IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 13 August 2025, accepted 2 September 2025, date of publication 5 September 2025,
date of current version 12 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3606518

RESEARCH ARTICLE

Real-Time, High-Fidelity Face Identity Swapping With a Vision Foundation Model

JONGMIN YU[1,2], (Member, IEEE), HYEONTAEK OH[1], (Member, IEEE), ZHONGTIAN SUN[2,3], YOUNKWAN LEE[4], AND JINHONG YANG[1,5], (Member, IEEE)

[1]ProjectG.AI Research, ProjectG.AI, Yuseong-gu, Daejeon 31412, South Korea
[2]Department of Applied Mathematics and Theoretical Physics, Centre for Mathematical Sciences, University of Cambridge, CB3 0WA Cambridge, U.K.
[3]School of Computing, University of Kent, Kent, CT2 7NZ Canterbury, U.K.
[4]Smart Factory Team, Global Technology Research, Samsung Electronics, Suwon-si 16677, South Korea
[5]Department of Medical Information Technology, Inje University, Gimhae-si 50834, South Korea

Corresponding author: Jinhong Yang (jinhong@inje.ac.kr)

This work was supported by the Institute of Information and Communications Technology Planning and Evaluation (IITP)-Innovative Human Resource Development for Local Intellectualization Program Grant funded by Korean Government through Ministry of Science and ICT (MSIT) under Grant IITP-2025-RS-2024-00436773.

ABSTRACT Many recent face-swapping methods based on generative adversarial networks (GANs) or autoencoders achieve strong performance under constrained conditions but degrade significantly in high-resolution or extreme pose scenarios. Moreover, most existing models generate outputs at limited resolutions (128 × 128), which fall short of modern visual standards. Diffusion-based approaches have shown promise in handling such challenges, but are computationally intensive and unsuitable for real-time applications. In this work, we propose FaceChanger, a real-time face identity swap framework designed to enhance robustness across various poses and outputs at 256 × 256 (double the linear resolution of typical 128 × 128 baselines). While maintaining compatibility with conventional GAN- and autoencoder-based pipelines, FaceChanger uniquely incorporates a vision foundation model (VFM) to extract richer semantic features, which can enhance identity preservation, attribute control, and robustness to variations. In this work, we employ the Contrastive Language-Image Pre-training (CLIP) model to obtain the features. These features guide identity preservation and attribute control through newly designed VFM-based visual and textual semantic contrastive losses. Extensive evaluations on benchmarks such as the FaceForensics++ (FF++) dataset, the Multiple Pose, Illumination, and Expression (MPIE) dataset, and the large-pose Flickr face (LPFF) dataset demonstrate that FaceChanger matches or exceeds state-of-the-art performance under standard conditions and significantly outperforms them in high-resolution, pose-intensive scenarios.

INDEX TERMS Face identity swap, face swap, vision foundation model, contrastive learning.

I. INTRODUCTION

Face swapping refers to replacing the identity of one person in an image or video with that of another. This technology holds significant promise in various computer vision applications, particularly within the visual arts and entertainment industries [1], [2]. However, it raises ethical concerns, with regard to its potential misuse in scams, identity theft, and the generation of non-consensual content [3]. Despite these risks, ongoing research on facial skin application remains essential due to its technical potential and substantial societal implications, as well as understanding the methodologies involved in the development of detection technologies for deep-fake content [4], [5].

The core objective of face identity swapping is to accurately transfer the identity of a source face image while preserving nonidentity-related attributes of the target face image, such as skin texture, lighting, hairstyle, and facial accessories [6]. With the rapid advancement of deep learning, even though their output resolutions are restricted to 128 × 128, face-swapping methods have become significantly more realistic [7], [8], [9]. In particular, the recently

The associate editor coordinating the review of this manuscript and approving it for publication was Turgay Celik.

proposed method based on diffusion models can generate extremely high-fidelity and high-resolution face swap results, which are often indistinguishable from authentic images in controlled settings [10], [11].

However, despite recent state-of-the-art methods showing incredible swapping results in their controlled environment, face swapping under extreme facial poses continues to pose a formidable challenge. Extreme facial poses distort facial geometry to the extent that traditional boundary-based representations fail to localise key features accurately. As illustrated in Figure 1, while frontal faces (e.g., 0°) provide well-defined boundaries that encapsulate all key landmarks (eyes, nose, mouth), large pose angles (e.g., ±90° or even ±45°) introduce significant occlusions and asymmetries. These effects complicate the identity mapping process, often leading to unnatural deformations when methods attempt to project the source identity onto a distorted or partially visible facial structure. As demonstrated in Figure 1, even recent methods such as [6], [7], and [10] struggle to produce seamless results under such conditions. Additionally, although it is a minor issue, logically, their output resolution is a critical one. The resolutions of recently released visual content are being increased. We can easily find 4K or 8K content. In this circumstance, the limited output resolution (128 × 128) of previous methods may not be sufficient.

Several approaches have been proposed to mitigate these issues by incorporating explicit geometric priors. Li et al. [12] utilised facial landmarks to encode component-level spatial information. Wang et al. [13] employed 3D supervision to model head pose and facial geometry. Rosberg et al. [14] proposed an interpretive regularisation scheme that emphasises pose preservation through feature similarity. As shown in Figure 1, those approaches seem to work well on extreme face poses; nevertheless, these methods often fail when facial silhouettes are heavily altered due to pose-induced distortions. For examples, see +90° and −90° cases of FaceDancer in Figure 1. The boundaries between face and background are corrupted. Diffusion model-based methods generate even worse results. Also, due to the iterative nature of their sampling process (for example, Diffswap [10], which utilises denoising UNet architectures, requires approximately 40 seconds to generate results), diffusion-based methods incur substantial computational costs that preclude real-time deployment.

To improve pose-robustness while providing fast execution speed, we introduce FaceChanger, a novel, real-time, and pose-robust face identity swapping framework. To maintain real-time performance, FaceChanger benchmarks the architectural characteristics of methods that do not utilise diffusion-based approaches. FaceChanger comprises three key modules: a source ID encoder, a progressive face encoder, and a face generator. The source ID encoder extracts latent features to describe the source identity. The progressive face encoder extracts latent features from an image and progressively combines the features with the source identity code extracted by the source ID encoder. FaceChanger

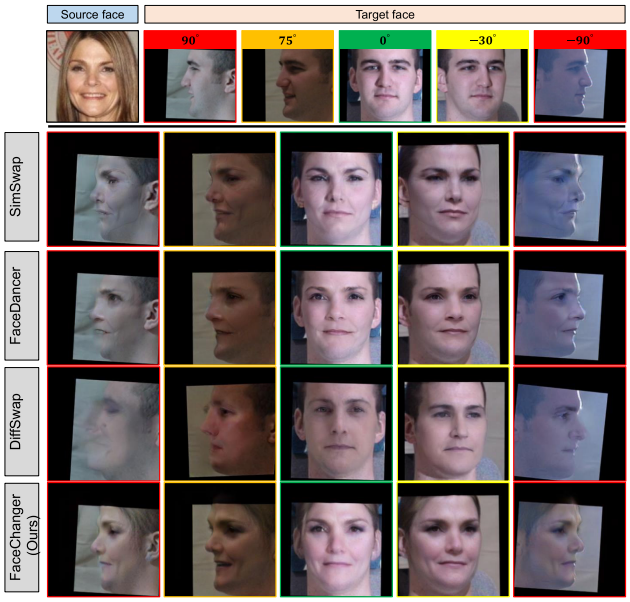

FIGURE 1. Examples of the results of face identity swapping obtained by recent SOTA methods [10], [13], [14], [17]. Compared to the frontal face image, the swapping results for extreme poses (greater than ±45 degrees) are still highly distorted. The proposed FaceChanger generates more detailed results that are robust to the face pose.

explicitly injects source identity information by multiplying and adding the source identity code to the latent features of multiple convolutional layers. After that, the face generator upscales outputs of the progressive face encoder and finally generates a swapped face image. In building the progressive face encoder and the swapped face generator, we consider a larger network scale to improve the output resolution which is 256 × 256. This size is twice as big as the majority of the resolutions of existing methods.

In this work, to enhance generalisation performance by feeding more semantic supervisory information while training the FaceChanger, we propose an approach that cooperates with vision foundation models (VFMs), such as Contrastive Language-Image Pre-training (CLIP) [15] or self DIstillation, NO labels (DINO) [16]. Vision foundation models are large-scale pretrained models that learn general-purpose visual representations, enabling strong performance on various downstream tasks with little to no task-specific training. In this work, we generate the text description of the target image and conduct text-image contrastive learning. Additionally, by using the CLIP's image encoder, we extract the source face and swapped face latent features and minimise the distance between both features. This approach enables FaceChanger to explicitly explore semantic representation by providing explicit semantic supervision, such as text description, thereby improving the description of source identity while preserving the target attribute.

The experimental results validate the effectiveness of the learning approach in conjunction with VFMs. FaceChanger is more robust to face poses without the use of explicit geometric features or diffusion models. Experimental results

in the FaceForensics++ (FF++) dataset [18] show that FaceChanger outperforms recent state-of-the-art (SOTA) methods, including diffusion-based methods. FaceChanger achieves 98.63 identity retrieval score, 1.31 pose error, and 2.11 expression error metrics, respectively. In extreme face pose experiments using the Multiple Pose, Illumination, and Expression (MPIE) dataset [19] and the large-pose Flickr face (LPFF) dataset [20] datasets, FaceChanger achieves a mean cosine similarity score of 0.457, with a pose error of 3.39 and expression error of 3.05, which are the best performances. The qualitative results presented in the Section IV also show that FaceChanger generates swapped faces of good quality under extreme facial pose cases. These results demonstrate that FaceChanger, trained with VFM, offers a more stable and accurate representation of source identity and preservation of target attributes, with improved robustness for face poses without increasing computational complexity, so that it can provide real-time performance. Overall, FaceChanger provides a promising foundation for future advances in pose-invariant face synthesis, deepfake detection, and identity-aware generative modelling.

