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

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

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

summertight.github.io/selfswapper

**Fig. 1: Real-world application of our model.** This figure showcases results with the in-the-wild samples, navy circles for the source, pink circles for the target, and overlaps showing the generated outputs. The second row displays one-source multi-target results. Our model accurately transforms the target face to match the source while faithfully preserving the target attributes such as the skin color, pose, expression, hair, background, and gaze. This showcases our model's robustness on in-the-wild samples and real-world applicability for diverse facial images. For resolutions beyond $256 \times 256$, an off-the-shelf super-resolution model [39] is used.

**Abstract.** Face swapping has gained significant attention for its varied applications. Most previous face swapping approaches have relied on the seesaw game training scheme, also known as the target-oriented approach. However, this often leads to instability in model training and results in undesired samples with blended identities due to the target identity leakage problem. Source-oriented methods achieve more stable training with self-reconstruction objective but often fail to accurately reflect target image's skin color and illumination. This paper introduces the Shape Agnostic Masked AutoEncoder (SAMAE) training scheme, a novel

self-supervised approach that combines the strengths of both target-oriented and source-oriented approaches. Our training scheme addresses the limitations of traditional training methods by circumventing the conventional seesaw game and introducing clear ground truth through its self-reconstruction training regime. Our model effectively mitigates identity leakage and reflects target albedo and illumination through learned disentangled identity and non-identity features. Additionally, we closely tackle the shape misalignment and volume discrepancy problems with new techniques, including perforation confusion and random mesh scaling. SAMAE establishes a new state-of-the-art, surpassing other baseline methods, preserving both identity and non-identity attributes without sacrificing on either aspect.

**Keywords:** Face Swapping · Face Disentanglement · Face Forensic

## 1 Introduction

Face swapping has gained significant attention for its ability to create virtual human avatars, digitally resurrect individuals, and develop virtual models. This growing interest is matched by extensive research in academia, driven by advancements in deep generative models [13, 14, 17, 19]. Face swapping aims to seamlessly merge the identity characteristics of a source face with the non-identity features (*e.g.*, skin tone, pose, and lighting) of a target face, thereby crafting a cohesive and realistic facial image.

The majority of previous approaches [3,11,18,21,22,30,32,38,45] often employ the *seesaw game* training scheme, a multi-task learning strategy that leverages both reconstruction loss and identity loss [6] to achieve dual objectives: maintaining non-identity attributes of the target image and capturing identity of the source image. Known as the **target-oriented** approach, this scheme is particularly effective in face swapping, where ground-truth images are unavailable. This approach efficiently incorporates knowledge from pre-trained face recognition models [6] for reflecting identity information. Additionally, reconstruction losses such as L1 and Perceptual [15] ensure comprehensive preservation of the target image's attributes, such as skin color and illumination.

However, training methods that depend on two loss kinds with distinct goals without clear ground truth often lead to instability and unreliability in the model. The conflicting nature of these losses can lead to unstable gradient flow [18], misleading the models to reflect imprecise identity information and reducing model consistency across different samples. Simple reconstruction losses may cause unwanted mixing of source and target identities, resulting in the identity leakage from the target image. Carefully tuned hyperparameters are needed to balance these competing objectives, which complicates the training process.

On the contrary, the recent **source-oriented** approaches [25, 27] adopt a self-supervised framework, incorporating a straightforward self-reconstruction task along with established head reenactment models [33, 36, 37] to morph the source image into the target pose. It then seamlessly integrates the reenacted

source image with the target background at the pixel level. However, its self-supervised training strategy overlooks the cross-identity inference phase, leading to difficulties with generalization. For instance, the model often transfers the color attributes of the source to the generated output, which, along with target body's heterogeneous skin tone and illumination, leads to unrealistic results.

Moreover, addressing shape and volume misalignment between the source and target face is crucial. The facial shape significantly influences facial identity, and precise alignment of facial volume is essential for achieving realistic results. Many previous methods overlook this aspect, and even those that consider it, such as DiffSwap [45], yield unsatisfactory outcomes due to vague guidance provided during contour generation in the seesaw game training scheme.

In this paper, we introduce the Shape Agnostic Masked AutoEncoder (SAMAE), a framework that combines the strengths of both target-oriented and source-oriented approaches. The SAMAE employs self-supervised training, which introduces clear ground truth through its self-reconstruction training regime, effectively addressing the conventional seesaw game issue. Additionally, it achieves high generalization performance including accurate incorporation of target skin color and illumination in the cross-view inference phase, where previous target-oriented methods have faltered. SAMAE also adeptly handles shape and volume misalignment problems.

This generalizable self-supervised training is achieved through 1) the model design and 2) novel techniques including *perforation confusion* and *random mesh scaling*. Our model features learnable modules that disentangle identity and non-identity features, along with a combination of pretrained models like 3D Morphable Models (3DMMs) [8, 28] and face identity encoders [6]. Our design also allows for accurate skin color and illumination extraction from the target.

We introduce perforation confusion and random mesh scaling to enhance the cross-identity inference capability. Perforation confusion creates shape-agnostic masks during training, addressing the shape misalignment problem. Random mesh scaling is designed to mitigate the model's excessive dependence on pixel-aligned information from the conditioned mesh, enabling the model to better manage volume discrepancies between the source and target faces.

