Identity-Preserving Face Swapping via Dual Surrogate Generative Models

ZIYAO HUANG, Institute of Computing Technology, Chinese Academy of Sciences, China
FAN TANG*, Institute of Computing Technology, Chinese Academy of Sciences, China
YONG ZHANG, AI Lab, Tencent, China
JUAN CAO, Institute of Computing Technology, Chinese Academy of Sciences, China
CHENGYU LI, Institute of Computing Technology, Chinese Academy of Sciences, China
SHENG TANG, Institute of Computing Technology, Chinese Academy of Sciences, China
JINTAO LI, Institute of Computing Technology, Chinese Academy of Sciences, China
TONG-YEE LEE, National Cheng Kung University, Taiwan

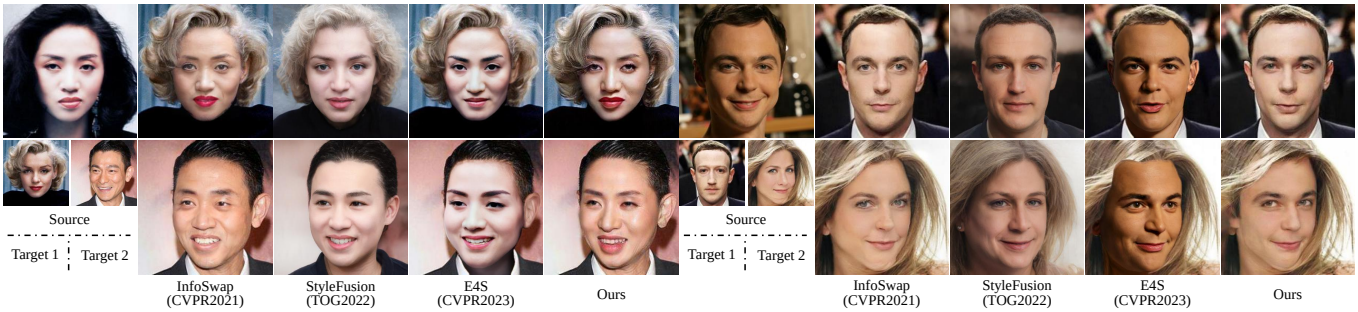


Fig. 1. Face-swapped images generated by our method, E4S [Liu et al. 2023], StyleFusion [Kafri et al. 2022], and InfoSwap [Gao et al. 2021]. From top-to-down and left-to-right, female-to-female, female-to-male, male-to-male, and male-to-female swapping conditions are shown. We are superior to these high-fidelity face-swapping models in terms of identity preserving.

In this study, we revisit the fundamental setting of face-swapping models and reveal that only using implicit supervision for training leads to the difficulty of advanced methods to preserve the source identity. We propose a novel reverse pseudo-input generation approach to offer supplemental data for training face-swapping models, which addresses the aforementioned issue. Unlike the traditional pseudo-label-based training strategy, we assume that arbitrary real facial images could serve as the ground-truth outputs for the face-swapping network and try to generate corresponding input <source, target> pair data. Specifically, we involve a source-creating surrogate that alters the attributes of the real image while keeping the identity, and a target-creating surrogate intends to synthesize attribute-preserved target images with different identities. Our framework, which utilizes proxy-paired data as explicit supervision to direct the face-swapping training process, partially fulfills a credible and effective optimization direction to boost the identity-preserving capability. We design explicit and implicit adaption strategies to better approximate the explicit supervision for face swapping. Quantitative and qualitative experiments on FF++, FFHQ, and wild images show that our framework could improve the performance of various face-swapping pipelines in terms of visual fidelity and ID preserving. Furthermore, we display applications with our method on re-aging, swappable attribute customization, cross-domain, and video face swapping.

CCS Concepts: • **Computing methodologies** → **Image manipulation**.

Additional Key Words and Phrases: Face swapping, image editing, digital face synthesis.

ACM Reference Format:
Ziyao Huang, Fan Tang, Yong Zhang, Juan Cao, Chengyu Li, Sheng Tang, Jintao Li, and Tong-yee Lee. 2024. Identity-Preserving Face Swapping via Dual Surrogate Generative Models. *ACM Trans. Graph.* 11, 1, Article 1 (January 2024), 19 pages. https://doi.org/XXXXXXX.XXXXXXX

*Corresponding author: Fan Tang.

Authors' addresses: Ziyao Huang, huangziyao19f@ict.ac.cn, Institute of Computing Technology, Chinese Academy of Sciences, China; Fan Tang, tfan.108@gmail.com, Institute of Computing Technology, Chinese Academy of Sciences, China; Yong Zhang, zhangyong201303@gmail.com, AI Lab, Tencent, China; Juan Cao, caojuan@ict.ac.cn, Institute of Computing Technology, Chinese Academy of Sciences, China; Chengyu Li, lichengyu23s@ict.ac.cn, Institute of Computing Technology, Chinese Academy of Sciences, China; Sheng Tang, ts@ict.ac.cn, Institute of Computing Technology, Chinese Academy of Sciences, China; Jintao Li, jtli@ict.ac.cn, Institute of Computing Technology, Chinese Academy of Sciences, China; Tong-yee Lee, tonylee@mail.ncku.edu.tw, National Cheng Kung University, Taiwan.

0730-0301/2024/1-ART1 $15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Face swapping [Bitouk et al. 2008], also referred to as face replacement, enables the automatic exchange of faces among different portraits. By changing the distinct shape and texture [Liu et al. 2023] of unique facial components (e.g., eyes, mouth, and face shape)

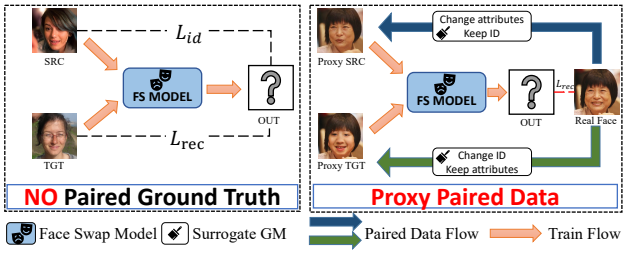


Fig. 2. Overviews of the Credible Supervision Completion via Surrogates framework. **Left**: the current face-swapping framework, where no ground truth paired data is used for training, and substituted by weakly ID loss and reconstruction loss. **Right**: the proposed framework, where dual-creating surrogates are applied to synthesize proxy-paired data to train the face-swapping.

while preserving non-identity attributes (e.g., pose, expression, illumination, and background), image and video face swapping methods [Chen et al. 2020; Mirsky and Lee 2021] have been widely used in the fields of movie and television creation, privacy protection, and virtual human creation.

Among fundamental requirements for face-swapping models, identity preservation is one of the most pivotal, where the swapped face should be faithful to the source person. Typical swapping methods [Chen et al. 2020; Gao et al. 2021; Kim et al. 2022; Li et al. 2020] utilize latent representations from pretrained face recognition models (e.g., ArcFace [Deng et al. 2019]) as identity prior to guiding the swapping model. Kafri et al. [2022]; Xu et al. [2022b,a]; Zhu et al. [2021] leverage the latent space of pretrained StyleGAN models to further support the control of identity and non-identity attributes during face swapping. Despite the impressive progress in face swapping, we argue that merely using latent embedding as implicit supervision will lead to "in-between" results. In these results, the swapped face appears as an interpolated identity between the source and the target. When the source and target images are not similar (i.e., of different genders), swapped results become difficult to recognize as the original individuals, as shown in Fig. 1.

This challenge can be dated back to the state of the ill-defined formulation for face-swapping tasks. Different from common data-driven methods where pre-collected <input, ground truth> could be used directly as **explicit supervision** for end-to-end training, finding a "real" swapped image for a given <source, target> is difficult. Considering the absence of explicit supervision, conventional face-swapping methods follow the **implicit supervision** in the form of an ID loss with a source image and a reconstruction loss with target image (Fig. 2 (Left)). However, the ID loss is calculated by a pretrained face classification model, which is not designed for face-swapping tasks [Liu et al. 2023]; the pixel-wise reconstruction loss may force the model to retain identity attributes of the target image [Kim et al. 2022]. Empirical studies in Sec. 3 show that implicit supervision is biased towards face-swapping tasks and may lead to inappropriate convergence of the swapping network.

In this research, we focus on boosting face-swapping models by adding explicit supervision for network training. Conceptually, we assume a given image as a ground-truth output for training face-swapping models. A novel Credible Supervision Completion via Surrogates (CS^2) framework is proposed to generate pseudo inputs reversely, as shown in Fig. 2 (Right): a source-creating surrogate model alters the attributes of the real image while keeping the original identity; a target-creating surrogate model changes the identity while maintaining non-identity attributes such as pose and expressions. An explicit adaptation is applied on the pseudo inputs to inject specific control (e.g., face shape) into the swapping process. Then, the proxy-paired data <pseudo source/target, ground truth> are directly fed to the network by pixel-wise explicit supervision. Furthermore, considering the inherent difference between face recognition and face-swapping tasks, implicit adaptation is employed to revise the identity representation by adding an ID encoder adapter trained on proxy-paired data. Extensive experiments with different backbones show that our method is effective in identity preservation while obtaining high-generation quality. Our contributions can be concluded as follows:

- We propose a novel face-swapping training framework to alleviate the absence of explicit supervision by introducing CS^2 framework, where proxy-paired data are constructed by dual-creating surrogates and explicit supervision is provided.
- We propose explicit adaption on proxy-paired data to further approximate ideal paired data, which enables face-shape transferring and other specific demands during face swapping for diverse real-world applications.
- We propose implicit adaptation, where a novel ID encoder adapter is employed to revise the identity representation according to proxy-paired data, and experiments demonstrate the improvement in identity preservation.
- Experiments show that the proposed CS^2 framework can generate high ID-preserved and shape-aware swapped face images. Various applications, including re-aging, customized swapping, cross-domain swapping, and video swapping are further proposed.

2 RELATED WORKS

2.1 Digital Face Generation

Digital face generation refers to the use of computer technology and algorithms to create virtual facial images. In the fields of computer graphics, artificial intelligence, and computer vision, researchers employ various techniques and methods, such as deep learning and generative adversarial networks (GANs) [Goodfellow et al. 2020], to produce realistic digital faces. These generated facial images find applications in various areas such as movie special effects, video games, virtual reality, and human-computer interface design. The research directions encompass modeling, editing, and animation.

2.1.1 Face Modeling. The foundation of digital face generation is the face modeling. 3D Morphable Model (3DMM) [Blanz and Vetter 1999], which is a statistical parameter face model, is widely applied across diverse applications. With the progress of artificial intelligence, deep learning-based face modeling [Ling et al. 2022; Zhang et al. 2023] and reconstruction [Li et al. 2022] methods have emerged, which yield outputs of notable quality. Furthermore, several methods have been introduced to systematically model distinct

facial components for comprehensively capturing diverse facial details. These facial components include eye [Francois et al. 2009], hair [Saito et al. 2018; Zheng et al. 2023], neck [Liu et al. 2021c] and face wrinkle [Deng et al. 2021]. Global and local face modeling plays a crucial role in achieving accurate facial generation.

2.1.2 Face Editing. Face editing is a prevalent application of digital face generation. Wang et al. [2022] guide face manipulation with 3D guidance, and Zhou et al. [2023] apply controllable face generation in neural radiance field [Mildenhall 2020]. In addition, certain works focus on editing specific facial components, such as hair editing [Song et al. 2023]. Face stylization [Han et al. 2021] translates the face into another style. It includes cartoons [Xiao et al. 2022] and caricature [Ye et al. 2021] and constitutes another pivotal research aspect item in face editing.

2.1.3 Face Animation. Animating facial poses and expressions is highly significant and extensively applicable. The research on transferring motion to the face has been a subject of extensive exploration since its early stages [Deng et al. 2006; Pei and Zha 2007]. As generative modeling advances, speech-driven face animation [Liu et al. 2021a; Wen et al. 2020; Yi et al. 2022] has become capable of producing realistic talking faces.

2.2 Face Swapping

Face swapping, as an application of digital face generation, aims to transfer identity from the source and non-identity attributes from the target to obtain the swapped face. Various approaches can be used to model identity information for face swapping, which will be discussed in the following sections.

2.2.1 3DMM-based Methods. In the classic methods, 3DMM [Blanz and Vetter 1999] is utilized to conduct face-swapping. 3DMM disentangles the face 3D structure into shape and texture, which contains ID information. 3DMM-based methods transform the ID-related 3D structure to achieve face-swapping. Classic 3DMM-based methods [Blanz et al. 2004; Nirkin et al. 2017] fit the source face into 3DMM to represent identity. Then, these methods transfer the pose, expression, and illumination according to the target image. Color transfer and segmentation are always applied to overcome skin or face differences. Face2Face [Thies et al. 2016] fits 3DMM first and then switches the ID-related information from the source to the target. Limited by 3DMM's capability, 3D-based methods suffer from unnatural results, and blending constraints the change in face shape.

