

# StyleSwap: Style-Based Generator Empowers Robust Face Swapping

Zhiliang Xu<sup>1\*</sup>, Hang Zhou<sup>1\*✉</sup>, Zhibin Hong<sup>1</sup>, Ziwei Liu<sup>2</sup>, Jiaming Liu<sup>1</sup>  
Zhizhi Guo<sup>1</sup>, Junyu Han<sup>1</sup>, Jingtuo Liu<sup>1</sup>, Errui Ding<sup>1</sup>, and Jingdong Wang<sup>1</sup>

<sup>1</sup> Department of Computer Vision Technology (VIS), Baidu Inc.

<sup>2</sup> S-Lab, Nanyang Technological University

{xuzhiliang,zhouhang09,liujiaming03,dingerrui,wangjingdong}@baidu.com  
ziwei.liu@ntu.edu.sg

**Abstract.** Numerous attempts have been made to the task of person-agnostic face swapping given its wide applications. While existing methods mostly rely on tedious network and loss designs, they still struggle in the information balancing between the source and target faces, and tend to produce visible artifacts. In this work, we introduce a concise and effective framework named **StyleSwap**. Our core idea is to *leverage a style-based generator to empower high-fidelity and robust face swapping, thus the generator’s advantage can be adopted for optimizing identity similarity*. We identify that with only minimal modifications, a StyleGAN2 architecture can successfully handle the desired information from both source and target. Additionally, inspired by the ToRGB layers, a *Swapping-Driven Mask Branch* is further devised to improve information blending. Furthermore, the advantage of StyleGAN inversion can be adopted. Particularly, a *Swapping-Guided ID Inversion* strategy is proposed to optimize identity similarity. Extensive experiments validate that our framework generates high-quality face swapping results that outperform state-of-the-art methods both qualitatively and quantitatively. Video, code, and models can be found at <https://hangz-nju-cuhk.github.io/projects/StyleSwap>.

**Keywords:** Face Swapping, Style-based Generator, GAN.

## 1 Introduction

The task of face swapping has drawn great attention [5,4,24,36,31,32,25,58,52,54] due to its wide applications in the fields of entertainment, film making, virtual human creation, privacy protection, *etc.* It aims at transferring the facial identity from a source person to a target frame or video while preserving attributes information including pose, expressions, lighting condition, and background [31].

With the development of deep learning, generative models have been leveraged to boost face swapping quality [24,30,32,25,52,55]. However, the task is still

\* Equal contribution. ✉ Corresponding author.



Fig. 1: **Qualitative results on  $512 \times 512$  resolution.** Our method is robust under complicated conditions and achieves high-fidelity face swapping results

challenging, particularly under the identity-agnostic setting where only a single frame is provided as the source image but targeting different sorts of scenarios. The key challenges lie in two essential parts: how to explicitly capture the identity information; and how to blend the swapped face into the target seamlessly while preserving the implicit attributes unchanged.

To tackle the above problems, previous studies take two different paths. **1)** Graphics-based methods [5,4] have involved the strong prior knowledge of intermediate structural representations such as landmarks and 3D models [13] into face swapping long ago. Recent researchers combine this information with generative adversarial networks (GANs) [15] for identity and expression extraction [33,32,55,52]. However, the inaccuracy of structural information would greatly influence the stableness and coherence of generated results, particularly in videos. **2)** Other methods explore pure learning-based pipelines [25,7,14,60]. Most of them rely on tedious loss and network structure designs [25,7,14] for balancing the information between source and target images. Such designs make training difficult and fail in expressing desired information, which leads to non-similar or non-robust results with visible artifacts.

Recently, StyleGAN architectures [20,21,19] and their variants have been verified effective on various facial generative tasks, including face attributes editing [1,2,41,42], face enhancement [51,56], and even face reenactment [6,57]. It is owing to style-based generator’s strong expressibility and its advantages in latent space manipulation. But the exploration of such architectures in face swapping [60,53] is still insufficient. Specifically, the lighting conditions are greatly condemned in Zhu *et al.* [60] due to the limited distribution covered by the fixed generator. The structure of their feature blending procedure is also designed in a hand-crafted and layer-specific manner, which requires complicated human tuning. Concurrently, Xu *et al.* [53] aggregate the StyleGAN2 features with another designed encoder and decoder. Thus, a natural question arises: can we avoid tedious layer-by-layer structure design by adopting a versatile style-based generator [20,21] with only minimal modifications?

To this end, we propose **StyleSwap**, a concise and effective pipeline that empowers face swapping by a style-based generator. It produces results with higher fidelity, identity similarity and is more robust (*i.e* avoids visible artifacts creation) under different scenarios compared with previous methods. Moreover, it is easy to implement and friendly for training. The key is to adapt StyleGAN2 architecture to face swapping data flows through simple modifications, and adopt the generator’s advantage for identity optimization. Detailedly, we first achieve the restoration of the target image’s attributes with a simple layer-fusion strategy. The same idea has been proven to maintain the original StyleGAN’s capability [56]. Then we argue that the identity information can be injected by mapping extracted identity features to the  $W$  space. In this way, the identity information can be implicitly blended into the attributes in the convolution operations. Additionally, we propose a *Swapping-Driven Mask Branch* which is identical to the *ToRGB* branch. It naturally enforces the network to focus less on the target’s high-level information and benefits final image blending.

We further illustrate the advantages of this architecture by involving a simple optimization strategy for improving identity similarity. As the identity feature is mapped to the  $W$  space, a natural inspiration from recent StyleGAN inversion studies [1,2] is to optimize a powerful  $W^+$  space through self-reconstruction. To avoid mode collapse, we introduce a novel Swapping-Guided ID Inversion strategy by iteratively performing feature optimization and face swapping. Armed with these tools, we show that our StyleSwap generates high-fidelity results with simple video training paradigms. It is particularly robust and can be supported with enhanced data for generating high-resolution results.

We summarize our contributions as follows: **(1)** We present the **StyleSwap** framework, which adopts a style-based generator into the person-agnostic face swapping task by simple modifications and the design of a Swapping-Driven Mask Branch. It is easy to implement and train. **(2)** By taking the advantage of the StyleGAN model, we design the novel *Swapping-Guided ID Inversion* strategy to improve the identity similarity. **(3)** Extensive experiments demonstrate that our method outperforms the state of the arts on person-agnostic face swapping. Particularly, it demonstrates great robustness and has the capability of generating high-quality results.

