IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 26 May 2025, accepted 11 June 2025, date of publication 26 June 2025, date of current version 10 July 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3583388

RESEARCH ARTICLE

Secure and Reversible Face De-Identification With Format-Preserving Encryption

HEEHWAN KIM[1], SUNGJUNE PARK[2], AND DAESEON CHOI[2], (Member, IEEE)
[1]Department of AI Security Convergence, Soongsil University, Seoul 07027, South Korea
[2]Department of Software, Soongsil University, Seoul 07027, South Korea

Corresponding author: Daeseon Choi (sunchoi@ssu.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant through the Korean Government [Ministry of Science and ICT (MSIT)] (Robust AI and Distributed Attack Detection for Edge AI Security) under Grant 2021-0-00511, and in part by IITP Grant funded by Korea Government (MSIT) (Development of Countermeasure Technologies for Generative AI Security Threats) under Grant RS-2024-00398353.

ABSTRACT With the rapid growth of digital services and transactions, concerns about personal information leakage and security threats have intensified. In particular, facial images stored in surveillance systems, security agency databases, and biometric authentication platforms can be exploited for identity theft, fraud, phishing, illegal marketing, or deepfake-based misinformation. These risks have increased the need for technologies that securely protect stored facial data while enabling authorized restoration when necessary. We propose a reversible face de-identification method using format-preserving encryption (FPE). Our method integrates symmetric-key FPE into two deep neural network (DNN)-based face swap models (FaceShifter and SimSwap) that separate identity and attribute information. This approach enables the generation of de-identified facial images and allows only authorized users to restore the original data. Experiments on the LFW, FFHQ, VGGFace2-HQ, and CelebA-HQ datasets showed that, on average, the FaceShifter model achieved a 98.64% de-identification success rate and 96.86% restoration rate, while SimSwap recorded 99.58% and 99.39%, respectively. Image quality was evaluated using SSIM, FID, LPIPS, PSNR, and BRISQUE, confirming that restored images closely resemble the originals. In conclusion, the proposed method provides a robust privacy-preserving solution for facial data in digital environments, balancing security and utility while supporting lawful restoration.

INDEX TERMS Artificial intelligence (AI), face de-identification, face privacy, format-preserving encryption (FPE), privacy protection.

I. INTRODUCTION

In recent years, advancements in computing hardware and software have significantly improved the efficiency and convenience of information processing across various digital environments. In particular, the rapid development of Deep Neural Networks (DNNs) and computer vision technologies has enabled the large-scale collection, analysis, modification, and feature extraction of various types of personal data—especially biometric data—which has raised growing concerns about data privacy and security.

Among various types of biometric data, facial images are especially vulnerable to misuse. When stored in surveillance systems, law enforcement databases, or biometric authentication platforms, these images, if leaked, can be exploited for a range of malicious purposes such as identity theft, financial fraud, phishing, unauthorized marketing, and the generation of misleading content through deepfake technologies [1]. These risks have intensified public and regulatory scrutiny of facial data protection, resulting in international data privacy frameworks such as the European Union's General Data Protection Regulation (GDPR) [2].

To mitigate such threats, facial de-identification techniques have emerged as a technical countermeasure. Traditional approaches, including noise injection, pixelation,

The associate editor coordinating the review of this manuscript and approving it for publication was Chien-Ming Chen.

© 2025 The Authors. This work is licensed under a Creative Commons Attribution 4.0 License.
For more information, see https://creativecommons.org/licenses/by/4.0/

and blurring, attempt to conceal identifiable facial features. However, these methods often degrade image quality and utility, and are increasingly susceptible to reversal by modern DNN-based face recognition models [3]. As a result, more advanced de-identification methods have been developed using deep learning, offering improved image realism while obscuring identity. Nevertheless, a critical drawback remains: most of these methods are irreversible by design, making it impossible to restore the original identity even under legally justified or user-authorized conditions [4].

To address this limitation, several reversible de-identification approaches have been explored [5], [6], [7], [8]. These methods typically use passwords or key-based identity encoding to allow restoration of the original face. While these solutions introduce the possibility of lawful recovery, they often rely on short or weak keys that are vulnerable to brute-force or dictionary attacks. Moreover, the lack of structural robustness in such password-based designs makes them difficult to deploy securely at scale in real-world systems.

In response, we propose a novel facial de-identification framework that incorporates symmetric-key format-preserving encryption (FPE) [9] into the identity representation process. FPE enables the encryption of the facial identity vector while preserving its numerical structure and dimensionality, ensuring compatibility with existing DNN architectures. This approach strengthens privacy protection by preventing unauthorized recovery, while also allowing legitimate restoration when the encryption key is available.

