# Supplementary Material for Learning to Wear: Details-Preserved Virtual Try-on via Disentangling Clothes and Wearer

## A Detailed Model Architectures

Architecture of the CFM (Clothes Fitting Module) follows the same structure of U-Net described in [11]. As described in Table i, it consists of an encoder (the left columns; from the Input to conv5b layers) and a decoder (the right columns; from upsample1 to Output layers). The encoder consists of 5 repeated blocks of two $3 \times 3$ conv-layers with stride and padding of size 1, respectively. After two convolution layers, a $2 \times 2$ max-pooling layer is applied for downsampling. For each convolution layer, we apply an instance normalization, and we use a Rectified Linear Unit (ReLU) for activation. Every layer in the decoder consists of an upsampling of the feature map with halved number of channels, a concatenation with the corresponding feature map from the encoding path, and one $3 \times 3$ convolution, each followed by a ReLU. At the final layer, a $3 \times 3$ convolution with stride and padding of size 1 is used, slightly modified from the original U-Net [11]. In total, the network contains 22 convolutional layers.

## B Baselines

There are additional current state-of-the-art works that we do not compare against for the following reasons. VITON-HD [1] shows impressive quality in high-resolution images, but we do not compare it qualitatively or quantitatively because their training dataset in high-resolution is not publicly available. Dress in Order [2] is another recent work, focusing on dressing clothes in a suggested order. Since the task is different from our standard virtual try-on, this model is not comparable to our work. DCTON [3] tries to address the lack of paired data by adopting a double-cyclic architecture. Since this model is similar in architecture and shows similar performance with PF-AFN [4], we choose to compare with PF-AFN instead of this work. LM-VTON [9] tries to express the characteristics of the clothes using landmarks of clothes. As this model utilizes an additional information that is not a part of the VITON dataset, we do not choose this as a baseline for fair comparison.

© 2022. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

| Layer | Details | Output Size | Layer | Details | Output Size |
|---|---|---|---|---|---|
| Input | Clothes image ($3 \times H \times W$)<br>Warped mask ($1 \times H \times W$) | $4 \times H \times W$ | upsample1 | $2 \times 2$ upsampling | $1024 \times H/8 \times W/8$ |
| conv1a | $3 \times 3 \times 64$<br>Instance Normalization<br>ReLU | $64 \times H \times W$ | conv6a | $3 \times 3 \times 512$<br>Instance Normalization<br>ReLU | $512 \times H/8 \times W/8$ |
| conv1b | $3 \times 3 \times 64$<br>Instance Normalization<br>ReLU | $64 \times H \times W$ | conv6b | $3 \times 3 \times 512$<br>Instance Normalization<br>ReLU | $512 \times H/8 \times W/8$ |
| pool1 | $2 \times 2$ max pool | $64 \times H/2 \times W/2$ | conv6c | $3 \times 3 \times 512$<br>Instance Normalization<br>ReLU | $512 \times H/8 \times W/8$ |
| conv2a | $3 \times 3 \times 128$<br>Instance Normalization<br>ReLU | $128 \times H/2 \times W/2$ | upsample2 | $2 \times 2$ upsampling | $512 \times H/4 \times W/4$ |
| conv2b | $3 \times 3 \times 128$<br>Instance Normalization<br>ReLU | $128 \times H/2 \times W/2$ | conv7a | $3 \times 3 \times 256$<br>Instance Normalization<br>ReLU | $256 \times H/4 \times W/4$ |
| pool2 | $2 \times 2$ max pool | $128 \times H/4 \times W/4$ | conv7b | $3 \times 3 \times 256$<br>Instance Normalization<br>ReLU | $256 \times H/4 \times W/4$ |
| conv3a | $3 \times 3 \times 256$<br>Instance Normalization<br>ReLU | $256 \times H/4 \times W/4$ | conv7c | $3 \times 3 \times 256$<br>Instance Normalization<br>ReLU | $256 \times H/4 \times W/4$ |
| conv3b | $3 \times 3 \times 256$<br>Instance Normalization<br>ReLU | $256 \times H/4 \times W/4$ | upsample3 | $2 \times 2$ upsampling | $256 \times H/2 \times W/2$ |
| pool3 | $2 \times 2$ max pool | $256 \times H/8 \times W/8$ | conv8a | $3 \times 3 \times 128$<br>Instance Normalization<br>ReLU | $128 \times H/2 \times W/2$ |
| conv4a | $3 \times 3 \times 512$<br>Instance Normalization<br>ReLU | $512 \times H/8 \times W/8$ | conv8b | $3 \times 3 \times 128$<br>Instance Normalization<br>ReLU | $128 \times H/2 \times W/2$ |
| conv4b | $3 \times 3 \times 512$<br>Instance Normalization<br>ReLU<br>Dropout (0.5) | $512 \times H/8 \times W/8$ | conv8c | $3 \times 3 \times 128$<br>Instance Normalization<br>ReLU | $128 \times H/2 \times W/2$ |
| pool4 | $2 \times 2$ max pool | $512 \times H/16 \times W/16$ | upsample4 | $2 \times 2$ upsampling | $128 \times H \times W$ |
| conv5a | $3 \times 3 \times 1024$<br>Instance Normalization<br>Relu | $1024 \times H/16 \times W/16$ | conv9a | $3 \times 3 \times 64$<br>Instance Normalization<br>ReLU | $64 \times H \times W$ |
| conv5b | $3 \times 3 \times 1024$<br>Instance Normalization<br>ReLU<br>Dropout(0.5) | $1024 \times H/16 \times W/16$ | conv9b | $3 \times 3 \times 64$<br>Instance Normalization<br>ReLU | $64 \times H \times W$ |
| | | | conv9c | $3 \times 3 \times 64$<br>Instance Normalization<br>ReLU | $64 \times H \times W$ |
| | | | Output | $3 \times 3 \times 64$<br>Instance Normalization<br>ReLU | $4 \times H \times W$ |