## 2 Related Work

### 2.1 Face Swapping

**Structural Prior-Guided Face Swapping.** The task of face swapping has long been a research interest for both the computer graphics and computer vision community. Structural information such as 3D models and landmarks provide strong prior knowledge. Blanz *et al.* [5] leverage 3DMM, and Bitouk *et al.* use 3D lighting basis to design adjustment-based methods. Both of them rely on manual interaction, and can hardly change the source’s expressions. Nirkin *et al.* [33] involve 3DMM with learned masks, but they render unrealistic results. Recent

studies [32,17,55,52] combine structural information with GANs for identity-agnostic face swapping. Xu *et al.* and Wang *et al.* both inject the parameters of 3DMM into self-designed architectures. Though high-fidelity results can be generated, the inaccuracy of 3D models and the need for inpainting greatly harm the temporal coherence and robustness of these methods under the video face swapping setting.

**Reconstruction-Based Face Swapping.** On the other hand, pure reconstruction based methods with GANs have also shown success. Korshunova *et al.* [24] train a network for swapping paired identities. The popular Deepfakes and DeepFaceLab [36] share the same setting. However, these methods cannot generalize to arbitrary identities, which limits their practical usage.

As for person-agnostic face swapping, Li *et al.* [25] build the Faceshifter network. SimSwap [7] improves the expression consistency. However, they generate low-quality results with visible artifacts under certain circumstances. Recently, InfoSwap [14] creates high quality results by building a pipeline that relies on careful loss designs. It involves multi-stage of finetuning on various datasets. Different from these methods, we would like to ease the network design procedure with a style-based generator.

Specifically, Wang *et al.* [50] firstly leverage a pretrained StyleGAN generator for high resolution face swapping. However, in order to adapt to the latent spaces of a pretrained StyleGAN generator, the authors design layer-specific fusion strategies, which involves a large number of hyper-parameters and ablative studies. Moreover, their method cannot keep the lighting conditions of the target frames. In our work, we retrain the style-based generator with simple modifications that preserve the attribute information better.

## 2.2 Facial Editing with Style-based Generator

The strong ability of StyleGAN [20,21,19] has been shown in various facial editing tasks including facial attributes editing [1,2,41,42,46,38], blind face restoration [51,56], face reenactment [6,57,28], hairstyle editing [59], and so on [59,46,3,45].

**StyleGAN Inversion.** Most face attribute editing framework [41,42] fix the pretrained generator unchanged and perform StyleGAN inversion. Abdal *et al.* [1,2] expand the original  $\mathcal{W}$  latent space to the  $\mathcal{W}^+$  space during inversion and achieve better image reconstruction results. Recent studies invert images with StyleGAN specific encoders [46,3,38] for fast inversion. In our work, we take the inspiration of StyleGAN inversion and expand our identity feature to  $\mathcal{W}^+$  space for boosting identity similarity. The usage of the StyleGAN specific encoders is left as a future work.

**Face Reenactment with Style-based Generator.** Face reenactment is very similar to face swapping and even serves as part of the face swapping procedure in certain methods [32,36,43]. The difference is that it aims at keeping the source image’s identity and background. Burkov *et al.* [6] encode identity and expression information into the  $\mathcal{W}$  space and re-train the generator in a simple pipeline. Later studies [57,28] then expand this pipeline to the audio-driven setting.

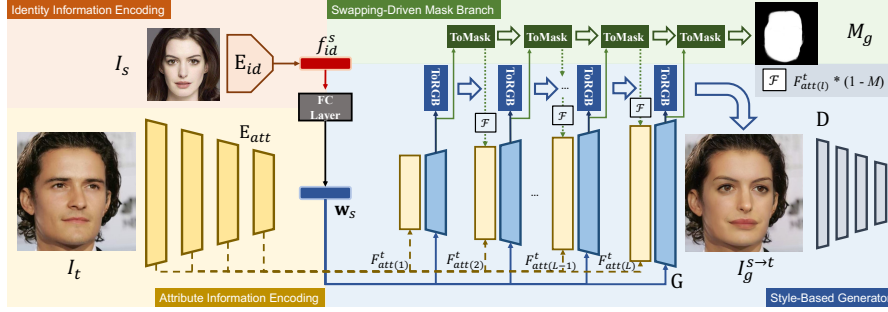


Fig. 2: **Our StyleSwap framework.** The building blocks in **Blue** indicate the original structure of StyleGAN2. The source  $I_s$  is encoded to  $f_{id}^s$  by  $E_{id}$  (**Red**). The attribute feature maps  $\mathbf{F}_{att}^t$  are encoded from  $I_t$  by  $E_{att}$  and concatenated to the StyleGAN2 generator blocks (**Yellow**). Specifically, we devise a Swapping-Driven Mask Branch (**Green**) to predict a mask  $M_g$  by leveraging the same structure of ToRGB layers

### 3 Our Approach

**Framework Overview.** The goal of person-agnostic face swapping is to swap the identity information from a source image  $I_s$  onto a target frame  $I_t$  while preserving  $I_t$ ’s attribute information unchanged.

In this section, we present our **StyleSwap** framework (illustrated in Fig. 2) which empowers face swapping by slightly modifying a style-based generator. In Section 3.1 we introduce how to adapt a style-based generator to the face swapping task. Section 3.2 illustrates the simple training paradigm. Importantly, in Section 3.3, we propose the *Swapping-Guided ID Inversion* which specifically takes the advantage of GAN inversion for optimizing identity similarity.

#### 3.1 Adapting Style-Based Generator to Face Swapping

**Revisiting StyleGAN2.** We first recap the general setting of StyleGAN2 [21]. The original generator takes a constant feature map at the lowest resolution, then a latent vector  $\mathbf{z}$  is sampled and mapped to a feature vector  $\mathbf{w}$  (lies in the  $\mathcal{W}$  space). Afterwards,  $\mathbf{w}$  is sent into each layer of the  $2L$ -layer generator by affine transformations as the “style” to modulate the convolutional kernels’ weights. At each resolution, a ToRGB layer is designed to render a three-channel RGB image progressively. The rough structures are depicted on Fig. 2 in blue.

The attribute disentanglement ability in StyleGAN2 is implicitly achieved in the  $\mathcal{W}$  or the expanded  $\mathcal{W}^+$  space (if feeding different  $\mathbf{w}$  features to different layers). As a result, operations on existing faces require inverting faces to latent vectors, which harms the preservation of spatial information. In our face swapping task, the problem lies in how to modify the generator so that the information of the target frame and identity can be sufficiently used.

**Infusing Attribute Information.** We propose to infuse the spatial information of the target frame as feature maps, rather than feature vectors, to preserve attribute information. A recent face restoration work [56] verifies two important properties of StyleGAN2. Firstly, concatenating a noise map to each layer of the generator would not affect the network’s generative ability. Secondly, such noise maps can be replaced by encoded spatial feature maps so that both the generative prior of StyleGAN2 model and the structural information of the input image can be kept.

