Supplementary Materials

1. User Study

The user study evaluates both the realism and accuracy of the generated image conditioned on correct layout. Because we want only aim to evaluate the image generator, we use the layout produced by the pre-trained human parser on the ground truth image (rather than the ones produced by the semantic layout generator). Because produced layout are often noisy, we choose examples where the layout is good, giving a fair representation of the top 20 percentile of our results.

We ran our user study with two populations of participants (vision researchers tested with Form B, and randomly selected people tested with Form A). Images are displayed using the highest possible resolution (512x512) and each participant is primed with two real and fake examples before the study. Each participant is then tested with 50 examples (25 real and 25 fake), without repeating products. During the study, the garment image and the model image are shown side-by-side, giving subject an easy way to determine whether the synthesized image accurately represent the garment – an important property for fashion e-commerce application.

For every model, we tested swapping a single garment. Each model in our dataset has a ground truth paired with only one garment (the other garments worn by the model are not matched with any product images). Both the real and fake images are shown with the same outfit. Form A and Form B mirror each other (i.e. if a garment is shown as the generated version in Form B, the real image will be shown in Form A).

The raw questions and responses are provided under the folder “user study”.

2. Semantic Layout and Pose Representation

2.1. Semantic Layout

We obtain our semantic layout using an off-the-shelf human parser [3]. Our Semantic Layout has 15 semantic labels to represent different human parts and garments. These are background, hair, face, neckline, right arm, left arm, right shoe, left shoe, right leg, left leg, bottoms, full-body, tops, outerwear, and bags. Among these semantic labels, bottoms, full-body, tops, and outerwear are garment labels. The semantic segmentation mask is $m \in R^{H \times W \times 15}$, where H and W correspond to the width and height of the image.

2.2. Incomplete Layout

During training, we create the incomplete layout m_i by hiding the target garment labels and relevant skin labels (by setting these labels to the background class). For tops and outerwear, we hide tops, outerwear, left arm, right arm and neckline; for bottoms, we hide bottoms, left leg and right leg; for full-body, we hide full-body, left leg, right leg, left arm, right arm and neckline. All original channels are still outputted as the incomplete layout ($m_i \in R^{H \times W \times 15}$).

Figure 1. A version of the Figure 1 without bounding boxes.

2.3. Pose Representation

We first apply the pre-trained Openpose [1, 6] model on the model images to obtain 18 key points. Following prior work [5, 2, 7], we convert the key points into a heatmap with 18 channels. Our key point heatmap becomes $p \in R^{H \times W \times 18}$. In each channel, we draw a 5×5 square centered at each keypoint's coordinate and set all values in the square to one. If a key point is missing, the channel will be set to all zeros.

3. Experiment Setups

3.1. Network Architecture

For both the semantic layout generator G_{layout} and the inpainting network of the multi-warp garment generator $G_{garment}$, we use a U-Net of 5 hidden layers. The channel sizes for the hidden layers are 64, 128, 256, 512, and 1,024 respectively. We downsample the image size by 2 at each layer, using bilinear interpolation.

For the warper module of the multi-warp garment generator $G_{garment}$, we use a ResNet-18 model with ImageNet pre-trained weights as the backbone.

3.2. Training Procedure

We train our network using Adam Optimizer with a learning rate of $1e^{-4}$ for the semantic layout generator G_{layout} and $2e^{-4}$ for the multi-warp garment generator $G_{garment}$. Both networks are trained on a Quadro RTX 6000 GPUs (24GB). G_{layout} is trained for 50k steps with a batch size of 16. $G_{garment}$ is trained for 100k steps with a batch size of 8.

For training the G_{layout}, λ_1 and λ_2 are set to 1 and 0.2 respectively. For training the $G_{garment}$, γ_1, γ_2, γ_3 and γ_4 are set to 5, 5, 3 and 1, respectively.

4. More Qualitative Comparisons

We show more qualitative comparisons between our method and O-VITON [4]. Notice O-VITON [4] is the only prior work that supports multi-garment try-on, but they did not release their dataset or implementation. For a fair comparison, we found garment images that most closely resemble the garments chosen in [4] in terms of style, color, and texture. Image results for O-VITON are directly taken from their paper and supplementary materials. Notice the substantial improvement in generation quality in Figure 2.

References

[1] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 1

[2] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R. Scott. Clothflow: A flow-based model for clothed person generation. In *ICCV*, 2019. 1

[3] Peike Li, Yunqiu Xu, Yunchao Wei, and Yi Yang. Self-correction for human parsing. *arXiv preprint arXiv:1910.09777*, 2019. 1

[4] Assaf Neuberger, Eran Borenstein, Bar Hilleli, Eduard Oks, and Sharon Alpert. Image based virtual try-on network from unpaired data. In *CVPR*, 2020. 2, 3

[5] Bochao Wang, Huabin Zheng, Xiaodan Liang, Yimin Chen, and Liang Lin. Toward characteristic-preserving image-based virtual try-on network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018. 1

[6] Shih-En Wei, Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Convolutional pose machines. In *CVPR*, 2016. 1

[7] 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

Figure 2. Qualitative comparison with O-VITON [4]. The top two rows in each cell show the garments in the outfit and the bottom row in each cell shows generated try-on results.