Table i: Full architectural details of the CFM.

| (↓) | LPIPS$^{\text{all}}$ | LPIPS$^{p}_{20}$ | LPIPS$^{p}_{25}$ | LPIPS$^{p}_{30}$ |
|---|---|---|---|---|
| CP-VTON+ | 0.133 | 0.067 | 0.114 | 0.115 |
| ACGPN | 0.406 | 0.237 | 0.349 | 0.358 |
| PF-AFN | 0.123 | 0.060 | 0.102 | 0.102 |
| Ours | **0.116** | **0.054** | **0.095** | **0.095** |

Table ii: Quantitative comparisons to state-of-the-art models with LPIPS using a pretrained AlexNet [8].

| CFM inputs | LPIPS$^{p}_{20}$(↓) |
|---|---|
| Warped clothes mask ($\hat{c}_m$) | **0.054** |
| Warped clothes ($\hat{c}$) | 0.147 |
| Both warped clothes ($\hat{c}$) and warped clothes mask ($\hat{c}_m$) | 0.124 |

Table iii: Comparison on various CFM input configurations with LPIPS using a pretrained AlexNet [8].

# C Additional Evaluation Results

Table ii compares LPIPS$_{\text{alex}}$ with other baselines, and Table iii compares LPIPS$_{\text{alex}}$ with various CFM input configurations, namely the warped clothes $\hat{c}$ in RGB and/or the warped clothes mask $\hat{c}_m$. The CFM may take as input either or both of these, together with the in-shop clothes image $c$.

Overall, we observe a similar pattern in both tables to the results in Table 1 in the main manuscript, reporting LPIPS$_{\text{VGG}}$, except for the LPIPS$^{\text{all}}$. Interestingly, the ACGPN turns out to perform the best in terms of LPIPS based on VGG (in Table 1), while it is the worst according to the metric based on AlexNet. This indicates that the metric based on the entire pixels is highly unstable. On the other hand, we observe that the proposed metrics LPIPS$^{p}_{\varepsilon}$ reserve the same ordering of all four methods across all $\varepsilon$s we report, for Table 1 (in the main manuscript) and Table ii.

# D More on the User Study

**Experimental Settings.** We conducted a user study to compare the quality of generation results. We randomly sampled 200 pairs from the VITON test set without cherry-picking and divided them into 2 sets, 100 pairs in each with 60 volunteers who are unfamiliar with virtual try-on techniques, and randomly assigned them to either set A or B. As illustrated in Fig. i, each question shows the reference person ($I$), the target clothes ($c$), and 4 randomly-ordered virtual try-on images generated by CP-VTON+, ACGPN, PF-AFN, and our method. The participant was asked to choose the best try-on result among them. Fig. ii shows more examples we used in the user study.

