SelfSwapper: Self-Supervised Face Swapping via Shape Agnostic Masked AutoEncoder - Supplementary Materials

Jaeseong Lee*, Junha Hyung*, Sohyun Jung, and Jaegul Choo

KAIST AI
{wintermad1245, sharpeeee, jsh0212, jchoo}@kaist.ac.kr
* indicates equal contributions.

1 CelebA-HQ Comparison with AFS, MegaFS, and ReliableSwap

In the Experiment section of our manuscript, we conduct a comprehensive comparison with additional baselines, including MegaFS [21], AFS [16], and ReliableSwap [19]. These baselines, although not widely accepted within the academic community, provide a valuable contrast in terms of their performance and capabilities. The quantitative results are presented in Table S1, extending the comparison beyond Table 1 in the main manuscript. Additionally, we introduce an overall score by standardized for each metric and averaged following the methodology of SmoothSwap [10], denoted *Overall*. Among the nine baselines, our approach achieves the highest overall score, indicating a substantial performance gap.

A closer examination of MegaFS and AFS, depicted in the 3rd and 4th columns of Fig. S1 and S2, reveals limitations in handling target attributes such as illumination and skin color. These methods rely on an encoder-based StyleGAN [9] inversion with a empirical layer-specific decomposition design. Moreover, they necessitate post-processing to address issues like hair and background, as StyleGAN inversion-based techniques often freeze and map latent vectors, leading to insufficient attribute preservation.

ReliableSwap, illustrated in the 5th columns of Fig. S1 and S2, exhibits unrealistic results with noticeable artifacts. While attempting to introduce reliable supervision through pseudo-triplets for face swapping training, ReliableSwap depends on the head reenactment model [17] to construct the synthetic dataset. This reliance on the head reenactment model introduces limitations, constraining the capacity of the trained model.

2 More Qualitative Comparison among Target-Oriented Baselines

We have posted eight sets of additional qualitative comparison among target-oriented baselines in Fig. S3 and S4.

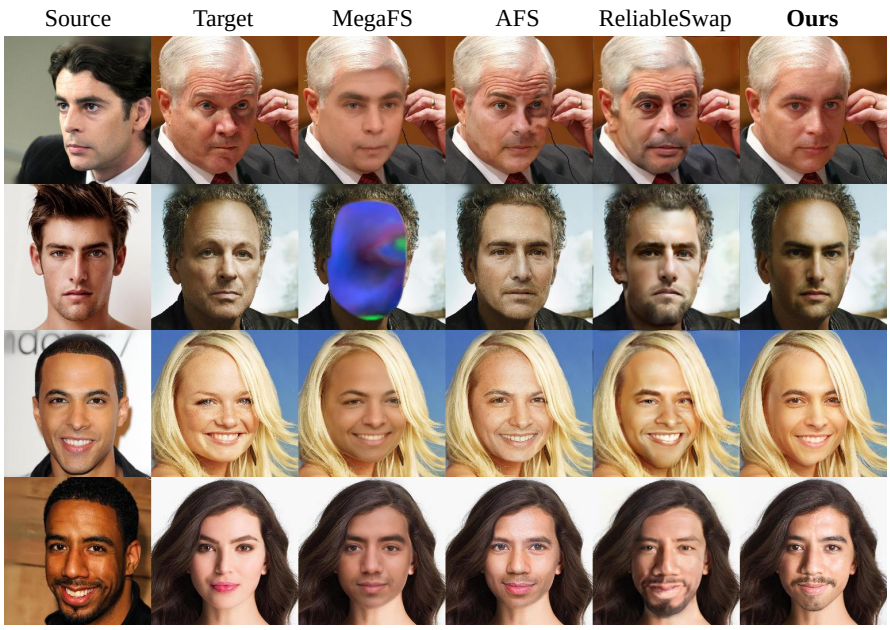

Fig. S1: Qualitative comparison of additional baselines (MegaFS [21], AFS [16], and ReliableSwap [19]).

