DeepFaceReshaping: Interactive Deep Face Reshaping via Landmark Manipulation —*Supplementary Material*

I. Additional Implementation Details

A. Network Architecture

a) Partial Refinement Network: Our partial refinement network comprises two parts: the appearance encoder and the backbone. We list the details of the backbone in Tab. I, which have been briefly discussed in the paper.

Layer	Normalization	Noise	Output Shape
input L^c_{out}	-	-	$512 \times 512 \times 3$
$Conv7 \times 7s1$	IN	-	$512 \times 512 \times 32$
$Conv3 \times 3s2$	IN	-	$256 \times 256 \times 64$
$Conv3 \times 3s2$	IN	-	$128 \times 128 \times 128$
$Conv3 \times 3s2$	IN	-	$64 \times 64 \times 256$
$Conv3 \times 3s2$	IN	-	$32 \times 32 \times 512$
ResBlk	SaBN	$\surd$	$32 \times 32 \times 512$
ResBlk	SaBN	$\surd$	$32 \times 32 \times 512$
ResBlk	SaBN	$\surd$	$32 \times 32 \times 512$
ResBlk	SaBN	$\surd$	$32 \times 32 \times 512$
ResBlk	SaBN	$\surd$	$32 \times 32 \times 512$
ResBlk	SaBN	$\surd$	$32 \times 32 \times 512$
ResBlk	SaBN	$\surd$	$32 \times 32 \times 512$
ResBlk	SaBN	$\surd$	$32 \times 32 \times 512$
$ConvT3 \times 3s2$	IN	-	$64 \times 64 \times 256$
$ConvT3 \times 3s2$	IN	-	$128 \times 128 \times 128$
$ConvT3 \times 3s2$	IN	-	$256 \times 256 \times 64$
$ConvT3 \times 3s2$	IN	-	$512 \times 512 \times 32$
$Conv7 \times 7s1$	-	-	$512 \times 512 \times 3$

TABLE I: Details of the backbone of our partial refinement network. The landmark graphs L^c_{out} of different facial components have different input shapes. Here we show the change of the output shape when $c = 4$ and L^c_{out} corresponds to the part of "face outline". $Conv3 \times 3s2$ denotes a 3×3 Convolution-InstanceNorm-ReLU layer with stride 2 and $ConvT3 \times 3s2$ denotes a 3×3 fractional-strided-Convolution-InstanceNorm-ReLU layer with stride 2. SaBN stands for the Sandwich Batch Normalization [1].

b) Global Fusion Network: Our global fusion network derives from the Unet adopted in Isola et al.[2], details are listed in Tab. II. It is noteworthy to mention that there exist skip connections between the down-sampling and upsampling layers i and $8-i$. Here the inputs are the concatenated features obtained from four local embedding networks.

B. Training Details

For the 512^2 resolution setting, we train each of our partial refinement network on a single NVIDIA 2080ti GPU for 12 days, and train the global fusion network on an NVIDIA 2080ti GPU for 2 days.

Layer	Normalization	Output Shape
input	-	$512 \times 512 \times 128$
$Conv3 \times 3s2$	IN	$256 \times 256 \times 64$
$Conv3 \times 3s2$	IN	$128 \times 128 \times 128$
$Conv3 \times 3s2$	IN	$64 \times 64 \times 256$
$Conv3 \times 3s2$	IN	$32 \times 32 \times 512$
$ConvT3 \times 3s2$	IN	$64 \times 64 \times 256$
$ConvT3 \times 3s2$	IN	$128 \times 128 \times 128$
$ConvT3 \times 3s2$	IN	$256 \times 256 \times 64$
$ConvT3 \times 3s2$	IN	$512 \times 512 \times 3$

TABLE II: Details of global fusion network. Specifically, $Conv3 \times 3s2$ denotes a 3×3 Convolution-InstanceNorm-ReLU layer with stride 2 and $ConvT3 \times 3s2$ denotes a 3×3 fractional-strided-Convolution-InstanceNorm-ReLU layer with stride 2.

