Style-Based Global Appearance Flow for Virtual Try-On
Supplementary Material

Sen He, Yi-Zhe Song, Tao Xiang
Center for Vision, Speech and Signal Processing, University of Surrey
iFlyTek-Surrey Joint Research Centre on Artificial Intelligence
{sen.he,y.song,t.xiang}@surrey.ac.uk

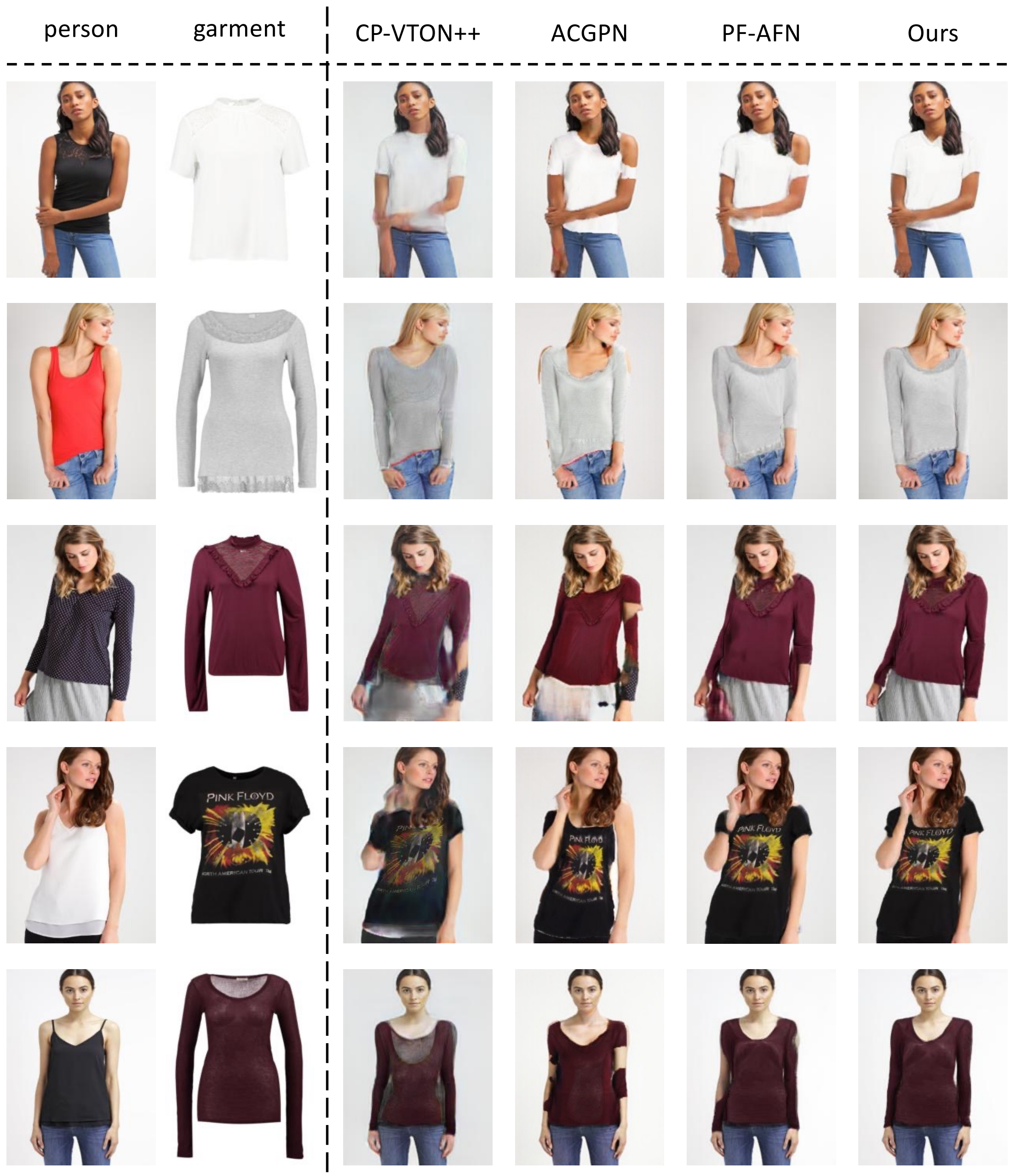

Figure 1. More qualitative results from different models (CP-VTON++ [8], ACGPN [11], PF-AFN [2] and ours) on VITON testing dataset.