2.2.2 Reenact-based Methods. This method utilizes a reenact model to transfer the source identity, which is then blended on the target background to swap faces. DeepFakes[1] and Deepfacelab [Perov et al. 2020] blend person-specific reenacted faces with mask and color correction, which results in failure in facial shape and texture preservation. [Naruniec et al. 2020] achieve high-resolution swapping results within this framework, and [Otto et al. 2022] extend it with the differentiable 3D network. FSGAN [Nirkin et al. 2019] reenacts via facial landmarks, inpaints, and blends with face segmentation. E4S [Liu et al. 2023] utilizes a reenact model to align the pose and expression. These methods could generate identity-preserving reenacted faces but fail to preserve after blending due to the difficulties in merging non-identity attributes.

2.2.3 Identity Encoder-based Methods. Identity encoder-based face-swapping approaches apply one encoder to extract identities, such as the pretrained face recognition model ArcFace [Deng et al. 2019], and another encoder to extract non-identity attributes. FaceShifter [Li et al. 2020] designs a two-stage network to integrate input attributes and revise output, while SimSwap [Chen et al. 2020] proposes a one-stage network to obtain high-fidelity swapped faces. In HifiFace [Wang et al. 2021a], a 3D face structure is additionally utilized to represent the identity information, and an auto-blending strategy is proposed. InfoSwap [Gao et al. 2021] disentangles identity attributes by information bottleneck. SmoothSwap [Kim et al. 2022] noticed the conflict between ID loss and reconstruction loss and applied supervised contrastive loss [Khosla et al. 2020] to smoothen ID feature space.

2.2.4 Latent-based Methods. In recent years, pretrained generative models, such as StyleGAN, have demonstrated remarkable disentanglement capabilities in latent space. MegaFS [Zhu et al. 2021] generates mega-pixel swapped faces by switching semantic appearance via StyleGAN2, while FSLSD [Xu et al. 2022a] transfers structural information. StyleFusion [Kafri et al. 2022] disentangles the composition of a face by fusing latent codes. RAFSwap [Xu et al. 2022b] uses face masks to guide the information interaction in StyleGAN2 latent space. E4S [Liu et al. 2023] treats face swapping as editing and demonstrates a novel regional inversion to transfer face shape and facial textures after reenacting. Latent space-based approaches are limited by the capability of the pretrained model and the inversion method to represent identities. These methods require a complex disentanglement mechanism to achieve attribute transfer, which still cannot get satisfactory results.

As progressive improvements have been made, face-swapping methods now use unintuitive implicit supervision guiding signals such as ID loss, reconstruction loss, 3D attributes, and face-mask information. However, they all suffer from no explicit supervision because of the inherent absence of ground-truth-swapped images.

3 BACKGROUND AND ANALYSIS

3.1 Notations

Given a source face x_{src} and a target face x_{tgt}, face-swapping transfers the identity from x_{src} and the non-identity attributes from x_{tgt} to generate the swapped face y_{out}. The current swapping framework, as visualized in Fig. 2, utilizes two basic losses to guide swapping.

- **ID loss** is commonly used to guide identity-swapping, which is calculated in pair (x_{src}, y_{out}) as $L_{id} = 1 - \cos(z_{src}, z_{out})$, where z_{src} and z_{out} are the ID features extracted by the ID encoder.
- **Reconstruction loss** is essential for the model to learn how to keep x_{tgt}'s attributes. L2 loss is utilized between x_{tgt} and y_{out} by $L_{rec} = ||y_{out} - x_{tgt}||_2^2$, when $x_{src} = x_{tgt}$.

The total loss L_{total} for face-swapping training is

1. https://github.com/deepfakes/faceswap

$$L = \lambda_{id} L_{id} + \lambda_{rec} L_{rec}, \tag{1}$$

where λ_{id} and λ_{rec} are the weights for L_{id} and L_{rec}.

We formulate the identity preservation problem for the face-swapping task. In face-swapping, the swapped face y_{out} is generated by the model M:

$$y_{out} = M(x_{src}, x_{tgt}). \tag{2}$$

When considering the visual attributes of each image,

$$\{Attr|y_{out}\} = \{\{Attr_{id}|x_{src}\}, \{Attr_{nid}|x_{tgt}\}\}, \tag{3}$$

where $Attr|y_{out} = \{a_1, a_2, \cdot, a_n\}$ is the visual facial features/attributes comprising the face image, and $Attr_{id}$ and $Attr_{nid}$ are identity and non-identity attributes, respectively. Eq. 3 represents the transfer identity and non-identity attributes from x_{src} and x_{tgt}. We posit that the setting of $Attr_{id}$, as an abstract concept, can vary across different scenarios. In the majority of instances, we assume that the visual features that would not be easily changed for individuals should be counted as $Attr_{id}$. Thus, the default setting of identity preservation is:

- *Default config*: $Attr^D_{id}$ = { distinct shape and texture of unique facial components (e.g., eyes, mouth, nose, and face shape)} and $Attr^D_{nid}$ = {pose, expression, illumination, hairstyle, face skin, occlusion, background }.

In the default setting, hairstyle and face skin belong to $Attr^D_{nid}$. Additional configurations of $Attr_{id}$ are discussed in Sec. 6.2.

3.2 Analysis on Explicit Supervision Absence

In the following passages, we delve into the reasons associated with the failure of identity preservation in face swapping. We explore the insufficient capability of the unintuitive identity loss and analyze the impact of implicit supervision caused by the absence of explicit supervision, that is, no ground-truth pair, which leads to identity failure.

We first conduct a brief experiment to investigate the capability of identity loss. We vary the weight of ID loss and reconstruction loss to train base face-swapping models, and the results are presented in Fig. 3. Our observations reveal that the output face is continuously varied from the target to the source, but the most ID-similarity swapped output is still difficult to recognize as x_{src}. We conclude that ID loss does not accurately reflect the faces for the entire control by ID loss.

Furthermore, we deform face shape on 1,000 images and compute the identity similarity using different face recognition models [Deng et al. 2019; Schroff et al. 2015; Wang et al. 2018] and a modified identity encoder SmoothID [Kim et al. 2022]. The results are shown in Fig. 4. All models exhibit similar results: the similarity far exceeds the threshold. This similarity results explicitly in the tolerance of identity loss toward the face-shape deformation and the gap between face recognition and face swapping, which certainly limits the capability of the face-swapping model to transfer face shape.

Based on the above-mentioned observation, we proposed an implicit supervision face-swapping training framework due to the absence of explicit supervision, which could result in identity preservation failure. We construct a counterfactual sample with different identities by modifying the face shape. As shown in Fig. 5a, we select

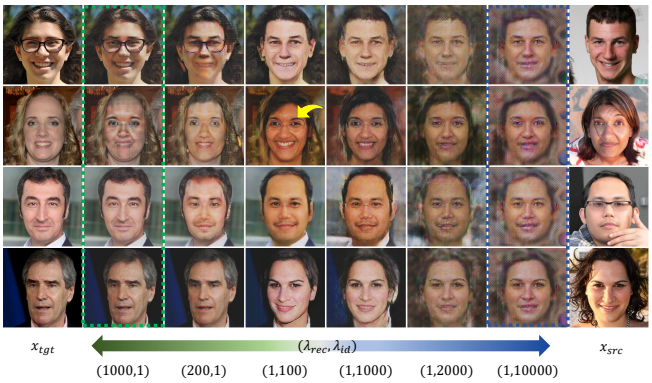


Fig. 3. Results of varying the weights of ID loss and reconstruction loss. The increase of ID loss to an extreme point (λ_{id}=10000) makes the outputs become stable. Meanwhile, drastic enhancement of reconstruction loss leads output images to tend to be target images. However, the balanced point to generate the most ID-similarity outputs (λ_{id}=100 in this picture) still results in source-unlikely, "in-between" [Liu et al. 2023] swapped images. In some cases, it brings errors such as 2nd row with an inexplicable glasses appearance and 3rd row with terrible color mutation.

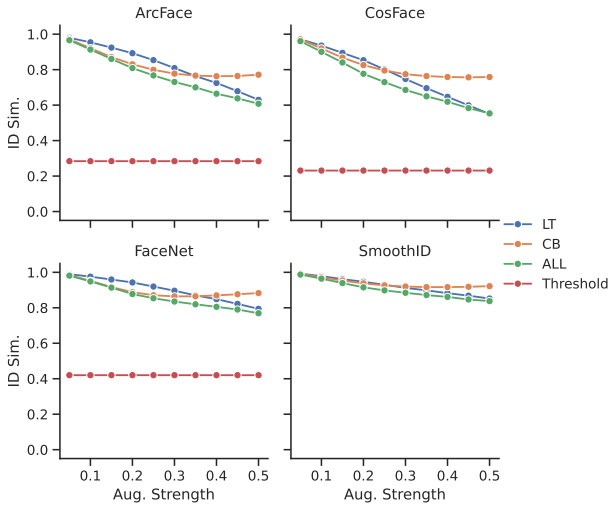
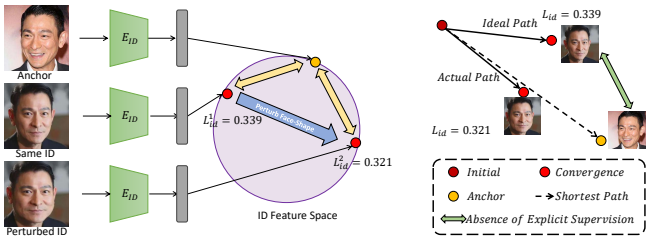


Fig. 4. Identity similarity on varying augmentation strength of shape-deformed faces and the corresponding face recognition thresholds. LT is linear transform, CB is face chubby and ALL means both are applied. Different identity encoders exhibit similar phenomena: similarity far exceeds the threshold even with a huge augmentation strength, indicating a mismatch between identity encoder and face swapping.

two real images of one identity and change the shape of one to an image with a perturbed identity in human perception. The identity losses by ArcFace are displayed in this image. Clearly, $L^1_{id} < L^2_{id}$, which means that, when the face-swapping training comes to this step, models guided by identity loss are more tended to converge by L^2_{id}, which results in identity change.



(a) (b)

Fig. 5. Analysis of identity loss. (a) Large changes in identification (e.g. face shape) by human perception could result in smaller identity loss from the pretrained face recognition model; (b) The phenomenon presented in (a) causes a wrong convergence path during face-swapping training because of the absence of explicit supervision.

ALGORITHM 1: Proxy-Paired Data Generation Pipeline

Input: Real Face Dataset D, Dual Surrogate Generative Models SM_{src} & SM_{tgt}, Shape Adaptation SA, Shape Adaptation Threshold th, Customized Adaptation CA, Dataset Size N
Result: Proxy pairs $\mathcal{M}$

$\mathcal{M} \leftarrow \{\}$;
for $n = 1, ..., N$ **do**
  Sampling a real face $y_{gt}^p \sim D$;
  $x_{src}^p \leftarrow SM_{src}(y_{gt}^p)$;
  $x_{tgt}^p \leftarrow SM_{tgt}(y_{gt}^p)$;
  $p \sim U(0, 1)$;
  if $p > th$ **then**
    $x_{tgt}^p \leftarrow SA(x_{tgt}^p)$;
  end
  $x_{src}^p, x_{tgt}^p, y_{gt}^p \leftarrow CA((x_{src}^p, x_{tgt}^p, y_{gt}^p))$;
  $\mathcal{M} \leftarrow \mathcal{M} \cup \{(x_{src}^p, x_{tgt}^p, y_{gt}^p)\}$;
end

Furthermore, when revisiting the common face-swapping training framework, reconstruction on the target image prompts the model to incorporate target identity attributes into the swapped face [Kim et al. 2022]. As a result, L_{id} tries to converge under the situation of identity change, while the loss could still decrease due to the above-mentioned analysis (Fig. 5b). Considering the combined effect of the two components of implicit supervision, the optimization is directly towards altering the identity. This conclusion inspires us to approach the identity preservation problem from the perspective of completing explicit supervision.