Our method is applied to two representative face-swapping networks: the Adaptive Embedding Integration Network (AEI-NET) of FaceShifter [10], and the ID-injection-based generator of SimSwap [11]. Both networks utilize ArcFace [12] as an identity encoder and are designed to separate identity from non-identity attributes such as pose, lighting, and expression. By encrypting the identity embedding vector using FPE and integrating it into these architectures, we enable high-quality, reversible facial de-identification that supports secure anonymization and accurate restoration within the same framework.

The main contributions of this study are as follows:

1) We propose a reversible facial de-identification method that employs format-preserving encryption (FPE) to ensure both strong security and compatibility with deep learning-based face-swapping architectures.
2) We conduct extensive experiments demonstrating that our method achieves high success rates in both de-identification and restoration, while preserving image quality comparable to the original.
3) We demonstrate the potential of our approach to protect personal information in digital services that utilize facial data, while maintaining data utility and practicality.

II. RELATED WORKS

A. TRADITIONAL FACIAL DE-IDENTIFICATION TECHNIQUES

Early research on facial de-identification primarily employed traditional methods such as noise addition, JPEG compression, pixelation, and blurring [13], [14], [15], [16]. Noise addition [13] distorts the original images by introducing random noise, with Gaussian noise being a common method. This involves adding random pixel values following a Gaussian distribution, which blurs the features of the original face and makes identification difficult. However, excessive noise can significantly degrade image quality, limiting usability.

JPEG compression [14] uses a lossy compression method to reduce image file sizes, adjusting compression levels to balance between facial de-identification and image quality. Pixelation [15] converts images into a low-resolution format where facial features are represented by large pixel blocks, obscuring details and hindering identification. While these methods offer quick and effective de-identification, they, like noise addition, compromise image utility by obscuring too much detail.

Blurring [16], particularly Gaussian blur, applies a smoothing effect to facial images, which blurs key features and reduces sharpness. Although this can make faces more difficult to identify, weak blurring may still allow for identification, while overly strong blurring degrades image utility, similar to the limitations of the previously mentioned methods. These traditional methods are useful for privacy-focused research but offer limited data utility owing to their overt anonymization and the ease of recognizing these anonymization efforts [17].

B. AI-BASED FACIAL DE-IDENTIFICATION TECHNIQUES

Advancements in artificial intelligence (AI) technologies, particularly deep learning methods, have revolutionized facial de-identification, balancing privacy protection with data utility. Generative adversarial networks (GANs) [18], with their adversarial learning framework of a generator and a discriminator, have enabled the creation of highly realistic and natural images, offering innovative approaches to facial de-identification. When applied to de-identification, GANs can transform original facial images to resemble those of other people or create transformations that are unrecognizable to machines but identifiable by humans.

Recent studies have explored the potential to align privacy protection with technical utility. For example, Hukkelas et al. [19] introduced DeepPrivacy, a GAN-based approach that generates entirely new facial regions while maintaining background details using the U-Net architecture [20] to produce high-quality de-identified images. This method detects facial landmarks, such as the eyes, nose, and mouth, using GANs to synthesize a new face while preserving image quality.

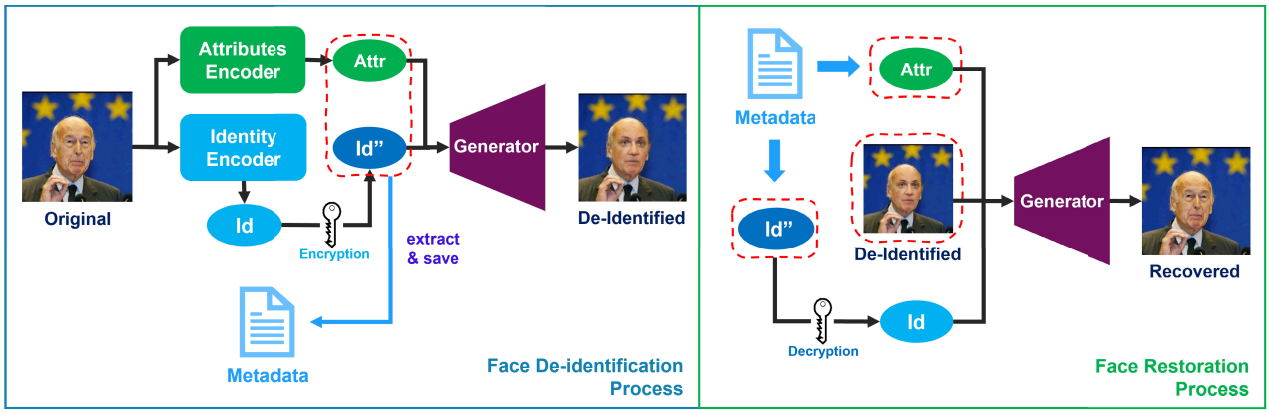


FIGURE 1. Framework of face de-identification process and restoration process.

