IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 29 May 2024, accepted 23 June 2024, date of publication 1 July 2024, date of current version 9 July 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3421528*

RESEARCH ARTICLE

# Face Swapping for Low-Resolution and Occluded Images In-the-Wild

**JAEHYUN PARK[1], (Graduate Student Member, IEEE), WONJUN KANG[2], (Student Member, IEEE), HYUNG IL KOO[3], (Member, IEEE), AND NAM IK CHO[1,2], (Senior Member, IEEE)**

[1]Interdisciplinary Program in Artificial Intelligence, Seoul National University, Seoul 08826, South Korea
[2]Department of Electrical and Computer Engineering, INMC, Seoul National University, Seoul 08826, South Korea
[3]Department of Electrical and Computer Engineering, Ajou University, Suwon-si 16499, South Korea

Corresponding author: Hyung Il Koo (hikoo@ajou.ac.kr)

This work was supported in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by Korean Government [Ministry of Science and ICT (MSIT)] under Grant 2021-0-01062; and in part by the BK21 FOUR Program of the Education and Research Program for Future ICT Pioneers, Seoul National University, in 2024.

**ABSTRACT** Safeguarding personal identity in various surveillance videos, dashcams, and on-street videos is crucial. One way to do this is to detect faces and blur them, but a better solution is to replace them with non-existent ones to maintain the naturalness of the videos. While face swapping methods have already been used in the media industry with high-quality faces, it is challenging to apply them for identity protection to faces in-the-wild where faces are often occluded and of low-resolution. Therefore, we propose a new framework for face swapping specifically designed to work with face images taken in real-world scenarios, making it useful as a privacy protection method. To tackle the issue of low-resolution images, we introduce a Cross-Resolution Contrastive Loss (CRCL) technique, which allows our neural network model to be trained using triplets of varying resolutions. This enables the model to learn and use identity information across different resolutions, thereby improving its accuracy. We also propose a plug-and-play framework that can be easily applied to existing face swapping models to handle occlusions. By explicit swapping of facial features and filling of occluded regions, our framework provides a more seamless blend. To demonstrate the effectiveness of our method in handling faces in-the-wild, we create an occluded VGGFace2 dataset consisting of face images augmented with various facial masks and hand occlusions. Through quantitative and qualitative assessments on this dataset, our proposed method demonstrates robust performance under low-resolution or occluded scenarios. Significant improvements are made in the quality of swapped faces while preserving their identity and attributes, highlighting the effectiveness of our framework in advancing face swapping as a reliable privacy protection measure.

**INDEX TERMS** Deep learning, de-identification, face swapping, in-the-wild, low-resolution, occlusion, privacy protection.

## I. INTRODUCTION

The objective of face swapping is to transfer the identity from a source face image onto a target face image while preserving the target's attributes, such as background, pose, and expression. This area has been extensively explored in computer vision and has applications in the media industry [10], [11], [17], [40], [41], [42], where high-quality images of source and target faces are crucial [10], [17], [18].

Face swapping can also be used as a de-identification method, which generally refers to anonymizing unlicensed individuals by removing personal information. This allows captured videos to be used as publicly available resources without privacy concerns. Specifically, face swapping involves replacing unlicensed face images with artificially generated non-existent identities that contain

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

no personal information [35], making it a powerful de-identification tool.

Face swapping offers several unique advantages compared to traditional image de-identification methods, such as blurring, pixelation, and masking. Unlike conventional methods that reduce image quality to remove identity information, face swapping synthesizes faces that remain within the natural distribution, preserving both the fidelity and integrity of the original image. Additionally, when done seamlessly, the swapped faces are imperceptible to the public and can accurately retain the target attributes of the original face, such as pose and expression. These advantages make the de-identified data more valuable as a publicly available training dataset.

However, the quality of faces in-the-wild can often be low, which makes it difficult to use existing face swapping methods developed for the media industry. Specifically, videos captured by surveillance cameras or public footage present two main challenges for face swapping: low-resolution and occlusions. Surveillance images and public footage are usually captured from a far distance using low-quality security cameras. As a result, the images are low-resolution, and the faces are often just a few pixels wide. Additionally, these captures are typically made in public settings, which commonly results in occlusions. This difference is illustrated in Fig. 1 (a), which are images used for face swapping in the media industry, having generally high-quality and are captured in a controlled setting [14], [15]. In contrast, the images for privacy protection [20], [25] in Fig. 1 (b) present varying resolution with occlusions. Existing face swapping methods developed for media applications [10], [11], [17], [40], [41], [42] are primarily trained with high-quality image datasets, such as CelebA-HQ [14] and FFHQ [15], and hence perform poorly on low-quality images. While there have been attempts to develop face swapping as a de-identification method [34], [35], [36], [37], [43], they focused on the architectural modifications for complete identity removal and less on the inherent challenges associated with low-resolution and occlusions in practical privacy protection settings.

Unlike the existing methods, we overcome these challenges by presenting a novel face swapping framework for privacy protection. First, we introduce a cross-resolution contrastive loss (CRCL) to enhance the robustness of the identity embedder in handling a wide range of resolutions. This loss allows the identity attributes of faces with varying resolution to share a single identity embedding space, making the face swapping model compatible with identity information across a range of resolutions.

Second, to handle facial occlusions, we develop a framework that decouples occlusion handling from the main face swapping process. Facial occlusions need to be preserved through the face swapping process. However, they also interrupt with facial feature extraction. Thus, we employ an occlusion parser and an inpainting module to explicitly extract and inpaint the occluded area of the face prior to face swapping. Once face swapping is performed on the inpainted image, it is followed by the refinement module to place the extracted mask back onto the faces seamlessly. The two approaches taken to address the challenges are independent of the face swapping process, and hence can be applied in a plug-and-play manner to enhance the robustness of existing face swapping models against low-resolution and occlusions.

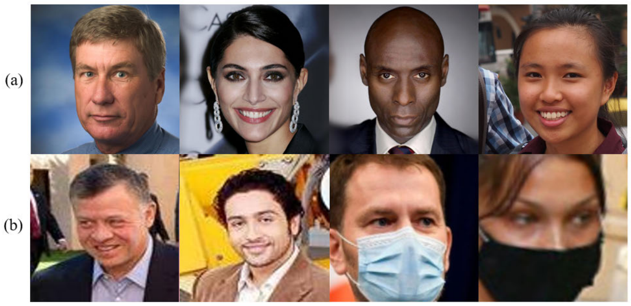

**FIGURE 1.** Examples of a) high-quality images [14], [15] used in media and b) in-the-wild images [20], [25] of varying resolution with occlusions.

In summary, our contributions can be summarized as:

- We have introduced a cross-resolution contrastive loss (CRCL) to create a shared embedding space for identity embeddings across different resolutions. This enhances the robustness of face swapping models against low-resolution images.
- We have developed an occlusion-handling framework that specifically deals with facial occlusions through extraction and swapping. This improves the robustness of face swapping models against various occlusions.
- We have also introduced an occluded VGGFace2 dataset to assess the proposed method on face images with synthetic occlusions. Our method has shown to be effective for privacy protection, remaining robust against faces in real-world scenarios, and handling low-resolution and occluded images exceptionally well.

The remaining paper is organized as follows: Section II discusses various related works on face swapping and methods for identity extraction. Section III introduces our method, largely divided into the cross-resolution contrastive loss and the occlusion handling framework. Section IV details the training parameters and settings. Section V presents quantitative and qualitative evaluation results, followed by the conclusion in Section VI.

## II. RELATED WORK

Since the proposed method is based on the face swapping algorithm, we first review face swapping methods. Then, we discuss methods to extract identity embeddings, which play a crucial role in many recent face swapping methods.

### A. FACE SWAPPING

Face swapping has received much attention due to its vast range of potential applications. As a result, numerous deep learning-based methods have been proposed. Korshunova et al. [2] first attempted to address this task

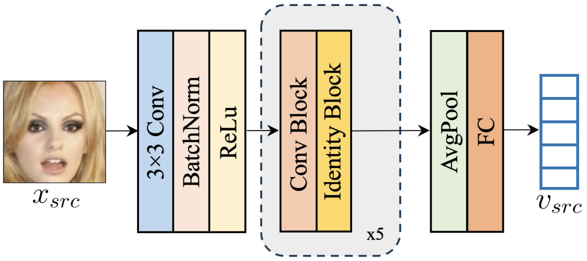

**FIGURE 2.** Architecture of the ID embedder $E$, which uses ArcFace [12] as its backbone.

using a simple convolutional network. Nirkin et al. [6] developed the convolutional neural network approach using face segmentation guidance. Soon after, the researchers shifted towards 3D-based methods [4], [6]. By fitting source and target faces to a 3D template, the authors in [7] were able to control expression and head poses. However, these methods showed limitations in manipulating attributes such as style and illumination, which led to GAN-based approaches. Amongst the many GAN-based approaches, RSGAN [3] carried out face swapping in the feature latent space. FSGAN [8] implemented a two-stage approach to carry out face swapping and reenactment simultaneously. Also, an efficient framework for generalized and high-fidelity swapped results was proposed in [1].

