Supplementary Materials for

"Controllable Person Image Synthesis with Attribute-Decomposed GAN"

[1]Yifang Men, [2]Yiming Mao, [2]Yuning Jiang, [2]Wei-Ying Ma, [1]Zhouhui Lian

[1]Wangxuan Institute of Computer Technology, Peking University, China
[2]Bytedance AI Lab

In this document we provide the following supplementary contents:

- Video demonstration of controllable person image synthesis.
- Applications of synthesizing person images in arbitrary poses and component attributes.
- Comparison with state-of-the-art methods.
- Style interpolation results.
- Experimental results with human parsers having different accuracies.

1. Video demonstration of controllable person image synthesis.

Our model constructs a complex manifold that is constituted of various human attributes of the person images, including pose, upper clothes, pants, head and so on. We can travel along this manifold of all human attributes to synthesize an animation from one attribute to another, thus visualizing the encoded low dimensional space. For component attributes, the person image with desired attributes can be synthesized by editing the style code in the latent space. For the pose attribute, with a series of target poses provided, our model can synthesize a motion for the source person. It is strongly recommended to see the supplemental video (https://youtu.be/hstN3lOWVHg) for the visualization, which proves the effectiveness of controllable human attributes and the continuity of our constructed manifold.

2. Applications

2.1. Person image synthesis in arbitrary poses

For arbitrary poses extracted from person images in the test set, our method can synthesize target person images with the source person in the target poses. And once the model is trained, any person in arbitrary poses can be generated.