4 CREDIBLE SUPERVISION COMPLETION VIA SURROGATES

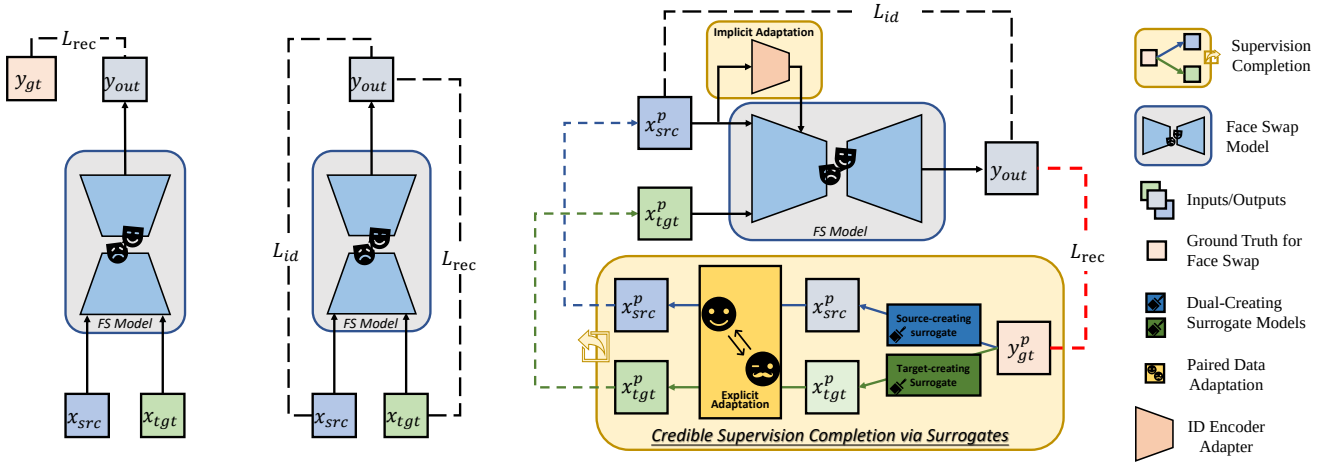
We propose the CS^2 framework to credibly complete the absence of explicit supervision in current face-swapping. To approximate the explicit supervision, our goal is to directly obtain real face-swapping images and construct the paired data (x_{src}, x_{tgt}, y_{gt}) to guide face-swapping training, as depicted in Fig. 6a. In the process of constructing paired data, the central challenge revolves around synthesizing these data and seamlessly incorporating them into the face-swapping task. We introduce dual-creating surrogates to generate credible proxy-paired data that can provide guidance for face-swapping training, which tackles the aforementioned issues. We also propose explicit and implicit adaptation strategies to revise the paired data and identity encoder. In this context, the surrogate [Papernot et al. 2017] refers to the pretrained models from other methods to serve our goal, and the proxy dataset is the dataset used as a replacement for the ground-truth dataset.

4.1 Dual-Creating Surrogates

The presence of a ground-truth swapped image, which is denoted as y_{gt} and illustrated in Fig. 6a, is indispensable to facilitate explicit supervision in face-swapping training. However, the acquisition of y_{gt} in real-life scenarios presents formidable challenges, which is mainly due to the dual dependencies of face swapping. Specifically, y_{gt} must mirror the exact identity of x_{src} while simultaneously sharing identical attributes with x_{tgt}. A pragmatic approach involves synthesizing y_{gt} by sequentially transferring attributes from two images. Established tools that transfer certain attributes, encompassing aspects, such as pose and expression [Wang et al. 2021c,d], hair style [Song et al. 2023], and background, can be utilized. However, this synthetic process is intricate and may entail error accumulation, particularly concerning identity preservation. Nevertheless, upon reconsidering the synthesis flow from a reverse perspective, we have discovered the feasibility of approximate ground-truth pairs without relying on long-term dependence.

We introduce *Dual-Creating Surrogates* to synthesize proxy-paired data x_{src}^p and x_{tgt}^p from a real face image y_{gt}^p. The generation pipeline for proxy-paired data is illustrated in Algorithm 1. In our condition, dual-creating surrogates are a couple of functionally symmetrical pretrained models: **source-creating surrogate** implements the function of changing the attributes of y_{gt}^p while keeping identity and output the x_{src}^p, such as reenact models like FaceVid2Vid [Wang et al. 2021c], LIA [Wang et al. 2021d], or rotation augmentation, can be applied to generate x_{src}^p with different pose and expression; x_{tgt}^p is synthesized by **target-creating surrogate**, by altering identity and preserving attributes, i.e., pretrained face-swapping models such as SimSwap [Chen et al. 2020] and InfoSwap [Gao et al. 2021], though insufficient at ID-preserving, are qualified to modify the identity of the real image. After generation with dual-creating surrogates, proxy-paired data (x_{src}^p,x_{tgt}^p, y_{gt}^p) are utilized as the training data for swapping, where the face-swapping model learns to transfer the identity of x_{src}^p to x_{tgt}^p to predict the ground-truth swapped face image y_{gt}^p.

The generation pipeline is visualized in Fig. 6c. Our pipeline has two advantages: **1)** our proxy-paired data provide credible supervision, a real face y_{gt}^p as ground-truth output, and a proxy source x_{src}^p with the same identity to guide swapping for completing the absence of explicit supervision. **2)** One-stage synthetic chain prevents the ID shift caused by sequential multiple steps, which also



(a) (b) (c)

Fig. 6. Comparison between different swapping training frameworks. (a) Face-swapping training with ground-truth swapped image y_{gt}; (b) Current swapping training with ID loss and reconstruction loss; (c) Our CS^2 synthesis proxy paired data by dual-creating surrogates from one real image, and adaptation is operated to get the required training data. Credible supervision is applied by L_{rec} between y_{out} and y_{gt} where a real image is utilized to guide swapping. The reserve design of dual surrogates prevents error accumulation from the multi-step synthesis process.

benefits identity-preserving swapping. The absence of any of these surrogates can potentially result in the failure to achieve explicit supervision, and a detailed analysis of this scenario will be presented in the following experimental Sec. 5.

4.2 Explicit Adaptation on Proxy-Paired Data

Directly applying proxy-paired data is insufficient for identity preservation, especially on face-shape transferring. This limitation stems from the shape-unawareness of the target-creating surrogate. To overcome this shortcoming, we propose an explicit adaptation on paired data strategy, where application-oriented modification is applied to revise the proxy-paired data, which helps face-shape transferring. Moreover, we have discovered that explicit adaptation enables corresponding adaptation to be applied to proxy-paired data, which results in adaptable guidance. This strategy yields customized face-swapping models to meet real-world diverse swapping needs.

4.2.1 Shape Adaptation. In face-shape transferring [Kim et al. 2022; Wang et al. 2021a], face-swapping models should reconstruct the face shape of x^p_{src} on swapped images rather than keep x^p_{tgt}. Face shape is an essential part of ID representation [Liu et al. 2023], and a model has difficulty to accurately transfer the face shape because of the reconstruction loss [Kim et al. 2022]. Researchers have explored various methods to guide face-shape transferring, including leveraging 3D information [Wang et al. 2021a], disentangled ID features [Gao et al. 2021], or smoothed identity representations [Kim et al. 2022]. However, face-shape transferring remains constrained by implicit guidance. In our framework, we simplify the face-shape transferring problem through explicit adaptation on paired data.

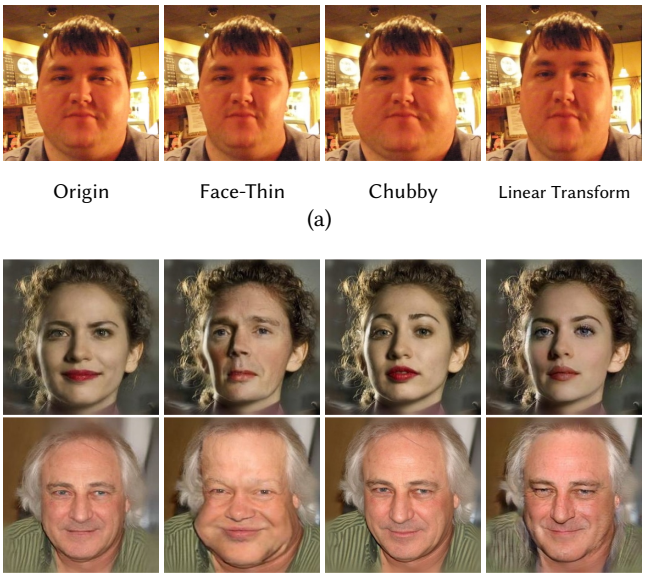


Fig. 7. Face-shape adaptation and proxy-paired data visualization. (a) Face warping on x^p_{tgt}; (b) Shape transferring on proxy paired data during training.

To guide the face-shape transferring, the ground-truth swapped face y^p_{gt} should have the same face shape as x^p_{src} while being different with x^p_{tgt}. To attain this requirement, we alter the face shape in x^p_{tgt},

which enlarges the difference between x^p_{tgt} and y^p_{gt} in terms of facial structure. We utilize three types of augmentations, namely, face-thin, chubby, and linear transform. The detailed visualization of different face-shape adaption and the final adapted paired data are shown in Fig. 7. This way allows the model to learn how to swap face-shape effectively.

4.2.2 Customized Adaptation. Customizing your face-swapping model to transfer or preserve certain attributes can significantly enhance the user experience. Customized face swapping is a problem that has been relatively underexplored in the past, which can be marginally achieved by E4S [Liu et al. 2023] through editing. Our framework possesses flexible adaptability, which enables it to accomplish various real-world applications. Customized requirements can be satisfied using carefully designed adaptation strategies on each component of proxy-paired data. For illustration, we provide two examples involving beard removal and glasses transfer, which may be practical in some special scenarios. Image editing method [Karras et al. 2020b; Shen et al. 2020] is utilized to construct adapted proxy-paired data. For beard removal, the objective is to instruct the swapping model to eliminate facial bread from x_{src}. For glasses transfer, we aim to selectively add glasses to facilitate the transfer process. Sec. 6.2. shows a more comprehensive understanding of these adaptation settings and visualizations.

4.3 Implicit Adaptation on ID Encoder

Considering the limitation of the identity encoder (i.e., ArcFace) discussed in Sec. 3, we propose a novel strategy to provide complementary identity information. Given that proxy-paired data offer more suitable guidance for face swapping than traditional face-recognition tasks, these data should be leveraged to enhance the performance of the identity encoder. However, a direct fine-tuning of the ID encoder on imperfect proxy-paired data could result in worse identity representation. As a result, our primary objective is to enhance the current identity representation while preserving the original capabilities of the identity encoder. To accomplish this objective, we train an adapter of the identity encoder on proxy-paired data during face-swapping training.

An adapter is a plug-and-play network architecture to modify the existing network. We initialize an adapter A_{id} with the same network architecture and weight from the identity encoder E_{id}. To retrain the inherent capabilities of the identity encoder, the weight of the identity encoder E_{id} is frozen, and the adapter A_{id} learns a residual to the identity embedding z during face-swapping training. For a face image x, adapter A_{id} extracts the residual identity embedding z_{rec}, and then adds it to the origin one z:

$$\hat{z} = z + z_{res}, \tag{4}$$

where $\hat{z}$ is the final identity embedding. Before adding, we apply the zero-convolution layer [Zhang and Agrawala 2023]:

$$z_{res} = zero\_conv(A_{id}(x_{src})), \tag{5}$$

which aims to totally preserve original identity embedding at the beginning of the training, and the identity residual would be learned during face-swapping training. The adapter is visualized in Fig. 8,

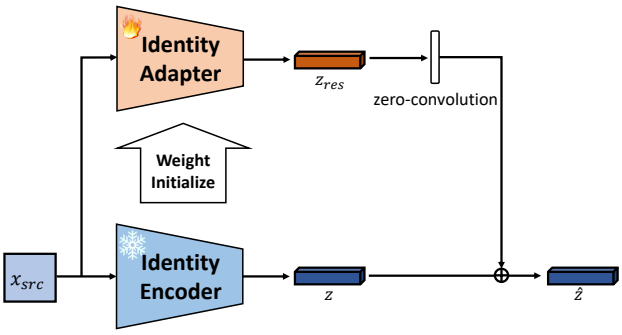


Fig. 8. Identity adapter architecture. A zero-convolution layer is applied to keep the identity encoder's ability at the beginning of training.

and the final equation is:

$$\hat{z_{src}} = E_{id}(x_{src}) + zero\_conv(A_{id}(x_{src})). \tag{6}$$

4.4 Joint Training with Proxy Paired Data.

We set the proxy-paired data as ancillary training data in normal steps. For each batch, half of the data are proxy-paired data, a quarter is normal swap data, and a quarter is for reconstruction. Besides the three two losses in Sec. 3, the others are defined in the following.

To guide the swapping by proxy-paired data, L2 loss is applied between y^p_{gt} and y^p_{out} for conducting supervision completion:

$$L^p_{rec} = ||y^p_{out} - y^p_{gt}||^2_2. \tag{7}$$

ID loss L^p_{id} is still calculated on pair (x^p_{src}, y^p_{out}):

$$L^p_{id} = 1 - \cos(z^p_{src}, z^p_{out}), \tag{8}$$

where z^p_{src} and z^p_{out} are the ID feature extracted by encoder. Adversarial loss is employed on y_{out} and y^p_{out}:

$$L_{adv} = -(\log(D(y_{out}), \tag{9}$$

$$L^p_{adv} = -(\log(D(y^p_{out})). \tag{10}$$

The total loss, which consists of the basics in Sec. 3, is:

$$L = \lambda_{id}L_{id} + \lambda_r L_{rec} + \lambda^p_{id}L^p_{id} + \lambda^p_r L^p_{rec} + L_{adv} + L^p_{adv}, \tag{11}$$

where $\lambda_{id} = 1$, $\lambda_r = 10$, $\lambda^p_{id} = 1$ and $\lambda^p_r = 20$. A high weight for λ^p_r boosts the effectiveness of proxy paired data.