Although Faceshifter [9] attached HEAR-Net to handle occlusions in the target image, it was limited to small occlusions that covered only a small portion of the face. Instead, most of the recent work focused on higher-resolution/high-quality image generation. MegaFS [10] and FSLSD [11] achieved state-of-the-art results on a set of higher resolution datasets (CelebA-HQ [14], and FFHQ [15]) by exploiting a pre-trained StyleGAN2 generator [13] along with latent space manipulation techniques. Disentanglement of features in the latent space has allowed precise control over specific identity and target attributes in the image space.

### B. IDENTITY EXTRACTION

The face swapping task involves transferring the identity of the source image to a target face while preserving the attributes of the target face. Hence the capability of the identity embedder to extract the identity features from the source face correctly is crucial. Identity features are the key traits that distinguish one face from another (e.g., the color and shape of eyes, nose, and lips). As these features coincide with features used in face recognition, many of the early face swapping works adopted the Arcface [12] model (from face recognition) as their identity embedder. However, Smoothswap [16] pointed out that identity embedding space trained on discriminative tasks is not continuous and proposed a smooth identity embedder with stable gradients for identity interpolation. BlendFace [26] re-designed the identity embedder and trained it on blended images to alleviate the problem of identity-attribute entanglement.

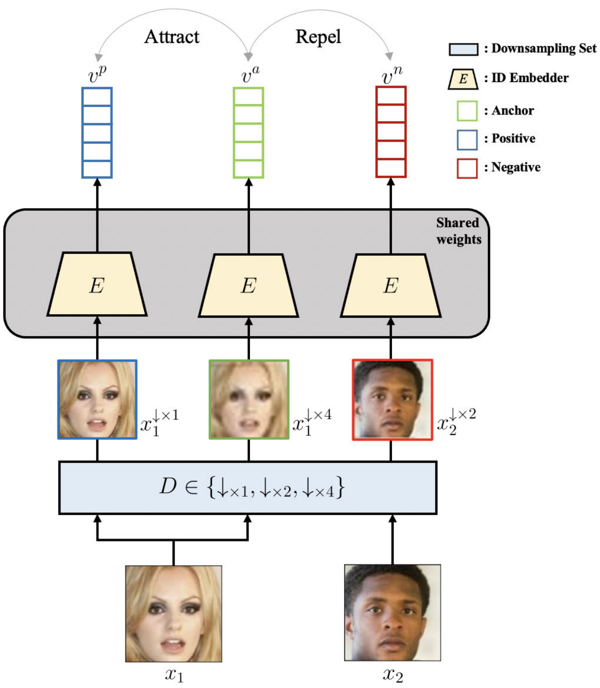
Attract
Repel
$v^p$
$v^a$
$v^n$
: Downsampling Set
E : ID Embedder
: Anchor
: Positive
: Negative
Shared weights
E
E
E
$x_1^{\downarrow\times 1}$
$x_1^{\downarrow\times 4}$
$x_2^{\downarrow\times 2}$
$D \in \{\downarrow_{\times 1}, \downarrow_{\times 2}, \downarrow_{\times 4}\}$
$x_1$
$x_2$

**FIGURE 3.** Finetuning procedure of our identity embedder $E$ using the cross-resolution contrastive loss. A shared identity embedding space is learned across resolution.

Identity in face swapping can also be handled by utilizing the GAN inversion methods: Identity and attribute features are fused and controlled in the StyleGAN latent space W++ [15]. RAFswap [17] used semantic labels to extract region-wise identity tokens. On the other hand, the authors in [18] relied on transferring target attributes to the source face without explicitly extracting identity features from the source face. Although these methods have shown high-quality results, conventional identity embedders have limitations in handling low-resolution images. The identity of swapped results, especially in low-resolution scenarios, may not accurately reflect the identity of the source images.

## III. PROPOSED METHOD

The face swapping problem is formulated as generating a swapped face, denoted as $x_{swap}$, with the identity from $x_{src}$ and the attributes (e.g., background, expression, and pose) of $x_{tgt}$. However, face swapping as a privacy protection measure needs to serve a very different scenario: replacing the identities of unwanted individuals in low-quality surveillance videos with identities of high-quality generated or licensed face images. Due to the quality disparity in the source and target images, the problem of face swapping for privacy protection is formulated asymmetrically. It aims to generate a realistic swapped image $x_{swap}$ given a high-quality source image $x_{src}$ and a low-quality in-the-wild target image $x_{tgt}$. Among the factors influencing the target image's quality, we specifically address low-resolution and facial occlusions through our cross-resolution contrastive loss and occlusion handling framework, respectively.

### A. CROSS-RESOLUTION CONTRASTIVE LOSS

To extract the identity of $x_{src}$, conventional face swapping models typically employ a pre-trained face-recognition model Arcface [12] as their identity embedder $E$. The Arcface identity embedder uses a ResNet50 [32] backbone, composed of convolution, identity blocks, and a final FC layer, to project the input face image $x_{src}$ to a 512-dim identity embedding vector $v_{src}$. The architecture of the Arcface identity embedder $E$ is shown in Fig. 2.

Unfortunately, for in-the-wild images, the resolution gap between the high-quality source $x_{src}$ and low-quality target image $x_{tgt}$ results in embeddings from differing resolutions that are incompatible, leading to artifacts around the facial regions. A straightforward remedy to this problem would be to reduce the resolution gap by resizing the source image to the resolution of the target image. However, the identity embeddings extracted from the Arcface identity embedder work poorly for low-resolution images, as it was primarily trained on discriminative tasks on a single resolution. Consequently, the identity embedder encounters difficulties in generalizing across varying resolutions.

To alleviate this problem, we fine-tune the pre-trained Arcface identity embedder on additional lower-resolution images with a cross-resolution contrastive loss (CRCL). To be precise, we fine-tune our identity embedder $E$ on 1/2 and 1/4 of the VGGFace2 dataset's [20] original input image size (224 × 224) to learn a shared identity embedding space across 56 × 56, 112 × 112, and 224 × 224 resolutions. The standard triplet loss function leverages the identity of the input face by pushing differing identities and pulling identical identities. In cross-resolution contrastive loss, the triplet loss is expanded to leverage both identities and resolutions. This fine-tuning procedure is illustrated in Fig. 3.

Given a pair of images of different identities $I_1$ and $I_2$, we use $I_1$ as our anchor and positive image and $I_2$ as our negative image to form an image triplet $(o_1(x_1), o_2(x_1), o_3(x_2))$, where $o_i$ is a randomly chosen downsampling operator from set $D \in \{\downarrow_{\times 1}, \downarrow_{\times 2}, \downarrow_{\times 4}\}$. We apply triplet losses across resolution to push and pull according to their identities regardless of the operator they go through. Hence, the identity embedder $E$ naturally learns a shared identity embedding space across multiple resolutions. The cross-resolution contrastive loss function is given by

$$\mathcal{L}(x_1, x_2) = \max\left(\|v^a - v^p\| - \|v^a - v^n\| + \alpha, 0\right) \quad (1)$$

where $\alpha$ is a positive constant and $(v^a, v^p, v^n)$ is the identity embedding obtained by $E$:

$$(v^a, v^p, v^n) = (E(o_1(x_1)), E(o_2(x_1)), E(o_3(x_2)))\,. \quad (2)$$

The resulting shared identity embedding space yields accurate identity embeddings across multiple resolutions, which is then leveraged by face swapping models to perform natural face swapping even in low-resolution scenarios.

### B. OCCLUSION HANDLING FRAMEWORK

As attributes of target images, occlusions in the target face should be retained throughout the face swapping process. Therefore, we develop a framework to handle occlusions through explicit steps of extraction, inpainting, and refinement. The framework is illustrated in Fig. 4 and the individual modules used are summarized in Table 1. To elaborate, the occlusion handling framework is broken down into the following steps.