II. Additional Discussions

a) The deformation of "face outline": Although our latent space optimization is performed component by component, the deformation of face outline is calculated on the landmarks of the whole face, causing global effects. In certain cases, it may leads to unpleasing results if we restrict the change of landmarks to a local inner part. For example, as shown in the results (b) Fig. 1, making a mouth open and smile will also affect the shape of the jaw, while the results (a) where we restrict the change of landmarks to the mouth are less natural.

b) The generation of wrinkles: As wrinkles belong to texture rather than shape in the image, our method does not support the direct control of wrinkles. However in the examples shown in Fig. 2, some wrinkles can be dynamically synthesized and removed. This is presumably because dimpled wrinkles are closely related to smiles in the dataset. Since there are too few elderly people in the dataset, thus a limited number of samples of wrinkles such as forehead wrinkles and nasolabial folds, it is difficult to synthesize various wrinkles just by manipulating the landmarks.

c) The changes of ears after jawline editing: Because the ears are very close to the cheeks, the generator will also automatically adjust the corresponding ear size when the jaw is modified greatly. When the jaw is stretched longitudinally, the ears tend not to change, as illustrated in Fig. 3.

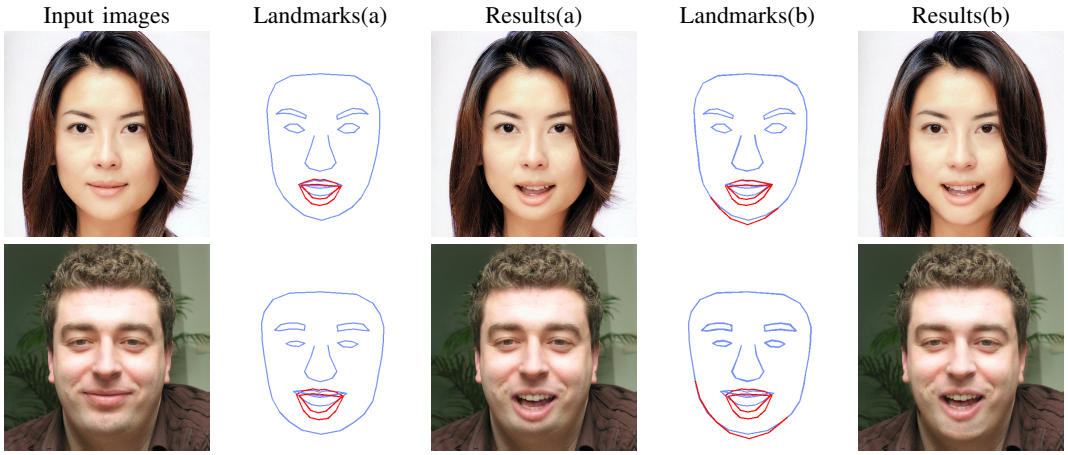

Fig. 1: Visual comparison of different conditions on opening the mouth. In results (a), we restrict the change of landmarks to the mouth and generate the results from the spatially restricted landmarks. In situation (b), we show the original results of our method, where the change of mouth landmarks also affect the shape of face outline. We highlight the changed lines of landmarks in red.

