Liquid Warping GAN: A Unified Framework for Human Motion Imitation, Appearance Transfer and Novel View Synthesis (Supplementary Material)

1. Details of Network Architecture

All codes and dataset are available on this site[1].

Generator. The generator, as shown in Fig. 1, consists of three streams. One of them is named G_{BG} for background inpainting, and the other two streams are source identity stream, namely G_{SID} and transfer stream, namely G_{TSF}. The G_{BG} is a ResNet [2] and it contains three parts, encoder, residual blocks and decoder. The encoder firstly uses a $7 \times 7 \times 64$ stride-1 convolution. Then, it follows three 3×3 stride-2 convolutions with 128, 256 and 512 filters, respectively. 6 residual blocks with $3 \times 3 \times 512$ convolutions are used. The decoder contains three 3×3 transposed convolutions whose number of filters are 256, 128 and 64 respectively, are thereby utilized to upscale the feature resolution. The G_{SID} and G_{TSF} have a similar architecture with G_{BG}, but they contain additional U-Net like skip connections. For G_{BG} stream, it directly regresses the inpainted images $\hat{I}_{bs}$ in the final convolution layer. For G_{SID} and G_{TSF} steam, they regress a color map P and an attention map A in the final convolution layer. We apply instance normalization and ReLU on all convolutions and transposed convolutions, except for the last layer.

Discriminator. It firstly uses a $4 \times 4 \times 64$ stride-1 convolution and is followed by three stride-2 convolutions. All these convolutions utilize instance normalization and LeakyReLU with 0.2 negative slopes. The last layer is $4 \times 4 \times 1$ convolution, which regresses the final score.

2. Dataset Details

We summarize the Impersonator (iPER) dataset in classes of actions, styles of clothes, weight and height distributions of actors. The details are illustrated in Fig. 3, and Fig. 4 shows some examples.

3. Human Motion Imitation

3.1. More Results of Human Motion Imitation

We illustrate more results of our methods in Fig. 5 and Fig. 6. Our method could produce high-fidelity images that preserve the face identity, shape consistency and clothes details of the source.

Table 1. User case study of iPER and DeepFashion datasets [4]. The numbers indicate the percentage of volunteers who favor the results of our proposed LWB over competing for other methods, including PG2 [5], SHUP [1], DSC [6] and our baselines, such as W_C, W_T and W_F.

W_{LWB} *vs.*	PG2	SHUP	DSC	W_C	W_T	W_F
iPER	0.91	0.79	0.82	0.89	0.75	0.74
DeepFashion	0.96	0.98	0.95	0.93	0.56	0.64

3.2. Comparison of Other Methods

We compare the performance of our method with that of existing methods, including PG2 [5], SHUP [1], and DSC [6]. More results are illustrated in Fig. 7. It reflects that our method could produce more realistic-looking results in the large layout of reference pose, and is more powerful to preserve the source information, in terms of clothes details, face identity and shape consistency.

3.3. Ablation and User Case Study

We design three baselines with different warping strategies, including early concatenation W_C (traditional CGAN), texture warping W_T and feature warping W_F . The details of these three baselines are shown in Figure 2 of the original paper. All baselines use the same 3D guided inputs with the same network architecture, except for the warping block. The quantitative results of our proposed dataset are shown in Table 1 of the original paper. For qualitative evaluation, we conduct a user case study on both our dataset and DeepFashion [4]. Specifically, we show the volunteers source image, reference image, and two generated outputs from different methods. We randomly sample 600 questions and let each question be answered by 3 different volunteers (30 in total) in each dataset. The results of the user case study are shown in Table 1 in this document, and our method with LWB outperforms other baselines. In addition, exemplar visualizations are shown in Figure 1 in this document.

[1] https://svip-lab.github.io/project/impersonator.html

3.4. Failure Case Analysis

There are two main types of failure cases of our methods. The one, as shown in the first two rows of Fig. 8, is that source image contains a large area of self-occlusion which introduces ambiguity, and thereby results in a bad synthesized image. The other occurs when Body Recovery Module (HMR) [3] fails, as illustrated in the last two rows of Fig. 8.

4. Human Appearance Transfer

We also illustrate more results of our methods in Fig. 9 and Fig. 10. Our method could produce high-fidelity and decent images that preserve the face identity and shape consistency of the source image and keep the clothes details of the reference image.

5. Discussion of Generalization

The ability of generalization of our method can be specified in the following two aspects: the foreground (human) and the background. For the foreground, our method has a certain degree of ability to generate a decent foreground part, as shown in Figure 1 in this document. While for the background, the background network G_{BG} is trained in a self-supervised way, which seems to overfit the background from the training set, as shown in Figure 1. One way to improve the ability of background generalization is to use additional images, such as Place2 dataset, as the auxiliary loss L_{aux} in the training phase. Specifically, in each training iteration, we sample mini-batch images from Place2 dataset, denoted as L_{aux}, add human body silhouettes to them, and denote the mask images as $\hat{I}_{aux}$. We use the paired ($\hat{I}_{aux}$, I_{aux}) images with a perceptual loss to update parameters in the G_{BG} network. The L_{aux} loss indeed improves the generalization of background inpainting, as shown in Figure 1. It is worth noting that for a fair comparison, we do not use this trick in experiments when comparing our method with other baselines.

References

[1] Guha Balakrishnan, Amy Zhao, Adrian V. Dalca, Frdo Durand, and John Guttag. Synthesizing images of humans in unseen poses. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[2] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778, 2016.