In summary, our contributions are:

- We introduce a novel training regime of the Shape-Agnostic Masked AutoEncoder (SAMAE) in face swapping, which effectively eliminates the unstable *seesaw game* and addresses the problem of target identity leakage through self-supervised learning, while disentangling target albedo and illumination.

- We address the shape-misalignment and facial volume discrepancy problem with two innovative yet straightforward techniques: *perforation confusion* and *random mesh scaling*.

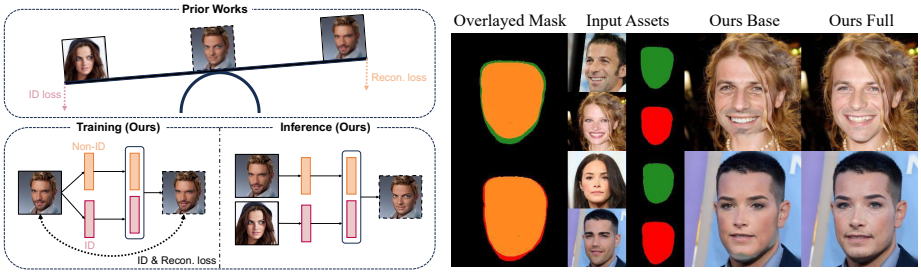



**Fig. 2:** (A) **Conceptual comparison between prior works and our method.** Prior works rely on a seesaw game of two potentially conflicting losses: reconstruction loss and identity loss. On the other hand, our method leverages a self-supervised approach with a clear ground truth, which allows for more stable training. (B) **Comparing our base approach (Ours Base) with our enhanced method (Ours Full)**, which includes techniques like perforation confusion and random mesh scaling. Green masks represent target-posed source 3DMM masks, red masks indicate target 3DMM masks, and orange masks denote their intersection. The first row shows that when the source face is larger than the target's, the jaw is cut off. The second row shows the opposite case, where the base model fails to inpaint the remaining regions effectively, while Ours Full generates realistic face-swapped outputs.

– Our new training approach establishes a new state-of-the-art, surpassing other baseline methods in terms of both identity and non-identity attributes without sacrificing on either aspect.

## 2 Related Work

**Target-Oriented Face Swapping.** The target-oriented face swapping refers to a group of face swapping methods that operate by manipulating the target spatial features and leveraging the off-the-shelf face recognition [6] models. This approach inherently relies on seesaw game between reconstruction loss and identity loss. DeepFakes [5] initially introduced an autoencoder-based algorithm capable of swapping faces between two specifically trained identities, but it lacked generalization capabilities. Faceshifter [22] proposes a Adaptive Embedding Intergration Network (AEI-Net) for merging target non-identity attributes with source identity. SimSwap [3] uses Weak Feature Matching (WFM) loss to maintain target attributes. InfoSwap [11] applies an information bottleneck principle to filter out target identity information. HifiFace [38] adopts a 3DMM parameter-based swapping method. SmoothSwap [18] focuses on identity space smoothness, improving source identity preservation at the cost of target attribute fidelity. MegaFS [46] and FSLSD [40] utilize StyleGAN for high-resolution outputs and structure disentanglement, respectively. RobustSwap [21] and BlendFace [32] address attribute leakage in face swapping. DiffSwap [45] tackles face swapping as an inpainting problem, leveraging

the latent diffusion models [31]. The work tackles the shape misalignment problem by using convex-hulls of keypoints from both source and target as perforation masks. However, it fails to explicitly disentangle facial contour information from the input conditions due to their seesaw training regime. WSC-swap [30] critiques the source-target disentanglement in existing methods. However, all these methods still engage in a seesaw game with the agenda of disentangling source-target information. In contrast, we propose a brand-new training paradigm for face swapping that is free from this seesaw game.

**Source-Oriented Face Swapping.** Face swapping methods with a source-oriented approach frequently integrate classical point mapping techniques [34] or employ face reenactment models [33, 36, 37] to warp the source image to match the pose of the target and blend the source face with target's background. DeepFaceLab [29] belongs to the former camp, employing similarity transformation-based warping and pasting followed by sharpening and blending strategies in post-processing. These methods typically excludes the target's facial region during network training, free from target identity leakage. FSGAN [27], belonging to the latter category, integrates face swapping and reenactment in a two-stage process. This is followed by the use of a face inpainting network to blend the reenacted source with the target images. The most recent work in the latter category, E4S [25], introduced a regional GAN inversion approach along with face reenactment and mask editing. However, these methods, dependent on reenactment models and point mapping methods, often struggle to capture the target's illumination and can result in identity shifts, particularly when there are large pose discrepancies between the source and the target. It arises from the fact that, despite their approaches originating from the self-supervised manner, their training schema is not enough to consider test phase scenarios. Furthermore, they show inferior performance when the source image includes self-occlusions or accessories.

**Face Disentanglement Learning.** DiscoFaceGAN [7] disentangles latent space of StyleGAN [17] into 3DMM parameters for controlled generation. GIF [12] combines 3DMM-rendered assets with StyleGAN's noise space to control expression and lighting. CONFIG [20] differentiates between real image latents and 3DMM parameters, while VariTex [1] uses variational space-based UV projection for manipulation. 3DFM-GAN [24] refines StyleGAN's latent space for controllable manipulation. However, these methods aggregate facial and non-facial elements into identity, complicating their use in face swapping.

