

Supplementary Material: Shape-Guided Clothing Warping for Virtual Try-On

Xiaoyu Han
xyhan@stu.hit.edu.cn
Harbin Institute of Technology
Weihai, China

Chenyang Wang
c.wang@stu.hit.edu.cn
Harbin Institute of Technology
Weihai, China

Shunyuan Zheng
sawyer0503@hit.edu.cn
Harbin Institute of Technology
Weihai, China

Xin Sun
sunxintyc@hit.edu.cn
Harbin Institute of Technology
Weihai, China

Zonglin Li
zonglin.li@hit.edu.cn
Harbin Institute of Technology
Weihai, China

Quanling Meng*
quanling.meng@hit.edu.cn
Harbin Institute of Technology
Weihai, China

This document presents the supplementary material omitted from the main paper. In Section 1, we provide a more detailed explanation of the semantic-replacement strategy. In Section 2, we provide more qualitative results. In Section 3, we present additional ablation studies on the global shape constraints and the co-training strategy.

1 Semantic-replacement Strategy

Before acquiring the target semantic layout S_t , we introduce target semantic guidance S_{rep} based on a semantic-replacement strategy, which removes the original clothing semantics while ensuring the continuity and pose consistency of the new semantics. As shown in Figure 1, the source semantic layout S_s is first separated into a multi-channel binary parsing map, where each channel corresponds to clothing or a part of the person’s body. Except for clothing and limbs, other contents of the person image should be preserved after try-on, so we only replace the clothing channel and limb channels of the source semantic layout S_s to get the target semantic guidance S_{rep} . Note that the contour of limbs reflects the shape of the clothing, so they also need to be replaced. For the clothing channel, we replace its content with the warped clothing mask M_w , which has been aligned with the person’s body in the dual-path clothing warping module. For the limb channels, we first extract the corresponding parts from the skeleton map P_s according to 6 key points of limb regions as the limb-skeleton map P_l . Then we mask P_l with $1 - M_w$ to discard the regions that conflict with the warped clothing due to the higher priority of clothing semantics, which is utilized to replace the original contents in the limb channels (with hand regions preserved). After the replacement, we recalibrate the last channel (representing the background region) of S_{rep}^i based on other channels. The overall semantic-replacement strategy can be presented as follows:

$$S_{rep}^i = \begin{cases} S_s^i & 0 \leq i \leq 2 \\ M_w & i = 3 \\ P_l \odot (1 - M_w) & 4 \leq i \leq 5 \\ 1 - \mathcal{T}(\sum_{k=0}^5 S_{rep}^k) & i = 6 \end{cases}, \quad (1)$$

where $i \in \{0, 1, \dots, 6\}$ denotes the channel index of S_s and S_{rep} , $\mathcal{T}(\cdot)$ is a truncated function that ensures the output value falls within the range of zero and one, and \odot is the element-wise multiplication. Then, the content of each S_{rep}^i is channel-wise merged again.

*Corresponding author.

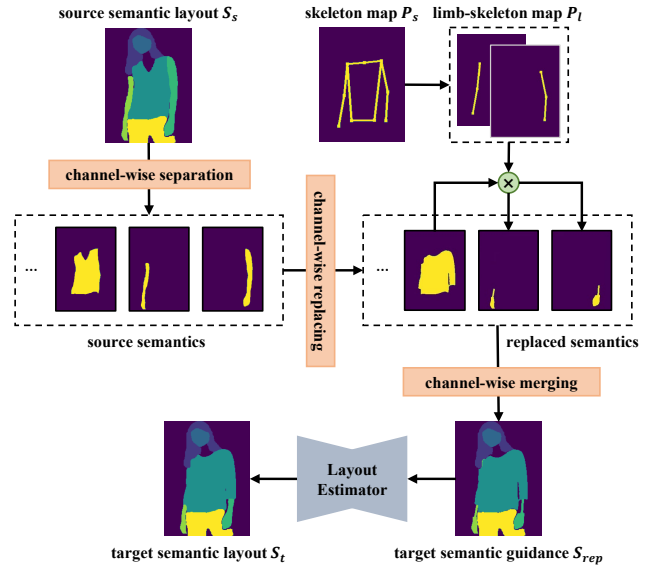


Figure 1: The schematic of the semantic-replacement layout estimation module. Based on the semantic-replacement strategy, this module produces the target semantic layout S_t that describes the person wearing new clothing.

Subsequently, we employ a UNet [8] model as the semantic layout estimator, which predicts the target semantic layout S_t by inputting the target semantic guidance S_{rep} . Since the pose information and the semantics of warped clothing have been integrated into the input beforehand, the estimator can more effectively discern the generation locations of target semantics based on the contents after replacement, which provides more accurate structural guidance for the subsequent try-on synthesis.