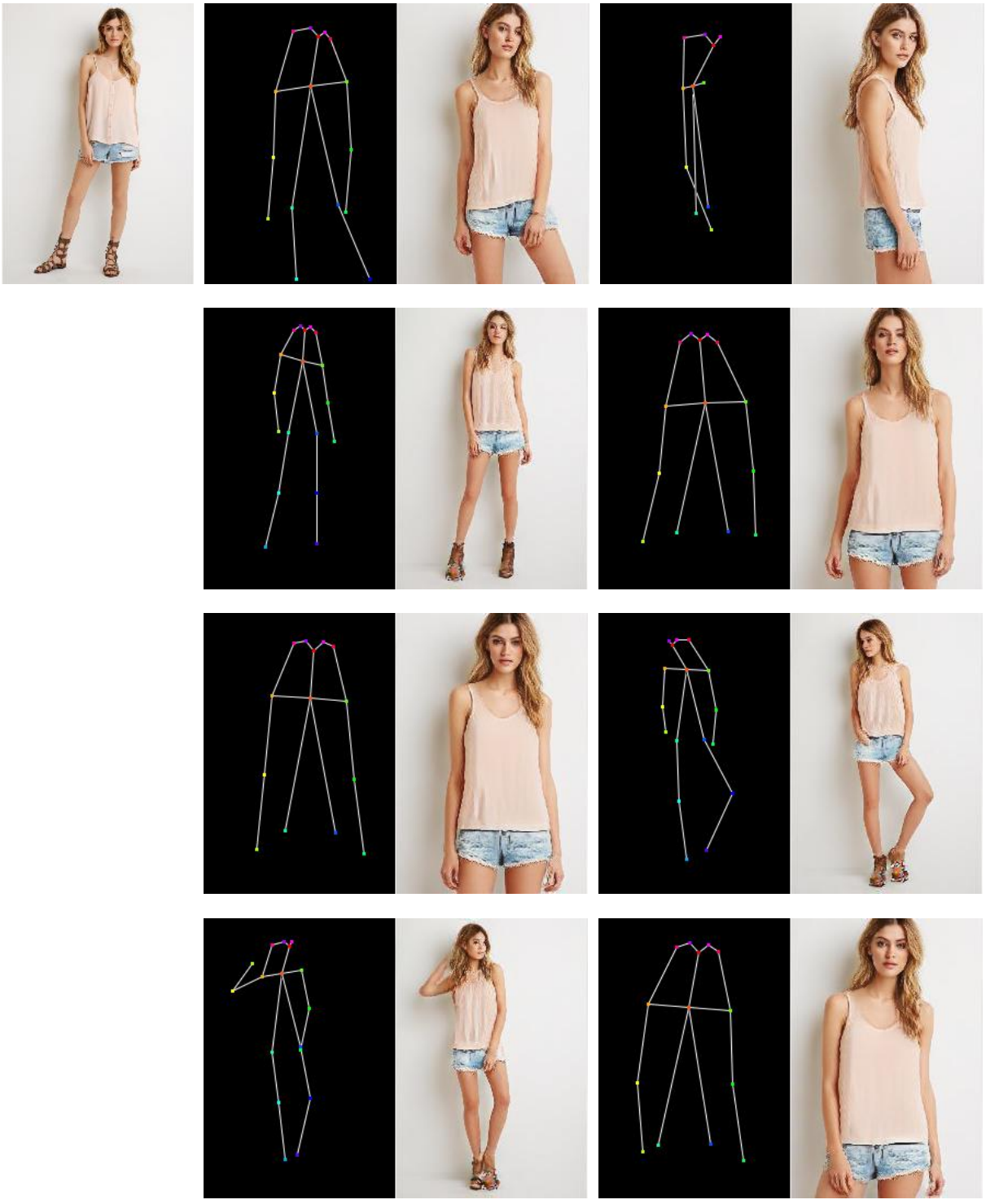

Figure 1: Results of synthesizing person images in arbitrary poses.

