Unsupervised Person Image Generation with Semantic Parsing Transformation (Supplementary Material)

Sijie Song [1], Wei Zhang [2], Jiaying Liu [1], Tao Mei [2]
[1] Institute of Computer Science and Technology, Peking University, Beijing, P.R. China
[2] JD AI Research

In this supplementary, we provide more details for implementation in Sec. 1 and Sec. 2. We then show additional pose-guided image generation comparisons in Sec. 3 and conduct a human perceptual study in Sec. 4. In Sec. 5, more application examples are presented. Sec. 6 discusses the effect of semantic map quality on the generated results.

1. Network Architecture

In Fig. 1, we provide details on network architectures of our semantic generative network (H_S) and appearance generative network (H_A) for low-resolution images (128×64 in Market-1501 [9]), respectively. For the discriminators, D_S, D_A and D_F share the same architecture $C^2_{64} - C^2_{128} - C^2_{256} - C^2_{512} - C^2_1$, where C^s_m indicates a convolution layer with m filters and stride s.

2. Training Scheme

In our work, we first pre-train the semantic generative network and appearance generative network in two-stage, then train them in an end-to-end manner.

DeepFashion [5]. In two-stage training, the semantic generative network is trained with images for $30k$ iterations in the resolution of 128×128. The appearance generative network is trained on images of 128×128 for $150k$ iterations. Then we adopt progressive training strategy [4] to train the appearance generative network with images of 256×256 for $300k$ iterations with upsampled semantic maps.

In end-to-end training, with fixed pre-trained semantic generative network, we train the appearance generative network for $70k$ iterations on the resolution of 128×128. Similarly, the appearance generative network is then trained for $70k$ iterations for images in 256×256 with upsampled semantic maps using progressive training strategy [4]. Finally, they are jointly trained for another $70k$ iterations. The mini-batch size for DeepFashion is 4.

Market-1501 [9]. In two-stage training, we train the semantic generative network for $100k$ iterations and then train the appearance generative network for $50k$ iterations.

In end-to-end training, with fixed pre-trained semantic generative network, the appearance generative network is trained for $50k$ iterations. The joint optimization is performed with another $50k$ iterations. The mini-batch size for Market-1501 is 16.

3. More Comparisons with State-of-the-Arts

In Fig. 2 to Fig. 9, we show more results compared with state-of-the-art methods on DeepFashion [5] and Market-1501 datasets [9]. Our method is especially superior in keeping clothing textures and generates better body shapes. Note that besides the state-of-the-art methods in the main paper, we further compare our method with V-UNet [2], which does not require paired data in the training process. However, their results are less faithful to the condition images in textures and colors (especially in the 2nd and 7th rows in Fig. 5), which is mainly because they are generated from highly compressed features.

4. User Study

We further implement a user study to evaluate our pose-guided person image generation results compared with other state-of-the-arts. For each dataset, we perform pairwise A/B tests to 35 volunteers, and everyone is given 200 pairs that are randomly selected from the results. In each pair, the images are in random order, one of which is our result while the other is from the compared method. Volunteers are asked to select which result is better without time limitation, considering: (1) correctly change the pose of the person in the condition images, (2) correctly preserve the clothing attributes (*i.e.*, textures, colors, clothing types) from the condition images to the target images, (3) natural and photo-realistic visual quality.

Table 1 shows our user-study results for mean and variance values. The mean values indicate the percentages of volunteers that select our method as better results in pairwise comparisons. For example, about 92.20% volunteers think our method generates better images than PG^2 [6]. The variance values indicate how volunteers think differently for the given pairs. The results in Table 1 illustrate that our

Table 1: Results for user study on pose-guided person image generation. The mean values indicate the percentages of volunteers that select our method as better results in pairwise comparisons.

	PG^2 [6]		Def-GAN [8]		UPIS [7]		V-UNet [2]	
	mean	var	mean	var	mean	var	mean	var
DeepFashion	92.20%	0.0070	64.61%	0.0364	97.90%	0.0011	68.55%	0.0069
Market-1501	78.55%	0.0188	67.23%	0.0215	86.17%	0.0117	73.31%	0.0463

