StyleSwap: Style-Based Generator Empowers Robust Face Swapping -Supplementary Materials-

Zhiliang Xu[1]*, Hang Zhou[1]*✉, Zhibin Hong[1], Ziwei Liu[2], Jiaming Liu[1]
Zhizhi Guo[1], Junyu Han[1], Jingtuo Liu[1], Errui Ding[1], and Jingdong Wang[1]

[1] Department of Computer Vision Technology (VIS), Baidu Inc.
[2] S-Lab, Nanyang Technological University
{xuzhiliang,zhouhang09,liujiaming03,dingerrui,wangjingdong}@baidu.com
ziwei.liu@ntu.edu.sg

1 Details of the Swapping-Driven ID Inversion

Strategy Revisit. Here we present more details of the *Swapping-Driven ID Inversion* strategy. For clearer representation, we re-illustrate the pipeline in Fig. 1. Inspired by the process of StyleGAN inversion, this strategy optimizes the features $\mathbf{W}_s^+ = \{\mathbf{w}_{s(1)}, \dots, \mathbf{w}_{s(2L)}\}$ in a total N iterations.

At the nth iteration of the optimization, we denote the source identity-related $\mathcal{W}^+$ space feature as $\mathbf{W}_s^{+\{n\}}$ and the desired output as $\mathbf{W}_s^{+\{n+1\}}$. $\mathbf{W}_s^{+\{1\}}$ is initialized as $\{\mathbf{w}_s, \dots, \mathbf{w}_s\}$. We randomly sample any image $I^{\{n\}}$, and firstly generate an intermediate frame $I_g^{r\to s} = \mathrm{G}(\mathbf{F}_{att}^s, f_{id}^r)$. Then it is taken as the target frame to generate the cycled-back image

$$I_g^{s\to r\to s} = \mathrm{G}(\mathrm{E}_{att}(I_g^{r\to s}), \mathbf{W}_s^{+\{n\}}) \tag{1}$$

The optimization is conducted using the identity loss $\mathcal{L}_{id}$ and the reconstruction loss $\mathcal{L}_{rec}$. We recap the two losses here. Given any source image I_s and our generated image I_g, the identity loss between the two images are:

$$\mathcal{L}_{id}(I_g, I_s) = 1 - D_{\cos}(f_{id}^s, f_{id}^g)), \tag{2}$$

where $D_{\cos}(f_a, f_b) = \frac{f_a \cdot f_b}{\|f_a\|_2 \|f_b\|_2}$ denotes the cosine distance, $f_{id}^i = \mathrm{E}_{id}(I_i)$ for any I_i. The reconstruction loss is:

$$\mathcal{L}_{rec}(I_g, I_s) = \|I_g - I_s\|_1 + \sum_{m=1}^{N_{vgg}} \|\mathrm{VGG}_m(I_g) - \mathrm{VGG}_m(I_s)\|_1. \tag{3}$$

Optimization Algorithm The choice of the final optimized $\mathcal{W}_s^+$ can be performed in two ways namely the *one-to-one* optimization and *one-to-many* optimization. The *one-to-one* optimization aims at finding the $\mathbf{W}_s^+$ for swapping a specific source-target image pair (I_s, I_t), while *one-to-one* optimization aims to find a general $\mathbf{W}_s^+$ that is suitable for swapping the identity of I_s to any face.

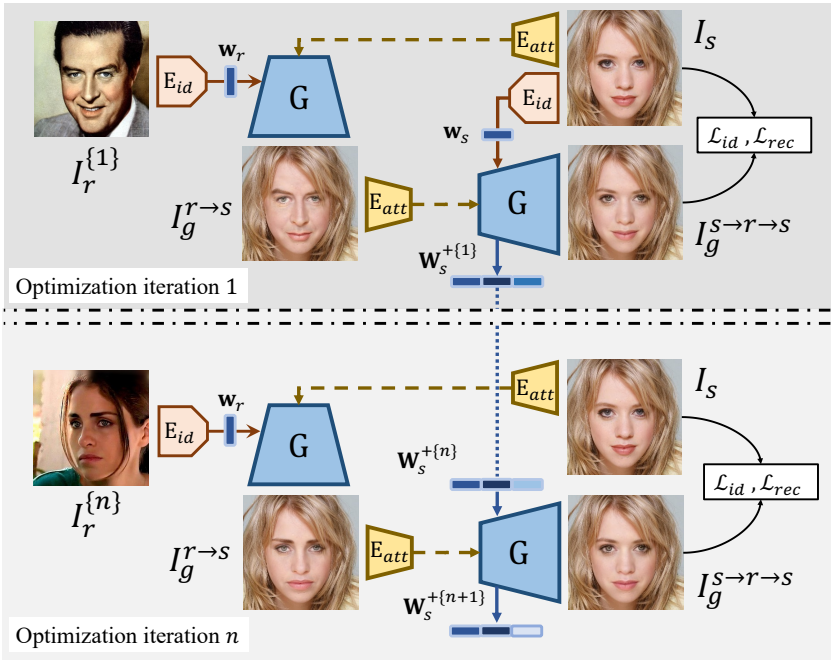

Fig. 1: The pipeline of the Swapping-Driven ID Inversion.