#### 1) OCCLUSION EXTRACTION

We first extract occlusions in the target face using an occlusion parser $P$. Given a target face image $x_{tgt}$, the occlusion parser outputs mask $m$ (1: occlusions and 0: others) corresponding to the occluded region of the target image. From the output $m$, we expand it by 10% to ensure complete coverage of the occlusion in the extracted mask since unmasked occlusions may interfere heavily with the following inpainting process. The extracted mask $m$ is used to extract the occlusion $x_{tgt}^{occ}$ in (4) and the occlusion-removed target image $x_{tgt}^{face}$ in (5).

$$m = P(x_{tgt}) \quad (3)$$

$$x_{tgt}^{occ} = m \odot x_{tgt} \quad (4)$$

$$x_{tgt}^{face} = (1 - m) \odot x_{tgt}\,. \quad (5)$$

As our occlusion parser, we modify the ResNet101 [32] backbone with fewer ResBlocks and a binary classification layer for our mask output. The described architecture is depicted in Fig 5.

#### 2) OCCLUSION INPAINTING

Then, we use an inpainting module $I$ to inpaint the occlusion region of the target image. The inpainting module is able to fill in missing facial landmarks and semantics while preserving facial symmetry.

$$x_{tgt}^{inpainted} = I(x_{tgt}^{face}). \quad (6)$$

This step is crucial as the face swapping module requires complete facial features to produce high-quality swapped results. Note that the similarity between the inpainted region and the original face is unimportant as the inpainted region is later replaced with the extracted occlusion $x_{tgt}^{occ}$.

#### 3) FACE SWAPPING

The inpainted image is then passed into the face swapping module $F$ along with the extracted identity embedding vector $v_{src}$ in (7) (from the CRCL fine-tuned identity embedder $E$) to produce the face swapped result $x_{swap}$ in (8).

$$v_{src} = E(x_{src}) \quad (7)$$

$$x_{swap} = F(v_{src}, x_{tgt}^{inpainted}) \quad (8)$$

Our occlusion handling framework is composed of the pre and post-processing step of the face swapping module. Hence, any existing face swapping methods can be used in place of our face swapping module $F$ in (8).

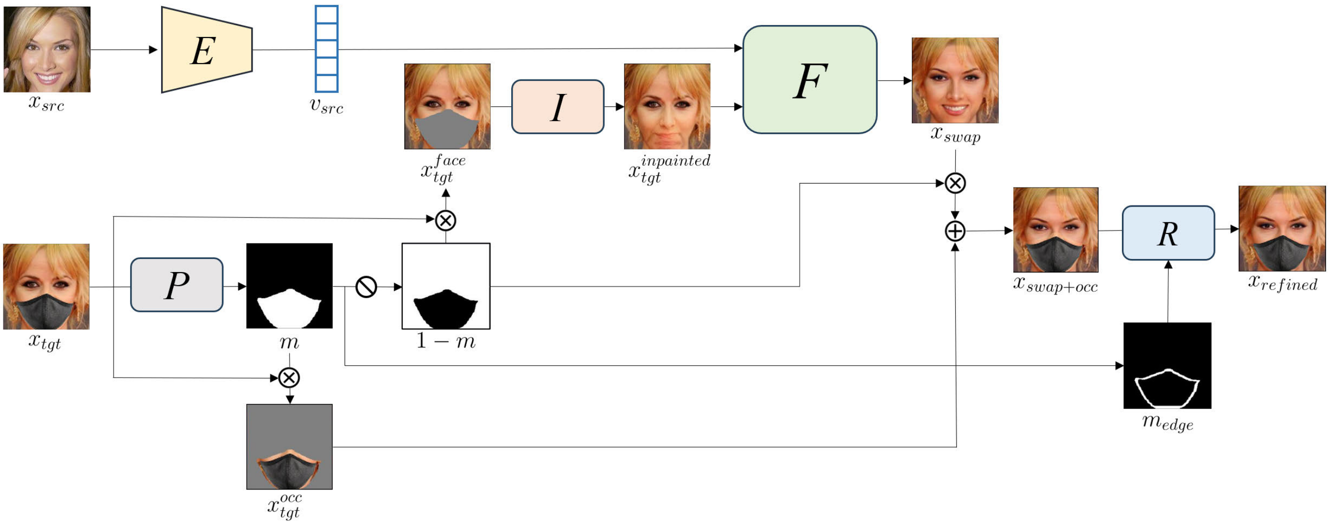

**FIGURE 4.** Overall structure of the proposed method. We first train the occlusion parser $P$ and the inpainting module $I$ to mask and inpaint the occluded target face $x_{tgt}$. We also adopt the CRCL finetuned identity embedder $E$ to extract the source identity embedding $v_{src}$. Then $E, P, I$ are used to train the face swapping module $F$ and refinement module $R$.

**TABLE 1.** Summary of the modules used in the proposed method.

| Network | Notation | Role | Model |
|---|---|---|---|
| CRCL Finetuned ID Embedder | $E$ | Extract resolution-robust identity embedding $v_{src}$ | Arcface [12] |
| Occlusion parser | $P$ | Extract occlusion mask $M$ from face image | Modified ResNet101 [32] |
| Inpainting module | $I$ | Inpaint occluded region of face | LaMa [24] |
| Face swapping module | $F$ | Generate swapped image $x_{swap}$ from $x_{src}$ and $x_{tgt}$ | Simswap [1] |
| Refinement module | $R$ | Refine the boundaries of the overlayed occlusion | U-Net [33] |

**TABLE 2.** Summary of the training parameters of the modules used. The modules are trained across two stages.

| Network | Stage 1 | Stage 2 | Loss | Batch size | Learning rate | Optimizer | GPU |
|---|---|---|---|---|---|---|---|
| $E$ | ✓ |  | CRCL (1) | 128 | 0.001 | Adam | GTX TITAN X |
| $P$ | ✓ |  | BCE | 16 | 0.001 | Adam | RTX 2080 Ti |
| $I$ | ✓ |  | $\mathcal{L}_{lama}$ [24] | 8 | 0.0001 | Adam | RTX 2080 Ti |
| $F$ & $R$ |  | ✓ | $\mathcal{L}$ (14) | 8 | 0.1 | Adam | RTX 2080 TI |

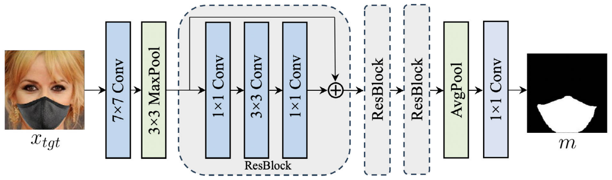

**FIGURE 5.** Architecture of our Occlusion Parser $P$. We use a modified version of ResNet101 [32].

### 4) STITCHING AND REFINEMENT

To restore occlusions in the target face, we transfer the occlusions $x_{tgt}^{occ}$ back onto the swapped face. Initially, the occlusion mask layer is simply overlayed above the swapped image layer using Gaussian blending in (9). However, to minimize the disparity between the layers, we post-process the occluded image $x_{swap+occ}$ with a refinement module $R$ to produce a natural result $x_{refined}$ in (10).

**TABLE 3.** Architecture of Refinement module $R$.

| $R$ |
|---|
| Conv(64, 4×4, 2×1) |
| Conv(128, 4×4, 2×1) |
| Conv(256, 4×4, 2×1) |
| Conv(512, 4×4, 2×1) |
| Conv(512, 4×4, 2×1) |
| ConvTranspose(512, 4×4, 2×1) |
| ConvTranspose(256, 4×4, 2×1) |
| ConvTranspose(128, 4×4, 2×1) |
| ConvTranspose(64, 4×4, 2×1) |
| Interpolate(Bi-linear, x2) |
| ConvTranspose(3, 4×4, 2×1) |
| Tanh |

$$x_{swap+occ} = x_{swap} \odot (1-m) + x_{tgt}^{occ} \tag{9}$$

$$x_{refined} = R(x_{swap+occ}, m) \tag{10}$$

The simple U-net architecture for our refinement module is listed in Table 3.

The occlusion handling framework is summarized in these four steps: explicit extraction, inpainting, swapping, and refinement. It can easily be applied to existing face swapping methods to enhance their robustness to facial occlusions.