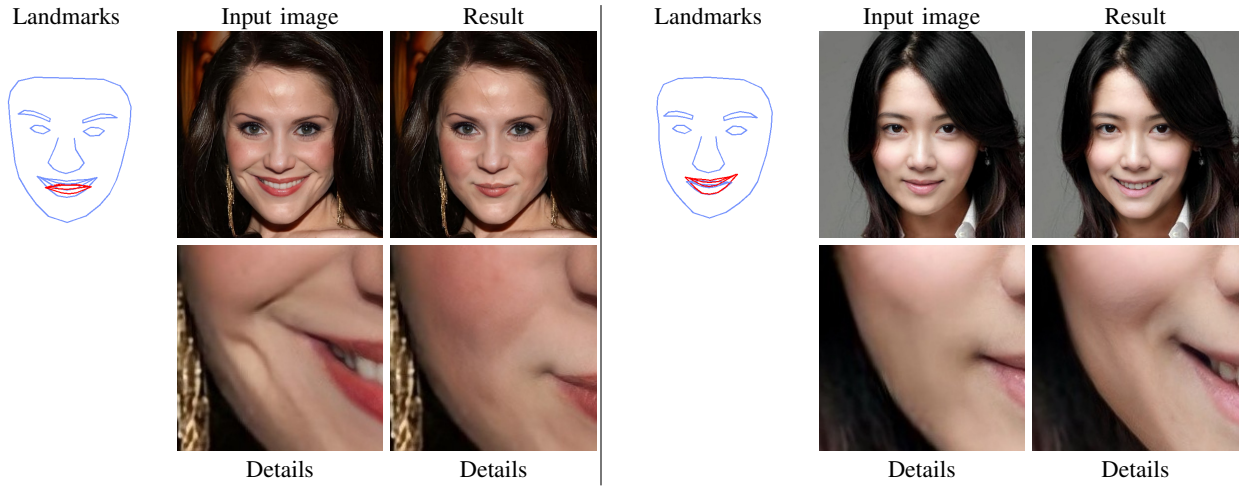

Fig. 2: Examples of removing and adding wrinkles with respect to smiling. We show the automatically generated and removed dimples.

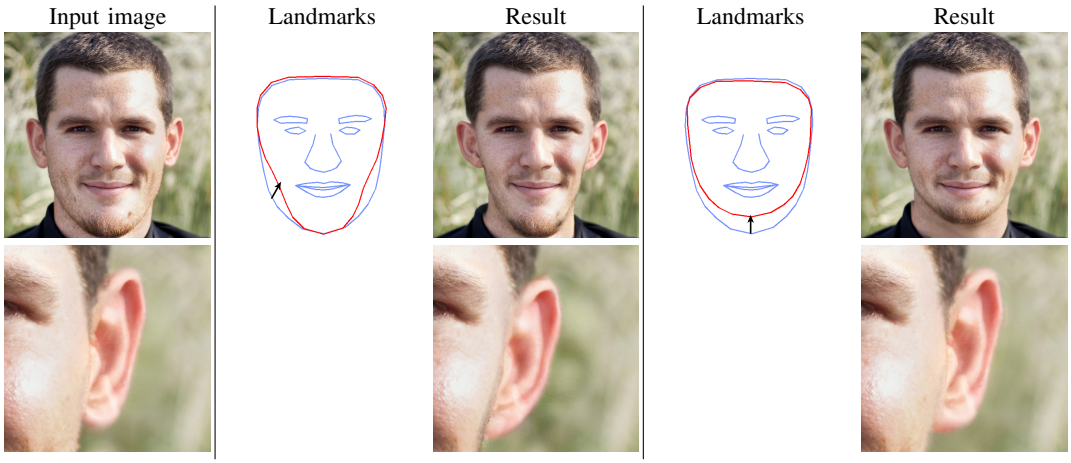

Fig. 3: Examples of the changes of ears before and after jawline editing.