2 More Qualitative Results

We conduct additional qualitative experiments to validate the diversity of the results generated by our method. First, we select in-shop clothing with different styles and person images with different poses from the testing set of the VITON-HD [2] dataset, which are combined into pairs to form multiple input groups, producing various try-on results. These results are depicted in Figure



Figure 2: More specific samples of our method.

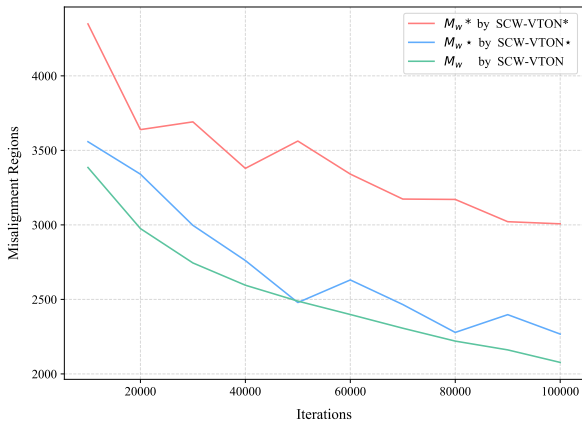


Figure 3: The number of pixels in the misalignment regions caused by M_w^* , M_w^* , and M_w as the training iteration increases.

4, showcasing the robustness and generalization capability of our SCW-VTON, which is not contingent upon specific paired data. Furthermore, as shown in Figure 2, we provide more specific samples of (a) male model, (b) turtleneck, (c) jacket, (d) back pose, (e) side pose, (f) children’s clothing, (g) big logo, (h) crop top, (i) shorts, (j) pants, (k) dress, and (l) coat.

Additionally, Figure 5 provides more qualitative comparison results of ACGPN [10], SDAFN [1], RMGN [7], DOC-VTON [11], and SCW-VTON on the VITON [4] dataset. Figure 6 exhibits further comparison results of HR-VTON [6], SAL-VTON [9], DCI-VTON [3], StableVITON [5], and SCW-VTON on the VITON-HD [2] dataset.

3 More Ablation Studies

We conduct additional experiments to demonstrate the necessity of incorporating extra global shape constraints and the co-training strategy. We adopt the same setup as in the main paper, defining

SCW-VTON* as a variant of SCW-VTON that ablates shape-guided cross-attention blocks, while SCW-VTON* is another variant that only ablates the co-training strategy. Specifically, we compare the ability of SCW-VTON, SCW-VTON*, and SCW-VTON* to capture the shape characteristics of clothing after deformation with increasing training iterations. To distinguish from the warped clothing mask M_w obtained from our SCW-VTON, we denote the output mask by SCW-VTON* as M_w^* , and the output mask by SCW-VTON* as M_w^* . We respectively calculate the misalignment regions of M_w^* , M_w^* , and M_w with the ground truth M_{gt} (i.e., the actual clothing regions in the person image) as the number of training iterations increases, and compare these misalignment regions in Figure 3. The results illustrate that M_w^* generates a considerable number of misalignment regions, and it is evident that the appearance flow predicted by SCW-VTON* is unstable. By incorporating shape-guided cross-attention blocks to provide global shape constraints in SCW-VTON*, M_w^* exhibits significantly less misalignment and better matches the person’s body. Moreover, with the implementation of the co-training strategy in our full model SCW-VTON, there is further improvement in both accuracy and stability. In summary, these results demonstrate the effectiveness of the shape-guided cross-attention blocks in providing global shape constraints for estimating the clothing shape aligned with the person’s body accurately and robustly, along with the beneficial impact of co-training between the two paths on the weight update of the flow decoder.

References

- [1] Shuai Bai, Huiling Zhou, Zhikang Li, Chang Zhou, and Hongxia Yang. 2022. Single stage virtual try-on via deformable attention flows. In *Proceedings of the European Conference on Computer Vision*. Springer, 409–425.
- [2] Seunghwan Choi, Sunghyun Park, Minsoo Lee, and Jaegul Choo. 2021. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 14131–14140.
- [3] Junhong Gou, Siyu Sun, Jianfu Zhang, Jianlou Si, Chen Qian, and Liqing Zhang. 2023. Taming the Power of Diffusion Models for High-Quality Virtual Try-On with Appearance Flow. In *Proceedings of the 31st ACM International Conference on Multimedia*. 7599–7607.