[3] Angjoo Kanazawa, Michael J Black, David W Jacobs, and Jitendra Malik. End-to-end recovery of human shape and pose. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[4] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[5] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Advances in Neural Information Processing Systems*, pages 405–415, 2017.

[6] Aliaksandr Siarohin, Enver Sangineto, Stphane Lathuilire, and Nicu Sebe. Deformable gans for pose-based human image generation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

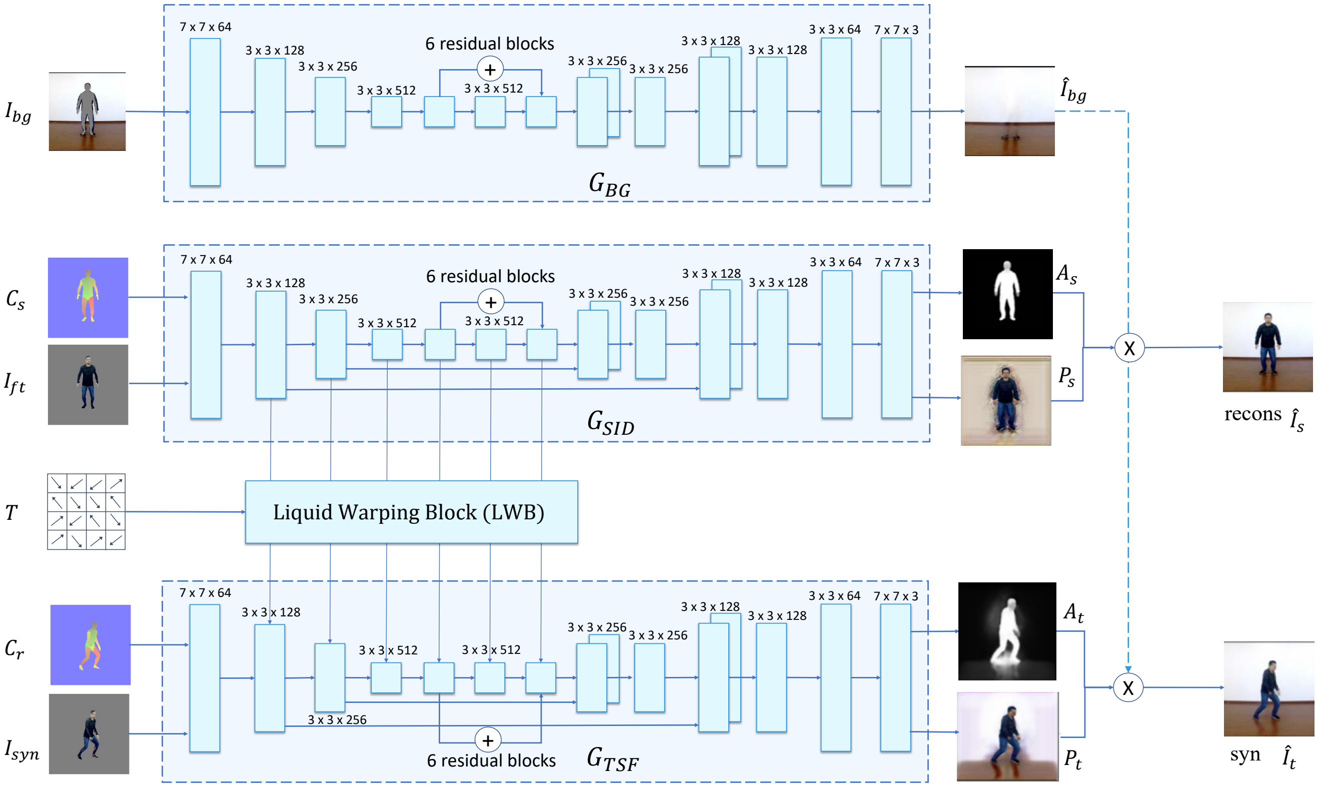


Figure 1. Details of our Liquid Warping GAN (generator). It consists of three streams, G_{BG}, G_{SID} and G_{TSF}. They have similar network architecture, and they do not share parameters.