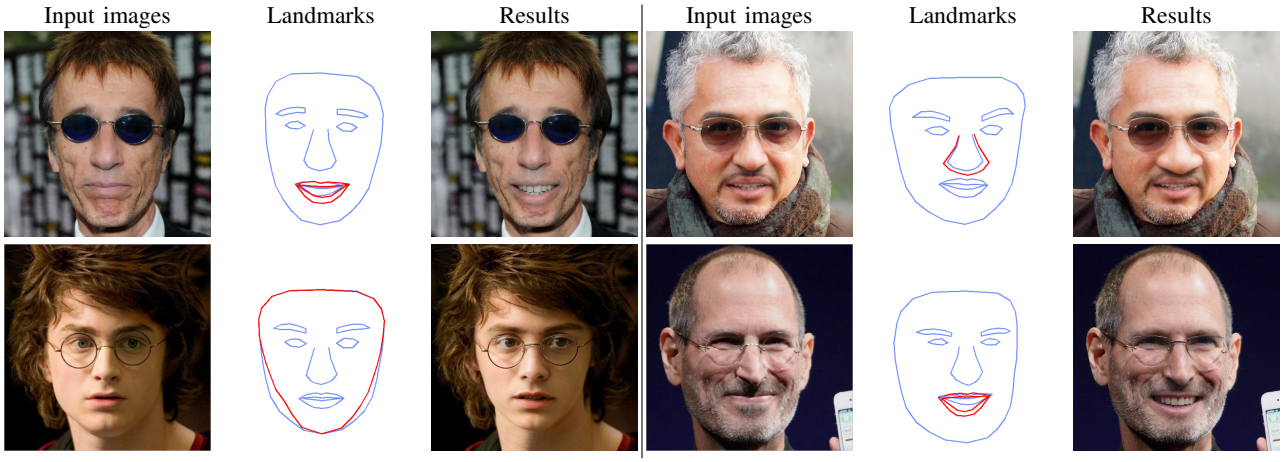

Fig. 4: Visual results of our method on faces with glasses.

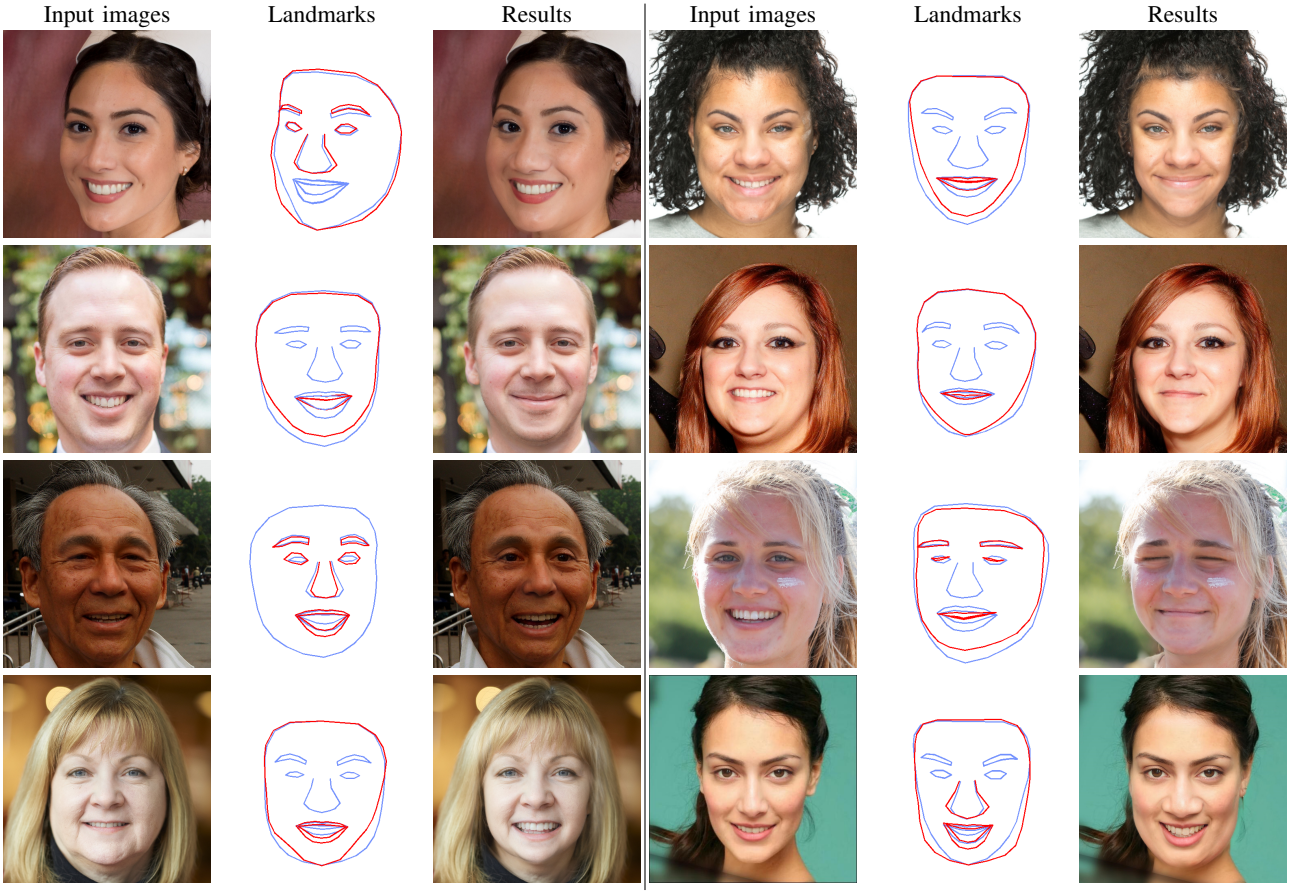

Fig. 5: Additional visual results of our method. We highlight the changed lines of landmarks in red.

