Multi-focal Conditioned Latent Diffusion for Person Image Synthesis

Supplementary Material

Method	NTED [6]	CASD [9]	PIDM [1]	CFLD [5]	Ours
Preferences	16.4%	11.0%	17.1%	12.9%	**42.6**%

Table 1. User Study about the preferences of generated images towards ground truths.

User Study. We conducted a user study to evaluate the image synthesis quality of various methods [1, 5, 6, 9], focusing on three key aspects: 1) texture quality, 2) texture preservation, and 3) identity preservation. We recruited 45 volunteers, most of whom are Ph.D. students specializing in deep learning and computer vision. Each participant was asked to answer 30 questions, selecting the method that best matched the ground truth based on the defined quality criteria. The results are listed in Tab. 1. Compared to other methods, our approach achieved the highest preference score of 42.6%, which is 25.5 percentage points higher than the second-best method. This indicates that our method excels in preserving identities and textures based on objective criterion.

Comparison of Editing. To compare the editing and its flexibility of our method with mask-based method [1, 3, 5], we build upon the concept of CFLD [5] to address the pose-variant appearance editing task, as demonstrated earlier. To enable the mask-based method to modify the corresponding regions, we introduce an additional denoising pipeline to blend the source image under a given pose. Initially, masks for the editing regions are extracted using a human parsing algorithm and then integrated into the sampling process. During sampling, the noise prediction, $\tilde{\epsilon}$, is decomposed into two components: ϵ^s, predicted by a UNet conditioned on the source image styles, and ϵ^{ref}, predicted by the same UNet conditioned on the target image styles. Both of the two componets is conditioned by the same give pose. Let $\tilde{\epsilon}_t$ be defined as follows:

$$\tilde{\epsilon}_t = m \cdot \epsilon_t^s + (1 - m) \cdot \epsilon_t^{ref}, \tag{1}$$

where ϵ_t^s and ϵ_t^{ref} is the noise at timestep t.

As shown in Fig. 2, 3, the method follows the same generation task by separately masking clothes, faces, and upper clothes. However, mask-based methods struggle to preserve facial and texture details under new poses.This limitation arises from the inherent inability of image-conditioned methods to accurately recover fine-grained details. Furthermore, the use of a provided mask introduces additional challenges, as generating precise masks for synthesized images remains non-trivial, often leading to artifacts at the edges in generated images. Moreover, restricting the masked region may adversely affect the preservation of cloth styles and categories, whereas our approach demonstrates superior retention of these attributes.

We present additional examples of our generated images in Fig. 4, 5, 6, illustrating the pose-variant editing setting. For clarity, the swapped texture maps are also provided to highlight the swapping procedures.

In addition, since the edited images feature combined clothing and identities, no ground truth exists in current datasets, making pixel-wise evaluation infeasible. Instead, we provide the quantitative comparison of our method using several perceptual benchmarks in Tab 2, which illustrates our method could generate more natural edited images.

Methods	Face Similarity ↑	Face Distance ↓	lpips ↓	CLIP score ↑
CFLD / Ours	0.274 / 0.341	28.2 / 26.6	0.273 / 0.268	0.893 / 0.903

Table 2. Perceptual comparison of Editing Performance

Generalization to diverse dataset. We further validate on 3 out-of-domain datasets without extra model training: 1) UBCFashion [8], 2) SHHQ [2], 3) Thuman [7]. We randomly selected some poses and characters from the datasets as input shown in 1. The results demonstrate consistent, appearance-preserving image generation.

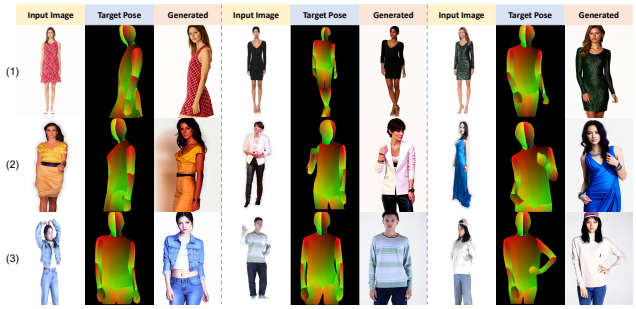

Figure 1. Results on other datsets.

Additional Qualitative Results. We conducted two additional qualitative experiments to demonstrate the generalization ability of our method. First, we generated person images under arbitrary poses randomly selected from the test set (Fig.7,8,9), and the results show that our method consistently preserves texture patterns and person identities from the source images, even retaining complex patterns and icons, with high-quality facial details. Second, we tested the method's adaptability to user-defined poses by rendering synthetic DensePose in real-time, where synthetic poses were rendered from SMPL [4] parameters estimated from the test set. The results (Fig.10,11) indicate that our method can generate plausible person images with correct textures and identities. Minor weaknesses were observed in the hands and boundary regions due to differences in the generated and estimated DensePose. This problem can be mitigated through finetuning.