## IV. TRAINING

As presented in the previous section, the proposed method consists of 5 neural network modules. In the first stage, the identity embedder $E$, occlusion parser $P$, and inpainting module $I$ are trained. Then, the trained modules are employed in stage 2 to jointly train the face swapping module $F$ and refinement module $R$. Especially for the joint training of $F$ and $R$, we propose a new loss function to yield a natural result. All modules used in this paper are summarized in Table 1, and each of its training parameters are summarized in Table 2.

### A. IDENTITY EMBEDDER

We first fine-tune the pre-trained Arcface identity embedder $E$ from [12] on low-resolution images. We use MS1MV2 [19], a large-scale face recognition dataset consisting of over 100K identities, augmented with low-resolution images as our dataset. To synthetically generate low-resolution images, we downsample the images to 112 × 112 and 56 × 56 images using bicubic interpolation and upsample them back to the original size. We use an ADAM optimizer with a learning rate of 0.001, scheduled to halve every 4 epochs. The identity embedder is trained with a batch size of 128 on GTX TITAN X GPU for about 4 days.

### B. OCCLUSION PARSER AND INPAINTING MODULE

The occlusion parser $P$ and inpainting module $I$ are trained using VGGFace2 [20], a dataset of 3.31 million face images of 9,131 different classes. Prior to training, the dataset is refined by removing images with sizes smaller than 250 × 250, then aligned and cropped to a size of 224 × 224. 80% of the dataset is used for training purposes, and the remaining 20% is used for validation.

To facilitate the training of the model on low-quality images, we further augment the VGGFace2 dataset [20]. First, we make low-resolution images following the method in Section IV-A. Then, we create and propose two synthetically occluded VGGFace2 datasets: Masked-VGGFace2 and Hand-occluded-VGGFace2, featuring VGGFace2 images overlayed with synthetic facial masks and hands, respectively. The Masked-VGGFace2 dataset uses MaskTheFace [25] to augment VGGFace2 images with synthetic facial masks of random types and colors. The Hand-occluded-VGGFace2 dataset uses 11k-hands [27] to augment VGGFace2 images with synthetic hand occlusions of random orientations and sizes. This multifaceted augmentation strategy contributes to the robustness and diversity of our dataset.

The objective of the occlusion parser $P$ is to extract a binary occlusion mask from a target face accurately. To do this, we adopt a modified ResNet101 [32] model as the backbone, as depicted in Fig 5. The model is supervised with a Binary Cross Entropy (BCE) loss and is trained with a batch size of 16 for 50 epochs on an NVIDIA 2080 TI GPU for a single day. For our face inpainting module $I$, we adopt LaMa [24], an inpainting network that uses Fast Fourier Convolutions [39]. It is known for its effectiveness at inpainting large masks and its ability to generalize well to images of varying resolution. To train the inpainting module $I$, we utilize large random masked regions of the downsampled VGGFace2 face image. The module is trained with a batch size of 8 on an RTX 2080 Ti GPU for 3 days, using an ADAM optimizer with a learning rate of 0.0001.

### C. FACE SWAPPING AND REFINEMENT MODULE

Finally, we jointly train $F$ and $R$ using the trained modules $E$, $P$, and $I$. The goal of the refinement module is to match the color characteristics and refine the unnatural blending artifacts occurring at the edges of the transferred occlusion mask.

To focus on the boundaries of the occlusion, we apply a perceptual loss [31] on the edges of the occlusion in (11). This is done by generating an occlusion edge mask $m_{edge}$ by subtracting a down-scaled mask $m_{-}$ from a slightly up-scaled mask $m_{+}$. The perceptual loss function is given by

$$\mathcal{L}_{edge} = ||\phi_{4\_2}(m_{edge} \odot x_{refined}) - \phi_{4\_2}(m_{edge} \odot x_{swap+occ})||_1 \tag{11}$$

where $\phi_{4\_2}$ denotes the feature map after relu4_2 layer of the VGG19 model [31] pretrained on ImageNet [30]. A pixel-wise reconstruction loss (12) is also used as a regularization term so that the refined face $x_{refined}$ does not change too much from input face $x_{swap+occ}$.

$$\mathcal{L}_{recon} = ||x_{refined} - x_{swap+occ}||_1 \tag{12}$$

$$\mathcal{L}_{R} = \lambda_{recon}\mathcal{L}_{recon} + \lambda_{edge}\mathcal{L}_{edge} \tag{13}$$

We set $\lambda_{recon} = 10^{-5}$ and $\lambda_{edge} = 10$.

This loss is combined with the standard losses (identity, reconstruction, adversarial, and weak feature matching loss) of Simswap [1] to form the entire loss function (14). We follow the settings in Simswap and set $\lambda_{id} = 10$, $\lambda_{recon} = 10$, $\lambda_{GP} = 10^{-5}$ and $\lambda_{wFM} = 10$:

$$\mathcal{L} = \lambda_{id}\mathcal{L}_{id} + \lambda_{recon}\mathcal{L}_{recon} + \mathcal{L}_{adv} + \lambda_{GP}\mathcal{L}_{GP} + \lambda_{wFM}\mathcal{L}_{wFM} + \mathcal{L}_{R} \tag{14}$$

While occlusions on face images come in various forms, we specifically focus on facial masks and hand occlusions, which are the most common types. Nevertheless, the framework can be extended to handle various occlusions by training the occlusion parser $P$ on additional occlusions. The two modules, $F$ and $R$, are jointly trained with a batch size of 8 across 4 RTX 2080 Ti GPUs for around 3 days using an ADAM optimizer with $B1 = 0$ and $B2 = 0.999$, similar to the hyperparameter settings in Simswap [1].

**TABLE 4.** Comparison against previous works [1], [9], [29] on VGGFace2 dataset [20]. Our method shows robustness to x1, x2, and x4 downsampled resolutions over multiple metrics. Red indicate the best score, Δ indicate improvement of Ours over runner-up.

| | $D$ | ID sim.↑ | ID ret.↑ | Pose↓ | Exp.↓ | FID↓ |
|---|---|---|---|---|---|---|
| **Simswap** [1] | $\downarrow_{\times 1}$ | 0.835 | 94.0 | 111.9 | 1.530 | 22.06 |
| | $\downarrow_{\times 2}$ | 0.821 | 93.9 | 110.3 | 1.530 | 25.21 |
| | $\downarrow_{\times 4}$ | 0.739 | 91.5 | 114.2 | 1.549 | 33.92 |
| **HifiFace** [29] | $\downarrow_{\times 1}$ | 0.675 | 73.4 | 126.5 | 1.906 | 52.04 |
| | $\downarrow_{\times 2}$ | 0.671 | 73.5 | 129.2 | 1.892 | 60.93 |
| | $\downarrow_{\times 4}$ | 0.671 | 71.7 | 133.0 | 1.900 | 70.80 |
| **Faceshifter** [9] | $\downarrow_{\times 1}$ | 0.749 | 85.5 | 140.2 | 2.164 | 24.69 |
| | $\downarrow_{\times 2}$ | 0.745 | 85.7 | 137.0 | 2.167 | 29.42 |
| | $\downarrow_{\times 4}$ | 0.749 | 86.6 | 137.3 | 2.162 | 38.19 |
| **Ours** | $\downarrow_{\times 1}$ | 0.824 | 98.5 $\Delta_{4.5}$ | 118.2 | 1.715 | 19.22 $\Delta_{2.84}$ |
| | $\downarrow_{\times 2}$ | 0.820 | 98.0 $\Delta_{4.1}$ | 117.4 | 1.712 | 23.16 $\Delta_{2.05}$ |
| | $\downarrow_{\times 4}$ | 0.772 $\Delta_{0.023}$ | 92.8 $\Delta_{1.3}$ | 117.6 | 1.715 | 25.77 $\Delta_{8.15}$ |

## V. EXPERIMENTAL RESULTS

The experimental result is largely comprised of the quantitative evaluation and qualitative evaluation results against multiple previous works [1], [9], [29]. Recent works [10], [11], [17], [40], [41], [42] that are designed and trained on a higher resolution (e.g., megapixels) dataset [14], [15] are excluded as they are not suited for comparison.

For the quantitative evaluation, we first introduce the five standard metrics used to evaluate face swapping models. Afterwards, we use these metrics to evaluate the model's robustness to resolution on the VGGFace2 test set, and its robustness to occlusion on the synthetic Masked and Hand-occluded-VGGFace2 dataset. Likewise, qualitative evaluations on robustness to low-resolution and occlusion are made. Additional qualitative evaluations on real-world occlusion are conducted to demonstrate our method's ability to generalize to unseen practical occlusions. This is followed by an ablation study where we examine the individual effects of each of the proposed modules. Our proposed method is labeled as **Ours** in the following evaluations.

