Style Transformer for Image Inversion and Editing

Xueqi Hu[1], Qiusheng Huang[1], Zhengyi Shi[1], Siyuan Li[1], Changxin Gao[3], Li Sun[1,2]*, Qingli Li[1]
[1]Shanghai Key Laboratory of Multidimensional Information Processing,
[2]Key Laboratory of Advanced Theory and Application in Statistics and Data Science,
East China Normal University, Shanghai, China
[3]Huazhong University of Science and Technology, Wuhan, China

A. Limitations

We now discuss limitations, which we have already realized, for our work. First, for the inversion task, although our proposed method achieves improved reconstruction quality, there are still some differences between the input and reconstructed images, especially for the out-of-domain input. We think it is mainly caused by the finite discriminative ability of W^+ space. As is described in [8], the distortion can be significantly reduced by adding more information from the source. Moreover, since we apply the multi-head attention, the training speed is slower due to the complex matrix multiplication. Second, for the reference-based editing task, we adopt a transformer-based module in the latent space, resulting in less diversity for some attributes compared with direct editing on the images, in which the mode seeking loss [4] can encourage the diversity in the pixel domain. But our method is lightweight and more flexible.

B. Ablations and Analysis

We further validate the benefit of transformer by comparing among pSp [5], our full model with both self- and cross-attention and ours w/o self-attention in Tab. 1. [5] maps image features to $w+$ by individual mapping networks, though $w+$ obtain the image features directly and completely, the relation between each w is not tightly enough. In our model, cross-attention is necessary to update queries by fusing image features, and self-attention is also important in constructing the potential relation between queries.

C. Label-based Editing Methods

We propose first- and second-order label-based editing methods in the main text. To give a detailed explanation, we provide the pseudo codes in PyTorch style. Algorithm 1 and Algorithm 2 illustrate the first- and second-order methods, respectively. Moreover, we measure the disentanglement of five attributes by Re-scoring [6] in Fig. 1. The top row lists edited attributes, and the scores are the classification logits

Method	MSE↓	LPIPS↓	Params(M)↓	FLOPs(G)↓	Time(s)↓
pSp	0.0373	0.1693	267.3	72.55	0.0668
Ours w/o self	0.0369	0.1716	**37.3**	**36.31**	**0.0429**
Ours full	**0.0363**	**0.1665**	40.6	36.37	0.0436

Table 1. Ablations of transformer structure. *Time* means the inference time of an iteration. The best results are indicated in **Bold**.

Method	Quality(%)			Disentanglement(%)		
	BA	GE	GO	BA	GE	GO
InterFaceGAN	15.00	7.50	9.17	11.67	1.67	8.33
StyleSpace	10.83	10.00	13.33	18.33	15.00	10.83
Ours-1	25.83	39.17	31.67	**35.83**	34.17	30.00
Ours-2	**48.33**	**43.33**	**47.50**	34.17	**49.17**	**49.17**

Table 2. User study of label-based editing compared with [6], [9]. **BA**, **GE** and **GO** represent 'Bangs', 'Gender' and 'Goatee' attributes.

	Smile	Bangs	Gender	Glass	Age
Smile	0.45	-0.02	0.00	-0.05	-0.03
Bangs	0.00	0.52	0.00	0.00	-0.01
Gender	-0.03	-0.03	0.54	0.02	0.03
Glass	0.00	0.00	0.01	0.52	0.01
Age	-0.05	-0.04	0.06	0.13	0.45

(a) Ours-1

	Smile	Bangs	Gender	Glass	Age
Smile	0.45	-0.01	0.00	-0.03	-0.02
Bangs	0.00	0.52	0.00	0.00	0.00
Gender	-0.01	-0.02	0.55	0.02	0.03
Glass	0.00	0.00	0.01	0.52	0.01
Age	-0.02	-0.03	0.04	0.12	0.45

(b) Ours-2

Figure 1. Re-scoring results of label-based editing on five attributes, Ours-1 and Ours-2 represent our first- and second-order methods, respectively.

changes between original and edited images. Considering human judgements, we further conduct a user study. We ask 60 volunteers to evaluate the methods in two aspects: image quality and disentanglement. Results are shown in Tab. 2.

D. Training Details

We adopt a pretrained StyleGAN2 [1] generator in our experiments, in which the synthesis network is fixed and

the mapping network (MLP) is trained. In the multi-head attention of the transformer block, the number of heads is set to 4, and the dimension of each head is 512. For inversion task, the Ranger optimizer is used in training, which is a combination of Rectified Adam [3] with the Lookahead technique [10]. We train the model for 6×10^5 iterations with a batch size of 8, the learning rate is set to 1×10^{-4}. For the reference-based editing task, we use the Adam [2] optimizer to train the model for 1×10^4 iterations with a batch size of 8, the learning rate is set to 1×10^{-3}. All experiments are implemented on 2 NVIDIA RTX 2080Ti GPUs.