Maximov et al. [21] proposed conditional identity anonymization generative adversarial networks (CIAGAN), a method similar to DeepPrivacy, which uses landmark detection to remove identifying features while retaining necessary details for facial and body detection. This approach produces high-quality de-identified images suitable for various applications.

Gafni et al. [22] introduced Live Face, a method capable of real-time video de-identification by separating facial identification information from attributes such as pose, lighting, and expression. This method generates new faces while preserving specific facial attributes and does not require retraining. Similarly, attribute-aware anonymization networks for face de-identification (A^3GAN) [23], proposed by Y. Liu et al., generates de-identified facial images by preserving attributes while removing personally identifiable information, balancing attribute retention with de-identification tasks.

Progressive growing of GANs (PGGAN) [24], introduced by Tero Karras et al., offers a novel de-identification method where image complexity and resolution grow progressively during training, resulting in high-resolution de-identified images. The method starts with low-resolution training, progressively doubling the resolution as the network stabilizes, and introduces blending stages to smooth the transition between different resolutions.

These studies have enhanced privacy protection by safeguarding facial identification information while maintaining data utility for applications such as facial recognition. However, these methods often fail to preserve personal attributes and identifiable features, making re-identification impossible and hindering the restoration of the original image when necessary.

C. RESTORABLE FACIAL DE-IDENTIFICATION TECHNIQUES

With the growing demand for privacy protection and the technological need to restore original images, several reversible facial de-identification methods have been proposed. Among them, Cao et al. [5] introduced a method that uses a six-digit password and a privacy protection parameter d to transform facial identity features for de-identification and restoration. This method allows for the restoration of the original face when the correct password and d value are provided, while an incorrect input generates a high-quality face image with a different identity.

Li et al. [6] proposed a method that utilizes randomly generated passwords to encrypt facial identity within the latent space of StyleGAN2 [25], thereby generating de-identified facial images. Xiao et al. [7] suggested a method that separates identity-related features from non-identity-related features in facial images and manipulates only the identity features through passwords to generate de-identified images in various ways. Both approaches, including that of Cao et al. [5], share the common characteristic that the original facial image can only be restored with the correct password.

IdentityMask, proposed by Wen et al. [8], focuses on analyzing motion flows between consecutive frames in facial videos, using a unique identifier, Ukey or password, to enhance the de-identification and restoration process. This method de-identifies the first frame and consistently applies de-identification across subsequent frames according to the motion flow module's guidance. The same Ukey is used during restoration to recover the original video, ensuring consistency across the entire video and allowing only authorized users to restore the original content.

These methods aim to remove personally identifiable information from facial images and videos while maintaining their utility for facial recognition tasks. They also seek to restore the original image as closely as possible when necessary. However, these methods are vulnerable to brute force and dictionary attacks due to their reliance on simple numerical or string-based passwords for de-identification. This vulnerability poses the risk of reverse-engineering the de-identified information or easily restoring the original content.

III. METHODS

The overall structure of our proposed method is illustrated in Figure 1. It consists of de-identification and

restoration stages, with encryption embedded within the de-identification process and decryption performed as an integral part of restoration.

A. FACIAL DE-IDENTIFICATION NETWORKS

For facial recognition, de-identification, and network training, we employ the FaceShifter model proposed by Li et al. [10] and the SimSwap model proposed by Chen et al. [11]. Unlike prior de-identification models [19], [21], [22], [23], [24], which generate random faces using a single input image, these models take both the source image X_S and the target image X_t as inputs to a de-identification network. These two models are particularly effective because they separate identity information from attribute features, allowing for more flexible and high-fidelity facial de-identification. FaceShifter is suitable for high-resolution face reconstruction and precise attribute handling, while SimSwap provides a generalized framework for arbitrary identity transformation.

1) FACESHIFTER

FaceShifter is a dual-stage framework designed for high-resolution face swapping with strong identity modification and attribute preservation. It separately extracts the identity features from the source image and multi-level attributes from the target image, and then merges them to generate a de-identified image. The face identity of the source image is first extracted using an ArcFace encoder:

$$z_{id} = ArcFace(X_s) \tag{1}$$

Then, facial attributes such as expression, pose, and lighting are extracted from the target image using a Multi-Level Attribute (MLA) encoder based on the U-Net architecture:

$$MLA(X_t) = z_{att} \tag{2}$$

$$z_{att} = \{z_{att}^1, z_{att}^2, \ldots, z_{att}^n\} \tag{3}$$

where z_{att}^k represents the feature map extracted from the k-th level of the U-Net, and n denotes the total number of feature levels. These features are combined by the AAD generator G, along with the source image X_s, to produce the de-identified image:

$$G(X_s, z_{att}, z_{id}) = X_{deid} \tag{4}$$

This structure enables the generation of images that are visually similar to the original while altering only the identity.