## 3 Backgrounds

**Essentials of Face Swapping.** High-fidelity face-swapped results, $\hat{I}^{swap}$, should adhere to the following conditions, inheriting identity from a source image $I^{src}$ and non-identity attributes from a target image $I^{tgt}$:

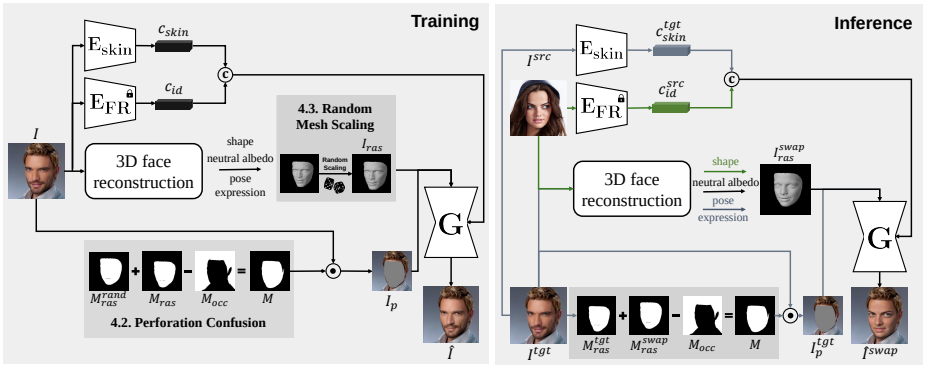



**Fig. 3: Overall pipeline of our method.** In the training phase (left), we employ self-reconstruction scheme with perforation confusion and random mesh scaling, enhancing shape agnostic robustness for SAMAE training. During inference (right), this training enables the model to efficiently perform cross-identity face swapping by disentangling ID and non-ID attributes.

- C1. They should preserve the non-identity attributes of $I^{tgt}$, which include non-facial attributes (*e.g.*, background and hair), facial posture (*e.g.*, expression and pose), and facial color (*e.g.*, skin color and lighting).
- C2. They must maintain the identity features such as facial contour, inner facial traits (*e.g.*, eyes, eyebrows, nose, lip, cheekbone, dimples, and their interrelations), and skin details (*e.g.*, freckles, moles, and wrinkles) of the source image $I^{src}$.
- C3. The resulting images should be indistinguishable from real images.

**3D Morphable Model (3DMM).** We employ an off-the-shelf 3DMM [28] and 3DMM parameter estimator $E_{3DMM}$ [8], coupled for 3D face reconstruction. The $E_{3DMM}$ estimates a tuple of the lighting and geometric parameters $v = (\alpha, \beta, \gamma, \delta, R, t_{xy}, s)$ of a facial image. The $v$ is composed of shape $\alpha$, expression $\beta$, albedo $\gamma$, lighting $\delta$, and camera parameters including rotation matrix $R$, xy-axial translation parameter $t_{xy}$ and object scale $s$.[1] The renderer $Rd$ outputs the rasterized face mesh image $I_{ras} = Rd(v)$. The foreground mesh mask $M_{ras}$ is automatically derived from the $I_{ras}$.

## 4 Method

### 4.1 Shape Agnostic Masked AutoEncoder

The goal of the Shape Agnostic Masked AutoEncoder (SAMAE) training regime is to reconstruct the ground truth image $I$ given the corresponding decomposed

[1] In most 3DMMs, as they adopt an orthographic camera model, there is no z-axis translation $t_z$.

components, which include identity (C2) and non-identity attributes (C1). To extract the non-facial attributes from the image, we utilize a foreground mesh mask $M_{ras}$ and an occlusion mask $M_{occ}$, obtained from the-off-the-shelf face parser [42]. This strategy is the most straightforward way to exclude the target facial identity information, aligning with the principles of source-oriented methods. By removing the occluded region $M_{occ}$ from $M_{ras}$, we construct the final facial mask $M$ as $M = M_{ras} - M_{occ}$. The non-facial attribute image is then defined by $I_p = (\mathbb{1} - M) \odot I$, where $\odot$ represents the Hadamard product and $\mathbb{1}$ is a 1-filled tensor matching the dimension of $M$.

The facial posture and facial color are represented by $(\beta, R)$ and $(c_{skin}, \delta)$, respectively. $c_{skin}$ is derived from the skin area of the image $I$ using the skin color encoder $E_{skin}$ (Sec. 4.4). For the identity, the facial contour can be expressed by the facial mesh outline of $I_{ras}$, and the inner facial traits and the skin details are represented by the combination of the shape parameter $\alpha$ and identity embedding $c_{id}$, where $c_{id}$ is extracted from the pretrained face recognition model $E_{FR}$.

These features are conditioned to the U-Net based generator $G$ [9] which outputs the reconstructed image $\hat{I}$:

$$\hat{I} = G(I_{ras}, I_p, c_{id}, c_{skin}) = G(Rd(v), I_p, E_{FR}(I), E_{skin}(I)), \tag{1}$$

where $v$ here is an estimated 3DMM parameter with albedo $\gamma$ replaced to a neutralized albedo $\gamma_{neu}$, which will be discussed in Sec. 4.4.