Inspired by this observation, we apply a similar modification. Specifically, we leverage a simple encoder  $E_{att}$  that encodes  $I_t$  to different scales of spatial feature maps  $\mathbf{F}_{att}^t = \{F_{att(l)}^t | l \in [1, L]\}$ . They are then infused into the StyleGAN2 architecture at each  $2l$ -th layer of the generator by concatenation following [56].

**Injecting Identity Feature.** In order to adapt to different target views, it is natural to encode the identity information into feature vectors. Here we use a pretrained identity encoder provided by ArcFace [11] to encode the identity feature  $f_{id}^s = E_{id}(I_s)$ .

As the attributes on the face are already fused, we identify that modulated convolutions in StyleGAN2 are naturally suitable for the blending and shape-shifting of facial organs. Thus we directly map  $f_{id}^s$  to  $\mathbf{w}_s$  in the  $\mathcal{W}$  space ( $\mathbf{w}_s = FC_{\mathbf{w}}(f_{id}^s)$ ) through fully connected layers  $FC_{\mathbf{w}}$ .

So far, a swapping result  $I_g^{s \rightarrow t} = G(\mathbf{F}_{att}^t, \mathbf{w}_s)$  from the source to the target frame can already be rendered by the generator.

**Swapping-Driven Mask Branch.** We then identify that learning a rough one-channel facial mask would benefit the whole face swapping step from two perspectives. 1) With a mask in the image domain, the areas that do not require modifications such as the background and hair can be directly kept unchanged. 2) If the mask can be progressively and implicitly learned along with the generator, coarse masks at lower resolutions can help balance attribute and identity information in a similar way as [25]. We thus propose an additional modification by devising a *Swapping-Driven Mask Branch* which takes the advantage of the StyleGAN2 model. Its structure is directly borrowed from the “ToRGB” branch, and denoted as “ToMask”. Here we leverage a soft mask with values between 0 and 1. The details of the mask branch are illustrated in Fig. 3 (a).

Let  $M'_{g(l)}$  denote the one-channel output of the  $l$ -th ToMask network, which has the same resolution with the output of the  $l$ -th ToRGB layer and the  $(l+1)$ -th  $\mathbf{F}_{att}^t$ . The non-normalized mask  $\tilde{M}_{g(l)}$  at the  $l$ -th layer of the mask branch is the combination of the  $(l-1)$ -th layer’s result and  $M'_{g(l)}$ .

$$\tilde{M}_{g(l)} = \text{upsample}(\tilde{M}_{g(l-1)}) + M'_{g(l)}, \quad (1)$$

where the bilinear *upsample* is used. The softmasks we used are the normalized results  $M_{g(l)} = \text{Sigmoid}(\tilde{M}_{g(l)})$ .  $M_{g(1)}$  is the normalized output of the first ToMask network, and  $M_g$  is the final predicted mask. The face swapping result can be updated as:

$$\hat{I}_g^{s \rightarrow t} = M_g * I_g^{s \rightarrow t} + (\mathbf{1} - M_g) * I_t, \quad (2)$$



where the  $*$  denotes element-wise multiplication with broadcasting and  $\mathbf{1}$  is the tensor with all ones. The mask repeats itself channel-wise three times to match the RGB channels.

**Masking Attribute Information.** Our design above infuses facial attribute information into all layers. However, spatial information provided at mid- and low-level layers might influence facial structures. Though we expect the network to automatically perform information balancing, we identify that this procedure can be eased by blocking the attribute information relying on an implicitly learned mask. Thus we multiply our learned mask  $M_{g(l)}$  at each resolution with the next layer’s attribute feature map  $F_{att(l+1)}^t$  to an updated version:

$$\hat{F}_{att(l+1)}^t = F_{att(l+1)}^t * (\mathbf{1} - M_{g(l)}). \quad (3)$$

Note that  $M_{g(l)}$  and  $F_{att(l+1)}^t$  share the same spatial resolution, and there is no masking operation on  $F_{att(1)}^t$  which provides the initial facial attributes. In this way, as  $M_{g(l)}$  progressively grows to reach the ground truth mask, it also implicitly prevents the attribute information from influencing the identity similarity.

### 3.2 Training Paradigm

With our StyleSwap architecture, person-agnostic face swapping results can be learned in a simple end-to-end training paradigm. Particularly, we propose to involve certain video data for training a more robust swapping model.

Given a source image  $I_s$  and a target video  $\{I_T | T \in \{t(1) \dots t(K)\}\}$ , we generate the following frames: **(1)** The face swapping results from the source to any target frame:  $I_g^{s \rightarrow t} = G(\mathbf{F}_{att}^t, \mathbf{w}_s)$ , where  $\mathbf{F}_{att}^t = E_{att}(I_t)$  and  $\mathbf{w}_s = \text{FC}_{\mathbf{w}}(f_{id}^s) = \text{FC}_{\mathbf{w}}(E_{id}(I_s))$ . **(2)** The self-reconstruction results on the source frame itself:  $I_g^{s \rightarrow s} = G(\mathbf{F}_{att}^s, \mathbf{w}_s)$ . **(3)** Particularly, we sample two target frames  $I_{t(a)}$ ,  $I_{t(b)}$  from a same video of a same person, and generate  $I_g^{t(a) \rightarrow t(b)} = G(\mathbf{F}_{att}^{t(b)}, \mathbf{w}_{t(a)})$ ,  $\mathbf{w}_{t(a)} = \text{FC}_{\mathbf{w}}(f_{id}^{t(a)})$ .

The training objectives consist of mainly four parts: the identity loss, feature matching loss, adversarial loss applied to all generated results, and the reconstruction loss applied to self- and cross-view reconstruction results.

**Identity Loss.** The identity loss is built upon the cosine distances  $D_{\cos}(f_a, f_b) = \frac{f_a^T \cdot f_b}{\|f_a\|_2 \|f_b\|_2}$  between extracted identity features from  $E_{id}$ . Given any sampled data  $I_i, I_j$  where  $i, j \in \{s, T\}$ :

$$\mathcal{L}_{id} = 1 - D_{\cos}(f_{id}^i, E_{id}(I_g^{i \rightarrow j})). \quad (4)$$

Note that in order to disentangle identity information with illuminations, all images are augmented with color jittering when sent to the identity encoder.

**Adversarial Loss and Feature Matching Loss.** We directly adopt the original discriminator  $D$  and adversarial loss functions of StyleGAN2 [20]. For any  $I_g^{i \rightarrow j}(i, j \in \{s, T\})$ ,  $I_j$  is provided as the real image when applying the adversarial training. We denote this loss function as  $\mathcal{L}_{adv}$  and omit the details.