The key contribution of our work is summarised as follows:

- FaceChanger is a real-time and high-fidelity face identity swapping method. FaceChanger provides real-time (similar to conventional face identity swap methods), high-fidelity (outperforming existing diffusion-based methods), higher resolution (256 × 256) face-swapping performance.
- A novel training approach for face identity swapping using a VFM. In particular, we formulate text2face contractive loss and identity-preserving loss using CLIP.
- Comprehensive experimental results that help in improving the understanding of how to leverage CLIP to enhance the face identity swapping tasks.

The remaining part of this paper is organised as follows. We introduce related work in Section II. In Section III, we provide detailed information about architectural details and loss functions of FaceChanger. We describe experimental settings and results in Section IV. This paper is concluded in Section V.

II. RELATED WORK

Face identity swapping has advanced tremendously in the deep learning era. In particular, Generative Adversarial Network (GAN) and autoencoder-based systems such as FaceShifter [21], Face Swapping GAN (FSGAN) [17], and SimSwap/SimSwap++ [7], [8] have demonstrated remarkable performance. Despite variations in their specific architectures, these methods share a common methodological paradigm: they first extract latent identity features from the source face image using an identity (ID) encoder, then combine these features with latent attributes from the target face image to generate a swapped face.

Within this paradigm, primary research contributions in the literature have focused on two main objectives: 1) enhancing source identity preservation, and 2) maintaining identity-irrelevant attributes of the target image. Addressing these objectives has involved developing advanced architectural components, such as the semantic-guided fusion module [22] and ID injection modules [7], [8], as well as formulating novel objective functions.

To enhance source identity preservation, FaceSwapper [22] regularised identity codes using a mixed-domain pre-training strategy. FaceSwapper also introduced a semantic-guided fusion module, integrating identity and attribute information via multiple semantic-guided face swapping blocks. BlendFace [6] highlighted that naïve identity encoders often absorb redundant contextual details (e.g., hair and background) and proposed synthetic hard-negative mining to disentangle identity-specific cues better.

Regarding the preservation of target identity-irrelevant attributes, such as skin tone, makeup, accessories, and illumination, research has predominantly focused on combining pixel and space reconstruction with perceptual losses [1], [6], [21], [22]. Typically, these methods utilised pre-trained Visual Geometry Group (VGG) networks to extract abstract feature representations for perceptual loss computation. The SimSwap series [7], [8] proposed a weakly supervised feature-matching loss that leverages adversarial learning with a discriminator, employing latent feature encoders for perceptual loss computation, thereby serving a similar role to the VGG-based approach.

Additionally, preserving the facial pose of the target face image remains a critical yet challenging issue. As illustrated in Figure 1, SimSwap, one of the recent SOTA methods, struggled to generate high-quality swapped faces under extreme pose conditions, particularly for faces rotated beyond 75 degrees. To mitigate these challenges, injecting explicit 2D or 3D geometric information is a commonly adopted approach. For instance, HiFiFace [13] warped local texture patches based on fitted 3D Morphable Models (3DMMs), while FaceDancer [14] introduced interpretability regularisation for maintaining pose consistency. Although these methods outperformed purely appearance-based models at moderate angles, issues, like silhouette tearing and texture stretching, persist at extreme angles exceeding $\pm 75°$, and still remain (see Figure 1).

Also, technically, the output resolution of those methods is starting to be a problem. All of the above methods take 128 × 128 as their output resolution. As the consideration of early 2010s' visual content resolution and scale of face, 128 × 128 was enough because the majority of visual content was smaller or the same as high definition (HD) (1280 × 720). However, we can easily find quad high definition (QHD) (2560 × 1440) contents, 4K, or even 8K contents are quite normal. In this trend, the output resolution of 128 × 128 is not enough. To provide more detailed texture and face component descriptions, it is essential to increase the output resolution.

More recently, diffusion models (DMs) have begun supplanting GANs in many generative tasks due to their

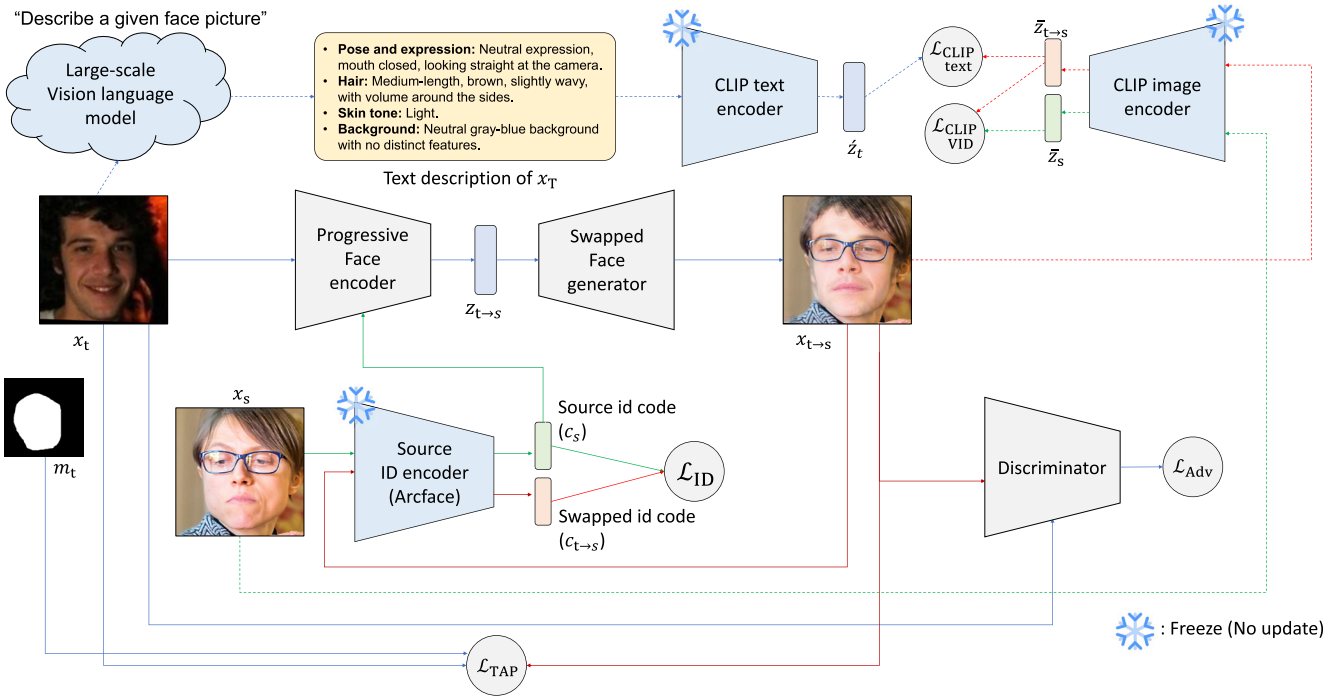

FIGURE 2. The workflow for training FaceChanger. The blue solid arrow lines define the training workflow using the target face image x_T and the corresponding facial mask m_t. The green solid arrow lines denote the workflow of the source face image x_S. The red solid arrow lines indicate the workflow of the swapped face image $x_{T\to S}$. Each coloured dotted arrow lines define the workflow for x_T, x_S, and $x_{T\to S}$ for applying the CLIP-based losses. The source ID encoder and the two CLIP encoders for text and image are frozen, so those are not updated during training.

superior generative quality and diversity at diverse high resolutions. Several diffusion-based face-swapping methods have been proposed [10], [11], [23]. DiffSwap [10] pioneered identity-conditional denoising diffusion probabilistic models (DDPMs) for swapping but requires approximately 50 sampling steps, limiting its practical applicability. Subsequent methods, such as DiffFace [23] and REFace [11], introduced guidance sampling, inpainting training, and triplet-ID supervision, achieving relatively high-resolution results compared to GAN and autoencoder-based approaches. However, as shown in Figure 1, these diffusion methods still struggle with extreme face pose robustness. Moreover, their performance significantly depends on the number of denoising steps, making them unsuitable for real-time applications due to the iterative inference procedure.

In this work, we propose a novel face-swapping method named FaceChanger, designed to deliver real-time performance, high fidelity, and robustness to extreme facial poses, with two times higher output resolution (256 × 256). To ensure real-time capabilities, we maintain compatibility with conventional GAN- and autoencoder-based pipelines that do not incorporate explicit geometric features, such as facial landmarks or 3D facial depth maps. Instead, we leverage rich semantic information extracted from vision foundation models (VFMs) during training. Encoders from VFMs such as CLIP and DINO provide geometry- and semantics-rich representations.

Several recent studies exploit CLIP embeddings as auxiliary losses to enhance source identity representation: REFace [11] embedded pose and expression information into CLIP's latent space to guide diffusion sampling; REFace uses CLIP as a conditioning module in generating swapped images using a diffusion model. ClipSwap [24] uses an image encoder of CLIP, and it applies contrastive learning using swapped images only. Also, after generating a swapped image, ClipSwap uses it to extract latent features and feeds them into the swapping process to improve the source identity representation.

However, REFace is built based on the diffusion model, so it can not provide a real-time process. Also, ClipSwap's recursive process in generating the swapped face image increases computational cost. In this work, we formulate two contrastive learning tasks using the text and image encoders of CLIP. Using the text encoder, we encode the description of the target face image into a latent feature space and conduct contrastive learning with the latent features of the swapped face, which is extracted by the image encoder of CLIP. Also, using the image encoder of the CLIP, we formulate an identity similarity loss between the source and swapped images. Those approaches increase computational cost during training, but since they can provide rich semantic information, they improve the performance of the model and will not increase the computational cost for the testing phase. Our work demonstrates the effectiveness

of VFM in enhancing the performance of mainstream face swapping methods, which are GAN-based or autoencoder-based, providing real-time processing performance. It also initiates a discussion on how to utilise the rich information obtained from VFM.

