Supplementary Material for " Fine-Grained Face Swapping via Regional GAN Inversion"

Zhian Liu[1†], Maomao Li[2†], Yong Zhang[2†], Cairong Wang[3], Qi Zhang[2], Jue Wang[2], and Yongwei Nie[1*]
[1]School of Computer Science and Engineering, South China University of Technology, China
[2]Tencent AI Lab, Shenzhen, China [3]Tsinghua Shenzhen International Graduate School, China
csliuzhian@mail.scut.edu.cn nieyongwei@scut.edu.cn wcr20@mails.tsinghua.edu.cn
{limaomao07, zhangyong201303, nwpuqzhang, arphid}@gmail.com

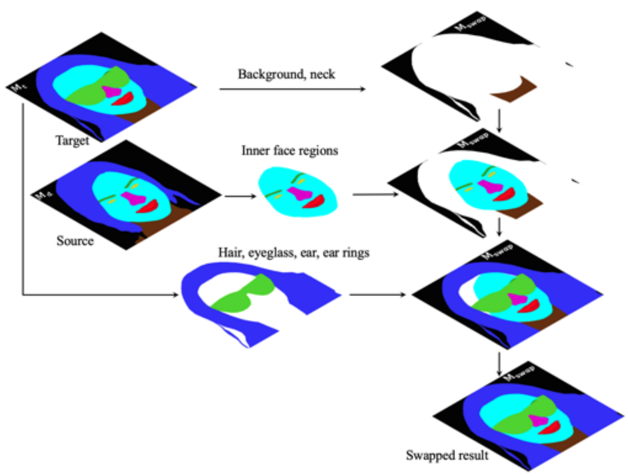

Figure 1. Detailed steps of the shape swapping process. We start with an empty mask as a canvas and then complete the mask recomposition gradually.

1. Shape Swapping Details

As described in Sec. 3.1 of the main paper, a shape swapping process is required to realize the aim of face swapping. Since facial masks represent the shape, the shape swapping is completed in a recomposition fashion. This process is illustrated in Fig. 1.

2. Loss Functions

Unlike most of the face swapping methods, we *do not* need to swap source and target face pairs in the training phase, but rely on the simple image reconstruction instead. We adopt the commonly-used loss functions in the GAN inversion literature.

Pixel-wise reconstruction loss. For the input image I, suppose the reconstructed image is $\hat{I}$, we use the Mean-Squared-Error (MSE) as the pixel-wise reconstruction loss:

$$\mathcal{L}_{mse} = \left\| \hat{I} - I \right\|_2^2 \tag{1}$$

Multi-scale LPIPS loss. Using MSE alone cannot produce sharp results. Inspired by [18], we use the multi-scale LPIPS [19] loss to encourage sharpness in the reconstructed images, which is expressed as:

$$\mathcal{L}_{ms\_lpips} = \sum_s \left\| \mathbf{V}(\lfloor \hat{I} \rfloor_s) - \mathbf{V}(\lfloor I \rfloor_s) \right\|_2^2 , \tag{2}$$

where $\mathbf{V}$ denotes the AlexNet [7] feature extractor pre-trained on ImageNet [8], $s \in \{256, 512, 1024\}$, and $\lfloor \hat{I} \rfloor_s$ represents the downsized input with the resolution of s.

Multi-scale face inversion loss. The ID loss was introduced in PSP [13] to preserve the identity of the input. Specifically, PSP uses a pre-trained face recognition network to maximize the cosine similarity between the input and the reconstructed face. Besides, the method [18] takes a step further to improve the ID loss within a multi-scale form, calculating the similarities in different feature levels. We follow these two works and apply the multi-scale ID loss constrain as:

$$\mathcal{L}_{ms\_id} = \sum_{i=1}^{5} \left(1 - \langle R_i(I), R_i(\hat{I}) \rangle \right) , \tag{3}$$

where R is the pre-trained ArcFace [4] model, and $\langle \cdot \rangle$ denotes the cosine similarity.

Moreover, we follow the work [18] to employ a multi-scale face parsing loss as:

$$\mathcal{L}_{ms\_parsing} = \sum_{i=1}^{5} \left(1 - \langle P_i(I), P_i(\hat{I}) \rangle \right) , \tag{4}$$

where P is the pre-trained face parser used in [9].

Zhian Liu does this work during an internship at Tencent AI Lab.
† Equal contribution. ∗ Corresponding Author.

Table 1. Quantitative comparison of our RGI under different ablative configurations. The reconstruction performance is measured.

Configurations	SSIM↑	PSNR↑	RMSE↓	FID↓
our RGI full model	**0.818**	**19.851**	**0.105**	**15.032**
(C) w/o $\mathcal{L}_{ms\_lpips}$	0.805	19.672	0.107	14.477
(D) w/o MS encoder	0.817	19.732	0.107	15.112

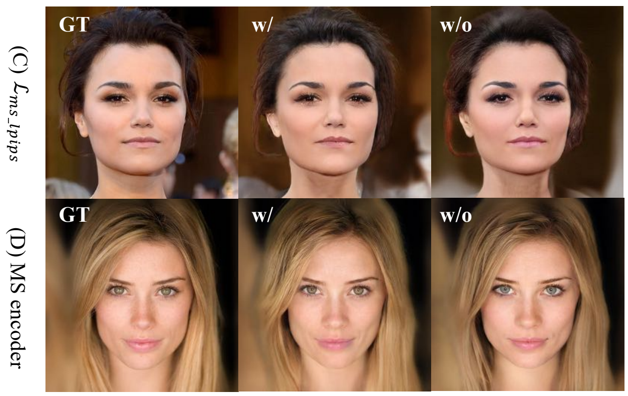

Figure 2. Qualitative comparison under different ablative configurations. For each row, we show the ground-truth, the reconstruction of our baseline, and the reconstruction of the corresponding configuration from left to right, respectively. (C) Without $\mathcal{L}_{ms\_lpips}$, the skin color and illumination are worse than the baseline. (D) The multi-scaled encoder helps to capture styles better (see the color of eyes).

