MagicMask: A Fast and High-fidelity Face Swapping Method Robust to Face Pose

Jongmin Yu[1,2] JY522@PROJECTG.AI
Anoushka Harit[3] ANOUSHKA.HARIT@CRUK.CAM.AC.UK
JianKang Deng[4] J.DENG16@IMPERIAL.AC.UK
Shan Luo[5] SHAN.LUO@KCL.AC.UK
Jinghang Yang[1,6] JINHONG@INJE.AC.KR
Zhongtian Sun[2,7,] Z.SUN-256@KENT.AC.UK
[1] *PGAI Research, ProjectG.AI, Daejeon, 34141, Republic of Korea*
[2] *Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, CB3 0WA, United Kingdom*
[3] *Cancer Research UK , University of Cambridge, Cambridge CB2 0RE, United Kingdom*
[4] *Department of Computing, Imperial College of London, London, SW7 2AZ, United Kingdom*
[5] *Department of Engineering, King's College London, London, WC2R 2LS, United Kingdom*
[6] *Department of Medical Information Technology, Inje University, Kimhae, 50834, Republic of Korea*
[7] *School of Computing, University of Kent, Kent, CT2 7NZ, United Kingdom*

Editors: Hung-yi Lee and Tongliang Liu

Abstract

Recent face-swapping methods excel under controlled conditions but often fail when presented with extreme facial poses. Diffusion-based approaches may be able to overcome these issues, but they still face significant computational costs. This paper introduces MagicMask, a novel face-swapping framework that robustly handles various poses in real time by fusing visual and geometric information. Our method incorporates explicit, identity-adapted geometric cues into the latent feature space via a multi-head attention mechanism. It employs an Adversarial Facial Silhouette Alignment (AFSA) loss to preserve detailed facial boundaries that are adapted to the source identity. Comprehensive experiments on multiple benchmarks demonstrate that MagicMask competes with state-of-the-art methods under standard conditions and significantly outperforms them in extreme pose scenarios.

Keywords: Face identity swap, face swap, pose robustness, generative adversarial network, transformer

1. Introduction

Face swapping is a technique that replaces the identity of one individual in an image or video with that of another. It offers promising applications in computer vision for the visual arts and entertainment industries (Perov et al., 2020); on the other hand, it poses risks of misuse in activities such as scams, abuse, and the creation of non-consensual pornography (Director, 2018). Despite these negative implications, research into face-swapping technology remains crucial due to its technological potential and significant societal impact.

© 2025 J. Yu, A. Harit, J. Deng, S. Luo, J. Yang & Z. Sun.

The primary objective in the face identity swapping literature is to make face identity more similar to a given source face while preserving non-identity-related attributes of the target images, such as skin texture, illumination, hairstyles, and facial accessories (Shiohara et al., 2023). Face identity-swapping methods have undergone considerable evolution with the rapid advancement of deep learning-driven computer vision. The quality of the results has improved to the extent that some methods (Chen et al., 2020, 2023; Wang et al., 2024) now yield seamless and photorealistic outcomes in controlled environments, making it increasingly challenging to distinguish swapped faces from the originals.

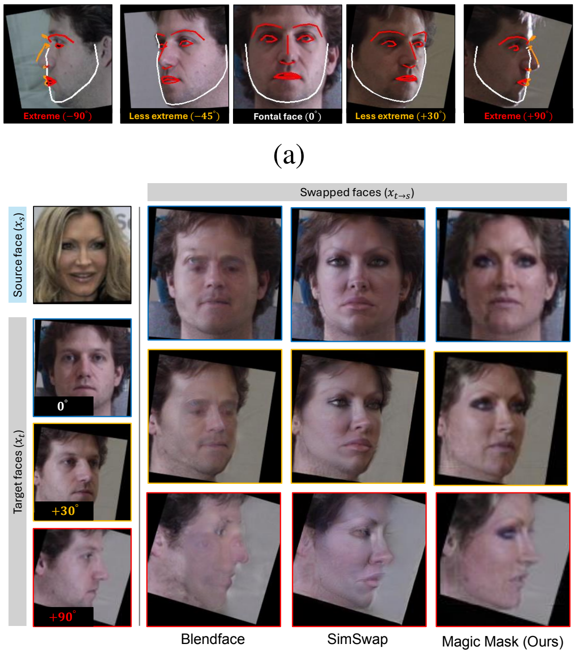
However, face identity swapping under extreme face poses remains a significant challenge. Extreme poses can be thought of as some angular ranges in which the shape of facial components changes the boundaries of the face, because the actual boundaries can not cover the components due to changes in the angle of the face. Figure 1 illustrates the morphological differences between face boundaries and face landmarks, depending on the poses. As shown in Figure 1(a), in frontal face (face angle = $0°$), where all key facial landmarks (e.g., both eyes, nose, and mouth) are positioned within the facial boundaries, the process can be conceptualised as reshaping and repositioning these components to reflect the source identity, with well-defined regions requiring modification.

In contrast, extreme face poses present challenges (See the extreme (face angle = $\pm 90°$) and less extreme (face angle = $-45°$ or $+30°$)cases in Figure 1(a)), such as occlusions and irregular facial boundaries caused by facial components, which collectively complicate the swapping process. As depicted in Figure 1(b), recent methods (Shiohara et al., 2023; Chen et al., 2020) are still intractable in generating seamless face-swapping results in certain extreme facial poses. Several studies have addressed face swapping under challenging face poses (Rosberg et al., 2023; Li et al., 2023; Wang et al., 2021b).

The predominant way is to leverage explicit geometric information. Li *et al.,*(Li et al., 2023) employs landmark features to encode the positional information of facial components. Wang *et al.,*(Wang et al., 2021b) use 3D face images as geometric supervision to understand the shape and pose of the face. Rosberg *et al.,*(Rosberg et al., 2023) presents an interpretive feature similarity regulisataion for preserving the pose of a target face. Nonetheless, as shown in Figure 1(b), they still produce unrealistic results if face poses cause significant modifications to the facial silhouette. Recently, several diffusion-based face-swapping methods (Zhao et al., 2023; Baliah et al., 2025) have shown remarkable achievements. Those methods produce exact target attribute reconstruction results while preserving the identity of the source image. However, due to the iterative nature of their sampling process, they incur substantial computational costs, which preclude real-time deployment.

