WaveIPT: Joint Attention and Flow Alignment in the Wavelet domain for Pose Transfer
Supplemental Material

Liyuan Ma[1,2*] Tingwei Gao[1*] Haitian Jiang[2] Haibin Shen[2] Kejie Huang[2†]
[1]Alibaba Group
[2]Zhejiang University, China
{mlyarthur, jianghaitian, shen_hb, huangkejie}@zju.edu.cn, tingwei.gtw@alibaba-inc.com

Overview

The overall supplementary material is organized as follows. First, we showcase more visual ablation results in Sec.1. A series of analyses on the parameter configurations and input variations of the ILC and IFI are presented, which includes Q and S input combinations (see Sec. 2.1), optimal size of the local correlation area (see Sec. 2.2) in the ILC, and the dilation rate in the IFI (see Sec. 2.3). Moreover, we introduce the training sample selection strategy for the prog flow loss (see Sec. 3) and further present additional qualitative comparison results under different poses in Sec. 4.

1. Additional Ablation Visual Results

Quantitative experimental results serve as an indicative measure of the model's performance, whereas qualitative results are more reflective of the model's efficacy and capability in generative tasks. Thus we present more results of ablation study from perspective of visualization. As shown in Figure 1, equipped with our newly designed module ILC and IFI, our method is able to recover low-frequency human body semantics and high-frequency edges by correlating texture information in different frequency bands adaptively. Furthermore, as shown in Figure 2, through comparative analysis of variant models that adopt different combinations of Q and S inputs, our method with the chosen input combination can better preserve the consistency between high-frequency details and low-frequency semantics. The comparison results in Figure 3 demonstrate the effectiveness of the progressive loss in enhancing texture deformation.

2. More Analysis for ILC and IFI

The proposed WaveIPT, comprising ILC and IFI, demonstrates efficient feature fusion through its explicit local information aggregation strategy and outperforms the mask-based fusion approach [2] in terms of computational efficiency,

2.1. Analysis for Q and S inputs in ILC

The supplement input S is utilized to compensate for the query input Q in ILC module. In our implementation, we use flow to supplement attention in the low frequency and attention to supplement flow in the high frequency. As shown in Table 1, to validate the rationality and effectiveness of this input configuration, we investigate different combinations of Q_{LF}/S_{LF} in low-frequency and Q_{HF}/S_{HF} in high-frequency.

2.2. Analysis for the local region radius r in ILC

ILC correlate each position in the query with a local region of size r. We analyze the impact of the local region size on model performance with multiple variant models in Table 2.

[*] Equal Contribution. [†] Corresponding author. This work was done when Liyuan Ma was an intern at Alibaba Group.

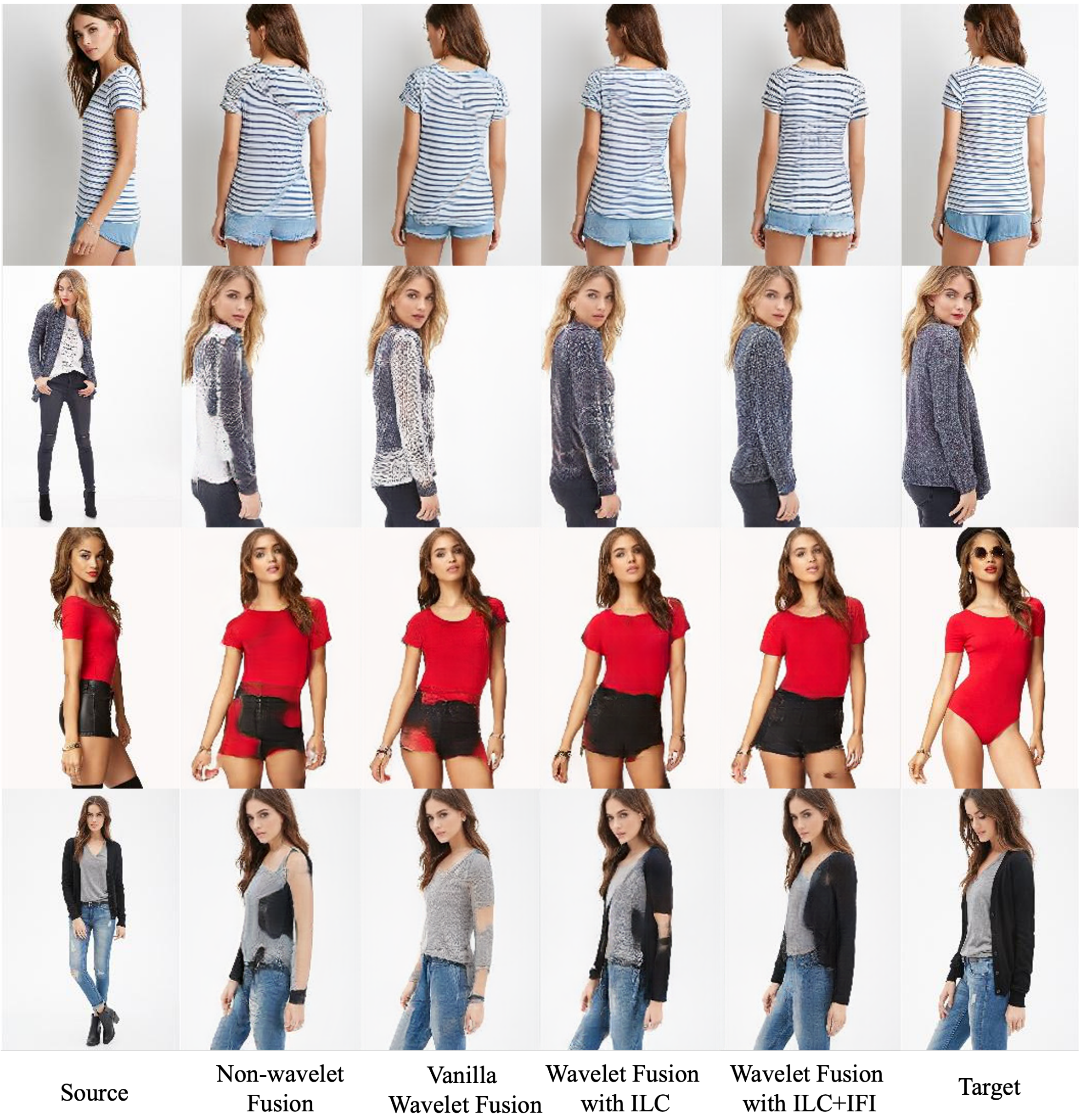


Figure 1. Ablation results of ILC and IFI modules. We can see that the visual quality is improved from left to right by adding ILC and IFI modules.

2.3. Analysis for dilated rates in IFI

To evaluate the efficacy of our dilated rates setting, we perform ablation experiments where we manipulate the dilated rates of the low frequency (d_{LF}) and high frequency (d_{HF}) components, and compare the results against our configuration where $d_{LF} = 3$ and $d_{HF} = 1$. Notably, we also conducte an ablation experiment where we swap the dilated rates of the low and high frequency components ($d_{LF} = 1$ and $d_{HF} = 3$) to verify the requirement of larger dilated rate for the low frequency component as opposed to the high frequency component. More details can be found in Sec. 3.

