Structure-Aware Flow Generation for Human Body Reshaping Supplementary Material

Jianqiang Ren, Yuan Yao, Biwen Lei, Miaomiao Cui, Xuansong Xie
DAMO Academy, Alibaba Group
{jianqiang.rjq, ryan.yy, biwen.lbw, miaomiao.cmm}@alibaba-inc.com, xingtong.xxs@taobao.com

This supplementary material provides additional details and results of our approach. We first discuss privacy and ethical concerns in Sec. 1. Then, we present details about the BR-5K dataset and our networks in Sec. 2. Lastly, Sec. 3 presents the settings of compared methods and more experimental results.

1. Privacy and Ethical Concerns

Usage of Person Images. Since our task aims to manipulate human bodies in images and we declare to release the BR-5K dataset, we carefully scrutinized our collection procedure for privacy and ethics concerns. As mentioned in the main paper, we collected all the images in our work from Unsplash website [3], which grants us an irrevocable, nonexclusive, worldwide copyright license (https://unsplash.com/license) to download, copy, modify, distribute, perform, and use photos from Unsplash for free, including for commercial purposes, without permission from or attributing the photographer or Unsplash. Therefore, The license of Unsplash approves us to use these photos in our research legally. The collection and use of these data have also been approved by IRB.

In addition, to better protect privacy, we conduct face obfuscation by blurring faces in the proposed BR-5K dataset, as we find that the blurred faces do not affect the body reshaping task.

Potential Societal Impact. The goal of our work is to achieve automatic adjustment of human bodies to generate shapely and attractive portrait images. Therefore, it can be used to improve the efficiency of portrait photography retouching pipeline, bringing a better experience for both photographers and customers.

Nevertheless, owing to the controllability of our method (Figure 8 in the main paper), misuse of the technology and dataset may lead to ethical concerns(e.g., misinformation). We propose some solutions to tackle these issues. Firstly, we will restrict the authentication for the BR-5K dataset, anyone who wants to acquire the dataset must download and sign an End User License Agreement and agree to use the dataset for resarch purposes only. Secondly, for company or organizations that want to employ our technology, our legal department will carefully check their certifications(e.g., business license, business type, customer group), and request them to obtain the permission from photo owner before they can edit the photo using our method.

2. Implementation Details

2.1. BR-5K Dataset

We initially collected more than 20,000 portrait photos from Unsplash [3]. After discarding photos with multiple persons, we abandoned photos with too tiny or heavily occluded human subjects. We then carefully selected photos whose figure we think can be more shapely after body retouching, and obtained 5,000 individual portrait photos. We invited three professional artists to retouch these 5,000 photos independently. Among the three results for each photo, we selected the one with the most believable and attractive figure as ground-truth. In Figure 5, we provide more samples (the first column) and their ground-truths (the last column) retouched by artists.

2.2. PAFs and Skeleton Maps

Part Affinity Fields (PAFs) are initially present to learn the association between body parts. Different from the original PAFs in [1], we customize our PAFs according to body reshaping task. Specifically, we modify the fields in the torso and remove fields on the head and shoulders, and then apply dilation operation to generate our PAFs. The comparison between original PAFs and ours is depicted in Figure 2. The skeleton maps are also designed in a similar way. As shown in Figure 3, the skeleton maps and PAFs have 12 and 10 channels respectively. We utilize the pose estimation code from https://github.com/tensorboy/pytorch_Realtime_Multi-Person_Pose_Estimation.

2.3. Structure Encoding in SASA

As mentioned in the main paper Sec. 4.2, the structure heatmaps Y are calculated from PAFs by *structure encoding*, which integrates related body part masks into one mask. Figure 4 demonstrates each channel in the structure heatmaps, which corresponds to Eq. 2 in the main paper. After *structure encoding*, each pixel is assigned with a 5-dimension code represented by structure heatmaps, thus the similarity between two pixels' codes can be used to measure the structural correlations between them.