III. FACECHANGER

A. METHODOLOGY DETAILS

FaceChanger is composed of the source identity (ID) encoder, a progressive face encoder, and a swapped face generator. The source ID encoder plays a role in extracting source ID information from source face images and feeding it into the main pipeline, which consists of the progressive face encoder and the swapped face generator, to swap face identities in target face images. The progressive face encoder extracts latent features from the target images and combines them with the source ID code. The output of the progressive face encoder is applied to the swapped face generator to generate the face image, which represents the source face identity, preserving the visual attributes that are irrelevant to describing the identity of the target face image. Figure 2 illustrates the workflow of FaceChanger. FaceChanger maintains the compatibility of conventional GAN- and autoencoder-based methods [17], [21], [25], which grant real-time capability. However, to improve the quality of swapping results and enhance robustness at higher resolutions, we newly apply complementary objective functions based on VFMs.

During training with FaceChanger, we utilise VFM's image and text encoders to provide FaceChanger with richer semantic information. VFMs employ broad, self-supervised, or contrastive objectives that encourage learning generalizable semantics. CLIP, for example, is trained with a contrastive image-text objective, aligning images and natural language descriptions in a shared embedding space [26]. By learning to match an image with its caption and distinguish it from others, CLIP' image encoder must capture high-level concepts that co-occur in text. This rich supervision via text imbues CLIP features with multi-faceted semantic information beyond a fixed label set. Particular domain models' objectives focus on a narrow label space (identity or fixed classes), whereas VFMs' objectives (contrastive, self-supervised) implicitly encourage learning a broader range of visual semantics (since the model must organise the data in terms of meaningful similarity beyond rigid class definitions). This fundamental difference is a key reason VFMs yield richer latent features. In this work, we use CLIP text and image encoders and formulate two contrastive learning losses: 1) CLIP-based text-to-image contrastive loss (CLIP-text) and 2) CLIP-based visual identity loss (CLIP-VID). We apply those losses as a complementarity term to improve the generalisation performances so that not only the visual quality of swapped faces but also the robustness of face pose.

B. ARCHITECTURAL DETAILS

As we mentioned above, FaceChanger is mainly composed of three key components: 1) the source ID encoder, 2) the progressive face encoder, and 3) the swapped face generator. In order to train FaceChanger, a discriminator and CLIP image and text encoders are also applied. However, those components do not involve inferring swapped faces; therefore, we do not consider them key components. In this section, we hence focus on introducing the three main components.

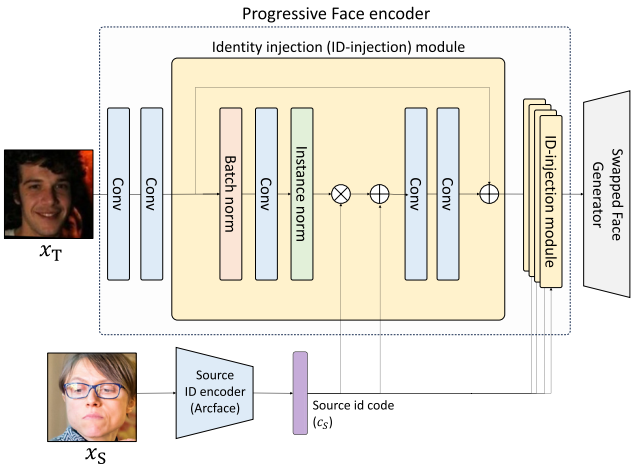

FIGURE 3. Structural details of the identity (ID) injection module and the progressive face encoder of FaceChanger.

Regarding the source ID encoder, it plays a role in delivering the latent feature extracted from the source face image to the main pipeline, which generates swapped faces. Therefore, the source ID encoder should provide a well-generalised but identity-distinguishable representation. A large-scale face dataset is essential to build a good source ID encoder. Most of the face identity swap studies consider a pre-trained network using a face recognition dataset as the encoder. We employ ArcFace [27] as the source ID encoder. ArcFace is the most popular model selected by various SOTA face identity swapping methods until now [7], [8], [11], [21], [23], [28]. It is even selected for developing a foundation model for face recognition [29].

The progressive face encoder extracts abstracted latent features, combining the source identity information and target identity-irrelevant information. The key function of the progressive face encoder is to inject source identity information. There are various approaches to injecting source information and combining it with the latent features extracted from target information. Simply concatenating [1], attention mechanism [14], [21], and a transformer [30] have been proposed.

Simply concatenating two latent features seems very straightforward and easy to apply without any major architectural revision. However, it is not suitable because it is highly coupled with identity information and some non-identity-related information. Concatenating two latent features may degrade the quality of face-swapping results. The transformer-based approaches are computationally cumbersome, which is not suitable for real-time performance.

Currently, attention mechanisms are a predominant approach for injecting source identity information. It is easy to implement and demonstrate the efficiency of various SOTA methods. As a result, we developed our ID injection module based on attention mechanisms.

FaceChanger's ID injection module is composed of convolutional layers and identity injection (ID-injection) modules. The ID injection module is designed to learn joint representation by combining the latent features extracted from the source face image and the target face image. In particular, the ID injection modules consist of two different types of normalisation layers and three convolutional layers. For the normalisation, the ID-injection module conducts batch normalisation (BN) first to improve the generalisation performances, and after that, the instance normalisation (IN) is produced. Using BN and IN together leverages the discriminative strength of batch statistics and the domain-invariant smoothing of instance statistics in a single layer. The result is a network that converges quickly like BN-only models but generalises across styles and domains like IN-only models.

The ID injection is as follows. First, we normalise the depth of the latent feature obtained by the source face image, which is matched with the latent features of the target face image. Next, we conduct channel-wise attention by multiplying the source and target latent features. After that, the output is added with the source identity feature again. The above ID-injection process is formulated as follows:

$$z_{\text{out}} = ((z_{\text{in}} \otimes c_{\text{S}}) \oplus c_{\text{S}}) \tag{1}$$

where z_{in} and z_{out} defines the input and the output for the ID injection operation. $\otimes$ and $\oplus$ denote the channel-wise multiplication and summation. z_{out} is applied to the convolutional layers to conduct further abstraction for the combined features.

These two steps make the output latent features take on a strong source identity, and it may disturb the swapped face by ignoring the attribute information of the target images. To find some balance between the source and target face image, the ID injection module carries out the residual operation using the input features at last. Figure 3 shows the architectural details of the ID injection module of the progressive face encoder. The output of the progressive face encoder is applied to the swapped face generator. The swapped face generator is built by stacking multiple deconvolutional layers to upscale the latent feature and generates swapped face images.

C. OBJECTIVE FUNCTIONS AND LEARNING

As shown in Figure 2, FaceChanger is trained by the objective functions combined with the five loss terms: 1) ID swapping loss, 2) Target attribute preserving loss, 3) Adversarial learning, 4) CLIP text encoder-based contrastive learning loss, and 5) CLIP image encoder-based ID swapping loss. The detailed explanation of those losses is as follows.

1) ID SWAPPING LOSS $\mathcal{L}_{\text{ID}}$

This encourages the swapped image $x_{\text{T}\to\text{S}}$ to have the same identity as x_{S}. We formulate the ID swapping loss using cosine angular similarity between $x_{t\to s}$ and x_s in the latent feature space, as follows:

$$\mathcal{L}_{\text{ID}}^{\text{t}\to\text{s}} = 1 - \frac{f_{\text{ID}}(x_s) \cdot f_{\text{ID}}(x_{t\to s})}{\|f_{\text{ID}}(x_s)\|_2 \|f_{\text{ID}}(x_{t\to s})\|_2}, \tag{2}$$

where f_{ID} indicates the face identity encoder defined by the Arcface [27]. $\cdot$ denotes the dot product between two latent features $f_{\text{ID}}(x_s)$ and $f_{\text{ID}}(x_{t\to s})$.

2) TARGET ATTRIBUTE PRESERVING LOSS $\mathcal{L}_{\text{TAP}}$

The perceptual fidelity of a face-swap result depends on two equally critical criteria: 1) the congruence of its identity with that of the source and 2) the faithful retention of identity-agnostic attributes such as illumination, cutaneous micro texture, and surrounding context. Attribute preservation cannot be guaranteed by identity losses alone. To promote the latter, we introduce a target attribute-preserving loss that combines two losses defined in a pixel space and one loss defined in the latent space.

The two loss functions defined under the pixel space are the masked reconstruction loss and cyclic reconstruction loss. The masked reconstruction loss is applied to explicitly learn visual information of the non-facial region of the target face image. Using the binary valued facial mask m_{t}, the masked reconstruction loss is defined by

$$\mathcal{L}_{\text{Recon}}^{\text{T}\to\text{S}}(x_{\text{t}\to\text{s}}, x_{\text{t}}) = \|(x_{\text{t}\to\text{s}} - x_{\text{t}}) \otimes (1 - m_{\text{t}})\|_1 . \tag{3}$$

However, strict spatial restrictions using the binary mask may cause performance degradation because it does not provide all the necessary information to rebuild the target appearance. In particular, to improve the pose robustness, it is essential to provide precise information for the natural boundary of the face and the background, which can not be provided by masked reconstruction loss.

To address this issue, we also add the cyclic reconstruction loss as follows:

$$\mathcal{L}_{\text{Cycle}}^{\text{T}\to\text{S}\to\text{T}}(x_{\text{t}\to\text{s}\to\text{t}}, x_{\text{t}}) = \|x_{\text{t}\to\text{s}\to\text{t}} - x_{\text{t}}\|_1 \tag{4}$$

where $x_{\text{t}\to\text{s}\to\text{t}}$ is face re-swapped results using the swapped image $x_{\text{t}\to\text{s}}$ as the target image x_{t} and the target image x_{t} as the source image x_{s}.