Similar to [7], we leverage a weak feature matching loss from the feature maps of the discriminator.

$$\mathcal{L}_{FM} = \sum_{m=n_D}^{N_D} (\|D_m(I_g^{i \rightarrow j}) - D_m(I_j)\|_1), \quad (5)$$

where  $D_m$  denotes the  $m$ -th layer’s output of the discriminator, and  $n_D$  is the layer that starts computing the feature matching loss. This loss accounts for preserving the expression and certain low-level attribute information.

**Reconstruction Loss.** The reconstruction loss consists of the  $L_1$  loss and the VGG perceptual loss [49,34] when the pixel-level supervision can be provided.

$$\mathcal{L}_{rec} = \|I_g^{i \rightarrow j} - I_j\|_1 + \sum_{m=1}^{N_{vgg}} \|\text{VGG}_m(I_g^{i \rightarrow j}) - \text{VGG}_m(I_j)\|_1, \quad (6)$$

where  $\text{VGG}_m$  denotes the  $m$ -th layer’s output of a pre-trained VGG19 network. When applying the reconstruction loss, we set  $i, j = s$  or  $i, j \in \{T\}$ . The reconstruction training provides supervision in the pixel space and has proven to be crucial in previous studies. Particularly, cross-view reconstruction is widely used in face reenactment training [22,6,57]. It benefits the attribute information’s preservation by providing samples with different expressions on the source and target, and forces the network to learn with strong supervision.

**Mask Loss.** We leverage a pretrained facial mask predictor [44] to predict only a rough facial mask for each image in the training set. For each generated image  $I_g^{i \rightarrow j} = G(\mathbf{F}_{att}^j, \mathbf{w}_i)$ , it is supervised with the binary mask of the target frame  $M_j$ . The supervision can be written as:

$$\mathcal{L}_{mask} = \text{BCELoss}(M_g, M_j), \quad (7)$$

where BCELoss denotes the point-wise binary cross entropy loss.

It is worth mentioning that the mask branch is a plug-in module that requires fine-tuning. When the **Masking Attribute Information** in Sec. 3.1 is activated, all generated images are updated to  $\hat{I}_g^{i \rightarrow j}$  as illustrated in Eq. 2 and all  $\mathbf{F}_{att}$  are updated according to Eq. 3.

**Overall Loss Function.** The overall loss function for training the StyleSwap framework is the combination of the losses introduced above. It can be represented as:

$$\mathcal{L}_{total} = \mathcal{L}_{adv} + \lambda_{id}\mathcal{L}_{id} + \lambda_{FM}\mathcal{L}_{FM} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{mask}\mathcal{L}_{mask}, \quad (8)$$

where the  $\lambda$ s are balancing coefficients.

### 3.3 Swapping-Guided ID Inversion

Inspired by StyleGAN inversion, we illustrate one interesting property of our  $\mathcal{W}$  space design: the identity similarity can be optimized in a GAN inversion manner



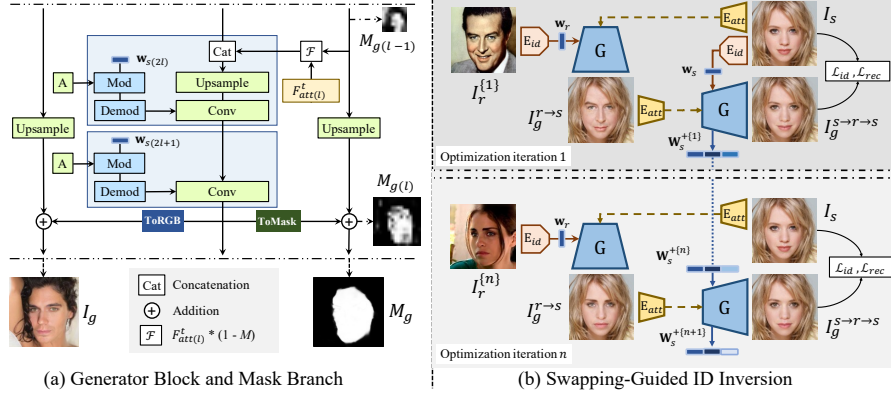


Fig. 3: (a) The details of the generator blocks and the *Swapping-Guided Mask Branch*. The figures show the block with visualized masks when  $l = 4$  and the final outputs.  $\mathbf{w}_{s(2l)} = \mathbf{w}_{s(2l+1)} = \mathbf{w}_s$  when the  $\mathbf{w}_s$  is not optimized. (b) The procedure for *Swapping-Guided ID Inversion*. At each iteration, a different random  $I_r^{\{n\}}$  is selected. The first and the  $n$ -th iterations are shown.  $\mathbf{w}_s$  is taken as initialization and updated to the  $\mathcal{W}^+$  space

through a *Swapping-Guided ID Inversion* strategy, where the generator is fixed. As the encoded  $\mathbf{w}_s$  feature from the source lies in the  $\mathcal{W}$  space, it can naturally be served as a good initialization. We expect to optimize  $\mathcal{W}$  or expanded  $\mathcal{W}^+$  space for a specific person’s identity. The optimization procedure on  $\mathcal{W}^+$  space is depicted in Fig. 3 (b). The reconstruction loss  $\mathcal{L}_{rec}$  and ID loss  $\mathcal{L}_{id}$  are used.

**Challenges.** We first take  $\mathcal{W}$  space optimization as an example, and expand it to the  $\mathcal{W}^+$  space. An intuitive operation is to directly optimize  $\mathbf{w}_s$  (to any  $\mathbf{w}_s + \Delta\mathbf{w}_s$ ) using gradient descent at each self-reconstruction step  $I_g^{s \rightarrow s} = G(\mathbf{F}_{att}^s, \mathbf{w}_s + \Delta\mathbf{w}_s)$ . However, such practice would easily lead to mode collapse, *i.e.*, direct reconstruction by ignoring the  $\mathbf{w}_s$  information. To tackle this problem, an alternative way is to perform face swapping before the optimization step (as shown in the optimization iteration 1 in Fig. 3). A randomly sampled face  $I_r$  of arbitrary identity is used to build  $I_g^{r \rightarrow s}$  which is later sent to the attribute encoder as the target frame. However, this way of restricting  $I_g^{s \rightarrow r \rightarrow s} = G(I_g^{r \rightarrow s}, \mathbf{w}_s)$  still relies on one fixed input image, which also leads to mode collapse during implementation.

