |G_x(x+i, y) - G_x(x, y)| \\ &\sum_{j=-1,1} \sum_x \sum_y |G_x(x, y+j) - G_x(x, y)| \end{aligned} \tag{3}$$

### 3.1.2 Blending Stage

Improvements to the TOM stage are three fold. Firstly, in order to retain the other human components other than the target clothing area, all the other areas, e.g., face, hair, lower clothes and legs are added to the human representation input of TOM. Secondly, in the mask loss term in the TOM loss function, we replace the Composition Mask with the supervised ground truth mask for a strong alpha mask.

$$\begin{aligned} L_{TOM}^{CP-VTON+} = &\lambda_1 ||I_0 - I_{GT}||_1 + \lambda_{VGG} L_{VGG} \\ &+\lambda_{mask} ||M_{GT} - M_o||_1 \end{aligned} \tag{4}$$

Finally, we added the binary mask of warped clothing to the TOM network input, since TOM could not recognize the white clothing area as being the same as the in-shop clothing image background.

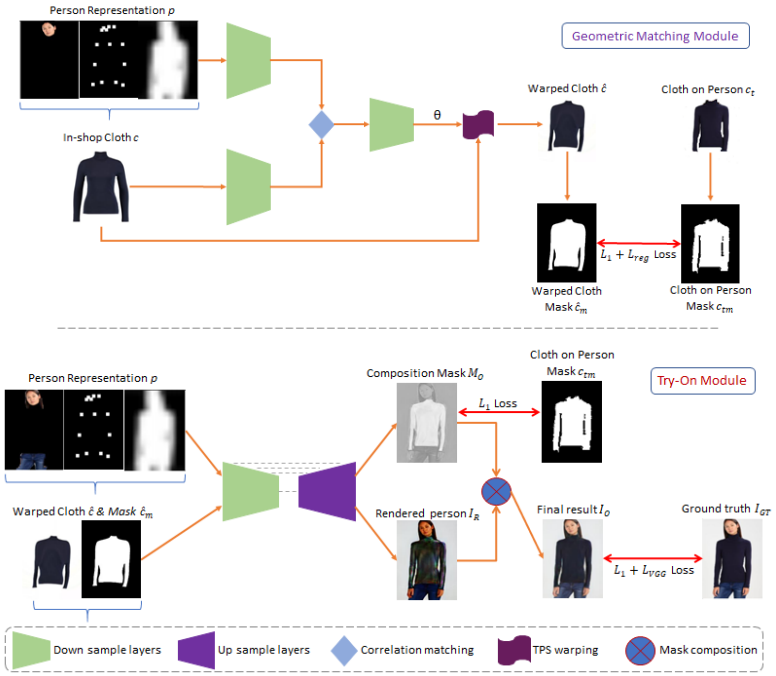


Figure 2. Full pipeline of our proposed CP-VTON+.

# 4. Experiments and Results

## 4.1. Implementation Details

We extended the existing CP-VTON implementation[1] as described above. We added automatic refinement for segmentation. For training, we used similar setting as [4] for comparison, i.e., $\lambda_1 = \lambda_{VGG} = \lambda_{mask} = 1$ and $\lambda_{reg} = 0.5$. We used the VITON clothing-human pair dataset for all experiments. We trained both networks for 200K steps with batch size 4, with the Adam optimizer with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The learning rate was first fixed at $0.0001$ for 100K steps and then linearly decays to zero for the remaining steps.

## 4.2. Results and Comparison

Table 1 shows numerical comparison between the baseline CP-VTON [4] and different versions of our proposed CP-VTON+. We use IoU, SSIM [5] and Learned Perceptual Image Patch Similarity (LPIPS) [6] metrics for the same clothing retry-on cases (when we have ground truths) for the warping stage and the blending stage, respectively. The original target human image is used as the reference image for SSIM and LPIPS (lower score means better), and the parsed segmentation area for the current upper clothing is used as the IoU reference. For different clothing try-on (where no ground truth is available), we used the Inception Score (IS) [3]. Our proposed CP-VTON+ outperforms CP-VTON on all measures.