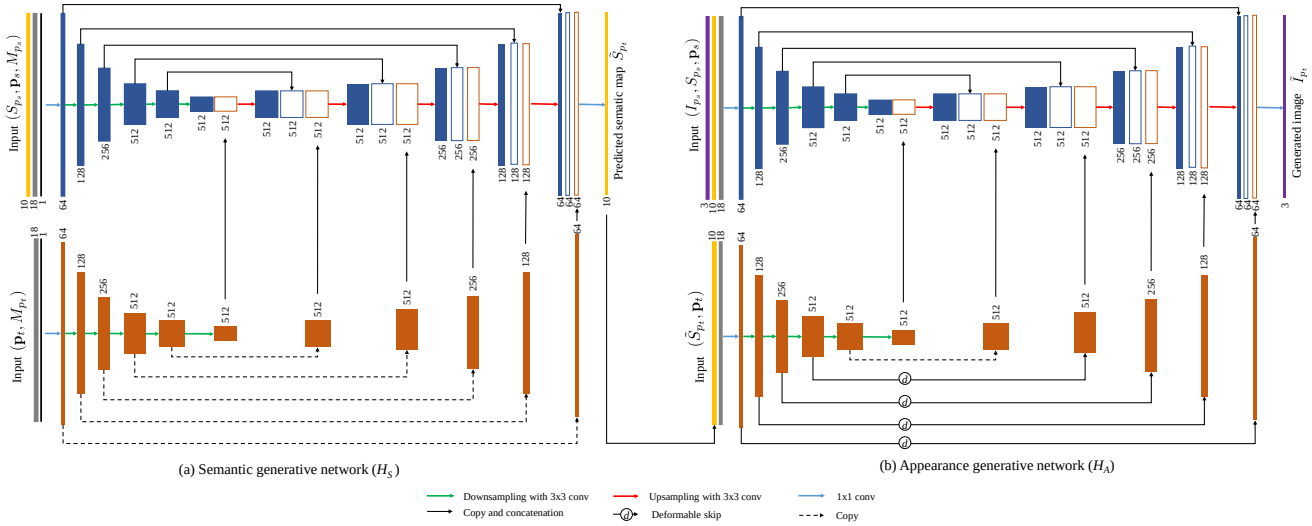

Figure 1: Detailed network architectures for images of 128×64 in Market-1501 [9]. In appearance generative network, we employ the deformable skips in [8].

method effectively generates images with more pleasant visual quality than state-of-the-arts.

5. More Application Examples

Clothing texture transfer. More bidirectional texture transfer results can be found in Fig. 10. We compare our results with image analogy [3] and neural doodle [1]. Our method successfully transfers the clothing textures while keeping the clothing types. Besides, natural faces can be generated with our method, which leads to more pleasant visual results.

Controlled image manipulation. In Fig. 11, we show more results on controlled image generation and comparisons with image analogy [3] and neural doodle [1]. By modifying semantic maps, we can change the clothing types to the desired layout while well keeping the clothing textures.

6. Effects of Semantic Map Quality

In this section, we discuss the effect of semantic map quality on the generated results. Since we do not have ground truth semantic labels for DeepFashion and Market-1501, we are not able to give quantitative analysis on how good the semantic maps have to be. Instead, we conducted experiments with semantic maps under different quality, by downsampling them with different scales, as shown below Fig. 12 (2nd row). Our generated results (1st row) rely on the semantic maps. The errors in semantic maps will result in unrealistic results, which also can be observed in Fig.10 in the main paper.

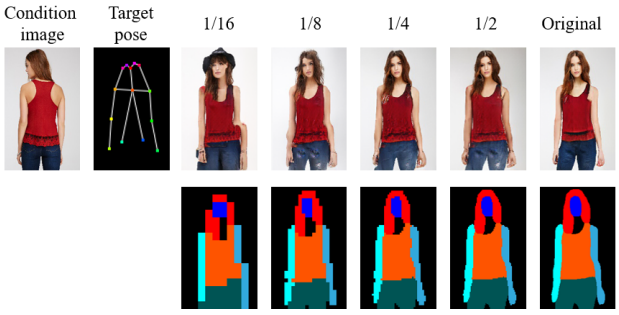

Figure 12: Effects on quality of semantic maps.

References

[1] Alex J. Champandard. Semantic style transfer and turning two-bit doodles into fine artworks. 2016. 2, 12, 13

[2] Patrick Esser, Ekaterina Sutter, and Björn Ommer. A variational u-net for conditional appearance and shape generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 1, 2, 7, 11

[3] Aaron Hertzmann. Image analogies. *Proc Siggraph*, 2001. 2, 12, 13

[4] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 1

[5] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 1, 4, 5, 6, 7

[6] Liqian Ma, Xu Jia, Qianru Sun, Bernt Schiele, Tinne Tuytelaars, and Luc Van Gool. Pose guided person image generation. In *Proc. Advances in Neural Information Processing Systems*, 2017. 1, 2, 4, 8

[7] Albert Pumarola, Antonio Agudo, Alberto Sanfeliu, and Francesc Moreno-Noguer. Unsupervised person image synthesis in arbitrary poses. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 6, 10

[8] Aliaksandr Siarohin, Enver Sangineto, Stéphane Lathuilière, and Nicu Sebe. Deformable gans for pose-based human image generation. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2, 5, 9

[9] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian. Scalable person re-identification: A benchmark. In *Proc. IEEE Int'l Conference on Computer Vision*, 2015. 1, 2, 8, 9, 10, 11