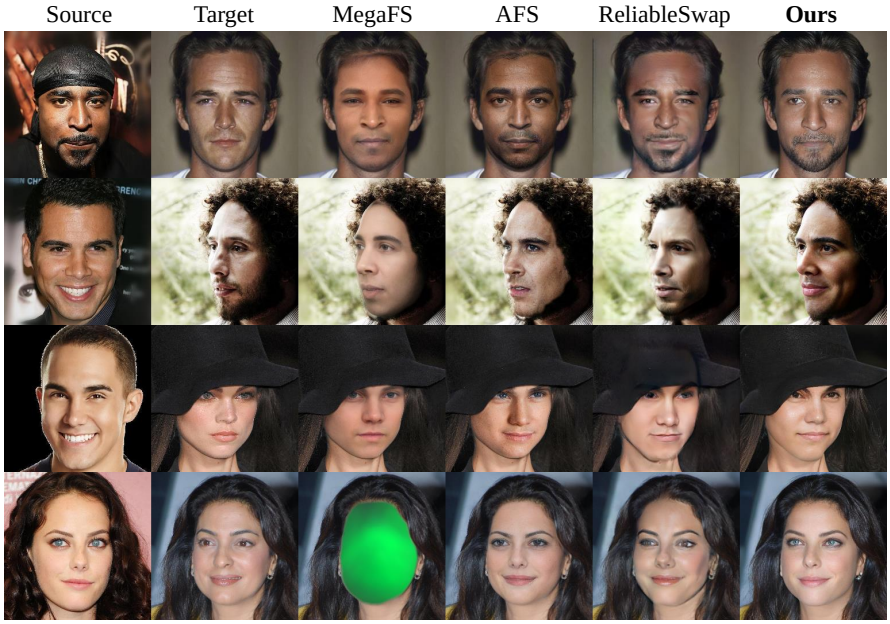

Fig. S2: Qualitative comparison of additional baselines (MegaFS [21], AFS [16], and ReliableSwap [19]).

Table S1: Quantitative comparison augmented with MegaFS [21], AFS [16], and ReliableSwap [19]. Bold text highlights the best scores. The Overall score indicates average standardized score of each method.

Categories	Methods	ID. Sim.↑	ID. Cons.↑	Shape ↓	Expression↓	Head Pose↓	FID↓	Overall↓
Target-Oriented	SimSwap [1]	0.525	0.543	0.128	0.204	0.014	26.77	-0.725
	InfoSwap [4]	0.527	0.583	0.126	0.233	0.019	32.21	-0.469
	FSLSD [18]	0.330	0.345	0.129	0.207	0.025	39.71	0.333
	BlendFace [15]	0.440	0.510	0.136	**0.189**	**0.013**	23.11	-0.593
	DiffSwap [20]	0.347	0.361	0.156	0.224	0.028	59.98	0.954
	MegaFS [1]	0.301	0.282	0.152	0.272	0.043	63.43	1.749
	AFS [4]	0.448	0.508	0.121	0.231	0.022	26.78	-0.277
	ReliableSwap [18]	0.492	0.532	0.126	0.238	0.022	44.76	-0.089
Source-Oriented	FSGAN [13]	0.338	0.403	0.164	0.193	0.016	42.56	0.354
	E4S [12]	0.501	0.588	0.118	0.262	0.028	52.90	0.089
SAMAE	**Ours**	**0.578**	**0.628**	**0.108**	0.190	**0.013**	**21.22**	**-1.326**

3 User Study

For the human evaluation about criteria in the Background section in the manuscript (C1, C2, and C3), we have conducted the user studies for three aspects, magnitude of identity reflection, attribute preservation, and naturalness. We nominate the top-3 overall scored baselines from Table S1 for the simplicity. We requested data from 33 subjects for these clauses. As illustrated in the table, our results exhibited the highest scores, with a significant margin over other baseline measures. We have also demonstrated that our method attains state-of-the-art quality in terms of human perceptual evaluation.

Table S2: User study with SimSwap [1], InfoSwap [4], and BlendFace [15] nominated as top-3 Overall scored from Table S1.

Methods	Identity ↑	Attributes ↑	Naturalness ↑
SimSwap [1]	2.50	3.34	2.65
InfoSwap [4]	2.93	3.29	2.95
BlendFace [15]	2.71	2.95	3.04
Ours	**4.18**	**3.81**	**3.97**

4 Architecture Details

We utilize ADM [3] U-Net architecture as the generator G. The model comprises a sequence of residual layers and downsampling convolutions, followed by another sequence of residual layers featuring upsampling convolutions, with skip connections linking layers that share identical spatial dimensions. For the E_{skin} architecture, we use a pretrained ResNet18 [6] and further train the model together with the generator. As for the discriminator, we directly adopt the architecture from the StyleGAN2 [9] discriminator.