In this work, we propose a novel face-swapping method called *MagicMask*, a real-time and pose-robust method for face identity swapping. The MagicMask comprises three main components: a visual representation module, a geometric representation module, and a decoder. The visual representation module extracts latent features from images. Coupled with a face identity encoder that derives the identity code from the source image, it produces identity-specific latent representations. To explicitly preserve and inject identity-specific characteristics into the latent space of the target face, we introduce a novel Attention-Residual Identity Embedding (ARIE) module. The ARIE adaptively enhances identity preservation through a synergistic attention-residual mechanism, ensuring robust feature refinement. The latent features obtained from both modules are then combined and fed into the decoder to generate the swapped face. Additionally, we introduce a novel complementary loss called *'Adversarial Facial Silhouette Alignment'* to ensure that the facial silhouette in extreme face poses is naturally aligned while preserving the source identity.

Experimental results demonstrate the effectiveness of the MagicMask in swapping face identities under extreme poses. MagicMask achieves scores of 98.41, 1.47, and 2.04 for identity (ID) retrieval, pose error, and expression error metrics, respectively, which surpass those of recent state-of-the-art (SOTA) methods. Also, the MagicMask achieves 27.9ms of execution speed, which is almost 36 frame-per-second (FPS) in our experiments. For face-swapping experiments for extreme poses using MPIE (Gross et al., 2010) and LPFF (Wu et al., 2023) datasets, the MagicMask produces a mean score of 0.372 cosine similarity with 2.94 pose error and 2.93 expression error. These results suggest that the MagicMask provides superior and more stable face-swapping performance compared to the existing SOTA approaches. Consequently, this paper opens a new avenue for identity-preserving face-swapping in extreme pose scenarios, laying a solid foundation for future advancements in pose-invariant facial synthesis, deepfake detection, and identity-aware generative models.



(a)

(b)

Figure 1: Illustration for the changes of shapes of facial boundaries and components and the identity swapping results of recently proposed methods and MagicMask depending on face poses, respectively. The images are obtained from MPIE dataset (Gross et al., 2010). (a) describes the change of facial component silhouette depending on face poses. The white lines define the face boundaries. The red lines show the silhouette of facial components such as the nose and eye, but if some parts of the facial components are outside the facial boundaries (white lines), the parts are drawn by the orange lines. (b) shows identity swapping results of Blendface (Shiohara et al., 2023), SimSwap (Chen et al., 2020), and the proposed MagicMask. MagicMask produces more natural face-swapping results across frontal to extreme face poses compared with others.

2. Related Works

The primary objective of face identity swapping research is not only just to swap the target face's identity to the source face but to preserve the target face's visual attributes that are irrelevant to identity, such as illumination conditions, hairstyles, makeup, skin texture, and facial accessories. Recently, proposed methods have addressed identity swapping by formulating the task to minimize the latent feature distances between the swapped face and the source face. Once the identity of the target image is swapped, methods such as Faceshifter (Li et al., 2019) and Faceswapper (Li et al., 2024) extract latent features using a face identity encoder and minimize the cosine angular distance between the latent features of the swapped face and those of the source face. In addition to cosine angular distance, Euclidean metrics, such as the l_1-distance and l_2-distance, are commonly used to assess the similarity between latent features (Nirkin et al., 2019).

Regarding the preservation of visual properties of target faces, Zhang *et al.,*(Zhang et al., 2021) employed a simple classifier with an associated classification loss to assess whether these properties are retained in the swapped face. However, this approach necessitates additional labelling to specify

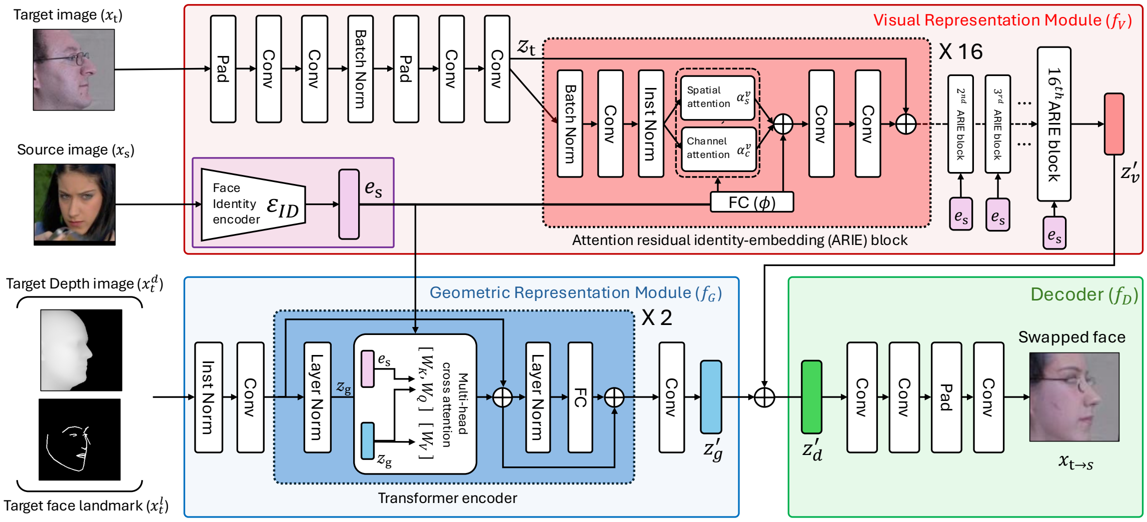


Figure 2: Architectural details of the proposed MagicMask. The visual representation module f_V extracts a visual latent feature $z_v^{'}$ using the target image x_t and the source identity code e_s extracted from source image x_s by using the identity encoder ε_{ID}. The e_s is continuously ensambled while extracting low-level to high-level features using the attention residual identity-embedding (ARIE) blocks to enhance the identity of the source image. The geometric representation module f_G draws out the geometric latent feature $z_g^{'}$ from a depth x_t and a facial landmark paired with x_t. $z_v^{'}$ and $z_g^{'}$ are combined and applied to the decoder to generate swapped face $x_{t \to s}$.

the types of properties present in the target face image, thus increasing the labour intensity. A more common strategy to preserve the visual properties of target images involves the use of loss functions inspired by perceptual loss (Gatys, 2015). In addition, reconstruction loss, which is applied between the swapped face and the target face image, is another widely used method for this purpose (Chen et al., 2020, 2023).

