Texture-Preserving Diffusion Models for High-Fidelity Virtual Try-On

Supplementary Material

This supplementary material includes four sections. Section 1 explains the details of the mask augmentation strategy. We provide more implementation details about applying our method to the person-to-person virtual try-on in Section 2. Section 3 conducts experiments to justify the efficiency of SATT. We present additional qualitative results and discuss our method's limitations in Section 4.

1. Details of the Mask Augmentation Strategy

Since our model needs to handle both bounding-box and precise garment based masks, we propose a mask augmentation strategy to produce masks in various shapes during the training phase. Specifically, we first obtain the outer contour of the original garment in the source person image using human parsing. Then we sample con points on the contour and shift each point outwards by s pixels. In this way, we can obtain precise masks in various shapes. In our experiments, we randomly select con from 20 to 50 and s from 30 to 70.

2. Details of the Person-to-Person Setting

For the person-to-person setting, we first use a human parsing tool to obtain the garment worn by the person in the reference image. Then, we concatenate the obtained garment image with the masked source person image along the spatial dimension. Moreover, since the person-to-person setting usually encounters significant pose differences, we need to provide the pose information of the person in the reference image. Therefore, we concatenate the pose image of the reference person on the right side of the pose image of the source person.

3. Comparisons in Model Complexity

We compare model complexities between SATT and two state-of-the-art warping-free methods, i.e., LaDI-VTON [8] and TryOnDiffusion [9]. All experiments are carried out with an A100 GPU.

We first compare the model sizes between LaDI-VTON, TryOnDiffusion, and our model. Since TryOnDiffusion is not open sourced, we re-implement it based on the Stable Diffusion [6] backbone. As presented in Table 1, our model has the smallest size. The reason is that LaDI-VTON adopts a VIT model [11] as the specialized garment encoder to obtain garment features for the cross-attention module. TryOnDiffusion duplicates the UNet [10] in the diffusion model to serve as the specialized garment encoder, which enlarges the model complexity. In comparison, our SATT does not require an additional specialized garment encoder. Furthermore, SATT is more efficient than the other methods. Specifically, they can process 2.85, 1.92, and 1.78 samples per second, respectively. These comparisons justify the superior efficiency of SATT.

Table 1. Comparisons in model complexity. We re-implement TryOnDiffusion based on the Stable Diffusion model.

Method	Parameters (Million)	Throughput (samples/s)
TryOnDiffusion	2176	1.78
LaDI-VTON	1920	1.92
SATT	945	2.85

4. Additional Results and Limitations

We show additional qualitative comparisons between our method and state-of-the-art methods including ACGPN [1], PF-AFN [7], SDAFN [3], VITON-HD [4], HR-VITON [2], and LaDI-VTON [8] in Figure 3 and Figure 4. Our method consistently outperforms the other methods in terms of the quality and fidelity of the try-on images. Moreover, we demonstrate the robustness of our method in handling challenging poses, as shown in Figure 5. It is shown that our method performs more stably compared with the other methods when encountering large pose changes.

Limitations This work also has certain limitations. For example, since images in nearly all databases for this task have single-color background, it is hard for our model to generalize to natural images. This is a commonly recognized problem in existing try-on works [9]. We provide some examples in Figure 1. Besides, we present some sub-optimal try-on results in Figure 2. It is shown that our method struggles to perfectly reproduce the tiny characters and patterns. We blame that to the infidelity latent space of the VAE. Since SATT is performed in the latent space, it cannot reproduce tiny textures which is lost during the compression process of the VAE.

Figure 1. Results where the source image contains natural background.

Figure 2. TPD struggles to perfectly reproduce the tiny characters and patterns.

References

[1] H. Yang, R. Zhang, X. Guo, W. Liu, W. Zuo, P. Luo. Towards photo-realistic virtual try-on by adaptively generating-preserving image content. In *CVPR*, 2020. 1

[2] S. Lee, G. Gu, S. Park, S. Choi, J. Choo. High-Resolution Virtual Try-On with Misalignment and Occlusion-Handled Conditions. In *ECCV*, 2022. 1

[3] S. Bai, H. Zhou, Z. Li, C. Zhou, H. Yang. Single stage virtual try-on via deformable attention flows. In *ECCV*, 2022. 1

[4] S. Choi, S. Park, M. Lee, J. Choo. Viton-hd: High-resolution virtual try-on via misalignment-aware normalization. In *CVPR*, 2021. 1, 3, 5