**Iterative Identity Optimization and Swapping.** The key to our method is to feed a different face to the source image at each iteration of the optimization procedure. Specifically, at each iteration  $n$ , we randomly sample any  $I_r^{\{n\}}$  to update the identity information on the source frame, so that the network perceives different identities.

Moreover, a unique advantage of Style-based generator is to optimize the set of  $2L$  different style features  $\mathbf{W}_s^+ = [\mathbf{w}_{s(1)}, \dots, \mathbf{w}_{s(2L)}]$  as performed in StyleGAN

inversion studies [1,2]. At each iteration, it is updated to  $\mathbf{W}_s^{+\{n\}}$ . After optimizing  $\mathbf{W}_s^+$ , it can be directly leveraged by the generator to create face swapping result:  $I_q^{s \rightarrow t} = G(\mathbf{F}_{att}^t, \mathbf{W}_s^+)$ . Such operation can hardly be performed on traditional face swapping pipelines, and has never been explored before. More details about the optimization algorithm are shown in the Supplementary Materials.

## 4 Experiments

**Datasets.** We train our model on the VGGFace [35] and a small part of VoxCeleb2 [9]. Due to the limitation of data quality, the original data from both datasets only supports  $256 \times 256$  resolution training. To show our capability of handling higher resolutions, we enhance the datasets with GPEN [56] and train a  $512 \times 512$  model. Part of the evaluation is conducted on FaceForensics++ (FF++) dataset [39] which contains 1,000 Internet face videos and 1,000 Deepfakes [10] and 1000 official FaceShifter results. Evaluations on high-quality images leverage CelebA-HQ [18,29].

**Implementation Details.** Our model is trained with batch size 64 on one NVIDIA Tesla A100 GPU for  $256 \times 256$  resolution and batch size 12 for  $512 \times 512$  resolution. We use the ADAM optimizer [23] with learning rate fixed at  $1 \times 10^{-4}$ . The  $\lambda_{id}$  is set at 10,  $\lambda_{rec}$  and  $\lambda_{FM}$  are set at 100. The other coefficients do not affect the generation results much thus empirically set at 1.

A total of  $2L = 14$  layers are used on  $256 \times 256$  resolution images, and  $2L = 16$  layers on  $512 \times 512$  resolution. We leverage the modified ResNet50 provided by Arcface [11] as the identity encoder and fix it through out the experiments. The number of iterations required for optimizing the  $\mathcal{W}$  and  $\mathcal{W}^+$  for one identity is set empirically at 50. Please refer to supplementary materials for more details.

**Competing Methods.** We compare our methods with previous state-of-the-arts reconstruction-based face swapping methods. They are **Faceshifter** [25] and **SimSwap** [7] which leverage self-designed network structures; **InfoSwap** [14] that relies on tedious training procedure for high-resolution results; **MegaFS** [60] which borrows a fixed StyleGAN2 pretrained model; and **Deepfakes** [10] which is a famous open-sourced tool. We use the official Deepfakes and FaceShifter results from the FF++ dataset and the officially released codes and models for producing swapped results of other methods. We refer our results without identity optimization as **StyleSwap** and depict the optimized one separately as **StyleSwap w/  $\mathcal{W}^+$** .

### 4.1 Quantitative Evaluation

**Evaluation Metrics.** The quantitative experiments are carried out on the FF++ dataset [39]. Following [25], 10 frames are uniformly sampled from the 1000 videos to get 10K faces for evaluation. We leverage five popularly used metrics. The *ID cosine similarity* and *ID retrieval scores* are measured between swapped results and the source using another pretrained identity-recognition network [48]. Particularly, the calculation process of the ID retrieval score is the

Table 1: **Quantitative Results.** We report the percentage of successfully retrieved images on the *ID Retrieval* metric. For the ID correlated metrics the higher the better, and it is the lower the better for other metrics

Method \ Metric	ID Retrieval $\uparrow$	ID Cosine $\uparrow$	Pose error $\downarrow$	Exp. error $\downarrow$	FID $\downarrow$
Deepfakes [10]	86.43%	0.438	3.96	8.98	4.07
FaceShifter [25]	90.04%	0.510	2.19	6.77	3.50
SimSwap [7]	93.07%	0.578	<b>1.36</b>	<b>5.07</b>	3.04
MegaFS [60]	89.12%	0.497	3.69	10.12	4.62
InfoSwap [14]	95.82%	0.635	2.54	6.99	4.74
<b>StyleSwap (Ours)</b>	<b>97.05%</b>	<b>0.677</b>	1.56	5.28	<b>2.72</b>
<b>StyleSwap w/ <math>\mathcal{W}^+</math></b>	<b>97.87%</b>	<b>0.706</b>	1.51	5.27	<b>2.58</b>

same as [25]. The pose error is evaluated on the poses’  $L_2$  distances produced by a pose estimator [40]. The expression error is the  $L_2$  distance between the results’ and targets’ expression embedding extracted from a facial expression extraction model [47].

**Quantitative Results.** We list the results in Table 1. As can be seen that our **StyleSwap** model directly achieves state-of-the-art results on ID-related metrics and FID compared to previous ones, which shows the high similarity and image quality of our method. The *Swapping-Driven ID Inversion* strategy further improves ID similarity. Meanwhile, we achieve comparable pose and expression errors with SimSwap [7] and outperform the others. The qualitative results further prove that we preserve the attribute information at a robust level.

## 4.2 Qualitative Evaluation

Subjective evaluation plays an important role in face swapping. Note that for fair comparisons, we do not perform  $\mathcal{W}^+$  space optimization. We provide image-based comparisons below. Demo videos and resources are available at <https://hangz-nju-cuhk.github.io/projects/StyleSwap>.

**Qualitative Results on  $256 \times 256$  Resolution.** Here we show the image-based comparisons on  $256 \times 256$  resolution in Fig. 4. The source and target figures are selected from FF++ [39]. It can be seen that MegaFS [60] keeps the lighting conditions and texture badly, and InfoSwap [14] also fails to preserve accurate information. Though for some cases, SimSwap renders the best expression (the first row), it sometimes produces visible artifacts. Our method generates robust results with the highest similarity and competitive attribute preservation.

**Qualitative Results on  $512 \times 512$  Resolution.** Particularly, we are able to generate high-resolution results with  $512 \times 512$  resolution. Some results are shown in Fig. 1. We compare our results with MegaFS [60] and InfoSwap [14] which are also able to produce high-resolution results. As shown in Fig. 5, MegaFS generates a false skin tone of the target person and InfoSwap generates visible artifacts. Our method clearly outperforms theirs on high-resolution results.