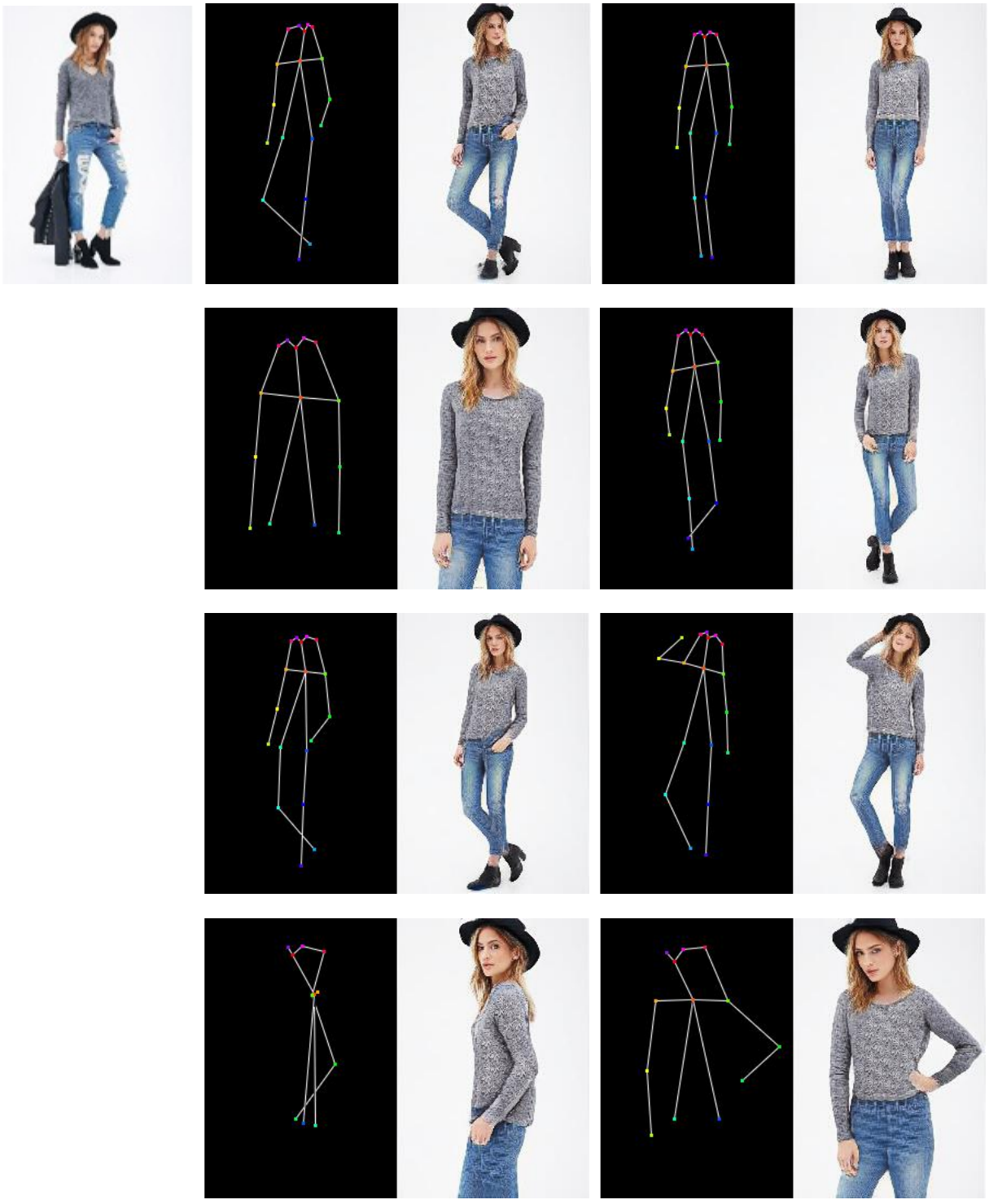

Figure 2: Results of synthesizing person images in arbitrary poses.