Although the methods described above demonstrate remarkable performance under controlled experimental conditions, it remains uncertain whether similar performance can be achieved in cases involving extreme poses. As shown in Figure 1(a), when face poses change drastically, the facial elements that are deformed to fit the source identity will affect the entire silhouette of the face, and if this is not handled properly, a deterioration in the quality of the distorted face will occur. Using explicit geometric features such as facial landmarks (Li et al., 2024) and face 3D maps (Wang et al., 2021b) can be thought of as a solution to this issue. However, as shown in Figure 1, recent SOTA methods still suffer from poor swap results on extreme poses. Consequently, it is essential to develop a robust method for a wide range of face pose variations.

3. The MagicMask

3.1. Methodology overview

We build a novel architecture called MagicMask, which can capture not only visual representation but also explicit geometric information. Figure 2 shows the architectural details of the MagicMask. The visual representation module f_V abstracts latent features $z_v^{'}$ extracted from the target image and integrates source identity code e_s extracted from the identity encoder ε_{ID} to grant the source identity information to latent features extracted from the target face. The geometric representation module f_G extracts source identity-adapted geometric latent features $z_g^{'}$ using e_s and explicit geometric information defined by depth maps x_t^d and landmark features x_t^l. Based on the multi-head cross-attention mechanism with e_s, $z_g^{'}$ is adapted to the source identity. $z_v^{'}$ and $z_g^{'}$ are combined and applied to the decoder f_D for generating the swapped face.

In addition, we present a complimentary loss term for improving the quality of facial boundaries and components, which are the regions where the most severe quality degradation occurs in extreme poses. As shown in Figure 1, depending on the face pose and facial component shape, the shapes of facial components and boundaries sometimes should be modified and also need some in-painting for the modified regions. We formulate '*Adversarial facial silhouette alignment*', which can reduce the image distortion and quality degradation according to the distortion caused by a changed shape of facial components and boundaries under some extreme poses. Detailed information for the architectural characteristics and loss terms are described in the further subsections.

3.2. Architectural details

Visual representation module: This module is established to obtain visual latent features which contain source identity information and target image attributes. It is best if we can only transfer the source identity information to some area that represents the target face identity in a latent feature space. However, it is almost intractable because the identity information is highly coupled with other information about identity-nonrelated attributes. Li *et al.,*(Li et al., 2024) use face segmentation masks of target face images to explicitly constrain spatial regions of latent features in combining identity codes and latent features (Li et al., 2024). However, the masks can contain a lot of noisy information when extreme face poses significantly vary face shapes, so strong constraints using a binary mask may degrade robustness on face pose, which may degrade the performance of swapping methods.

Instead of leveraging binary masks to constrain feature-combining areas for the identity code injection, we use attention mechanisms, a more natural way to combine identity information with target latent features. We use an objective function to constrain the region to show the identity and to preserve the attributes of target images. We present an attention residual identity-embedding (ARIE) block to integrate identity information into the latent features. Figure 2 shows the architectural details of the module. Initially, the ARIE block conducts batch normalisation to improve the generalisation performance and stabilise the training. After that, we apply a convolutional layer and Instance Normalisation (IN) to enhance the invariance about image textures, which are irrelevant in recognising face identity, as follows:

$$IN(z_{\mathrm{v}}) = \frac{z_{\mathrm{v}} - \mu_c(z_{\mathrm{v}})}{\sqrt{\sigma_c(z_{\mathrm{v}})^2 + \epsilon}}, \tag{1}$$

where μ_c and σ_c indicate the mean and standard deviation computed separately for each instance and each channel of a given feature computed with respect to the feature's height and width axes. ϵ defines a small constant added for numerical stability.

Then, we compute spatial α_s^v and channel α_c^v attention matrixes using the output of the instant normalisation and e_{s} as follows:

$$\alpha_s^v = \text{softmax}\left(\left(\frac{1}{C}\sum_{c=0}^{C} IN(z_t)_c\right)\phi(e_s)\right), \tag{2}$$

$$\alpha_c^v = \text{softmax}\left(\left(\frac{1}{WH}\sum_{w=0}^{W}\sum_{h=0}^{H} IN(z_{\text{t}})_{w,h}\right)\otimes\phi(e_s)\right), \tag{3}$$

where W, H, and C denote the width, height, and channels of the output of the instance normalisation. w, h, and c define the indices for the width, height, and channels, respectively. $\phi(*)$ defines linear kernels for scaling e_s and projecting into the latent feature space of each ARIE block. $\otimes$ indicates element-wise multiplication. z_t and $\phi(e_s)$ take 1024 dimensionalities.

At last, we integrate source identity information using the above attention tensors and the identity code into latent features as follows:

$$z_{\text{v}} = [\alpha_s^v \otimes IN(z_{\text{t}}) + IN(z_{\text{t}})\alpha_c^v] \oplus e_{\text{s}}, \tag{4}$$

where $\oplus$ denotes the channel-wise summation.

At last, we pass the modified features through additional convolutional layers and then combine them with the original z_t. As shown in Figure 2, by sequentially placing the ARIE blocks, we aim to strengthen the identity information of the source face in all layers of latent features.

Geometric representation module: The geometric representation module f_G learns explicit geometric information such as locations and depths of facial components and their positioning. Li *et al.,*(Li et al., 2024) and Wang *et al.,*(Wang et al., 2021b) suppose that the usage of explicit geometric information improves the robustness of face identity swaps when face poses are dynamically changed. The architectural details of f_G are shown in Figure 2. We extract latent features from depth map and landmark features using convolutional layers and apply it into the multiple transformer encoders with e_{s}. The transformer encoders are used to capture more global information, such as relative positional information of facial components, which the convolutional layer may miss.

Geometric latent feature z_g using multi-head attention is produced as follows. At the beginning, we obtain the key K_g, query Q_g, and values V_g using the three independent linearly kernels W_K, W_Q, and W_V as follows:

$$K_g = z_g W_K, \quad Q_g = z_g W_Q, \quad V_g = z_g W_V. \tag{5}$$

Usually, K_g and Q_g are used to compute the attention matrix normalised by the Softmax function, and it is combined with V_g to derive a self-attention output. In addition to the self-attention on the geometric features, we compute an additional sharing key-query attention matrix using W_Q, W_K, and e_{s}: $K_e = e_s W_K$ and $Q_e = e_s W_Q$ and combined with the self-attention. We derive the final output of the transformer as follows:

$$z_g = \text{softmax}\left(\frac{Q_g K_g^{\text{T}}}{\sqrt{d_g}}\right)V_g + \text{softmax}\left(\frac{Q_e K_e^{\text{T}}}{\sqrt{d_g}}\right)V_g, \tag{6}$$

where d_g denotes the dimensionality of the transformer encoder. z_g abstracts information from a geometric feature adapted with the source identity code using the attention mechanism so that it automatically learns the correspondence between geometric information and the source identity.