**Results and Discussions.** As mentioned in Sec. 6.4, in total 60 users participated in the survey, and 65.3% of them chose our method most frequently as the best performing one. Fig. iii shows the percentage of participants sorted by the number of questions they chose ours as the best. We see that more than half of the participants chose ours as the best on $\geq 60$ questions (out of 100). All participants chose ours as the best for at least 35 questions (out of 100), which is significantly higher than 25, the expected number if randomly chosen from 4 choices. From the question-centric view, as described in Table iv, participants answered that our model generated the best results compared to the other 3 baselines in 153 (76.5%)

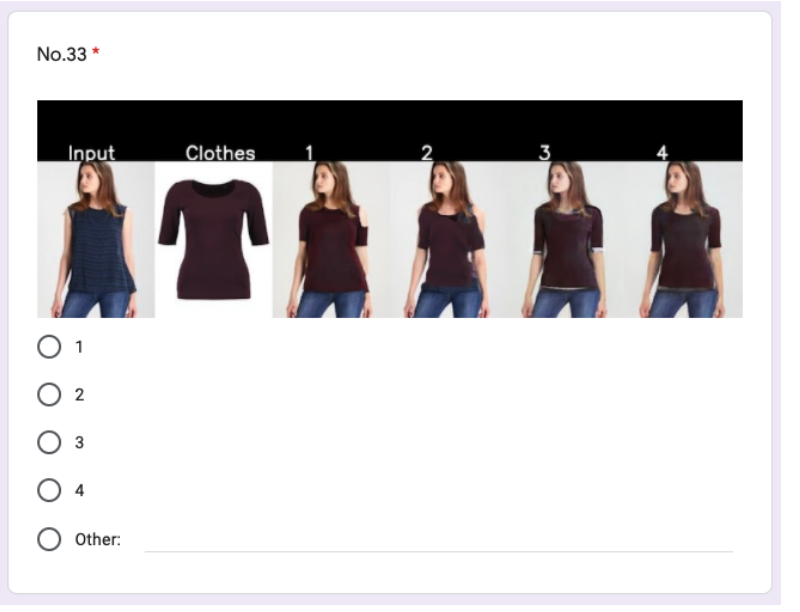


Figure i: A screenshot of a survey question.

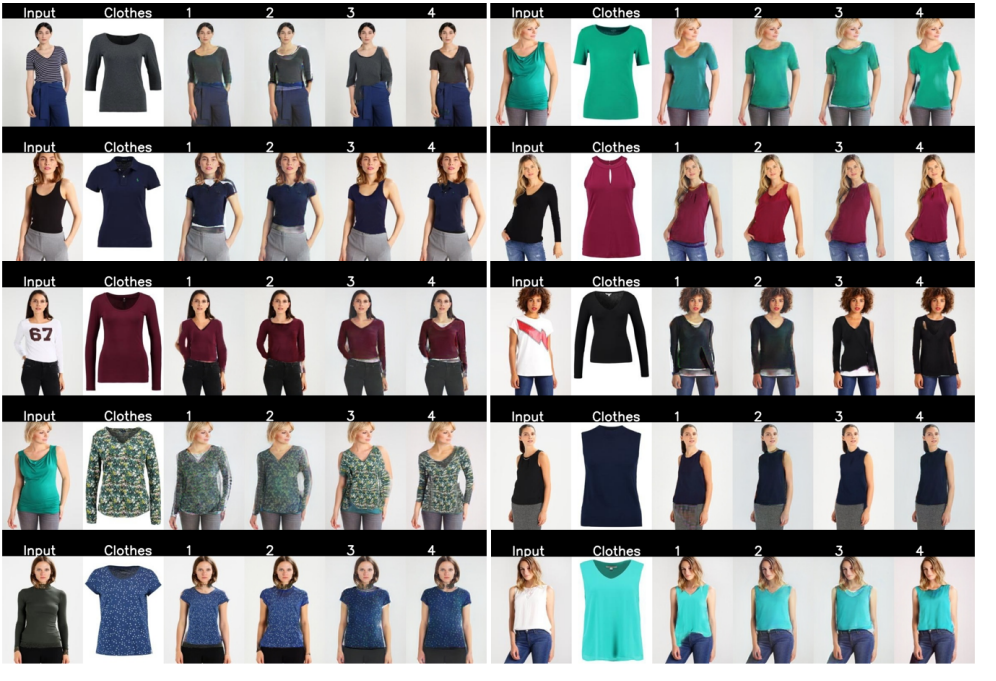


Figure ii: Examples of user study questions.