We start by introducing the *one-to-one* optimization. Within the total optimization iteration N, we select the $\mathbf{W}_s^+$ as the feature that achieves the lowest $\mathcal{L}_{id}(I_g^{s\to t}, I_s)$. The whole optimization algorithm is depicted as follows: N is empirically selected as 200. After the optimization, the optimized $\mathbf{W}_s^+$ can be sent into the generator for swapping any target face $I_g^{s\to t} = \mathrm{G}(\mathrm{E}_{att}(I_t), \mathbf{W}_s^+)$ given the source image I_s.

As for the *one-to-many* optimization, all parts related to $\hat{\mathcal{L}}_{id}$ are not required. Thus $\mathbf{W}_s^+$ is set as $\mathbf{W}_s^{+\{N+1\}}$. According to empirical studies carried out on the *one-to-one* setting, the inversion procedure normally optimizes the identity similarity around the first 50 iterations. Thus we set $N = 50$ in the *one-to-many* setting, and this is the standard setting in our experiments.

2 Experiments on Face Forgery Detection

We conduct face forgery detection experiments with backbone Xception [3] that has been widely used as baseline in previous face forgery detection methods [12,8,10]. The experiments are carried out on the following datasets. **1)** FaceForensics++ (FF++) [12] that has been introduced in the main paper. It is the most popular dataset used in face forgery detection. **2)** WildDeepfake (Deepwild) [16] contains 3805 real clips and 3509 fake clips. All these videos are manually collected from the Internet. **3)** Celeb-DF (CDF) [9] which contains

Algorithm 1: The algorithm of Swapping-Guided ID Inversion

Input: A set of images with random identities $\{I_r^{\{1\}}, \ldots, I_r^{\{N\}}\}$;
The source image I_s; The trained encoders E_{id}, E_{att} and G;
The gradient-based optimizer $\mathcal{O}$. The target image I_t.
Output: The $\mathcal{W}^+$ space feature $\mathbf{W}_s^+$.
Initialize $\mathbf{W}_s^+ = \mathbf{W}_s^{+\{1\}} = \{\mathbf{w}_{s(1)}, \ldots, \mathbf{w}_{s(2L)}\}$, $\hat{\mathcal{L}}_{id} = 100$ and $n = 1$.
while $n \leq N$ **do**
  $I_g^{r\to s} \leftarrow \mathrm{G}(\mathrm{E}_{att}(I_s), \mathrm{E}_{id}(I_r^{\{n\}}))$;
  $I_g^{s\to r\to s} \leftarrow \mathrm{G}(\mathrm{E}_{att}(I_g^{r\to s}), \mathbf{W}_s^{+\{n\}})$;
  $\mathcal{L} \leftarrow \lambda_{rec}\mathcal{L}_{rec}(I_g^{s\to r\to s}, I_s) + \lambda_{id}\mathcal{L}_{id}(I_g^{s\to r\to s}, I_s)$;
  $I_g^{s\to t} \leftarrow \mathrm{G}(\mathrm{E}_{att}(I_t), \mathbf{W}_s^{+\{n\}})$;
  if $\mathcal{L}_{id}(I_g^{s\to t}, I_s) < \hat{\mathcal{L}}_{id}$ **then**
    $\mathbf{W}_s^+ \leftarrow \mathbf{W}_s^{+\{n\}}$;
    $\hat{\mathcal{L}}_{id} \leftarrow \mathcal{L}_{id}(I_g^{s\to t})$
  $\mathbf{W}_s^{+\{n+1\}} \leftarrow \mathbf{W}_s^{+\{n\}} - \eta * \mathcal{O}(\nabla_{\mathbf{W}}\mathcal{L})$;
  $n \leftarrow n + 1$;

Table 1: **Face Forgery Detection Experiments.** The cross-dataset evaluation of enlarging the training set with FaceShifter's and our results

Training set \ Test set	FF++	Deepwild	CDF	DFDC	Kodf
FF++	87.66	65.89	66.60	67.62	66.00
FF++ w/ FaceShifter	87.82	70.95	71.92	69.35	70.85
FF++ w/ Ours	**88.16**	**70.96**	**73.59**	**70.07**	**72.13**

high-quality face-swapped videos. **4)** DeepFake Detection Challenge (DFDC) [5] which is one of the most challenging datasets.

As for evaluation metrics, we use Area Under the Receiver Operating Characteristic Curve (AUC). The final confidence score of one video comes from the average of the first 80 frames. The baseline model is trained with 0/1 label (0 for real, 1 for fake, and p-fake) supervision using binary cross-entropy loss.