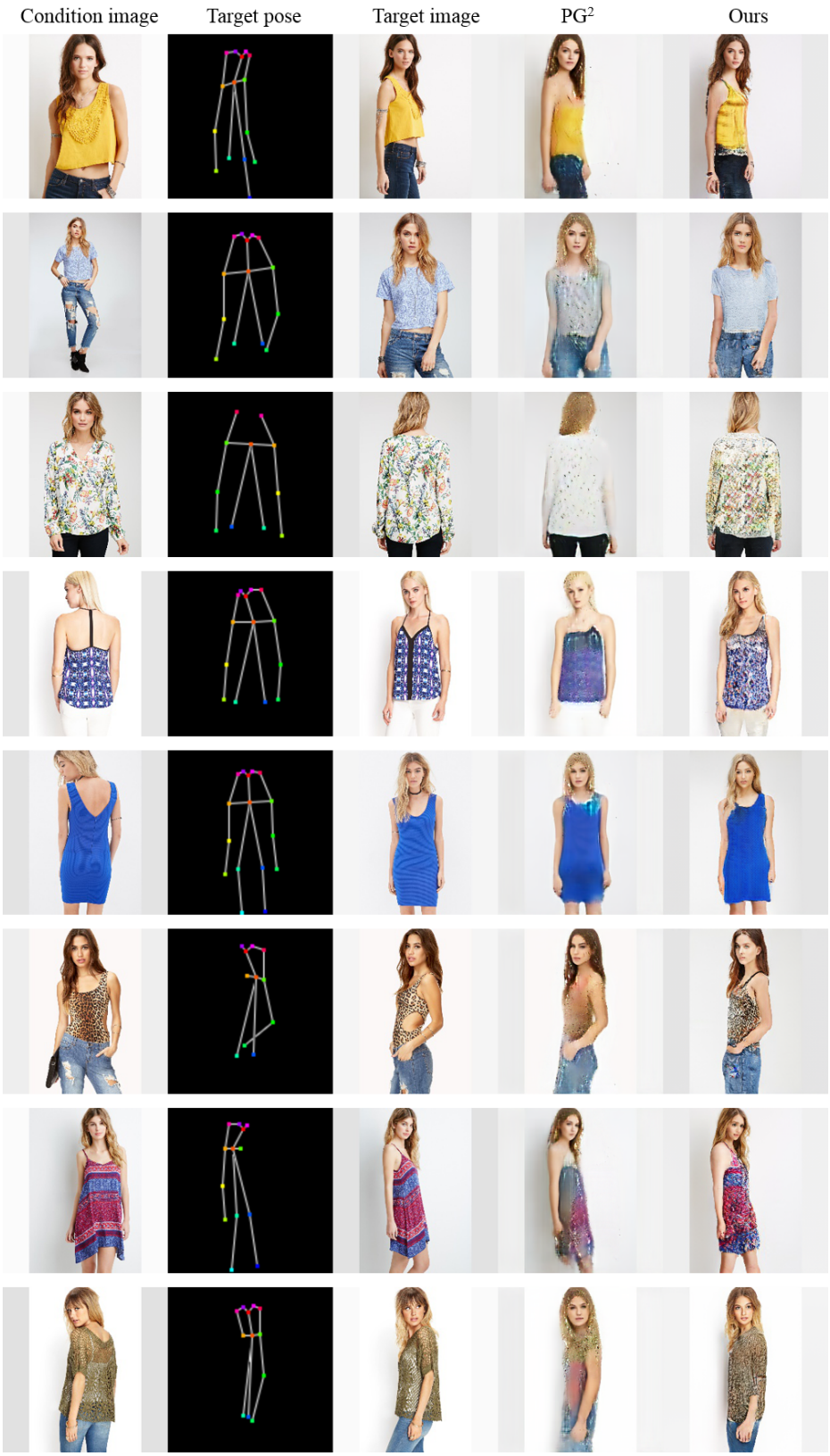

Figure 2: Comparisons with PG^2 [6] on DeepFashion [5]. Our method generates more natural images with less artifacts, and is able to preserve clothing textures. Zoom in for details.