We sum up the above loss functions as our reconstruction loss $\mathcal{L}_{recon}$, which can be expressed as:

$$\mathcal{L}_{recon} = \mathcal{L}_{mse} + \lambda_1 \mathcal{L}_{ms\_lpips} + \lambda_2 \mathcal{L}_{ms\_id} + \lambda_3 \mathcal{L}_{ms\_parsing}, \quad (5)$$

where $\lambda_1 \sim \lambda_3$ are trade-off hyperparameters.

Adversarial loss. Imposing the reconstruction loss $\mathcal{L}_{recon}$ solely cannot produce realistic reconstruction results. Thus, we additionally leverage the adversarial training to help improve the final image quality, which is defined as:

$$\mathcal{L}_{adv} = \mathbb{E}[1 - \log D(\hat{I})] + \mathbb{E}[\log D(I)], \quad (6)$$

where D is initialized with the pre-trained StyleGAN discriminator. Finally, the overall loss function of our RGI can be expressed as:

$$\mathcal{L} = \mathcal{L}_{recon} + \lambda_{adv} \mathcal{L}_{adv}. \quad (7)$$

In all the experiments we set the hyperparameters $\lambda_1, \lambda_2, \lambda_3$, and λ_{adv} as 0.8, 0.1, 0.1, and 0.01, respectively.

3. More Ablation Studies

We conduct more ablation studies on other components and show the quantitative result in Tab. 1, where the configuration (C) and (D) denote multi-scale LPIPS loss and multi-scale feature encoder are discarded during training, respectively. The reconstruction performance drops a little when the multi-scale LPIPS loss is discarded. The corresponding reconstruction quality also supports this observation, which can be seen from the first row in Fig. 2 (see the skin color and illumination). We employ a similar encoder network as in [13] for our F_θ, which is used to extract multi-scale feature maps and the subsequent per-region style codes. For configuration (D), only the last level of feature maps produced by F_ϕ is used. Compared with our full model, the performance of the single-scale encoder is worse, which is consistent with the qualitative comparison shown in the second row in Fig. 2 (see the color of eyes).

4. Additional Swapping Results

We present more qualitative comparisons with state-of-the-art face swapping methods FSGAN [11], SimSwap [2], FaceShifter [10] and HifiFace [16], where the results are shown in Fig. 4 and Fig. 5. The generated faces of FSGAN are blurry. Besides, artifacts can be found in the results of SimSwap and Hififace (see the 1st and 4th row of Fig. 4, and the 6th row of Fig. 5). The swapped faces of Faceshifter and ours are visually pleasing; however, Faceshifter may focus too much on the target (see the 1st and 7th row of Fig. 4, and the 2nd and 7th row of Fig. 5). Our method is able to produce high-fidelity and high-resolution (1024^2) results, which preserve the identity information from the source image better and show a similar pose and expression as the target image. It is worth noting that only our method can keep the skin tone of the source, which is also an identity-related attribute. More realistic and high-quality face swapping results achieved by our method are shown in Fig. 6 and Fig. 7, compared with some StyleGAN-based methods (MegaFS [22], StyleFusion [6] and HiRes [17]) for reference. In each row, we display the source and target faces in the first column, and show the swapped results of each method in sequence. We can clearly see that MegaFS would produce unexpected artifacts when the source and target show different poses (see the 2nd row of Fig. 6 and 4th row of Fig. 7). While for the HiRes, we can find some distortions on the swapped face, especially for the skin and teeth regions. Though StyleFusion can generate visually satisfying results, the resulting faces show a bit of over-smoothing, and the hair looks unnatural. On the contrary, our results are much sharper and retain detailed textures better. Our approach can handle more challenging cases where the occlusion exists in the source and target faces (see the 2nd and 4th row of Fig. 6).

5. Reconstruction Visual Comparison

In Fig. 8, we display the high-fidelity reconstruction results achieved by our RGI. Here, we compare our re-

https://github.com/mindslab-ai/hififace

sults with the leading fine-grained face editing methods: SPADE [12], Mask-guided GAN [5], SEAN [21], MaskGAN [9] and SofGAN [1]. Note that SofGAN is an optimization-based method and for a fair comparison, we also add an optimization phase to our RGI (*i.e.*, RGI-Optim.). Whenever the optimization stage is applied, we fix the learning rate as $1e^{-2}$. We empirically find our RGI-Optim. only needs about 50 iterations (∼16s) to achieve satisfying results, while SofGAN [1] roughly needs about 1000 iterations (∼2 mins).

As can be seen, SEAN sometimes produces artifacts on hair. In contrast, our RGI keeps identity and texture information better, such as skin tone, eye color, and illumination. Although both SofGAN and our RGI-Optim. apply style code optimization during the reconstruction, Our RGI-Optim. is able to preserve details better (e.g., the curly degree of hair, the thickness of beard, dimples, and background), while SofGAN suffers from losing identity. Please zoom in and pay attention to the red rectangles of each example for a better understanding.

6. More Applications

6.1. Face beautification

Inspired by SEAN [21], we develop a user-interface system to perform face beautification. A screenshot is shown in Fig. 9. A control panel (Fig. 9(a)) is placed on the left side, where some necessary editing functionalities like brush, fill and undo are included. On the right side, we provide a reference image gallery (Fig. 9(b)) for users to choose from. Generally, users can perform two kinds of editing, *i.e.*, shape editing and texture editing. For shape editing, one can directly modify the facial mask, such as cutting some hair, enhancing the eyebrows, etc. For texture editing, a reference image is needed to perform style code swapping or interpolation. Users can select the interested facial region(s) in the top checkbox panel and the bottom colored button panel (Fig. 9(c)). We show the input image, mask, and the edited result in the main panel (Fig. 9(d)) from left to right. Note that the incremental editing feature is supported. That is, users can choose different reference images and edit contiguously until the result is satisfying. We also provide an interactive editing video demo, which is placed at "/interactiveEditing/systemDemo.mp4".