Specifically, our baseline model is trained on FF++ [12] without involving FaceShifter [7] data. Then we additional involve 50,000 fake images generated by our method and 50,000 fake images from the results of FaceShifter to enlarge the training set to **FF++ w/ Ours**, and **FF++ w/ FaceShifter** respectively. The results are shown in the Table 2.

It can be seen that the model trained with the assistant of our method outperforms the model trained assisted with FaceShifter, which proves that our model could be more useful to the deepfake detection community. We suppose that it is because our model creates fewer artifacts and appears to be more realistic. Thus the forgery detection model trained on our data has more generalization ability.

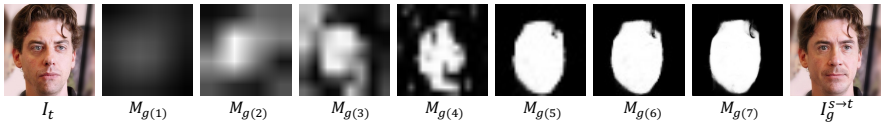

Fig. 2: The visualization of the mask branch

Fig. 3: Comparison with results trained without color jittering

3 More Visualization

Visualization of the ToMask Branch. We visualize a whole set of masks learned in the Swapping-Driven Mask Branch on Fig. 2. All masks are resized to the same resolution (256×256). It can be seen that the contour is gradually learned in the low-level resolutions and finetuned at higher levels.

We also show the results trained without color jittering in Fig. 3 to verify its effectiveness.

More Results. We illustrate more results in this section. For video results, please refer to our video. Specifically, we compare with DeepFakes [4], FaceShifter [7], SimSwap [2], MegaFS [15], InfoSwap [6] in the same way as described in the main paper. We particularly compare with the authors' released results of HifiFace [14]. Their model leverages 3D models for shape changing. However, it leads to unstable results on videos. It can be seen that our results are not only of higher similarity but are also extremely stable and robust compared with previous state-of-the-art methods. Then we show the visualization of an ablation study, where the result not trained with our video-based training paradigm is involved. Finally, we show a clip of swapping different celebrity faces to the same target video.

4 Limitations and Future Work

In order to achieve higher robustness, our method cannot change facial shapes. Also, the optimization of the ID inversion is time-consuming. As a result, we identify certain easy-to-implement directions for improving our method. For example, after the optimization of the $\mathcal{W}^+$ space, recent advances in training StyleGAN encoders [13,1,11] can be directly leveraged to learn an additional encoder in order to get rid of the optimization steps. We leave them to the future.

Fig. 4: **More Qualitative Results**

References

1. Alaluf, Y., Patashnik, O., Cohen-Or, D.: Restyle: A residual-based stylegan encoder via iterative refinement. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6711–6720 (2021) 4
2. 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) 4
3. Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1251–1258 (2017) 2
4. Deepfakes: Faceswap. https://github.com/deepfakes/faceswap 4
5. Dolhansky, B., Bitton, J., Pflaum, B., Lu, J., Howes, R., Wang, M., Ferrer, C.C.: The deepfake detection challenge (dfdc) dataset. arXiv preprint arXiv:2006.07397 (2020) 3
6. 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) 4
7. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. CVPR (2020) 3, 4
8. Li, L., Bao, J., Zhang, T., Yang, H., Chen, D., Wen, F., Guo, B.: Face x-ray for more general face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5001–5010 (2020) 2
9. Li, Y., Yang, X., Sun, P., Qi, H., Lyu, S.: Celeb-df: A large-scale challenging dataset for deepfake forensics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3207–3216 (2020) 2
10. Qian, Y., Yin, G., Sheng, L., Chen, Z., Shao, J.: Thinking in frequency: Face forgery detection by mining frequency-aware clues. In: European Conference on Computer Vision. pp. 86–103. Springer (2020) 2
11. Richardson, E., Alaluf, Y., Patashnik, O., Nitzan, Y., Azar, Y., Shapiro, S., Cohen-Or, D.: Encoding in style: a stylegan encoder for image-to-image translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2287–2296 (2021) 4

12. Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., Nießner, M.: FaceForensics++: Learning to detect manipulated facial images. In: International Conference on Computer Vision (ICCV) (2019) 2, 3
13. Tov, O., Alaluf, Y., Nitzan, Y., Patashnik, O., Cohen-Or, D.: Designing an encoder for stylegan image manipulation. ACM Transactions on Graphics (TOG) (2021) 4
14. Wang, Y., Chen, X., Zhu, J., Chu, W., Tai, Y., Wang, C., Li, J., Wu, Y., Huang, F., Ji, R.: Hififace: 3d shape and semantic prior guided high fidelity face swapping. IJCAI (2021) 4
15. 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 (CVPR). pp. 4834–4844 (2021) 4
16. Zi, B., Chang, M., Chen, J., Ma, X., Jiang, Y.G.: Wilddeepfake: A challenging real-world dataset for deepfake detection. In: Proceedings of the 28th ACM International Conference on Multimedia. pp. 2382–2390 (2020) 2