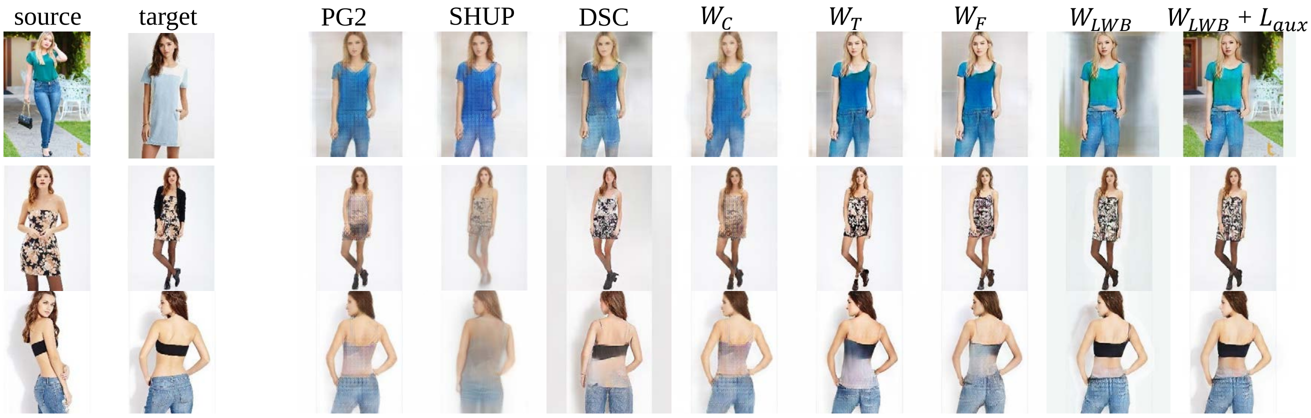
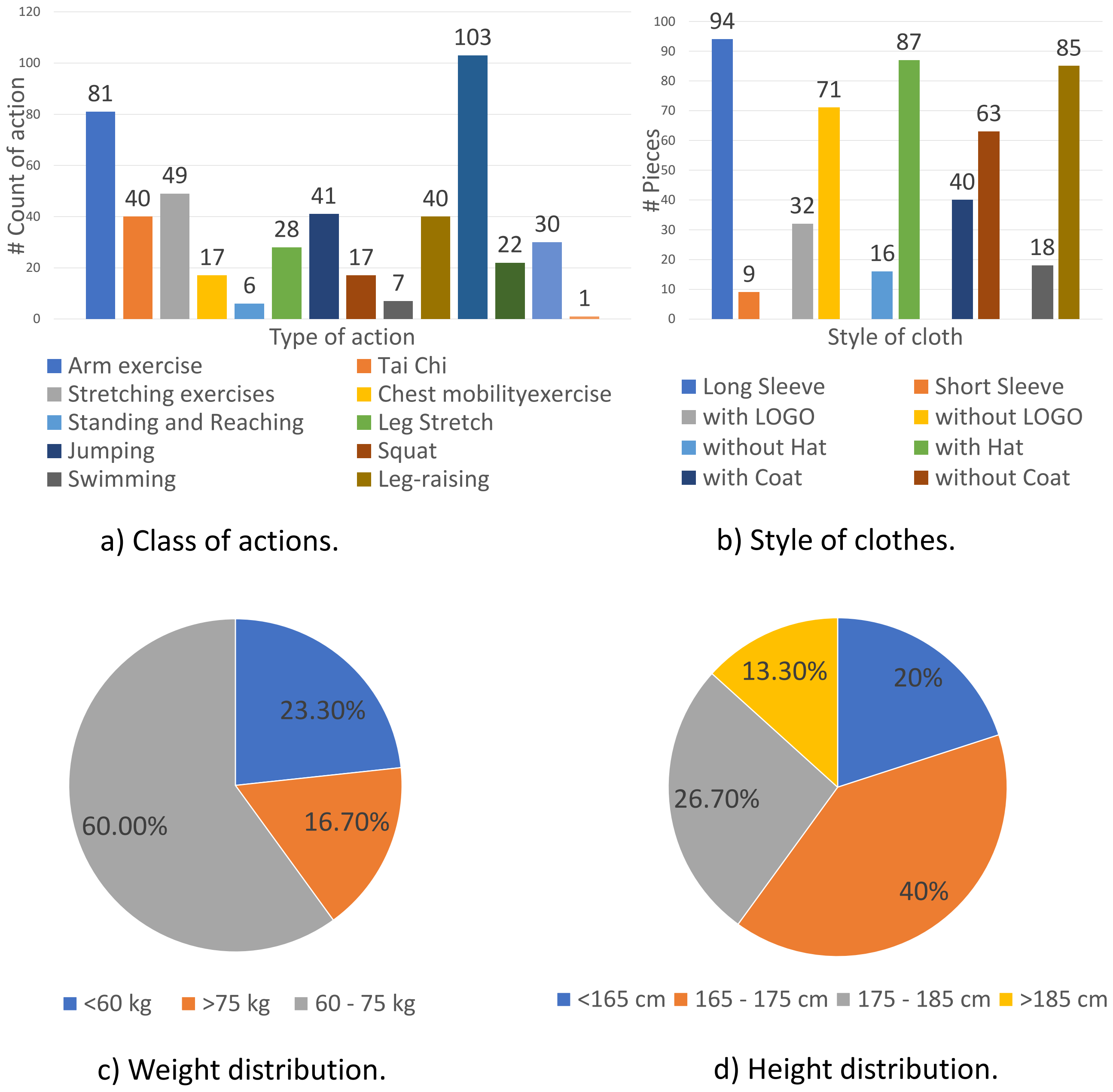


Figure 2. Examples of our method of human appearance transfer. Source images come from iPER dataset and reference images come from DeepFashion dataset [4].



a) Class of actions.

b) Style of clothes.

c) Weight distribution.

d) Height distribution.

Figure 3. Details of Impersonator (iPER) dataset. a) shows the class of actions and their number of occurrences. b) shows the styles of clothes. c) and d) are the distributions of weight and height of all actors. There are 30 actors in total.