Considering the ability of metrics to effectively evaluate the quality of generated images, LPIPS and FID were selected to compare the performance of models with varying inflation rates in Figure 4. It can be found that our setting outperforms others in both metrics.

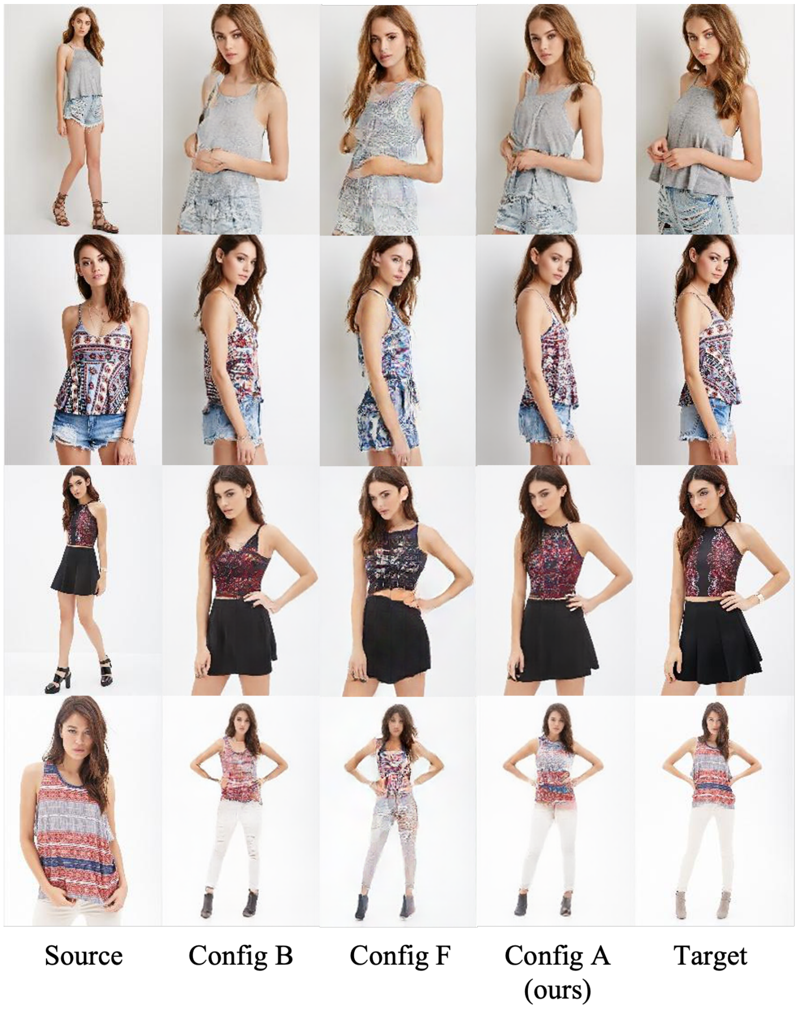


Figure 2. Ablation results for Q and S inputs in ILC with different configurations as illustrated in Table 1. Config B reverses the Q and S inputs in the high frequency, resulting in the deterioration of the sharp details (see the cloth pattern in the 4th row). Config F fails to recover the semantic information such as cloth color (2nd row) or haircut (4th row), which is caused by the different Q and S inputs compared with our Config A setting in the low frequency.

Config		LF_{attn}	HF_{attn}	LF_{flow}	HF_{flow}		LF_{attn}	HF_{attn}	LF_{flow}	HF_{flow}	SSIM↑	FID↓	LPIPS ↓		Reid Score(%)↑		
													AlexNet	VGG	Topk-1	Topk-5	Topk-10
A (ours)	Q_{LF}	✓				S_{LF}			✓		**0.7801**	**8.8259**	**0.1955**	**0.2348**	**99.05**	**99.96**	**99.99**
	Q_{HF}				✓	S_{HF}		✓									
B	Q_{LF}	✓				S_{LF}			✓		0.7800	9.7890	0.1989	0.2356	98.24	99.79	99.89
	Q_{HF}		✓			S_{HF}				✓							
C	Q_{LF}	✓				S_{LF}		✓			0.7679	11.9708	0.2329	0.2655	81.42	92.74	96.15
	Q_{HF}				✓	S_{HF}			✓								
D	Q_{LF}	✓				S_{LF}				✓	0.7743	11.0636	0.2136	0.2496	94.22	98.63	99.28
	Q_{HF}		✓			S_{HF}			✓								
E	Q_{LF}			✓		S_{LF}	✓				0.7775	10.4844	0.2052	0.2417	96.23	99.31	99.66
	Q_{HF}		✓			S_{HF}				✓							
F	Q_{LF}			✓		S_{LF}	✓				0.7637	12.9238	0.2431	0.2738	69.78	86.24	92.02
	Q_{HF}				✓	S_{HF}		✓									
G	Q_{LF}			✓		S_{LF}				✓	0.7754	10.8537	0.2100	0.2477	95.20	98.97	99.49
	Q_{HF}		✓			S_{HF}	✓										
H	Q_{LF}			✓		S_{LF}		✓			0.7732	11.3360	0.2207	0.2542	89.65	96.88	98.48
	Q_{HF}				✓	S_{HF}	✓										

Table 1. Quantitative analysis of Q and S inputs in ILC.