4.5 Analysis of Identity Preservation

Based on the configurable $Attr_{id}$ discussed in Sec. 3.1, despite the challenges associated with direct encoding of $\{Attr_{id}, Attr_{nid}\}$ through face-swapping model, CS^2 constructs proxy-paired data (x^p_{src}, x^p_{tgt}, y^p_{out}) with:

$$\begin{aligned} \{Attr|y^p_{out}\} &= \{Attr_{id}, Attr_{nid}\}, \\ \{Attr|x^p_{src}\} &= \{Attr_{id}, \hat{Attr}_{nid}\}, \\ \{Attr|x^p_{tgt}\} &= \{\hat{Attr}_{id}, Attr_{nid}\}, \end{aligned} \tag{12}$$

where $\hat{Attr}_{id}$ and $\hat{Attr}_{nid}$ are attributes that do not need to be transferred by the face-swapping model M. The constructed proxy-paired data (x^p_{src}, x^p_{tgt}, y^p_{out}) fit specific $\{Attr_{id}, Attr_{nid}\}$ through chosen

surrogates and explicit adaptation, which tells the model which elements should be transferred with configurable $Attr_{id}$.

5 EXPERIMENTS

5.1 Setup

5.1.1 Implementation Details. We train our model on FFHQ [Karras et al. 2019] and CelebaHQ [Karras et al. 2018] dataset. A total of 90% of these datasets are used as training data, and the remaining 10% are set as test data. The proxy-paired data are synthesized from FFHQ and CelebaHQ, which results in a total of 90K pairs. Our main setting uses InfoSwap [Gao et al. 2021] as the Target-Creating Surrogate and FaceVid2Vid [Wang et al. 2021c] as the Source-Creating Surrogate. When generating x^p_{src} by the Source-Creating Surrogate, the angle between x_{gt} and the referenced face image is limited to under 30° by using a face angle estimator for selection. Some face-warping methods are utilized to vary the face shape in x^p_{tgt}. In detail, the pychubby library[1] makes the face chubby or randomly changes the face scale in the x and y axes. We also apply a face-thin algorithm. During training, employing the shape warped x^p_{tgt} has 50% probability.

We train our model on 256×256 resolution. The FFHQ preprocess method is utilized to align and crop given that it holds more background. The batch size is set as 32, 16 of which are for proxy paired data, 8 for reconstruction, and 8 for normal swap training. Our model is trained on one NVIDIA Tesla A100 GPU. The training steps are set as 200K to be the same with Li et al. [2020]. The learning rate is fixed at 4×10^{-4} for the generator and discriminator with the ADAM optimizer [Kingma and Ba 2014].

5.1.2 Network Architecture. We exploit the classic face-swapping model as our network architecture. Arcface [Deng et al. 2019] is utilized as our source encoder and in ID loss calculation. We apply our framework on three face-swapping backbones, namely, AEI-Net borrowed from FaceShifter [Li et al. 2020] (main setting), SimSwap [Chen et al. 2020], and HifiFace [Wang et al. 2021a] without 3D face extractor. For adversarial training, we use the StyleGAN2's [Karras et al. 2020b] discriminator, and we follow Karras et al. [2020b]; Kim et al. [2022] in using R1 regularizer [Mescheder et al. 2018]. Our framework is used as a plug-and-play component in classic face-swapping training and is easy to use and easy to control.

5.1.3 Comparison Methods. We compare several face-swapping models that share official implementation and weights or uploaded swapped FF++ [Rössler et al. 2019] videos. Specifically, **SimSwap** [Chen et al. 2020], **InfoSwap** [Gao et al. 2021], **MegaFS** [Zhu et al. 2021], **StyleFusion** [Kafri et al. 2022], **FSLSD** [Xu et al. 2022a], **RAFSwap** [Xu et al. 2022b] and **E4S** [Liu et al. 2023] share code and checkpoints, and the results of **DeepFakes**, **FaceShifter** [Li et al. 2020] and **HifiFace** [Wang et al. 2021a] come from the open official FF++ swapped videos. Uploaded checkpoints or our implementations are utilized for comparing with wild images and FFHQ test split.

5.2 Experiments on FF++

5.2.1 Quantitative Results. Quantitative experiments are organized following previous works [Gao et al. 2021; Kim et al. 2022; Li et al. 2020]. FF++ [Rössler et al. 2019] is a video dataset containing 1,000 real and corresponding face-swapping videos. We evenly sampled ten frames from each video and obtained a total of 10K frames. The swapping pairs comply with the official setting, and finally, we build 10K image pairs for swapping. For each source video, the first selected frame is set as the x_{src}. For the ID-preserving performance, we use CosFace [Wang et al. 2018] and FaceNet [Schroff et al. 2015] to compute the ID Retrieval rate and ID cosine similarity between x_{src} and y_{out}. As discussed in the Sec. 3.2, metrics based on these two ID encoders do not entirely align to ID-preserving performance. Nevertheless, there are currently no better methods for quantitative evaluation. Pose error and expression error are calculated to measure the attribute consistency between x_{tgt} and y_{out}. Following InfoSwap [Gao et al. 2021] , 3DFFA-v2 [Guo et al. 2020] is employed to predict the parameters of face structure, and L2 distance is computed as pose error and expression error.

As presented in Table 1, our method results in the highest ID cosine similarity and comparable ID retrieval rate, especially when conducting experiments on different backbones. The results show the ability of our method to preserve identity. InfoSwap obtains the second highest ID similarity, given that it disentangles source-related ID features and target-related ID features and obtains a high ID-preserving performance. SimSwap, HifiFace, and FaceShifter are comparable to one another, while ours and InfoSwap are obviously better. When combined with ours, their numerical results at identity are improved obviously. MegaFS performs comparable results to other SOTAs, while the results of FSLSD and RAFSwap are inferior to those of others. This performance may be caused by the domain gap brought by the GAN-inversion method[1] . The results obtained with E4S suffer from poor illumination, which is primarily attributable to the reenactment-based framework.

For pose and expression ("Exp." in Table 1), our method is comparable to other methods. The reason is that ours changes face shape largely and affects the results. Another shape-aware method HifiFace [Wang et al. 2021a], which introduces a 3D face representation to guide the shape transferring, obtains higher pose error than ours, and ours is better at ID preserving. DeepFakes obtains the lowest pose and expression performance given that it only blends a square of the target face. SimSwap is the best model for pose and expression preserving, but this method swaps a little in perception compared with the target face image due to high reconstruction.

5.2.2 Qualitative Results. As shown in Fig. 9, we list the swapped results on the FF++ dataset. Our method performs the most similarity identity with the source face. When compared with the 3D guided shape-aware method HifiFace [Wang et al. 2021a], we observe that our swapped face shape is transferred more similarly.

5.3 Results on Wild Images and FFHQ Test Split

5.3.1 Performance on Wild Images. We collect celebrity images for swapping to demonstrate the identity-preserving ability of our

1. https://github.com/jankrepl/pychubby
1. https://github.com/cnnlstm/FSLSD_HiRes/issues/10

Table 1. Quantitative results on FF++ dataset with six metrics. ↑ indicates that, when the score is higher, the model performance is better, and vice versa. The best scores of each metric are in bold, and the second scores are underlined.

Method	CosFace Ret.↑	FaceNet Ret.↑	CosFace Sim.↑	FaceNet Sim.↑	Pose↓	Exp.↓
DeepFakes	0.818	0.462	0.449	0.577	8.852	0.189
FaceShifter [Li et al. 2020]	0.798	0.702	0.518	0.673	2.513	<u>0.054</u>
MegaFS [Zhu et al. 2021]	0.874	0.818	0.499	0.754	3.511	0.123
InfoSwap [Gao et al. 2021]	<u>0.993</u>	0.915	0.655	0.760	2.463	0.079
StyleFusion [Kafri et al. 2022]	0.284	0.169	0.288	0.430	6.452	0.080
FSLSD [Xu et al. 2022a]	0.303	0.144	0.271	0.446	3.838	0.149
RAFSwap [Xu et al. 2022b]	0.043	0.009	0.167	0.135	2.693	0.072
E4S [Liu et al. 2023]	0.945	0.845	0.548	0.684	3.868	0.087
AEI-Net [Li et al. 2020]	0.835	0.730	0.525	0.641	7.912	0.349
+ Ours	**0.994** (+0.159)	**0.972** (+0.242)	**0.711** (+0.186)	**0.815** (+0.174)	2.801 (-5.111)	0.088 (-0.261)
SimSwap [Chen et al. 2020]	0.958	0.874	0.602	0.720	**1.260**	**0.034**
+ Ours	0.990 (+0.032)	<u>0.968</u> (+0.094)	<u>0.674</u> (+0.072)	<u>0.785</u> (+0.065)	<u>2.159</u> (+0.898)	0.063 (+0.029)
HifiFace [Wang et al. 2021a]	0.878	0.772	0.561	0.692	3.025	0.084
+ Ours	0.986 (+0.108)	0.954 (+0.182)	0.652 (+0.091)	0.767 (+0.075)	2.615 (-0.410)	0.083 (-0.001)

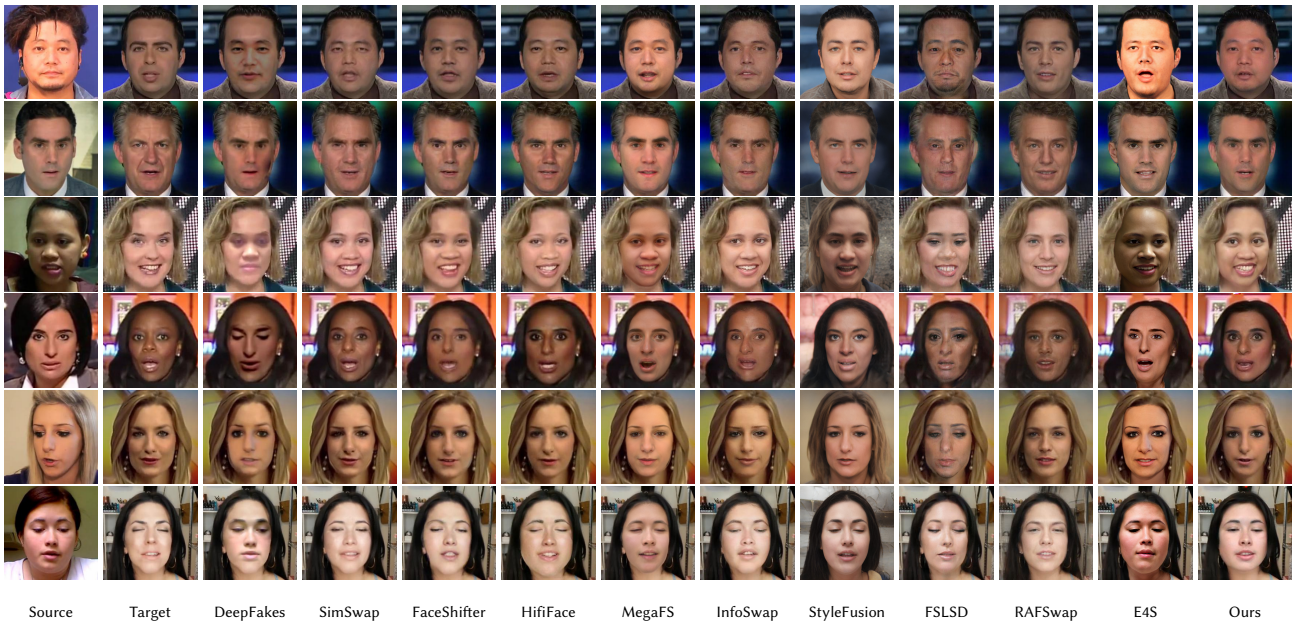


Fig. 9. Qualitative comparison results on FF++ dataset. Ours keeps abundant facial features such as face shape to obtain a high identity-preserving effect. Besides, our model preserves pose and expression consistent with the target face.

method. As displayed in Figs. 1 and 10, the swapped celebrities can be recognized at a glance more than other methods. This finding shows the superior identity-preserving capability of our method.

5.3.2 Performance on FFHQ Test Split. Results on the FFHQ test split are displayed in Fig. 11. Compared with others, ours has a significant advantage in face-shape transferring. These results indicate that, when applied to cross-gender or cross-age swapping, our method can keep the identity-related global features of the source, such as face shape, gender, and facial attributes.

5.4 Ablation Study

We conduct a detailed ablation study on CS^2 to verify the effectiveness of our framework. The study is unfolded from four aspects: the entire pipeline, the surrogates, explicit adaptation, and the ID adapter.