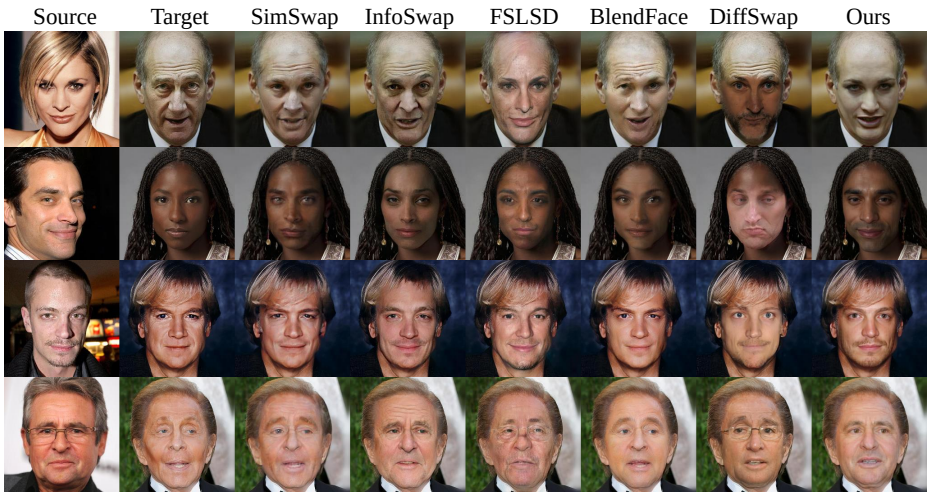

Fig. S3: Additional qualitative comparison of target-oriented baselines.

Fig. S4: Additional qualitative comparison of target-oriented baselines.

5 Training Objectives

The predicted image $\hat{I}$ is generated as described in the main paper: $\hat{I} = G(I_{ras}, I_p, c_{id}, c_{skin})$. We devise reconstruction loss between $\hat{I}$ and I with $L1$ loss and perceptual loss [7, 8],

$$\mathcal{L}_{REC} = \mathcal{L}_{L1} + \mathcal{L}_{VGG}(\hat{I}, I), \tag{1}$$

where $\mathcal{L}_{L1} = \mathbb{E}[||I - \hat{I}||_1]$. We also utilize identity loss [2] to maximize identity similarity:

$$\mathcal{L}_{ID} = 1 - \text{cossim}(E_{FR}(\hat{I}), E_{FR}(I)), \tag{2}$$

where cossim refers to cosine similarity. For improved realism of the generated images, we employ a non-saturating GAN loss [5, 9]:

$$\mathcal{L}_{GAN}(G, E_{skin}, D) = \min_{G, E_{skin}} \max_{D} \mathbb{E}[f(D(-\hat{I}))] + \mathbb{E}[f(D(I))], \tag{3}$$

where $f(u) = -\log(1 + \exp(-u))$. Additionally, for the eye-gaze controllability of the model, the rendered iris keypoints image from the real image I is concatenated with I itself, and this pair is regarded as the real sample. Other cases including a real image concatenated with perturbed keypoints image or keypoints image concatenated with the generated images are regarded as fake samples. The total loss function is composed as a weighted sum of these loss functions:

$$\mathcal{L} = \lambda_{REC}\mathcal{L}_{REC} + \lambda_{ID}\mathcal{L}_{ID} + \lambda_{GAN}\mathcal{L}_{GAN}. \tag{4}$$

For all experiments in our paper, we set hyperparameters $\lambda_{REC} = \lambda_{ID} = \lambda_{GAN} = 1$.

Table S3: Ablation study on dimensionality of c_{skin}.

dim. of c_{skin}	ID. Sim. ↑
8	0.580
32	0.580
64 (Ours)	0.578
128	0.568
512	0.566

6 c_{skin} dimensional ablation

Determining the dimensionality of c_{skin} is empirical, but thorough analysis establishes that a 64-dimensional representation is most suitable for our model, as depicted in Table S3. Increasing the dimensionality of c_{skin} beyond 64 may lead to potential issues, such as the leakage of target identity through the portrayal of wrinkles or dimples in the target's skin, while compromising identity preservation.

7 Disentangled Control

To validate the effect of c_{skin}, we modified the skin condition image while maintaining the same source and target images. As depicted in Fig. S5, we substituted the skin condition image with different images or altered the original target image's brightness. The last row of Fig. S5 displays results with varied skin color. This experiment indicates that our approach effectively separates skin color without affecting other facial features.

Additionally, we evaluated the pose and expression adaptability of our model using various target frames from a video. The top row of Fig. S6 illustrates a source image alongside different target frames with varying poses and expressions. The bottom row presents the swapped results, where the images successfully emulate the target's poses and expressions while preserving the source identity. This demonstrates our model's capability to disentangle the poses and expressions from other facial attributes.