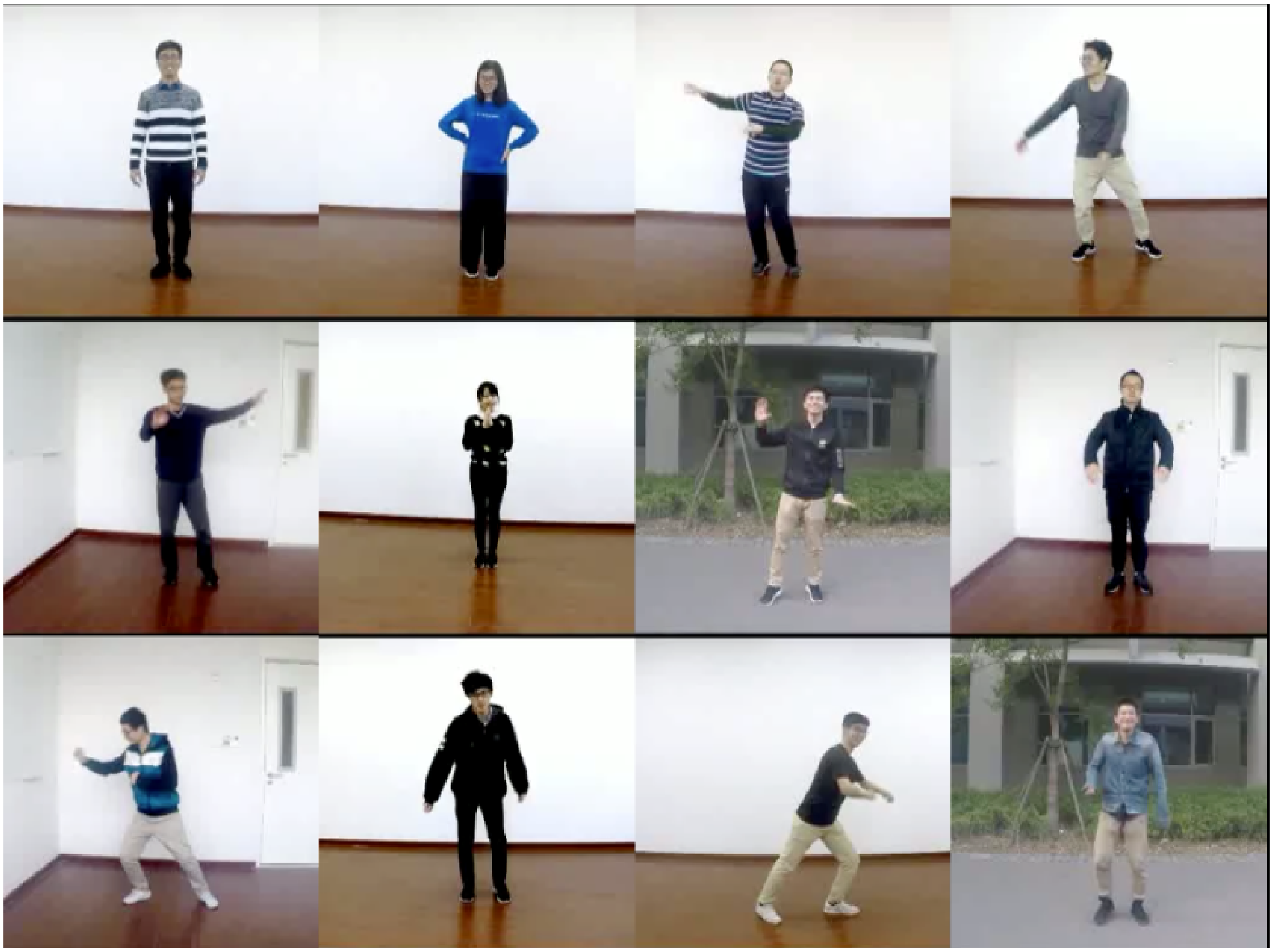


Figure 4. Example of frames in Impersonator (iPER) dataset.