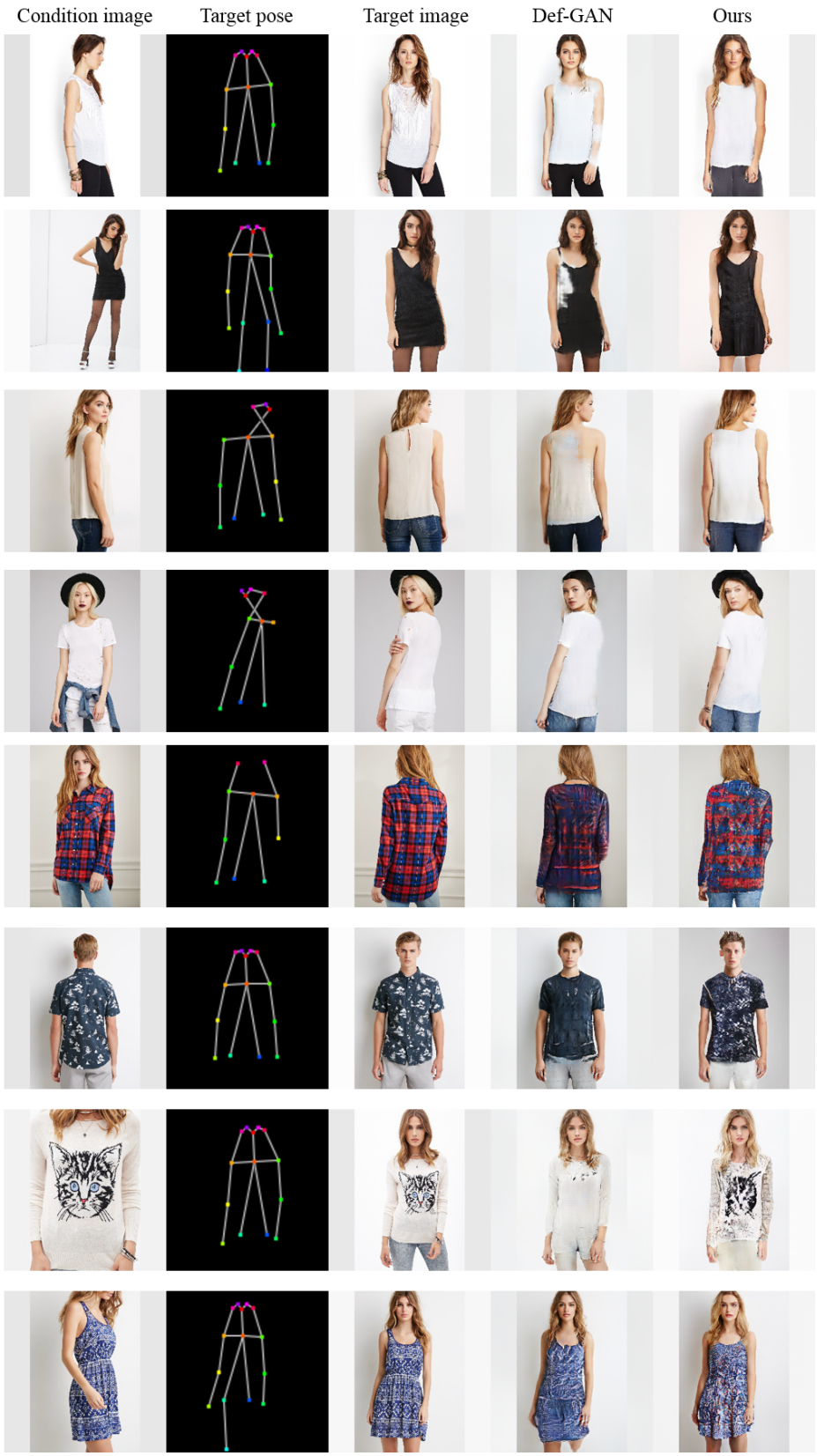

Figure 3: Comparisons with Def-GAN [8] on DeepFashion [5]. Our method successfully renders the textures for the unseen parts in the condition images (the 1st, 3rd rows) and is superior in preserving clothing textures (4th-8th rows). Zoom in for details.

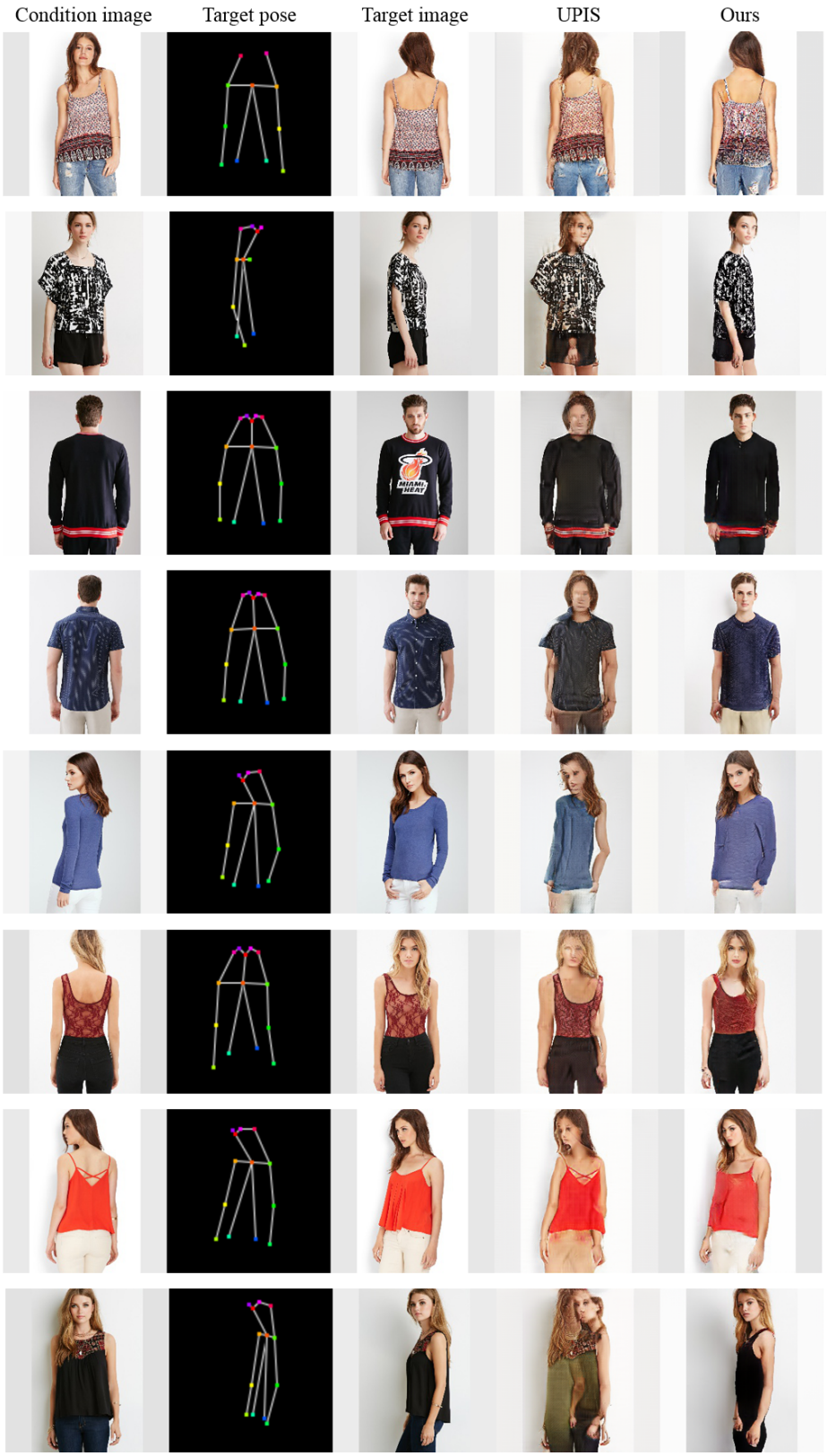

Figure 4: Comparisons with UPIS [7] on DeepFashion [5]. Our method generates more natural images with less artifacts. Zoom in for details.