5.4.1 Validation of the Pipeline. We conduct an ablation study on each component of our method to measure the entire pipeline. The quantitative results are shown in Fig. 2, and qualitative results are shown in Fig. 12a. We observe that, without proxy-paired data, our training setting can only obtain poor-quality output and numerical

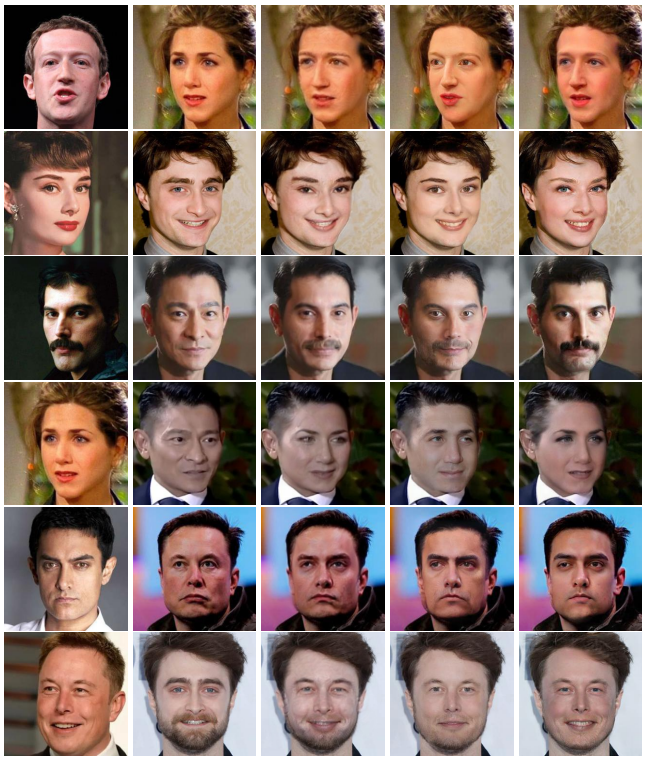


Fig. 10. Qualitative comparison results of SimSwap, InfoSwap, and ours on wild images. InfoSwap generates shape-aware realistic images but not as good as ours, while SimSwap changes slightly compared with the target image, particularly in face shape.

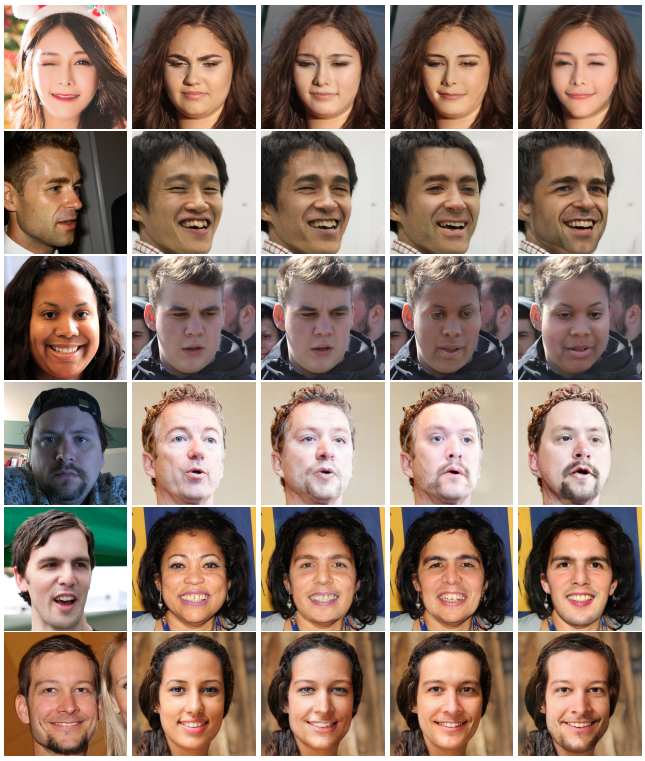


Fig. 11. Qualitative comparison results of SimSwap, InfoSwap, and ours on FFHQ test split. InfoSwap generates shape-aware realistic images but not as good as ours, while SimSwap changes slightly compared with the target image, particularly in face shape.

Table 2. Quantitative ablation study of our method.

Method	CosFace Sim.↑	FaceNet Sim.↑	Pose↓	Exp.↓
Ours	0.711	0.815	2.801	0.088
w/o ID Adapter	0.672	0.782	2.885	0.095
w/o Adapatation	0.657	0.768	2.674	0.088
w/o Proxy Paired Data	0.525	0.641	7.912	0.349
AEI-Net w shape Aug.	0.546	0.663	11.165	0.310
Only Proxy Data w/o ID loss	0.398	0.509	2.111	0.067
Only Proxy Data w ID loss	0.637	0.744	2.393	0.087
SimSwap [Chen et al. 2020] as Surrogate	0.672	0.788	2.234	0.086
PIRender [Ren et al. 2021] as Surrogate	0.640	0.746	3.557	0.118
LIA [Wang et al. 2021d] as Surrogate	0.682	0.797	2.638	0.083
Rotation as Surrogate	0.691	0.809	2.449	0.083
Single Target Surrogate	0.699	0.810	3.456	0.117
Single Source Surrogate	0.445	0.573	9.017	0.226
Bad Target Surrogate	0.648	0.763	3.864	0.152
Bad Source Surrogate	0.640	0.746	3.557	0.118
Fine-tune ID Encoder	0.018	0.039	6.282	0.291
ID Adapter with "Concat"	0.726	0.822	3.720	0.135

Fig. 12. Ablation of the pipeline. **P** means proxy paired data, **A** means shape adaptation and **I** means ID Adapter. In our training set, the model outputs poor-quality images without proxy-paired data. Shape adaptation enhances face-transferring capability. By introducing an ID adapter, face shape, and identity attributes are preserved better.

results, which shows the capability of speedy convergence. Moreover, the ID similarity and attribute measurements in Table 2a are enhanced largely, which reflects the advantage of CS^2. In Fig. 12a, the face-shape transferring is clear to observe, which indicates that paired data adaptation helps the model extract the face-shape attributes of the source face and successfully transfers shape to swapped faces. In addition, shape adaptation on paired data benefits the identity of qualitative results.

We demonstrate more results among different network architectures on wild images and FFHQ test split in Fig. 13 to verify the advantages of our method as a novel framework. The results

show the effectiveness of our framework apart from the results in Sec. 5.2. For swapped results on wild images and FFHQ test split, we utilize our implemented HifiFace and AEI-Net models due to the inaccessible weight. Our framework significantly improves the identity-preserving capability compared with base models among different face domains.

We further explore the effectiveness of our methods on convergence in Fig. 14, where we list comparison results of continuous training steps. We observe that training with proxy-paired data helps the model output high-quality results and keep attributes like expression. The reason is that proxy-paired data provides intuitive supervision, which benefits the convergence speed of face-swapping training.

Another important concern is about the training and inference cost. The inference is the same as the baseline model given that no additional modules are proposed. The training dataset and steps are the same as the baselines. The only additional cost is the time to generate proxy-paired data, which takes about 16 to 24 hours in our experiments.

5.4.2 Validation of Surrogates. To ensure the reliability and credibility of dual-creating surrogates, we conduct evaluations from the following aspects: the necessity of both surrogates, the choice of each one, and the persistence of identity loss on proxy-paired data.

First, we evaluate the setting of a single surrogate, which means only target or source-creating surrogate is applied. For a single target-creating surrogate, we set x^p_{src} equals y^p_{gt}. In this setup, proxy-paired data can also be generated, which yields a complete training pipeline. For a single source-creating surrogate, the proxy target faces x^p_{tgt} cannot be prepared, and our framework fails to operate properly. In this scenario, we use the proxy source face x^p_{src} as a kind of data augmentation. The results are shown in Fig. 15 and Table 2. In the setting of a single target-creating surrogate, the $x^p_{src} = y^p_{gt}$ setup gives the same identity for training and obtains high numerical ID-preserving results. However, the qualitative results show overfitting and apparent color artifacts. Single source-creating surrogate gives similar results with no proxy-paired data setting, as discussed in the abovementioned paragraph because of the absence of explicit supervision.

Besides the necessity of both surrogates, we validate the choice of each one. The choice of source-creating surrogate depends on its identity-preserving capability. Although imperfect, reenact models provide complementary information to guide face swapping in our framework. We vary the selection of source-creating surrogate while Face-Vid2Vid [Wang et al. 2021c] keeps as target-creating surrogate. In this experiment, another reenact model, namely, LIA [Wang et al. 2021d], and rotation augmentation (rotate between -30° and 30°) are chosen as source-creating surrogates. Face-Vid2Vid is based on implicit 3D facial landmarks, while LIA utilizes latent motion representation. Apart from the complex approach, the simple rotation operation provides x^p_{src} with a slightly modified pose, which can also serve as our source surrogate given that it keeps the identity of y^p_{gt}. The comparable results are shown in Fig. 15. The quantitative findings and identity-preserving results reveal that different source surrogates work similarly in our framework. We vary the surrogate models with others. We replace InfoSwap [Gao et al. 2021] with SimSwap [Chen et al. 2020] as the target-creating surrogate, and the results are shown in Fig. 17. SimSwap supplies marginally altered x^p_{tgt} compared with InfoSwap. However, the results in Fig. 15 show that the swapping results obtained by replacing different pretrained models for the surrogate model are consistent. The comparison results demonstrate the effectiveness of our methods as a framework, rather than relying on the specific surrogate model used.

Another consideration in this section is the persistence of identity loss on proxy-paired data. Although proxy-paired data provide credible supervision guidance toward repairing the weakness of identity loss, CS^2 is not intended as its replacement. Common face recognition models are trained on millions of faces, such as MS1M [Guo et al. 2016] while face swapping at a smaller scale of training data. Therefore, they are powerful but have imperfect capability in the face-swapping task. Accordingly, we believe that joint training with common identity loss and corresponding samples could work better with our framework. Table 2 and Fig. 15 show the results for training only on proxy-paired data with and without identity loss. Without identity loss, the training convergence becomes hard, which reveals the identity loss capability. Furthermore, with only proxy data, the changes in identity become less apparent.

5.4.3 Validation of Explicit Adaptation. Explicit adaptation on paired data is another crucial module within CS^2, which provides flexible and carefully revised guidance for face swapping. In this part, face-shape adaptation, which is closely associated with identity preservation, will be primarily discussed. Other face-swapping applications accomplished by specific adaptation operations are demonstrated in the following Sec. 6.

The detailed visualization of face-shape adaptation can be found in Sec. 4.2. We have shown an improvement in face-shape transferring due to this adaptation. Moreover, we display the swapped results on shape-adapted proxy-paired data in Fig. 16 compared with InfoSwap and SimSwap. InfoSwap and SimSwap keep the face shape from the target face, which indicates the disadvantage in face-shape transferring by other methods compared with ours.

The other meaningful investigation is to compare our framework with face-shape augmentation directly applied to training data. We randomly augment the face shape on our training set to reach the data scale of our setting and train our baseline swapping model on it. Experiment results are displayed in Fig. 17, where baseline model training with augmentation is insufficient to generate high-quality and shape-awareness results compared with ours. The reason is that augmentation only diversifies the dataset and does not support the more credible guidance. Meanwhile, CS^2, gives the appropriate guide by creating proxy ground-truth pairs through reversely utilizing the augmentations. Furthermore, training the face-swapping model with only identity loss is unstable, and joint training with proxy-paired data significantly improves the swapping results.

5.4.4 Validation of ID Encoder Adapter. To investigate the impact of the configuration of the ID adapter, we conduct a comparative analysis using two different settings: i) Fine-tuning of ID Encoder: In this setting, we fine-tune the ID encoder during training; ii) “Concat" Operation Replacement: This setting involves substituting the “Add"

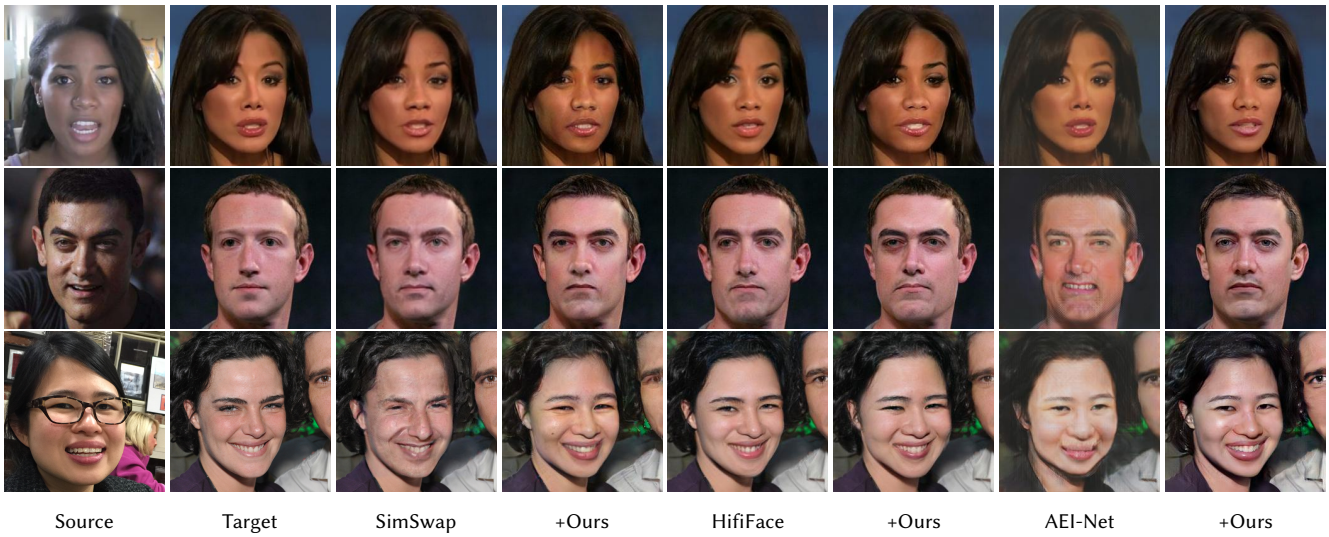


Fig. 13. Qualitative results with different backbones. All of them achieve high ID-preserving and the ability to transfer face-shape than base model results.