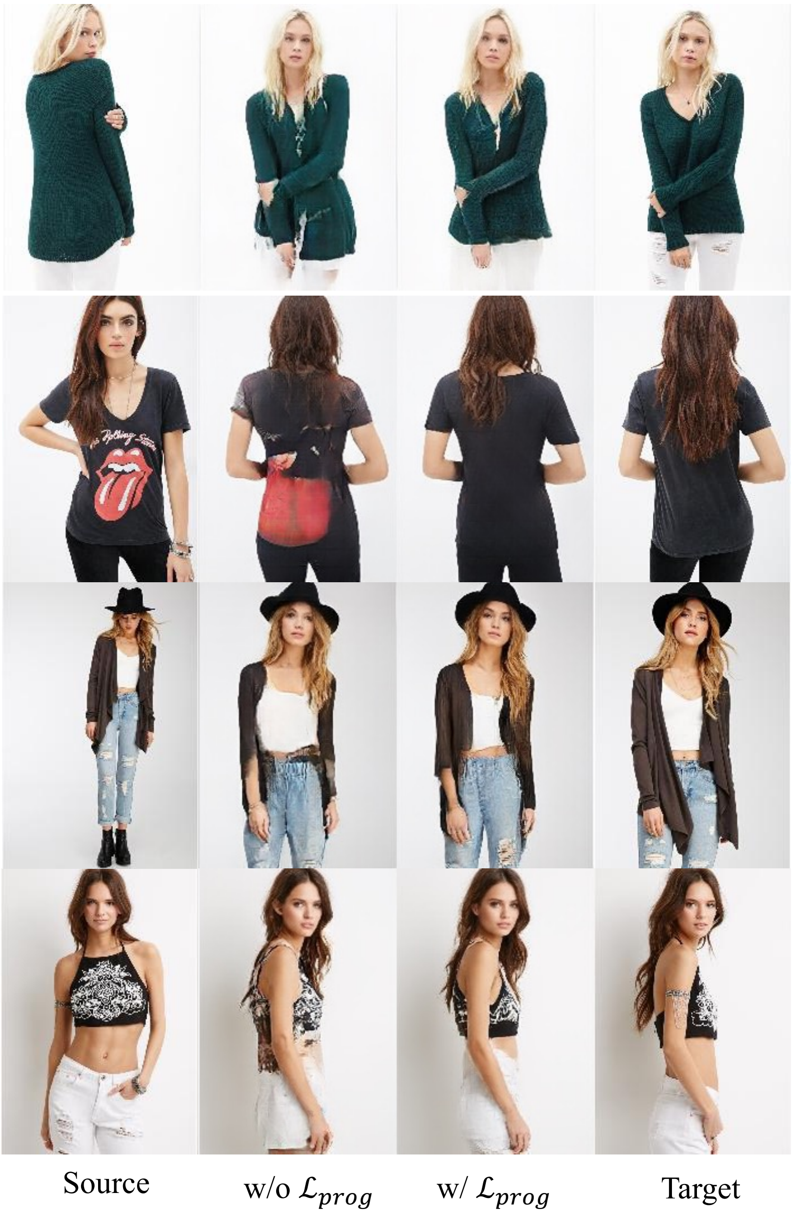
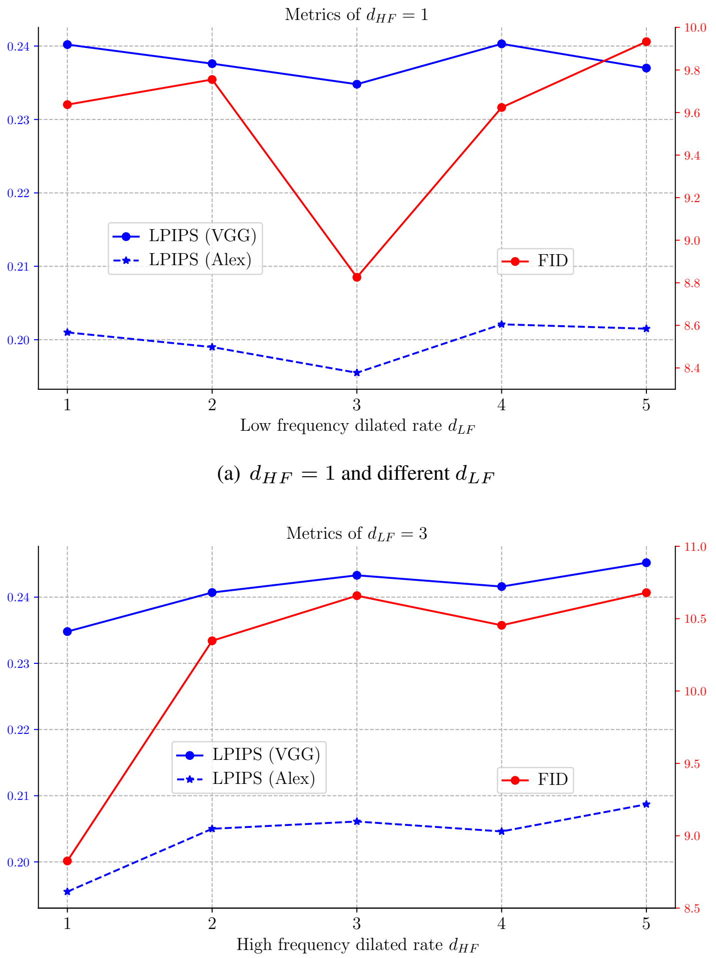


Figure 3. Ablation results of progressive loss. by incorporating the progressive flow loss, the model can effectively enhance its ability to deform texture precisely.

		SSIM↑	FID↓	LPIPS ↓		Reid Score(%)↑		
				AlexNet	VGG	Topk-1	Topk-5	Topk-10
ILC local size r	r=3 (ours)	**0.7801**	**8.8259**	**0.1955**	**0.2348**	**99.05**	**99.96**	**99.99**
	r=5	0.7800	9.9715	0.2005	0.2372	97.77	99.71	99.84
	r=7	0.7770	10.3544	0.2084	0.2425	96.46	99.43	99.68
	r=9	0.7775	10.3147	0.2094	0.2432	96.63	99.54	99.70
	r=11	0.7775	10.4246	0.2061	0.2435	96.76	99.53	99.79

Table 2. Quantitative analysis for the local region radius r in ILC. Best result is achieved in $r = 3$.



(a) $d_{HF} = 1$ and different d_{LF}

(b) $d_{LF} = 3$ and different d_{HF}

Figure 4. Analysis for dilated rates in IFI. Our model achieves lowest LPIPS and FID when $d_{LF} = 3$ and $d_{HF} = 1$, which proves that larger dilated rate in low frequency is reasonable.

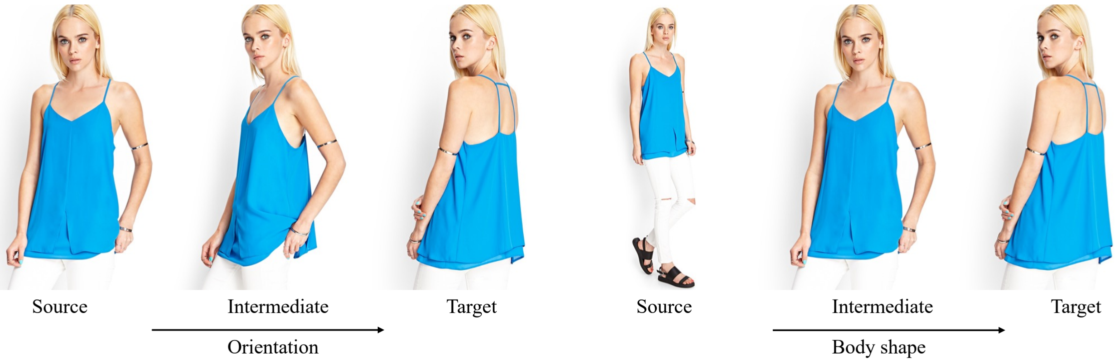


Figure 5. Illustration of intermediate pose selection.

3. Training Details about Progressive Flow Regularization

The training of Progressive Flow Regularization involves the selection of intermediate pose between source and target poses. The selection of intermediate pose must ensure that it can alleviate the differences in orientation and shape between source and target poses as shown in Figure 5. The intermediate poses serve as the transition between poses with different

			SSIM↑	FID↓	LPIPS ↓		Reid Score(%)↑		
					AlexNet	VGG	Topk-1	Topk-5	Topk-10
IFI dilated rate	Interchange ours setting		0.7740	10.1769	0.2160	0.2494	93.87	98.73	99.40
	$d_{LF} = 3$	$d_{HF} = 1$ (ours)	**0.7801**	**8.8259**	**0.1955**	**0.2348**	**99.05**	**99.96**	**99.99**
		$d_{HF} = 2$	0.7781	10.3471	0.2050	0.2407	96.84	99.43	99.74
		$d_{HF} = 3$	0.7767	10.6587	0.2061	0.2433	97.67	99.75	99.88
		$d_{HF} = 4$	0.7768	10.4541	0.2046	0.2416	98.02	99.75	99.88
		$d_{HF} = 5$	0.7759	10.6795	0.2087	0.2452	97.07	99.59	99.84
	d_{HF}=1	$d_{LF} = 1$	0.7788	9.6371	0.2011	0.2402	98.53	99.86	99.92
		$d_{LF} = 2$	0.7794	9.7551	0.1991	0.2376	98.52	99.86	99.91
		$d_{LF} = 3$ (ours)	**0.7801**	**8.8259**	**0.1955**	**0.2348**	**99.05**	**99.96**	**99.99**
		$d_{LF} = 4$	0.7785	9.6238	0.2021	0.2403	97.84	99.74	99.85
		$d_{LF} = 5$	0.7793	9.9330	0.2016	0.2370	97.88	99.71	99.87

Table 3. Quantitative analysis for dilated rates in IFI. We interchange ours setting to validate the rationality of large dilated rate in low frequency.

orientations (see the left part in Figure 5) and shapes from full to half body (see the right part in Figure 5).