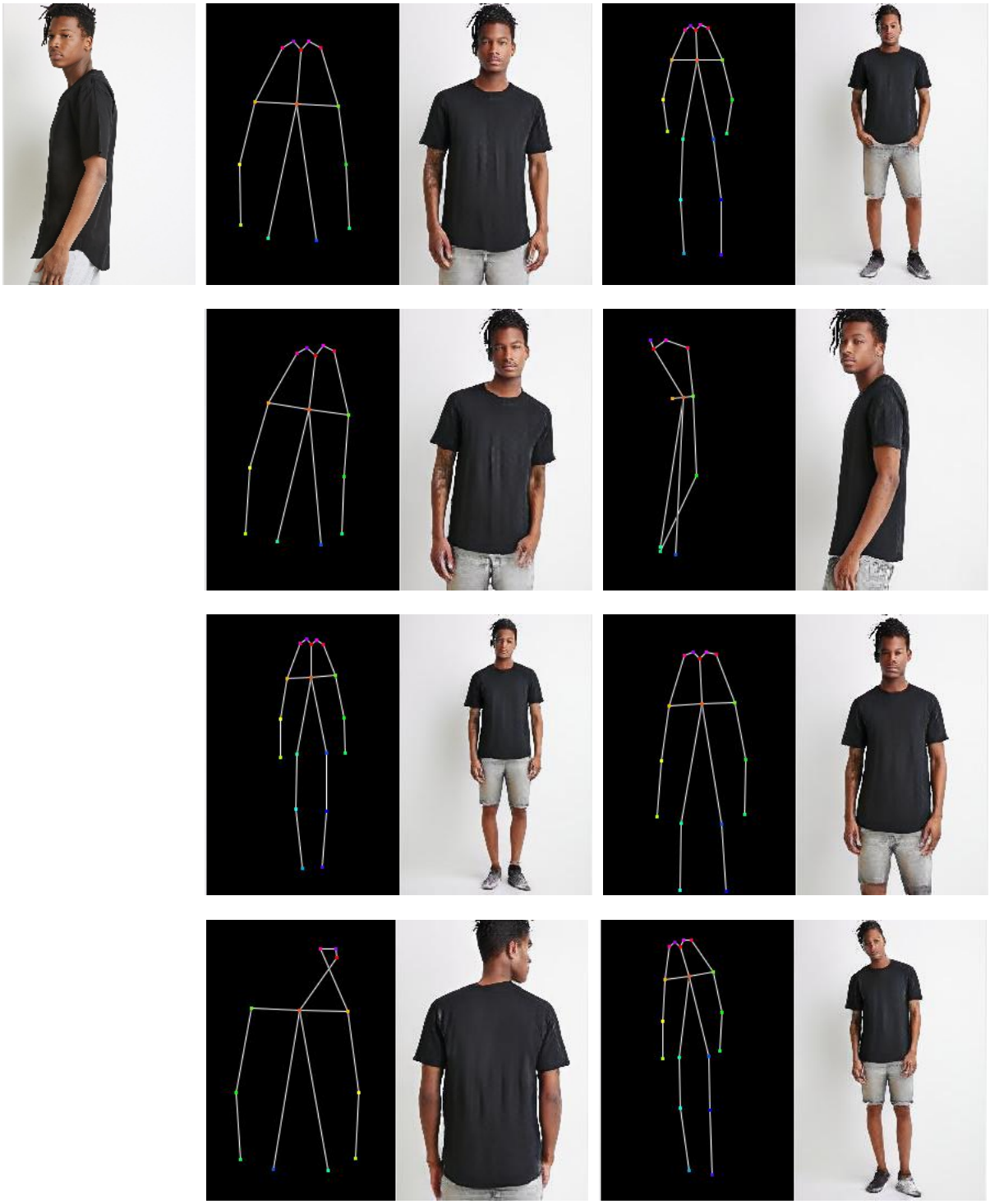

Figure 3: Results of synthesizing person images in arbitrary poses.

Figure 4: Results of synthesizing person images in arbitrary poses.

2.2. Person image synthesis in controllable component attributes

For the original person image, our method can change its component attributes (e.g., upper clothes, pants and head) with another person image containing the desired attribute is provided.

Controllable upper clothes.

