Person Image Synthesis via Denoising Diffusion Model
(Supplementary Material)

Ankan Kumar Bhunia[1] Salman Khan[1,2] Hisham Cholakkal[1] Rao Muhammad Anwer[1,4]
Jorma Laaksonen[4] Mubarak Shah[5] Fahad Shahbaz Khan[1,3]
[1]Mohamed bin Zayed University of AI, UAE [2]Australian National University, Australia
[3]Linköping University, Sweden [4]Aalto University, Finland [5]University of Central Florida, USA

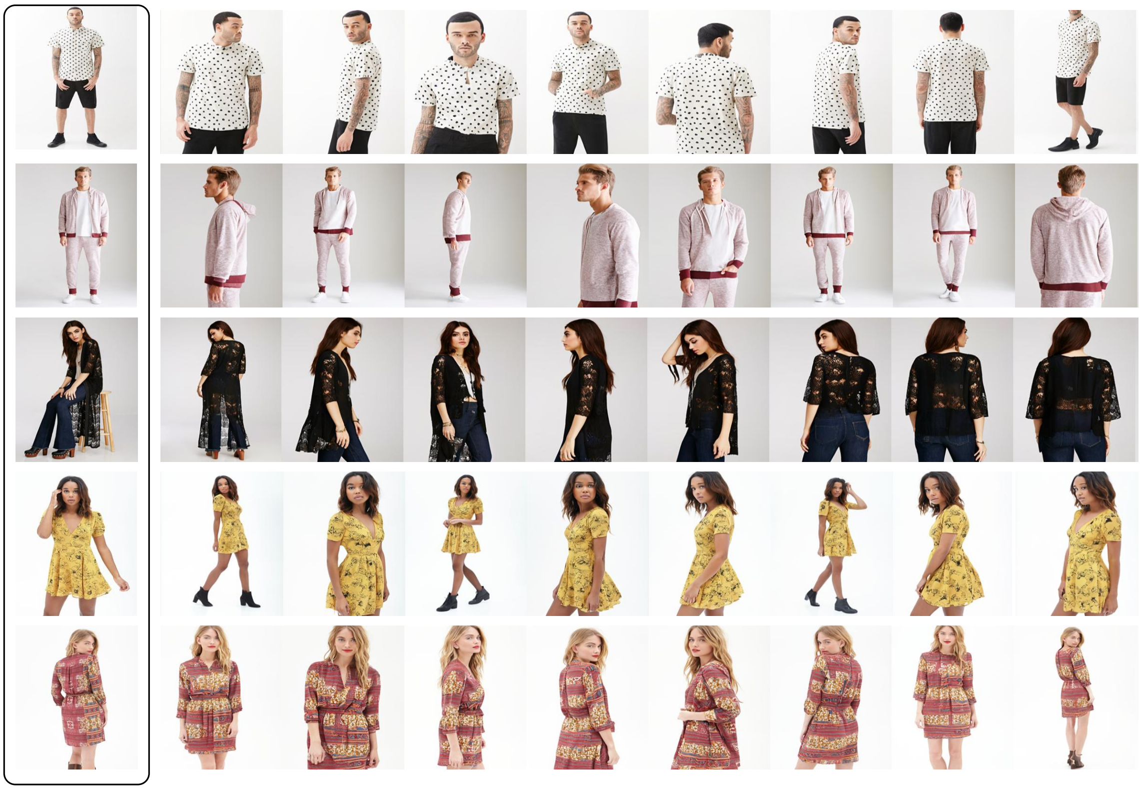

Figure 1. Results of synthesizing person images at arbitrary poses using our proposed PIDM. The left column contains the source images.

In this supplementary material, we present additional qualitative results of our proposed PIDM.

1. Additional Qualitative Results

In Fig. 5-7, we present a comprehensive visual comparison of our method with other state-of-the-art frameworks on DeepFashion dataset. We compare our method with ADGAN [1], PISE [5], GFLA [4], DPTN [6], CASD [7] and NTED [2]. In comparison to the existing methods, our proposed PIDM accurately retains the appearance of the source while also producing images that are more natural and sharper. Moreover, even if the target pose is complex, our method can still generate it precisely.