- [4] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. 2018. VITON: An Image-Based Virtual Try-On Network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7543–7552.
- [5] Jeongho Kim, Gyojung Gu, Minhoo Park, Sunghyun Park, and Jaegul Choo. 2024. StableVITON: Learning Semantic Correspondence with Latent Diffusion Model for Virtual Try-On. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.
- [6] Sangyun Lee, Gyojung Gu, Sunghyun Park, Seunghwan Choi, and Jaegul Choo. 2022. High-resolution virtual try-on with misalignment and occlusion-handled conditions. In *Proceedings of the European Conference on Computer Vision*. Springer, 204–219.
- [7] Chao Lin, Zhao Li, Sheng Zhou, Shichang Hu, Jialun Zhang, Linhao Luo, Jiarun Zhang, Longtao Huang, and Yuan He. 2022. Rmgn: A regional mask guided network for parser-free virtual try-on. *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence (2022)*.
- [8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015: 18th International Conference, Munich, Germany, October 5–9, 2015, Proceedings, Part III* 18. Springer, 234–241.
- [9] Keyu Yan, Tingwei Gao, Hui Zhang, and Chengjun Xie. 2023. Linking Garment With Person via Semantically Associated Landmarks for Virtual Try-On. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 17194–17204.
- [10] Han Yang, Ruimao Zhang, Xiaobao Guo, Wei Liu, Wangmeng Zuo, and Ping Luo. 2020. Towards Photo-Realistic Virtual Try-On by Adaptively Generating-Preserving Image Content. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7850–7859.
- [11] Zhijing Yang, Junyang Chen, Yukai Shi, Hao Li, Tianshui Chen, and Liang Lin. 2023. OccluMix: Towards De-Occlusion Virtual Try-on by Semantically-Guided Mixup. *IEEE Transactions on Multimedia* (2023).

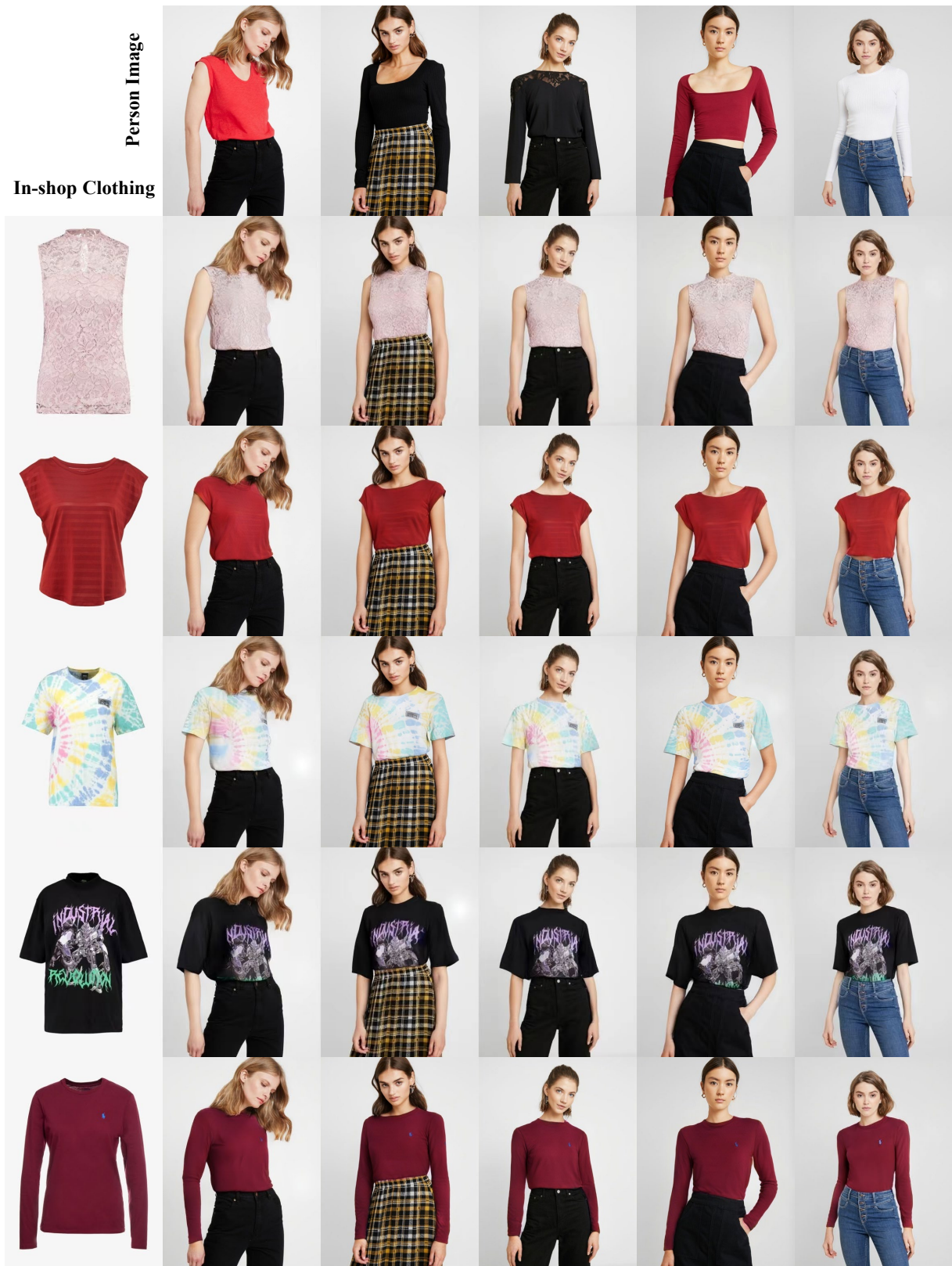


Figure 4: In-shop clothing with different styles and person images with different poses are selected from the testing set of the VITON-HD [2] dataset to form multiple input groups and generate diverse try-on results.