8 Time-Space Complexity Comparision

Our model has comparable number of trainable parameters and Multiply–Accumulate (MAC) operations (Tab. S4). We also tested the robustness of our method by reducing the number layers and channel sizes (*Ours Small*). Even with decreased parameter, the performance remains competitive to the original model.

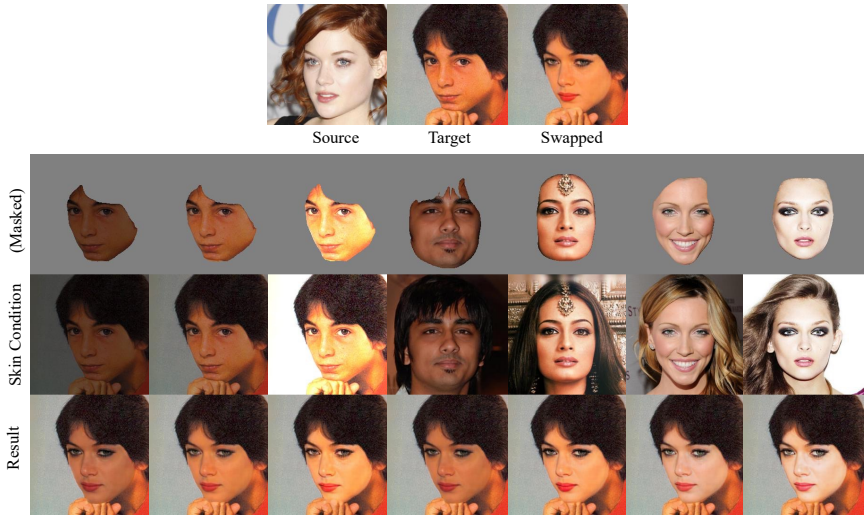

Fig. S5: Skin condition image is replaced with other images or with the original target image with different brightness ((Masked), Skin Condition). The last row shows the results.

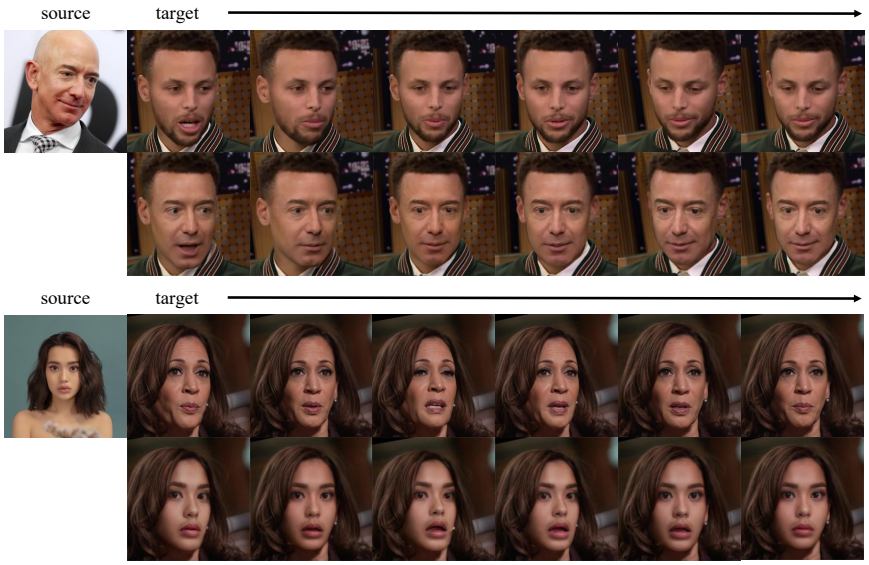

Fig. S6: Controlling pose and expression with different target frames from a video.

Table S4: Complexity comparison with ID. Sim. and Expression (Exp.) performance.

Methods	# trainable parameters ↓	MACs ↓	ID. Sim. ↑	Exp. ↓
SimSwap	55M	**193B**	0.525	0.204
InfoSwap	203M	240B	0.527	0.233
FSLSD	163M	1.9T	0.330	0.207
BlendFace	420M	359B	0.440	**0.189**
DiffSwap	169M	22T	0.347	0.193
FSGAN	200M	847B	0.338	0.193
E4S	162M	44T	0.501	0.262
Ours-ab.	144M	559B	0.477	0.237
Ours Small	**35M**	369B	0.566	0.193
Ours	144M	559B	**0.578**	0.190

9 Quantitative evaluation of Ours-ab.

As described in the main manuscript, we demonstrate the effectiveness of self-supervised training of SAMAE (Ours) against Ours-ab. in Tab. S4.