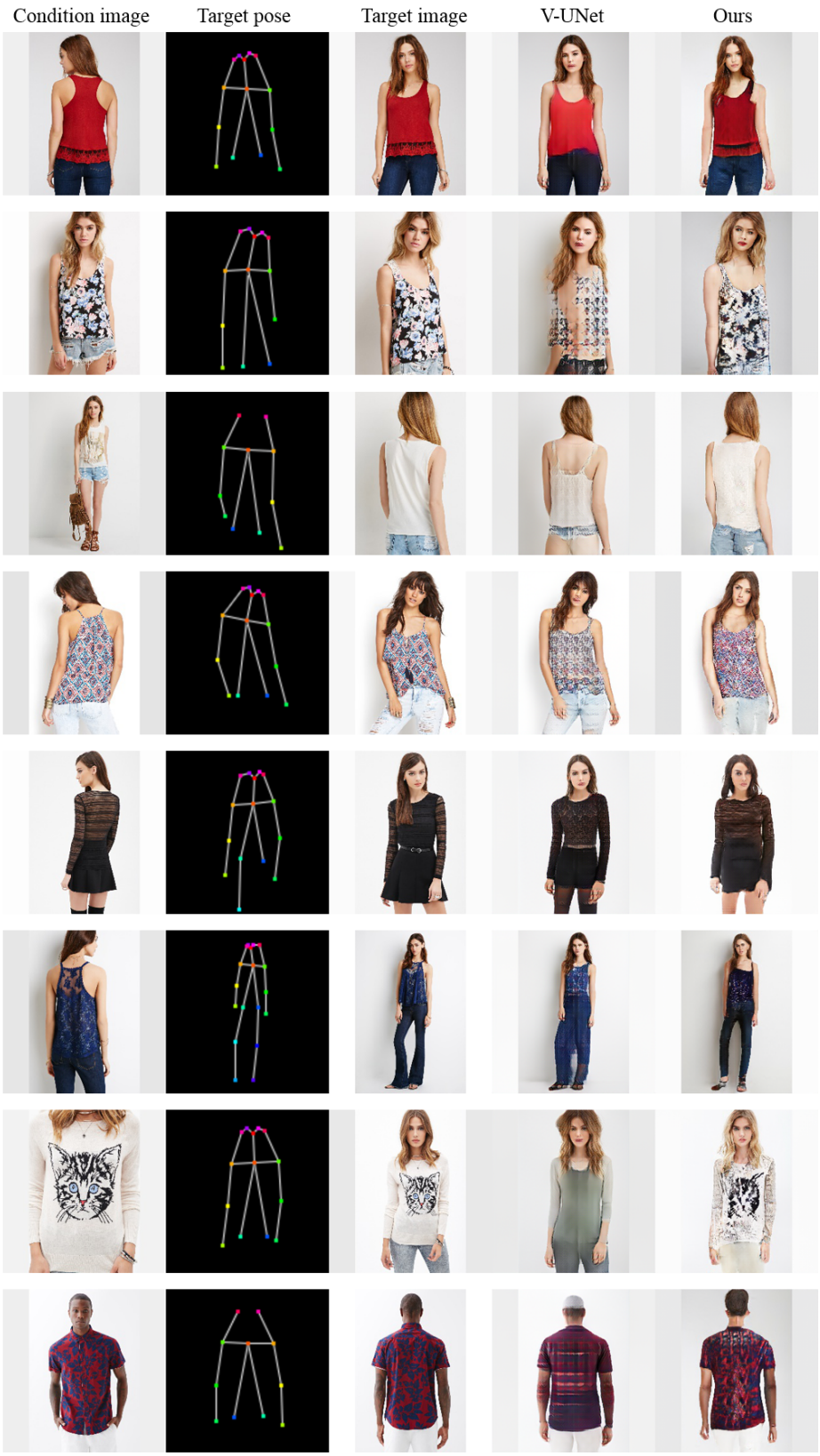

Figure 5: Comparisons with V-UNet [2] on DeepFashion [5]. Our method keeps the clothing textures better. Zoom in for details.