Fig. 14. Ablation analysis with proxy paired data at different training steps (twice the steps of *w/o proxy paired data* setting is equivalent to *ours*). Training with proxy-paired data speeds up the convergence effectively and contributes to image quality and attribute preservation.

operation in Eq. 6 with the “Concat" operation and removing the unnecessary zero convolution layer. We present the results of this comparison in Table 2 and illustrate them in Fig. 18. In the case of the fine-tuning setting, the quantitative identity results are nearly zero, and the qualitative assessment indicates that the swapped faces do not resemble either the source or target face. This finding underscores the fact that direct fine-tuning leads to overfitting the training data, which results in incorrect outcomes. Although the “Concat" setting achieves a high degree of identity similarity, it fails in preserving pose and expression, as evidenced by quantitative and qualitative results. Ultimately, the “Add" operation exhibits a superior balance among the adapter settings, which offers improved outcomes in terms of identity preservation and maintaining pose and expression.

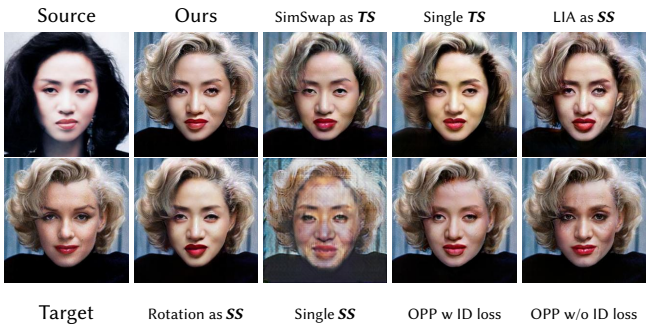


Fig. 15. Evaluation on Surrogates. ***SS*** stands for source surrogate, ***TS*** stands for target surrogate and OPP stands for training only on proxy-paired data.

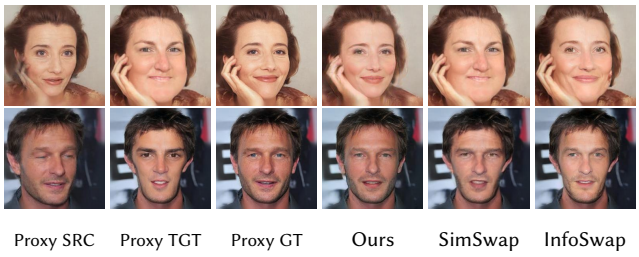


Fig. 16. Face-shape adaptation and comparison. After adaptation, ours generate shape-preserving results compared with others.

5.5 User Study

We conduct a user study on FF++. We randomly sample 100 swapping pairs from the abovementioned experiment on FF++, and corresponding swapped results are gathered to operate the user study. We consider three kinds of measurements, namely, identity preserving,

Fig. 17. Results compared with augmentation. Face-shape augmentation diversifies training data distribution, which is insufficient to enable shape transferring without proxy-paired data.

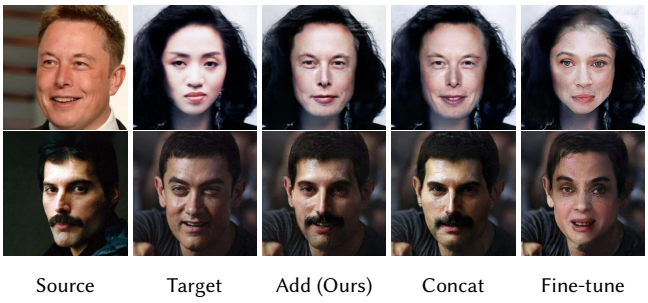
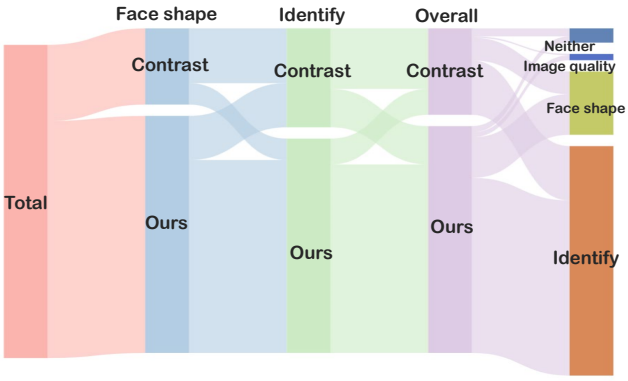


Fig. 18. Analysis of different ID adapter configurations. Fine-tuned ID encoder generates wrong swapped faces, while "Concat" fails to preserve expression. "Add (Ours)" outputs the best results considering identity preserving and attribute transferring.

face-shape transferring, and overall quality. We ask participants to choose the more suitable swapped image from two results: ours and a randomly selected comparison method. Moreover, we ask users what factors influence the choice of overall quality, including identity, shape, image quality, or neither. A total of 45 users are organized, and each is asked to answer 30 question pairs. The results are presented in Fig. 19.

Fig. 19a presents a Sankey diagram, which demonstrates the choice flow. Specifically, most users who select the face shape also choose the identity, and a few select the comparison method. The same phenomenon also occurs between choice identity and overall quality. When users are requested to answer what factors influence the overall selection, most of them choose the identity, and face shape is the second most chosen. Results indicate that most people are impressed by CS^2's identity-preserving capability.

The detailed user preferences are listed in Fig. 19b. The results show that our method is superior to other methods in all measurements. In detail, HifiFace [Wang et al. 2021a] obtains the nearest users' preference with ours, particularly in overall quality and face shape. This phenomenon is due to the 3D-guided and auto-blending strategy in HifiFace, while we still have 80% votes, which demonstrates our method's face-shape transferring ability.



(a) Proportion of user preferences

Method	Identify	Face shape	Overall
Ours vs DeepFakes	90.31	95.47	96.02
Ours vs SimSwap	87.29	89.60	96.88
Ours vs FaceShifter	84.84	91.21	87.19
Ours vs HifiFace	85.41	82.73	76.61
Ours vs MegaFS	85.86	89.00	87.65
Ours vs InfoSwap	93.74	95.66	96.16
Ours vs StyleFusion	93.92	82.69	83.55
Ours vs FSLSD	90.06	91.22	95.35
Ours vs RAFSwap	94.81	89.34	94.72
Ours vs E4S	88.38	88.40	93.88

(b) Proportion of wins (%)

Fig. 19. User study results. (a) The Sankey diagram displays the users' preference between ours and other methods, (b) detailed results in each comparison.

5.6 Discussion: Influence of Surrogates on Identity

The generative capability scope of the chosen surrogates decides the attributes of proxy-paired data, which could influence the performance of the final obtained model, especially for identity. To investigate the influence, we show the detailed performance upon the choice of surrogates.

5.6.1 Influence of Target Surrogates. As shown in Fig. 20, InfoSwap has a certain capability for facial-shape transferring, while SimSwap barely modifies facial shapes. Therefore, the performance of models trained by the two surrogates shows a noticeable difference in facial appearance. Fig. 21 illustrates that models trained using InfoSwap as a surrogate exhibit pronounced facial-shape transferring capability, while SimSwap does not. This finding is also reflected in numerical results, with the former maintaining better identity preservation, while the latter, holds a certain advantage in preserving pose and expression in Table 2 due to lesser variations compared to the target. On the one hand, when constructing proxy-paired data, the surrogate used must consider some of its own performance. On the other hand, explicit adaptation is important, because it can mitigate the shortcomings of the surrogate's own capabilities to some extent.

Fig. 20. Comparisons of proxy data generated by different surrogates. InfoSwap has a more shape-transferring capability compared with SimSwap, where LIA performs similarly to FaceVid2Vid.

5.6.2 Influence of Source Surrogates. We use the proxy source images x^p_{src} generated by Face-Vid2Vid and LIA in Fig. 20 to investigate the influence of **source surrogate** for identity preservation. Compared with the target surrogate, the source surrogates require that the generated proxy image x^p_{src} maintains consistency with the original image in terms of identity, which does not explicitly affect the ID functionally. As shown in Fig. 20, although the two methods exhibit varying abilities in preserving the expressions of the driver, they are essentially consistent in terms of ID. Therefore, the constructed x^p_{src} are all usable. As a result, both source surrogates show significant improvements in numerical results. Furthermore, we provide new qualitative results in Fig. 20, including rotation as a source surrogate. The qualitative results show that LIA produces similar results to Face-Vid2Vid, while the ability of rotation to simulate the proxy source is inferior, which results in slightly poorer performance in the results, although it also exhibits decent performance. Overall, as long as the source surrogate can achieve the goal of maintaining ID to construct the proxy source x^p_{src}, it can yield comparable results.

5.7 Discussion: Influence of Poor Surrogates

We conduct two additional sets of experiments focusing on the artifact issues in the source and target surrogates, respectively, to investigate the impact of artifacts present in proxy data on the results. Specifically, we artificially create two models with evident artifacts in their generated results. For the source surrogate, we employ PIRender [Ren et al. 2021] due to its significant deviations in aligning key points, which result in noticeable artifacts in x^p_{src} on our proxy pairs. As for the target surrogate, we utilize a base model with poorly optimized training parameters, which leads to artifacts in the results. Images generated from the two poor surrogates are illustrated in Fig. 22.

We replace the according portion of the original proxy-paired dataset with proxy data generated by these two poor surrogates, each exhibiting evident artifacts. Subsequently, we conduct two sets of experiments, and the qualitative and quantitative results of these experiments can be observed in Fig. 23 and Table. 2, respectively. Next, we analyze the results of the two sets of experiments.

5.7.1 Influence of Poor Source Surrogate. For the poor source surrogate, although recognizing the proxy source as a person in Fig. 22 is difficult, the final results in Fig. 23 show that high-quality generations are still obtained. The reason is that the robust ID encoder of the face-swapping model is not tuned. However, the identity-preserving capability has weakened. Despite the decreased identity similarity in Table 2, the third row of Fig. 23 shows that the model generates an identity (Asian) that is completely unrelated to the source (African) and target (Caucasian). The reason is that x^p_{src} constructed through a poor source surrogate exhibits significant artifacts. In terms of pairing, it differs markedly from y^p_{out}. Thus, the exchange of identity information between them is invalid. Consequently, the model learns erroneous representations of identity.

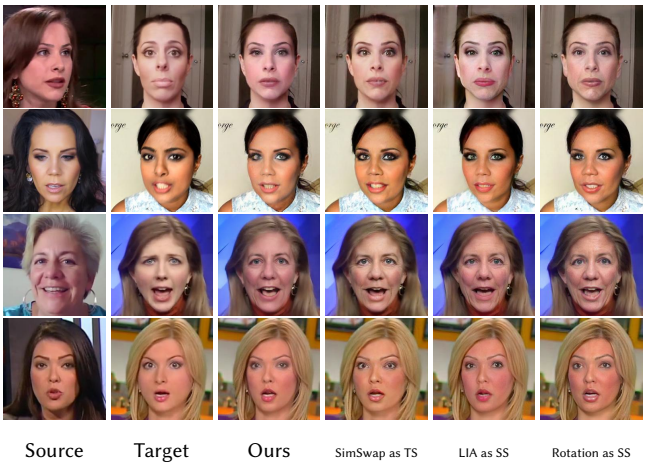


Fig. 21. Comparing different surrogates of their impact on final results. With more capability of shape transfer, the target surrogate InfoSwap performs better than SimSwap, especially on shape. Meanwhile, the results of LIA as SS are similar to ours, while rotation as SS performs slightly worse. TS means target surrogate and SS means source surrogate.

5.7.2 Influence of Poor Target Surrogate. For the poor target surrogate, given that a complete (x^p_{tgt}, y^p_{out}) pair cannot be constructed, the quality of the generated images exhibits a noticeable decrease. In addition, the generated results exhibit better identity preservation than the poor source surrogate, as shown in Fig. 23 and Table 2, due to the undisturbed flow of identity transmission. Meanwhile, compared with the target surrogate (a face-swapping model) used to generate x^p_{src} as illustrated in Fig. 22, our approach significantly improves the quality of the generated images. This finding provides further evidence of the effectiveness of CS^2.

The two experiments indicate that the artifacts stemming from the surrogate significantly impact the training results. Therefore, careful consideration is required when selecting surrogates to ensure alignment with expected performance. Nevertheless, even in scenarios where the constructed proxy-paired data exhibit noticeable artifacts, CS^2 exhibits inherent robustness, which is ensured by the use of real images as ground truth.

6 APPLICATIONS

The practice of face swapping has a diverse utility within the realm of digital face generation. We will illustrate the applications of our

Fig. 22. Proxy paired data produced by our setting and surrogates with apparent artifacts. SS means source surrogate; TS means target surrogate.