We present visual comparison among VITON [1] and CP-VTON [4] (please refer to Figure 1). Subjective evaluation shows significant visual improvements: 1) the warped clothes do not have severe distortion, 2) the lower clothes are retained, 3) the new clothing collar shape is not affected by the current clothing's collar shape, and 4) clothing textures such as logos and patterns are clearer.

## 4.3. Ablation Study

In figure 3, we highlight the impacts of the identified problems and improvements of the proposed method step-by-step through the ablation study of CP-VTON+. First and second columns are target humans and try-on clothes, respectively. The third column has the CP-VTON [4] results. The fourth column shows when unchanged clothes and body parts are added to the reserved region inputs of TOM, retaining the original pants texture. The fifth column is when the mask loss function of TOM is updated with the target clothing area, making the texture and color of clothing sharp and vivid. Finally, the sixth/last column is when the body masks are updated, where the skin areas wrongly labeled as background and hair are removed from the reserved region input of GMM, having GMM with a better clothing-and-hair-agnostic human representation.

## 4.4. Discussions

Although CP-VTON+ improves the quality, it does not always produce successful results for long sleeved, com-

[1] https://github.com/sergeywong/cp-vton

| Method | Warped (IoU) | Blended SSIM | Blended LPIPS | Blended IS (mean ± std.) |
|---|---|---|---|---|
| CP-VTON[4] | 0.7898 | 0.7798 | 0.1397 | 2.7809 ± 0.0594 |
| CP-VTON+ (w/o GMM regularization & mask loss) | 0.7602 | 0.8076 | 0.1263 | 3.0735 ± 0.0531 |
| CP-VTON+ (w/o GMM mask loss) | 0.7920 | 0.8077 | 0.1231 | **3.1312 ± 0.0837** |
| CP-VTON+ (Ours) | **0.8425** | **0.8163** | **0.1144** | 3.1048 ± 0.1068 |

Table 1. Quantitative comparison results between the state-of-the-art CP-VTON [4] and our proposed CP-VTON+. All evaluation metrics are measured on the testing dataset from Han et al. [1], where input pairs are same-clothing for IoU SSIM [5] & LPIPS [6], and different-clothed for IS [3].

Figure 3. Ablation study of CP-VTON+. From left to right column: target humans, try-on clothes, CP-VTON, with corrected human representation in TOM, warped clothing mask and mask loss function updated, and CP-VTON+ (with corrected human representation in GMM)

plicated shaped, or textured clothing, and target humans with complex postures. Figure 4 shows two typical failure cases due to the clothing warping, where arms cover the body area, warped clothing does not match the human body, and TOM fails in hiding the warping error. Any 2D transform including the non-rigid TPS algorithm cannot handle the strong 3D deformations of clothing. Also, 3D poses induce self-occlusions. The TOM network should recognize the clothing area and skin areas, like naked arms. Sometimes, even for simple poses, warped clothing shows unrealistic results. Additionally, better human parsing is crucial for better try-on results.

## 5. Conclusion

We proposed a refined image-based VTON system, CP-VTON+, solving issues in previous approaches: errors in human representation and the dataset, network design and loose cost function. Even though CP-VTON+ improves the performance, we find that a 2D image-based approach has inherent limitations for coping with diversely posed target


Figure 4. Failures of our CP-VTON+

human cases. Therefore, the application would be limited to simple clothing and standard posed target humans. For more diverse cases, 3D reconstruction would be more suitable.

## References

[1] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. Viton: An image-based virtual try-on network. *CVPR*, pages 7543–7552, 2018. 1, 2, 3, 4

[2] Ignacio Rocco, Relja Arandjelovic, and Josef Sivic. Convolutional neural network architecture for geometric matching. In *CVPR*, pages 6148–6157, 2017. 2

[3] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *NeurIPS*, pages 2234–2242, 2016. 3, 4

[4] Bochao Wang, Hongwei Zhang, Xiaodan Liang, Yimin Chen, Liang Lin, and Meng Yang. Toward characteristic-preserving image-based virtual try-on network. In *ECCV*, 2018. 1, 2, 3, 4

[5] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 13(4):600–612, 2004. 3, 4

[6] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018. 3, 4