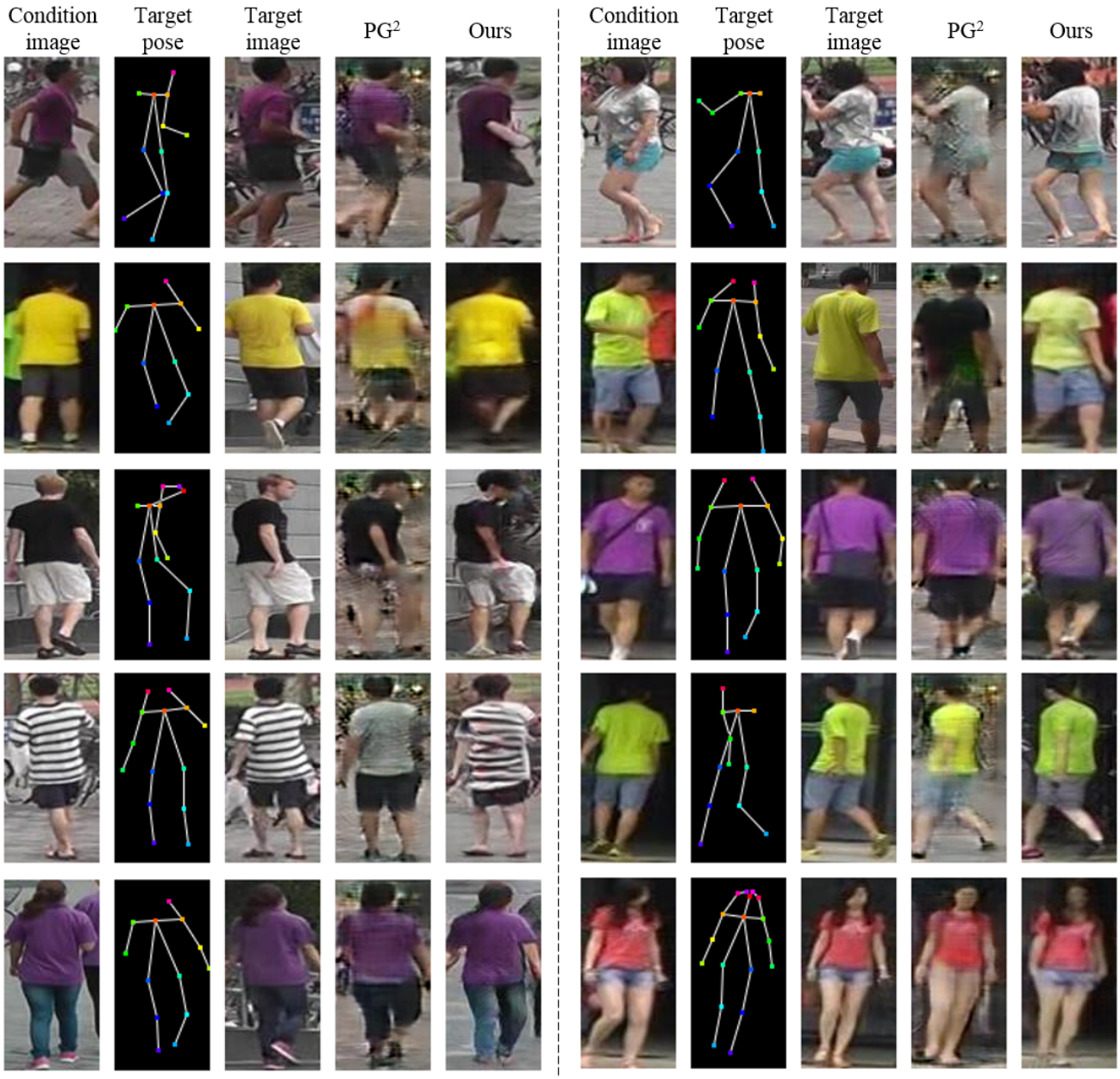

Figure 6: Comparisons with PG^2 [6] on Market-1501 [9]. Our method generates clearer body shapes with less artifacts. Zoom in for details.