2) SIMSWAP

SimSwap is a high-fidelity face swapping framework that generates de-identified images by combining the identity of the source image X_s with the attributes of the target image X_t. While SimSwap uses the notation f_{id} for the identity vector, we adopt z_{id} for consistency in this paper. In this network, the identity vector is extracted using a pretrained ArcFace model:

$$z_{id} = ArcFace(X_s) \tag{5}$$

The target image is encoded into a feature map F_t containing implicit attribute information:

$$F_t = Encoder(X_t) \tag{6}$$

The identity vector z_{id} is then injected into the target feature map, and the generator G takes the source image X_s, the modified target feature map, and the identity vector to reconstruct the de-identified image:

$$G(X_s, F_t, z_{id}) = X_{deid} \tag{7}$$

Although SimSwap does not explicitly extract attribute vectors, it effectively preserves non-identity features through feature-level identity injection and the use of Weak Feature Matching Loss. This architecture enables robust and generalized de-identification while maintaining visual fidelity across various identity-attribute combinations.

B. FORMAT-PRESERVING ENCRYPTION

FPE is an encryption method that retains the format of the original data even after encryption. It shares similarities with symmetric key encryption, as both use a single encryption key; however, FPE differs from conventional methods, such as the Advanced Encryption Standard (AES) [26], by preserving the data structure. For example, when encrypting numeric data such as "21312345678," the result is also a numeric string of the same length, such as "49538274192". This makes FPE particularly useful for protecting information such as social security numbers and credit card numbers, which must retain their format.

In this study, we employ the FF3 (Feistel Finite Field) algorithm [27], a popular FPE method. FF3 is based on AES-128, providing a 128-bit key space that ensures sufficient cryptographic strength for secure identity encoding in our framework. The FF3 algorithm is based on the Feistel network and follows a sequence of steps involving input splitting, application of a round function, and partial recombination. Initially, the input numeric string is split into two parts, followed by a string concatenation operation, expressed as

$$X = L||R \tag{8}$$

where X is the numeric string, L and R represent the left and right parts of the string, respectively, and || denotes the string concatenation operator. In the next step, the round function is applied. For each round, the Feistel network operates as follows:

$$L_i = R_i, R_i = L_{i-1} \oplus F(R_{i-1}, K_i) \tag{9}$$

Here, F represents the round function, K_i is the round key, $\oplus$ denotes the XOR operation, and Li and R_i correspond to the left and right halves at the ith round, respectively. After r rounds, the final ciphertext is formed by concatenating the two parts, represented as:

$$X' = L_r||R_r \tag{10}$$

Consequently, the encryption and decryption processes can be simplified to

$$FF3_{encrypt}(X, key) = X', FF3_{decrypt}(X', key) = X \quad (11)$$

In these formulas, the symmetric key *key* is used in both the encryption and decryption processes, serving as an input to the round function F to determine the output. Even with the same input, different keys produce different encryption and decryption results. Through this process, the FF3 algorithm ensures that the format of the input data is preserved, resulting in the final encrypted output, X'.

C. SECURE AND REVERSIBLE FACE DE-IDENTIFICATION WITH FORMAT-PRESERVING ENCRYPTION

1) DE-IDENTIFICATION PROCESS

In this research, we propose a novel reversible de-identification method by integrating format-preserving encryption into a de-identification network. For both de-identification and restoration, we apply our method to two representative de-identification networks: AEI-NET from FaceShifter and the ID-injection-based generator from SimSwap. Unlike previous approaches that generate random faces from a single input image, this approach uses the same source image X_s and target image X_t to extract face identity and attributes for encryption and de-identification. Figure 1 illustrates how the face identity z_{id} is extracted from the input image. The process generates a de-identified image with a new identity while preserving non-identifiable features from the original image.

We extracted the face identity information z_{id} from the source image using an encoder. The vector z_{id} contains values between -0.1 and 0.1, representing normalized feature points from the source image. The dimensionality of the vector is $(1, 512)$.

To prepare these values for encryption with the FF3 algorithm, we define a digit slicing function that converts each floating-point element into a 6-digit string based on its decimal portion. Specifically, ExtDigits(z_{id}) refers to the element-wise application of the slicing operation to each component of the identity vector z_{id}. That is, for each element $z_{id}^{(i)} \in z_{id}$, we extract the six digits starting from the second digit after the decimal point to the seventh digit (i.e., positions 2 through 7) of each $|z_{id}^{(i)}|$. For example, if $z_{id}^{(i)} = 0.0234589$, then ExtDigits($z_{id}^{(i)}$) = ''234589''. This range was chosen because float32 encoding preserves approximately seven decimal digits, and FF3 requires a minimum 6-digit input string.