6.2. Hair transferring.

Thanks to the fine-grained editing capability of our RGI, we can exchange the style code of hair between the source and reference images to realize hair style transferring. Beside ours, we show the results of StyleFuison [6], RetrieveInStyle [3] and Barbershop [20] for reference. As demonstrated in Fig. 10, we can obverse that our results are visually pleasing and look realistic. The StyleFusion and RetrieveInStyle are methods based on the $\mathcal{S}$ latent space of StyleGAN. We find their results struggle to maintain the original hair shape due to the entanglement between the shape and texture of the $\mathcal{S}$ space. Our method is superior in terms of hair textures and source identity preservation. Compared with Barbershop, our method shows comparable results. However, it takes about 400s to finish the hair transferring process of the Barbershop since it is an optimization-based method. In contrast, we only need a simple forward pass that takes about 0.3s. Interestingly, other than color, the curly degree and splitting are also captured in our texture.

6.3. Controllable face swapping.

Our proposed fine-grained face swapping approach is flexible to control the amount of the face swapping by the interpolation of style codes of two faces. This concept is illustrated in Fig. 11. We first perform the shape swapping to obtain the recomposed masks of facial components (see Sec 3.1 in our main paper). With the fixed masks, we then perform the texture code interpolation of eyebrows, eyes, nose, mouth, lips, face skin, neck, and ears. The amount of face swapping is determined by an interpolation ratio λ. $\lambda = 0$ means the full face swapping, *i.e.*, the texture codes of the swapped components come from the source. $\lambda = 1$ means the texture codes of the components come from the target. As shown in Fig. 11, we can produce high-quality swapped faces with smooth texture transition from the source to the target. Please pay attention to the skin tone, beard, nose, eyebrows, and eye color.

6.4. Video face swapping.

Our E4S framework can also be applied for video face swapping. We follow STIT [14] to crop and align the source image and the target video beforehand, obtaining the source face S and an n-frame target face video $\{T_i\}_{i=1}^n$. We first reenact the source S towards the target video to obtain the driven video $\{D_i\}_{i=1}^n$. Then, we can achieve face swapping by swapping each driven and target pair $\{(D_i, T_i)\}_{i=1}^n$, as described in Sec 3.1 in our main paper. However, we find this will bring some temporal inconsistency, which is an expected result since our RGI is only trained with images. To mitigate this, we opt to fine-tune the generator of our RGI on all the driven frames, where the $\mathcal{L}_{recon}$ is leveraged as the loss function. We update the parameters 200 times for each frame, and set the learning rate to 10^{-3}. After fine-tuning, we can adopt the frame-by-frame swapping strategy and then blend with the background of the target face video to obtain a temporally consistent result. For reference, we compare our results with FaceShifter [10] and HiRes [17] in Fig. 12, please check the folder "/videoSwap/" to see the videos. We find HiRes struggles to generate a wink in the swapped video (see the frames in Fig. 12 highlighted

Figure 3. Failure cases of our E4S approach. First row: the currently used pretrained reenactment model sometimes fails to keep the similar gaze direction as the target. Second row: the illumination is also included in our per-region style code, causing inconsistency when the source and target show large lighting difference.

in red). On the other hand, FaceShifter sometimes fails to transfer the source identity (see the frames in Fig. 12 highlighted in purple). As can be observed from the image and video comparisons, our results show better visual quality and temporal consistency.

7. Limitations and Discussion

Our proposed $E4S$ framework has several limitations. First, we rely on a reenactment model to obtain a similar pose and expression as the target face. In our current implementation, we employ the open-source pre-trained model FaceVid2Vid [15]. We find it fails to keep a similar gaze direction sometimes; however, the pose and expression of the swapped face mainly depend on the output of the reenactment model. Here, we show an example in the first row of Fig. 3. Second, the overall inference time of our method is about 0.97s for swapping a source and target pair on a Tesla A100 GPU. We inspect the running time of each inside step, finding the reenactment and the swapping cost 0.51s and 0.23s, respectively. In a word, a better and more efficient reenactment model will alleviate the above two limitations. Third, illumination is another challenging problem for E4S. Specifically, we have not specially considered illumination in our framework, while the illumination is one kind of information in the proposed per-region texture. It may cause some inconsistency when the source and target show large lighting differences, with an example shown in the second row of Fig. 3. Decoupling the illumination from the texture or harmonizing the swapped result will be explored in future work.

References

[1] Anpei Chen, Ruiyang Liu, Ling Xie, Zhang Chen, Hao Su, and Jingyi Yu. Sofgan: A portrait image generator with dynamic styling. *ACM transactions on graphics*, 2021. 3, 10

[2] Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2003–2011, 2020. 2, 6, 7

[3] Min Jin Chong, Wen-Sheng Chu, Abhishek Kumar, and David Forsyth. Retrieve in style: Unsupervised facial feature transfer and retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3887–3896, 2021. 3, 11

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1

[5] Shuyang Gu, Jianmin Bao, Hao Yang, Dong Chen, Fang Wen, and Lu Yuan. Mask-guided portrait editing with conditional gans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3436–3445, 2019. 3

[6] Omer Kafri, Or Patashnik, Yuval Alaluf, and Daniel Cohen-Or. Stylefusion: A generative model for disentangling spatial segments. *arXiv preprint arXiv:2107.07437*, 2021. 2, 3, 8, 9, 11

[7] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. 1

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[9] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5549–5558, 2020. 1, 3, 10

[10] Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019. 2, 3, 6, 7, 12

[11] Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7184–7193, 2019. 2, 6, 7

[12] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 3, 10

[13] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding

in style: a stylegan encoder for image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2287–2296, 2021. 1, 2

[14] Rotem Tzaban, Ron Mokady, Rinon Gal, Amit H Bermano, and Daniel Cohen-Or. Stitch it in time: Gan-based facial editing of real videos. *arXiv preprint arXiv:2201.08361*, 2022. 3

[15] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10039–10049, 2021. 4

[16] Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. *arXiv preprint arXiv:2106.09965*, 2021. 2, 6, 7