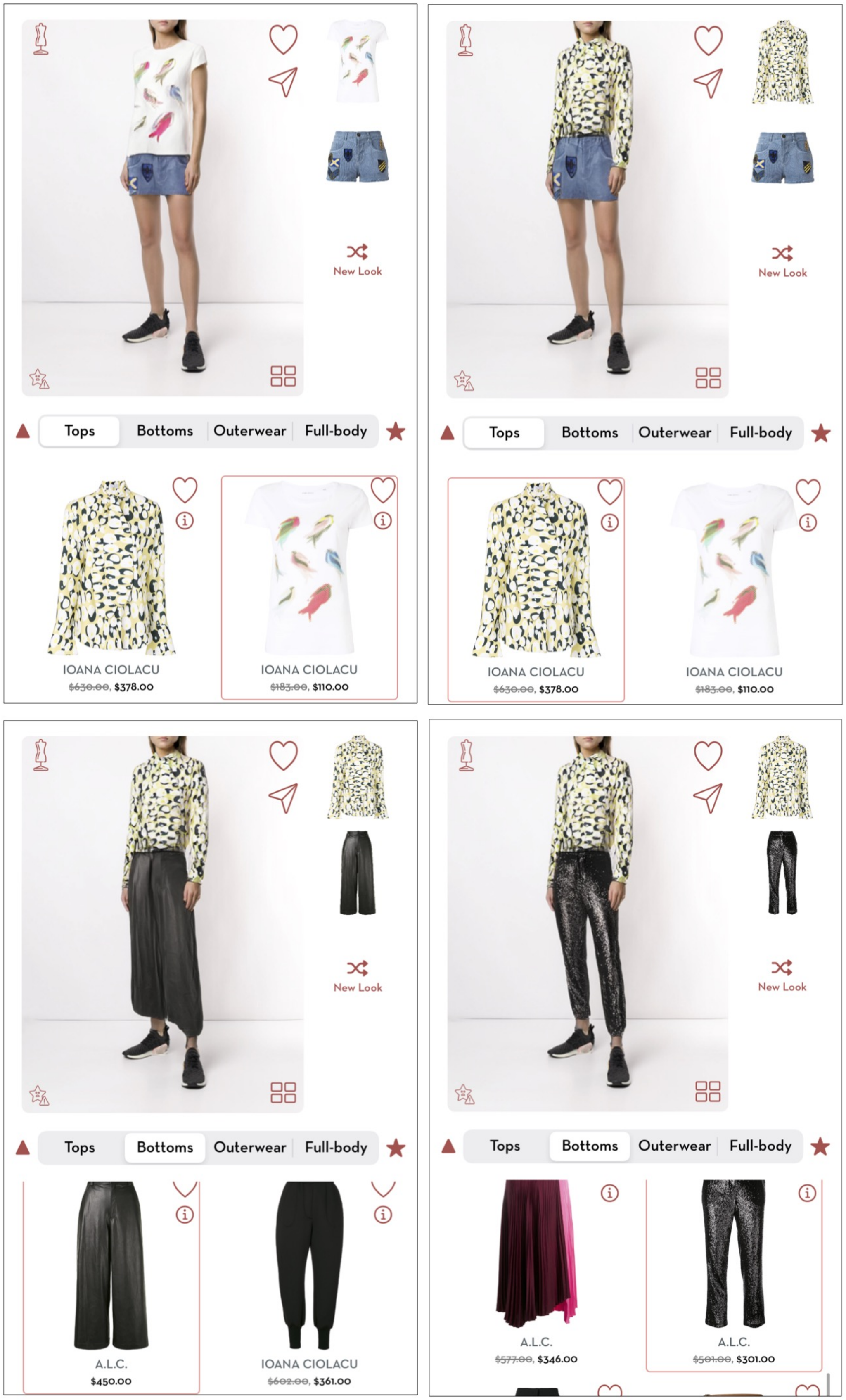

Figure 3. The figure shows examples of an interactive interface powered by our method. A user can select a garment listed below and our method will produce a visualization of the model wearing the selected garment in real-time. Garment images are products listed on an e-commerce site.

Figure 4. The figure shows outfits curated by users through the interactive interface (male models).

Figure 5. The figure shows outfits curated by users through the interactive interface (male models).

Figure 6. The figure shows outfits curated by users through the interactive interface (female models).

Figure 7. The figure shows outfits curated by users through the interactive interface (female models).