d) Reshaping effects on faces with glasses: We show our reshaping effects on faces with glasses in Fig. 4. Our method cannot reshape the eyes with glasses properly due to the unsuccessful landmark detection for eyes, but can still reshape the other regions because of our adopted local-to-global structure.

III. Additional Visual Results

In this section we add more visual results. We show more visual results in Fig. 5 and Fig. 10. More results of comparison and ablation study are shown in Fig. 6 and Fig. 7. We also show users' editing results in our usability study in Fig. 8.

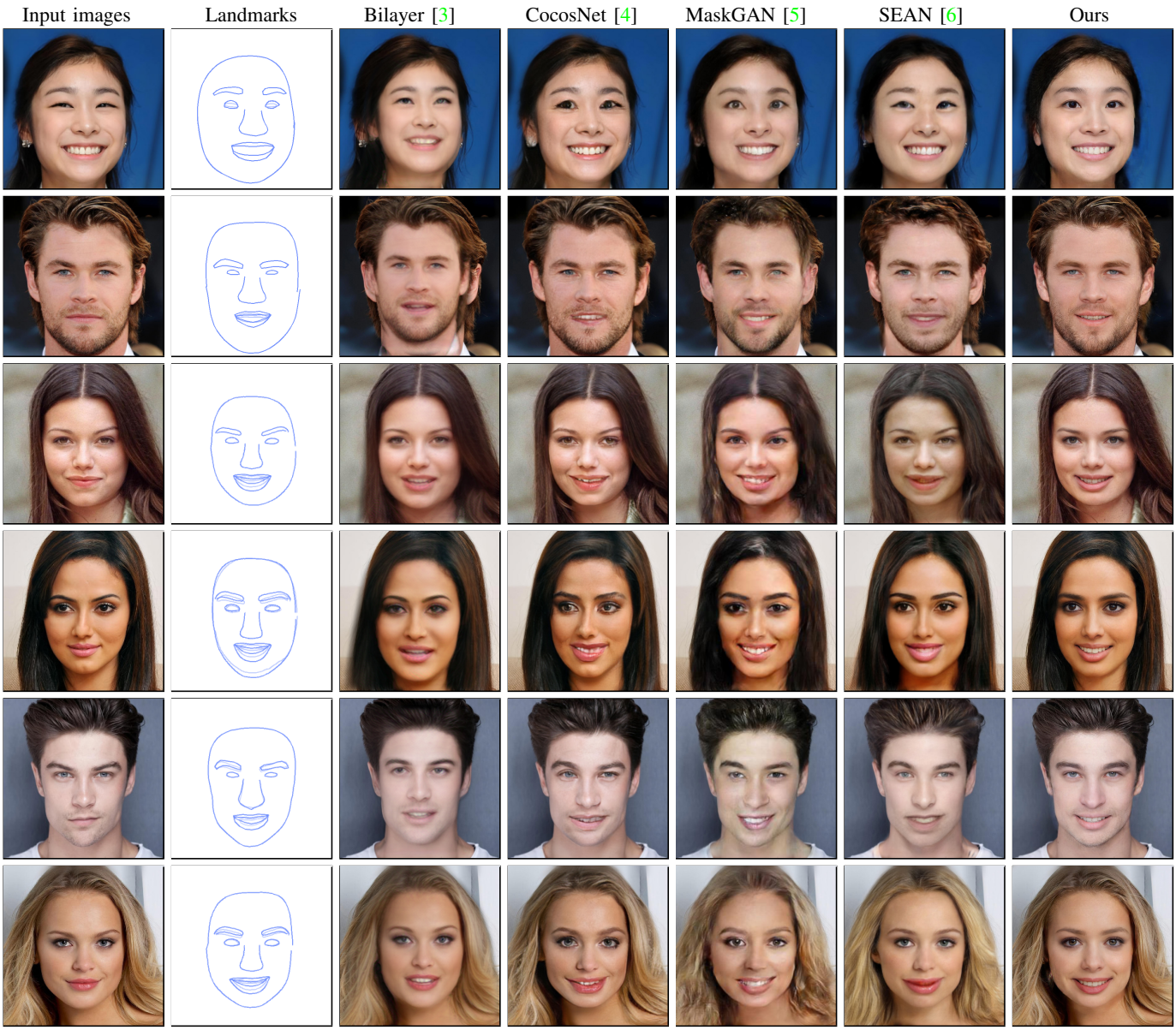

Fig. 6: Additional comparisons with the state-of-the-art methods given the same deformed landmarks (the second column). The original landmarks are marked in light blue with the deformed landmarks overlapped in blue. For fair comparison, the background is added by one copy-paste-blend step for the results of all the face generation methods.