2.4. Warping Operation

Flow fields can be used to deform images, which we denote as warping operation $\mathcal{W}$ in the main paper Sec. 4.3. The two channels of flow field F represent the deformation offsets on horizontal and vertical directions, respectively. Thus the warping operation:

$$O = \mathcal{W}(I; \mu F) \tag{1}$$

is equivalent to:

$$O(x, y) = I(x + \mu F_x(x, y), y + \mu F_y(x, y)) \tag{2}$$

where I and O denote the input and warped images respectively. Since the flow values are floating-point numbers, the target pixel values are generally calculated by bilinear interpolation in practice.

2.5. Local Adjustment for Certain Body Parts

As PAFs could locate which body areas to be manipulated, a flow mask can be easily derived from PAFs to wipe out flow on certain parts via multiplication with the full flow, and achieve local adjustment(e.g., wiping out flow except for arms in Fig. 1).

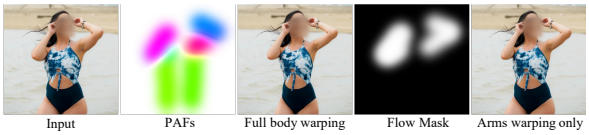


Figure 1. Result of local adjustment for just warping the arms.

3. Experiments and Results

3.1. Compared Methods

We employ four state-of-the-art methods for image translation and manipulation (FAL [4], ATW [6], pix2pixHD [5], GFLA [2]) and evaluate their applicability on body reshaping task. To make a fair comparison, we use their released codes and retrain these methods on the BR-5K dataset with default configurations. As the original ATW method is designed for facial expression animation, it needs an additional driving signal as input. In our implementation, we slightly modify the ATW by eliminating the driving signal and driving loss, which makes it more suitable for our task.

3.2. Additional Results

More Visual Comparisons. We show more qualitative comparison results in Figure 5. Compared with other methods, our approach can produce high-resolution, consistent, and visually pleasing body editing results.

High-Resolution Image Manipulation. To demonstrate the capability of our method to edit high-resolution photos, we present the body reshaping results of 6K-resolution images in Figure 6 and Figure 7. The deformation flows are upsampled from 256 $\times$ 256 pixels and perform well in handling high-resolution body reshaping.

Weight and Height Manipulation. Since we take human skeletons as priors, our method concentrates on body weight editing only (without changing skeletal lengths). By simply combing with non-uniform scaling on body length direction, we can achieve comprehensive reshaping on human bodies as shown in Figure 8.

References

[1] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 1, 3

[2] Yurui Ren, Xiaoming Yu, Junming Chen, Thomas H Li, and Ge Li. Deep image spatial transformation for person image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7690–7699, 2020. 2

[3] Unsplash. Unsplash. https://unsplash.com/, 2021. 1

[4] Sheng-Yu Wang, Oliver Wang, Andrew Owens, Richard Zhang, and Alexei A Efros. Detecting photoshopped faces by scripting photoshop. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10072–10081, 2019. 2

[5] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 2

[6] Zili Yi, Qiang Tang, Vishnu Sanjay Ramiya Srinivasan, and Zhan Xu. Animating through warping: An efficient method for high-quality facial expression animation. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 1459–1468, 2020. 2