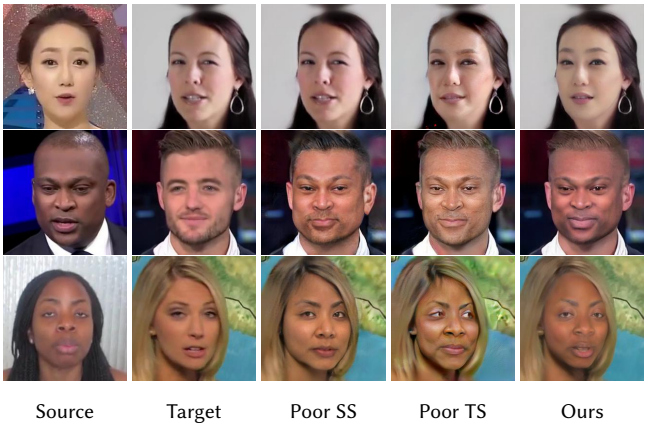


Fig. 23. Comparisons on surrogates with apparent artifacts. SS means source surrogate; TS means target surrogate.

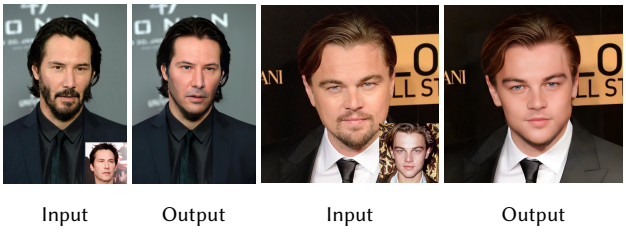


Fig. 24. Application I: re-aging. Re-aging results for Keanu Reeves and Leonardo DiCaprio. GFPGAN [Wang et al. 2021b] is applied to enhance the face quality. Note that the face restoration model may hurt identity preservation slightly.

approach involving specifically adapted swapping, re-aging, cross-domain swapping, and video face swapping.

6.1 Re-Aging

An important application of face-swapping is re-aging, which involves transforming the age of one face to become older or younger. Face re-aging is applied in the film industry, and face-swapping has certain advantages compared with other methods as it utilizes an image of the actor at a certain age for the purpose of re-aging. We collect some celebrities' pictures and apply our model for re-aging, and the results are illustrated in Fig. 24. To enhance image quality because of the limitation on 256 pixel resolution of our model, we utilize GFPGAN [Wang et al. 2021b] to restore the face resolution. This approach slightly wealems the identity-preserving capability. However, our method still obtains satisfactory results.

Fig. 25. Application-II: Customized adaptation on proxy-paired data for beard removal and glasses transfer. The first row displays the adapted pair for beard removal, as x^p_{src} has been edited to add more beard, while y^p_{gt} is removed. The second row is the pair for glasses transfer, where x^p_{src} and y^p_{gt} are both added glasses while x^p_{tgt} at the opposite.

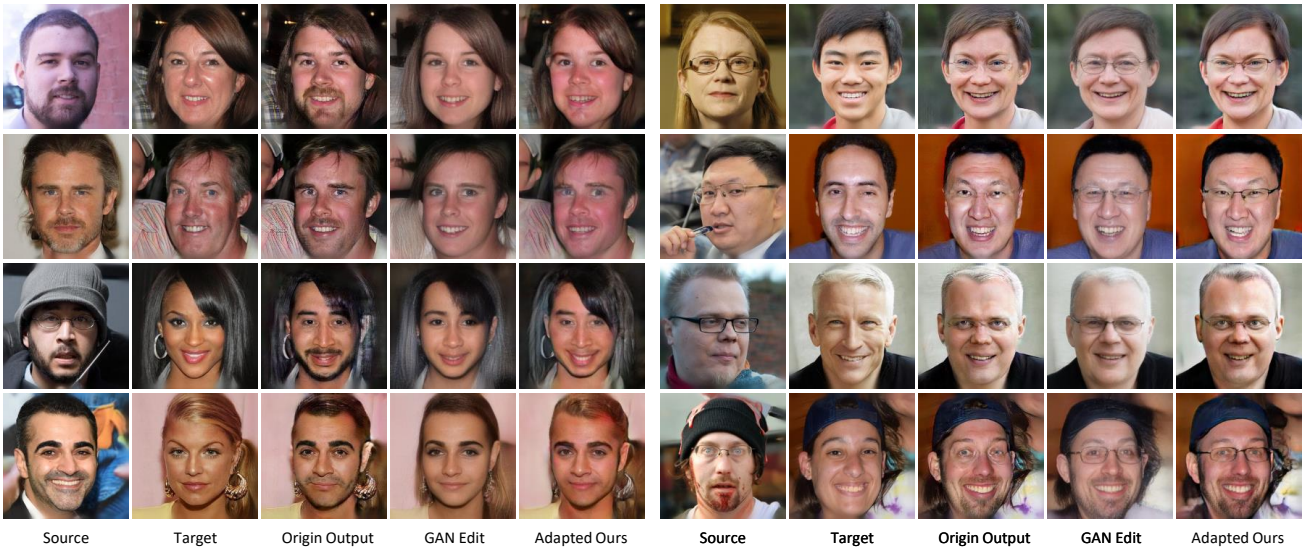
Table 3. Customized swapping compared with editing. Face editing models lower down the retrieval and similarity numbers, while our adapted models could obtain comparable identity-preserving results.

Method	CosFace Ret.↑	FaceNet Ret.↑	CosFace Sim.↑	FaceNet Sim.↑
Beard remove comparison				
Origin Ours	0.967	0.905	0.697	0.755
GAN Edit	0.169	0.074	0.221	0.375
Adapted Ours	0.966	0.888	0.623	0.726
Glasses transfer comparison				
Origin Ours	0.962	0.853	0.623	0.732
GAN Edit	0.317	0.256	0.270	0.463
Adapted Ours	0.973	0.895	0.658	0.756

6.2 Swappable Attribute Customization

Compared with current face-swapping methods, we propose a configurable formulation for face-swapping tasks from the perspective of identity preservation. To investigate the potential applications of our framework, we set two different real-world customized settings, namely, beard removal and glasses transfer. We also implement suitable adaptations to adapt to these conditions, as described in Sec. 3.1 and Sec. 4.2. Following the default configuration of identity in Sec. 3.1, the two configurations can be formulated as:

- *Config of bread removal*: $Attr_{id} = Attr^D_{id} - \{\text{ bread }\}$ and $Attr_{nid} = Attr^D_{nid} \cup \{\text{ bread }\}$,
- *Config of glasses transfer*: $Attr_{id} = Attr^D_{id} \cup \{\text{ glasses }\}$ and $Attr_{nid} = Attr^D_{nid} - \{\text{ glasses }\}$.



(a) Beard remove results. (b) Glasses transfer results.

Fig. 26. Application-II: swappable attribute customization. On the foundation of the model's capacity to accomplish these specific functions like beard removal and glasses transfer, ours get better identity-preserving results compared with other GAN editing methods.

In detail, we add beard on x^p_{src} and remove it on y^p_{gt}, as shown in the first row of Fig. 25. For the glasses transfer, x^p_{src} and y^p_{gt} are attached to glasses in the second row of Fig. 25. The face-swapping model is fine-tuned with adapted data for five epochs, and the results are shown in Figs. 26a and 26b.

We compare our fine-tuned models with the image editing methods by applying them to the origin output, and the qualitative and quantitative results are displayed. To compute quantitative results, we conduct 2K swapping pairs in the FFHQ test split and compute the ID retrieval rate and cosine similarity. The results show that our adapted model allows for better preservation of identity information while completing the intended usage. In some cases, it may obtain higher identity numerical results in Fig. 26b, which reflects that glasses transfer could benefit identity preservation as well.

6.3 Cross-Domain Results

Another usual application is cross-domain face swapping. Many image domains of the face, excluding realism, include art paintings, sculptures, and cartoons. Face swapping between different face domains has a board application prospect, and it poses a further challenge to the capability of the swapping model. Moreover, preserving identity between different domains in swapping is an unexplored question.

We present face-swapping results in different domains. MetFaces [Karras et al. 2020a], an image dataset of human faces extracted from works of art, is applied in this setting. We swapped real faces with artistic faces, or vice versa, and the results are illustrated in Fig. 27. Surprisingly, ours keeps a superior identity-preserving capability to comparison methods regardless of whether the source face is real or artwork.

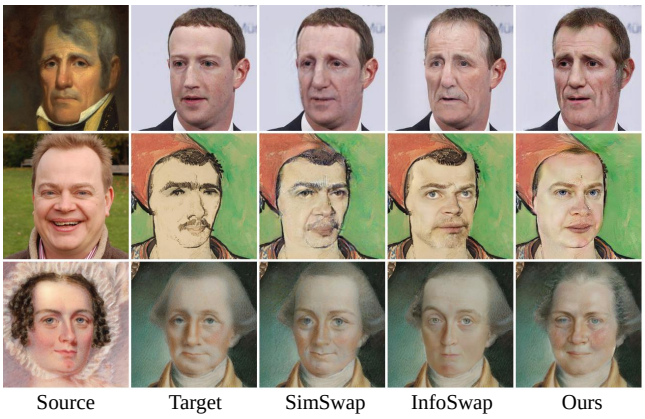


Fig. 27. Application-III: art-to-photo/photo-to-art/art-to-art swapping. Despite having different source and target domains, our method maintains a superior capability for preserving identity when compared qualitatively with others.

6.4 Video Face Swapping

Face swapping needs to obtain stable video results, especially in identity preserving for videos. We add two video-level losses during training to improve video stability. The stitch tuning loss is borrowed from STIT [Tzaban et al. 2022]. We simplify it as

$$L_{stitch} = L_{rec}(x_{tgt} \cdot m_{bak}, y_{out} \cdot m_{bak}), \tag{13}$$

where m_{bak} is the background mask generated by a segmentation model. In this case, the head mask is applied. The loss is utilized

Fig. 28. Application-IV: video face swapping. The first row of each video result is the swapped face and the target frames are listed in the second row. Our method could get consistent results with pose and expression from the target preserved as described in Sec. 6.4.

after the face-swapping training when the capability of shape transferring is already ready. We apply element-wise *or* operator to the mask from x_{tgt} and y_{out} to obtain m_{bak} for face-shape transferring. Another loss is the cross-frame similarity loss [Liu et al. 2021b]:

$$L_{is} = L_{rec}(cos(f_{tgt}^{t}, f_{tgt}^{t+1}), cos(f_{out}^{t}, f_{out}^{t+1})), \tag{14}$$

where f_{tgt}^{t}, f_{tgt}^{t+1}, f_{out}^{t}, and f_{out}^{t+1} represent the model's feature map of x_{tgt} and y_{out} for frame t and $t+1$, which aims to keep the attribute difference between consecutive frames for enhancing frame consistency. To further reduce flickering defects, we apply a deflickering method [Lei et al. 2023] on the swapped video. Our results are illustrated in Fig. 28, where the identity preservation in videos is well performed and has a good performance on frame consistency. The video results can be found in the supplementary materials.

7 LIMITATION AND FUTURE WORK

Although the CS^2 framework achieves better identity-preserving face-swapping results, the proxy-paired data utilized in the framework remain imperfect and could introduce errors. Despite constructing better proxy pairs in one shot, training in a progressive approach to iteratively refine the model based on ongoing results could yield further improvements.

Our model occasionally produces unnatural results in the eye area, which is possibly attributable to the limited resolution of our model. It is set at 256 pixels, which may result in inadequate attention to small-scale eye regions. To alleviate this problem, we train a new model with 512 pixels, and the result is illustrated in Fig. 29. The comparison shows that the 512-pixel model performs better in small regions with natural eyes.

Another limitation is that while we have defined identity preservation and our framework allows for the configuration of identity as needed by users, certain attributes, such as hairstyle and skin color, cannot be fully generated according to the configuration in some cases due to the lack of suitable tools for constructing proxy-paired data. These issues might be alleviated with better portrait editing tools.

Fig. 29. Comparison with higher-resolution model. The eye region becomes better.

8 BORDER IMPACT

Face-swapping could be used maliciously to create DeepFake videos, which may bring adverse social impacts. We deeply understand the possible negative influence of face-swapping technology, and will strictly prevent the spread of our method. Besides, studying the face-swapping methods can help researchers build effective DeepFake detection tools [Gao et al. 2021; Rössler et al. 2019]. Apart from the negative usages, there still exists positive applications for face-swapping, including privacy protection and entertainment.

9 CONCLUSION

In this study, we propose the CS^2 framework to add explicit supervision to alleviate the identity-preserving problem caused by implicit supervision in face swapping. Our CS^2 approximates ground-truth paired data from a real face image by dual-creating surrogates. It sets the proxy-paired data as credible explicit supervision to guide

face-swapping training, which gives credible and effective direction to boost the identity-preserving capability. We further propose explicit adaptation on paired data and implicit adaptation on ID encoder, where explicit adaptation with handy adaptation enhances the capability for face-shape transferring and customized swapping. Moreover, implicit adaptation with a novel ID adapter architecture narrows the gap between face recognition and face swapping. Our method is easy to use and can be used as a plugin on different network architectures. Extensive experiments and user studies show that the face-swapping model trained using our framework achieves high identity-preserving and face-shape transferring results compared with state-of-the-art methods. It has comparable performance in target attribute consistency and generation quality. Furthermore, our method has a diverse range of applications.