As a result, z_g explicitly models the alignment between the source identity and the target's geometric information, enabling better disentanglement and more robust synthesis under pose variation. z_g is applied to the decoder to generate an identity-swapped face.

Consequently, the input to the visual representation module is RGB images containing rich visual information. The visual representation module encodes texture-rich RGB features and injects identity via channel & spatial attention. The geometric representation module encodes topology-rich depth and landmark tensors, aligning them with the identity code through cross-attention, thereby learning the correspondence between 3-D structure and identity tokens.

Decoder: After obtaining the two latent features z_v and z_g, we pass a combined feature defined as the summation of the two features $z_v + z_g$ through the Decoder to generate the swapped face. The decoder will focus only on restoring the image from the features and leave the identity modification mission to the identity code embedding in the visual and geometric representation modules. In this work, we just use the padding operation and convolutional layers to upscale the latent feature and generate the swapped face.

3.3. Objective function

Identity swapping loss: This encourages the swapped image $x_{t\to s}$ to have the same identity as x_s. In this paper, the loss term for identity swap is employed to minimise the cosine angular similarity between $x_{t\to s}$ and x_s in the latent feature space, as follows:

$$\mathcal{L}_{\text{IS}}^{\text{t}\to\text{s}} = 1 - \frac{\varepsilon_{ID}(x_s)^T \varepsilon_{ID}\left(x_{t\to s}\right)}{\|\varepsilon_{ID}\left(x_s\right)\|_2 \|\varepsilon_{ID}\left(x_{t\to s}\right)\|_2}, \tag{7}$$

where ε_{ID} indicates the face identity encoder. Our work uses the pre-trained face recognition model (Deng et al., 2019) as the identity encoder, which is the most frequently used network for various face swapping methods (Shiohara et al., 2023; Chen et al., 2020, 2023; Wang et al., 2024; Li et al., 2019). The parameters on the identity encoder are frozen during the optimisation.

Attribute preserving and visual quality loss: The visual quality of the swapped face images is determined not only by their identity similarity to the source face but also by how well they preserve non-identity-related attributes such as illumination, skin texture, and background details. To encourage attribute preservation, we formulate reconstruction loss using image and latent features as follows:

$$\mathcal{L}_{\text{Recon}}^{\text{t}\to\text{s}} = \|(x_{t\to s} - x_t) \otimes (1 - m_t)\|_1 + \sum_{i=1}^{\mathcal{N}_D} \left\|\left(z_{t\to s}^i - z_t^i\right) \otimes \left(1 - \bar{m}_t^i\right)\right\|_2^2, \tag{8}$$

where $\mathcal{N}_D$ defines the number of convolutional layers in the decoder. z_*^i denotes the latent feature taken from i^{th} convolutional layers in the decoder. m_t indicates a binary-valued face mask, and $\bar{m}_t^i$ denotes the mask that linearly rescaled from m_t according to the width and height of i^{th} convolutional layer's latent feature.

The reconstruction loss is motivated by the perceptual loss of neural style transfer (Gatys et al., 2016). The gradients of this loss are only computed from a non-facial area of the target face image marked by the mask. Because we don't explicitly limit how the target latent feature and source identity code overlap, integrating the source identity code influences the entire target feature space; Thus, we apply a loss function that explicitly enhances the visual attribute of the target face.

In addition to $\mathcal{L}_{\text{Recon}}^{\text{t}\to\text{s}}$, we employ the cyclic reconstruction loss function $\mathcal{L}_{\text{Cycle}}^{\text{t}\to\text{t}}$ to improve the image generation performance regardless of identity swapping, represented by

$$\mathcal{L}_{\text{Cycle}}^{\text{t}\to\text{t}} = \|x_{t\to t} - x_t\|_2^2 \,, \tag{9}$$

where $x_{t\to t}$ defines the swapped image using x_t only *i.e.,*we assign x_t not only the target face but a source face also. We employ this loss because m and $\bar{m}$ in $\mathcal{L}_{\text{Recon}}^{\text{t}\to\text{s}}$ consists of binary values (*i.e.,*1 is face area and 0 is non-face area), it may output unnatural results of facial boundaries which are potentially harmful in training the MagicMask. Cyclic reconstruction loss is employed to decay the disadvantage of strong constraints by using a binary value facial mask and improve general performance in generating face images.

Adversarial facial silhouette alignment: Adversarial loss is a commonly used loss function to improve the quality of swapped face images by helping to increase performance in recovering high-frequency features such as image sharpness (Li et al., 2024, 2019; Chen et al., 2020). The commonly used adversarial loss for the face identity swap is formulated as follows:

$$\mathcal{L}_{Adv}^{t\to s} = \mathbb{E}\left[\log\left(1 - D\left(x_{t\to s}\right)\right)\right] + \mathbb{E}\left[\log D\left(x_t\right)\right], \tag{10}$$

where D indicates the discriminator for the adversarial loss.

Existing methods have demonstrated that the above loss function works well when face poses are normal. However, as shown in Figure 1(b), extreme face poses can lead to artefacts, such as synthesising a frontal face feature onto a side face, because the discriminator D only evaluates whether an entire image is real or generated, without enforcing local consistency of facial components.

To address this issue and explicitly improve the visual quality of facial components and face boundaries, we propose an adversarial loss called *Adversarial Facial Silhouette Alignment* (AFSA). The idea is to perform adversarial learning on local image patches corresponding to facial components. Wang *et al.,*(Wang et al., 2021a) applied similar ideas for face restoration and it's demonstrate that adversrial learning using facial components will improve the visual quality the outputs.

Given the target image x_t, the swapped image $x_{t\to s}$, and the face landmarks x_t^l (which provide coordinates for key facial components such as the eyes, nose, and mouth), we extract local patches by defining a bounding box for each component using its extreme coordinates (top, bottom, left, and right) with a small margin. In this work, the set of facial component patches is defined as:

$$X^l = \{x^{\text{eye}}, x^{\text{nose}}, x^{\text{mouth}}, x^{\text{jaw}}\}.$$

After we obtain the two facial component image sets X_t^l and $X_{t\to s}^l$ from x_t and $x_{t\to s}$, we conducted an adversarial learning as follows:

$$\mathcal{L}_{AFSA}^{t\to s} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathbb{E}_{x_{t\to s,i}^l \in X_{t\to s}^l}\left[\log\left(1 - D\left(x_{t\to s,i}^l\right)\right)\right] + \mathbb{E}_{x_{t,i}^l \in X_t^l}\left[\log D\left(x_{t,i}^l\right)\right]\right) \tag{11}$$

where i and N are the index and the number of facial component patch (e.g., $N = 4$ for eye, nose, mouth, jaw and the index of the eye is 1), and $x_{t,i}^l$ and $x_{t\to s,i}^l$ denote the i^{th} patch extracted from x_t and $x_{t\to s}$, respectively.