Switching to the cross-identity inference (swap) regime, we condition the model with the identity-relevant information extracted from the source image $I^{src}$ to generate the swapped result $\hat{I}^{swap}$,

$$\begin{aligned} \hat{I}^{swap} &= G(I^{swap}_{ras}, I^{tgt}_{p}, c^{src}_{id}, c^{tgt}_{skin}) \\ &= G(Rd(v^{swap}), I^{tgt}_{p}, E_{FR}(I^{src}), E_{skin}(I^{tgt})), \end{aligned} \tag{2}$$

where $v^{swap} = (\alpha^{src}, \beta^{tgt}, \gamma_{neu}, \delta^{tgt}, R^{tgt}, t^{tgt}_{xy}, s^{tgt})$ and superscripts *src* and *tgt* indicates whether the features are from $I^{src}$ or $I^{tgt}$. Furthermore, since 3DMMs offer limited representation for eye-gazing, we leverage the positions of the target image's iris keypoints from [4]. Keypoints are represented as a stickmen-like rendering (as described in [44]), and we utilize it as a spatial input of the model, concatenating it with $I_{ras}$ and $I_p$. For brevity, we omit notions of this input in this manuscript. For overall pipeline of training and inference, please refer to Fig. 3.

### 4.2 Perforation Confusion

In the training regime, $M$ and $I_{ras}$ can both carry the same information of the facial contour of $I$. This means that the model can attend to either $M$ or $I_{ras}$ to reconstruct the facial contour of $I$. However, in the inference phase, the final mask $M$ is constructed from the target image $I^{tgt}$, whereas $I^{swap}_{ras}$ contains the facial contour of the source image. Since facial contour is part of the identity (C2), $I_{ras}$ (or $I^{swap}_{ras}$ in the inference phase) should be the sole factor for carrying the contour information.

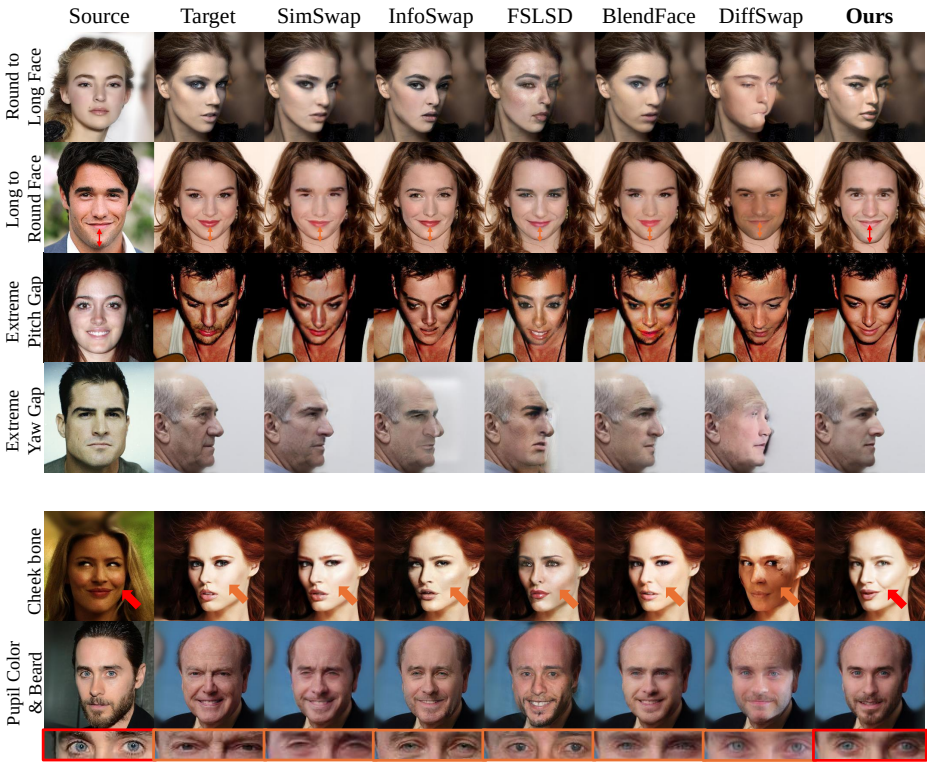



**Fig. 4: Comparison among target-oriented baselines**. (Top) Other baselines struggle to replicate the source's facial features such as facial contours and volumes (*e.g.*, jaw shape and facial scale) and (Bottom) inner facial traits (*e.g.*, pupil color, beard, and cheekbone). In contrast, **Ours** conveys these with high-fidelity. Pay attention to the red and orange indicators for detailed comparison.

To address this problem, we introduce a novel technique called *perforation confusion* which randomly augments the final mask $M$ so that the model learns not to retrieve any contour related information from the masked image $I_p$ (or $I_p^{tgt}$ in the inference phase). Specifically, the perforation confusion is performed by the following sequence:

1. Sample a random shape parameter $\alpha^{rand}$ from a facial image dataset [17].
2. Replace $\alpha$ of $v$ to $\alpha^{rand}$ to construct a random 3DMM parameter tuple $v^{rand} = (\alpha^{rand}, \beta, \gamma_{neu}, \delta, R, t_{xy}, s)$, and extract a random mask $M_{ras}^{rand}$ from $I_{ras}^{rand} = Rd(v^{rand})$.
3. $M_{ras}$ is summated with $M_{ras}^{rand}$ and exclude non-facial region $M_{occ}$ to generate the final mask $M = M_{ras} + M_{ras}^{rand} - M_{occ}$.