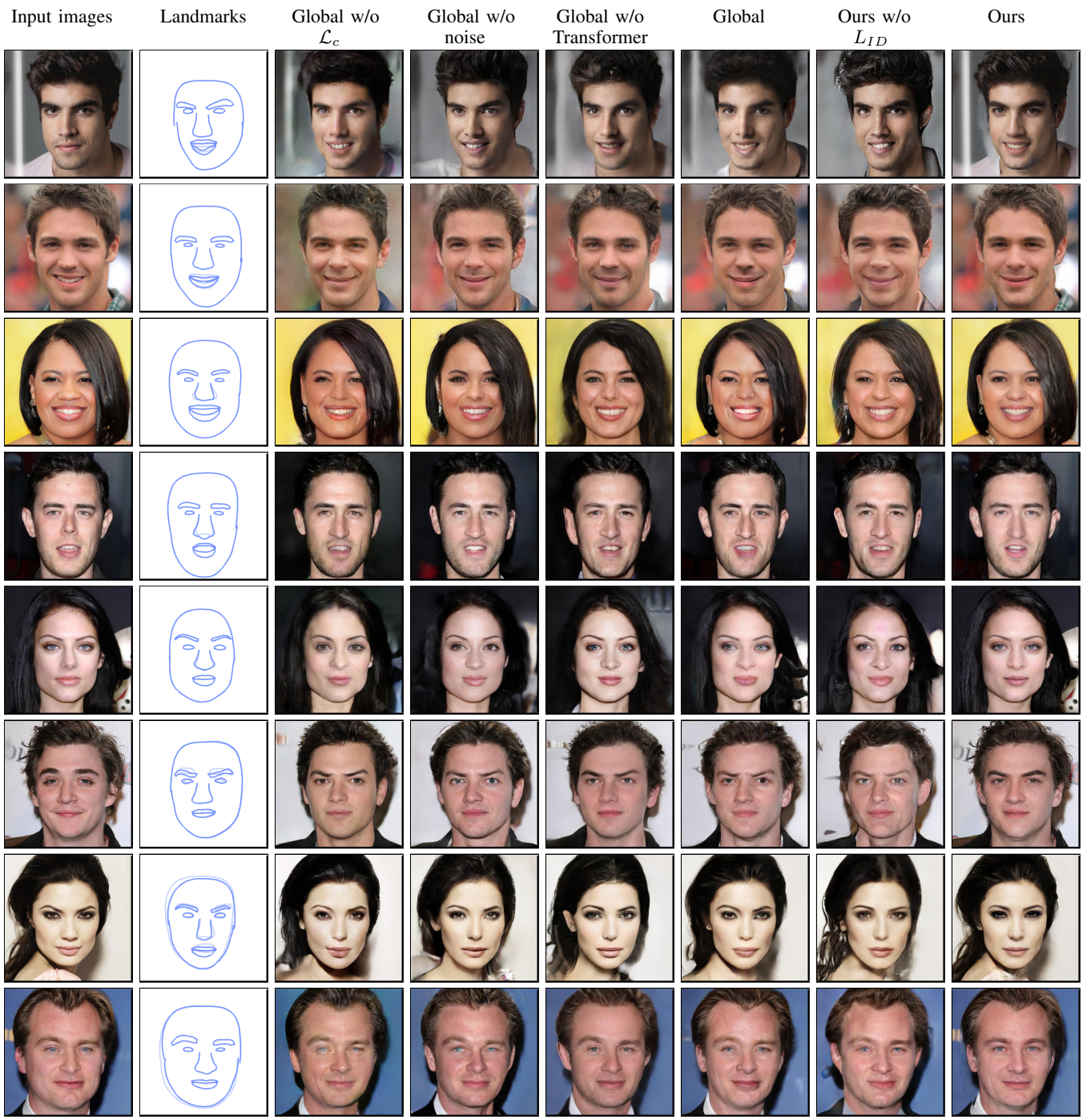

Fig. 7: Additional results of ablation study. We visualize the reshaped generated results under different settings. Our full framework achieves the best results.