However, the above pixel-space reconstruction losses enforce exact colour-level fidelity. Still, they are overly sensitive to tiny misalignments, encourage blur by averaging plausible solutions, and lack semantic understanding of facial structures. Adding a perceptual loss computed on deep features (from a trained recognition network) supplies semantics-aware gradients that are robust to small shifts, directly align identity vectors, preserve fine textures, and regularise the generator against mode collapse. Combining both objectives leverages their complementary strengths: pixel losses maintain overall photometric consistency, while

perceptual loss guides the model to produce identity-faithful, sharp, and visually realistic swaps even under varying pose, illumination, or expression. Thus, we employ a perceptual loss using VGG network f_{VGG} defined as follows:

$$\mathcal{L}^{\text{T}\to\text{S}}_{\text{Percept}}(x_{\text{t}\to\text{s}}, x_{\text{t}}; f_{\text{VGG}}) = \sum_{i=1}^{\mathcal{N}_{\text{VGG}}} \left\| f^{i}_{\text{VGG}}(x_{t\to s}) - f^{i}_{\text{VGG}}(x_{\text{t}}) \right\|_2^2, \tag{5}$$

where $\mathcal{N}_{\text{VGG}}$ denotes the number of layers of the VGG network. The index i on f^{i}_{VGG} indicates i^{th} layer of the VGG network.

The total target attribute-preserving loss function is defined by combining the above three loss terms, and it is finally defined by

$$\mathcal{L}_{\text{TAP}} = \mathcal{L}^{\text{T}\to\text{S}}_{\text{Recon}} + \mathcal{L}^{\text{T}\to\text{S}\to\text{T}}_{\text{Cycle}} + \mathcal{L}^{\text{T}\to\text{S}}_{\text{Percept}}. \tag{6}$$

3) ADVERSARIAL LEARNING LOSS

Adversarial objectives are routinely employed to elevate the visual fidelity of identity-swapped faces, chiefly by restoring high-frequency cues, such as edge acuity and fine textural details, which govern visual sharpness [7], [21], [22]. We apply the adversarial loss represented by:

$$\mathcal{L}^{\text{T}\to\text{S}}_{\text{Adv}} = \mathbb{E}\left[\log\left(1 - D\left(x_{\text{t}\to\text{s}}\right)\right)\right] + \mathbb{E}\left[\log D\left(x_{\text{t}}\right)\right], \tag{7}$$

where D indicates the discriminator for the adversarial loss. In this work, we utilise the PatchGAN [31] to enhance the visual quality of the swapped face. Instead of normal GANs, which apply adversarial learning to single images, the PatchGAN formulates adversarial learning into multiple small-sized patches extracted from the single images, allowing it to derive more precise visual content.

4) CLIP BASED LOSSES $\mathcal{L}_{\text{CLIP-TEXT}}$ AND $\mathcal{L}_{\text{CLIP-VID}}$

The losses mentioned above are commonly used in the face identity swapping domains. Various SOTA methods derived from those losses achieve outstanding performances; however, as shown in Figure 1, these methods sometimes generate distorted results when obstacles cover the face and are also not robust to facial poses, despite structural advancements. Several reasons, including architectural intrinsic limitations and the quality of the dataset, contribute to these issues.

In this work, we address these issues by leveraging an additional supervisory signal, one that remains discriminative under occlusion, pose variance, and lighting drift. We posit that a contrastive loss is derived from VFMs. In particular, CLIP's joint vision–language embedding space fulfils this role. Unlike face image encoders such as ArcFace [27], which are trained solely on canonical, front-facing portraits, the image encoder of CLIP is exposed to billions of web-scale image–text pairs. Also, the text encoder learns a great amount of description to explain the detailed information, both the face identity and identity-irrelevant attributes, paired with the face image. As a result, CLIP learns semantically rich, pose-invariant descriptors that correlate global shape with fine-grained appearance.

Our CLIP-based contrastive learning is composed of two loss functions. The first is to align the image embedding extracted from the swapped image with a text prompt embedding obtained from the description of the target attribute (*e.g.,* descriptions for pose and expression, hair, skin tone, and background). This CLIP-based image-to-text contrastive learning loss $\mathcal{L}_{\text{CLIP-text}}$ is formulated as follows:

$$\mathcal{L}_{\text{CLIP-text}} = 1 - \frac{\left\langle \phi_{\text{img}}\left(x_{\text{t}\to\text{s}}\right),\ \phi_{\text{text}}\left(x_{\text{T}}\right)\right\rangle}{\left\|\phi_{\text{img}}\left(x_{\text{t}\to\text{s}}\right)\right\| \left\|\phi_{\text{text}}\left(x_{\text{T}}\right)\right\|}, \tag{8}$$

where ϕ_{img} and ϕ_{text} is the image and text encoders of CLIP. x_{T} is the description of the target face image. The description is automatically obtained by large vision models (LVMs) such as GPT or LLaMA. In this work, we use the GPT-4o [32] to obtain the description.

The second CLIP-based loss is similar to the ID swapping loss, but it is derived by using the image encoder of CLIP. The first CLIP-based loss (Eq. (8)) guides FaceChanger to learn richer semantic representations about identity-irrelevant information of the target face images. The second loss, called CLIP-based visual ID swapping loss $\mathcal{L}_{\text{CLIP-VID}}$, is applied as a complementary loss term to improve the representation of source identity by using the image encoder of CLIP. $\mathcal{L}_{\text{CLIP-VID}}$ is represented as follows:

$$\mathcal{L}_{\text{CLIP-VID}} = 1 - \frac{\left\langle \phi_{\text{img}}\left(x_{\text{t}\to\text{s}}\right),\ \phi_{\text{img}}\left(x_{\text{s}}\right)\right\rangle}{\left\|\phi_{\text{img}}\left(x_{\text{t}\to\text{s}}\right)\right\| \left\|\phi_{\text{img}}\left(x_{\text{s}}\right)\right\|}, \tag{9}$$

By using the above CLIP-based losses, we can feed richer semantic information during FaceChanger training, which consequently improves the visual quality and face pose robustness of the face identity-swapping model. We demonstrate the effectiveness of the two CLIP-based contrastive learning losses in our ablation study.

5) TOTAL OBJECTIVE

The overall optimisation criterion is constructed by linearly combining the previously defined losses, each modulated by a dedicated balancing weight, and by appending an l_1 weight-decay term to discourage overfitting. It is expressed as

$$\begin{aligned}\mathcal{L}_{\text{Total}} = {} & \lambda_{\text{ID}}\mathcal{L}_{\text{ID}} + \lambda_{\text{TAP}}\mathcal{L}_{\text{TAP}} + \lambda_{\text{Adv}}\mathcal{L}_{\text{Adv}} \\ & + \lambda_{\text{CLIP-text}}\mathcal{L}_{\text{CLIP-text}} \\ & + \lambda_{\text{CLIP-VID}}\mathcal{L}_{\text{CLIP-VID}},\end{aligned} \tag{10}$$

where λ_{IS}, λ_{TAP}, λ_{Adv}, $\lambda_{\text{CLIP-text}}$, and $\lambda_{\text{CLIP-VID}}$ define balancing weights for each term, respectively. The specific values adopted during training are detailed in Section IV.

During each optimisation step, the same loss is also computed for the reverse swap $x_{s\to t}$, effectively doubling the diversity of training instances and enhancing the network's ability to generalise across a wide spectrum of face-swap scenarios.

IV. EXPERIMENTS

A. DATASET AND EXPERIMENT PROTOCOL

1) DATASETS

In our experiments, we employ VGGFace2[1] [33] dataset and CelebA-HQ[2] [34] dataset for trainig FaceChanger. In performance estimation and comparison with existing SOTA methods, we use FaceForensics++[3] (FF++) [18] dataset, Multi Pose, Illumination, Expressions (MPIE)[4] [19] dataset and Large-pose Flickr Face (LPFF)[5] [20] dataset. The detailed information about the datasets is as follows.

- **VGGFace2** contains 3.31 million images of 9131 subjects (identities), with an average of 362.6 images for each subject. Images are downloaded from Google Image Search and have large variations in pose, age, illumination, ethnicity and profession (*e.g.,* actors, athletes, politicians). The whole dataset is split into a training set (including 8631 identities) and a test set (including 500 identities).
- **CelebA-HQ** is a visually enhanced version of the CelebFaces Attributes dataset (CelebA) [35], and it provides 30,000 images with 1024 × 1024 resolution.
- **FF++** is a forensics dataset consisting of 1000 original video sequences that have been manipulated with four automated face manipulation methods: Deepfakes, Face2Face, FaceSwap and NeuralTextures. The data has been sourced from 977 YouTube videos, and all videos contain a trackable, mostly frontal face without occlusions, which enables automated tampering methods to generate realistic forgeries.
- **LPFF** comprises 19,590 high-quality, numerous identities, and extensive-pose diversity images. They first collect 155,720 raw portrait images from Flickr, then they remove all the raw images that already appeared in FFHQ [36]. After that, they align the remaining facial images and remove low-resolution images as well as noisy and blurred images.
- **MPIE** contains over 750,000 images of 337 individuals. Each subject was photographed under 15 poses and 19 illumination conditions while exhibiting a range of facial expressions.

2) HOW TO OBTAIN TEXT DESCRIPTION T_T

To train our FaceChanger, it is essential to obtain the text description for each face image. As we mentioned above, we use GPT-4o [32] to obtain the description. All text descriptions are obtained by using the prompt *"Describe about pose, expression, facial attribute, and background about the given face image. The length of the description is 70 words."*. We generate txt files that match each face image file, storing the corresponding description. Since it is impractical to check the text descriptions of all images, we did not check the response of GPT-4o manually. However, since the text encoder of CLIP only accepts text less than 77 words, we conduct some filtering based on the word length, and the description containing only less than 70 words is accepted as a normal text description.