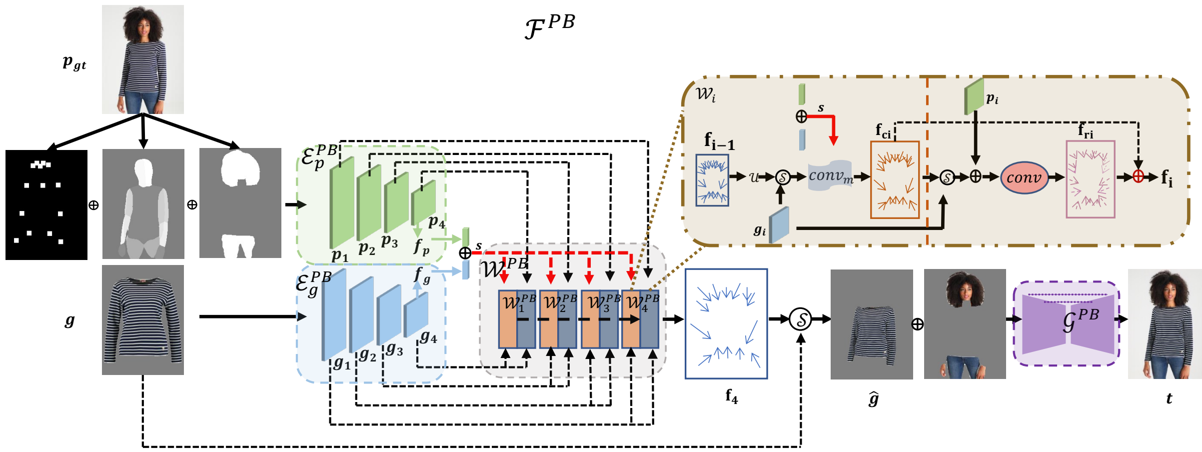

Figure 2. A schematic of our parser-based model $\mathcal{F}^{PB}$.

1. Introduction

This supplementary material provides: (1) more qualitative results from different models; (2) the training details of the parser based model $\mathcal{F}^{PB}$.

2. More Qualitative Results

More qualitative results from different models are illustrated in Fig. 1. Overall, our model generates better try-on images.

3. Training Details of Parser Based Model

The architecture of $\mathcal{F}^{PB}$ is illustrated in Fig. 2. It shares the same inner architecture with $\mathcal{F}$. The only difference is that the person encoder in $\mathcal{F}^{PB}$ takes as inputs de-clothed person representation (pose, dense pose and human segmentation map) and its generator takes as inputs the masked person image and the warped garment.

$\mathcal{F}^{PB}$ is trained with paired person and garment images. More specifically, we use the off-the-shelf pose detection model [1], dense pose model [5] and the human parser [4] to extract the pose, dense pose and human segmentation map for the person image. These extracted representations subsequently are concatenated and then fed into the person encoder. All other information flows are the same as that in the parser free model $\mathcal{F}$.

Similar to $\mathcal{F}$, $\mathcal{F}_{PB}$ is trained with three losses.

We first apply a perceptual loss [7] between the output of $\mathcal{F}^{PB}$ and the ground truth person image p_{gt}:

$$L_p = \sum_i \|\phi_i(t) - \phi_i(p_{gt})\|, \tag{1}$$

where ϕ_i is the i^{th} block of the pre-trained VGG network [9].

To supervise the training of the warping model $\mathcal{W}^{PB}$, we apply a loss on the warped garment:

$$L_g = \|\hat{g} - m_g \cdot p_{gt}\|, \tag{2}$$

where m_g is the garment mask of p_{gt} predicted by the off-the-shelf human parsing model [4].

As per standard in previous appearance flow methods [3,6], we also apply a smoothness regularization on the predicted flow from each block in $\mathcal{W}$:

$$L_R = \sum_i \|\nabla \mathbf{f_i}\|, \tag{3}$$

where $\|\nabla \mathbf{f_i}\|$ is the generalized charbonnier loss function [10].

The overall learning objective is:

$$L = \lambda_p L_p + \lambda_g L_g + \lambda_R L_R, \tag{4}$$

where λ_p, λ_g ,and λ_R denote the hyperparameters for balancing the three objectives.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *CVPR*, 2017. 2

[2] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *CVPR*, 2021. 1

[3] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, 2021. 2

[4] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In *CVPR*, 2017. 2

[5] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos. Densepose: Dense human pose estimation in the wild. In *CVPR*, 2018. 2

[6] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, 2019. 2

[7] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016. 2

[8] Matiur Rahman Minar, Thai Thanh Tuan, Heejune Ahn, Paul Rosin, and Yu-Kun Lai. Cp-vton+: Clothing shape and texture preserving image-based virtual try-on. In *CVPRW*, 2020. 1

[9] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 2

[10] Deqing Sun, Stefan Roth, and Michael J Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *IJCV*, 106(2):115–137, 2014. 2

[11] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, 2020. 1