Each digit string is encrypted using the FF3 algorithm with a pre-defined key *key*, producing a transformed 6-digit string. For example, when using a specific key, the string ''234589'' is deterministically mapped to a format-preserving ciphertext such as ''594932''. This ciphertext is then interpreted as an integer and scaled by 10^{-7} to obtain the floating-point value 0.0594932.

However, early experiments revealed that applying FF3 encryption and scaling alone was often insufficient to produce faces that were visually distinct from the originals. To address this, we introduce a sign inversion step. By multiplying the encrypted values by -1, we significantly shift the identity vector in the embedding space. This modification yields outputs that are visually and algorithmically different from the original identity, enhancing the effectiveness of de-identification.

This process is applied to every element of z_{id} to generate the encrypted identity vector z'_{id}. That is, for each $z_{id}^{(i)} \in z_{id}$:

$$z_{id}^{\prime(i)} = (-1) \times FF3_{encrypt}(\text{ExtDigits}(z_{id}^{(i)}), \text{key}) \times 10^{-7} \quad (12)$$

This process ensures that the essential features of the original facial image are preserved while effectively anonymizing identity in a cryptographically sound and visually effective manner. The encrypted identity vector z'_{id} is then used to generate the de-identified image, and its inverse transformation enables restoration.

The de-identified image is produced by combining the encrypted identity vector z'_{id} and the attribute vector z_{att} as inputs to the generator, along with the original source image X_s. The complete process is described as follows:

$$G(X_s, z_{att}, z'_{id}) = X_{deid} \quad (13)$$

The full de-identification workflow is summarized in the following pseudocode. This step-by-step outline demonstrates how the identity vector is extracted, encrypted using the FF3 algorithm, and then combined with attribute vectors to synthesize the de-identified image. Additionally, metadata required for restoration is securely stored alongside the output. The detailed algorithm is presented below.

Algorithm 1 De-Identification Algorithm

Require: X_s (source image), X_t (target image), *key*
Ensure: X_{deid} (de-identified image), metadata: $[z'_{id}, z_{att}]$
  $z_{id} \leftarrow ArcFace(X_s)$
  $z_{att} \leftarrow MLA(X_t)$ or $Encoder(X_t)$
  for each $z_{id}^{(i)}$ in z_{id} **do**
    $str \leftarrow ExtDigits(z_{id}^{(i)})$
    $enc \leftarrow FF3_{encrypt}(str, key)$
    $z_{id}^{\prime(i)} \leftarrow (-1) \times enc \times 10^{-7}$
  end for
  $X_{deid} \leftarrow G(X_s, z_{att}, z'_{id})$
  /* Store metadata for restoration */
  Store z'_{id} and z_{att} securely as metadata
  return X_{deid}, metadata $[z'_{id}, z_{att}]$

This approach results in a final de-identified image that not only encrypts and protects the identity information but also preserves the visual attributes of the original image. This provides secure, natural-looking de-identified

images suitable for privacy-sensitive applications, such as surveillance or biometric data sharing.

2) RESTORATION PROCESS

The restoration process follows a similar procedure to de-identification, as illustrated in Figure 1. During restoration, the encrypted facial identity vector z'_{id}, which was stored during de-identification, is decrypted element-wise using the FF3 decryption process outlined in equation 11. Specifically, for each encrypted component $z'^{(i)}_{id}$, the restored identity element is computed as:

$$z''^{(i)}_{id} = (-1) \times FF3_{decrypt}(ExtDigits(z'^{(i)}_{id}), key) \times 10^{-7} \tag{14}$$

The *ExtDigits* function is applied similarly as in the de-identification process, extracting relevant decimal digits from each component. This element-wise decryption ensures that the restored identity vector z''_{id} closely approximates the original identity vector z_{id}, thereby ensuring accurate recovery of the facial identity. The restored identity vector z''_{id} is then combined with the attribute vector z_{att} and the de-identified image X_{deid} as inputs to the generator, reconstructing the final restored image. This method preserves the original facial identity while maintaining critical identity attributes from the source. The entire restoration can be summarized as:

$$G(X_{deid}, z_{att}, z''_{id}) = X_{restored} \quad \text{where} \quad X_{restored} \approx X \tag{15}$$

Lastly, the full restoration procedure is expressed in the following pseudocode.