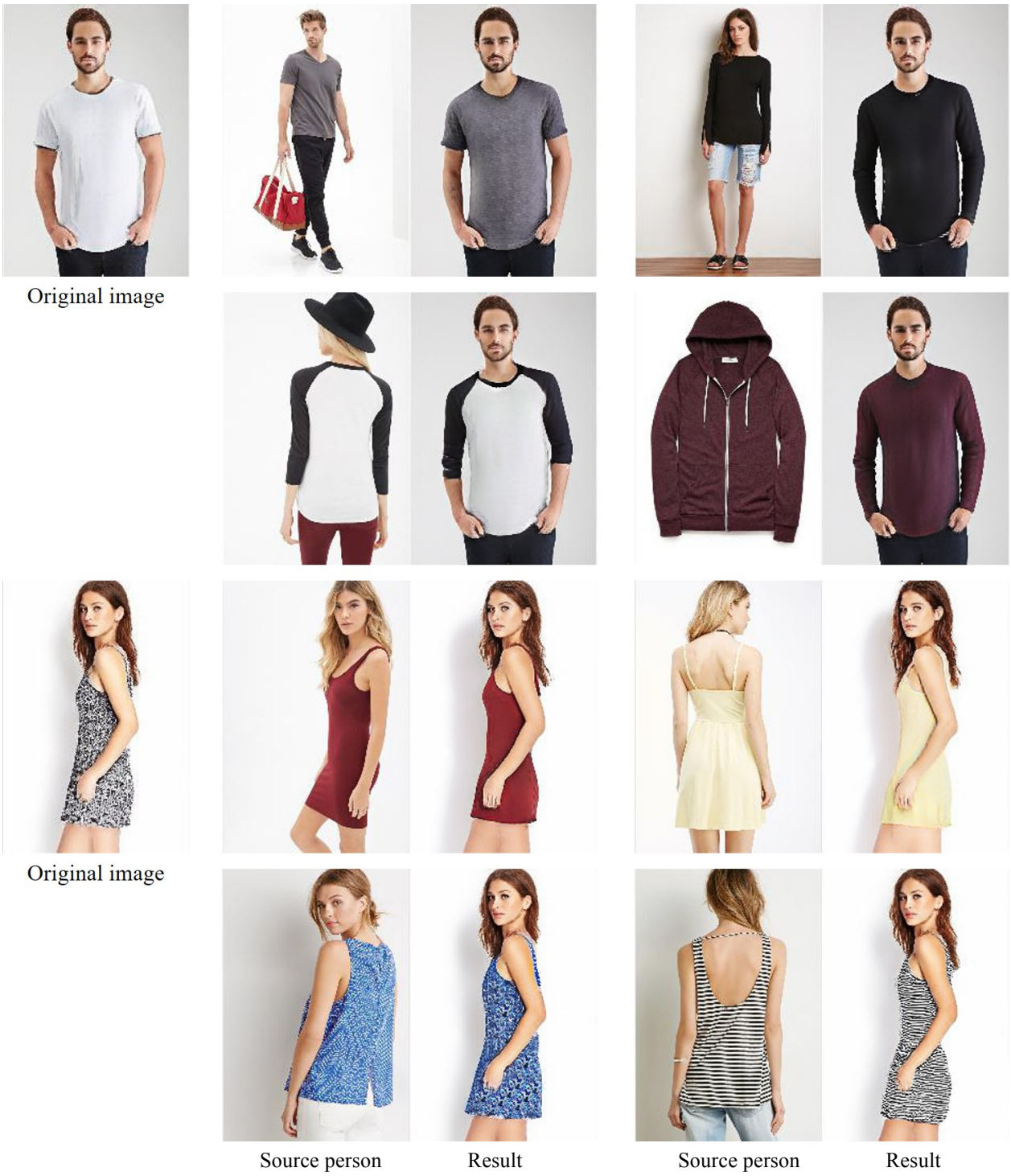

Figure 5: Results of synthesizing person images with controllable upper clothes.

Controllable pants.