During the inference phase, it's important to highlight that we substitute $M_{ras}^{rand}$ with $M_{ras}^{swap}$ which is directly derived from the foreground region of $I_{ras}^{swap}$, and

**Table 1: Quantitative comparison** with both target-oriented and source-oriented baselines. **Bold** text highlights the best scores. Our method outperforms other baselines in Identity Similarity (ID. Sim.), Identity Consistency (ID. Cons.), Head Pose, and Fréchet Inception Distance (FID). Expression distance is on par with the best performing models.

| Categories | Methods | ID. Sim.↑ | ID. Cons.↑ | Shape ↓ | Expression↓ | Head Pose↓ | FID↓ |
|---|---|---|---|---|---|---|---|
| **Target-Oriented** | SimSwap [3] | 0.525 | 0.543 | 0.128 | 0.204 | 0.014 | 26.77 |
| | InfoSwap [11] | 0.527 | 0.583 | 0.126 | 0.233 | 0.019 | 32.21 |
| | FSLSD [40] | 0.330 | 0.345 | 0.129 | 0.207 | 0.025 | 39.71 |
| | BlendFace [32] | 0.440 | 0.510 | 0.136 | **0.189** | **0.013** | 23.11 |
| | DiffSwap [32] | 0.347 | 0.361 | 0.156 | 0.224 | 0.028 | 59.98 |
| **Source-Oriented** | FSGAN [27] | 0.338 | 0.403 | 0.164 | 0.193 | 0.016 | 42.56 |
| | E4S [25] | 0.501 | 0.588 | 0.118 | 0.262 | 0.028 | 52.90 |
| **SAMAE** | **Ours** | **0.578** | **0.628** | **0.108** | 0.190 | **0.013** | **21.22** |

the final mask is generated as $M = M_{ras}^{tgt} + M_{ras}^{swap} - M_{occ}$. Note that $M_{ras}^{tgt}$ is a foreground mesh mask derived from the target facial mesh $I_{ras}^{tgt}$ of the target image $I^{tgt}$. The necessity of the perforation confusion technique for handling the shape-misalignment problem of source and target face is demonstrated by the results in Fig. 2 (B) and Sec. 5.3.

### 4.3 Random Mesh Scaling

Empirically, during the inference phase, we observed that when the facial volume of the source is significantly smaller (or larger) than that of the target, the swapped face appears awkward, either shrunken or dilated. This issue arises because the model, trained to self-reconstruct the input image with a consistently sized facial region, tends to over-rely on the pixel-aligned information from $I_{ras}$. This reliance hinders generalization to cross-identity inferences.

To handle this problem, we propose the *random mesh scaling* technique, which allows the model to generate realistic face images using randomly scaled $I_{ras}$, thereby enhancing the model's ability to generalize to facial priors of varying scales during inference. Specifically, in the train regime, the scale parameter $s$ is substituted with a random scale parameter $s^{rand} \sim U(-4, 1)$, leading to randomly scaled $I_{ras}$. It is important to note that random mesh scaling is not used during the inference phase; instead, the estimated target scale $s^{tgt}$ is utilized.

### 4.4 Disentangling Albedo Condition

The estimated 3DMM albedo parameter $\gamma$ contains both non-identity attributes such as skin color (C1), and identity features, including skin details and certain inner facial traits like eye color (C2). Due to the entangled nature of these attributes within $\gamma$, it is unsuitable for our task that requires distinct separation of identity and non-identity features. Given the necessity of the albedo parameter in rendering $I_{ras}$, we have devised a workaround. We transform $\gamma$ into its

neutralized form, $\gamma_{neu}$, which incorporates a white-colored albedo map. This modification eliminates any albedo-related information from $I_{ras}$. Furthermore, $I_{ras}$ undergoes min-max normalization to isolate the remaining brightness differences, which serve as illumination information in a facial image.

Instead of relying on albedo parameter $\gamma$, we find that using identity embeddings $c_{id}$ extracted from $E_{FR}$ is sufficient, as they contain rich identity information. However, directly using these can lead to unintended transfer of skin color from the source face, which should be derived from the target face. To address this, we apply random color-jittering to the image before processing it with $E_{FR}$, encouraging our generator to disregard skin color information in $c_{id}$.

In parallel, we train an additional trainable encoder $E_{skin}$ to capture skin color information as a vector-formed neural albedo. Specifically, we mask out non-skin areas from $I$ (or $I^{tgt}$ during inference) with an off-the-shelf face parser [42], and encode the masked image with $E_{skin}$. To prevent $c_{skin}$ from incorporating irrelevant features such as skin detail or facial geometry, we empirically choose a low-dimensional embedding, ensuring it has only the necessary capacity to capture skin color information. More details and experiments are provided in the supplementary materials.

### 4.5 Training Objectives

We employ reconstruction losses (L1 loss and Perceptual loss [15]) and an identity loss [6] between the estimated image $\hat{I}$ and the ground-truth image $I$. To enhance the realism of the output image, we also incorporate a non-saturating adversarial loss [17]. For additional details, please refer to the supplementary materials.

## 5 Experiments

**Datasets.** The model is trained on FFHQ dataset [17] with images at a resolution of $256 \times 256$. For evaluation, we use 1K randomly sampled source-target pairs, from the CelebA-HQ dataset [16]. It is noteworthy that, unlike our method, numerous approaches [22,30,41] extend their training datasets beyond FFHQ by integrating multiple datasets, including identity-labeled datasets and/or video datasets [2,26] to stabilize the training. Our method, on the other hand, delivers state-of-the-art performance without such additions.