3) EVALUATION METRICS AND PROTOCOL

We assessed performance with four standard metrics in face-identity-swapping research: 1) cosine-similarity score (CSIM), 2) identity-retrieval accuracy, 3) pose-alignment error, and 4) expression-matching error. For the FF++ dataset, we replicated the evaluation protocol of Li et al. [21] and Chen et al. [7] to ensure a direct comparison. The MPIE and LPFF datasets lack widely accepted quantitative testbeds for this task, so our analysis on these datasets is primarily qualitative. We show and visualise various face swapped results and analyse their visual components. Nevertheless, because MPIE contains multiple images per subject, we adapted the FF++ evaluation pipeline as follows. One thousand source images were sampled at random from CelebA-HQ, while every MPIE portrait served as a target. The procedure was repeated ten times with different random seeds to mitigate identity-specific bias and yield more reliable statistics.

Additionally, since not many methods report their performance on MPIE and LPFF datasets, we compared our performance with that of methods that have released their projects in public repositories. FSGAN[6] [17], SimSwap[7] [7], BlandFace[8] [6], HifiFace[9] [13], Diffswap[10] [23], and FaceDancer[11] [14] are selected.

B. IMPLEMENTATION DETAILS

Data preprocessing proceeds in four stages. Firstly, we conduct face detection, localisation, and alignment. We detect facial bounding boxes with YOLO5Face [37] and refine them by five-point landmark alignment following Bulat and Tzimiropoulos [38]. Secondly, face geometry and mask extraction are produced. For every aligned crop, we estimate a depth map using DECA [39] and obtain a binary facial mask with the BiSeNet-based segmenter of Yu et al. [40]. At last, we carry out the resolution normalisation. All face images are resized to 256 × 256 to remove resolution variance. The source face images destined for the identity encoder are further down-sampled to 112 × 112. This is because we need to match the input dimensionality of Arcface [27], which we have selected for our source ID encoder. For the setting of the balancing weights, we set 1.0, 0.5, and 1.0 for λ_{ID}, λ_{TAP}, λ_{Adv}, respectively. For the balancing weight of the two CLIP-based losses, we set 1.0 and 1.0 for $\lambda_{\text{CLIP-text}}$ and $\lambda_{\text{CLIP-VID}}$,

[1] https://www.robots.ox.ac.uk/ vgg/data/vgg_face2/
[2] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html
[3] https://github.com/ondyari/FaceForensics
[4] https://www.kaggle.com/datasets/aliates/multi-pie
[5] https://github.com/oneThousand1000/LPFF-dataset
[6] https://github.com/YuvalNirkin/fsgan
[7] https://github.com/neuralchen/SimSwap
[8] https://github.com/mapooon/BlendFace
[9] https://github.com/maum-ai/hififace
[10] https://github.com/wl-zhao/DiffSwap
[11] https://github.com/felixrosberg/FaceDancer

respectively. Model training is carried out for 500 epochs on two NVIDIA RTX A6000 GPUs.

C. ABLATION STUDIES ON NORMALISATION LAYERS

Table 1 summarises the identity retrieval accuracy, pose error, and expression error measured on the FF++ dataset in terms of the normalisation layer settings. The variant of FaceChanger that uses Batch Normalisation (BN) or Instance Normalisation (IN) only yields relatively low results. FaceChanger without IN attains an ID retrieval score of 97.82, with pose and expression errors of 3.15 and 2.51, respectively. Conversely, excluding BN but retaining IN produces an ID retrieval score of 98.01, alongside pose and expression errors of 1.67 and 2.55. The configuration incorporating both BN and IN achieves the most favourable results, with an ID retrieval score of 98.63, a pose error of 1.31, and an expression error of 2.11.

Qualitative comparisons, illustrated in Figure 4, corroborate these findings. The FaceChanger using both IN and BN produces the best face swapping results. As shown in Figure 4, when IN and BN are applied together, the visual representation of the eye and mouth parts is most natural and similar to the source face image while it preserves the colour spectrum of the target face. It can be interpreted that the synergy of BN and IN enhances both identity preservation and attribute consistency, substantiating the effectiveness of the proposed FaceChanger architecture for high-fidelity face swapping.

D. ABLATION STUDIES ON CLIP-BASED LOSSES

Table 2 presents the quantitative evaluation results comparing two architectural variants of FaceChanger on the FF++ and MPIE datasets. Specifically, the table contrasts the performance of the baseline architecture ($\text{FaceChanger}_{\text{normal}}$) with the enhanced variant that incorporates CLIP-based features ($\text{FaceChanger}_{\text{CLIP}}$). Three metrics are reported for each dataset to assess model performance comprehensively: identity preservation accuracy, pose consistency, and expression consistency. For the FF++ dataset, ID retrieval accuracy evaluates how effectively the model maintains the source identity, while pose error and expression error measure the differences in facial pose and expression between the swapped output and the target face. Similarly, on the MPIE dataset, CSIM quantifies identity preservation as the cosine similarity between the identity embeddings of the swapped and source faces, complemented by pose and expression error metrics to ensure structural and semantic alignment.

The experimental results clearly demonstrate that incorporating CLIP features into the FaceChanger architecture yields consistent and notable improvements on pose and expression across all evaluation criteria. On the FF++ dataset, $\text{FaceChanger}_{\text{CLIP}}$ achieves an ID retrieval accuracy of 98.63, outperforming the baseline's 97.56, while also reducing the pose error from 2.82 to 1.31 and the expression error from 3.52 to 2.11. Similarly, on the MPIE dataset, the CSIM improves from 0.418 to 0.457, with pose error reduced from 4.18 to 3.39 and expression error reduced from 3.52 to 3.05. These results provide robust evidence that leveraging CLIP's rich semantic representations enhances the model's ability to preserve source identity features while simultaneously ensuring faithful reproduction of target facial poses and expressions, thereby demonstrating the effectiveness and generalisability of the proposed architectural modification across diverse datasets and evaluation protocols. Further experiments for comparing the performance of FaceChanger to the recently proposed SOTA methods have been done with FaceChanger trained with CLIP losses.

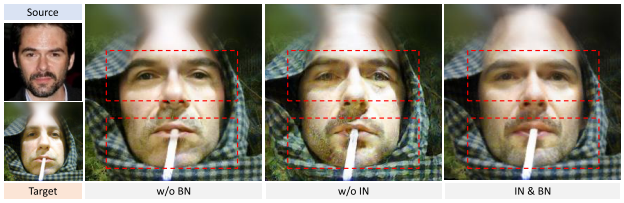

FIGURE 4. Qualitative results of face swapping using FaceChanger, depending on the normalisation layer settings. FaceChanger, when using IN and BN together, produces a more enhanced source identity representation and a more detailed target image attribute.

TABLE 1. Quantitative face swapping performance depending on the setting of instant normalisation (IN) and batch normalisation (BN) layers. The performances are evaluated using the FF++ dataset.

Settings	ID retrieval ↑	pose error ↓	expression error ↓
w/o BN & IN	93.05	3.85	3.96
w/o BN	97.82	3.15	2.51
w/o IN	98.01	1.67	2.55
BN & IN	**98.63**	**1.31**	**2.11**

TABLE 2. The quantitative results on the FF++ dataset [18] and MPIE dataset [19] regarding CLIP-based losses (Eq. 8 and Eq. 9) of the FaceChanger.

Experimental results on FF++ dataset			
Architectural setting	**ID retrieval ↑**	**pose error ↓**	**expression error ↓**
$\text{FaceChanger}_{\text{normal}}$	97.56	2.82	3.52
$\text{FaceChanger}_{\text{CLIP}}$	**98.63**	**1.31**	**2.11**
Experimental results on MPIE dataset			
Architectural setting	**CSIM ↑**	**pose error ↓**	**expression error ↓**
$\text{FaceChanger}_{\text{normal}}$	0.418	4.18	3.52
$\text{FaceChanger}_{\text{CLIP}}$	**0.457**	**3.39**	**3.05**

E. PERFORMANCE COMPARISON

1) GENERAL FACE POSE CASES

It is essential to confirm that FaceChanger provides competitive performance compared to existing SOTA methods based on commonly used benchmarks to demonstrate that our contribution using CLIP embedding does not bias the performance of the face identity swap for the extreme face angle cases only. Table 3 presents a detailed quantitative comparison of FaceChanger against various state-of-the-art face identity swapping methods on the FF++ dataset. Four