[5] X. Han, Z. Wu, Z. Wu, R. Yu, L. Davis. Viton: An image-based virtual try-on network. In *CVPR*, 2018. 4

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 1

[7] Y. Ge, Y. Song, R. Zhang, C. Ge, W. Liu, P. Luo. Parser-free virtual try-on via distilling appearance flows. In *CVPR*, 2021. 1

[8] D. Morelli, A. Baldrati, G. Cartella, M. Cornia, M. Bertini, R. Cucchiara. LaDI-VTON: Latent Diffusion Textual-Inversion Enhanced Virtual Try-On. arXiv:2305.13501, 2023. 1

[9] L. Zhu, D. Yang, T. Zhu, F. Reda, W. Chan, C. Saharia, M. Norouzi, I. Kemelmacher-Shlizerman. TryOnDiffusion: A Tale of Two UNets. In *CVPR*, 2023. 1

[10] O. Ronneberger, P. Fischer, T. Brox. U-net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015. 1

[11] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv:2010.11929, 2020. 1

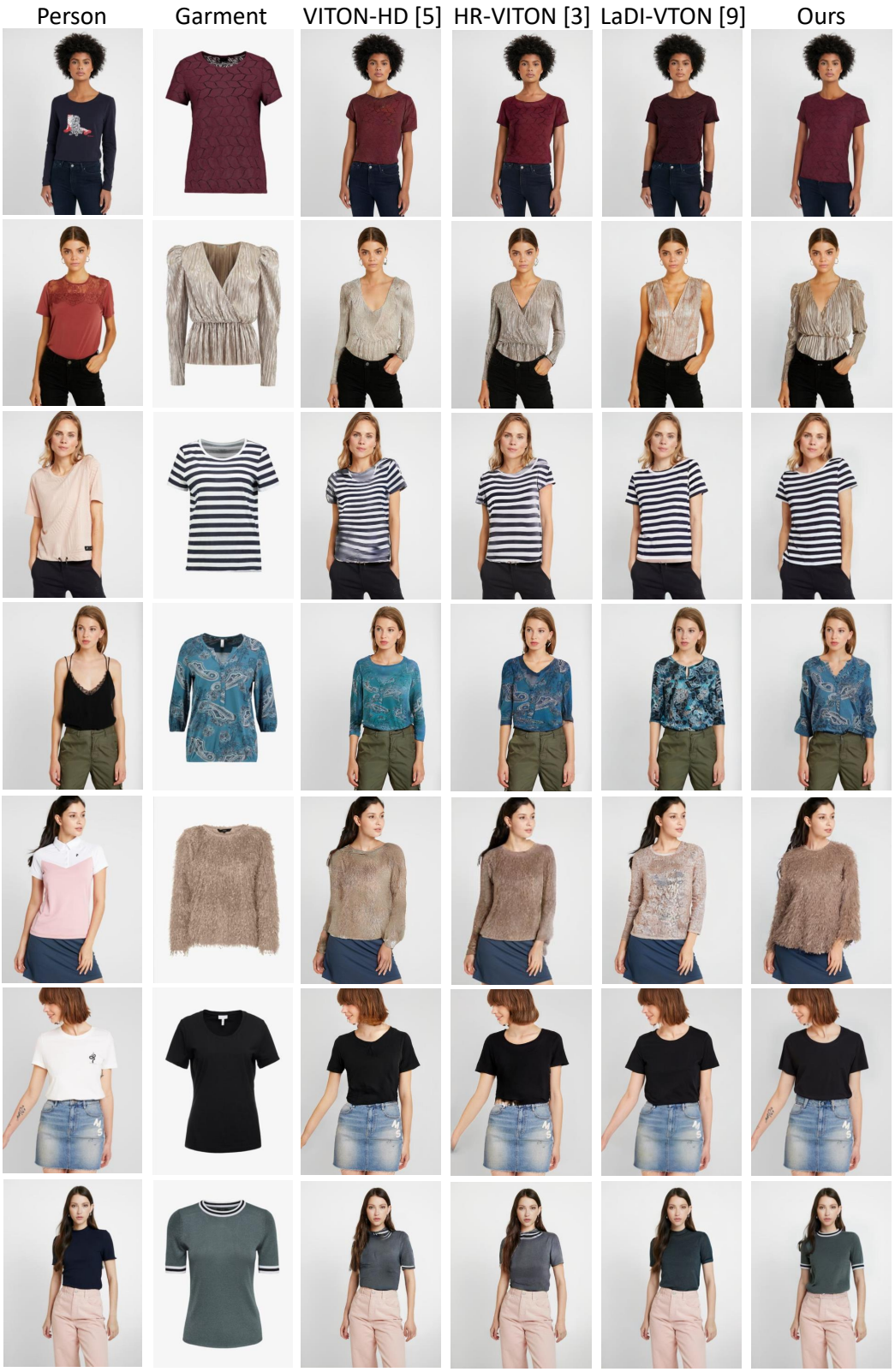

Figure 3. The additional results on VITON-HD [4] database.