E. More Results

In this section, we provide more results of inversion, label-based editing and reference-based editing in Fig. 2, Fig. 3, Fig. 4.

Algorithm 1 First-order Label-based Editing

```
# w: input latent code (18, 512), C: latent
    classifier, y_t: target label

predicted = C(w)
loss = torch.nn.BCELoss(predicted, y_t)
loss.backward()
direct = w.grad
direct = direct / torch.norm(direct, dim=1)
w_edit = w - alpha * direct # alpha is a
    scaling factor.
```

Algorithm 2 Second-order Label-based Editing

```
# w: input latent code (18, 512), C: latent
    classifier, y_t: target label

r_d = torch.randn(18, 512)
r_0 = torch.zeros(18, 512)
w_d = w + kasi * r_d # kasi is a small number
    , we set it to 10e-4.
w_0 = w + r_0
predicted_d = C(w_d)
loss = torch.nn.BCELoss(predicted_d, y_t)
loss.backward()
direct_d = r_d.grad

C.zero_grad()
predicted_0 = C(w_0)
loss = torch.nn.BCELoss(predicted_0, y_t)
loss.backward
direct_0 = r_0.grad
direct = direct_d - direct_0
direct = direct / torch.norm(direct, dim=1)
w_edit = w - alpha * direct # alpha is a
    scaling factor.
```

References

[1] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 1

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[3] Liyuan Liu, Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Jiawei Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019. 2

[4] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1429–1437, 2019. 1

[5] 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, 3

[6] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 2020. 1

[7] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *ACM Transactions on Graphics (TOG)*, 40(4):1–14, 2021. 3

[8] Tengfei Wang, Yong Zhang, Yanbo Fan, Jue Wang, and Qifeng Chen. High-fidelity gan inversion for image attribute editing. *arxiv:2109.06590*, 2021. 1

[9] Zongze Wu, Dani Lischinski, and Eli Shechtman. Stylespace analysis: Disentangled controls for stylegan image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12863–12872, 2021. 1

[10] Michael R Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back. *arXiv preprint arXiv:1907.08610*, 2019. 2