10 Video

Although our model was not trained with a video dataset, it can still generate convincing swapped videos. Please refer to the attached video and image files for the results.

11 Evaluation on FF++ benchmark

Given the absence of official code, we perform a qualitative and quantitative comparison with FaceShifter [11] based on the provided results in FaceForensics++ (FF++) [14]. The qualitative assessment is illustrated in Fig.S7, while the quantitative results are summarized in Table S5. Notably, our approach attains superior scores across all metrics with a substantial margin compared to FaceShifter on other evaluation protocols.

Fig. S7: Qualitative comparison with FaceShifter on FF++.

Table S5: Quantitative comparision with FaceShifter on FF++.

Methods	ID. Sim. ↑	Shape. ↓	Expression ↓	Pose ↓
FaceShifter	0.541	0.122	0.213	**0.019**
Ours	**0.646**	**0.108**	**0.189**	**0.019**

12 Self-Supervised (SAMAE) vs. Target-Oriented Regime (Seesaw Game)

In this subsection, we will discuss the pros and cons of SAMAE and the seesaw training regime. As mentioned in the manuscript, target-oriented training's reconstruction losses, such as L1 and Perceptual [7], ensure comprehensive preservation of the target image's attributes, including skin color and even extreme illumination. However, this reconstruction loss acts as a double-edged sword by globally imposing these attributes on the swapped images, making it difficult to manually exclude the target's shape leakage. Conversely, our self-supervised training regime excels in preventing the target's shape leakage. However, this regime also has the drawback of being unable to guarantee preservation under harsh illumination conditions, as the c_{skin} is highly compressed vectorized information. This presents an ill-posed problem of decomposing the albedo and light environment from a single field of view in a single image. In future research, we believe that combining single-shot portrait relighting techniques with SelfSwapper would be a promising direction.

13 Code Release

We have included the implementation of our work in the attached .zip file, and we intend to release the refined code publicly in the near future.

References

1. Chen, R., Chen, X., Ni, B., Ge, Y.: Simswap: An efficient framework for high fidelity face swapping. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2003–2011 (2020)
2. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4690–4699 (2019)
3. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
4. Gao, G., Huang, H., Fu, C., Li, Z., He, R.: Information bottleneck disentanglement for identity swapping. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 3404–3413 (2021)
5. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
7. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)
8. Johnson, J., Alahi, A., Fei-Fei, L.: Perceptual losses for real-time style transfer and super-resolution. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14. pp. 694–711. Springer (2016)
9. Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., Aila, T.: Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8110–8119 (2020)
10. Kim, J., Lee, J., Zhang, B.T.: Smooth-swap: a simple enhancement for face-swapping with smoothness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10779–10788 (2022)
11. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. arXiv preprint arXiv:1912.13457 (2019)
12. Liu, Z., Li, M., Zhang, Y., Wang, C., Zhang, Q., Wang, J., Nie, Y.: Fine-grained face swapping via regional gan inversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8578–8587 (2023)
13. Nirkin, Y., Keller, Y., Hassner, T.: Fsgan: Subject agnostic face swapping and reenactment. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 7184–7193 (2019)
14. Rossler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: Faceforensics++: Learning to detect manipulated facial images. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 1–11 (2019)
15. Shiohara, K., Yang, X., Taketomi, T.: Blendface: Re-designing identity encoders for face-swapping. arXiv preprint arXiv:2307.10854 (2023)
16. Vu, T., Do, K., Nguyen, K., Than, K.: Face swapping as a simple arithmetic operation. arXiv preprint arXiv:2211.10812 (2022)
17. Wang, Y., Yang, D., Bremond, F., Dantcheva, A.: Latent image animator: Learning to animate images via latent space navigation. arXiv preprint arXiv:2203.09043 (2022)

18. Xu, Y., Deng, B., Wang, J., Jing, Y., Pan, J., He, S.: High-resolution face swapping via latent semantics disentanglement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7642–7651 (2022)
19. Yuan, G., Li, M., Zhang, Y., Zheng, H.: Reliableswap: Boosting general face swapping via reliable supervision. arXiv preprint arXiv:2306.05356 (2023)
20. Zhao, W., Rao, Y., Shi, W., Liu, Z., Zhou, J., Lu, J.: Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8568–8577 (2023)
21. Zhu, Y., Li, Q., Wang, J., Xu, C.Z., Sun, Z.: One shot face swapping on megapixels. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4834–4844 (2021)