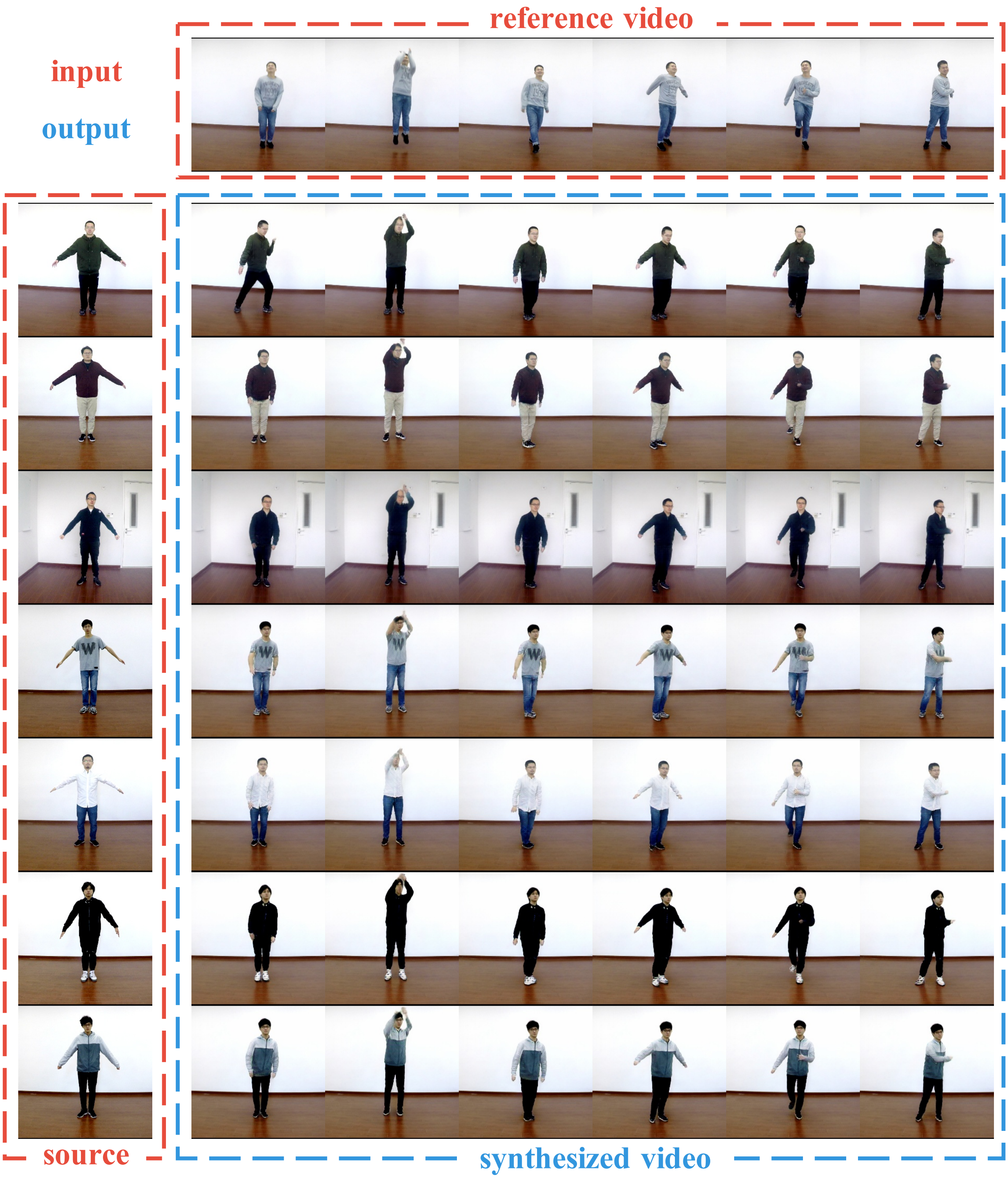


Figure 5. Examples of motion imitation from our proposed methods on the iPER dataset (zoom-in for the best of view).

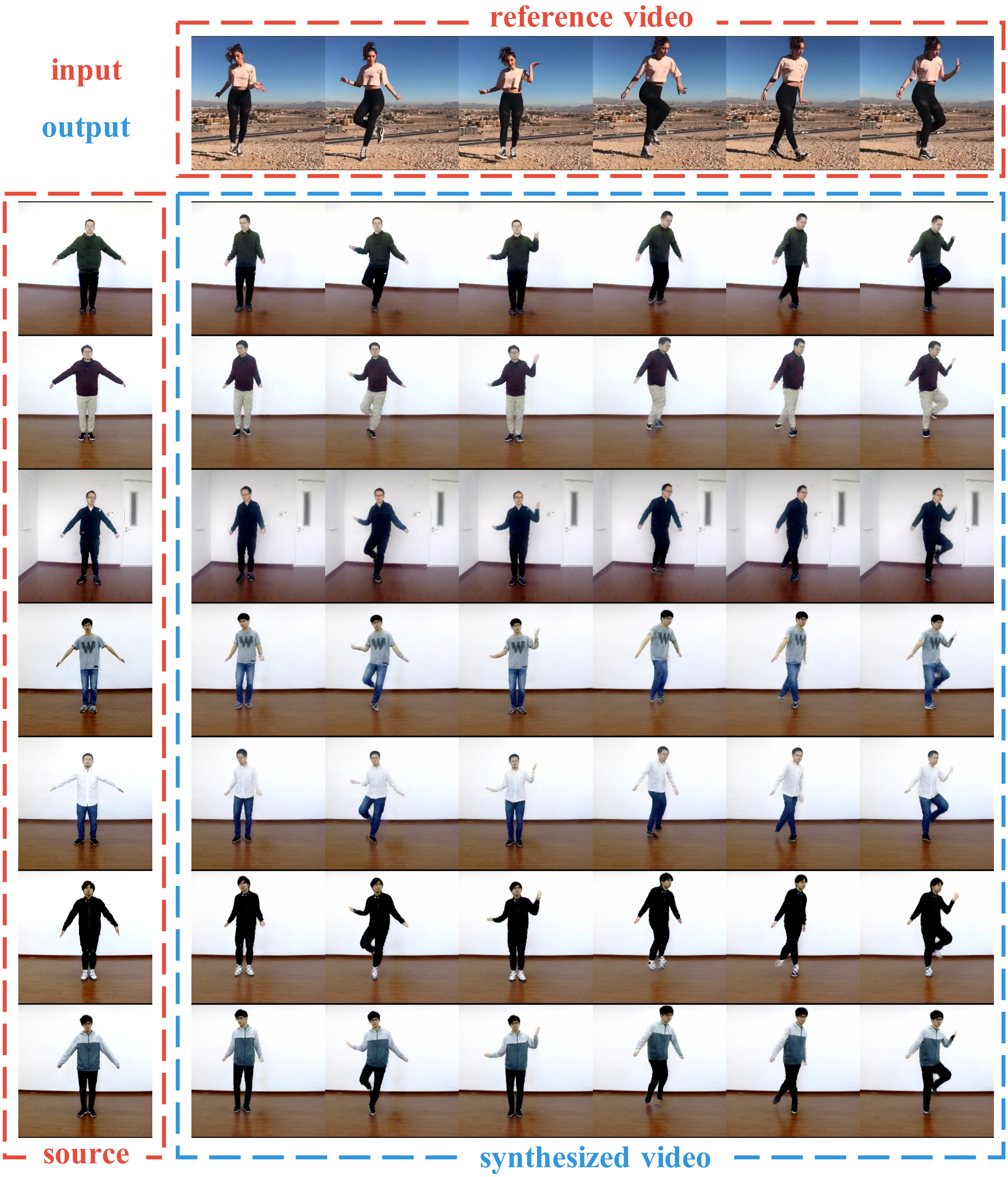


Figure 6. Examples of motion imitation from our proposed methods on the iPER dataset (zoom-in for the best of view).