[17] Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7642–7651, 2022. 2, 3, 8, 9, 12

[18] Xu Yao, Alasdair Newson, Yann Gousseau, and Pierre Hellier. Feature-style encoder for style-based gan inversion. *arXiv e-prints*, pages arXiv–2202, 2022. 1

[19] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 1

[20] Peihao Zhu, Rameen Abdal, John Femiani, and Peter Wonka. Barbershop: Gan-based image compositing using segmentation masks. *arXiv preprint arXiv:2106.01505*, 2021. 3, 11

[21] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 3, 10

[22] Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4834–4844, 2021. 2, 8, 9

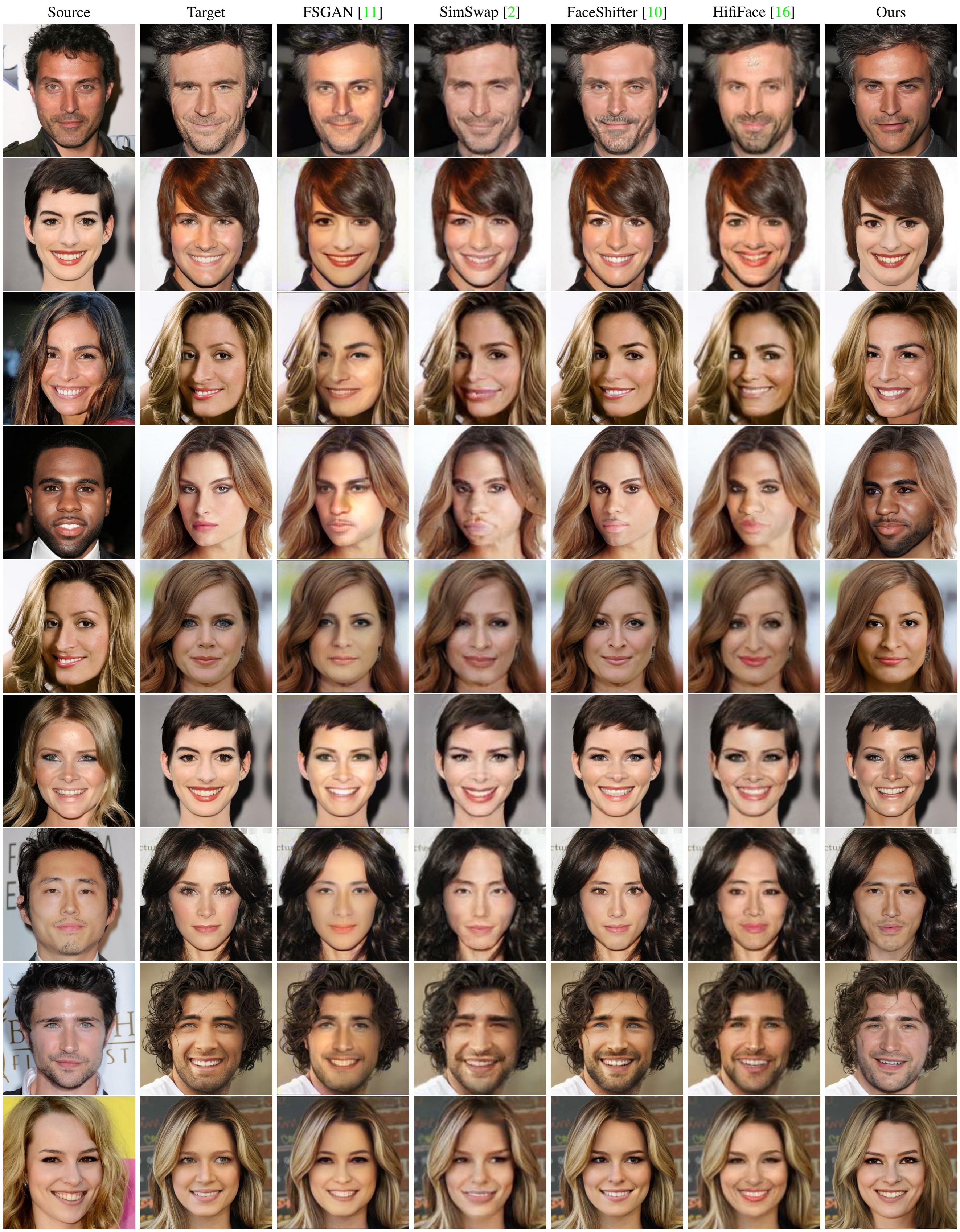

Figure 4. Additional qualitative comparisons of our results with state-of-the-art face swapping methods. Zooming-in is recommended to better observe fine details in all figures.

Figure 5. Additional qualitative comparisons of our results with state-of-the-art face swapping methods. Zooming-in is recommended to better observe fine details in all figures.