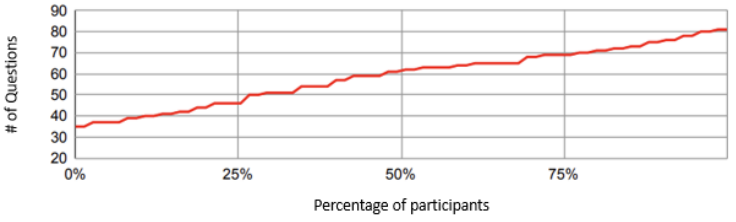


Figure iii: Number of questions in which participants chose our result as the best.

| | #Questions | Ratio |
|---|---|---|
| CPVTON+ | 5 | 2.5% |
| ACGPN | 22 | 11.0% |
| PF-AFN | 20 | 10.0% |
| Ours | **153** | 76.5% |

Table iv: Number of questions that each method was chosen as the best.

questions among 200 questions.

For reference, ACGPN [13] conducted a user study comparing with CP-VTON [12], VITON [6], and VTNFP [14]. Unlike our study, they asked questions in an A/B (binary) manner; that is, each question had two choices, one by their ACGPN method and the other by one of the baselines. 66.7%, 89.8%, and 76.6% among the participants chose the proposed method (ACGPN) against VTNFP, CP-VTON, and VITON, respectively. PFAFN [4] also conducted a similar user study with 50 volunteers, comparing against CP-VTON, ClothFlow [8], CP-VTON+ [10], ACGPN, and WUTON [7]. 84.3% of their participants chose the proposed method against a single baseline, just as in the user study conducted by [13]. We emphasize that these two user studies asked the participants to choose one out of the two candidates, one from the proposed method and the other from another baseline. Thus, the arithmetic expectation of their experiment was 50%. On the other hand, our user study showed all four images at the same time, making the expectation 25%, while we achieved a 65.3% (participant-centric) or 76.5% (question-centric) winning rate.

# E More Qualitative Results

Fig. iv shows additional virtual try-on results using our proposed approach and three baselines [4, 10, 13]. Fig. v illustrates our and baselines' generation results focusing on various body shapes. Table v–vi list additional results on various poses using ACGPN, PF-AFN, and ours.