**Baselines.** For both quantitative and qualitative comparisons, we establish our baselines with FSGAN [27], SimSwap [3], InfoSwap [11], FSLSD [40], E4S [25], DiffSwap [45], and BlendFace [32]. We exclude HifiFace [38] and FaceShifter [22] from our main analysis, as they lack official open-sourced codes. Additionally, in the supplementary materials, we present comparisons with other methods such as AFS [35], MegaFS [46], and ReliableSwap [43], which either demonstrate subpar quality or have not been accepted to the academic community.

**Implementation Details.** Our model is trained on two NVIDIA RTX 3090 GPUs with a batch size of 8, completing 500k iterations in about 6 days. We use the Adam optimizer with a learning rate of $2 \times 10^{-4}$ for both the generator

and the discriminator. We leverage the ADM [9] U-Net for the generator, and StyleGAN2 [17] architecture for the discriminator.

### 5.1 Qualitative Comparisons

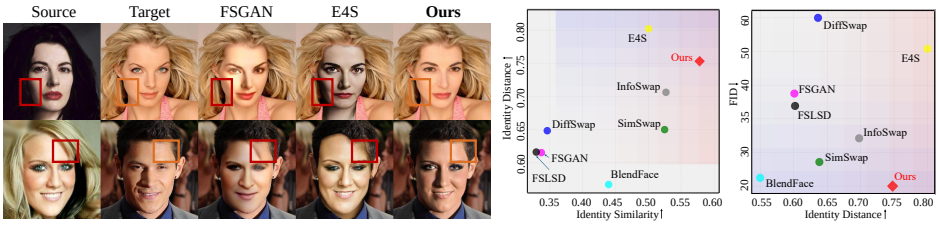
We categorize the baselines into two groups: target-oriented approaches including SimSwap [3], InfoSwap [11], FSLSD [40], BlendFace [32], and DiffSwap [45] and source-oriented approaches FSGAN [27] and E4S [25].

**Target-oriented Baselines.** Target-oriented baselines directly employ reconstruction losses between generated outputs and target images, effectively preserving target attributes like lighting conditions, expressions, and head poses, as shown in Fig. 4. However, this straightforward application of reconstruction loss often leads to the leakage of the target's identity, resulting in a blend of source and target identities. These models frequently struggle to accurately represent the source's facial contours and skin details. In contrast, our model moves beyond this trade-off, adeptly reproducing the source's facial contours, skin details, and inner facial features, while still maintaining the target's non-facial attributes, and poses. Further results can be found in the supplementary materials.

**Source-oriented Baselines.** We present a comparison of our method with source-oriented baselines in Fig. 5a. These methods involve merging the reenacted source face with the target image, effectively reducing the risk of target identity leakage. However, their effectiveness is limited by the reenactment models' performance, often leading to inaccuracies in source identity preservation and difficulties in handling pose variations. Furthermore, the blending process can result in unnatural skin tones and inadequate replication of the target's lighting, sometimes introducing noticeable artifacts. For example, the shadow from the source image can be carried over into the swapped image, as depicted in the second row of Fig. 5a. In contrast, our method is adept at matching the target's skin color and lighting conditions, producing outputs that blend seamlessly with the target images.

### 5.2 Quantitative Comparisons

We conduct a quantitative comparison of our method against both target-oriented and source-oriented approaches. The evaluation involves several metrics: Identity Similarity (ID. Sim.), Identity Consistency (ID. Cons.), Shape, Expression and Head Pose distance, and the Fréchet Inception Distance (FID) score. Identity Similarity (ID. Sim.) calculates the cosine similarity between the identity embeddings of the source and swapped images, using a separate face recognition model [6] that was not used in the training phase. Identity Consistency (ID. Cons.) [30] measures the consistency of identity among the swapped images, using one source image and various target images. Further, we measure the Expression and Head Pose Distance between the target and swapped faces, as well as the Shape Distance between the source and the swapped results, using a 3DMM predictor [10] that was also excluded from the training process. The FID score evaluates the overall quality of the images.




**(a) Comparison among source-oriented baselines.** These baselines exhibit issues with leakage of the source's illumination. Observe red and orange indicators. Our model effectively avoids source illumination leakage, thanks to our method's finely disentangled features.

**(b)** 2-dimensional graph comparison on Identity Similarity and Distance (left), where positioning in the upper-right corner signifies a good model. In the Identity Distance - FID graph (right), a location in the lower-right side indicates favorable model performance.

**Fig. 5:** Comparison among source-oriented baselines and 2-dimensional graph comparison on Identity scores and FID.

As shown in Table 1, our method achieves the state-of-the-art quality without sacrificing certain metrics, unlike other methods. InfoSwap ranks second-best in terms of ID. Sim. but falls behind in other metrics. While DiffSwap tackles the shape misalignment problem, the Shape Distance results suggest shortcomings in the method. Regarding Expression distance, although BlendFace achieves the highest score, it exhibits a lower ID. Sim. score. Additionally, its low ID. Cons. score suggests a leakage of target attributes. This indicates that BlendFace potentially compromises identity preservation in favor of maintaining the target's pose and expression.