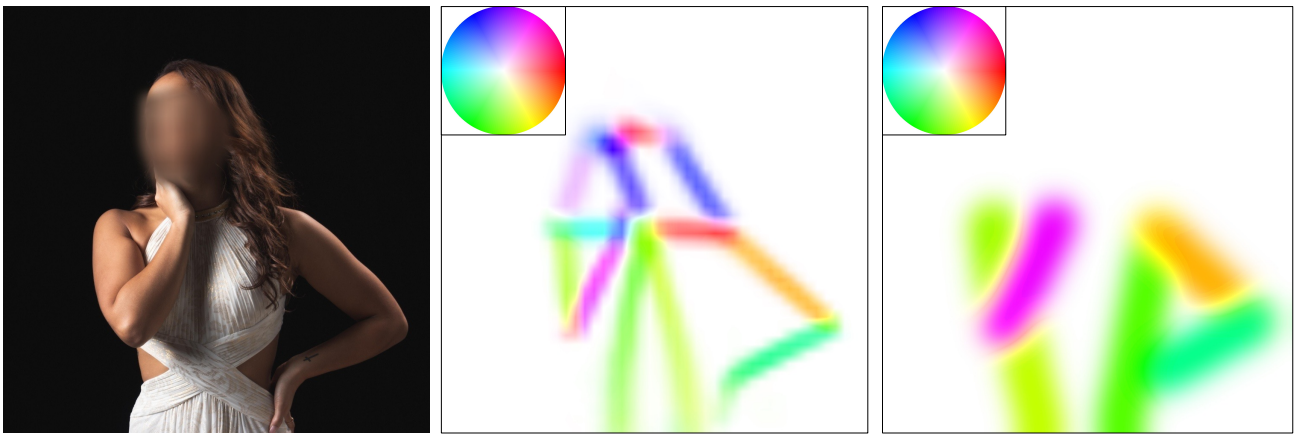


Figure 2. Different from the original PAFs in [1], we modify the fields in the torso and remove fields on the head and shoulders, and then apply dilation operation to generate our PAFs.

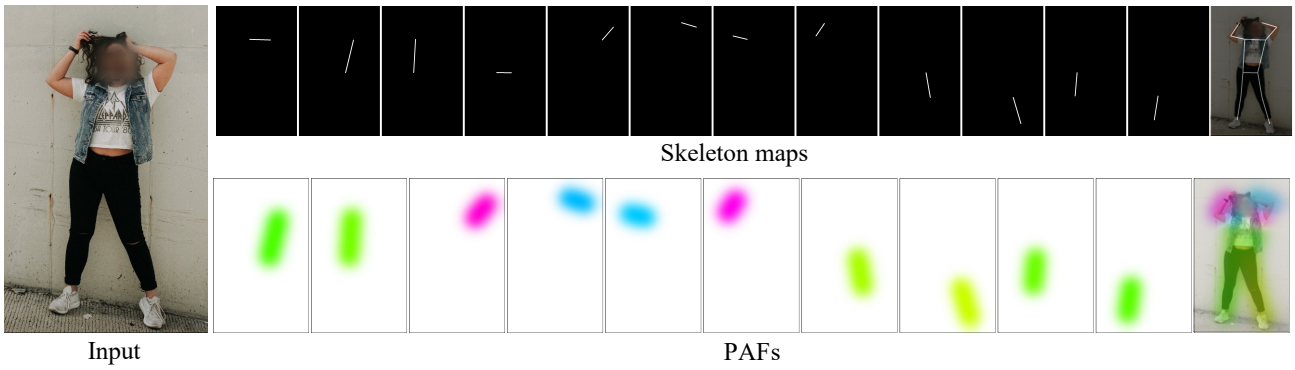


Figure 3. Visualization of each channel in skeleton maps and PAFs. The skeleton maps and PAFs have 12 and 10 channels respectively.

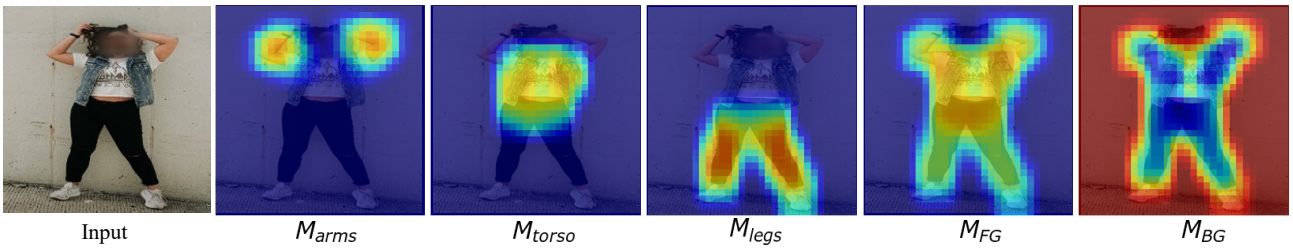


Figure 4. Visualization of each channel in structure heatmaps. The masks of body parts are obtained according to PAFs magnitude.