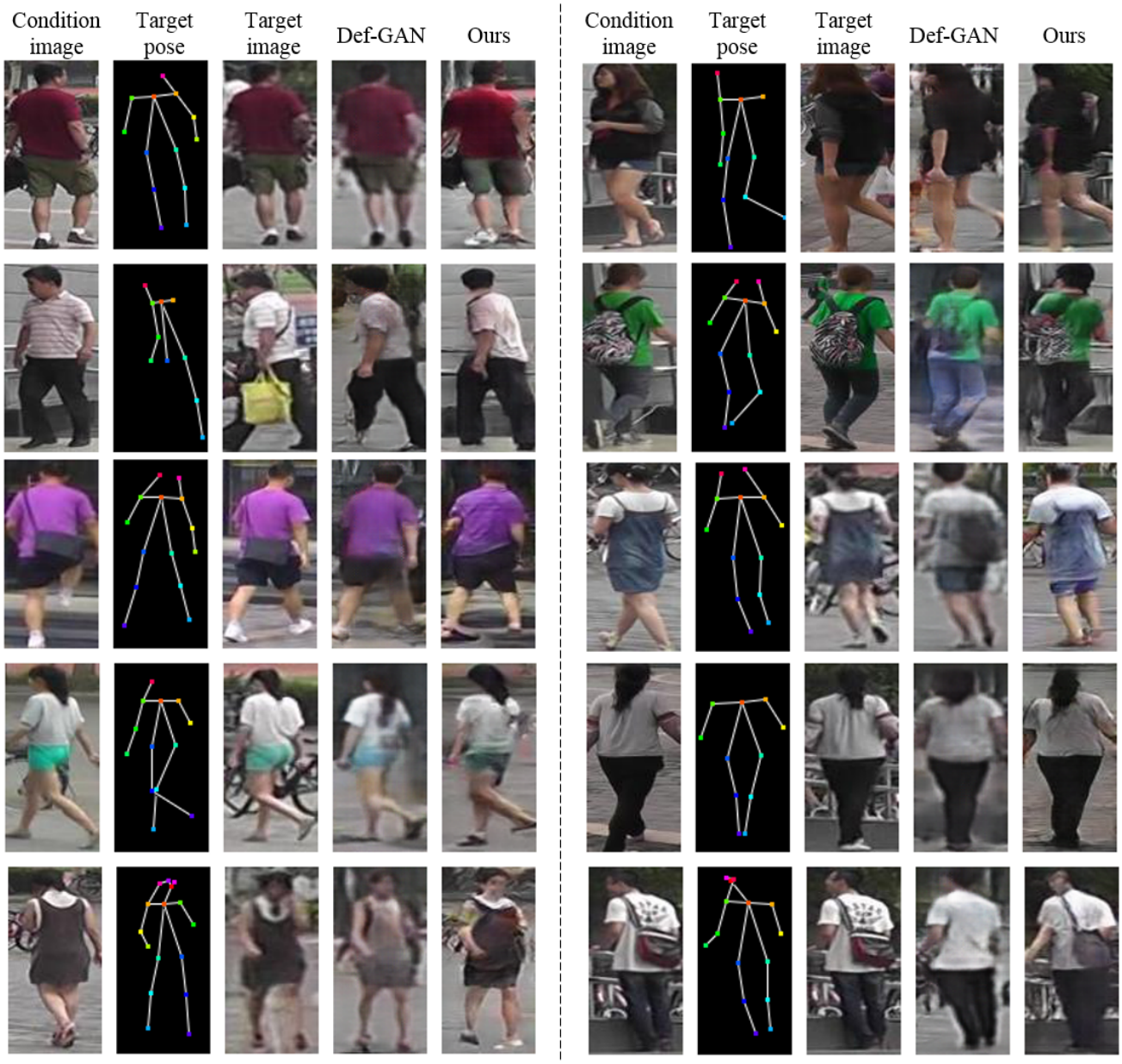

Figure 7: Comparisons with Def-GAN [8] on Market-1501 [9]. Our method generates clearer body shapes, and keeps the clothing attributes better. Zoom in for details.

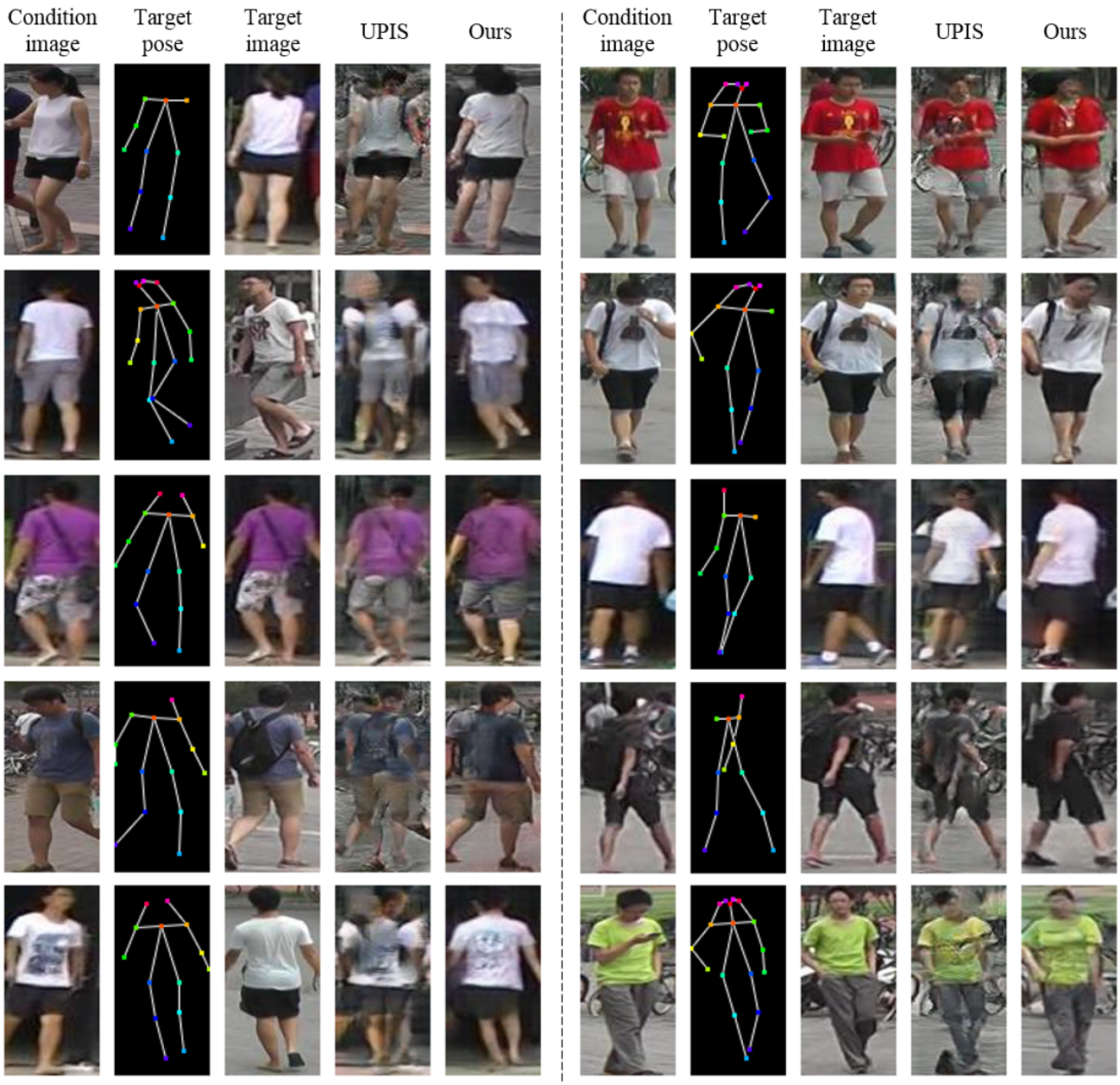

Figure 8: Comparisons with UPIS [7] on Market-1501 [9]. Our method generates more natural images with less artifacts. Zoom in for details.