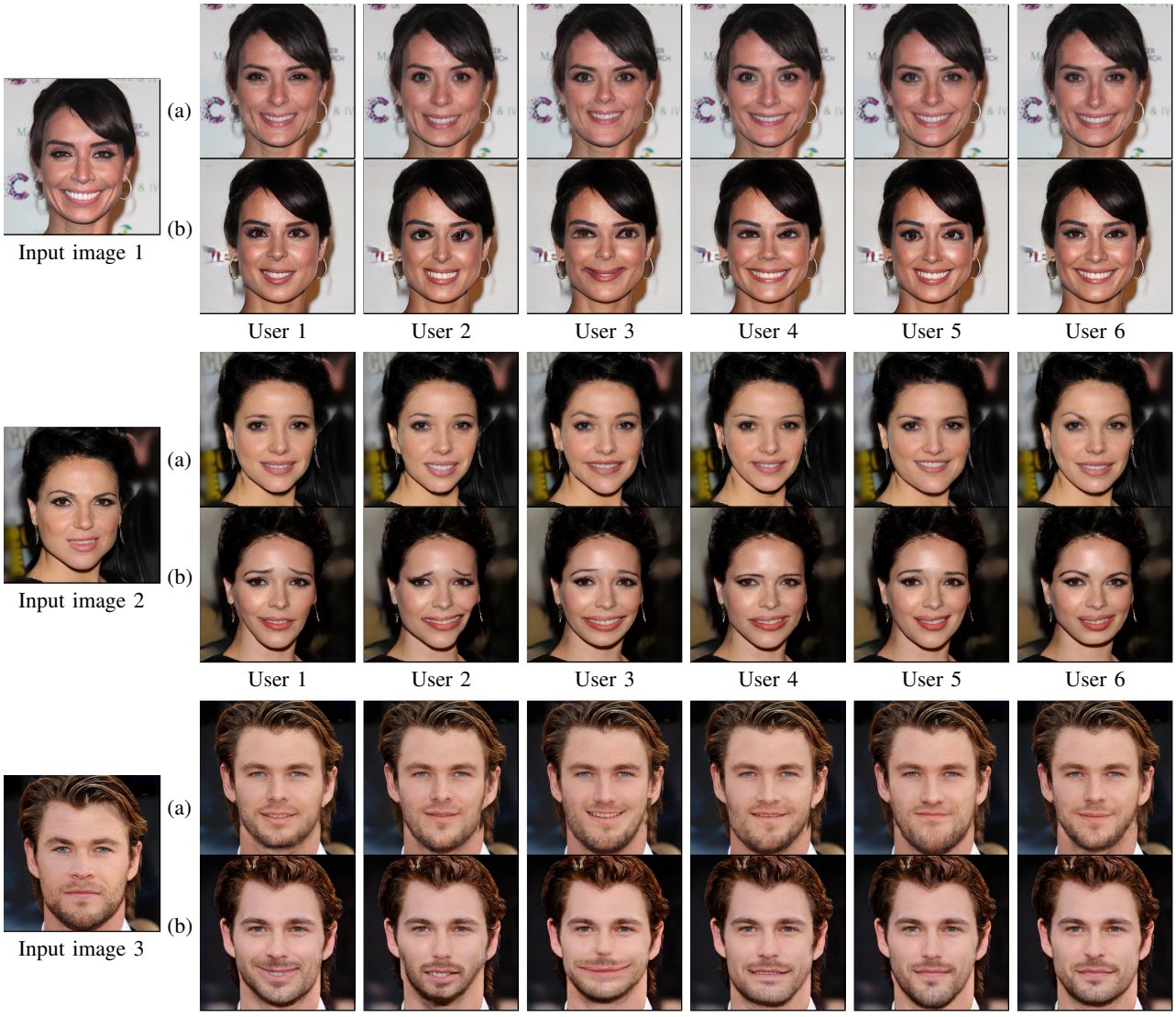

Fig. 8: Users' editing results using our system (a), the official drawing interface (b) in [5]. Our method has better control over identity.

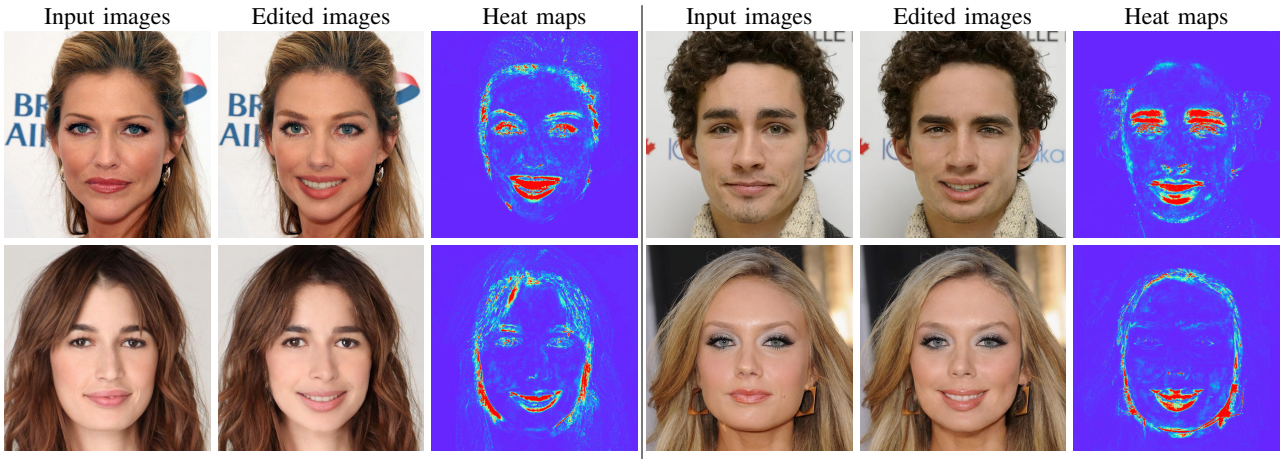

Fig. 9: Heat maps corresponding to our results of Fig. 7 of the main paper. It can be seen that the edited areas differ greatly, while other areas such as skin actually have no obvious color difference.

Fig. 10: More Visual Results.

REFERENCES

[1] X. Gong, W. Chen, T. Chen, and Z. Wang, "Sandwich batch normalization," *arXiv preprint arXiv:2102.11382*, 2021. 1

[2] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 1

[3] E. Zakharov, A. Ivakhnenko, A. Shysheya, and V. Lempitsky, "Fast bi-layer neural synthesis of one-shot realistic head avatars," in *European Conference of Computer vision (ECCV)*, August 2020. 4

[4] P. Zhang, B. Zhang, D. Chen, L. Yuan, and F. Wen, "Cross-domain correspondence learning for exemplar-based image translation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5143–5153. 4

[5] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5549–5558. 4, 6

[6] P. Zhu, R. Abdal, Y. Qin, and P. Wonka, "Sean: Image synthesis with semantic region-adaptive normalization," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 4