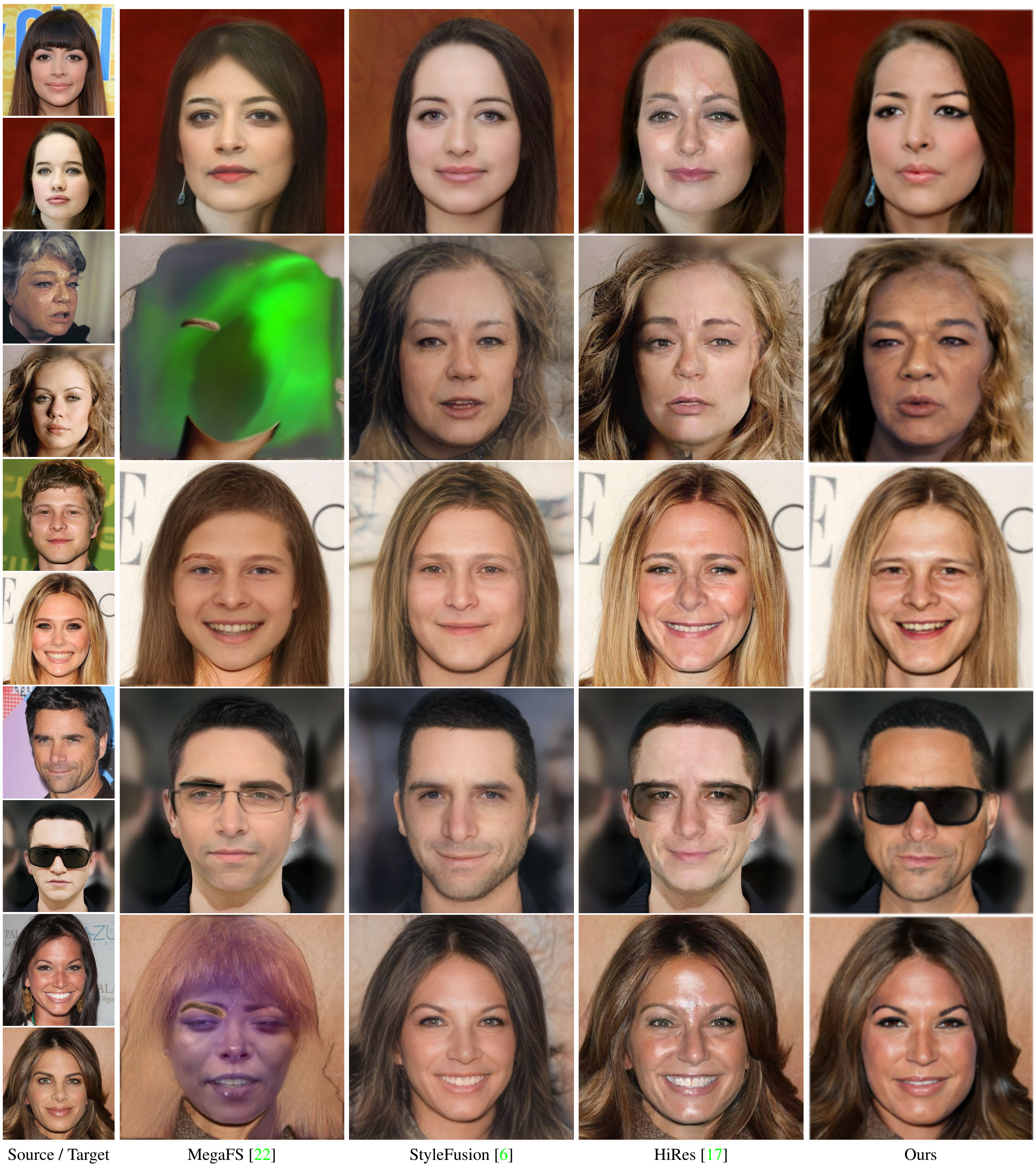

Figure 6. Compared with the existing StyleGAN-based face swapping approaches (MegaFS [22], StyleFusion [6] and HiRes [17]), our proposed method can achieve high-fidelity results that show better identity keeping from the source, while keeping the similar pose and expression as the target. Note that skin color preservation and proper occlusion handling are our advantages over others. All the facial images are at 1024×1024.