ACKNOWLEDGMENTS

We thank the reviewers for their insightful comments and suggestions. This work was partly supported by the National Natural Science Foundation of China under No. 62102162, Beijing Science and Technology Plan Project under No. Z231100005923033, and the National Science and Technology Council under No. 111-2221-E-006-112-MY3, Taiwan.

REFERENCES

Dmitri Bitouk, Neeraj Kumar, Samreen Dhillon, Peter Belhumeur, and Shree K. Nayar. 2008. Face Swapping: Automatically Replacing Faces in Photographs. In *ACM SIGGRAPH 2008 Papers* (Los Angeles, California) *(SIGGRAPH '08)*. Association for Computing Machinery, New York, NY, USA, Article 39, 8 pages. https://doi.org/10.1145/1399504.1360638

Volker Blanz, Kristina Scherbaum, Thomas Vetter, and Hans-Peter Seidel. 2004. Exchanging Faces in Images. *Computer Graphics Forum* 23 (2004), 669–676.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*. ACM Press/Addison-Wesley Publishing Co., USA, 187–194.

Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. 2020. SimSwap: An Efficient Framework For High Fidelity Face Swapping. In *Proceedings of the 28th ACM International Conference on Multimedia (MM)*. Association for Computing Machinery, New York,NY,United States, 2003–2011.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 4690–4699.

Qixin Deng, Luming Ma, Aobo Jin, Huikun Bi, Binh Huy Le, and Zhigang Deng. 2021. Plausible 3d face wrinkle generation using variational autoencoders. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 28, 9 (2021), 3113–3125.

Zhigang Deng, U. Neumann, J.P. Lewis, Tae-Yong Kim, M. Bulut, and S. Narayanan. 2006. Expressive Facial Animation Synthesis by Learning Speech Coarticulation and Expression Spaces. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 12, 6 (2006), 1523–1534. https://doi.org/10.1109/TVCG.2006.90

Guillaume Francois, Pascal Gautron, Gaspard Breton, and Kadi Bouatouch. 2009. Image-Based Modeling of the Human Eye. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 15, 5 (2009), 815–827. https://doi.org/10.1109/TVCG.2009.24

Gege Gao, Huaibo Huang, Chaoyou Fu, Zhaoyang Li, and Ran He. 2021. Information bottleneck disentanglement for identity swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Nashville, TN, USA, 3404–3413.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2020. Generative adversarial networks. *Commun. ACM* 63, 11 (2020), 139–144.

Jianzhu Guo, Xiangyu Zhu, Yang Yang, Fan Yang, Zhen Lei, and Stan Z Li. 2020. Towards Fast, Accurate and Stable 3D Dense Face Alignment. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer International Publishing, Cham, 152–168.

Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao. 2016. MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition. In *European Conference on Computer Vision (ECCV)*. Springer International Publishing, Cham, 87–102.

Fangzhou Han, Shuquan Ye, Mingming He, Menglei Chai, and Jing Liao. 2021. Exemplar-based 3d portrait stylization. *IEEE Transactions on Visualization and Computer Graphics* 29, 2 (2021), 1371–1383.

Omer Kafri, Or Patashnik, Yuval Alaluf, and Daniel Cohen-Or. 2022. Stylefusion: Disentangling spatial segments in stylegan-generated images. *ACM Transactions on Graphics (TOG)* 41, 5 (2022), 1–15.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2018. Progressive Growing of GANs for Improved Quality, Stability, and Variation. In *Proceedings of International Conference on Learning Representations (ICLR) 2018*. https://iclr.cc/Conferences/2018 International Conference on Learning Representations, ICLR ; Conference date: 30-04-2018 Through 03-05-2018.

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020a. Training generative adversarial networks with limited data. *Advances in neural information processing systems (NeurIPS)* 33 (2020), 12104–12114.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Long Beach, CA, USA, 4401–4410.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2020b. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 8110–8119.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), 18661–18673.

Jiseob Kim, Jihoon Lee, and Byoung-Tak Zhang. 2022. Smooth-Swap: A Simple Enhancement for Face-Swapping with Smoothness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, New Orleans, LA, USA, 10779–10788.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

Chenyang Lei, Xuanchi Ren, Zhaoxiang Zhang, and Qifeng Chen. 2023. Blind Video Deflickering by Neural Filtering with a Flawed Atlas. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. 2020. Advancing high fidelity identity swapping for forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, Seattle, WA, USA, 5074–5083.

Moran Li, Haibin Huang, Yi Zheng, Mengtian Li, Nong Sang, and Chongyang Ma. 2022. Implicit Neural Deformation for Sparse-View Face Reconstruction. In *Computer Graphics Forum*, Vol. 41. Wiley Online Library, 601–610.

Jingwang Ling, Zhibo Wang, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. 2022. Semantically disentangled variational autoencoder for modeling 3d facial details. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2022).

Jingying Liu, Binyuan Hui, Kun Li, Yunke Liu, Yu-Kun Lai, Yuxiang Zhang, Yebin Liu, and Jingyu Yang. 2021a. Geometry-guided dense perspective network for speech-driven facial animation. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 28, 12 (2021), 4873–4886.

Songhua Liu, Tianwei Lin, Dongliang He, Fu Li, Meiling Wang, Xin Li, Zhengxing Sun, Qian Li, and Errui Ding. 2021b. AdaAttN: Revisit Attention Mechanism in Arbitrary Neural Style Transfer. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)* (2021), 6629–6638.

Yilong Liu, Chengwei Zheng, Feng Xu, Xin Tong, and Baining Guo. 2021c. Data-Driven 3D Neck Modeling and Animation. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 27, 7 (2021), 3226–3237. https://doi.org/10.1109/TVCG.2020.2967036

Zhian Liu, Maomao Li, Yong Zhang, Cairong Wang, Qi Zhang, Jue Wang, and Yongwei Nie. 2023. Fine-Grained Face Swapping via Regional GAN Inversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8578–8587.

Lars Mescheder, Andreas Geiger, and Sebastian Nowozin. 2018. Which training methods for GANs do actually converge?. In *International conference on machine learning (ICML)*. PMLR, 3481–3490.

B Mildenhall. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In *European conference on computer vision (ECCV)*.

Yisroel Mirsky and Wenke Lee. 2021. The creation and detection of deepfakes: A survey. *ACM Computing Surveys (CSUR)* 54, 1 (2021), 1–41.

Jacek Naruniec, Leonhard Helminger, Christopher Schroers, and Romann M Weber. 2020. High-resolution neural face swapping for visual effects. In *Computer Graphics Forum (CGF)*, Vol. 39. Wiley Online Library, 173–184.

Yuval Nirkin, Yosi Keller, and Tal Hassner. 2019. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 7184–7193.

Yuval Nirkin, Iacopo Masi, A. Tran, Tal Hassner, and Gérard G. Medioni. 2017. On Face Segmentation, Face Swapping, and Face Perception. *2018 13th IEEE International*

Conference on Automatic Face & Gesture Recognition (FG 2018) (2017), 98–105. https://api.semanticscholar.org/CorpusID:15169802

Christopher Otto, Jacek Naruniec, Leonhard Helminger, Thomas Etterlin, Graziana Mignone, Prashanth Chandran, Gaspard Zoss, Christopher Schroers, Markus Gross, Paulo Gotardo, et al. 2022. Learning Dynamic 3D Geometry and Texture for Video Face Swapping. In *Computer Graphics Forum (CGF)*, Vol. 41. Wiley Online Library, 611–622.

Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. 2017. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*. 506–519.

Yuru Pei and Hongbin Zha. 2007. Transferring of Speech Movements from Video to 3D Face Space. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 13, 1 (2007), 58–69. https://doi.org/10.1109/TVCG.2007.22

Ivan Perov, Daiheng Gao, Nikolay Chervoniy, Kunlin Liu, Sugasa Marangonda, Chris Umé, Mr Dpfks, Carl Shift Facenheim, Luis RP, Jian Jiang, et al. 2020. DeepFaceLab: Integrated, flexible and extensible face-swapping framework. *arXiv preprint arXiv:2005.05535* (2020).

Yurui Ren, Ge Li, Yuanqi Chen, Thomas H Li, and Shan Liu. 2021. Pirenderer: Controllable portrait image generation via semantic neural rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 13759–13768.

Andreas Rössler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. 2019. FaceForensics++: Learning to Detect Manipulated Facial Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Shunsuke Saito, Liwen Hu, Chongyang Ma, Hikaru Ibayashi, Linjie Luo, and Hao Li. 2018. 3D hair synthesis using volumetric variational autoencoders. *ACM Transactions on Graphics (TOG)* 37, 6 (2018), 1–12.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 815–823.

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. 2020. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44 (2020), 2004 – 2018. Issue 4.

Xinhui Song, Chen Liu, Youyi Zheng, Zunlei Feng, Lincheng Li, Kun Zhou, and Xin Yu. 2023. HairStyle Editing via Parametric Controllable Strokes. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2023), 1–14. https://doi.org/10.1109/TVCG.2023.3241894

Justus Thies, Michael Zollhofer, Marc Stamminger, Christian Theobalt, and Matthias Nießner. 2016. Face2face: Real-time face capture and reenactment of rgb videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2387–2395.

Rotem Tzaban, Ron Mokady, Rinon Gal, Amit Haim Bermano, and Daniel Cohen-Or. 2022. Stitch it in Time: GAN-Based Facial Editing of Real Videos. *SIGGRAPH Asia 2022 Conference Papers* (2022).

Can Wang, Menglei Chai, Mingming He, Dongdong Chen, and Jing Liao. 2022. Cross-domain and disentangled face manipulation with 3d guidance. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 29, 4 (2022), 2053–2066.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5265–5274.

Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021c. One-shot free-view neural talking-head synthesis for video conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10039–10049.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. 2021b. Towards Real-World Blind Face Restoration with Generative Facial Prior. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. 2021a. HifiFace: 3D Shape and Semantic Prior Guided High Fidelity Face Swapping. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, (IJCAI)*, Zhi-Hua Zhou (Ed.). International Joint Conferences on Artificial Intelligence Organization, 1136–1142. https://doi.org/10.24963/ijcai.2021/157 Main Track.

Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. 2021d. Latent Image Animator: Learning to Animate Images via Latent Space Navigation. In *International Conference on Learning Representations (ICLR)*.

Xin Wen, Miao Wang, Christian Richardt, Ze-Yin Chen, and Shi-Min Hu. 2020. Photorealistic audio-driven video portraits. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* 26, 12 (2020), 3457–3466.

Wenpeng Xiao, Cheng Xu, Jiajie Mai, Xuemiao Xu, Yue Li, Chengze Li, Xueting Liu, and Shengfeng He. 2022. Appearance-preserved Portrait-to-anime Translation via Proxy-guided Domain Adaptation. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2022), 1–17. https://doi.org/10.1109/TVCG.2022.3228707

Chao Xu, Jiangning Zhang, Miao Hua, Qian He, Zili Yi, and Yong Liu. 2022b. Region-aware face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7632–7641.

Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. 2022a. High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7642–7651.

Zipeng Ye, Mengfei Xia, Yanan Sun, Ran Yi, Minjing Yu, Juyong Zhang, Yu-Kun Lai, and Yong-Jin Liu. 2021. 3D-CariGAN: an end-to-end solution to 3D caricature generation from normal face photos. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2021).

Ran Yi, Zipeng Ye, Ruoyu Fan, Yezhi Shu, Yong-Jin Liu, Yu-Kun Lai, and Paul L Rosin. 2022. Animating portrait line drawings from a single face photo and a speech signal. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–8.

Jie Zhang, Kangneng Zhou, Yan Luximon, Tong-Yee Lee, and Ping Li. 2023. MeshWGAN: Mesh-to-Mesh Wasserstein GAN With Multi-Task Gradient Penalty for 3D Facial Geometric Age Transformation. *IEEE Transactions on Visualization and Computer Graphics (TVCG))* (2023), 1–14. https://doi.org/10.1109/TVCG.2023.3284500

Lvmin Zhang and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543* (2023).

Yujian Zheng, Zirong Jin, Moran Li, Haibin Huang, Chongyang Ma, Shuguang Cui, and Xiaoguang Han. 2023. Hairstep: Transfer synthetic to real using strand and depth maps for single-view 3d hair modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12726–12735.

Wen-Yang Zhou, Lu Yuan, Shu-Yu Chen, Lin Gao, and Shi-Min Hu. 2023. LC-NeRF: Local Controllable Face Generation in Neural Radiance Field. *IEEE Transactions on Visualization and Computer Graphics (TVCG)* (2023), 1–12. https://doi.org/10.1109/TVCG.2023.3293653

Yuhao Zhu, Qi Li, Jian Wang, Cheng-Zhong Xu, and Zhenan Sun. 2021. One shot face swapping on megapixels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4834–4844.

Received 31 September 2023; revised 3 April 2024; accepted 24 June 2024