Algorithm 2 Restoration Algorithm

Require: X_{deid} (de-identified image), metadata $[z'_{id}, z_{att}]$, key
Ensure: $X_{restored}$ (restored image)
  Load metadata for restoration */
   Retrieve encrypted identity vector z'_{id} and attributes z_{att} from metadata
  for each $z'^{(i)}_{id}$ in z'_{id} **do**
    $str \leftarrow ExtDigits(z'^{(i)}_{id})$
    $dec \leftarrow FF3_{decrypt}(str, key)$
    $z''^{(i)}_{id} \leftarrow (-1) \times dec \times 10^{-7}$
  end for
  $X_{restored} \leftarrow G(X_{\text{deid}}, z_{att}, z''_{id})$
  return $X_{restored}$

Through this approach, our method offers enhanced security compared to traditional de-identification and restoration techniques by incorporating FPE. This enables secure encryption and decryption, providing a higher level of privacy protection while ensuring that the original facial identity can be restored when necessary for identification purposes, thereby significantly strengthening overall security.

IV. EXPERIMENT

A. EXPERIMENT SETTINGS

In this study, we evaluate the proposed facial de-identification method, which integrates FPE into two representative de-identification networks: AEI-NET of the FaceShifter model and the ID-injection-based generator of SimSwap. The same symmetric key was used for all de-identification and restoration experiments. For training two de-identification networks of the FaceShifter and SimSwap model, as well as for experiments involving facial image de-identification, restoration, and recognition, we utilized three datasets: Labeled Faces in the Wild (LFW) [28], FlickrFaces-HQ (FFHQ) [29], VGGFace2-HQ [30], and CelebA-HQ [24], [31].

The LFW dataset contains a wide variety of facial images captured under real-world conditions and is widely used to evaluate the performance of facial recognition algorithms. The FFHQ dataset provides high-resolution facial images with extensive variability across races, ages, and genders. The VGGFace2-HQ dataset includes facial images of over 9,000 individuals, offering high-resolution images captured under different conditions, such as varying angles, lighting, and expressions. Lastly, the CelebA-HQ dataset contains 30,000 high-quality facial images generated from the original CelebA dataset, offering improved resolution and visual fidelity.

The specifications of each dataset used in this study are summarized as follows. The LFW dataset contains approximately 13,000 facial images at a resolution of 250×250. The FFHQ dataset includes around 70,000 high-quality facial images with a resolution of 1024×1024, and the VGGFace2-HQ dataset contains over 1.1 million facial images at a resolution of 512×512. The CelebA-HQ dataset includes 30,000 high-quality facial images at a resolution of 1024×1024, derived from the original CelebA dataset with enhanced visual quality and diversity. These datasets capture various facial details under diverse conditions, making LFW and FFHQ particularly suitable for evaluating the real-world performance of de-identification and restoration algorithms.

For training the de-identification network, the original image datasets were used. Whereas, in the de-identification and restoration experiments, 13,000 high-quality facial images were randomly selected from each dataset, and all images were resized to a resolution of 256×256 to enhance performance. For facial verification, we employed the ArcFace [12], FaceNet512 [32], and VGG-Face [33] models. To ensure reliability and accuracy, we set threshold values for determining whether two faces belonged to the same person at 0.68 for ArcFace, 0.3 for FaceNet512, and 0.4 for VGG-Face. These models were used to compare the original images with the de-identified and restored images, evaluating facial similarity throughout the de-identification and restoration processes. For facial recognition, we utilized YOLOv8 [34], known for its high accuracy in real-time object detection, particularly in quickly and accurately detecting