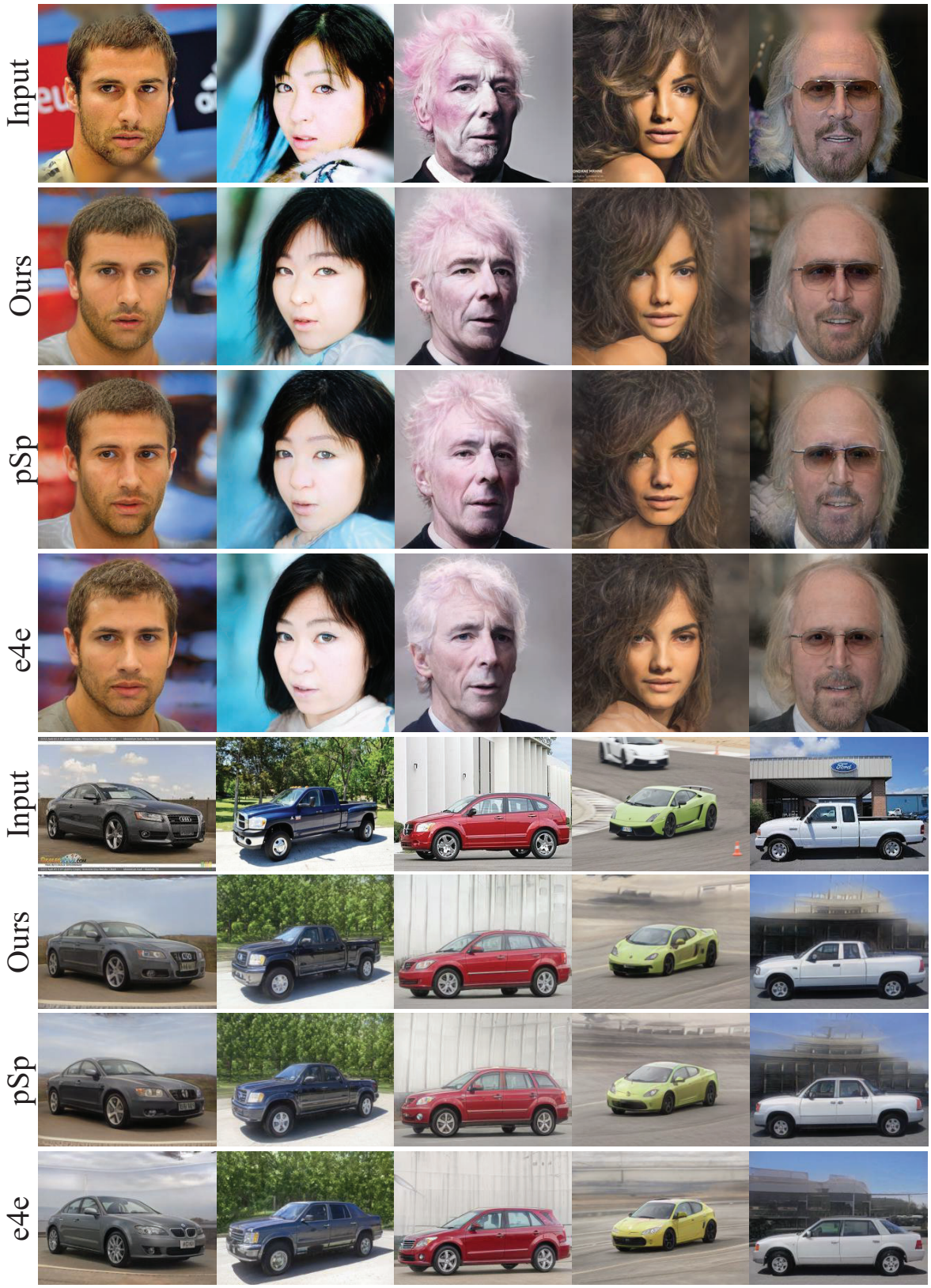

Figure 2. More results of inversion compared with [5] and [7].

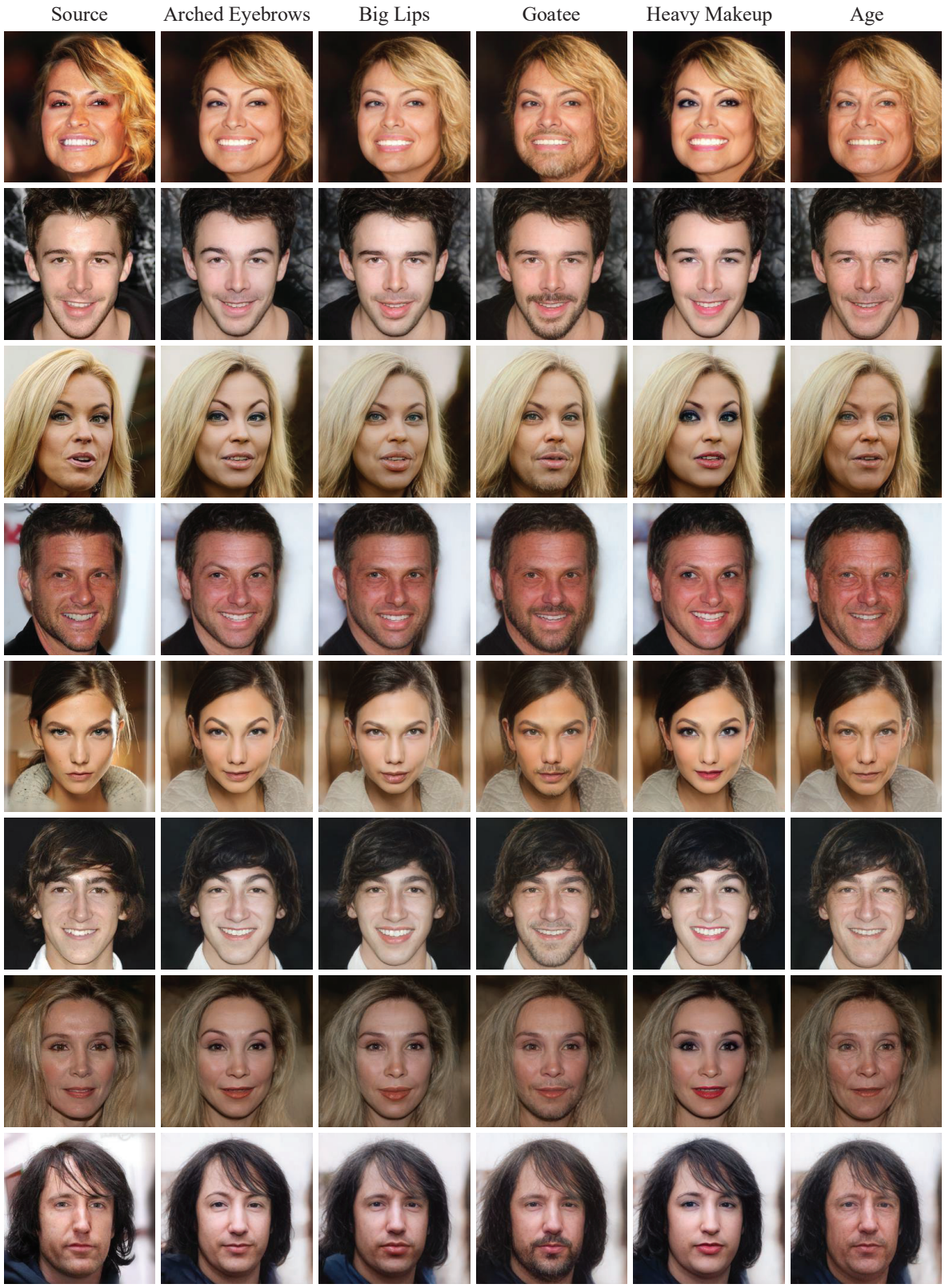

Figure 3. More results of label-based editing on five attributes.