Fig. 1 shows qualitative results of synthesizing person images at arbitrary poses using our proposed PIDM. For each source image, we generate 8 samples of the same person in various poses. Our proposed PIDM accurately retains the appearance of the source while also generating consis-

Figure 2. We visualize the gradual transfer of appearance at selected timesteps from $t = T$ to $t = 1$.

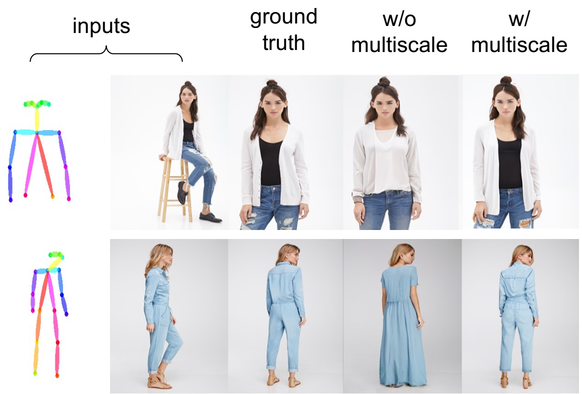

Figure 3. Qualitative analysis of multiscale fusion with texture diffusion blocks. The introduction of multiscale fusion significantly enhances the effectiveness of appearance transfer.

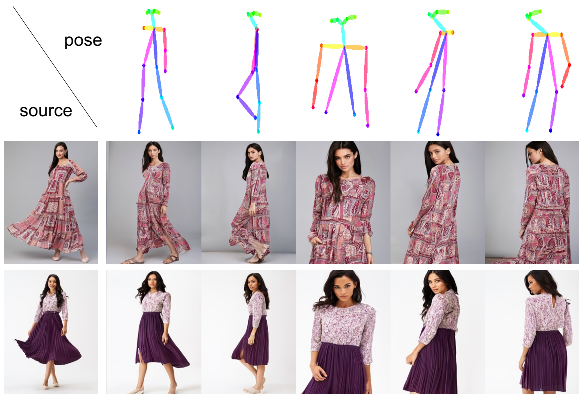

Figure 4. Qualitative analysis of PIDM in-the-wild setting. We demonstrate the robustness of our approach by testing it on images collected from a fashion e-commerce site. Given a source image (shown in the left column), PIDM attempts to synthesize the person image in five different poses.

tent patterns across different poses.

Fig. 3 shows that multiscale fusion aids in generating photo-realistic images, in which the output image style tightly aligns with the source image appearance. Fig. 2 illustrates a visualization of the gradual transfer of appearance at different timesteps. In particular, we visualize the prediction of $\boldsymbol{x}_0$ at selected timesteps from $t = T$ to $t = 1$. The visualization demonstrates the importance of gradually transferring the source appearance to generate the final output image. We verify the robustness of our approach by testing it on images collected from a fashion e-commerce site. Fig. 4 presents a few generated samples that demonstrate the generalization capability of PIDM in-the-wild scenarios.

References

[1] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *CVPR*, 2020. 1, 3, 4, 5

[2] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H Li. Neural texture extraction and distribution for controllable person image synthesis. In *CVPR*, 2022. 1

[3] Yurui Ren, Xiaoqing Fan, Ge Li, Shan Liu, and Thomas H. Li. Neural texture extraction and distribution for controllable person image synthesis. In *CVPR*, 2022. 3, 4, 5

[4] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *CVPR*, 2020. 1, 3, 4, 5

[5] Jinsong Zhang, Kun Li, Yu-Kun Lai, and Jingyu Yang. Pise: Person image synthesis and editing with decoupled gan. In *CVPR*, 2021. 1, 3, 4, 5

[6] Pengze Zhang, Lingxiao Yang, Jian-Huang Lai, and Xiaohua Xie. Exploring dual-task correlation for pose guided person image generation. In *CVPR*, 2022. 1, 3, 4, 5

[7] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In *ECCV*, 2022. 1

[8] Xinyue Zhou, Mingyu Yin, Xinyuan Chen, Li Sun, Changxin Gao, and Qingli Li. Cross attention based style distribution for controllable person image synthesis. In *ECCV*, 2022. 3, 4, 5