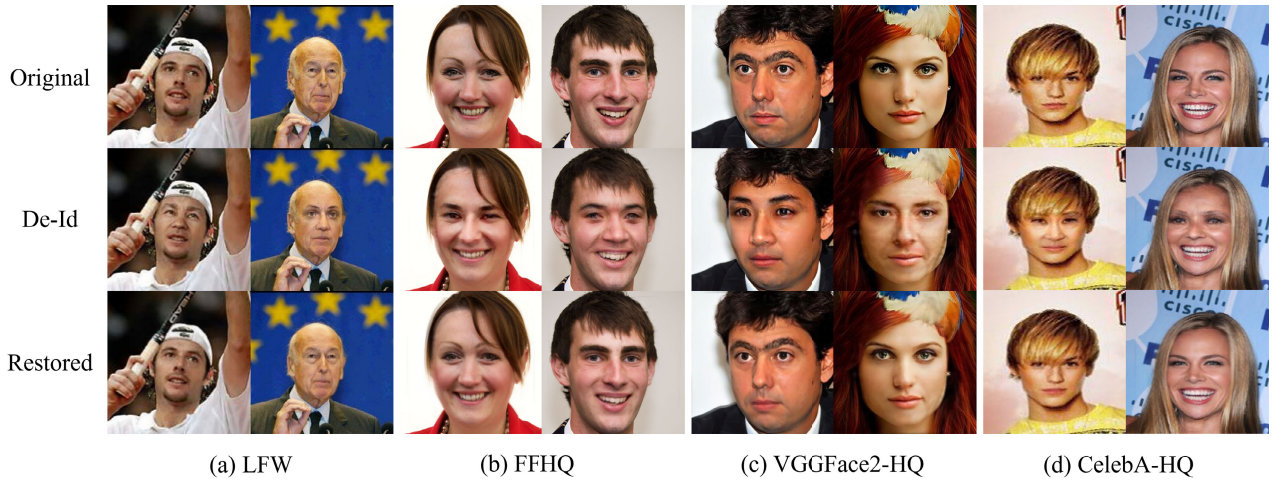


FIGURE 2. Visual results of the de-identification and restoration process using FaceShifter across four datasets: (a) LFW, (b) FFHQ, (c) VGGFace2-HQ, and (d) CelebA-HQ. Each row shows: original, de-identified, and restored images.

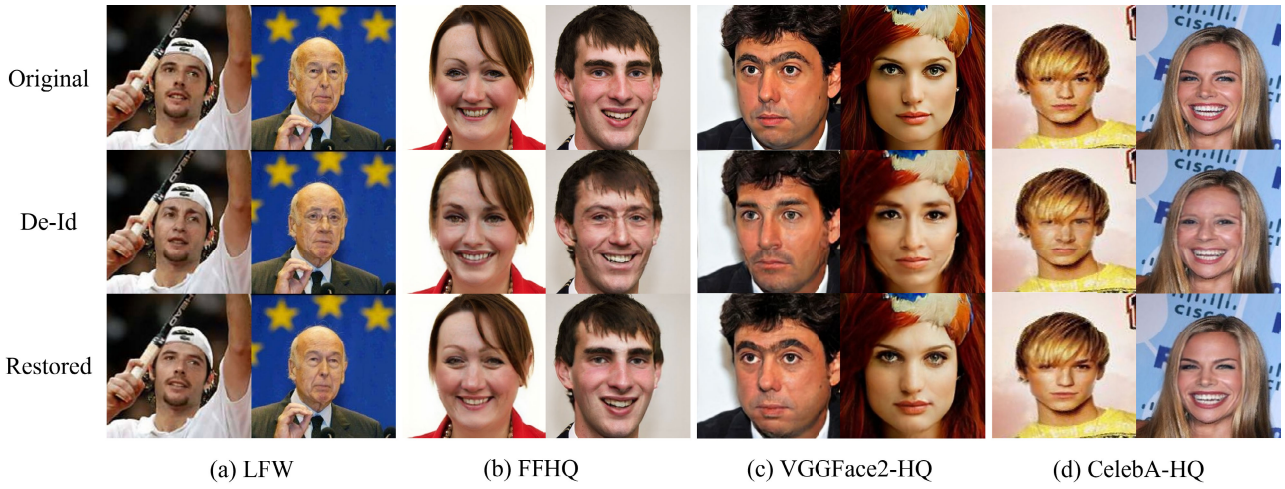


FIGURE 3. Visual results of the de-identification and restoration process using SimSwap. The datasets and image arrangement are consistent with those shown in Figure 2.

facial regions. In our experiments, YOLOv8 played a key role in detecting faces, assisting the aforementioned facial verification models.

For quality assessment, we used structural similarity index (SSIM) [35], blind/referenceless image spatial quality evaluator (BRISQUE) [36], Frechet inception distance (FID) [37], learned perceptual image patch similarity (LPIPS) [38] and the peak signal-to-noise ratio (PSNR). SSIM measures the structural similarity between two images, analyzing brightness, contrast, and structure to evaluate the similarity between the images. It is typically used to compare the quality of a reconstructed image with its original, with values ranging from 0 to 1, where 1 indicates high similarity. BRISQUE, in contrast, is a no-reference quality metric that assesses quality of a single image without requiring a reference, where lower values indicate better quality. FID measures the similarity between two image datasets, with lower values indicating a smaller quality gap between images, and is commonly used to compare the quality of generated images to real ones. LPIPS assesses the perceptual similarity between two images by comparing deep feature representations extracted from a neural network; lower LPIPS values correspond to higher perceptual similarity. PSNR evaluates the pixel-wise reconstruction quality by computing the logarithmic ratio between the maximum possible pixel value and the mean squared error, with higher values indicating better fidelity. Based on the above experimental settings, we conducted all subsequent de-identification and restoration experiments.