4. Additional Results

Additional results of human pose transfer on the DeepFashion [1] dataet are given in Figure 6, 7, 8, 9, 10, 11, 12, 13, 14.

References

[1] Ziwei Liu, Ping Luo, Shi Qiu, Xiaogang Wang, and Xiaoou Tang. Deepfashion: Powering robust clothes recognition and retrieval with rich annotations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1096–1104, 2016. 6

[2] Liyuan Ma, Tingwei Gao, Haibin Shen, and Kejie Huang. Freqhpt: Frequency-aware attention and flow fusion for human pose transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3490–3495, 2023. 1

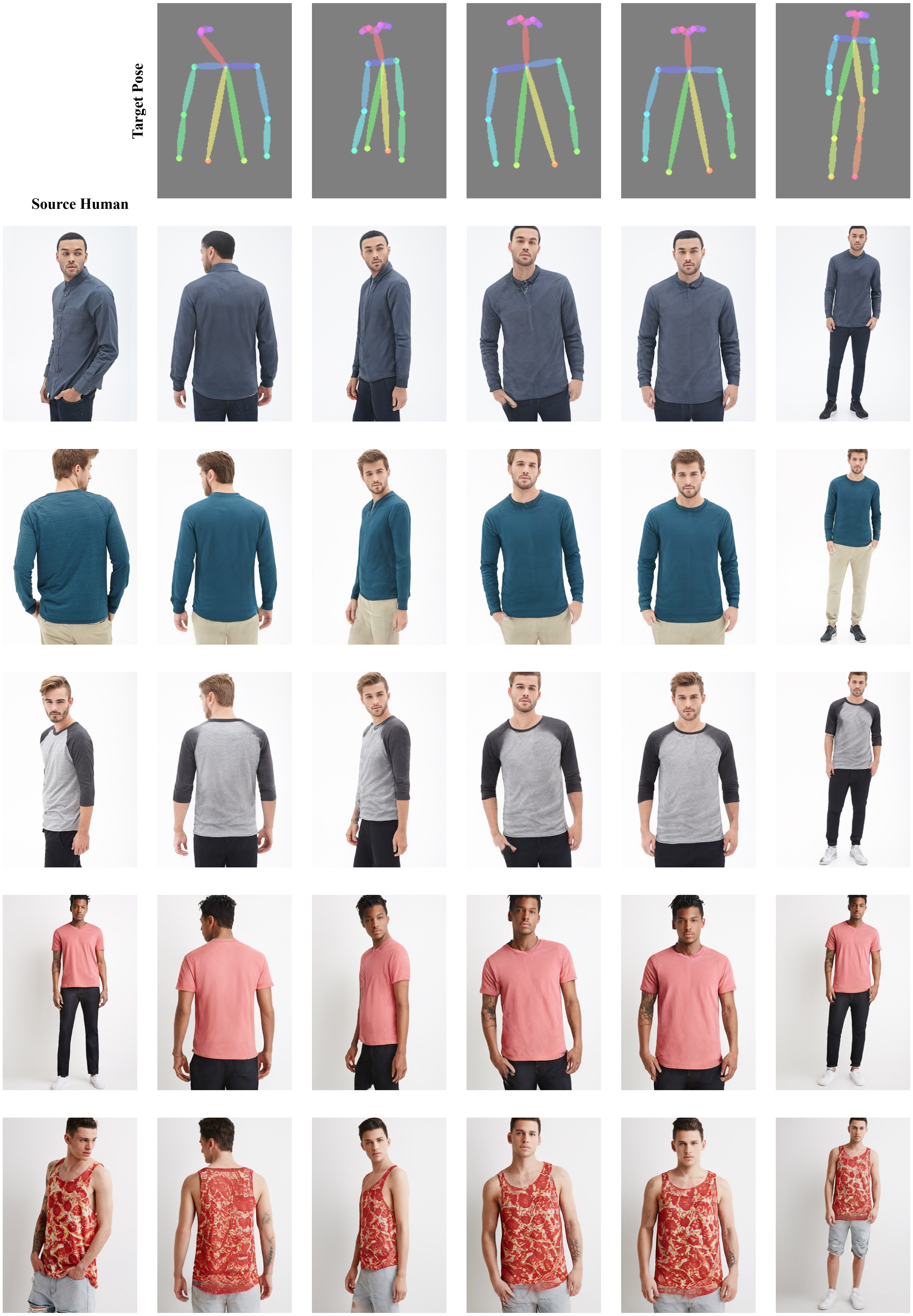


Figure 6. Pose transfer results under different poses.