### A. QUANTITATIVE EVALUATION

Comparisons are made with respect to five widely used metrics: FID, ID similarity, ID retrieval, Pose error, and Expression error. The Frechet Inception Distance (FID) [28] is a measure of the quality of the data and specifically measures the Wasserstein-2 distance between the distribution of swapped images and the target face. Hence, it is a measure of the naturalness of the generated face.

ID similarity and retrieval measures the identity transfer capability of the model and hence indicates whether the identity from the source faces $x_{src}$, is accurately reflected onto the swapped face $x_{swap}$. ID similarity is measured by the cosine similarity between the identity embedding of the swapped and the source face. ID retrieval uses identity embedding extracted using our CRCL fine-tuned identity embedder to measure the top-1 matching rate amongst the source images.

Pose and expression errors represent the attribute preservation capability of the model and are measured by the L2 distance between the pose and expression vectors yielded by HopeNet [23]. Hence, a low pose and expression error indicates that the pose and expression of the swapped face have remained consistent with those of the target face, resulting in a seamless swapped face.

#### 1) RESULTS ON LOW-RESOLUTIONS

We first evaluate our proposed method on low-resolutions. As summarized in Table 4, previous works [1], [9], [29] have difficulty retrieving the correct identity in low-resolutions ($\downarrow_{\times 2}$, $\downarrow_{\times 4}$), underlined by ID Sim. and ID ret. scores dropping significantly on lower resolutions. The extraction of incorrect identity can lead to face swapped results that fail to reflect the identity of the source image. Moreover, it may also lead to blur artifacts and resolution discrepancies around the contours of the transferred identity properties and the retained target attributes. This observation is further supported by a noticeable decline in the FID score metric, which indicates that the distribution of the swapped face has deviated from the distribution of the target face. As target faces consist of real natural faces, the deviation indicates that previous works encountered difficulties generating a natural face in lower-resolution settings.

On the other hand, the proposed method shows robustness to various resolutions. The ID sim. score remains more consistent, indicating successful learning of the cross-resolution embedding space. Consequently, this leads to improved ID ret. and FID scores, indicating that the extracted identity is accurate and has been successfully transferred even in low-resolution settings. Note that while the ID sim. scores are higher for Simswap [1] for $\downarrow_{\times 1}$ and $\downarrow_{\times 2}$ resolutions, the ID ret. score is low. This translates to a concentrated distribution of the ID embedding resulting in high ID sim. scores on average, that is not well differentiated resulting in a low ID ret. score. In privacy protection, this implies the better removal of the target identity, which is the unlicensed identity we aim to remove, and a better reflection

**TABLE 5. Comparison against previous works [1], [9], [29] on Masked-VGGFace2 [20]. We use MaskTheFace [25] to apply synthetic masks of various types, color and texture on the VGGFace2 dataset. Red indicate the best score, Δ indicate improvement of Ours over runner-up.**

| | *D* | ID sim.↑ | ID ret.↑ | Pose↓ | Exp.↓ | FID↓ |
|---|---|---|---|---|---|---|
| **Simswap** [1] | ↓×1 | 0.573 | 87.0 | 128.7 | 1.726 | 29.77 |
| | ↓×2 | 0.568 | 84.3 | 128.3 | 1.723 | 29.89 |
| | ↓×4 | 0.578 | 85.4 | 130.2 | 1.755 | 39.20 |
| **HifiFace** [29] | ↓×1 | 0.438 | 55.1 | 135.1 | 2.059 | 49.35 |
| | ↓×2 | 0.435 | 53.0 | 142.2 | 2.092 | 59.95 |
| | ↓×4 | 0.434 | 51.9 | 143.0 | 2.121 | 88.25 |
| **Faceshifter** [9] | ↓×1 | 0.514 | 78.3 | 142.3 | 2.282 | 49.84 |
| | ↓×2 | 0.511 | 76.7 | 142.6 | 2.297 | 48.01 |
| | ↓×4 | 0.515 | 76.8 | 146.0 | 2.300 | 49.99 |
| **Ours** | ↓×1 | 0.748 Δ0.175 | 97.3 Δ10.3 | 126.4 Δ2.3 | 1.888 | 16.47 Δ13.30 |
| | ↓×2 | 0.740 Δ0.172 | 96.7 Δ12.4 | 127.7 Δ0.6 | 1.904 | 20.21 Δ9.68 |
| | ↓×4 | 0.716 Δ0.138 | 92.2 Δ6.8 | 129.3 Δ4.5 | 1.911 | 26.31 Δ12.89 |

**TABLE 6. Comparison against previous works [1], [9], [29] on the Hand-occluded-VGGFace2 dataset [20]. We use 11k-hands dataset to add synthetic hands of random orientation and sizes on the VGGFace2 dataset. Red indicate the best score, Δ indicate improvement of Ours over runner-up.**

| | *D* | ID sim.↑ | ID ret.↑ | Pose↓ | Exp.↓ | FID↓ |
|---|---|---|---|---|---|---|
| **Simswap** [1] | ↓×1 | 0.531 | 75.1 | 154.3 | 1.750 | 38.31 |
| | ↓×2 | 0.529 | 73.0 | 152.7 | 1.746 | 37.70 |
| | ↓×4 | 0.534 | 75.2 | 152.9 | 1.750 | 43.60 |
| **HifiFace** [29] | ↓×1 | 0.434 | 61.8 | 168.5 | 2.108 | 69.48 |
| | ↓×2 | 0.436 | 65.2 | 168.9 | 2.108 | 78.09 |
| | ↓×4 | 0.440 | 64.6 | 168.5 | 2.127 | 89.96 |
| **Faceshifter** [9] | ↓×1 | 0.464 | 67.5 | 170.3 | 2.265 | 39.60 |
| | ↓×2 | 0.480 | 71.1 | 171.4 | 2.299 | 36.50 |
| | ↓×4 | 0.485 | 71.9 | 173.6 | 2.304 | 44.88 |
| **Ours** | ↓×1 | 0.626 Δ0.095 | 75.5 Δ0.4 | 153.7 Δ0.6 | 1.855 | 23.62 Δ14.69 |
| | ↓×2 | 0.639 Δ0.110 | 77.3 Δ4.3 | 155.3 | 1.854 | 28.55 Δ7.95 |
| | ↓×4 | 0.667 Δ0.133 | 82.3 Δ7.1 | 151.9 Δ1.0 | 1.863 | 35.23 Δ8.37 |

of the generated artificial identity, which we aim to transfer over. As remaining unlicensed identity can lead to privacy protection problems, our framework's high retrieval rate underlines its effectiveness as a reliable privacy protection measure. As for pose and expression scores, we find that they remain relatively low and competitive with other models.

#### 2) RESULTS ON OCCLUSIONS

We also evaluate the robustness of our model on two types of occlusions: facial masks and hands. In Table 5, we compare the results on the Masked-VGGFace2 dataset. As shown, metric scores deteriorate compared to the original VGGFace2 dataset, indicating the challenges faced with occlusions. The drop becomes more noticeable in the pose and expression metrics as these metrics are evaluated in [23] by leveraging the facial geometry and landmark information of the input face. As occlusions such as facial masks largely occlude this information, occlusions heavily interfere with the extraction of pose and expression, leading to inaccurate preservation of target face attributes. Even steeper drops are observed on the Hand-occluded-VGGFace2 dataset as shown in Table 6. This is explained by the difficulty associated with differentiating hand occlusions due to their identical color with the face and their irregular shape.

On the other hand, our method works robustly in both occluded settings compared to previous works. As indicated by the low FID scores, the occlusion handling framework retains the occlusion to produce natural swapped results. As target face consists of occlusions in this case, the low FID score not only indicates the natural generation of the swapped face but also indicates that occlusions are preserved in a realistic manner. This is important as an unnatural swapped face harms the fidelity and integrity of the original data, rendering the method less effective in privacy protection. Moreover, it shows strong identity preservation capabilities through high ID sim. and ID ret. scores at the slight expense of its pose and expression preservation ability. Overall, upon quantitative evaluation using face swapping benchmark metrics, our proposed method displays strong performance, especially in low-resolution and occluded in-the-wild scenarios, in comparison to previous works [1], [9], [29].