Figure 9: Comparisons with V-UNet [2] on Market-1501 [9]. Our method is especially superior in keeping clothing attributes, such as colors (the 2nd row), pants lengths (the 3rd and last rows). Zoom in for details.

Figure 10: More bidirectional transfer results on clothing texture transfer. Left: condition and target images. Middle: transfer clothing textures from A to B. Right: transfer clothing textures from B to A. We compare our methods with image analogy [3] and neural doodle [1]. Our method successfully transfers the clothing textures and generates photo-realistic faces. (Empty spaces are cut off for visualization.)

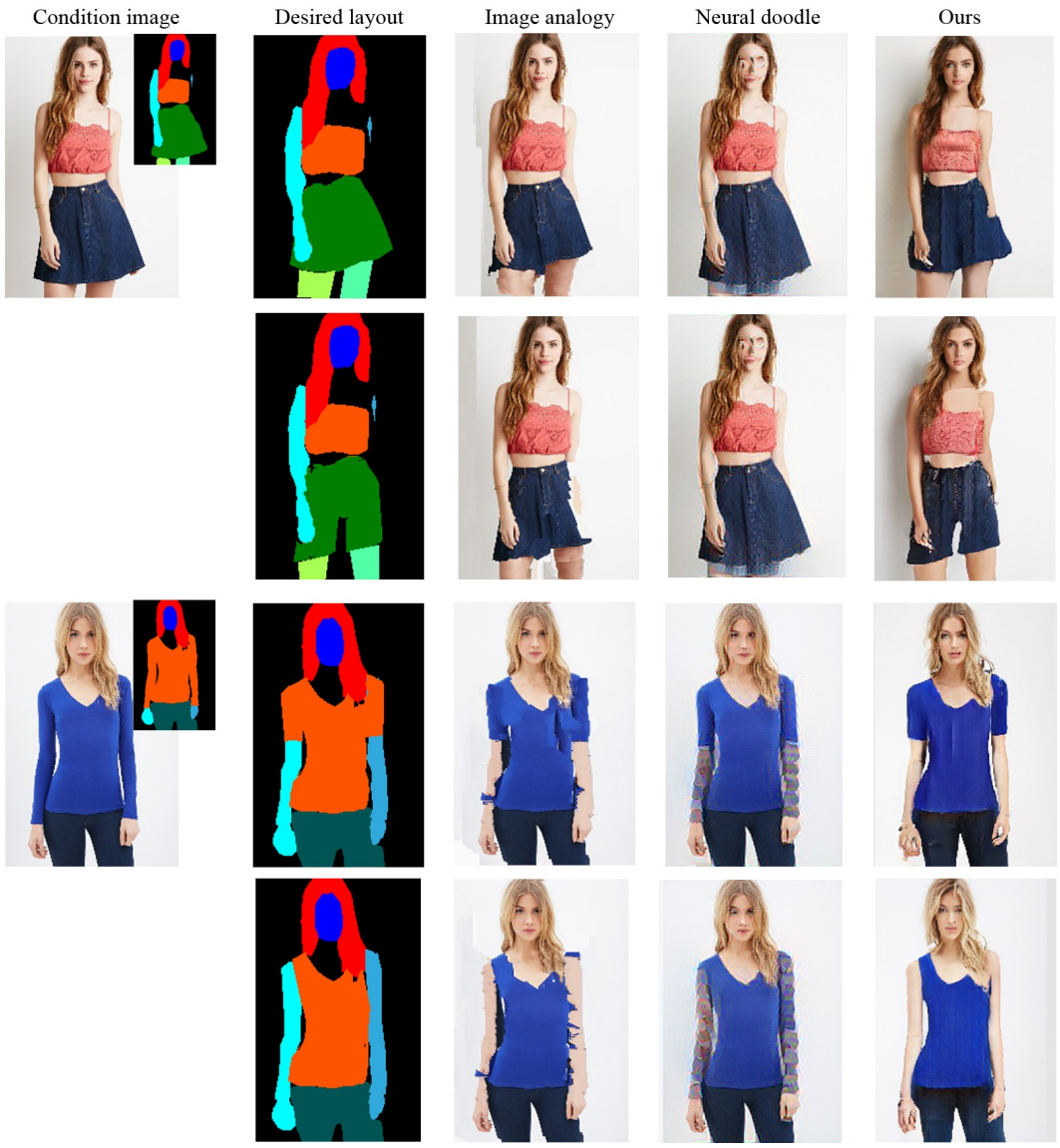

Figure 11: More results on controlled image generation. By modifying the semantic maps, we can control the image generation in desired layout. We compare our methods with image analogy [3] and neural doodle [1]. Our method shows more visual pleasant results. (Empty spaces are cut off for visualization.)