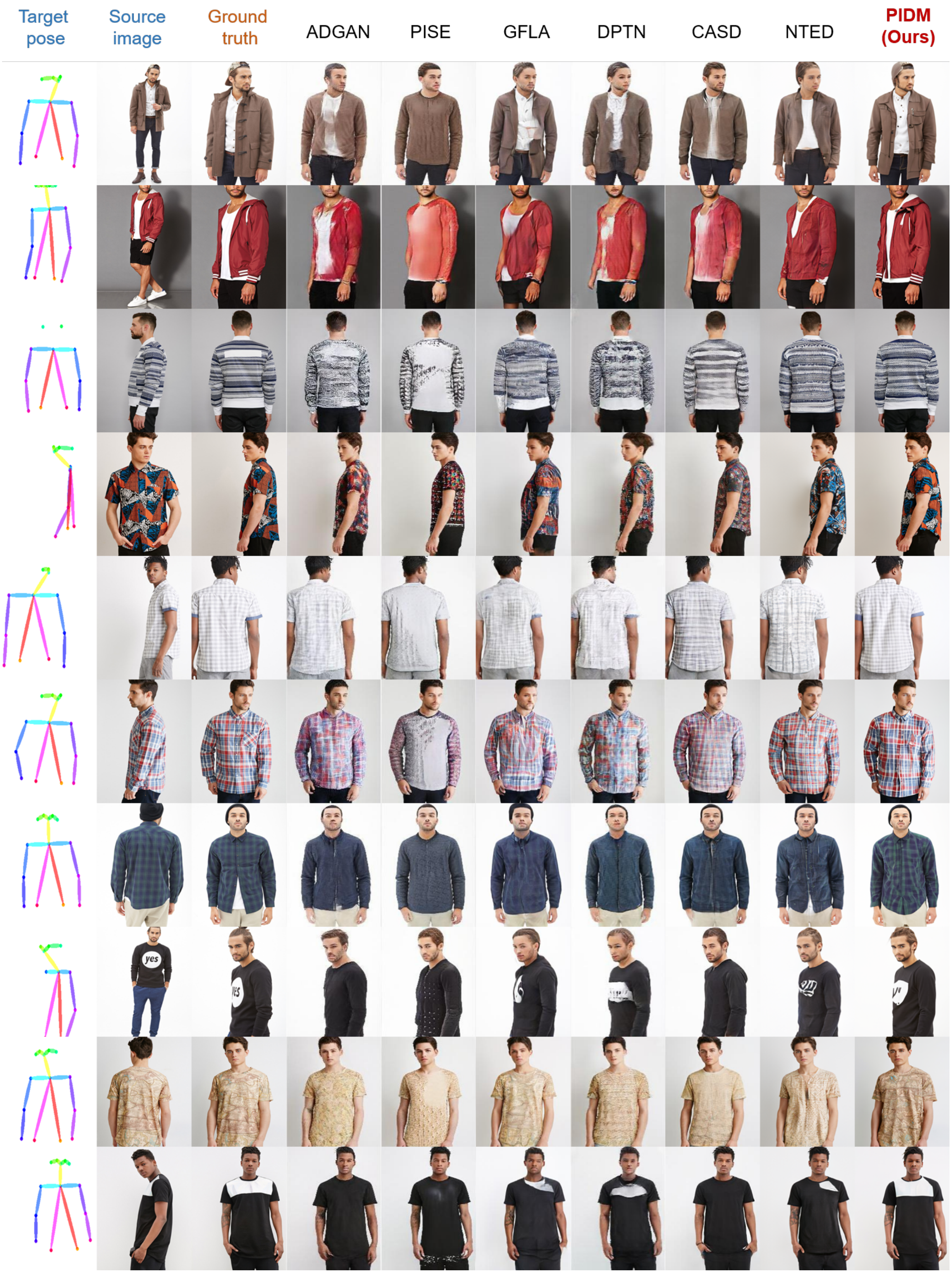

Figure 5. Additional qualitative comparisons with several state-of-the-art models such as ADGAN [1], PISE [5], GFLA [4], DPTN [6], CASD [8], NTED [3] and Ours on the DeepFashion dataset.