Fig. 4: **Qualitative results on FF++ with  $256 \times 256$  resolution.** Please pay attention to the identity information at the **red** arrows, and the attribute information at the **yellow** arrows

Table 2: **User Study’s Ranking Scores.** Larger is higher, with the maximum value to be 5

Perspective \ Method	FaceShifter [25]	SimSwap [7]	MegaFS [60]	InfoSwap [14]	Ours
ID Similarity	3.05	2.59	2.80	2.89	<b>3.68</b>
Att. Preservation	3.16	2.94	2.21	3.36	<b>3.82</b>
Naturalness	2.95	2.62	2.43	3.11	<b>3.88</b>

**User Study.** We further conduct a user study for subjective evaluations. A total of 30 users are involved to discriminate 20 different samples from the FF++ dataset. The users are asked to rank the quality of the fake images from the following perspectives: **1) Id similarity** with the source image; **2) Attribute (Att) Preservation** including expressions and backgrounds referring to the target; and **3) Naturalness**. Whether there are visible artifacts on the face and does the figure look like a real person. Detailed instructions and training are provided. As Deepfakes cannot produce plausible results, we only conduct the user study on the other 4 methods and our StyleSwap. Thus we define the highest score to be 5 and the lowest score to be 1 for each case.

The results are shown in Table 2. All ranks are given corresponding scores with 5 being the highest. It can be seen that our method performs the best on

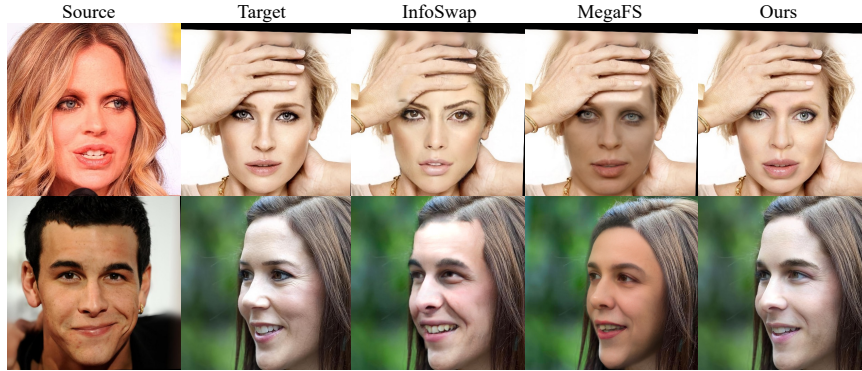


Fig. 5: **Qualitative results** on CelebA-HQ with  $512 \times 512$  resolution

Table 3: **Quantitative Ablation Study.** The StyleSwap model with  $\mathcal{W}^+$  space optimization achieves the best results on all metrics

Method \ Metric	ID Retrieval $\uparrow$	ID Cosine $\uparrow$	Pose error $\downarrow$	Exp. error $\downarrow$	FID $\downarrow$
StyleSwap (vector)	96.68%	0.668	2.81	8.44	4.54
StyleSwap w/o Mask	95.56%	0.653	1.80	5.78	3.18
StyleSwap	97.05%	0.677	1.56	5.28	2.72
StyleSwap w/ $\mathcal{W}$	97.54%	0.693	1.52	5.28	2.60
<b>StyleSwap w/ <math>\mathcal{W}^+</math></b>	<b>97.87%</b>	<b>0.706</b>	<b>1.51</b>	<b>5.27</b>	<b>2.58</b>

given all metrics. Specifically, the low score of SimSwap with respect to attribute preservation is led by their artifacts.

### 4.3 Further Analysis

**Ablation Studies.** We conduct ablation studies on several important designs of our network. (1) The effectiveness of the  $\mathcal{W}^+$  space optimization and its comparison with the  $\mathcal{W}$  space optimization **StyleSwap w/  $\mathcal{W}$** ; (2) Our StyleSwap without the Mask Branch **StyleSwap w/o Mask**. (3) Our design of concatenating the target’s spatial feature maps into the style-based generator. An alternative way is to map both source and target information into the  $\mathcal{W}$  as performed in [6, 57]. This model is denoted as **StyleSwap (vector)**. The cross-reconstruction training paradigm does not affect numerical results and visualization in most cases, thus the comparison will be given.

We show the quantitative ablation results in Table 3 under the same evaluation protocols on FF++, and the qualitative ablation results on Fig. 6 which is generated on CelebA-HQ [18]. It can be seen that without the spatial feature maps, StyleSwap (vector) cannot preserve the attribute information correctly and thus loses details. Moreover, the identity similarity improves along with the addition of the *Mask Branch*, the  $\mathcal{W}$  and  $\mathcal{W}^+$  space optimization.



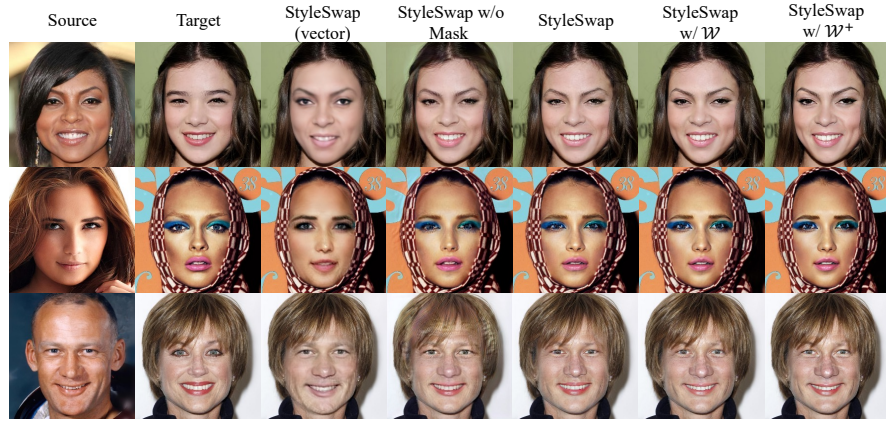


Fig. 6: Qualitative Ablation Study on CelebA-HQ.

**Effectiveness on Face Forgery Detection.** Importantly, our method can contribute to the face forgery detection community [39,31,16]. We conduct experiments under the standard face forgery detection pipeline and additionally provide the same amount of generated data by our method and FaceShifter [25]. The evaluation results on 4 different datasets validate that our method assists the forgery detection better. The details can be found in the Supplementary Materials.

## 5 Conclusion and Discussion

**Conclusion.** In this paper, we propose **StyleSwap**, a concise and effective framework that adapts a style-based generator for high-fidelity face swapping. We emphasize several key properties of our method: 1) With only minor modifications to the StyleGAN2 generator, our method is easy to implement and friendly to train, which saves a lot of human labor. 2) With the strong capability of the style-based generator and the simple design of the *Swapping-Guided Mask Branch*, our results are not only with high quality, similarity but enjoy high robustness. 3) Our method can take the advantage of GAN inversion and optimize the  $W^+$  space for improving identity similarity. 4) Our method can benefit face forgery detection by providing realistic fake results.