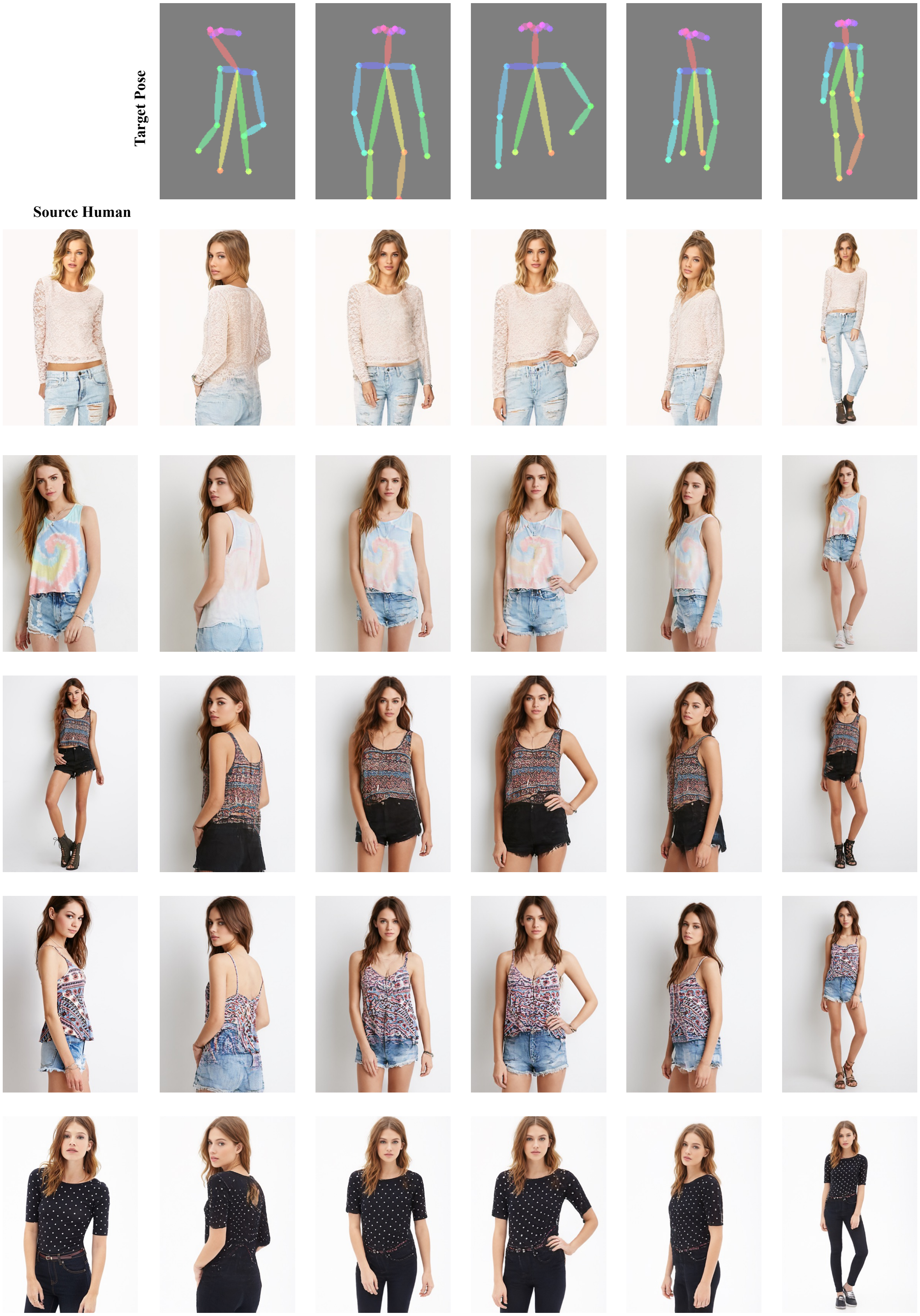


Figure 7. Pose transfer results under different poses.

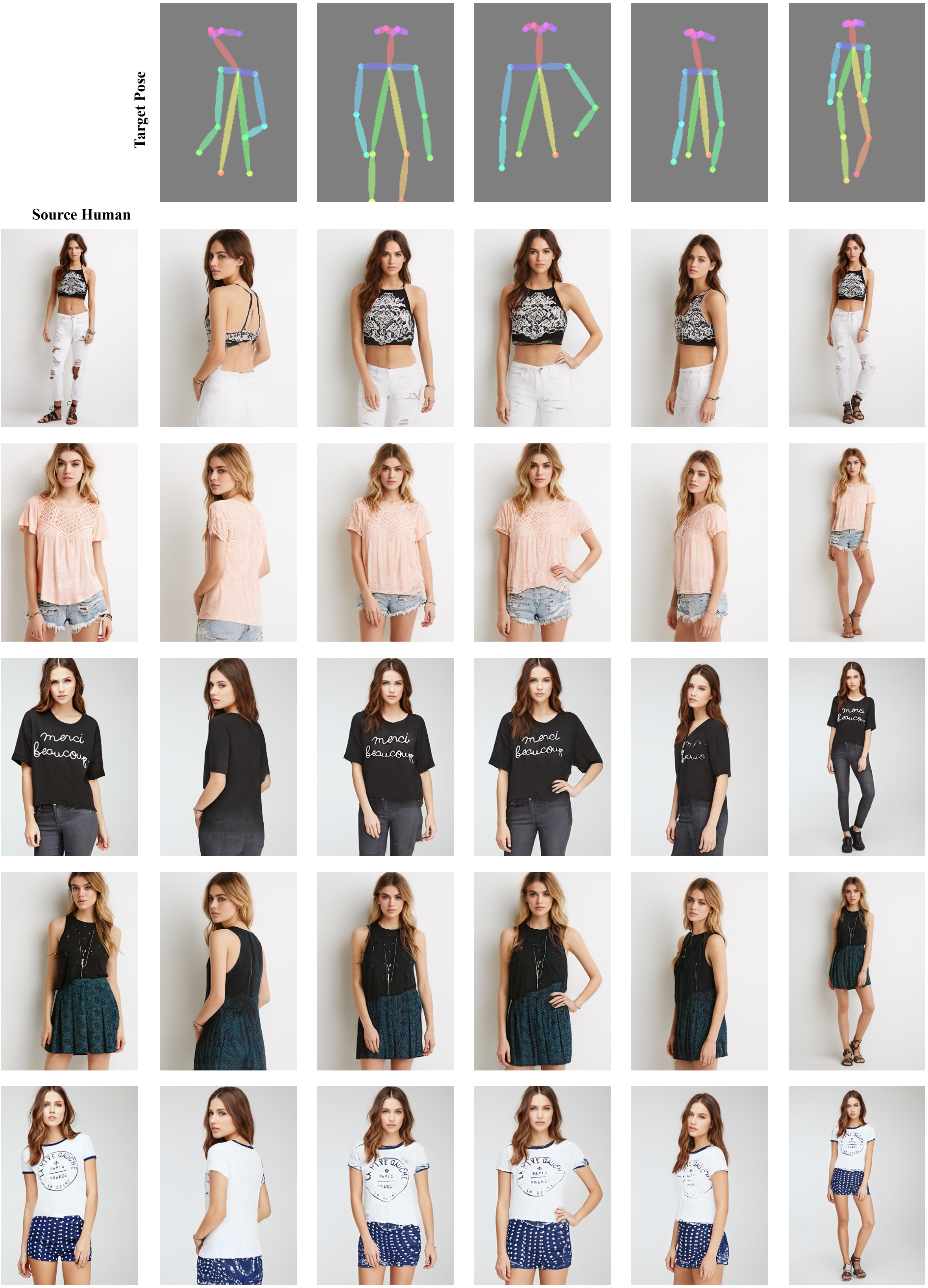


Figure 8. Pose transfer results under different poses.