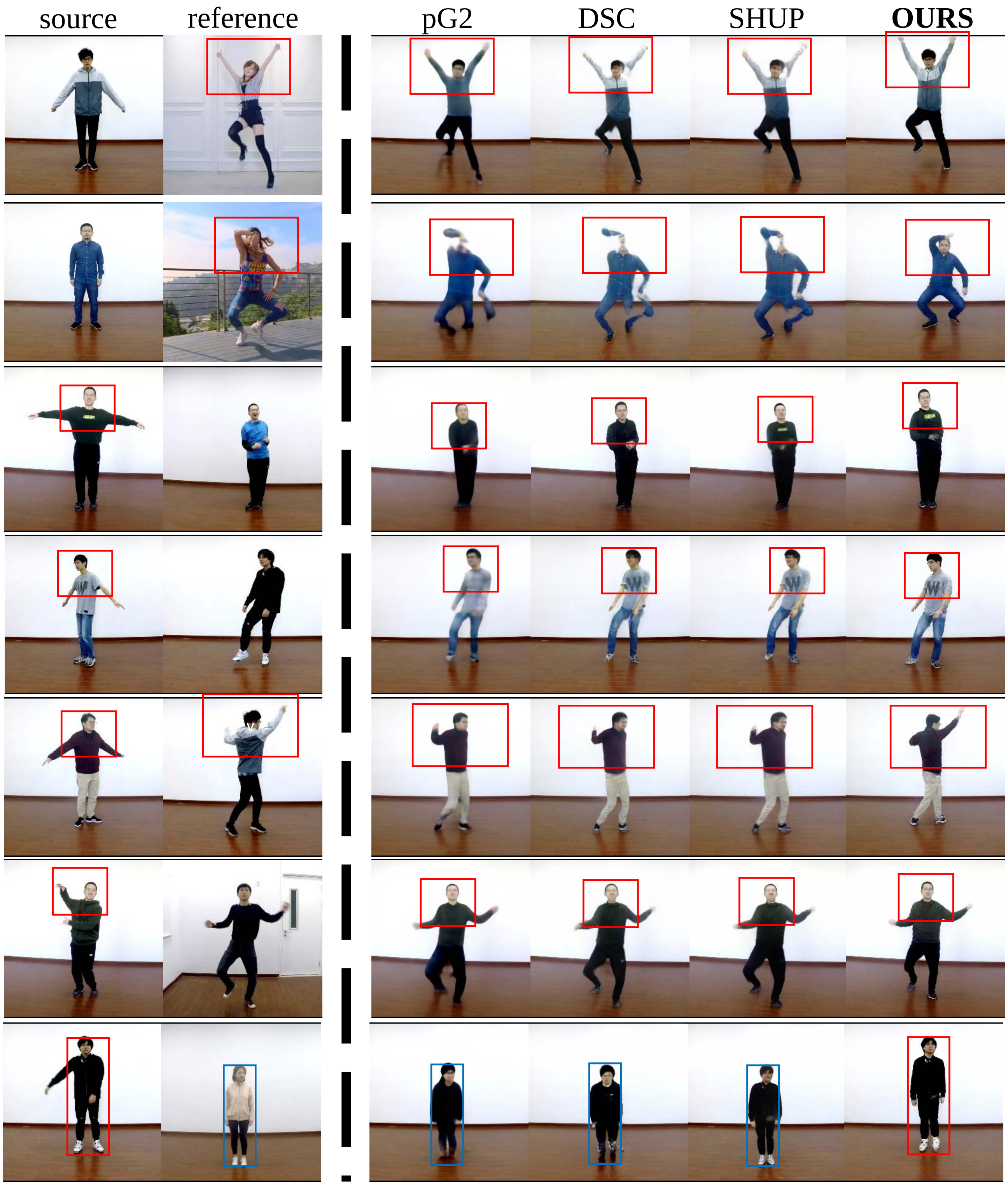


Figure 7. Comparison of our method with others of motion imitation on the iPER dataset (zoom-in for the best of view). 2D pose-guided methods pG2 [5], DSC [6] and SHUP [1] cannot preserve the clothes details, face identity and shape consistency of source images. We highlight the details by red and blue rectangles.