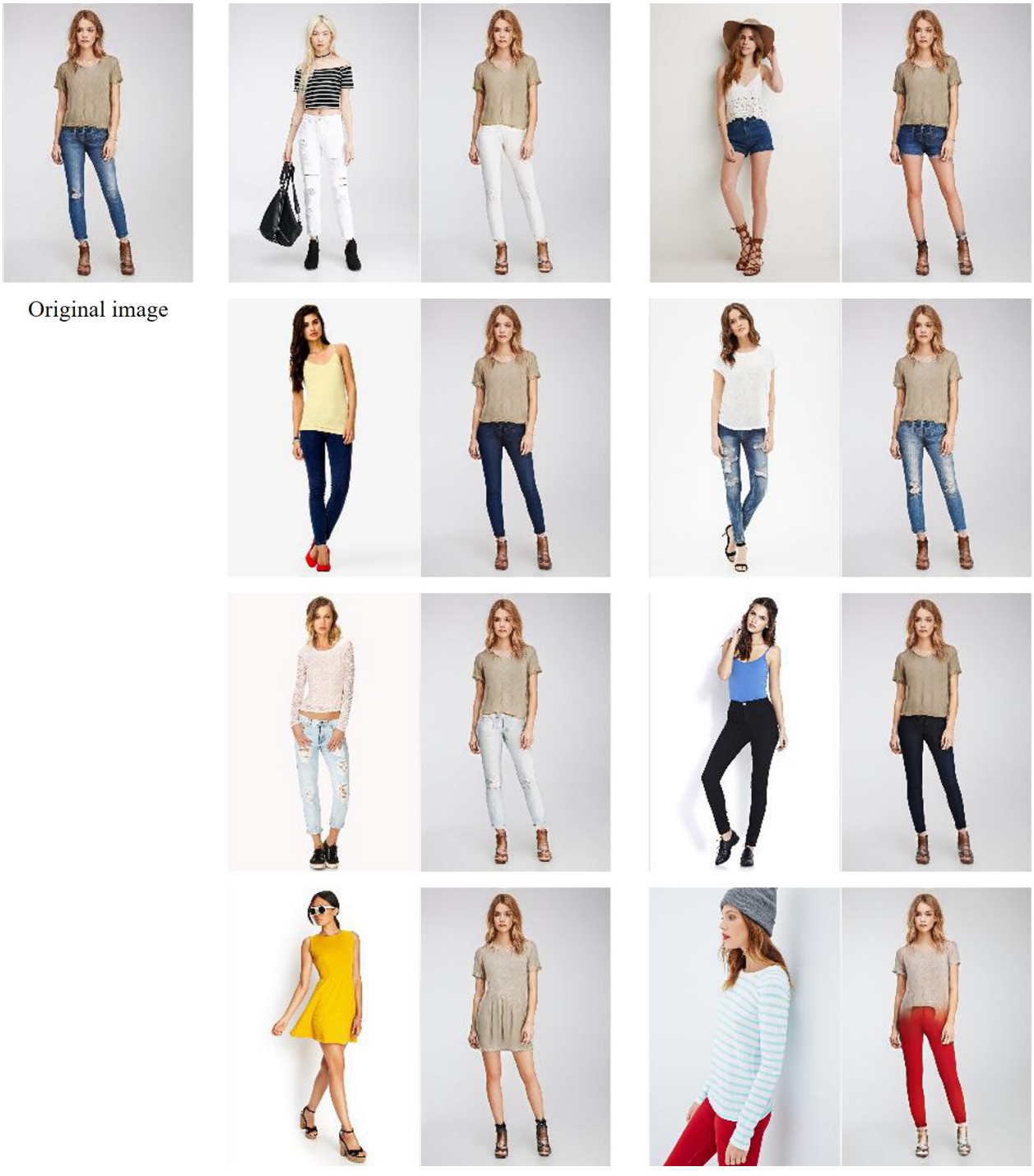

Figure 6: Results of synthesizing person images with controllable pants.

Controllable head attributes.