The adversarial facial silhouette alignment aims to compare facial components and boundaries of target and swapped faces using an adversarial manner and reduce some distortion or blurred parts, which are poorly generated areas.

Total objective: The total objective function is defined by the summation of the above loss functions with balancing weights for each term and additional regularisation term defined using $l1$-norm on trainable parameters, and it is represented as follows:

$$\mathcal{L}_{\text{Total}} = \lambda_{\text{IS}}\mathcal{L}_{\text{IS}} + \lambda_{\text{Recon}}(\mathcal{L}_{\text{Recon}} + \mathcal{L}_{\text{Cycle}}) + \lambda_{\text{AFSA}}\mathcal{L}_{\text{AFSA}}, \tag{12}$$

where λ_{IS}, λ_{Recon}, and λ_{AFSA} indicate the balancing weights for $\mathcal{L}_{\text{IS}}$, $\mathcal{L}_{\text{Recon}}$ and $\mathcal{L}_{\text{Cycle}}$, and $\mathcal{L}_{\text{AFSA}}$, respectively. Section 4 describes the setting of those balancing weights for training.

Each training sample for optimizing MagicMask consists of a pair of source and target face images, along with their corresponding depth maps and landmark features. The loss function is also computed for the reverse identity swap $x_{s\to t}$. This not only enhances computational efficiency but also increases the diversity of face swap scenarios, improving the model's adaptability.

4. Experiments

4.1. Dataset and Experiment protocol

Datasets: In our experiments, we employ VGGFace2 (Cao et al., 2018) dataset and CelebA-HQ (Karras et al., 2018) dataset for model training. In performance estimation and comparison with existing SOTA methods, we use the FaceForensics++ (FF++) (Rossler et al., 2019) dataset, 4) Multi Pose, Illumination, Expressions (MPIE) (Gross et al., 2010) dataset and 5) Large-pose Flickr Face (LPFF) (Wu et al., 2023) datasets.

VGGFace2, CelebA-HQ, and FF++ datasets are well-known and frequently used datasets in the face identity swap literature. In particular, the FF++ dataset is one of the most popular benchmarks. However, the FF++ dataset does not focus on the extreme face poses. It is essential to use datasets that encompass a wide variety of face orientations to demonstrate the effectiveness of MagicMask in handling extreme face angle cases. Thus, we employ the LPFF and MPIE datasets, which provide numerous variations for facial poses. We provide detailed information and the accessible URLs of those datasets in Appendix A.

Evaluation metrics and protocol: We employ four evaluation criteria which are commonly used for face identity swap literature: 1) Cosine Similarity Metrics (CSIM), 2) ID retrieval, 3) pose error, and 4) expression error. We follow the same setting as in Li *et al.,*(Li et al., 2019) and Chen *et al.,*(Chen et al., 2020) for the performance comparison using the FF++ data sets.

About the MPIE and LPFF datasets, due to the lack of established quantitative benchmarks for face identity swapping using these two datasets, our analysis in these cases focuses primarily on qualitative results. However, since the MPIE dataset provides multiple images for each identity, we revised the evaluation process for the FF++ dataset and applied it to the MPIE dataset. However, there is no prepared source and target image pair. Therefore, we randomly selected 1000 images from the CeleA-HQ dataset as source faces and used all subjects in the MPIE dataset as target faces. We repeated those processes 10 times to reduce bias for some particular identities in our experiments.

4.2. Implementation details

We conduct data preprocessing as follows. First, we detect face regions using a face detector (Qi et al., 2021). For the detected results, face alignment is performed using the face landmark detection and alignment algorithm proposed by Bulat *et al.,*(Bulat and Tzimiropoulos, 2017). Depth maps and facial masks for the pre-processed images are obtained by the facial depth extraction (Feng et al.,

Architectural setting	CSIM ↑	pose error ↓	expression error ↓	Execution speed (ms) ↓
Normal adversarial loss (eq. 10)				
f_V	0.291	4.21	4.13	27.9
$f_V + f_G$	0.435	3.97	3.51	31.6
Adversarial facial silhouette alignment (eq. 11)				
f_V	0.427	3.82	3.22	27.93
$f_V + f_G$	**0.463**	**3.35**	**2.91**	31.6

Table 1: The quantitative results on the MPIE dataset (Gross et al., 2010) regarding on the architecture setting and loss functions of the MagicMask.

2021) and the face segmentation (Yu et al., 2018). Initially, all face images are resized to 256 × 256 to enhance the dataset's quality. Subsequently, the images are aligned and cropped to a uniform resolution of 224 × 224 for use in the visual representation module. To generate the source identity code, the source face images are resized to 112 × 112. λ_{IS}, λ_{Recon}, and λ_{AFSA} are set by 5.0, 2.0, and 1.0. We employ a pre-trained ArcFace model (Deng et al., 2019) as the face identity encoder. Training and testing are conducted on two RTX A6000 GPUs over 500 epochs. The source code of MagicMask is publicly available[1].

4.3. Ablation studies

Loss function settings: Performance metrics indicate a significant impact when transitioning from a normal adversarial loss to AFSA loss. Specifically, for the architecture using only f_V, the CSIM metric increased from 0.291 to 0.427, while both pose and expression errors saw notable reductions. This suggests that the AFSA loss more effectively preserves structural and identity details in the generated outputs, likely due to its focus on the facial outline and alignment. Overall, this loss function provides a more robust supervision signal that leads to improved face identity swap results.
Geometric representation module: The incorporation of the geometric representation module f_G consistently boosts performance across both loss function settings. When collaborated with f_V, f_G contributes to higher CSIM values and lower pose and expression errors. Under normal adversarial loss, the CSIM improved from 0.291 to 0.435 , and both pose and expression errors decreased. Combined with the AFSA loss, the CSIM improved from 0.427 to 0.463. These improvements indicate that the geometric representation module effectively complements the visual features, refining the output quality and ensuring a more accurate face swap by capturing and preserving important geometric details in diverse pose variations.