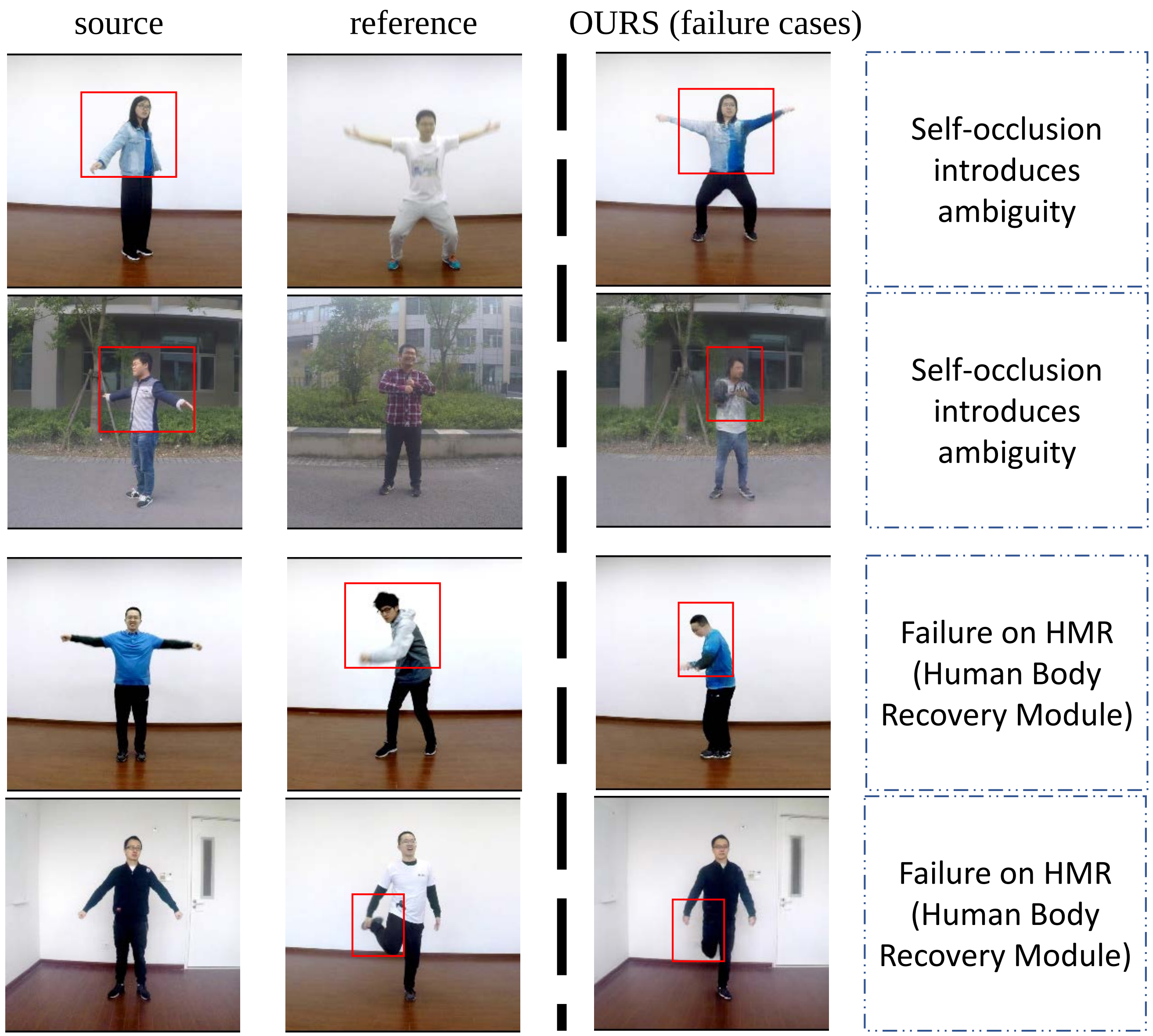

source
reference
OURS (failure cases)
Self-occlusion introduces ambiguity
Self-occlusion introduces ambiguity
Failure on HMR (Human Body Recovery Module)
Failure on HMR (Human Body Recovery Module)

Figure 8. Failure cases of our methods. They occur when the source image contains a large area with ambiguous self-occlusion or the HMR [3] fails.