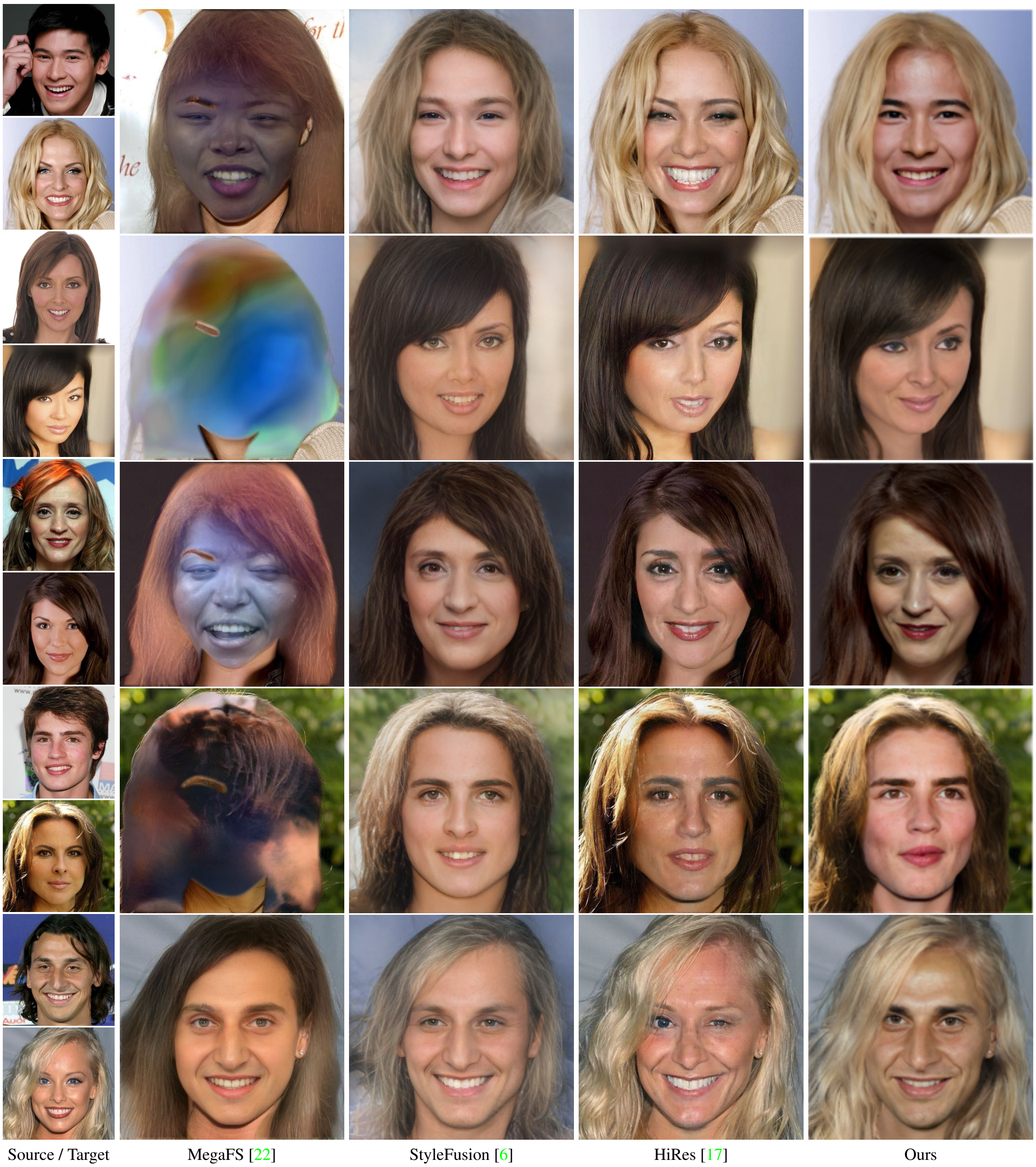

Figure 7. Compared with the existing StyleGAN-based face swapping approaches (MegaFS [22], StyleFusion [6] and HiRes [17]), our proposed method can achieve high-fidelity results that show better identity keeping from the source, while keeping the similar pose and expression as the target. Note that skin color preservation and proper occlusion handling are our advantages over others. All the facial images are in 1024×1024.

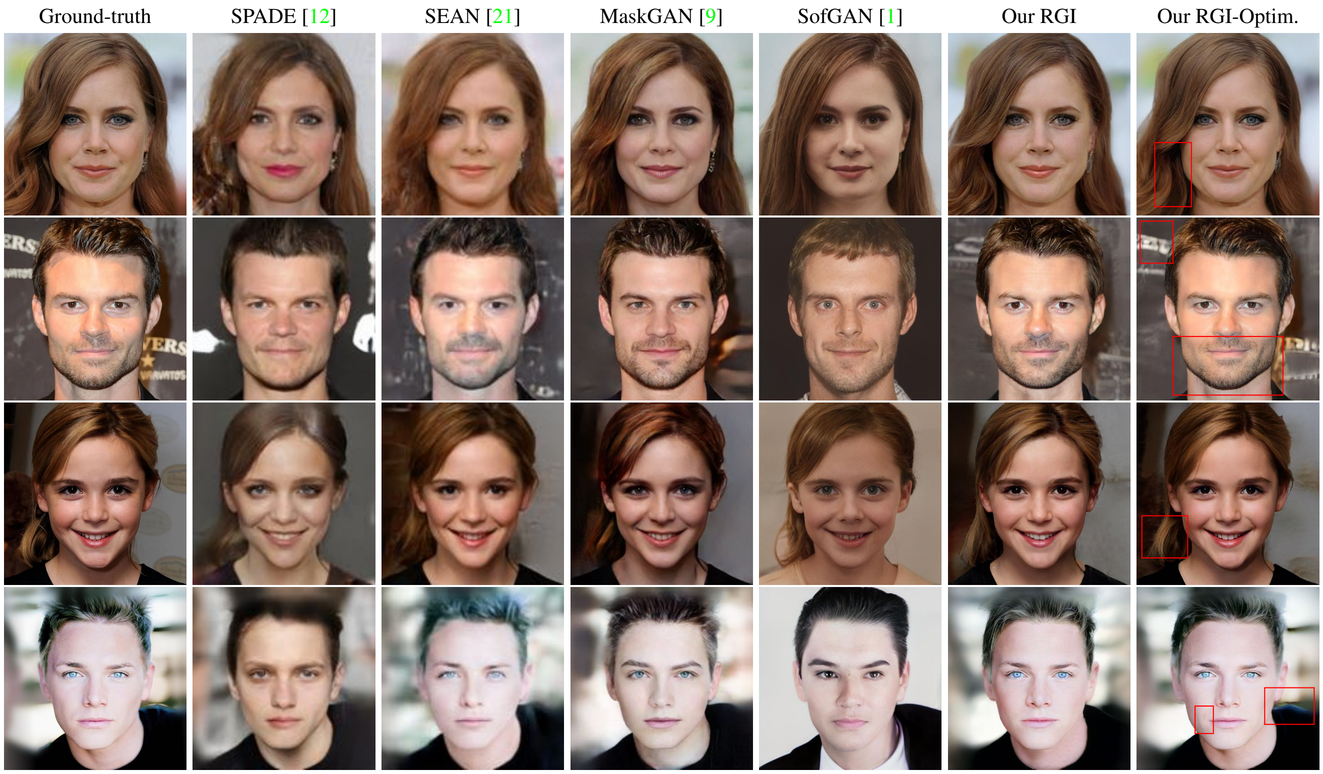

Figure 8. Reconstruction comparisons of our results with state-of-the-art fine-grained face editing methods. Our method can achieve high-fidelity reconstructed results. We also show our results with style code optimization where the details (*e.g.*, curly degree of hair, thickness of the beard, dimple, and background) are preserved better.