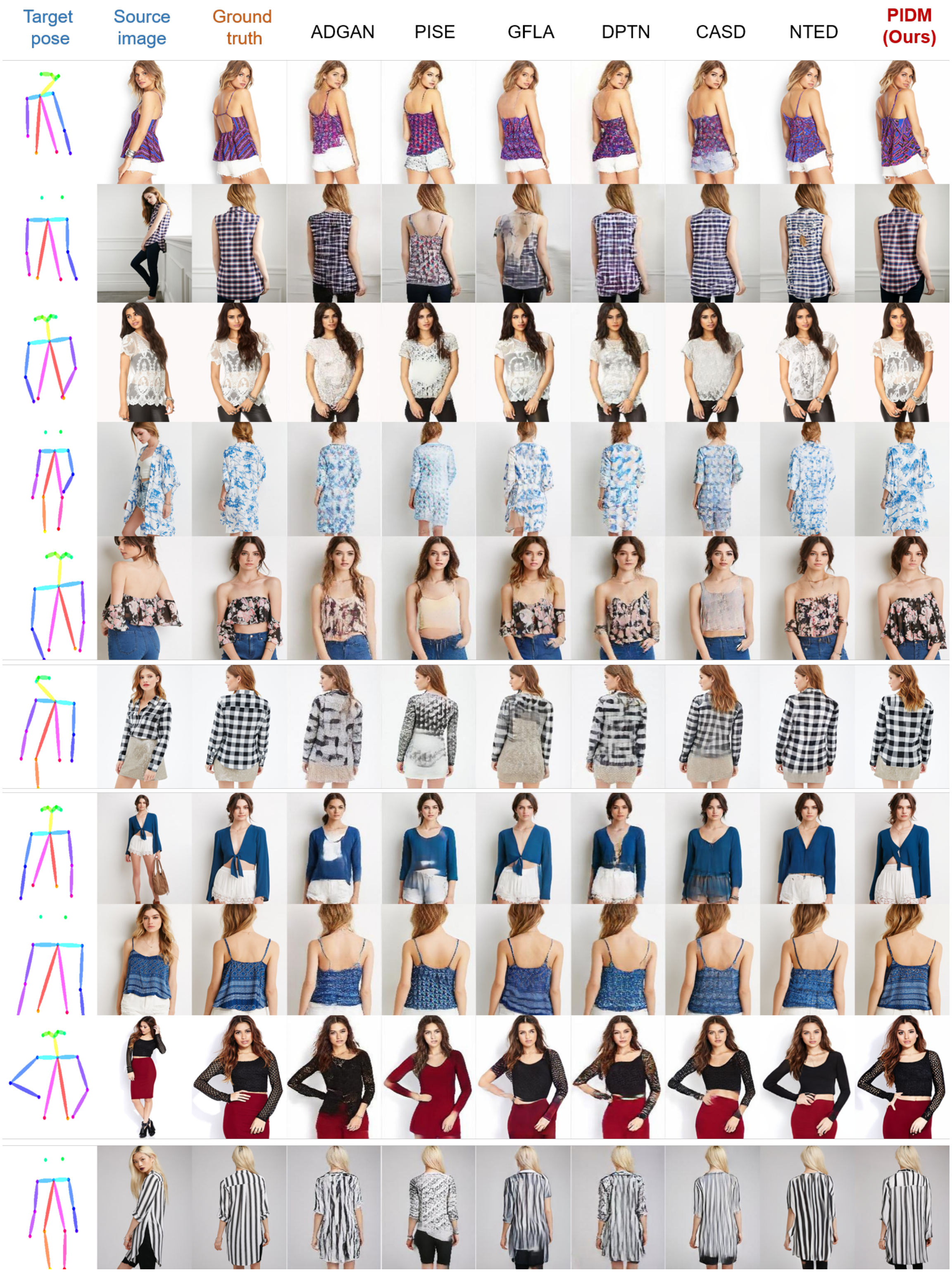

Figure 6. Additional qualitative comparisons with several state-of-the-art models such as ADGAN [1], PISE [5], GFLA [4], DPTN [6], CASD [8], NTED [3] and Ours on the DeepFashion dataset.