4.4. Performance Comparison

General face pose cases: It is essential to confirm that MagicMask provides competitive performance compared to existing SOTA methods based on commonly used benchmarks to demonstrate that the structural and objective functional contribution does not bias the performance of the face identity swap for the extreme face angle cases only. Figure 3 and Table 2 represent qualitative and quantitative results on the FF++ dataset (Rossler et al., 2019), respectively. MagicMask achieves the best pose and expression errors and a competitive ID retrieval score compared with existing SOTA methods: the 98.41 ID retrieval score, 1.47 pose error, and 2.04 expression error. FaceDancer (Rosberg et al., 2023) produces the best IR retrieval, which is 98.84. This figure is higher than the DiffSwap Zhao

[1] https://github.com/andreYoo/magicmask_official

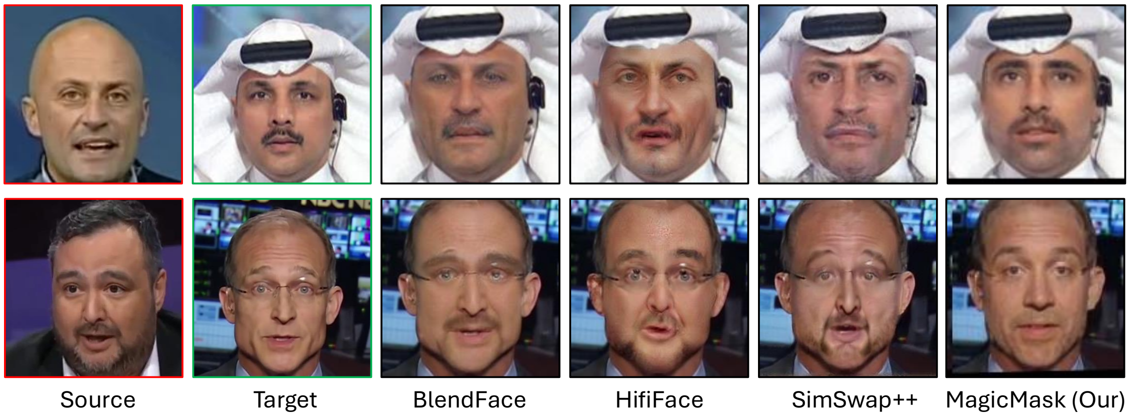


Figure 3: Qualitative results on the FF++ dataset (Rossler et al., 2019).

Method	ID retrieval ↑	pose error ↓	expression error ↓	FID ↓	Execution speed (ms) ↓
FaceSwap (Rossler et al., 2019)	72.69	2.58	2.89	-	-
DeepFakes (DeepFakes, 2020)	88.39	4.64	3.33	-	-
FaceShifter (Li et al., 2020)	90.68	2.55	2.82	-	-
MegaFS (Zhu et al., 2021)	90.83	2.64	2.96	-	-
FSLSD (Xu et al., 2022b)	90.05	2.46	2.79	-	-
RAFSwap (Xu et al., 2022a)	92.54	3.21	3.60	-	-
FaceSwapper (Li et al., 2024)	94.48	2.10	2.69	-	-
DiffSwap (Zhao et al., 2023)	98.54	2.45	5.35	-	-
FSGAN† (Nirkin et al., 2019)	61.07	3.31	3.02	-	**21.5**
SimSwap† (Chen et al., 2020)	93.01	1.53	2.84	13.8	27.1
BlendFace† (Shiohara et al., 2023)	97.02	3.07	2.14	-	24.7
HifiFace† (Wang et al., 2021b)	98.01	2.84	2.51	11.58	22.3
FaceDancer† (Rosberg et al., 2023)	**98.84**	2.04	7.97	18.34	72.6
MagicMask (Our)	98.41	**1.47**	**2.04**	13.4	31.6

Table 2: Quantitative results of on the FF++ dataset (Rossler et al., 2019). Results of FaceSwap (FaceSwap), DeepFakes (DeepFakes, 2020), FaceShifter (Li et al., 2020) and MegaFS (Zhu et al., 2021) are obtained from their websites. † denotes that we ran officially released source codes to obtain the results.

et al. (2023), which is a diffusion-based method. About the execution speed, MagicMask achieves a slightly slow but still real-time execution speed (27.9 ms), which is around 28 FPS. This figure is slightly slower than others (21.5 ms of FSGAN and 22.3ms of FaceSwapper) that use the face images only. But compared with the FaceDancer (72.6 ms), it's much faster and good enough for a real-time performance. About the FID, while MagicMask is slightly behind FaceDancer on raw ID accuracy and behind HifiFace on FID, it simultaneously delivers strong perceptual quality and the best geometric consistency, yielding the most balanced overall performance. The experimental results on the FF++ dataset demonstrate that the MagicMask can produce face-swapping performance comparable to that of the existing state-of-the-art methods. Consequently, even though the model complexity is being increased by using the geometric representation module, and parameters are additionally optimised with the AFSA loss, that modification is helpful in improving the performance.

Extreme face pose cases: Since not many methods report their performance on the MPIE and LPFF datasets, we compared the performance with methods that released their project in public repositories. FSGAN (Nirkin et al., 2019), SimSwap (Chen et al., 2020), BlandFace (Shiohara et al.,