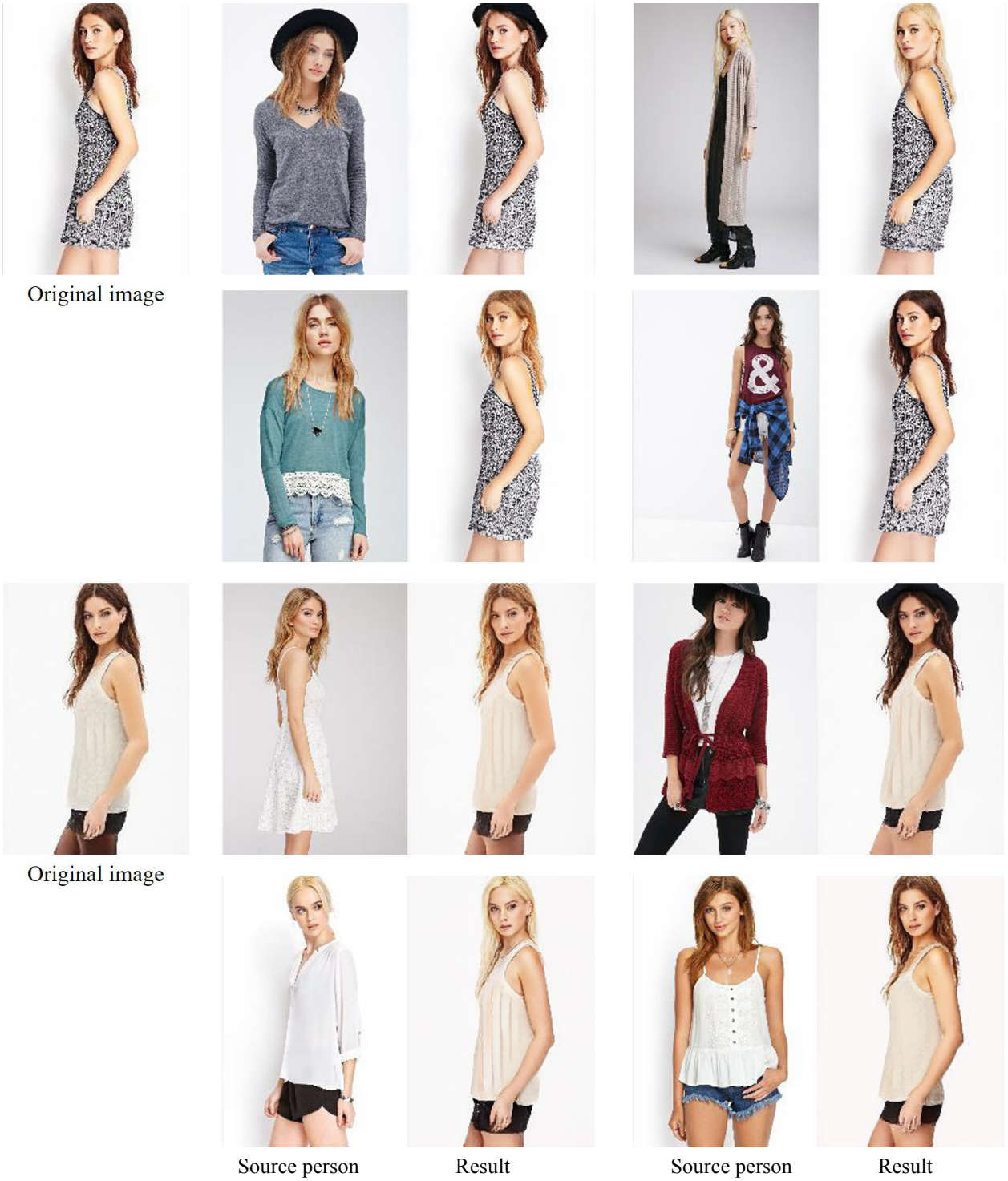

Figure 7: Results of synthesizing person images with controllable head attributes.

3. Comparisons with state-of-the-art methods

Figure 8: Comparison with state-of-the-art pose transfer methods.