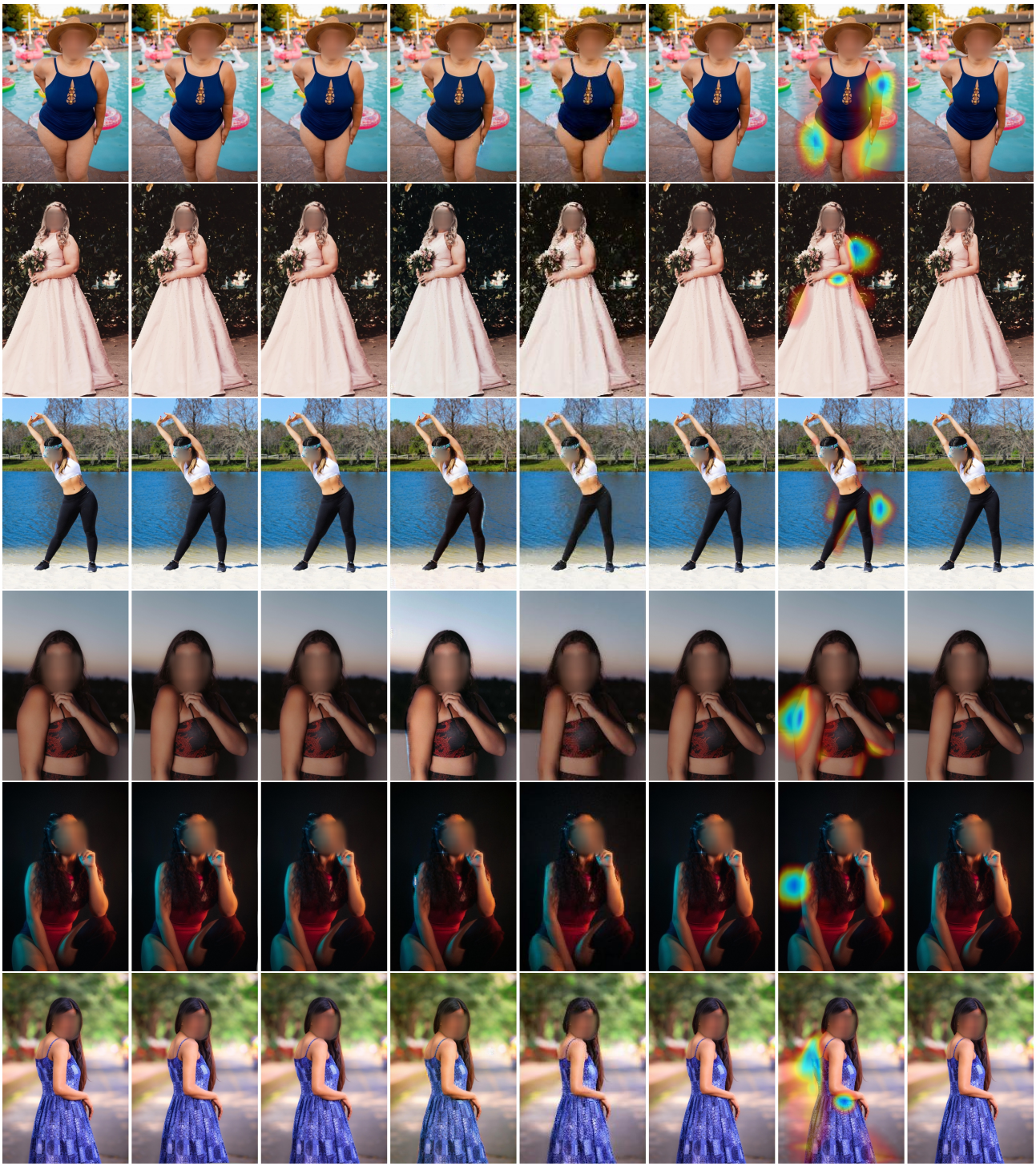


Figure 5. More visual comparisons among different methods. Our method can produce high-resolution, believable, and consistent body reshaping results. Zoom in for details.