Computation Complexity of MFCA module. Our method improves performance by introducing only 5.8% more trainable parameters compared to baseline B1, where the MFCA modules only extend around **19M** parameters and the face projector introduces **76M** parameters. We have provided additional information , validating on A6000 GPUs. We also compare the cost with CFLD in Tab.3. Our method adopts a ControlNet-like structure, which minimally increases the inference cost while reducing the training time.

Methods	GPU hours (H)	inference memory (G)	inference time (s)
CFLD [5]	353.6	6.5	3.55
B1	45.2	10.8	4.35
MCLD (Ours)	54.0	13.9	4.46

Table 3. Complexity Comparison of baselines.

References

[1] Ankan Kumar Bhunia, Salman Khan, Hisham Cholakkal, Rao Muhammad Anwer, Jorma Laaksonen, Mubarak Shah, and Fahad Shahbaz Khan. Person image synthesis via denoising diffusion model. In *CVPR*, 2023. 1

[2] Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. In *ECCV*, 2022. 1

[3] Xiao Han, Xiatian Zhu, Jiankang Deng, Yi-Zhe Song, and Tao Xiang. Controllable person image synthesis with pose-constrained latent diffusion. In *ICCV*, 2023. 1

[4] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J. Black. SMPL: A skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):248:1–248:16, 2015. 1

[5] Yanzuo Lu, Manlin Zhang, Andy J Ma, Xiaohua Xie, and Jianhuang Lai. Coarse-to-fine latent diffusion for pose-guided person image synthesis. In *CVPR*, 2024. 1, 2

[6] 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

[7] Tao Yu, Zerong Zheng, Kaiwen Guo, Pengpeng Liu, Qionghai Dai, and Yebin Liu. Function4d: Real-time human volumetric capture from very sparse consumer rgbd sensors. In *CVPR*, 2021. 1

[8] Polina Zablotskaia, Aliaksandr Siarohin, Bo Zhao, and Leonid Sigal. Dwnet: Dense warp-based network for pose-guided human video generation. *BMVC*, 2019. 1

[9] 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*. Springer, 2022. 1

Figure 2. Comparison of appearance editing between ours and CFLD. The 2nd, 3rd, 4th rows show the editing of clothes, face, upper cloth region, respectively.

Figure 3. Comparison of appearance editing between ours and CFLD. The 2nd, 3rd, 4th rows show the editing of clothes, face, upper cloth region, respectively.

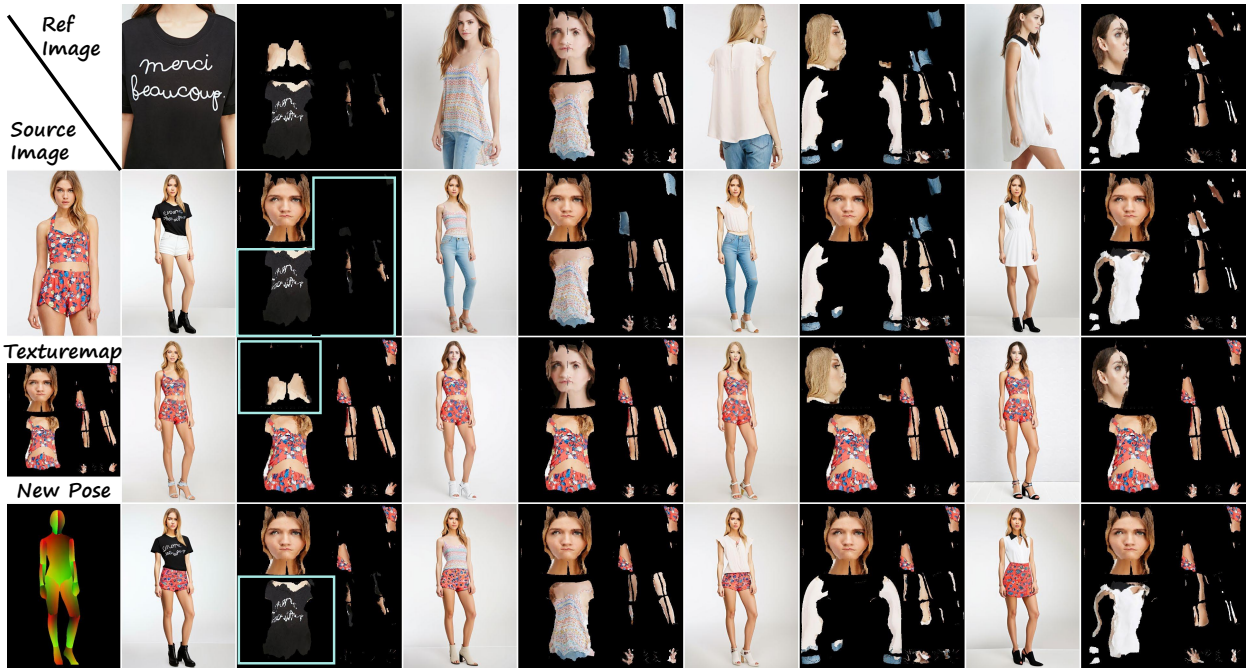

Figure 4. Additional results of our editings. We show the texture map on the right to illustrate our swapped regions in texture map. The editing regions are labelled using light green bounding box.