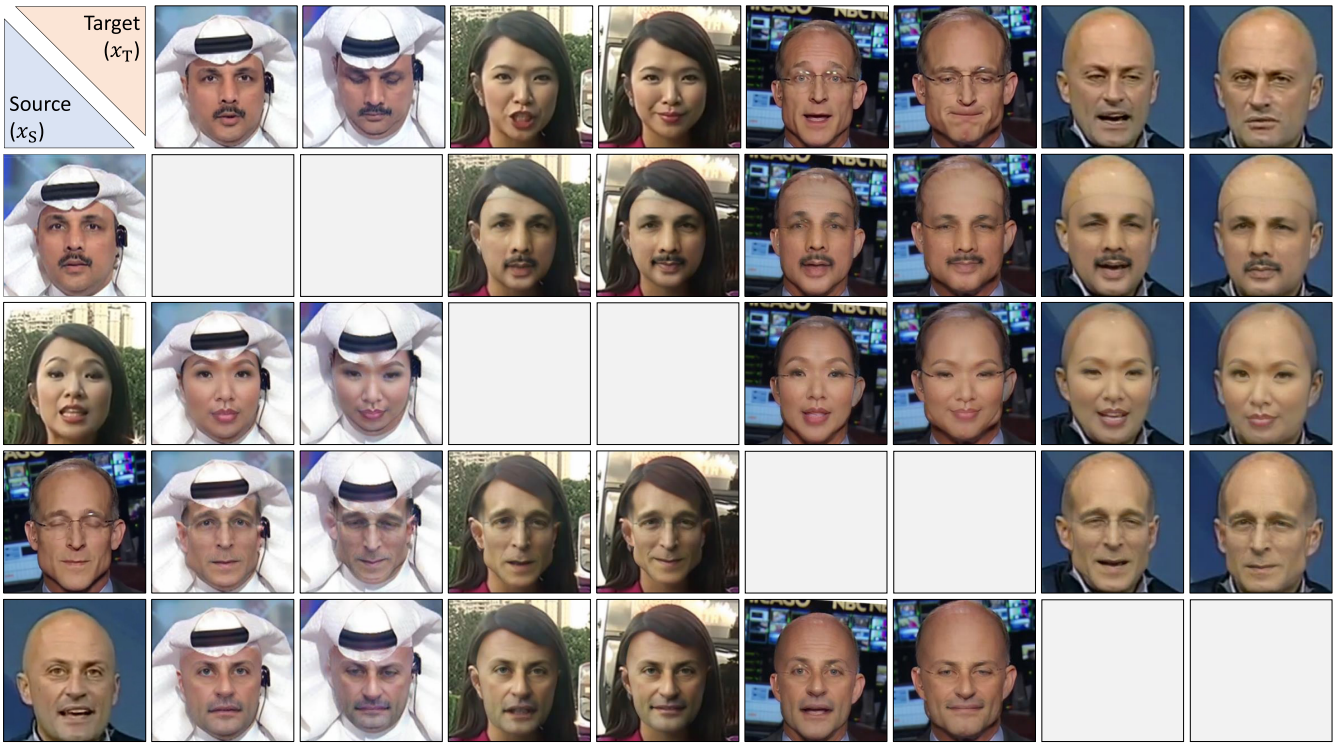

FIGURE 5. Qualitative results of FaceChanger on FF++ dataset [18].

TABLE 3. Quantitative examples of on the FF++ dataset [18]. Results of FaceSwap [41], DeepFakes [25], FaceShifter [42], MegaFS [43], FSLSD [44], RAFSwap [45], and FaceSwapper [22] are obtained from their websites or papers. † denotes that we ran officially released source codes to obtain the results.

Method	ID retrieval ↑	pose error ↓	expr error ↓	output resolution	exec speed (ms) ↓
FaceSwap [18]	72.69	2.58	2.89	128×128	-
DeepFakes [25]	88.39	4.64	3.33	128×128	-
FaceShifter [41]	90.68	2.55	2.82	128×128	-
MegaFS [42]	90.83	2.64	2.96	128×128	-
FSLSD [43]	90.05	2.46	2.79	128×128	-
RAFSwap [44]	92.54	3.21	3.60	128×128	-
FaceSwapper [22]	94.48	2.10	2.69	128×128	-
ClipSwap [22]	**98.91**	1.75	5.76	256×256	-
FSGAN† [17]	61.07	3.31	3.02	128×128	**21.5**
SimSwap† [7]	93.01	1.53	2.84	128×128	27.1
BlendFace† [6]	97.02	3.07	2.14	128×128	24.7
HifiFace† [13]	98.01	2.84	2.51	128×128	22.3
FaceDancer† [14]	98.84	2.04	7.97	128×128	726.5
DiffSwap† [10]	98.54	2.45	5.35	**512×512**	46000
FaceChanger (Our)	98.63	**1.31**	**2.11**	256×256	23.5

evaluation metrics are reported: ID retrieval accuracy, pose error, expression error, and execution speed. Figure 5 shows the quantitative results of FaceChanger on FF++ dataset.

Among all evaluated methods, FaceChanger offers the best balance of identity, pose, expression and speed: although ClipSwap attains the highest ID score (98.91), its expression error is 5.76 and can not ensure that it provides real-time processing or not; whereas FaceChanger delivers markedly lower pose/expression errors (1.31/2.11) in 23.5 ms The second-ranked method, FaceDancer, achieves a slightly higher ID retrieval accuracy of 98.84 but at the cost of a significantly higher expression error (7.97) and an execution speed of 726.5 milliseconds, rendering it impractical for real-time applications. The third-ranked method, DiffSwap, achieves an ID retrieval accuracy of 98.54, with pose and expression errors of 2.45 and 5.35, respectively. However, it requires 46 seconds per image, indicating prohibitively slow inference for practical deployment. In comparison, FaceChanger not only achieves competitive or superior identity preservation but also produces the lowest pose and

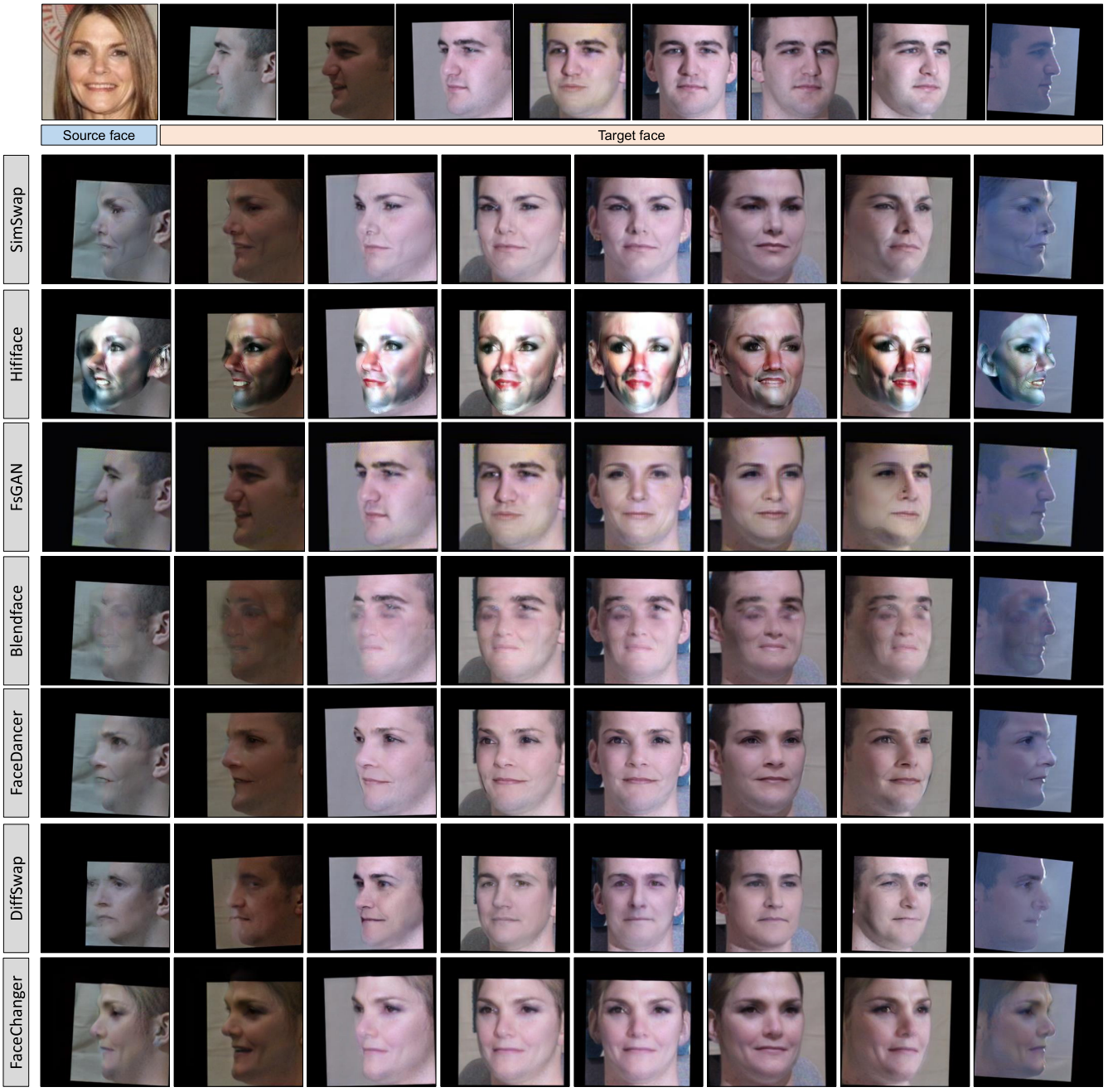

FIGURE 6. The face identity swapping results of the FaceChanger and other methods [6], [7], [10], [13], [14], [17] on MPIE dataset [19]. The black coloured area was generated during the data preprocessing using the face detection and alignment module.

expression errors across all methods evaluated, demonstrating its robustness in preserving target facial geometry and expressions. Furthermore, its fast inference speed positions it as a highly effective and practical solution for real-time face identity swapping tasks without compromising quantitative performance.

2) EXTREME FACE POSE CASES

Figure 6 and Table 4 summarise, respectively, the qualitative and quantitative outcomes on the MPIE benchmark. Table 4 presents a quantitative comparison of FaceChanger with existing state-of-the-art face identity swapping methods evaluated on the MPIE dataset. Among the evaluated baseline methods, FSGAN and HifiFace demonstrate low CSIM scores (0.105 and 0.092, respectively) with high pose and expression errors, indicating limited identity preservation and structural fidelity. SimSwap shows moderate performance with a CSIM of 0.180 and pose and expression errors of 3.92 and 3.81, respectively. More advanced methods, such as BlendFace and FaceDancer, achieve higher CSIM scores (0.392 and 0.401, respectively) and relatively lower pose and expression errors compared to early GAN-based approaches.