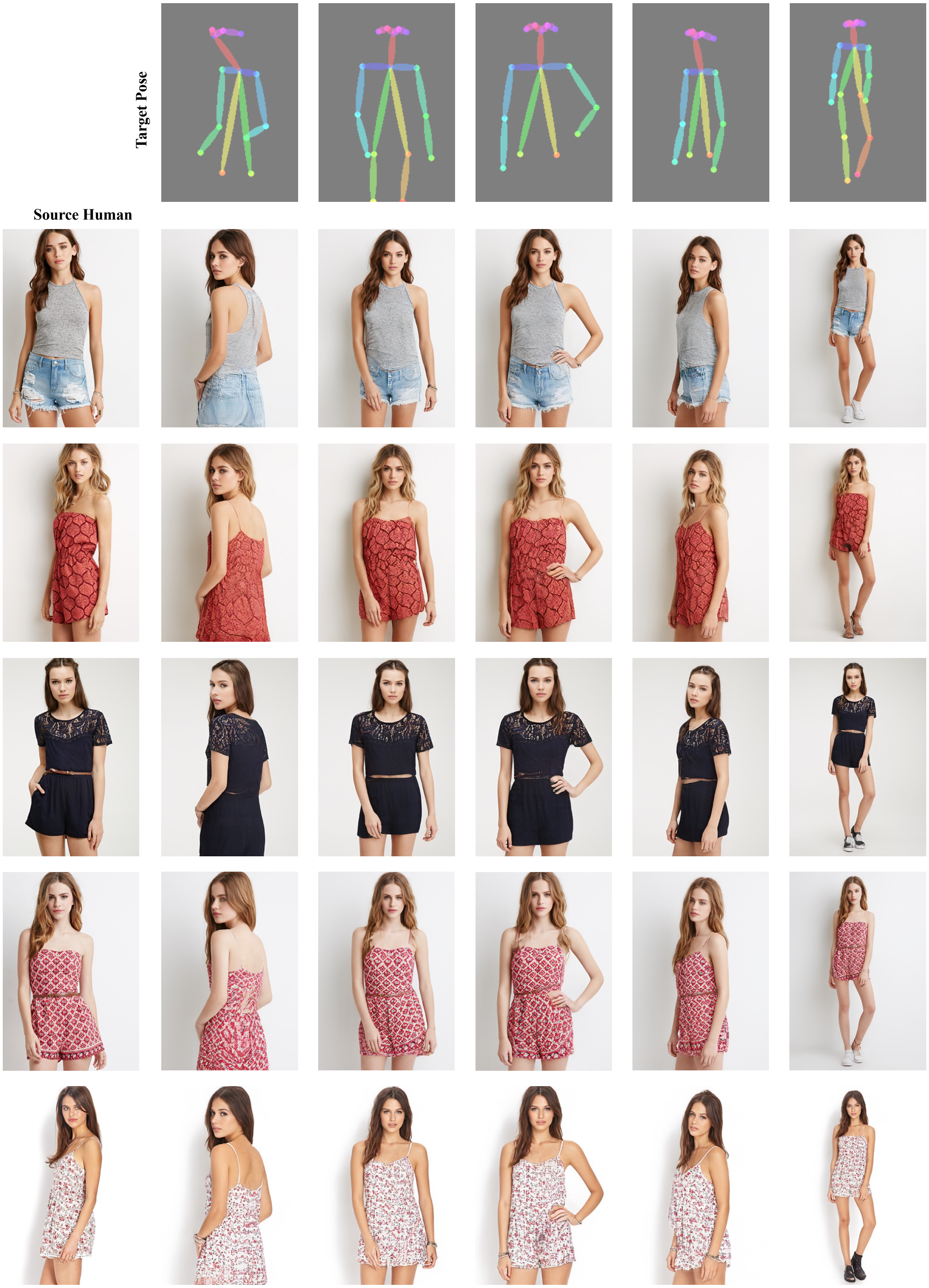


Figure 9. Pose transfer results under different poses.

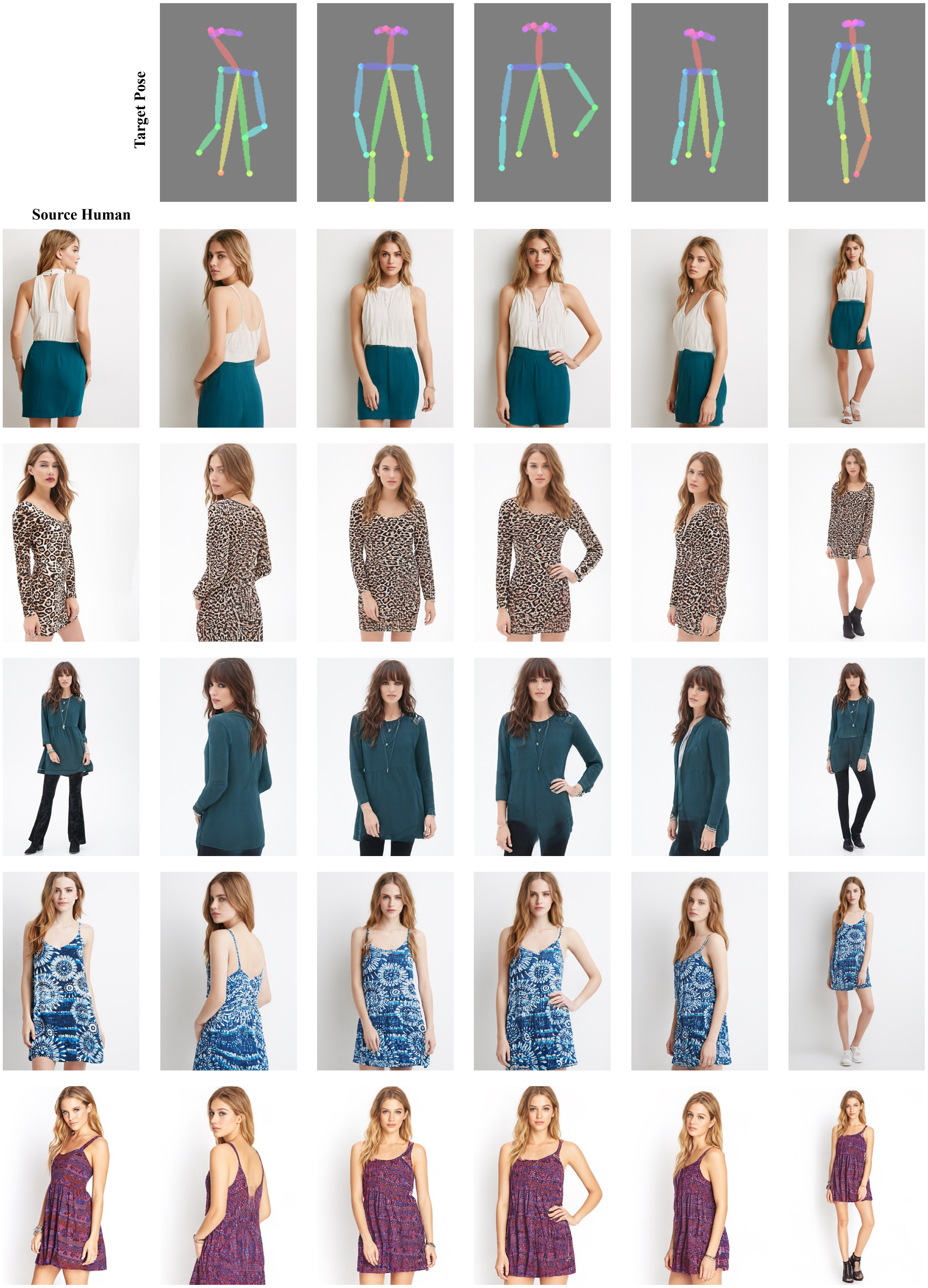


Figure 10. Pose transfer results under different poses.