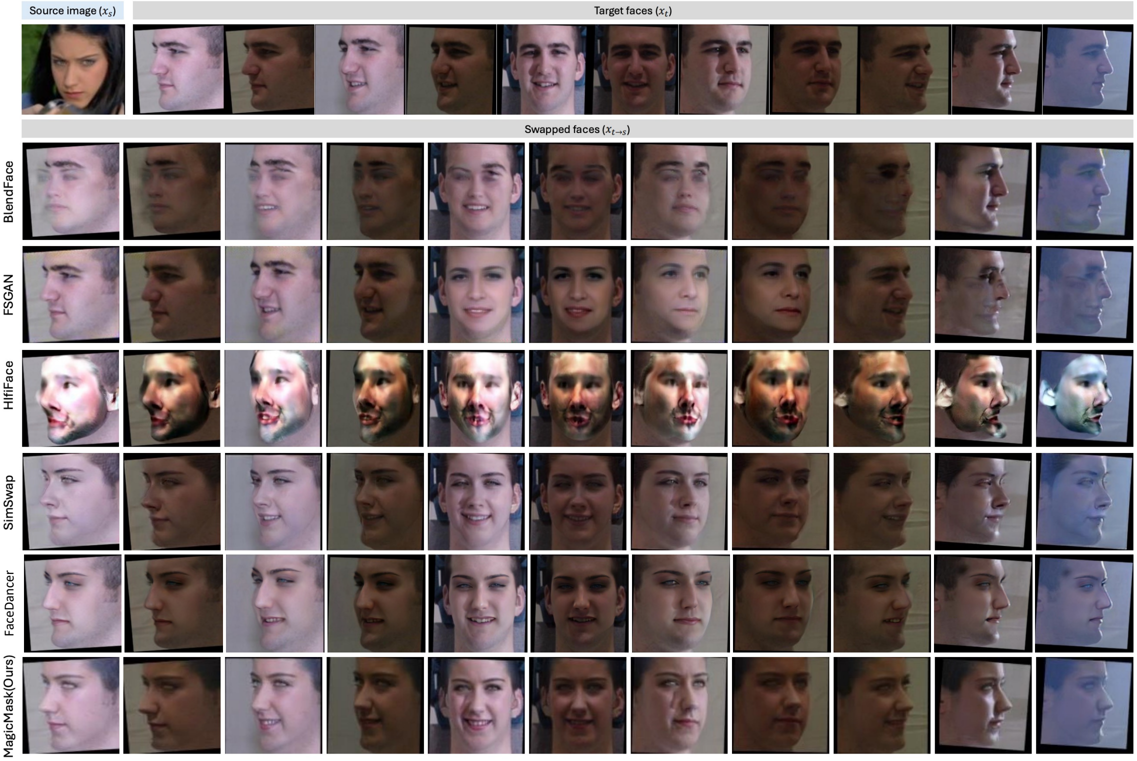


Figure 4: The face identity swapping results of the MagicMask and other methods (Shiohara et al., 2023; Nirkin et al., 2019; Wang et al., 2021b; Chen et al., 2020; Rosberg et al., 2023) on the MPIE dataset (Gross et al., 2010). Appendix D provides extended results on the MPIE dataset.

Method	**CSIM** ↑	**pose error** ↓	**expression error** ↓
FSGAN† (Nirkin et al., 2019)	0.105	5.31	4.02
SimSwap† (Chen et al., 2020)	0.180	3.92	3.81
BlendFace† (Shiohara et al., 2023)	0.392	3.71	3.18
HifiFace† (Wang et al., 2021b)	0.092	5.01	4.65
FaceDancer† (Rosberg et al., 2023)	0.401	4.72	3.31
MagicMask (Ours)	**0.463**	**3.35**	**2.91**

Table 3: Quantitative results on the MPIE dataset. † denotes that we ran officially released source codes to obtain the results.

2023), HifiFace (Wang et al., 2021b), and FaceDancer (Rosberg et al., 2023) are selected. We provide the URLs that we used to obtain their source codes in Appendix B.

Figure 4 and Table 3 show quantitative and qualitative results for the MPIE dataset, respectively. The results of BlendFace and FSGAN (second and third row of Figure 4) are mismatched with actual face areas or could not even generate suitable swapped faces. Their results in extreme poses contain a lot of image corruption. Hififace output is totally disrupted; its facial components are not properly swapped, and boundaries are also corrupted. SimSwap (5th row) and FaceDancer (6th row) achieve better results than others, and those are even compatible with MagicMask. However, SimSwap's

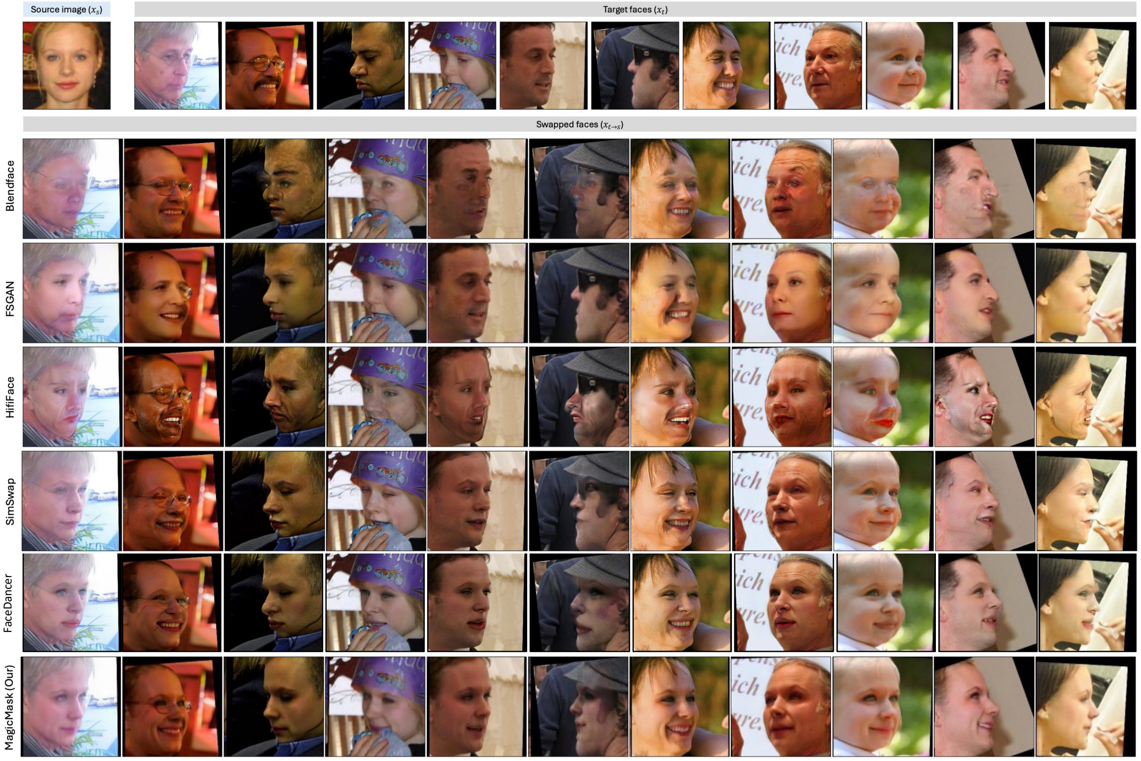


Figure 5: The face identity swapping result of the MagicMask and other methods (Shiohara et al., 2023; Nirkin et al., 2019; Wang et al., 2021b; Chen et al., 2020; Rosberg et al., 2023) on LPFF dataset (Wu et al., 2023). Appendix E provides extended results on the LPFF dataset. You can see results on more varied facial poses, including rotating and tilting.

facial silhouettes are more blurred in extreme pose cases. Also, FaceDancer's output identity is sometimes closer to the target. For example, the nose shape looks more like the target face.