**Broader Impact.** Face swapping technique could create deepfake results for malicious purposes. We also take this issue into serious consideration and show the effectiveness of our method in the face forgery detection community. We will strictly limit the usage of this work for research purposes only.

**Acknowledgements.** This work is supported by NTU NAP, MOE AcRF Tier 1 (2021-T1-001-088), and under the RIE2020 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) Funding Initiative, as well as cash and in-kind contributions from the industry partner(s).

## References

1. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan: How to embed images into the stylegan latent space? In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4432–4441 (2019) [2](#), [3](#), [4](#), [10](#)
2. Abdal, R., Qin, Y., Wonka, P.: Image2stylegan++: How to edit the embedded images? In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8296–8305 (2020) [2](#), [3](#), [4](#), [10](#)
3. 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](#)
4. Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., Nayar, S.K.: Face swapping: automatically replacing faces in photographs. In: ACM SIGGRAPH 2008 papers, pp. 1–8 (2008) [1](#), [2](#)
5. Blanz, V., Scherbaum, K., Vetter, T., Seidel, H.P.: Exchanging faces in images. In: Computer Graphics Forum. vol. 23, pp. 669–676. Wiley Online Library (2004) [1](#), [2](#), [3](#)
6. Burkov, E., Pasechnik, I., Grigorev, A., Lempitsky, V.: Neural head reenactment with latent pose descriptors. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13786–13795 (2020) [2](#), [4](#), [8](#), [13](#)
7. 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](#), [4](#), [8](#), [10](#), [11](#), [12](#)
8. 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) [20](#)
9. Chung, J.S., Nagrani, A., Zisserman, A.: Voxceleb2: Deep speaker recognition. arXiv preprint arXiv:1806.05622 (2018) [10](#)
10. Deepfakes: Faceswap. <https://github.com/deepfakes/faceswap> [10](#), [11](#)
11. 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) [6](#), [10](#)
12. 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) [20](#)
13. Egger, B., Smith, W.A., Tewari, A., Wuhler, S., Zollhoefer, M., Beeler, T., Bernard, F., Bolkart, T., Kortylewski, A., Romdhani, S., et al.: 3d morphable face models—past, present, and future. ACM Transactions on Graphics (TOG) **39**(5), 1–38 (2020) [2](#)
14. 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) [2](#), [4](#), [10](#), [11](#), [12](#)
15. 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) [2](#)
16. Guan, J., Zhou, H., Hong, Z., Ding, E., Wang, J., Quan, C., Zhao, Y.: Delving into sequential patches for deepfake detection. Advances in Neural Information Processing Systems (2022) [14](#)
17. Jiang, L., Li, R., Wu, W., Qian, C., Loy, C.C.: Deepforensics-1.0: A large-scale dataset for real-world face forgery detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2889–2898 (2020) [4](#)



18. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. arXiv preprint arXiv:1710.10196 (2017) [10](#), [13](#)
19. Karras, T., Aittala, M., Laine, S., Härkönen, E., Hellsten, J., Lehtinen, J., Aila, T.: Alias-free generative adversarial networks. In: Proc. NeurIPS (2021) [2](#), [4](#)
20. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 4401–4410 (2019) [2](#), [4](#), [7](#)
21. 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) [2](#), [4](#), [5](#)
22. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Niessner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. ACM Transactions on Graphics (TOG) (2018) [8](#)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014) [10](#)
24. Korshunova, I., Shi, W., Dambre, J., Theis, L.: Fast face-swap using convolutional neural networks. In: Proceedings of the IEEE international conference on computer vision. pp. 3677–3685 (2017) [1](#), [4](#)
25. Li, L., Bao, J., Yang, H., Chen, D., Wen, F.: Faceshifter: Towards high fidelity and occlusion aware face swapping. CVPR (2020) [1](#), [2](#), [4](#), [6](#), [10](#), [11](#), [12](#), [14](#), [20](#)
26. 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) [20](#)
27. 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) [20](#)
28. Liang, B., Pan, Y., Guo, Z., Zhou, H., Hong, Z., Han, X., Han, J., Liu, J., Ding, E., Wang, J.: Expressive talking head generation with granular audio-visual control. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3387–3396 (June 2022) [4](#)
29. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: Proceedings of the IEEE international conference on computer vision. pp. 3730–3738 (2015) [10](#)
30. Natsume, R., Yatagawa, T., Morishima, S.: Rsgan: face swapping and editing using face and hair representation in latent spaces. arXiv preprint arXiv:1804.03447 (2018) [1](#)
31. Nguyen, T.T., Nguyen, Q.V.H., Nguyen, C.M., Nguyen, D., Nguyen, D.T., Nahavandi, S.: Deep learning for deepfakes creation and detection: A survey. arXiv preprint arXiv:1909.11573 (2019) [1](#), [14](#)
32. 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) [1](#), [2](#), [4](#)
33. Nirkin, Y., Masi, I., Tuan, A.T., Hassner, T., Medioni, G.: On face segmentation, face swapping, and face perception. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 98–105. IEEE (2018) [2](#), [3](#)
34. Park, T., Liu, M.Y., Wang, T.C., Zhu, J.Y.: Semantic image synthesis with spatially-adaptive normalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2019) [8](#)

35. Parkhi, O.M., Vedaldi, A., Zisserman, A.: Deep face recognition (2015) 10
36. Perov, I., Gao, D., Chervoniy, N., Liu, K., Marangonda, S., Umé, C., Dpfks, M., Facenheim, C.S., RP, L., Jiang, J., et al.: Deepfacelab: Integrated, flexible and extensible face-swapping framework. arXiv preprint arXiv:2005.05535 (2020) 1, 4
37. 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) 20
38. 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
39. 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) 10, 11, 14, 20
40. Ruiz, N., Chong, E., Rehg, J.M.: Fine-grained head pose estimation without keypoints. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 2074–2083 (2018) 11
41. Shen, Y., Yang, C., Tang, X., Zhou, B.: Interfacegan: Interpreting the disentangled face representation learned by gans. TPAMI (2020) 2, 4
42. Shen, Y., Zhou, B.: Closed-form factorization of latent semantics in gans. In: CVPR (2021) 2, 4
43. Shu, C., Wu, H., Zhou, H., Liu, J., Hong, Z., Ding, C., Han, J., Liu, J., Ding, E., Wang, J.: Few-shot head swapping in the wild. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10789–10798 (June 2022) 4
44. Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J.: High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 (2019) 8
45. Sun, Y., Zhou, H., Liu, Z., Koike, H.: Speech2talking-face: Inferring and driving a face with synchronized audio-visual representation. In: IJCAI. vol. 2, p. 4 (2021) 4
46. 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
47. Vemulapalli, R., Agarwala, A.: A compact embedding for facial expression similarity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5683–5692 (2019) 11
48. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5265–5274 (2018) 10
49. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2018) 8
50. Wang, T.C., Mallya, A., Liu, M.Y.: One-shot free-view neural talking-head synthesis for video conferencing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10039–10049 (2021) 4
51. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2021) 2, 4