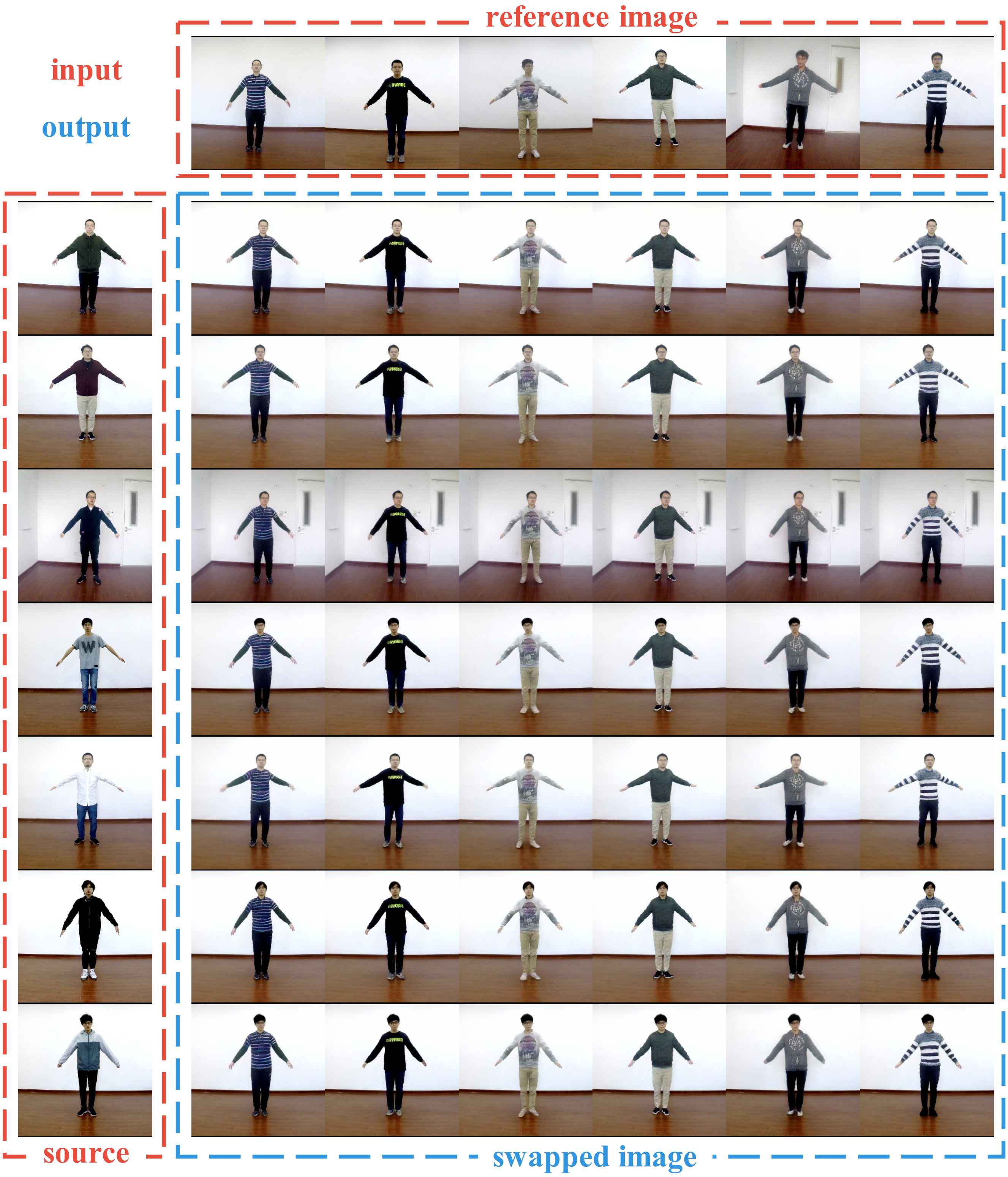


Figure 9. Examples of our method of human appearance transfer in iPER dataset (zoom-in for the best of view).

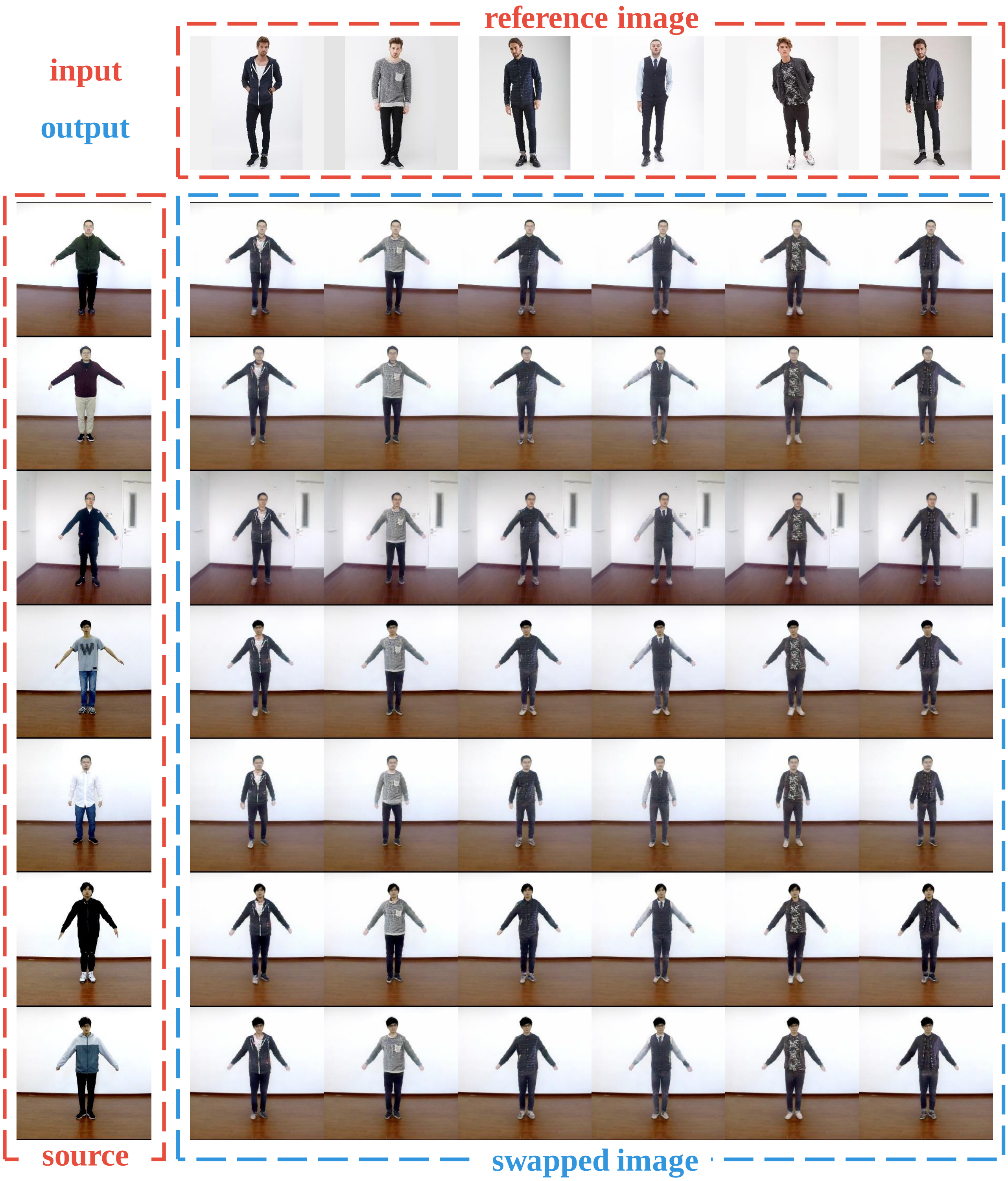


Figure 10. Examples of our method of human appearance transfer. Source images come from iPER dataset and reference images come from DeepFashion dataset [4].