### *B. QUALITATIVE EVALUATION*

#### 1) RESULTS ON LOW-RESOLUTIONS

Face swapping results on standard resolution (224 × 224) are shown in Fig. 6. In addition to the correct extraction of the source identity, our model is able to preserve the pose and expression of the target image accurately. Also, our results on

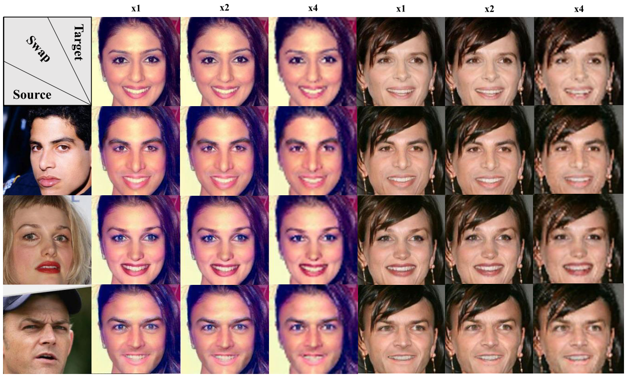

**FIGURE 6.** Face swapping results of our proposed method on the VGGFace2 test set [20] on low-resolution target images. Our method successfully retains the downsampled target image fidelity in the swapped result. (Best viewed zoomed in).

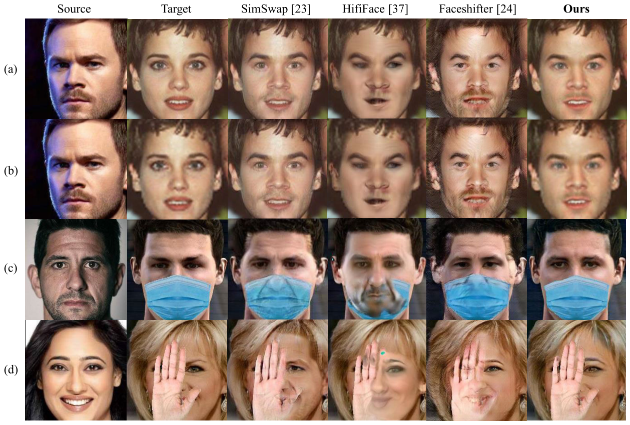

**FIGURE 7.** Comparison against previous works [1], [9], [29] on the VGGFace2 test set [20]. Rows (a) and (b) show face swapping performed at x2 and x4 resolutions. Rows (c) and (d) shown face swapping performed in occluded scenarios. Unlike previous models, our method remains faithful to the fidelity and the occlusions of the target face.

$\downarrow_{\times 2}$ and $\downarrow_{\times 4}$ downsampled resolutions ($112 \times 112$ and $56 \times 56$) demonstrate the method's robustness to low-resolution. Even in the presence of resolution discrepancy (between the source and the target image), our method generates naturally swapped results that accurately reflect the target image resolution. On the other hand, previous works [1], [9], [29] struggle to maintain the fidelity of the target image, as seen in Fig. 7 (a) and (b). [1] produces excessively high-resolution swapped results despite the low-resolution target images. Meanwhile, [9], [29] failed to naturally fuse attributes extracted from low-resolution target images with identity embeddings from high-resolution source images. As a result, unnatural swapped results are generated with artifacts around the contours of the face.

### 2) RESULTS ON OCCLUSIONS

We also evaluate our model on occluded settings where faces are occluded with facial masks or hands using our proposed occluded VGGFace2 datasets. Through the occlusion handling framework, our model is able to preserve the whole occlusion in the swapped results, as illustrated in Fig. 8. Moreover, our model is able to perform high-fidelity face swapping in the remaining unoccluded regions. On the other hand, previous works perform poorly and thus fail to preserve the occlusions. The comparison in Fig. 7 (c) and (d) shows that occlusions are only partially reconstructed, or facial features are partially visible through the occlusion. This experiment result emphasizes the difficulty of handling occlusions using the face swapping model and demonstrates that an explicit module is better suited for this purpose.

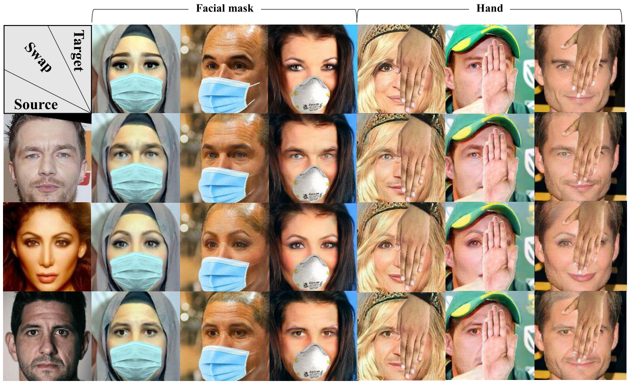

**FIGURE 8.** Face swapping results of our proposed method on the occluded VGGFace2 test set [20]. Our method successfully preserves the target face occlusions, producing diverse results.

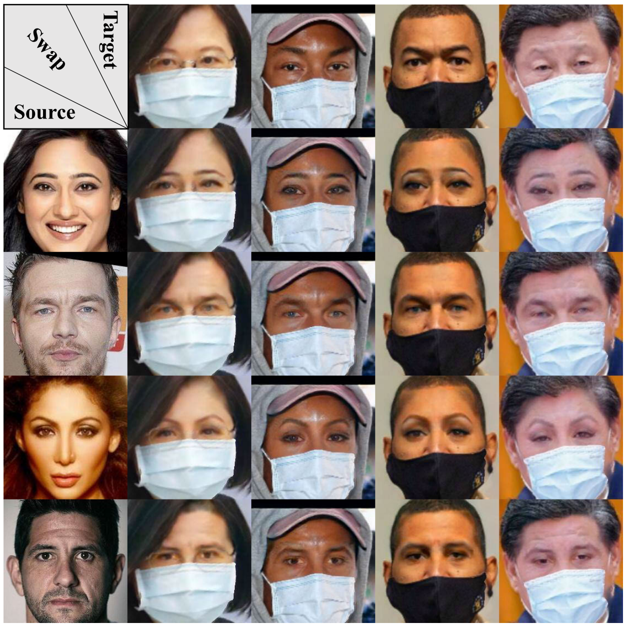

**FIGURE 9.** Face swapping results of our proposed method for real-world MFR2 dataset [25]. Our method works robustly for real-world images of facial masks.

### 3) RESULTS ON REAL-WORLD OCCLUSIONS

Finally, we evaluate the robustness of our method on real-world occlusions. For evaluation on real-world facial mask occlusions, we apply the model to the MFR2 [25] dataset,

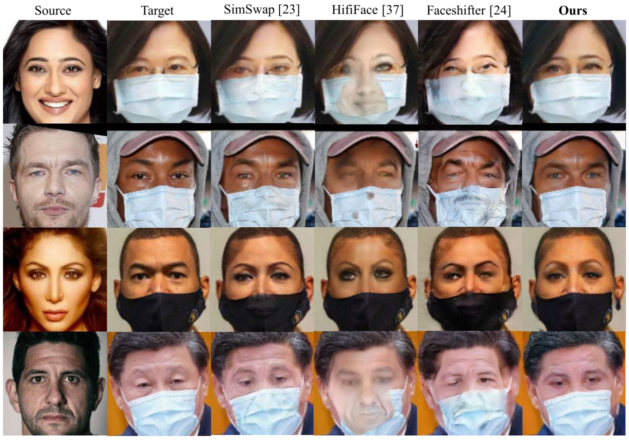

**FIGURE 10. Comparison against previous works [1], [9], [29] on real-world mask-occluded images. Facial mask both interfere with face swapping and are not preserved.**

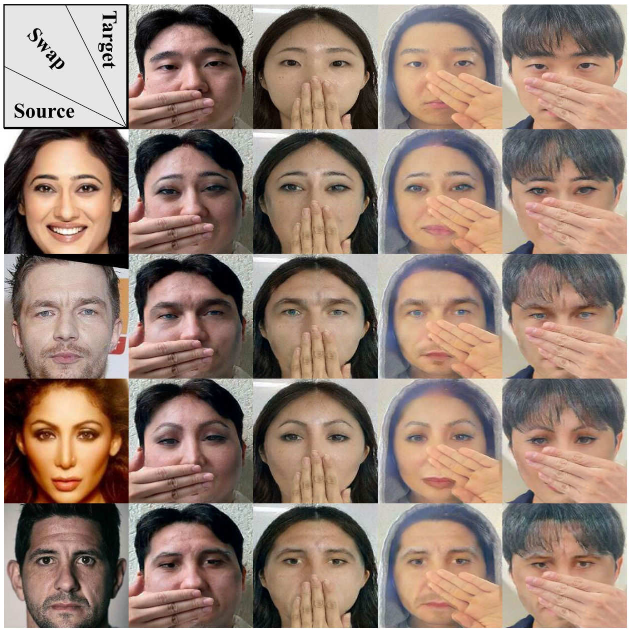

**FIGURE 11. Face swapping results of our proposed method on real-world hand-occluded images. The method successfully performs face swapping while preserving challenging hand occlusions.**

a dataset of 269 images consisting of real-world images of public figures wearing facial masks. As seen in Fig. 9, our model produces natural swapped results even for large real-world facial masks generally covering half of the target face. Moreover, in comparison to conventional face swapping models in Fig. 10, the entirety of the facial mask is well preserved both in terms of its shape and opacity. On the other hand, conventional models either fail to completely preserve the occlusions or perform natural face swapping.