Figure 5. Additional results of our editings. We show the texture map on the right to illustrate our swapped regions in texture map. The editing regions are labelled using light green bounding box.

Figure 6. Additional results of our editings. We show the texture map on the right to illustrate our swapped regions in texture map.The editing regions are highlighted with light green bounding boxes.

Figure 7. Additional results on arbitrary poses from the test set.

Figure 8. Additional results on arbitrary poses from the test set.

Figure 9. Additional results on arbitrary poses from the test set.

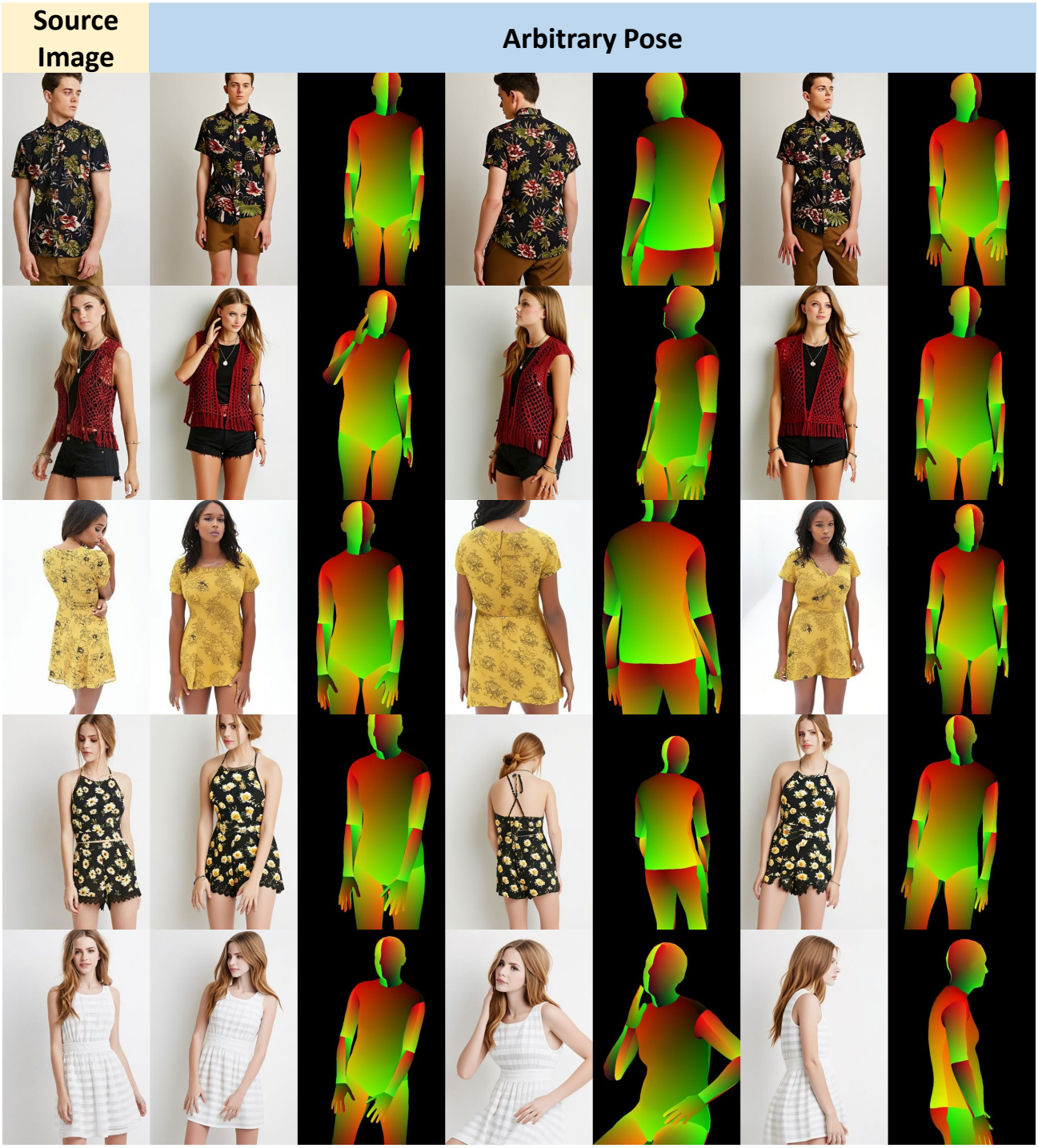

Figure 10. Additional results on rendered Densepose. The Densepose is rendered by user-defined SMPL parameters.

Figure 11. Additional results on rendered Densepose. The Densepose is rendered by user-defined SMPL parameters.