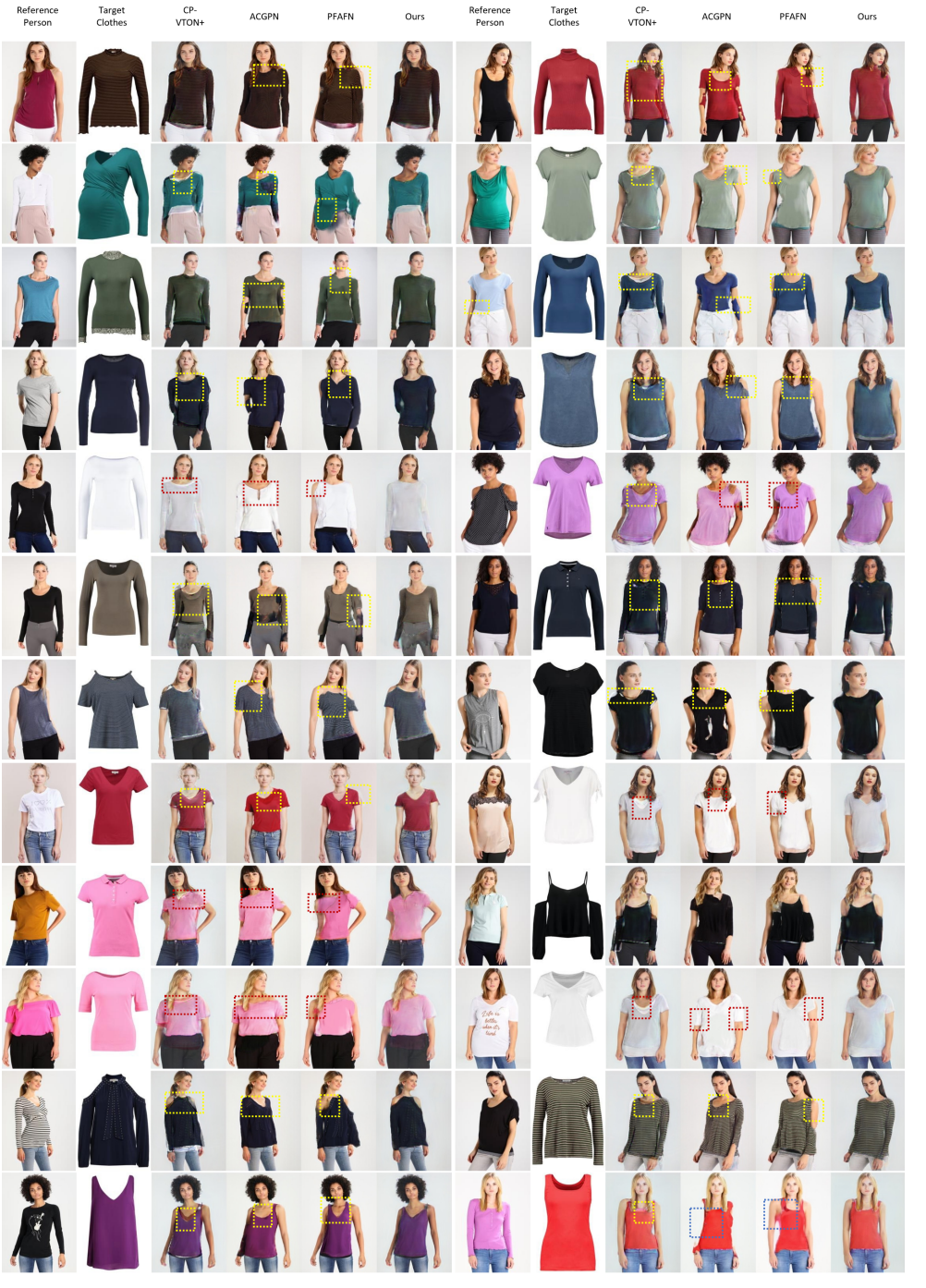


Figure iv: Additional qualitative virtual try-on results.