The quantitative results also show similar circumstances. MagicMask achieves the highest CSIM score (0.463), significantly outperforming competing methods: FaceDancer (0.401), SimSwap (0.180), BlendFace (0.392), and HifiFace (0.092). This indicates that MagicMask excels in preserving the identity of the source face. Additionally, MagicMask achieves the lowest pose error (3.35) and expression error (2.91). Those results demonstrate the superiority of MagicMask in face identity-swapping tasks under diverse facial poses.

The quantitative results on the LPFF dataset shown in Figure 5 also suggest that the MagicMask provides more robust performances in swapping face identity in various pose variations. Similar to the results on the MPIE dataset, the swapped results were obtained by Blendface and FSGAN. The results of Hififace appear better than those of the MPIE dataset, but it still outputs disrupted facial boundaries and components. SimSwap and FaceDancer sometimes generated more detailed textures than the MagicMask, but in extreme pose cases, the results of the MagicMask look more natural and better represent the source identity. In general, MagicMask consistently outperforms state-of-the-art methods across both datasets, demonstrating its effectiveness and reliability in face identity-swapping tasks. Its superior performance in identity similarity (CSIM), pose preservation, and expression accuracy establishes it as a robust and reliable solution for handling challenging conditions in face-swapping applications.

5. Conclusion

In this work, we introduced MagicMask, a groundbreaking face-swapping framework designed to maintain identity consistency and visual fidelity under extreme pose variations. By integrating visual and geometric cues through an attention-based mechanism and introducing the Adversarial Facial Silhouette Alignment (AFSA) loss, our approach improves robustness on face poses and outperforms existing state-of-the-art methods on multiple benchmarks containing large facial pose variations.

Despite the additional computational cost and dependency on explicit geometric inputs, MagicMask significantly enhances identity preservation and pose accuracy, addressing key limitations in existing face-swapping techniques. These advancements establish a strong foundation for future research in pose-invariant facial synthesis and identity-aware generative models. Future research will focus on reducing reliance on explicit geometric inputs, improving computational efficiency, and extending this framework to multi-view and real-time settings, further expanding the possibilities of high-fidelity face manipulation.

Acknowledgement

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation(IITP)-Innovative Human Resource Development for Local Intellectualisation program grant funded by the Korea government(MSIT)(IITP-2025-RS-2024-00436773)

References

Sanoojan Baliah, Qinliang Lin, Shengcai Liao, Xiaodan Liang, and Muhammad Haris Khan. Realistic and efficient face swapping: A unified approach with diffusion models. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1062–1071. IEEE, 2025.

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE international conference on computer vision*, pages 1021–1030, 2017.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018.

Renwang Chen, Xuanhong Chen, Bingbing Ni, and Yanhao Ge. Simswap: An efficient framework for high fidelity face swapping. In *Proceedings of the ACM International Conference on Multimedia*, pages 2003–2011, 2020.

Xuanhong Chen, Bingbing Ni, Yutian Liu, Naiyuan Liu, Zhilin Zeng, and Hang Wang. Simswap++: Towards faster and high-quality identity swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

DeepFakes. faceswap. https://github.com/deepfakes/faceswap, 2020.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.

Tamara Newlands-Executive Director. Submission by echildhood reviews of the enhancing online safety act 2015 and the online content scheme. 2018.

FaceSwap. https://github.com/ondyari/FaceForensics/tree/master/dataset/FaceSwapKowalski. Accessed: 2022-02-14.

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM Transactions on Graphics (ToG), Proc. SIGGRAPH*, 40(4):88:1–88:13, August 2021.

Leon A Gatys. A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*, 2015.

Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2414–2423, 2016.

Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and vision computing*, 28(5):807–813, 2010.

Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *International Conference on Learning Representations*, 2018.

Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1867–1874, 2014.

Lingzhi Li, Jianmin Bao, Hao Yang, Dong Chen, and Fang Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

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

Qi Li, Weining Wang, Chengzhong Xu, Zhenan Sun, and Ming-Hsuan Yang. Learning disentangled representation for one-shot progressive face swapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Yixuan Li, Chao Ma, Yichao Yan, Wenhan Zhu, and Xiaokang Yang. 3d-aware face swapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12705–12714, 2023.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

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

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

Delong Qi, Weijun Tan, Qi Yao, and Jingfeng Liu. Yolo5face: Why reinventing a face detector. 2021.

Felix Rosberg, Eren Erdal Aksoy, Fernando Alonso-Fernandez, and Cristofer Englund. Facedancer: Pose-and occlusion-aware high fidelity face swapping. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3454–3463, 2023.

Andreas Rossler, Davide Cozzolino, Luisa Verdoliva, Christian Riess, Justus Thies, and Matthias Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1–11, 2019.

Kaede Shiohara, Xingchao Yang, and Takafumi Taketomi. Blendface: Re-designing identity encoders for face-swapping. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7634–7644, 2023.

Tianyi Wang, Zian Li, Ruixia Liu, Yinglong Wang, and Liqiang Nie. An efficient attribute-preserving framework for face swapping. *IEEE Transactions on Multimedia*, 2024.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9168–9178, 2021a.

Yuhan Wang, Xu Chen, Junwei Zhu, Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Yongjian Wu, Feiyue Huang, and Rongrong Ji. Hififace: 3d shape and semantic prior guided high fidelity face swapping. *arXiv preprint arXiv:2106.09965*, 2021b.

Yiqian Wu, Jing Zhang, Hongbo Fu, and Xiaogang Jin. Lpff: A portrait dataset for face generators across large poses. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20327–20337, 2023.

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

Yangyang Xu, Bailin Deng, Junle Wang, Yanqing Jing, Jia Pan, and Shengfeng He. High-resolution face swapping via latent semantics disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7642–7651, 2022b.

Changqian Yu, Jingbo Wang, Chao Peng, Changxin Gao, Gang Yu, and Nong Sang. Bisenet: Bilateral segmentation network for real-time semantic segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 325–341, 2018.

Longhao Zhang, Huihua Yang, Tian Qiu, and Lingqiao Li. Ap-gan: Improving attribute preservation in video face swapping. *IEEE transactions on circuits and systems for video technology*, 32(4):2226–2237, 2021.

Wenliang Zhao, Yongming Rao, Weikang Shi, Zuyan Liu, Jie Zhou, and Jiwen Lu. Diffswap: High-fidelity and controllable face swapping via 3d-aware masked diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8568–8577, 2023.

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