Also to comprehensively evaluate our model's performance, including aspects such as target identity leakage and image realism, we present a two-dimensional graph in Fig. 5b. In Fig. 5b (left), the $x$-axis shows the identity similarity score between the source and the generated images, while the $y$-axis indicates the identity distance between the target and the generated images. Optimal models without target identity leakage should exhibit both high identity similarity scores and high identity distances, positioning them in the upper-right corner of the graph. The figure implies that our method excels in reflecting the source identity and preventing target identity leakage. E4S is also positioned at the upper-right corner of the figure. However, right panel of Fig. 5b indicates that E4S yields the lowest FID score, suggesting it creates source-like images with reduced target identity leakage, but with poor realism. In contrast, our model is situated in the bottom-right corner of the graph, demonstrating superior performance in image realism and robustness against target identity leakage, outperforming other baselines.

### 5.3 Ablation Study

We conducted an ablation study on various components of our method, including perforation confusion, random mesh scaling, and disentangled albedo conditions. The results are shown in Fig. 6a. Column (A) displays the base model

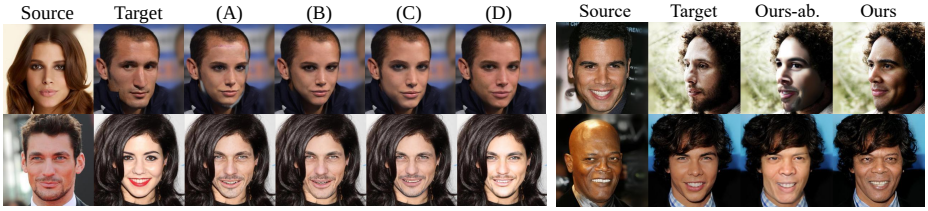
**Table 2: Quantitative evaluation in ablation study.** (B) indicates Ours Base. Perforation confusion (P) enhances realism (FID), reducing unpainted regions or cut-off facial contours. Random mesh scaling (R) further boosts the realism and identity scores. Incorporating skin color condition (S) ensures skin color and illumination closely match the target images, yielding naturally blended images, and optimally improves all metrics.

| Methods | ID. Sim. ↑ | Expression ↓ | Head Pose ↓ | FID ↓ |
|---|---|---|---|---|
| B | 0.570 | 0.208 | 0.014 | 24.05 |
| B+P | 0.568 | 0.209 | 0.014 | 23.92 |
| B+P+R | 0.573 | 0.220 | 0.015 | 22.99 |
| B+P+R+S | **0.578** | **0.190** | **0.013** | **21.22** |

without these techniques. Without perforation confusion, the model struggles with cross-identity swapping and fails to properly inpaint the gaps caused by shape misalignment between the source and target faces. With perforation confusion, as shown in column (B), the model handles the shape misalignment and successfully fills non-facial areas, such as the neck. However, this column also reveals that without random mesh scaling, the facial volume in the target image is not preserved, leading to swapped images with unnaturally shrunken or enlarged faces. Random mesh scaling allows the model to adaptively fit the target priors, maintaining the facial volume of the target image (column (C)). Finally, matching the skin color of the generated results with the target images is crucial for seamless blending. By employing the disentangled skin color embedding $c_{skin}$, we see in column (D) that the skin color and illumination of the swapped faces closely match those of the target, enhancing the realism of the blended images.

Also to assess the effectiveness of SAMAE's self-supervised training approach compared to the widely used seesaw game training regime, we conducted a comparison against a model trained using a multi-task learning strategy similar to target-oriented methods. Specifically, the model generates a swapped image from randomly selected source and target images, and both the identity loss and reconstruction loss are applied to the output. We utilized the same network architecture and hyperparameters as our original model for this experiment. The resulting output generally exhibits blurriness and incorrect identity reflectance and illumination, as demonstrated in Fig. 6b, labeled as Ours-ab. We also provide quantitative evaluation results in the supplementary materials.

We also compare two different conditioning methods: "parameter" and "mesh". The mesh condition, as described in our paper, uses $I_{ras}^{swap}$ for the geometry condition, fully utilizing 3DMM renderer and thus providing the model with enhanced spatial information. The parameter method directly employs 3DMM coefficients $v^{swap}$ in vector form without rendering the mesh. As shown in the first row of Fig. 7a, we find that the mesh condition is vital for preserving the source's facial features. Furthermore, the results in the second row indicate that the mesh condition more accurately guides the model in learning the head pose.




**(a) Qualtitative evaluation in ablation study.** (A) Ours Base, (B) Ours Base + perforation confusion (P), (C): Ours Base + perforation confusion (P) + random mesh scaling (R), and (D): Ours Base + perforation confusion (P) + random mesh scaling (R) + skin color condition (S).

**(b)** Qualtitative comparison between the seesaw game training scheme (Ours-ab.) and the self-supervised training scheme (Ours).

**Fig. 6:** Qualitative results for various ablation studies.


**(a) Another option on the geometry condition.** Models trained with direct 3DMM parameters in vector form, without rendering (Parameter), often fail to accurately capture the precise facial geometries of the source.

**(b) Limitations**. When 3DMM estimators are unable to accurately capture exaggerated expressions or tongue, as shown in the "Mesh" visualization, our model is constrained by the mesh condition, reflecting these limitations.

**Fig. 7:** Selection of geometry conditions and limitations of our model.

On the other hand, the parameter condition results in an artifact on the nose with messy geometry.

## 6 Discussion

**Limitations and Future Work.** As can be seen in Fig. 7b, our model sometimes fails to capture detailed expressions, such as exaggerated grin, due to the limited representation capacity of 3DMM. Advanced 3DMMs and their prediction models [10, 23] can be utilized for future works. We posit that incorporating strong generative priors like StyleGAN or the latest diffusion models into our training pipeline could enhance the face swapping quality and will be an interesting research direction.