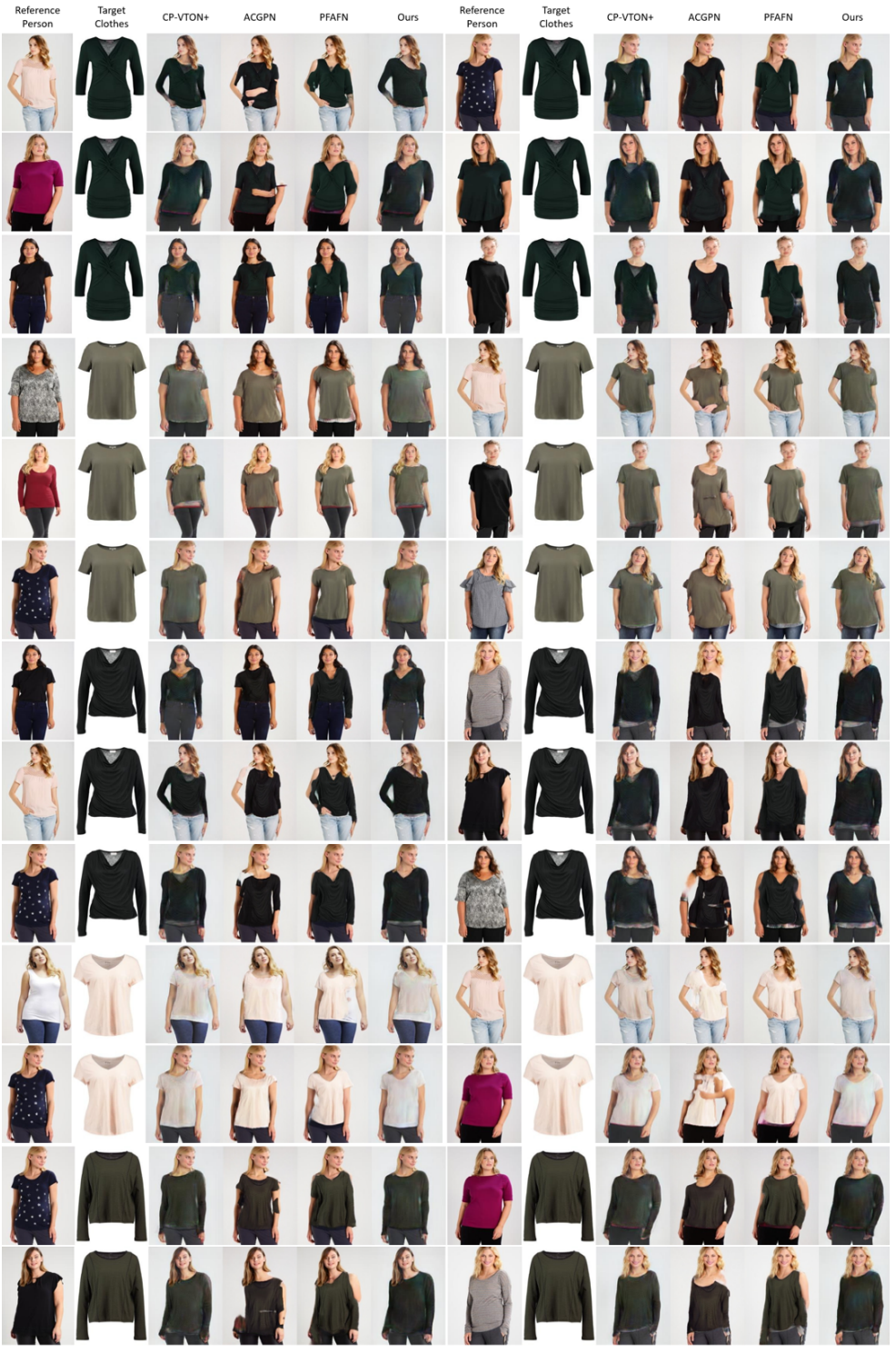


Figure v: Additional results on various body shapes.