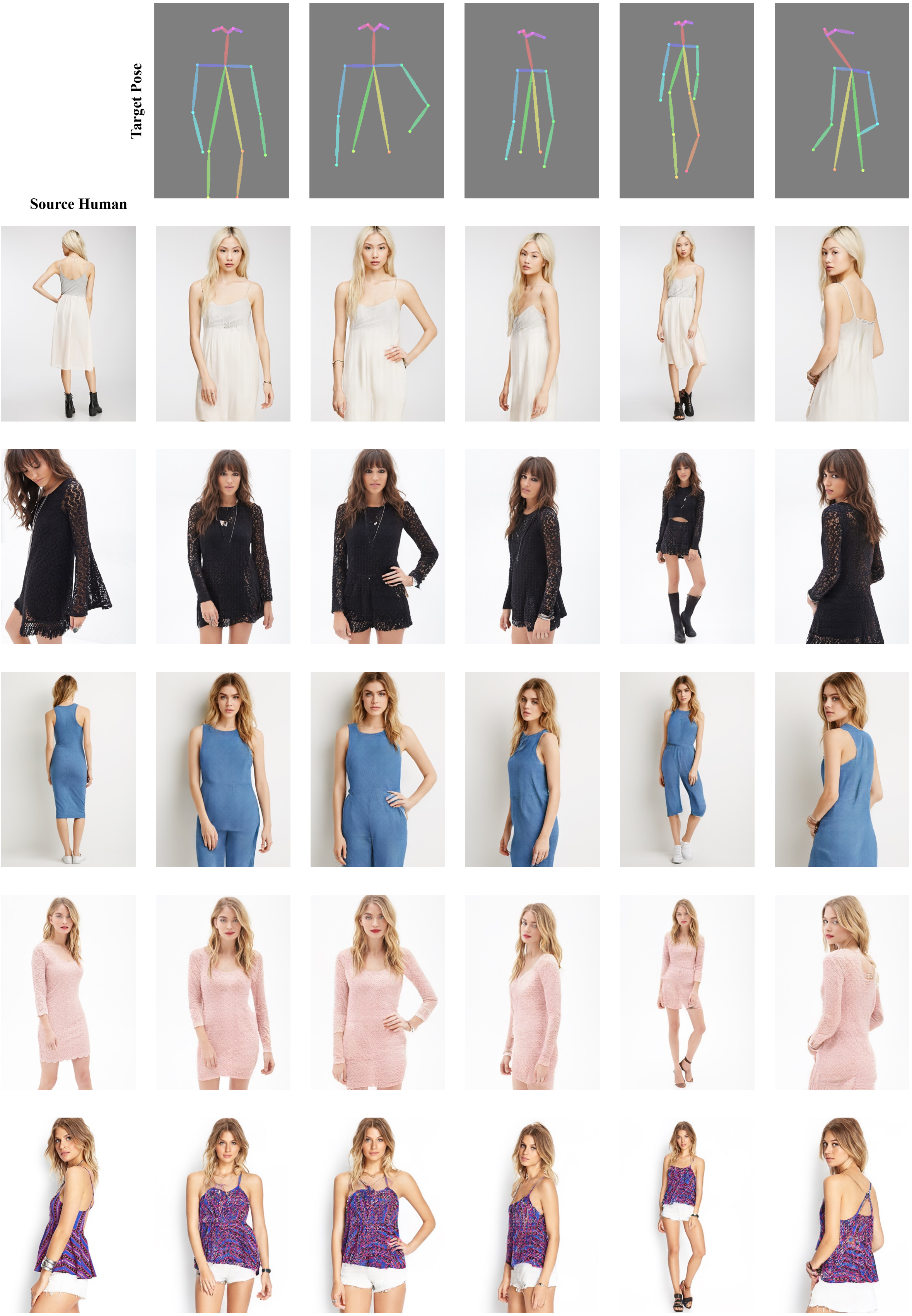


Figure 11. Pose transfer results under different poses.

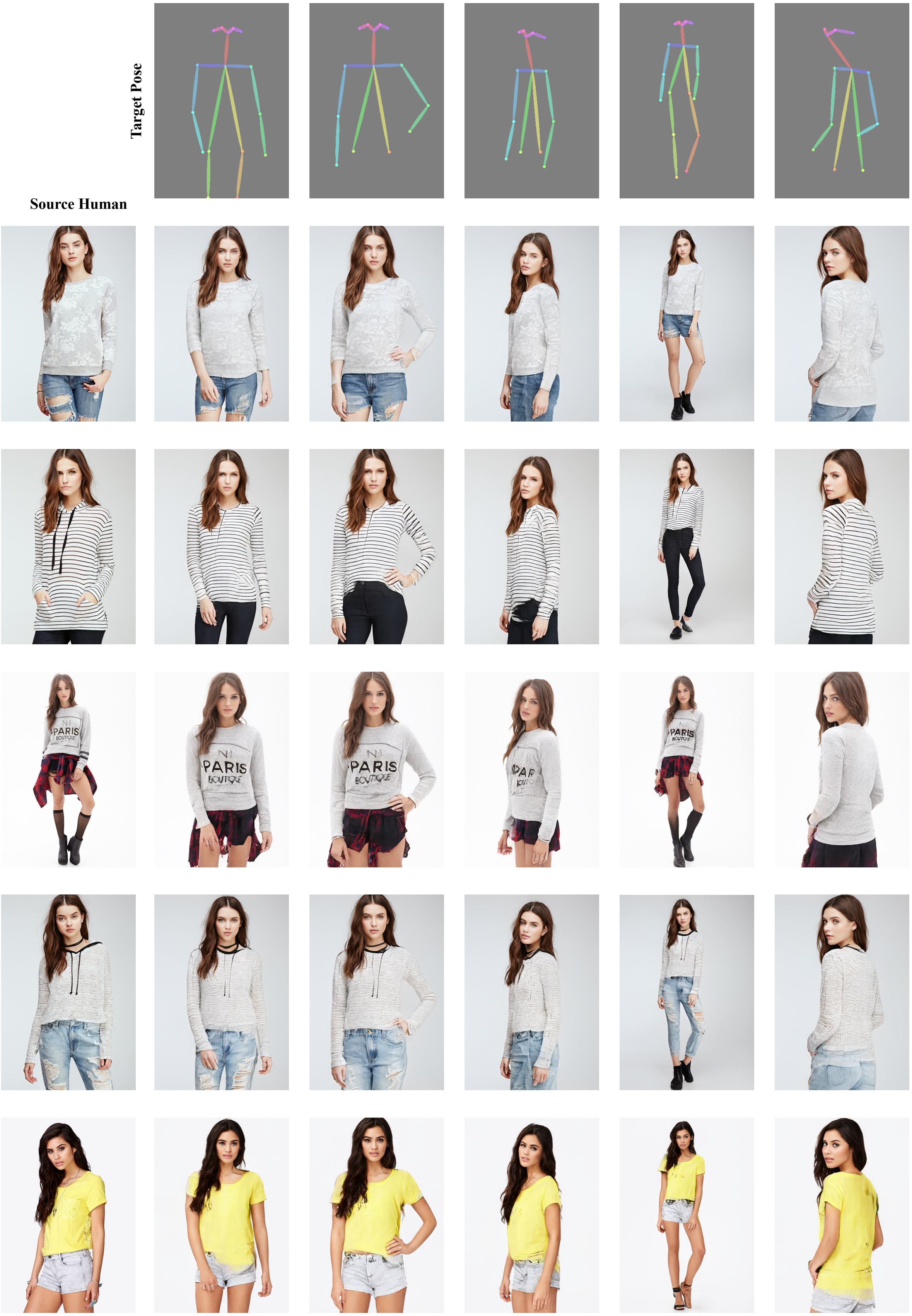


Figure 12. Pose transfer results under different poses.