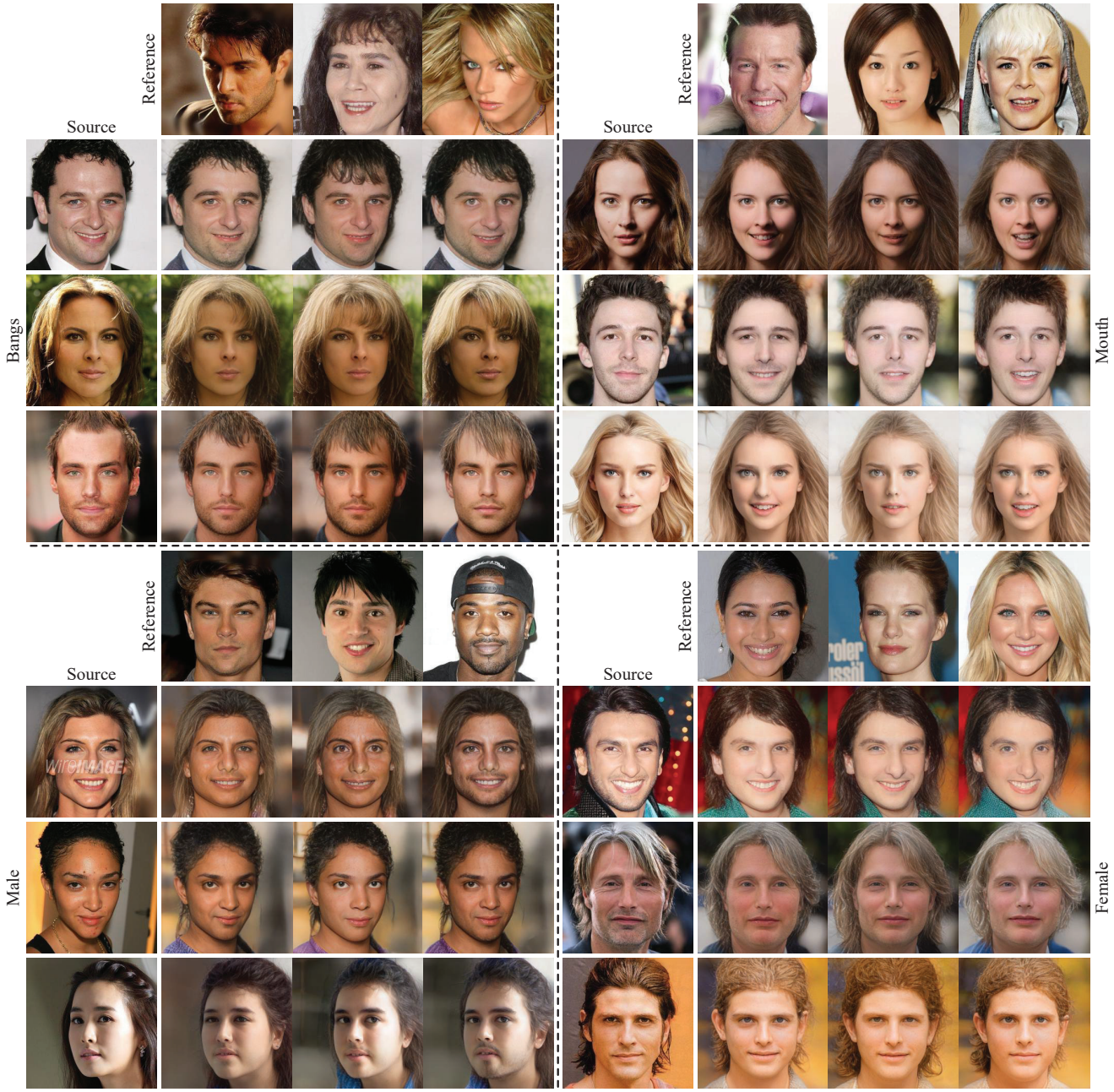

Figure 4. More results of reference-based editing on three attributes. The edited images take the style of *Bangs*, *Mouth* and *Gender* from different reference images.