Figure 5: Qualitative results of our SCW-VTON and baseline methods on the VITON [4] dataset.

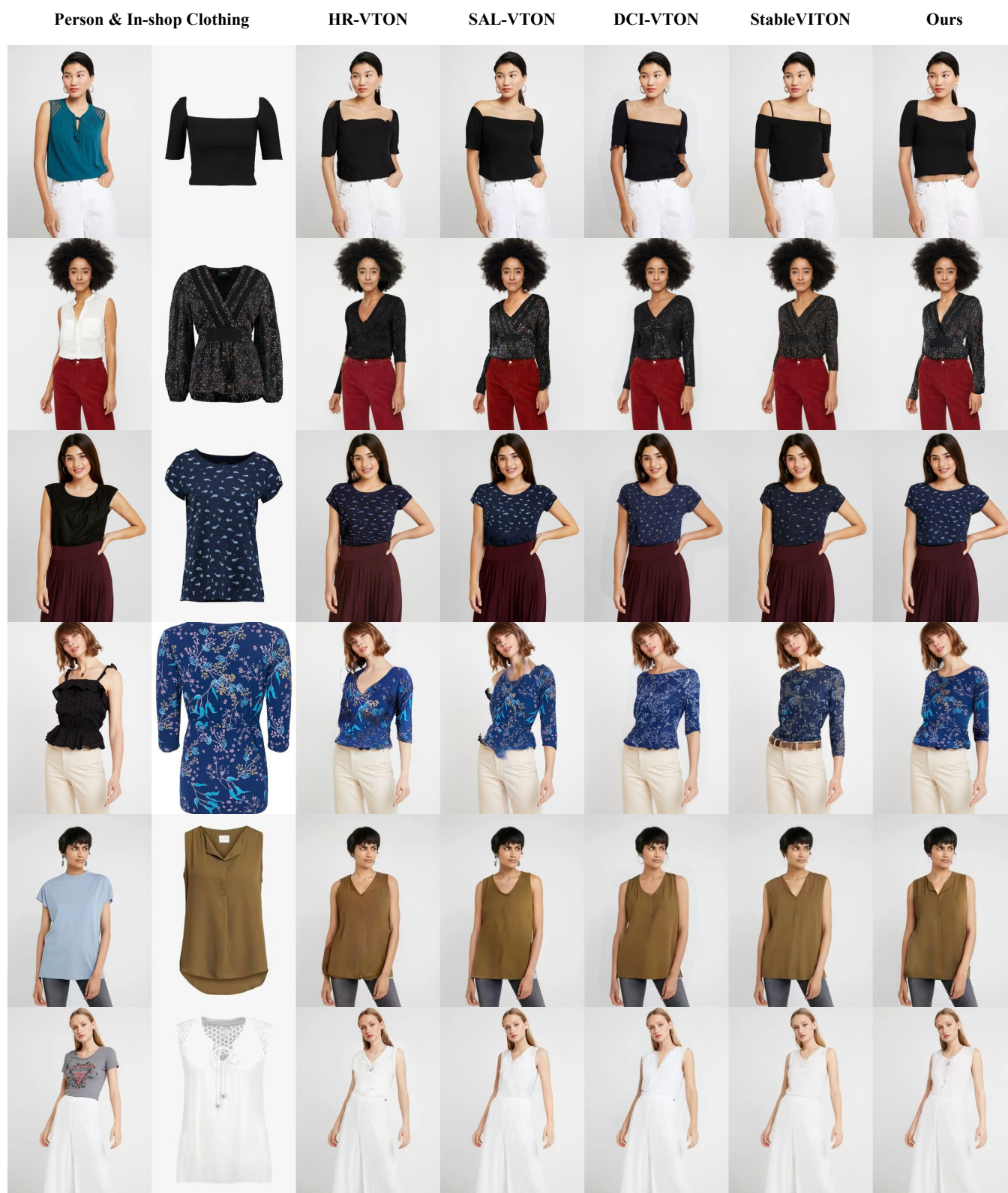


Figure 6: Qualitative results of our SCW-VTON and baseline methods on the VITON-HD [2] dataset.