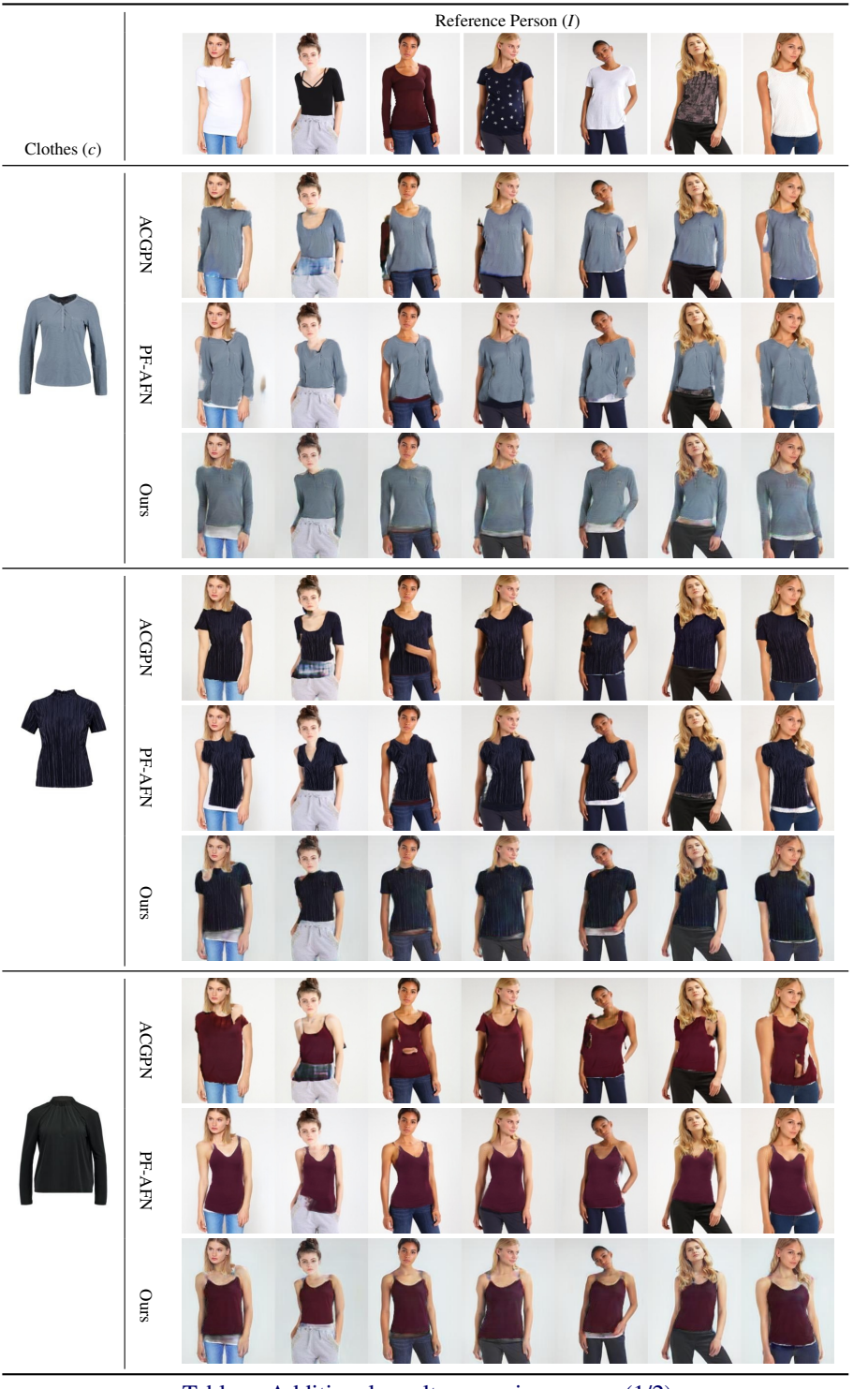


Table v: Additional results on various poses (1/2)

Table vi: Additional results on various poses (2/2)

# F Failure cases

We observe the proposed method struggles when arms are folded (marked with red box in Fig. vi). Although it still shows better synthesized results than baselines, better handling such complicated poses is an interesting direction for future work.

Figure vi: Qualitative comparisons with various poses.

# References

[1] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. Viton-HD: High-resolution virtual try-on via misalignment-aware normalization. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[2] Aiyu Cui, Daniel McKee, and Svetlana Lazebnik. Dressing in order: Recurrent person image generation for pose transfer, virtual try-on and outfit editing. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[3] Chongjian Ge, Yibing Song, Yuying Ge, Han Yang, Wei Liu, and Ping Luo. Disentangled cycle consistency for highly-realistic virtual try-on. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[4] Yuying Ge, Yibing Song, Ruimao Zhang, Chongjian Ge, Wei Liu, and Ping Luo. Parser-free virtual try-on via distilling appearance flows. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[5] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. VITON: An image-based virtual try-on network. In *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[6] Xintong Han, Xiaojun Hu, Weilin Huang, and Matthew R Scott. ClothFlow: A flow-based model for clothed person generation. In *Proc. of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2019.

[7] Thibaut Issenhuth, Jérémie Mary, and Clément Calauzènes. Do not mask what you do not need to mask: a parser-free virtual try-on. In *Proc. of the European Conference on Computer Vision (ECCV)*, 2020.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[9] Guoqiang Liu, Dan Song, Ruofeng Tong, and Min Tang. Toward realistic virtual try-on through landmark-guided shape matching. In *Proc. of the AAAI Conference on Artificial Intelligence*, 2021.

[10] 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 *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshop*, 2020.

[11] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *Proc. of the International Conference on Medical Image Computing and Computer-assisted Intervention (MICCAI)*, 2015.

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

[13] 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 *Proc. of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[14] Ruiyun Yu, Xiaoqi Wang, and Xiaohui Xie. VTNFP: An image-based virtual try-on network with body and clothing feature preservation. In *Proc. of the IEEE/CVF Conference on International Conference on Computer Vision (ICCV)*, 2019.