As for hand occlusions, as there are no public datasets for face images with hand occlusions, we perform face swapping on our collected real-world dataset and display our results in Fig. 11. Although the occlusion parser sometimes yields inaccurate occlusion masks, our method shows satisfactory swapping performance while preserving hand occlusions for most cases. This is further underlined when compared to other face swapping models in Fig. 12, where blur artifacts around the face contours occur in [1] and [9] and the hand occlusions are not properly preserved in [1] and [29]. As real-world datasets more accurately reflect the nature of in-the-wild images, the evaluation on real-world datasets again demonstrates the effectiveness of our proposed method as a privacy protection measure against real-world occlusions.

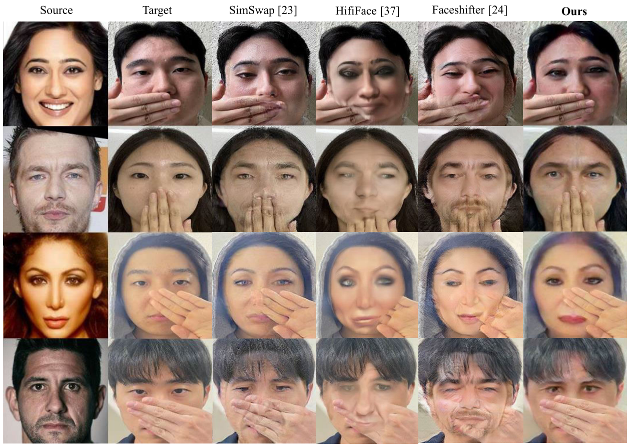

**FIGURE 12. Comparison against previous works [1], [9], [29] on real-world hand-occluded images. Conventional methods struggle to either preserve occlusion or successfully perform face swapping.**

### C. ABLATION STUDY

To validate the contribution of each module in our method, we conduct an ablation study on a combined occluded-VGGFace2 dataset that contains both facial masks and hand occlusions. As depicted in Table 7, we begin with Simswap [1] as our base model and progressively add our proposed modules to examine their effects. We report the averaged score for $\downarrow_{\times 1}, \downarrow_{\times 2}$ and $\downarrow_{\times 4}$ downsampled images and further conduct ablation experiments with sample progress images in Fig. 13 and an extensive experiment on the effect of CRCL in Table 8.

First, in Ablation 1 of Table 7, the addition of the occlusion parser $B$ alone does not improve the face swapping performance. In Fig. 13, it is shown that the output of the occlusion parser $B$ is an incomplete face image $x^*_{occ}$. Naturally, using an incomplete face image as the input to the face swapping model leads to decreased performance. However, in combination with the inpainting module $P$, $x^*_{occ}$ is inpainted to a complete face $x^*_p$ with complete facial features. As shown in Ablation 2, face swapping with a face image with complete facial features brings dramatic improvements to most metrics. However, the FID score remains high as the inpainted region generates unnatural color and contour artifacts, as shown in $x_{swap+occ}$.

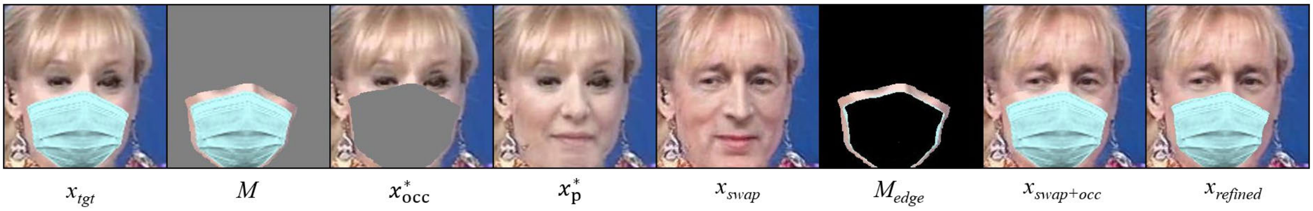

**FIGURE 13.** Face and mask images for each step of the occlusion handling framework. We extract the mask explicitly and transfer the extracted mask back onto the swapped face.

**TABLE 7.** Ablation study of the proposed method. We examine the effect of each module on occluded (facial mask + hand) VGGFace2 [20]. We measure the performance on x1, x2 and x4 downsampled resolutions and report its average.

| | Binary parser $B$ | Inpainting Module $P$ | Refinement Network $R$ | CRCL | ID Sim.↑ | ID ret.↑ | Pose↓ | Exp.↓ | FID↓ |
|---|---|---|---|---|---|---|---|---|---|
| **Simswap** [1] | | | | | 0.528 | 75.8 | 141.6 | 1.888 | 39.36 |
| Ablation 1 | ✓ | | | | 0.458 | 60.0 | 157.1 | 1.989 | 45.94 |
| Ablation 2 | ✓ | ✓ | | | 0.480 | 66.6 | 150.4 | 1.973 | 49.04 |
| Ablation 3 | ✓ | ✓ | ✓ | | 0.477 | 67.6 | 150.4 | 1.941 | 34.23 |
| **Ours**<br>(Ablation 4) | ✓ | ✓ | ✓ | ✓ | 0.650 | 84.6 | 142.9 | 1.847 | 26.00 |

**TABLE 8.** Ablation study on the effect of CRCL fine-tuning on VGGFace2 [20]. CRCL indicate the CRCL fine-tuned Arcface [12] identity embedder. Red indicate the best score.

| | $D$ | ID sim.↑ | ID ret.↑ | Pose↓ | Exp.↓ | FID↓ |
|---|---|---|---|---|---|---|
| **Arcface**<br>[12] | $\downarrow\times1$ | 0.827 | 94.2 | 118.2 | 1.713 | 20.92 |
| | $\downarrow\times2$ | 0.802 | 94.1 | 117.2 | 1.719 | 25.01 |
| | $\downarrow\times4$ | 0.735 | 91.1 | 117.7 | 1.714 | 31.44 |
| **Ours**<br>**(CRCL)** | $\downarrow\times1$ | 0.824 | 98.5 | 118.2 | 1.715 | 19.22 |
| | $\downarrow\times2$ | 0.820 | 98.0 | 117.4 | 1.712 | 23.16 |
| | $\downarrow\times4$ | 0.772 | 92.8 | 117.6 | 1.715 | 25.77 |

To address this limitation, the refinement module $R$ refines the final result $x_{swap+occ}$ to $x_{refined}$, removing major artifacts around the overlayed occlusion. This produces natural face images, yielding further improvements in the FID score as shown in Ablation 3. The lower FID score indicates a swapped face distribution closer to a natural face distribution, implying stronger similarity and improved naturalness. Moreover, a better ID retrieval score implies that the identity of the source face is well transferred even in the presence of a facial occlusion. This can also be examined visually in $x_{swap+occ}$ and $x_{refined}$ of Fig. 13. Without the refinement module, the blending of the explicit mask are unnatural with artifacts around the contours of the facial mask. The refinement module corrects the color and the lighting around the facial mask to make a natural-looking result.

Finally, Ablation 4 shows that leveraging a CRCL fine-tuned identity embedder results in a significant improvement in the ID Sim. and ID ret. scores due to the accurate identity embeddings extracted from cross-resolution embedding space. The effect of CRCL fine-tuning on the identity embedder can further be examined in Table 8. Compared to the original Arcface [12] identity embedder, our model's attribute preservation capability (represented by Pose and Exp. metric) remains unaffected. On the other hand, it shows vastly improved identity transfer capability, especially on low-resolutions. This verifies the accuracy of the fine-tuned identity embedder across various resolutions, which translates to the robust performance of the face swapping model on faces in-the-wild. This extensive ablation study shows that individual modules improve certain aspects of face swapping with definite trade-offs. However, when all the modules are combined into a single framework, it forms a robust model for in-the-wild images.