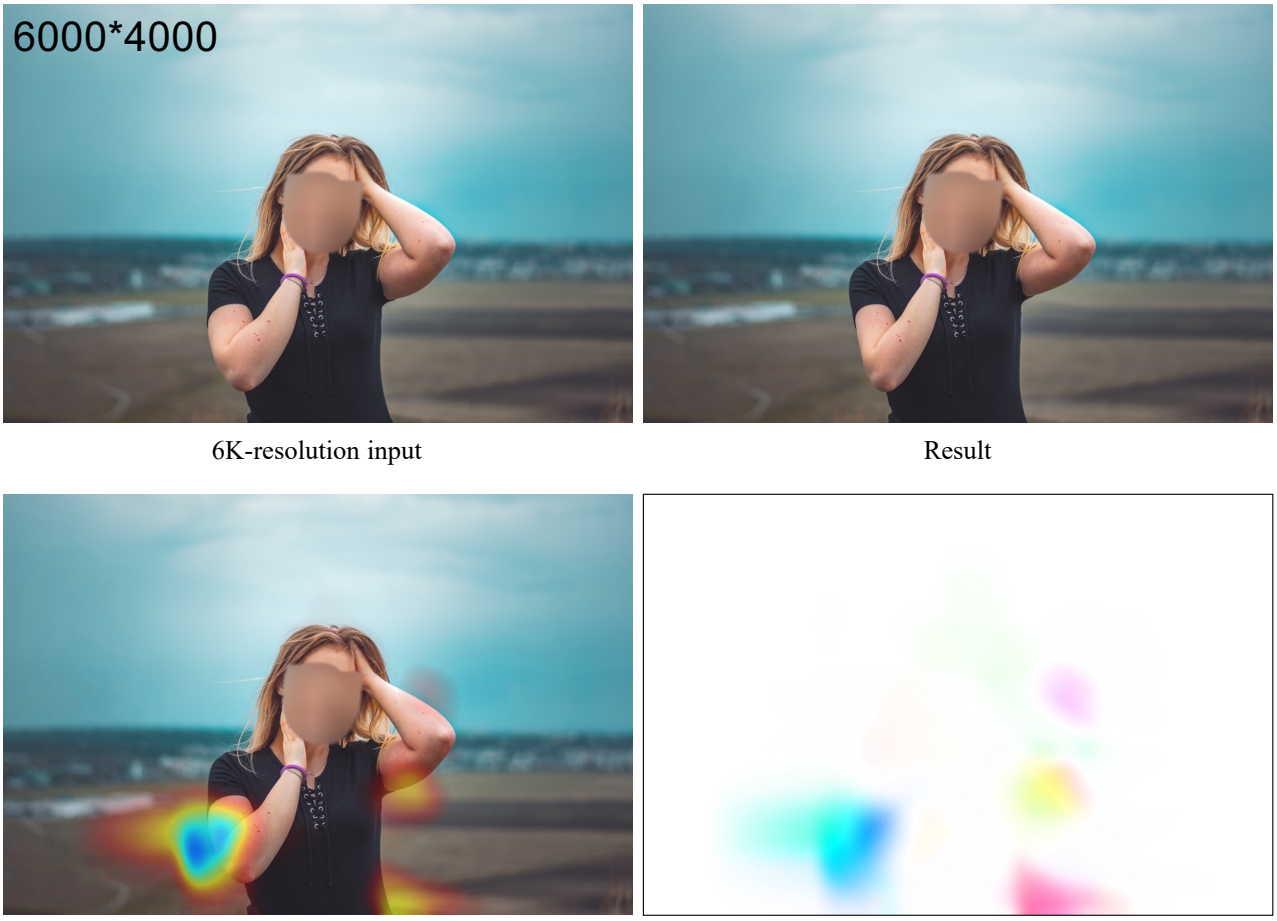
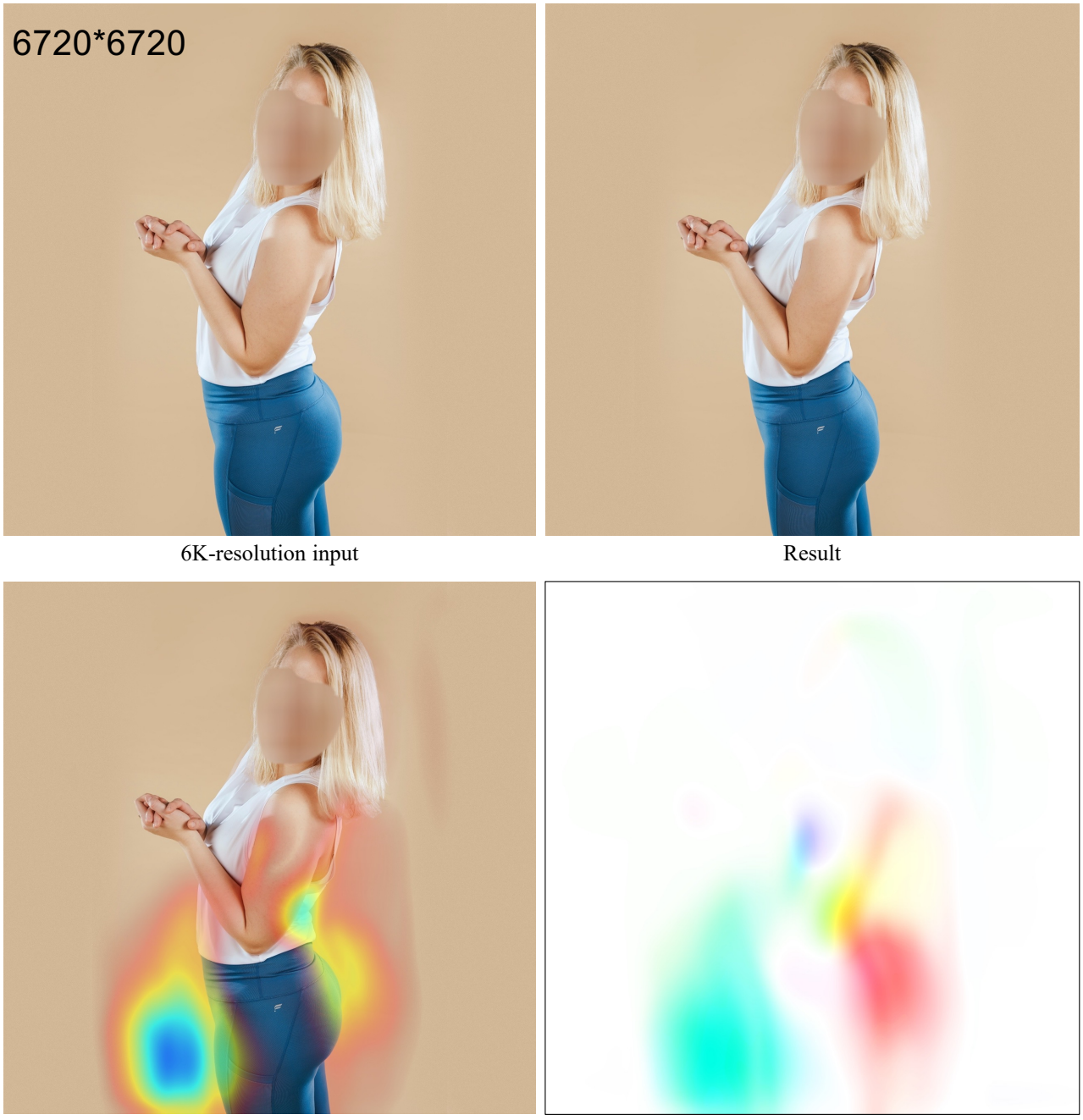


Figure 6. Visual examples of 6K-resolution photos and their reshaping results by our methods. The generated flow fields are smooth enough to handle high-resolution images.



6K-resolution input | Result

Flow overlay | Deformation flow

Figure 7. Visual examples of 6K-resolution photos and their reshaping results by our methods. The generated flow fields are smooth enough to handle high-resolution images.

Input | Losing weight | Losing weight + Lengthening height

Figure 8. Since we take human skeletons as priors, our reshaping method concentrates on weight editing only (without changing skeletal lengths). By simply combing with non-uniform scaling on body length direction, we can achieve comprehensive reshaping on human bodies.