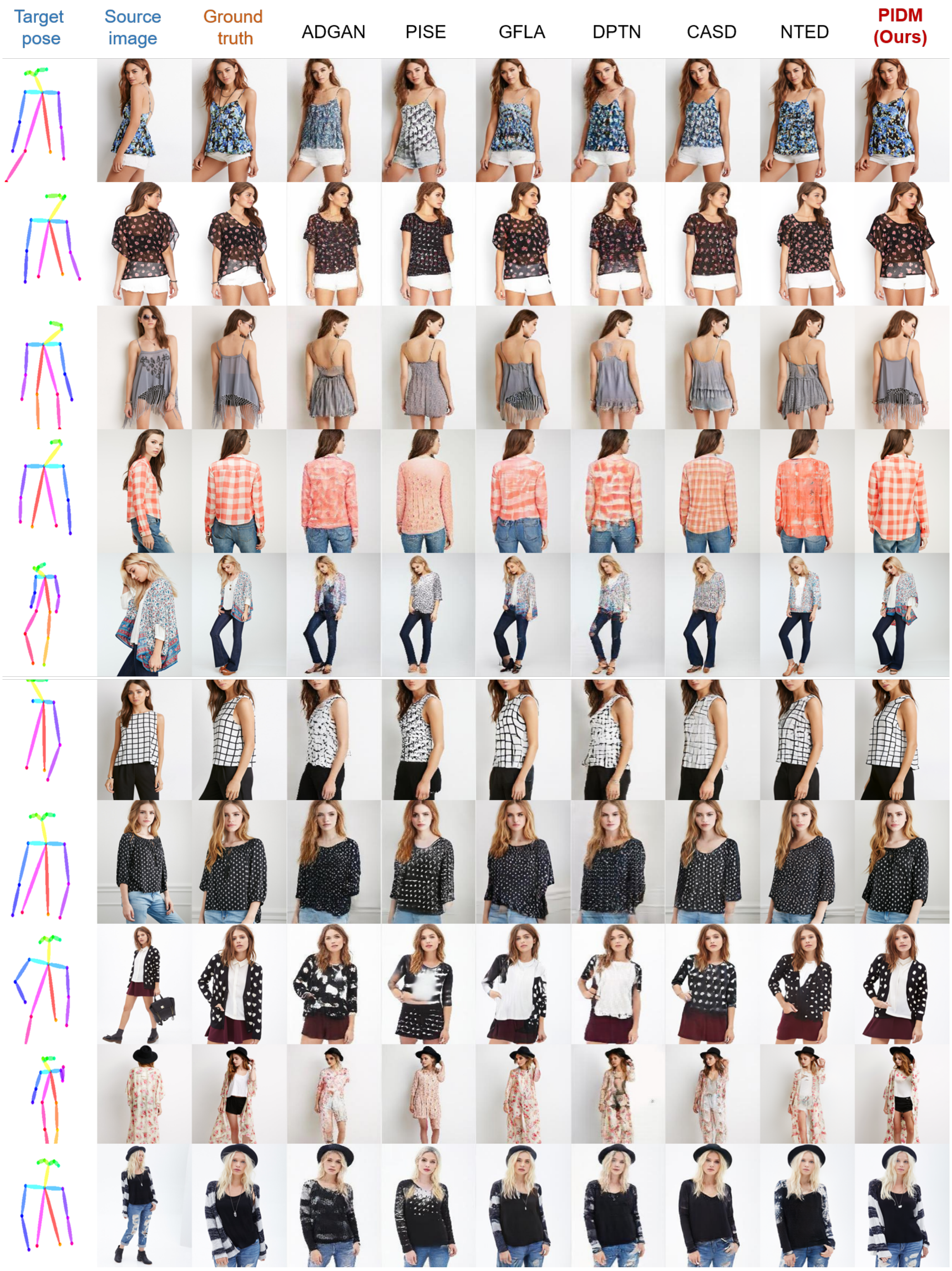

Figure 7. Additional qualitative comparisons with several state-of-the-art models such as ADGAN [1], PISE [5], GFLA [4], DPTN [6], CASD [8], NTED [3] and Ours on the DeepFashion dataset.