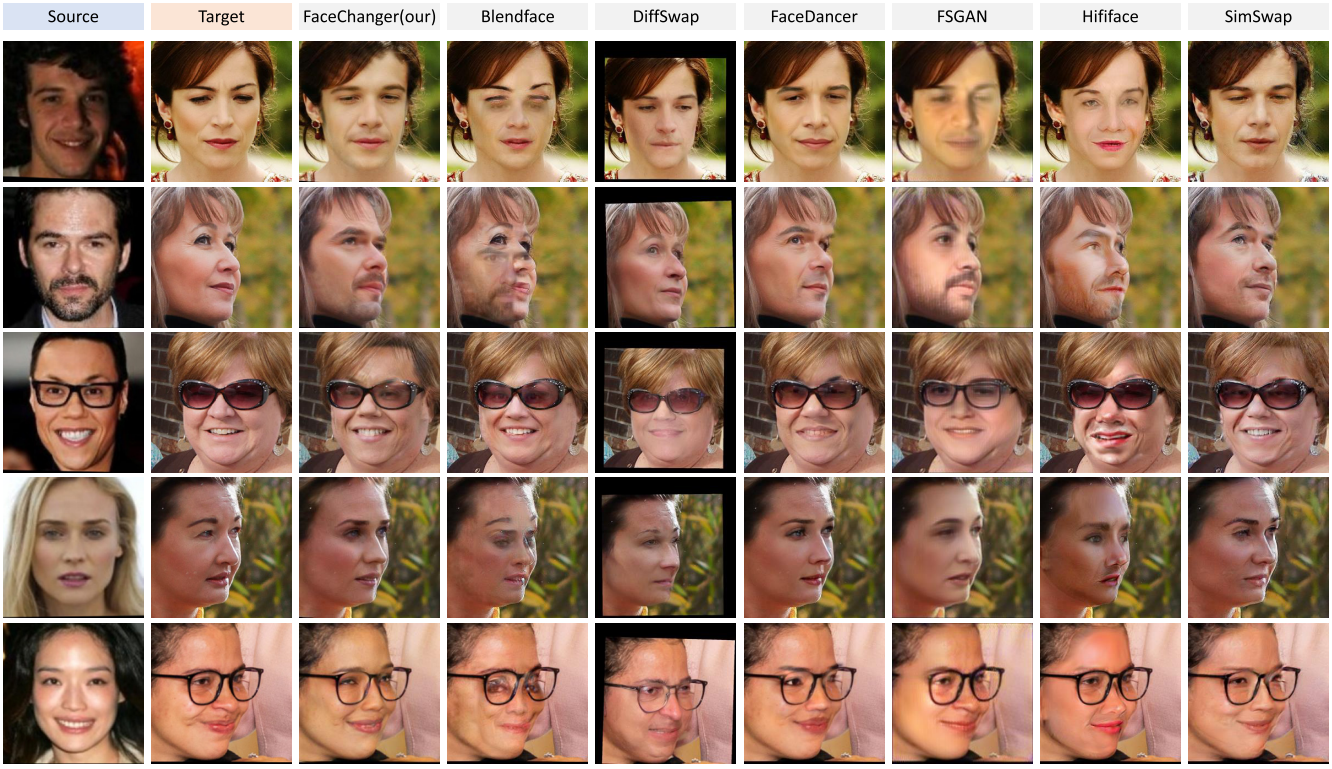

FIGURE 7. The face identity swapping result of the FaceChanger and other methods [6], [7], [10], [13], [14], [17] on LPFF dataset [20]. The black coloured areas on the results of DiffSwap are created because the publicly available source code for DiffSwap uses their own data preprocessing pipeline.

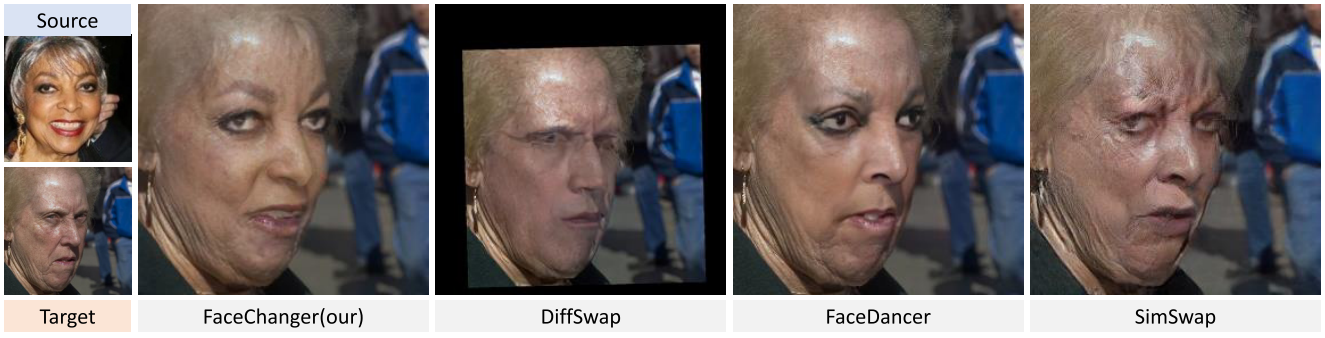

FIGURE 8. The face identity swapping result of the FaceChanger and some selected methods [7], [10], [13], [14]. FaceDancer [14] is selected because it is a method aiming to improve the pose robustness, and it produced a higher ID retrieval score in the experimental results on the FF++ dataset. DiffSwap [10] is one of the representative face swapping methods using a diffusion model. SimSwap [7] is selected because it is one of the most commonly chosen methods for comparing face swap performance.

FaceChanger outperforms all compared methods, achieving the highest CSIM of 0.457, indicating superior identity preservation on the MPIE dataset. Additionally, it records the lowest pose error of 3.39 and the lowest expression error of 3.05, demonstrating its strong ability to maintain target face geometric alignment and semantic expressions. The second-ranked method, FaceDancer, achieves a CSIM of 0.401 with pose and expression errors of 4.72 and 3.31, respectively, indicating good identity preservation but higher errors in pose and expression consistency compared to FaceChanger. The third-ranked method, BlendFace, achieves a CSIM of 0.392 with pose and expression errors of 3.71 and 3.18, respectively, performing slightly better in pose and expression than FaceDancer but with lower identity preservation. These results demonstrate that FaceChanger not only achieves superior identity similarity but also preserves target pose and expression attributes more faithfully than the existing state-of-the-art models, underscoring its effectiveness and generalisability for robust face identity swapping in practical settings.

The qualitative results also suggest that FaceChanger outperforms the recently proposed SOTA methods. Figure 6

TABLE 4. Quantitative results of on the MPIE dataset. † denotes that we ran officially released source codes to obtain the results.

Method	CSIM ↑	pose error ↓	expression error ↓
FSGAN† [17]	0.105	5.31	4.02
SimSwap† [7]	0.180	3.92	3.81
BlendFace† [6]	0.392	3.71	3.18
HifiFace† [13]	0.092	5.01	4.65
FaceDancer† [14]	0.401	4.72	3.31
DiffSwap† [10]	0.278	4.58	4.12
FaceChanger (Ours)	**0.457**	**3.39**	**3.05**

and Figure 7 show qualitative results on the MPIE and LPFF datasets, respectively. In Figure 6, FSGAN and BlendFace (the third and fourth rows of Fig. 6) frequently mis-localise the facial region or fail outright to synthesise a plausible swap; extreme-pose inputs lead to pronounced artefacts and distorted textures. HiFiFace (the second row) fares even worse, as the facial components are misplaced, and boundary transitions collapse altogether. SimSwap (the first row) and FaceDancer (the fifth row) produce noticeably cleaner results and, in many instances, approach the visual quality of FaceChanger. Yet, their silhouettes still blur under severe pose variation. Interestingly, the DiffSwap (the sixth row), which is a diffusion model-based method, results show low identity similarity. Also, at extreme angles, the facial boundaries of the DiffSwap's results are very blurred.

Comparable patterns emerge on the LPFF dataset (see Fig. 7). BlendFace, FSGAN, and HifiFace again struggle. Those exhibit fewer breakdowns than on MPIE but continue to suffer from the representation of facial boundaries. SimSwap and FaceDancer occasionally reproduce fine textures better than FaceChanger; however, under extreme yaw or pitch, the latter yields more natural and coherent faces. Those observations are also found in Figure 8, which shows larger-scale face images of FaceChanger, SimSwap, FaceDancer, and DiffSwap. DiffSwap, which is a diffusion model-based method, produces relatively stable swapped faces compared with other conventional methods, such as FSGAN and Blendface; however, their performance is not as good as that of FaceDancer or Simswap.

Across both benchmarks, FaceChanger consistently outperforms the current state of the art in identity similarity, pose fidelity, and expression accuracy. Also, compared with most methods which constrain their output resolution to 128 × 128, which may not be suitable for media contents having resolutions over full high-definition scale (1280 × 720), FaceChanger provides 256 × 256, which is a doubled output resolution. These results establish FaceChanger as a reliable solution for face-swapping applications that must cope with substantial pose diversity and other adverse conditions.

V. CONCLUSION

We have introduced FaceChanger, a real-time face-identity–swapping framework that combines a lightweight, GAN-based generator with two CLIP-driven contrastive objectives. Ablation studies confirm that the dual CLIP objectives are chiefly responsible for the pose-robustness gains. By aligning pose and expression representations while reinforcing identity similarity using rich semantic supervisory signals extracted from CLIP, our method achieves robust swapping under extreme facial poses and expressions, operating at 40 FPS on a single RTX 4090 for 256 × 256 images.

Extensive experiments across three publicly available benchmarks indicate that FaceChanger delivers consistently robust face-swapping performance, surpassing existing state-of-the-art (SOTA) approaches in stability. On the FF++ dataset, FaceChanger attains an ID retrieval score of 98.63, which is very competitive with the first-ranked method, while achieving SOTA results for both pose and expression error metrics. Notably, in scenarios involving extreme facial poses, FaceChanger exhibits a substantial performance margin over competing methods. These gains are particularly significant given that the method sustains an inference speed comparable to real-time systems, thereby balancing accuracy and efficiency without sacrificing computational practicality.