## VI. CONCLUSION

We have proposed a new method for swapping faces to protect privacy, especially for low-resolution and occluded faces commonly found in real-world videos. Our approach uses innovative techniques such as CRCL for low-resolution faces and a robust occlusion-handling framework to provide better privacy protection while preserving image quality. We also created the occluded-VGGFace2 dataset to assess our method on synthetic occlusions, which includes face images overlaid with synthetic facial masks and hand occlusions. As far as we know, this is the first work to customize face swapping techniques to address privacy issues. Through comprehensive experiments, we have validated our framework's superiority in both synthetic and real-world in-the-wild scenarios. By overcoming the difficulties posed by occlusions and low-resolution images, our framework successfully preserved the target image resolution and occlusion, yielding natural face-swapped results. Specifically, our method achieves impressive retrieval rates of 92.8%, 92.2%, and 82.3% in low-resolution scenarios across normal, masked, and hand-occluded conditions, respectively. These results represent a substantial improvement over previous works, highlighting the effectiveness of our framework in real-world privacy protection scenarios. Our method not only outperforms existing solutions by a large margin but also ensures reliable privacy protection even in adverse conditions, making it a highly effective tool for safeguarding identities in various practical applications. By demonstrating such high performance, our framework sets a new standard for face swapping techniques in privacy protection, providing a viable alternative to unlicensed identities through artificially generated replacements. Furthermore, we expect this work to pave the way for the safe release of public datasets and to encourage further research and development in this field.

## REFERENCES

[1] 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.

[2] I. Korshunova, W. Shi, J. Dambre, and L. Theis, "Fast face-swap using convolutional neural networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 3697–3705.

[3] R. Natsume, T. Yatagawa, and S. Morishima, "RSGAN: Face swapping and editing using face and hair representation in latent spaces," 2018, *arXiv:1804.03447*.

[4] Y.-T. Cheng, V. Tzeng, Y. Liang, C.-C. Wang, B.-Y. Chen, Y.-Y. Chuang, and M. Ouhyoung, "3D-model-based face replacement in video," in *Proc. SIGGRAPH*, Aug. 2009, p. 1.

[5] 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.

[6] Y. Nirkin, I. Masi, A. Tran Tuan, T. Hassner, and G. Medioni, "On face segmentation, face swapping, and face perception," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, May 2018, pp. 98–105.

[7] J. Thies, M. Zollhöfer, M. Stamminger, C. Theobalt, and M. Nießner, "Face2Face: Real-time face capture and reenactment of RGB videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2387–2395.

[8] 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.

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

[10] 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.

[11] C. Lin, P. Hu, C. Shen, and Q. Li, "End-to-end face-swapping via adaptive latent representation learning," 2023, *arXiv:2303.04186*.

[12] 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.

[13] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8107–8116.

[14] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," 2017, *arXiv:1710.10196*.

[15] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4396–4405.

[16] J. Kim, J. Lee, and B.-T. Zhang, "Smooth-swap: A simple enhancement for face-swapping with smoothness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 10769–10778.

[17] 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.

[18] Z. Liu, M. Li, Y. Zhang, C. Wang, Q. Zhang, J. Wang, and Y. Nie, "Fine-grained face swapping via regional GAN inversion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8578–8587.

[19] Y. Guo, L. Zhang, Y. Hu, X. He, and J. Gao, "MS-Celeb-1M: A dataset and benchmark for large-scale face recognition," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 87–102.

[20] 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.

[21] 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.

[22] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong, J. Zhou, Z. Li, and W. Liu, "CosFace: Large margin cosine loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5265–5274.

[23] N. Ruiz, E. Chong, and J. M. Rehg, "Fine-grained head pose estimation without keypoints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 2074–2083.

[24] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky, "Resolution-robust large mask inpainting with Fourier convolutions," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2022, pp. 3172–3182.

[25] A. Anwar and A. Raychowdhury, "Masked face recognition for secure authentication," 2020, *arXiv:2008.11104*.

[26] K. Shiohara, X. Yang, and T. Taketomi, "BlendFace: Re-designing identity encoders for face-swapping," 2023, *arXiv:2307.10854*.

[27] M. Afifi, "11K hands: Gender recognition and biometric identification using a large dataset of hand images," *Multimedia Tools Appl.*, vol. 78, no. 15, pp. 20835–20854, Aug. 2019, doi: 10.1007/s11042-019-7424-8.

[28] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "GANs trained by a two time-scale update rule converge to a local Nash equilibrium," in *Proc. Adv. In Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5–6.

[29] 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*.

[30] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[31] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. 14th Eur. Conf. Comput. Vis.*, Amsterdam, The Netherlands, Oct. 2016, pp. 694–711.

[32] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[33] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. 18th Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, vol. 9351, Munich, Germany, Oct. 2015, pp. 234–241.

[34] B. Meden, M. Gonzalez-Hernandez, P. Peer, and V. Štruc, "Face deidentification with controllable privacy protection," *Image Vis. Comput.*, vol. 134, Jun. 2023, Art. no. 104678.

[35] S. Yang, W. Wang, Y. Cheng, and J. Dong, "A systematical solution for face de-identification," in *Proc. 15th Chin. Conf. Biometric Recognit. (CCBR)*, Shanghai, China, Sep. 2021, pp. 20–30.

[36] J. Cao, B. Liu, Y. Wen, R. Xie, and L. Song, "Personalized and invertible face de-identification by disentangled identity information manipulation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3314–3322.

[37] H. Li, Y. Li, J. Liu, Z. Hong, T. Hu, and Y. Ren, "Zero-shot face swapping with de-identification adversarial learning," in *Proc. Int. Conf. Parallel Distrib. Comput., Appl. Technol.*, 2021, pp. 101–112.

[38] A. Maity, R. More, G. Kambli, and S. Ambadekar, "Preserving privacy in video analytics: A comprehensive review of face de-identification and background blurring techniques," *TechRxiv*, pp. 8–14, Nov. 2023.

[39] L. Chi, B. Jiang, and Y. Mu, "Fast Fourier convolution," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 4479–4488.

[40] G. Gao, H. Huang, C. Fu, Z. Li, and R. He, "Information bottleneck disentanglement for identity swapping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 3403–3412.

[41] D. Jiang, D. Song, R. Tong, and M. Tang, ‘‘StyleIPSB: Identity-preserving semantic basis of StyleGAN for high fidelity face swapping,’’ in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 352–361.

[42] 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.

[43] Y. Wu, F. Yang, and H. Ling, ‘‘Privacy-Protective-GAN for face de-identification,’’ 2018, *arXiv:1806.08906*.

**JAEHYUN PARK** (Graduate Student Member, IEEE) received the B.S. degree in electrical engineering from Seoul National University, Seoul, South Korea, in 2021, where he is currently pursuing the Ph.D. degree with the Interdisciplinary Program in Artificial Intelligence. His research interests include image processing, computer vision, and machine learning.

**WONJUN KANG** (Student Member, IEEE) received the B.S. degree in electrical and computer engineering from Seoul National University, Seoul, South Korea, in 2020, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interests include image processing, computer vision, and machine learning.

**HYUNG IL KOO** (Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the Department of Electrical Engineering and Compute Science, Seoul National University, Seoul, South Korea, in 2002, 2004, and 2010, respectively. From 2010 to 2012, he was a Research Engineer with Qualcomm Research Korea. He joined the Department of Electrical and Computer Engineering, Ajou University, in 2012, where he is currently an Associate Professor. His research interests include computer vision and machine learning.

**NAM IK CHO** (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in control and instrumentation engineering from Seoul National University, Seoul, South Korea, in 1986, 1988, and 1992, respectively. From 1991 to 1993, he was a Research Associate with the Engineering Research Center for Advanced Control and Instrumentation, Seoul National University. From 1994 to 1998, he was with the University of Seoul, Seoul, as an Assistant Professor of electrical engineering. He joined the Department of Electrical and Computer Engineering, Seoul National University, in 1999, where he is currently a Professor. His research interests include image processing, adaptive filtering, and computer vision. He is an Associate Editor of IEEE Transactions on Image Processing and a Handling Editor of *Signal Processing*.