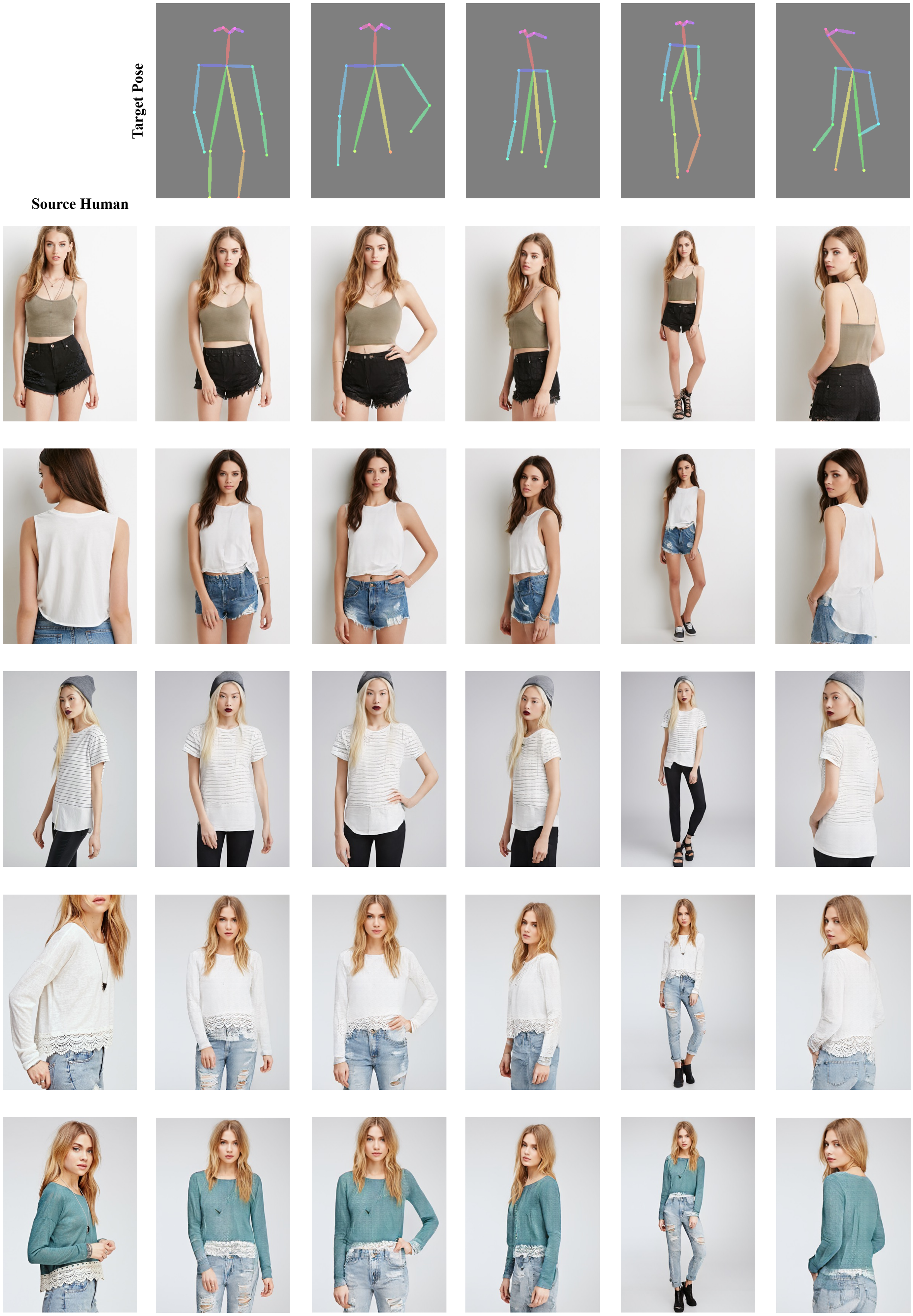


Figure 13. Pose transfer results under different poses.

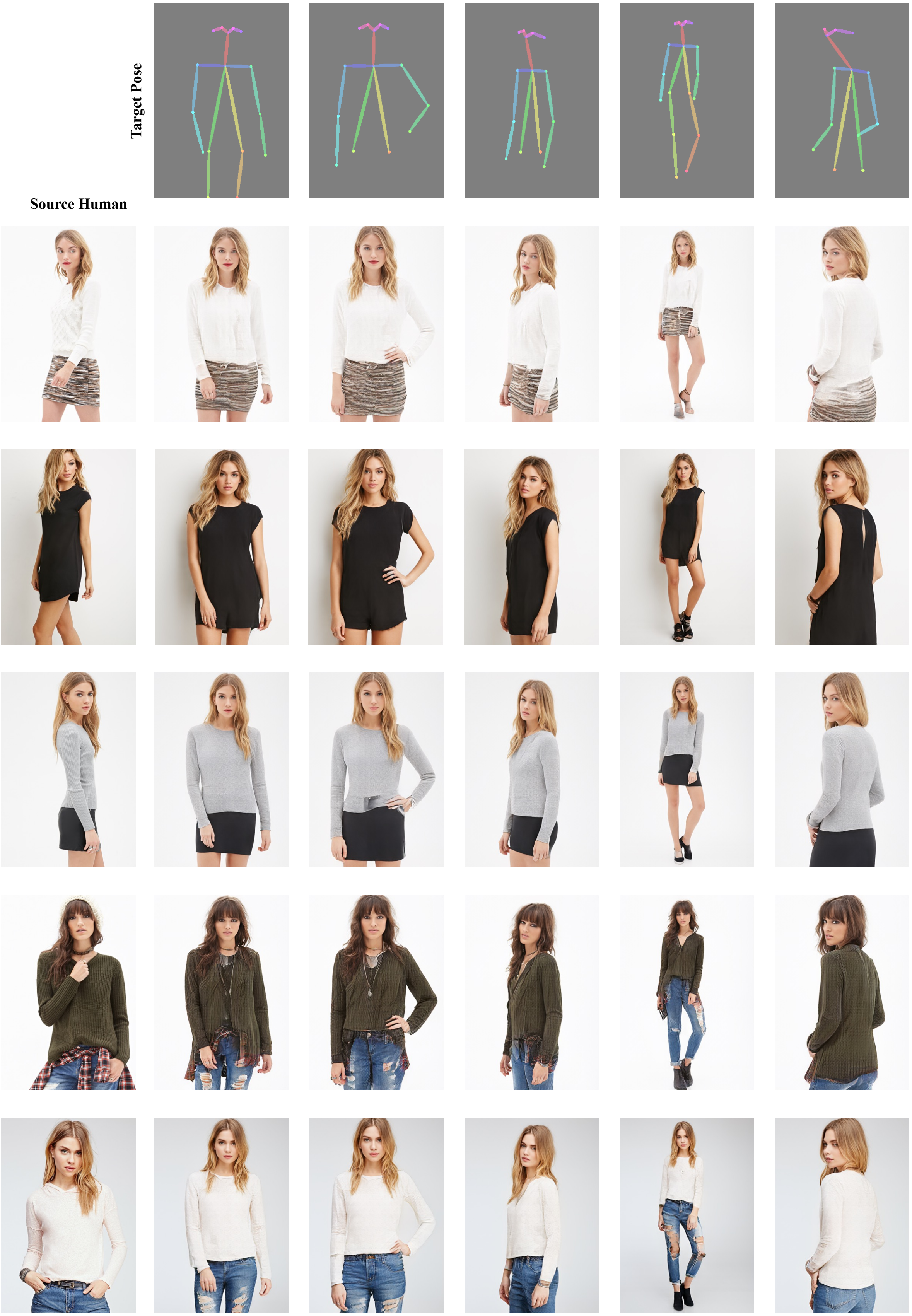


Figure 14. Pose transfer results under different poses.