52. Wang, Y., Chen, X., Zhu, J., Chu, W., Tai, Y., Wang, C., Li, J., Wu, Y., Huang, F., Ji, R.: Hiface: 3d shape and semantic prior guided high fidelity face swapping. *IJCAI* (2021) [1](#), [2](#), [4](#)
53. 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 (CVPR)*. pp. 7642–7651 (June 2022) [2](#)
54. Xu, Z., Hong, Z., Ding, C., Zhu, Z., Han, J., Liu, J., Ding, E.: Mobilefaceswap: A lightweight framework for video face swapping. *AAAI* (2022) [1](#)
55. Xu, Z., Yu, X., Hong, Z., Zhu, Z., Han, J., Liu, J., Ding, E., Bai, X.: Facecontroller: Controllable attribute editing for face in the wild. *AAAI* (2021) [1](#), [2](#), [4](#)
56. Yang, T., Ren, P., Xie, X., Zhang, L.: Gan prior embedded network for blind face restoration in the wild. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 672–681 (2021) [2](#), [3](#), [4](#), [6](#), [10](#)
57. Zhou, H., Sun, Y., Wu, W., Loy, C.C., Wang, X., Liu, Z.: Pose-controllable talking face generation by implicitly modularized audio-visual representation. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2021) [2](#), [4](#), [8](#), [13](#)
58. Zhu, H., Fu, C., Wu, Q., Wu, W., Qian, C., He, R.: Aot: Appearance optimal transport based identity swapping for forgery detection. *Advances in Neural Information Processing Systems* (2020) [1](#)
59. Zhu, P., Abdal, R., Femiani, J., Wonka, P.: Barbershop: Gan-based image compositing using segmentation masks. *arXiv preprint arXiv:2106.01505* (2021) [4](#)
60. 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) [2](#), [10](#), [11](#), [12](#)
61. 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) [20](#)

## Appendices

### A 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. 3. 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  $n$ th 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  $I_g^{r \rightarrow s} = G(\mathbf{F}_{att}^s, f_{id}^r)$ . Then it is taken as the target frame to generate the cycled-back image

$$I_g^{s \rightarrow r \rightarrow s} = G(\mathbf{E}_{att}(I_g^{r \rightarrow s}), \mathbf{W}_s^{+\{n\}}) \quad (9)$$

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:

$$\mathcal{L}_{id}(I_g, I_s) = 1 - D_{\cos}(f_{id}^s, f_{id}^g), \quad (10)$$

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

$$\mathcal{L}_{rec}(I_g, I_s) = \|I_g - I_s\|_1 + \sum_{m=1}^{N_{vgg}} \|\text{VGG}_m(I_g) - \text{VGG}_m(I_s)\|_1. \quad (11)$$

**Optimization Algorithm** The choice of the final optimized  $\mathbf{W}_s^+$  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-many* optimization aims to find a general  $\mathbf{W}_s^+$  that is suitable for swapping the identity of  $I_s$  to any face.

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  $\mathcal{L}_{id}(I_g^{s \rightarrow t}, I_s)$ . 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  $I_g^{s \rightarrow t} = G(\mathbf{E}_{att}(I_t), \mathbf{W}_s^+)$  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.

**Algorithm 1:** The algorithm of Swapping-Guided ID Inversion

---

**Input:** A set of images with random identities  $\{I_r^{\{1\}}, \dots, I_r^{\{N\}}\}$ ;  
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  $\mathbf{W}_s^+ = \mathbf{W}_s^{+\{1\}} = \{\mathbf{w}_{s(1)}, \dots, \mathbf{w}_{s(2L)}\}$ ,  $\hat{\mathcal{L}}_{id} = 100$  and  $n = 1$ .

**while**  $n \leq N$  **do**

$I_g^{r \rightarrow s} \leftarrow G(E_{att}(I_s), E_{id}(I_r^{\{n\}}));$   
 $I_g^{s \rightarrow r \rightarrow s} \leftarrow G(E_{att}(I_g^{r \rightarrow s}), \mathbf{W}_s^{+\{n\}});$   
 $\mathcal{L} \leftarrow \lambda_{rec} \mathcal{L}_{rec}(I_g^{s \rightarrow r \rightarrow s}, I_s) + \lambda_{id} \mathcal{L}_{id}(I_g^{s \rightarrow r \rightarrow s}, I_s);$   
 $I_g^{s \rightarrow t} \leftarrow G(E_{att}(I_t), \mathbf{W}_s^{+\{n\}});$   
**if**  $\mathcal{L}_{id}(I_g^{s \rightarrow t}, I_s) < \hat{\mathcal{L}}_{id}$  **then**  

$\mathbf{W}_s^+ \leftarrow \mathbf{W}_s^{+\{n\}};$   
 $\hat{\mathcal{L}}_{id} \leftarrow \mathcal{L}_{id}(I_g^{s \rightarrow t})$

 $\mathbf{W}_s^{+\{n+1\}} \leftarrow \mathbf{W}_s^{+\{n\}} - \eta * \mathcal{O}(\nabla_{\mathbf{W}} \mathcal{L});$   
 $n \leftarrow n + 1;$

---

Table 4: **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
<b>FF++ w/ Ours</b>	<b>88.16</b>	<b>70.96</b>	<b>73.59</b>	<b>70.07</b>	<b>72.13</b>

## B Experiments on Face Forgery Detection

We conduct face forgery detection experiments with backbone Xception [8] that has been widely used as baseline in previous face forgery detection methods [39, 26, 37]. The experiments are carried out on the following datasets. **1)** FaceForensics++ (FF++) [39] that has been introduced in the main paper. It is the most popular dataset used in face forgery detection. **2)** WildDeepfake (Deepwild) [61] contains 3805 real clips and 3509 fake clips. All these videos are manually collected from the Internet. **3)** Celeb-DF (CDF) [27] which contains high-quality face-swapped videos. **4)** DeepFake Detection Challenge (DFDC) [12] 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++ [39] without involving FaceShifter [25] data. Then we additionally 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 [A](#).

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.