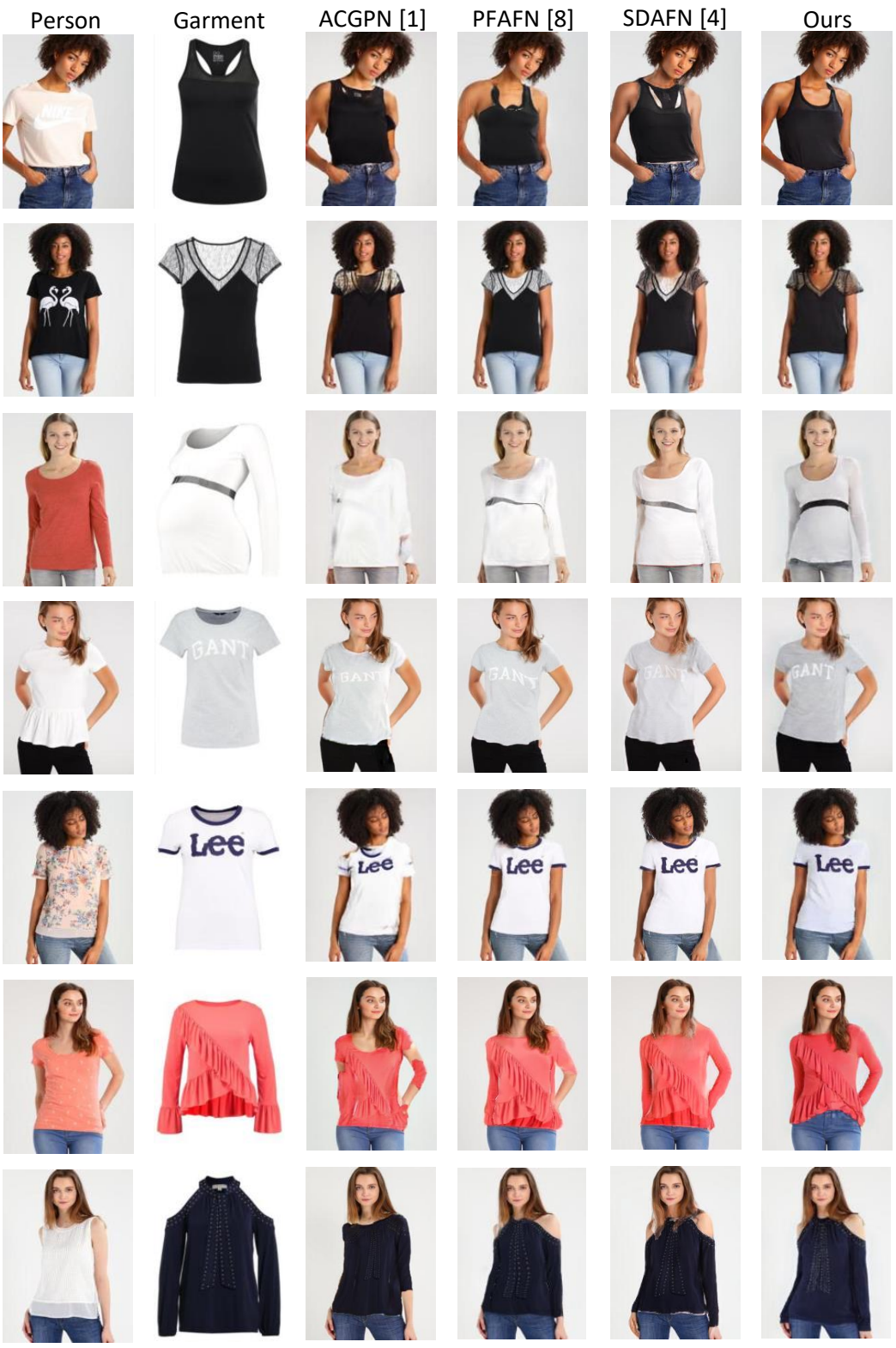

Figure 4. The additional results on VITON [5] database.

Figure 5. The additional results with challenging poses on VITON-HD [4] database.