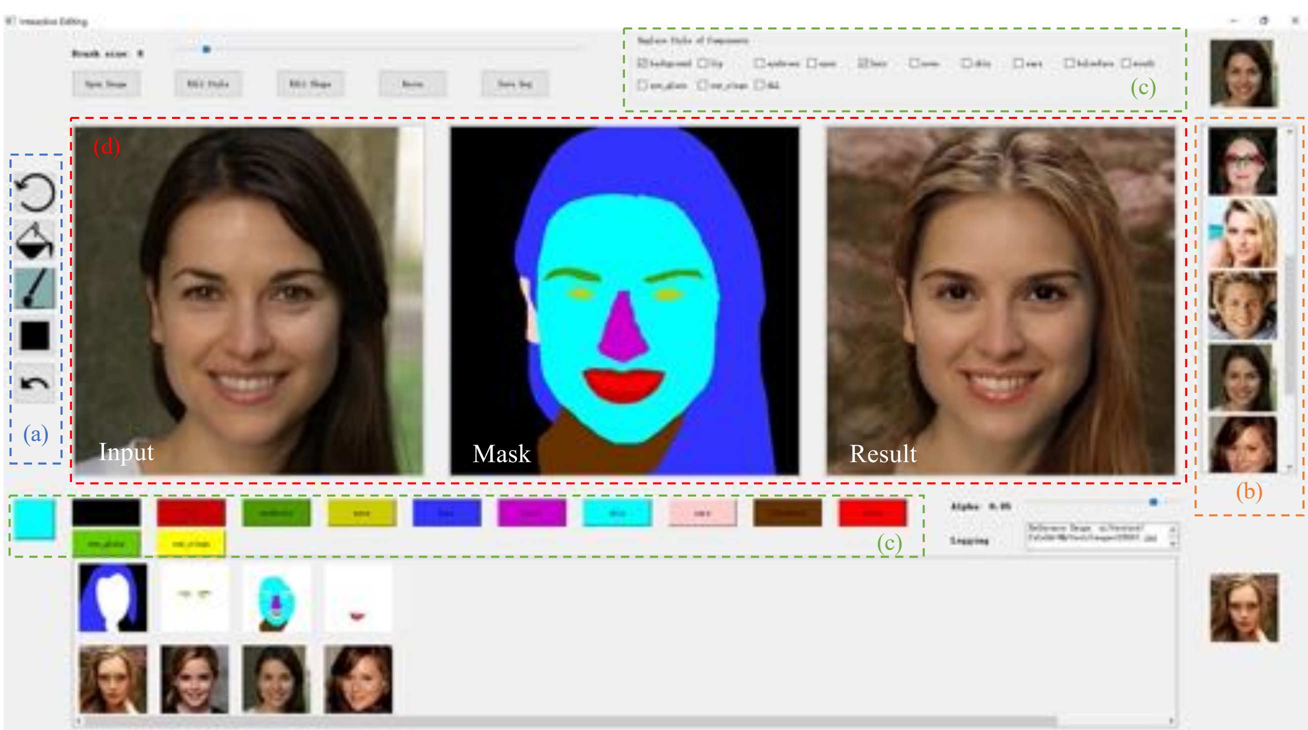

Figure 9. A screenshot of our interactive editing system.

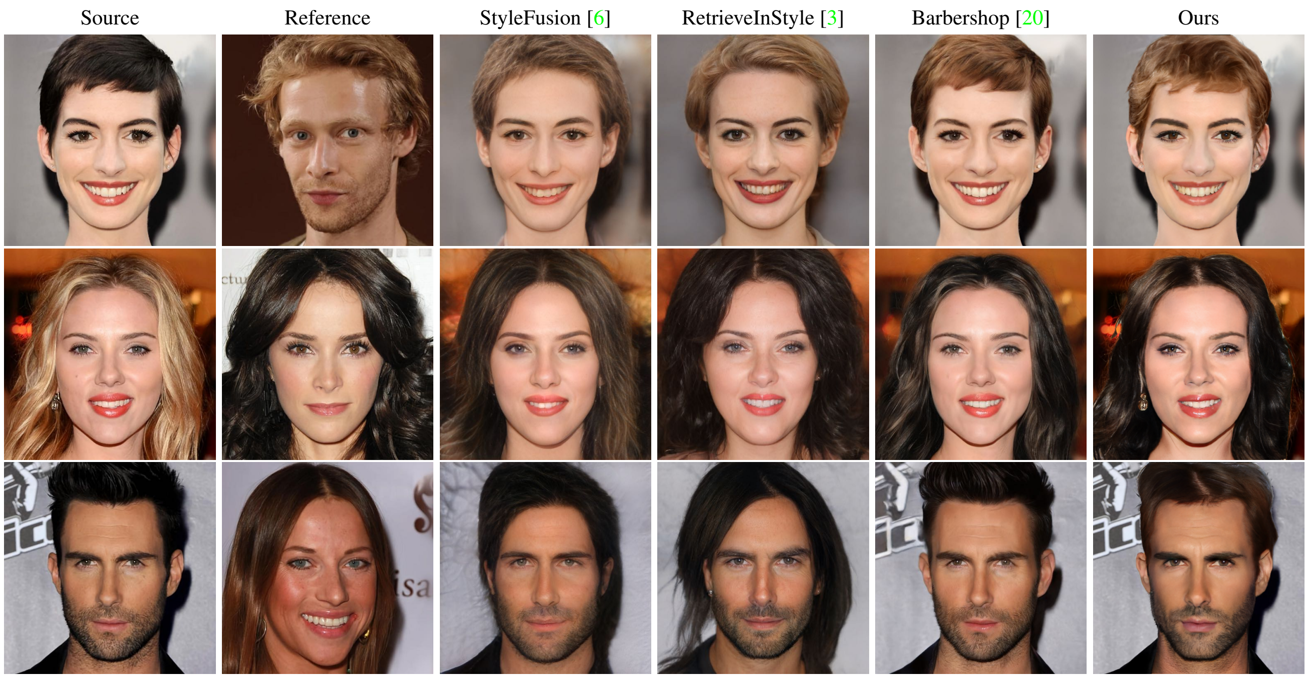

Figure 10. Hair style transferring examples achieved by our RGI. The results are in high-quality and look realistic.