However, even though the above contribution is significant, some limitations must be mentioned. The current implementation is constrained to a 256 × 256 resolution, which may not be enough for extremely high-resolution content, such as 4K or 8K. Moreover, we have not explored explicit temporal consistency for video sequences. In scenarios where facial regions undergo significant movement or become partially occluded, the visual quality can deteriorate markedly. While face identity swapping methods are typically developed under the assumption that only clean, high-quality images are available, this assumption does not hold in the video domain. In practice, video frames frequently contain motion blur, introducing noise into the facial features and consequently impairing the fidelity of the swapped results. Addressing these adverse conditions is therefore essential to enhance the robustness and overall performance of our approach.

Future work will focus on extending the framework to higher resolutions (512 × 512 and beyond) via progressive training and integrating recurrent or transformer-based modules to ensure temporally coherent video face swaps. We believe these directions will broaden the applicability of FaceChanger and further advance robust, high-fidelity face-swapping technology.

REFERENCES

[1] K. Liu, I. Perov, D. Gao, N. Chervoniy, W. Zhou, and W. Zhang, "Deepfacelab: Integrated, flexible and extensible face-swapping framework," *Pattern Recognit.*, vol. 141, Sep. 2023, Art. no. 109628.

[2] S. Waseem, S. A. R. S. Abu Bakar, B. A. Ahmed, Z. Omar, T. A. E. Eisa, and M. E. E. Dalam, "DeepFake on face and expression swap: A review," *IEEE Access*, vol. 11, pp. 117865–117906, 2023.

[3] *Submission by Echildhood Reviews of the Enhancing Online Safety Act 2015 and the Online Content Scheme*, Dept. Commun. Arts, Tamara Newlands-Executive Director, Canberra, ACT, Australia, 2018.

[4] A. Malik, M. Kuribayashi, S. M. Abdullahi, and A. N. Khan, "DeepFake detection for human face images and videos: A survey," *IEEE Access*, vol. 10, pp. 18757–18775, 2022.

[5] R. Mubarak, T. Alsboui, O. Alshaikh, I. Inuwa-Dutse, S. Khan, and S. Parkinson, "A survey on the detection and impacts of deepfakes in visual, audio, and textual formats," *IEEE Access*, vol. 11, pp. 144497–144529, 2023.

[6] K. Shiohara, X. Yang, and T. Taketomi, "BlendFace: Re-designing identity encoders for face-swapping," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 7600–7610.

[7] R. Chen, X. Chen, B. Ni, and Y. Ge, "SimSwap: An efficient framework for high fidelity face swapping," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 2003–2011.

[8] X. Chen, B. Ni, Y. Liu, N. Liu, Z. Zeng, and H. Wang, "SimSwap++: Towards faster and high-quality identity swapping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 1, pp. 576–592, Jan. 2024.

[9] T. Wang, Z. Li, R. Liu, Y. Wang, and L. Nie, "An efficient attribute-preserving framework for face swapping," *IEEE Trans. Multimedia*, vol. 26, pp. 6554–6565, 2024.

[10] W. Zhao, Y. Rao, W. Shi, Z. Liu, J. Zhou, and J. Lu, "DiffSwap: High-fidelity and controllable face swapping via 3D-aware masked diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8568–8577.

[11] S. Baliah, Q. Lin, S. Liao, X. Liang, and M. H. Khan, "Realistic and efficient face swapping: A unified approach with diffusion models," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Feb. 2025, pp. 1062–1071.

[12] Y. Li, C. Ma, Y. Yan, W. Zhu, and X. Yang, "3D-aware face swapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 12705–12714.

[13] Y. Wang, X. Chen, J. Zhu, W. Chu, Y. Tai, C. Wang, J. Li, Y. Wu, F. Huang, and R. Ji, "HifiFace: 3D shape and semantic prior guided high fidelity face swapping," 2021, *arXiv:2106.09965*.

[14] F. Rosberg, E. E. Aksoy, F. Alonso-Fernandez, and C. Englund, "FaceDancer: Pose- and occlusion-aware high fidelity face swapping," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 3443–3452.

[15] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8748–8763.

[16] M. Caron, H. Touvron, I. Misra, H. Jegou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9650–9660.

[17] Y. Nirkin, Y. Keller, and T. Hassner, "FSGAN: Subject agnostic face swapping and reenactment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7183–7192.

[18] A. Rössler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Niessner, "FaceForensics++: Learning to detect manipulated facial images," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1–11.

[19] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2009.

[20] Y. Wu, J. Zhang, H. Fu, and X. Jin, "LPFF: A portrait dataset for face generators across large poses," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 20270–20280.

[21] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "FaceShifter: Towards high fidelity and occlusion aware face swapping," 2019, *arXiv:1912.13457*.

[22] Q. Li, W. Wang, C. Xu, Z. Sun, and M.-H. Yang, "Learning disentangled representation for one-shot progressive face swapping," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 8348–8364, Dec. 2024.

[23] K. Kim, Y. Kim, S. Cho, J. Seo, J. Nam, K. Lee, S. Kim, and K. Lee, "DiffFace: Diffusion-based face swapping with facial guidance," *Pattern Recognit.*, vol. 163, Jul. 2025, Art. no. 111451.

[24] P. T. Yee, S. Mishra, and A. Dhall, "ClipSwap: Towards high fidelity face swapping via attributes and CLIP-informed loss," in *Proc. IEEE 18th Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2024, pp. 1–10.

[25] DeepFakes. (2020). *Faceswap*. [Online]. Available: https://github.com/deepfakes/faceswap

[26] N. T. Luu, "CLIP unreasonable potential in single-shot face recognition," 2024, *arXiv:2411.12319*.

[27] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4685–4694.

[28] Y. Lin, S. Wang, Q. Lin, and F. Tang, "Face swapping under large pose variations: A 3D model based approach," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jul. 2012, pp. 333–338.

[29] F. P. Papantoniou, A. Lattas, S. Moschoglou, J. Deng, B. Kainz, and S. Zafeiriou, "Arc2Face: A foundation model for ID-consistent human faces," in *Proc. Eur. Conf. Comput. Vis.*, 2024, pp. 241–261.

[30] K. Cui, R. Wu, F. Zhan, and S. Lu, "Face transformer: Towards high fidelity and accurate face swapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 668–677.

[31] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, 2017, pp. 1125–1134.

[32] A. Hurst et al., "GPT-4o system card," 2024, *arXiv:2410.21276*.

[33] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 67–74.

[34] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent.*, 2017.

[35] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.

[36] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1867–1874.

[37] D. Qi, W. Tan, Q. Yao, and J. Liu, "YOLO5Face: Why reinventing a face detector," 2021, *arXiv:2105.12931*.

[38] A. Bulat and G. Tzimiropoulos, "How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks)," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1021–1030.

[39] Y. Feng, H. Feng, M. J. Black, and T. Bolkart, "Learning an animatable detailed 3D face model from in-the-wild images," in *Proc. ACM Trans. Graph.*, vol. 40, Aug. 2021, pp. 1–13.

[40] C. Yu, J. Wang, C. Peng, C. Gao, and G. Yu, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 325–341.

[41] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen, "Advancing high fidelity identity swapping for forgery detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5073–5082.

[42] Y. Zhu, Q. Li, J. Wang, C. Xu, and Z. Sun, "One shot face swapping on megapixels," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 4832–4842.

[43] Y. Xu, B. Deng, J. Wang, Y. Jing, J. Pan, and S. He, "High-resolution face swapping via latent semantics disentanglement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7632–7641.

[44] C. Xu, J. Zhang, M. Hua, Q. He, Z. Yi, and Y. Liu, "Region-aware face swapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 7622–7631.

[45] *FaceSwap*. Accessed: Feb. 14, 2022. [Online]. Available: https://github.com/ondyari/FaceForensics/tree/master/dataset/FaceSwapKowalski

JONGMIN YU (Member, IEEE) received the joint Ph.D. degree from the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju, Republic of Korea, and the School of Electrical Engineering, Computing and Mathematical Sciences, Curtin University, Perth, Australia. He was a Research Associate at the Department of Applied Mathematics and Theoretical Physics, University of Cambridge. He worked at King's College London and the Institute of IT Convergence, Korea Advanced Institute of Science and Technology. He is currently a Principal Scientist at ProjectG.AI. His research interests include artificial intelligence, machine learning, pattern recognition, and the mathematical understanding of these.

HYEONTAEK OH (Member, IEEE) received the B.S. degree (summa cum laude) in computer science and the Ph.D. degree in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), in 2012 and 2020, respectively. From 2020 to 2023, he was a Team Leader at the KAIST Institute for Information Technology Convergence. He is currently a Research and Development Team Leader at ProjectG.AI. His research interests include the applications of artificial intelligence and agent systems.

ZHONGTIAN SUN is currently a Lecturer (an Assistant Professor) with the School of Computing, University of Kent, and a Visiting Fellow with the Department of Computer Science and Technology, University of Cambridge. Previously, he was a Research Associate at the University of Cambridge and the University of Oxford. He also has industry experience in asset management. He is the Co-Founder and the CTO of the AI4finance startup. His research interests include enhancing the expressiveness and interpretability of deep learning models with applications, including healthcare, finance, education, and recommender systems. His current research interests include representation learning, large language models, multi-agent systems, causal inference, and symbolic reasoning.

YOUNKWAN LEE received the B.S. degree in computer science from Korea Aerospace University, Gyeonggi, South Korea, in 2016, and the Ph.D. degree from the School of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2022. He is currently a Staff Researcher with Global Technology Research, Samsung Electronics, South Korea. His current research interests include vision foundation models, multi-modal AI, and computer vision.

JINHONG YANG (Member, IEEE) received the Ph.D. degree in information and communication engineering from KAIST, Daejeon, in 2017, where he is currently pursuing the Ph.D. degree with the Information and Communication Engineering Department. From 2005 to 2008, he was a Research Assistant with HERIT Inc. He is an Associate Professor with the Department of Medical IT, Inje University. He holds more than 300 domestic registered patents and more than 30 international registered patents in the field of information and communications technology. His research interests include reinforcement learning, physical AI, and embodied AI.