**Ethical Considerations.** Face swapping is useful in areas like digital resurrection and telepresence but also poses risks of privacy invasion and misinformation. We are dedicated to prevent the potential misuse our model, and plan to release our model exclusively for research purposes. Additionally, we will provide a benchmark dataset to support research in face forensics and privacy protection.

**Acknowledgements.** This work was supported by Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II190075 Artificial Intelligence Graduate School Program(KAIST)), the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. NRF-2022R1A2B5B02001913), and KAIST-NAVER hypercreative AI center.

## References

1. Bühler, M.C., Meka, A., Li, G., Beeler, T., Hilliges, O.: Varitex: Variational neural face textures. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13890–13899 (2021)
2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). pp. 67–74. IEEE (2018)
3. 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. digital-nomad cheng: Dense iris landmarks. `https://github.com/digital-nomad-cheng/Iris_Landmarks_PyTorch`, accessed: 2024-03-07
5. deepfakes: Deepfakes. `https://github.com/deepfakes/faceswap`, accessed: 2024-03-07
6. 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)
7. Deng, Y., Yang, J., Chen, D., Wen, F., Tong, X.: Disentangled and controllable face image generation via 3d imitative-contrastive learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5154–5163 (2020)
8. Deng, Y., Yang, J., Xu, S., Chen, D., Jia, Y., Tong, X.: Accurate 3d face reconstruction with weakly-supervised learning: From single image to image set. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. pp. 0–0 (2019)
9. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
10. Feng, Y., Feng, H., Black, M.J., Bolkart, T.: Learning an animatable detailed 3d face model from in-the-wild images. ACM Transactions on Graphics (ToG) **40**(4), 1–13 (2021)
11. 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)
12. Ghosh, P., Gupta, P.S., Uziel, R., Ranjan, A., Black, M.J., Bolkart, T.: Gif: Generative interpretable faces. In: 2020 International Conference on 3D Vision (3DV). pp. 868–878. IEEE (2020)
13. 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)
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

15. 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)
16. 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)
17. 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)
18. 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)
19. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
20. Kowalski, M., Garbin, S.J., Estellers, V., Baltrušaitis, T., Johnson, M., Shotton, J.: Config: Controllable neural face image generation. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 299–315. Springer (2020)
21. Lee, J., Kim, T., Park, S., Lee, Y., Choo, J.: Robustswap: A simple yet robust face swapping model against attribute leakage. arXiv preprint arXiv:2303.15768 (2023)
22. 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)
23. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. ACM Transactions on Graphics, (Proc. SIGGRAPH Asia) **36**(6), 194:1–194:17 (2017), `https://doi.org/10.1145/3130800.3130813`
24. Liu, Y., Shu, Z., Li, Y., Lin, Z., Zhang, R., Kung, S.: 3d-fm gan: Towards 3d-controllable face manipulation. In: European Conference on Computer Vision. pp. 107–125. Springer (2022)
25. 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)
26. Nagrani, A., Chung, J.S., Xie, W., Zisserman, A.: Voxceleb: Large-scale speaker verification in the wild. Computer Speech & Language **60**, 101027 (2020)
27. 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)
28. Paysan, P., Knothe, R., Amberg, B., Romdhani, S., Vetter, T.: A 3d face model for pose and illumination invariant face recognition. In: 2009 sixth IEEE international conference on advanced video and signal based surveillance. pp. 296–301. Ieee (2009)
29. 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)
30. Ren, X., Chen, X., Yao, P., Shum, H.Y., Wang, B.: Reinforced disentanglement for face swapping without skip connection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20665–20675 (2023)
31. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)

32. Shiohara, K., Yang, X., Taketomi, T.: Blendface: Re-designing identity encoders for face-swapping. arXiv preprint arXiv:2307.10854 (2023)
33. Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., Sebe, N.: First order motion model for image animation. Advances in neural information processing systems **32** (2019)
34. Umeyama, S.: Least-squares estimation of transformation parameters between two point patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence **13**(4), 376–380 (1991). https://doi.org/10.1109/34.88573
35. Vu, T., Do, K., Nguyen, K., Than, K.: Face swapping as a simple arithmetic operation. arXiv preprint arXiv:2211.10812 (2022)
36. 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)
37. 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)
38. 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. arXiv preprint arXiv:2106.09965 (2021)
39. Wang, Z., Zhang, J., Chen, R., Wang, W., Luo, P.: Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 17512–17521 (2022)
40. 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)
41. Xu, Z., Zhou, H., Hong, Z., Liu, Z., Liu, J., Guo, Z., Han, J., Liu, J., Ding, E., Wang, J.: Styleswap: Style-based generator empowers robust face swapping. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV. pp. 661–677. Springer (2022)
42. Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N.: Bisenet: Bilateral segmentation network for real-time semantic segmentation. In: Proceedings of the European conference on computer vision (ECCV). pp. 325–341 (2018)
43. Yuan, G., Li, M., Zhang, Y., Zheng, H.: Reliableswap: Boosting general face swapping via reliable supervision. arXiv preprint arXiv:2306.05356 (2023)
44. Zakharov, E., Shysheya, A., Burkov, E., Lempitsky, V.: Few-shot adversarial learning of realistic neural talking head models. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9459–9468 (2019)
45. 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)
46. 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)