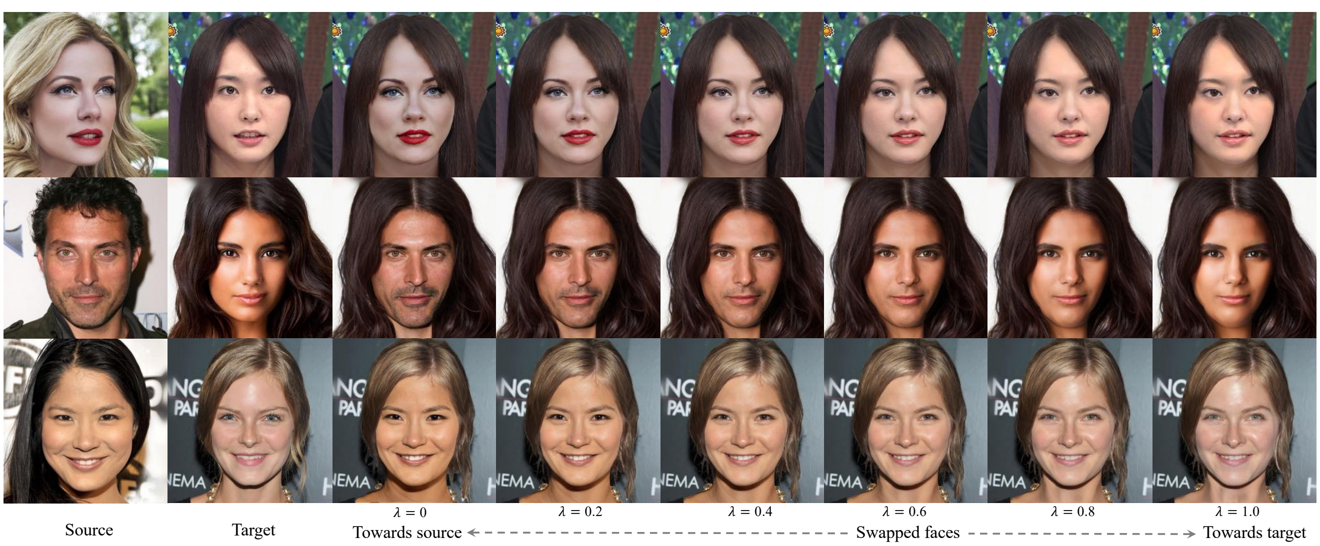

Figure 11. Controllable face swapping examples achieved by our E4S. We can smoothly generate some transitional faces between target and swapped face via style codes interpolation. The λ under each image denotes the interpolation coefficient.

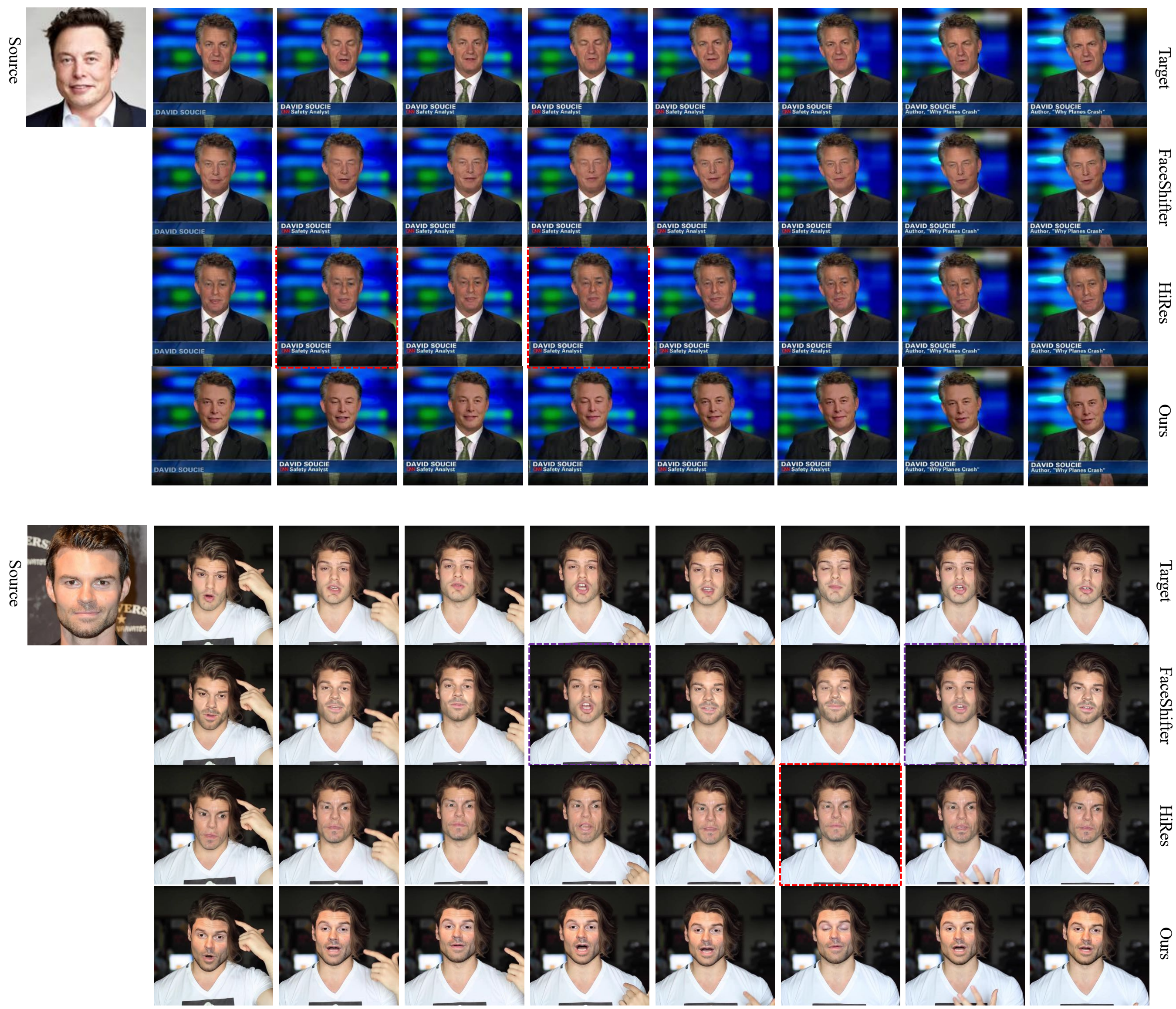

Figure 12. Video face swapping comparisons of our results with FaceShifter [10] and HiRes [17]. Our method shows the better capability of source identity transferring and target attribute preservation (*e.g*., pose, expression, wink). The visual quality and temporal consistency of our results also surpass previous methods.