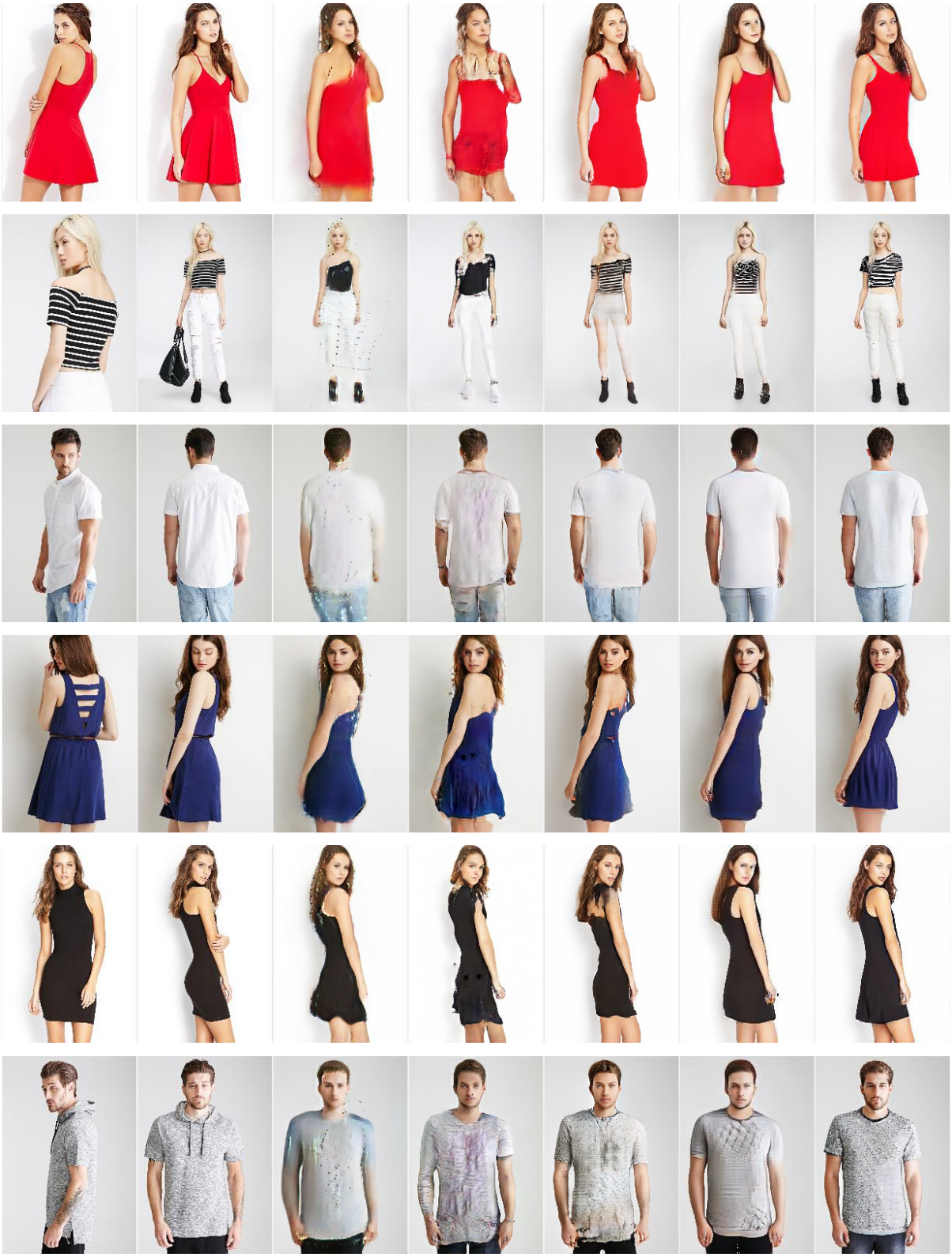

Figure 9: Comparison with state-of-the-art pose transfer methods.

4. Style interpolation results.

Results of some representative frames are depicted in Figure 10 and more detailed results are available in the supplemental video.

Figure 10: Style interpolation results of some representative frames as β decreases.

5. Experimental results with human parsers having different accuracies.

Since the model decompose the person image to components using extracted semantic layouts, we explore the influence caused by two human parsers with different accuracies.

We use SegNet [5], a classic segmentation model (18.2% IoU in LIP [6]), and Graphonomy [7], a state-of-the-art mothed proposed recently (58.58% IoU in CIHP [8]) to train our model, respectively. The evaluation results in Table1 show that there is no obvious difference with these two parsers. Since our model integrate the original labels into 8 main categories, which relaxes the requirements of parsing accuracy to an extent. Rough segmentation borders do not have obvious influence but different classification results by parsers for the same category may need some heuristic merging schemes.

Parser	IS	SSIM
SegNet [5]	3.364	0.772
Graphonomy [7]	3.367	0.764

Table 1: The performance of models trained with different human parsers.

Bibliography

[1] 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 406-416, 2017.

[2] Liqian Ma, Qianru Sun, Stamatios Georgoulis, Luc Van Gool, Bernt Schiele, and Mario Fritz. Disentangled person image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 99-108, 2018.

[3] Aliaksandr Siarohin, Enver Sangineto, Stephane Lathuiliere, and Nicu Sebe. Deformable gans for pose-based human image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3408-3416, 2018.

[4] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 2347-2356, 2019.

[5] Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE transactions on pattern analysis and machine intelligence, 39(12):2481-2495, 2017.

[6] Ke Gong, Xiaodan Liang, Dongyu Zhang, Xiaohui Shen, and Liang Lin. Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 932-940, 2017.

[7] Ke Gong, Yiming Gao, Xiaodan Liang, Xiaohui Shen, Meng Wang, and Liang Lin. Graphonomy: Universal human parsing via graph transfer learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 7450-7459, 2019.

[8] Ke Gong, Xiaodan Liang, Yicheng Li, Yimin Chen, Ming Yang, and Liang Lin. Instance-level human parsing via part grouping network. In Proceedings of the European Conference on Computer Vision (ECCV), pages 770-785, 2018.