B. EXPERIMENT RESULT

The visual results of the experiments conducted using the LFW, FFHQ, VGGFace2-HQ, and CelebA-HQ datasets are illustrated in Figure 2 and Figure 3, which show the outcomes produced by the FaceShifter and SimSwap models, respectively. In addition, the de-identification and restoration success rates for each dataset are summarized in Table 1, and their corresponding visual comparisons are presented

TABLE 1. De-identification and restoration accuracy (%) of FaceShifter and SimSwap across four datasets (LFW, FFHQ, VGGFace2-HQ, and CelebA-HQ). Results are evaluated using three face verification models (ArcFace, FaceNet512, and VGG-Face). Higher de-identification accuracy indicates better anonymization, while higher restoration accuracy reflects more effective identity recovery.

(a) FaceShifter

Dataset	Verification Model	De-Identified	Restored
LFW	ArcFace	99.38%	99.99%
	FaceNet512	99.81%	99.87%
	VGG-Face	99.85%	99.99%
FFHQ	ArcFace	99.18%	99.99%
	FaceNet512	99.99%	99.81%
	VGG-Face	99.05%	99.99%
VGGFace2-HQ	ArcFace	96.87%	98.96%
	FaceNet512	99.92%	97.53%
	VGG-Face	97.91%	99.08%
CelebA-HQ	ArcFace	96.53%	100.00%
	FaceNet512	98.22%	99.73%
	VGG-Face	97.02%	99.99%

(b) SimSwap

Dataset	Verification Model	De-Identified	Restored
LFW	ArcFace	99.04%	99.98%
	FaceNet512	99.95%	99.56%
	VGG-Face	91.91%	99.98%
FFHQ	ArcFace	99.10%	99.89%
	FaceNet512	99.95%	97.49%
	VGG-Face	90.96%	99.86%
VGGFace2-HQ	ArcFace	98.41%	99.85%
	FaceNet512	99.98%	97.71%
	VGG-Face	92.75%	99.90%
CelebA-HQ	ArcFace	97.84%	99.91%
	FaceNet512	99.95%	98.51%
	VGG-Face	92.51%	99.98%

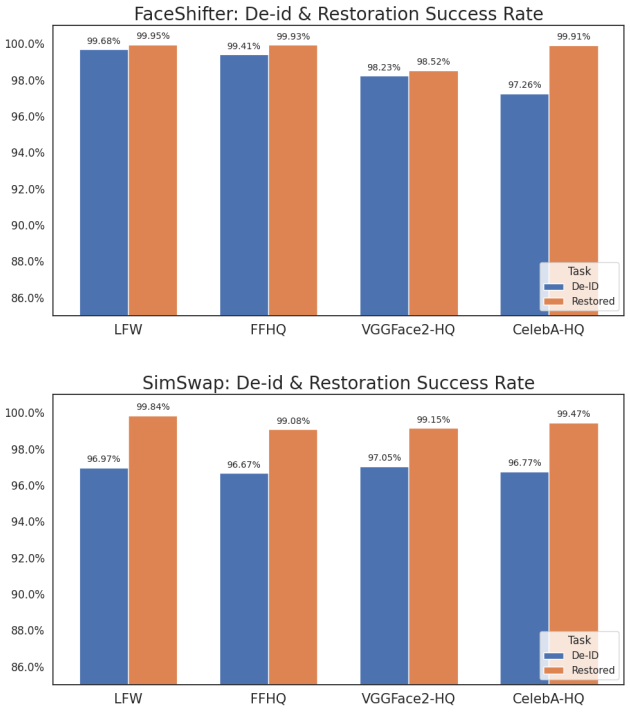


FIGURE 4. Visualization of de-identification and restoration success rates from Table 1, averaged over ArcFace, FaceNet512, and VGG-Face verification models for (a) FaceShifter and (b) SimSwap across four datasets: LFW, FFHQ, VGGFace2-HQ, and CelebA-HQ. Higher values indicate better anonymization and restoration performance.

in Figure 4. Furthermore, the results of quality assessment for both models are provided in Table 2, with Figure 5 offering a graphical representation of five widely used image quality metrics: SSIM, FID, LPIPS, PSNR, and BRISQUE.

1) DE-IDENTIFICATION AND RESTORATION (FaceShifter)

a: LFW DATASET

The results of de-identification and restoration using the LFW dataset indicated extremely high success rates for each facial verification model in recognizing de-identified images as belonging to different individuals (i.e., the de-identification success rate). The ArcFace model achieved a success rate of 99.33% in identifying image pairs as different individuals. The FaceNet512 model recorded the highest de-identification success rate at 99.81%, while the VGG-Face model showed a de-identification success rate of 98.85%. For the restoration of de-identified images, ArcFace reached 99.99%, FaceNet512 achieved 99.87%, and VGG-Face also reached 99.99% restoration success.

b: FFHQ DATASET

In the experiments using the FFHQ dataset, the de-identification success rates were similarly high, with the ArcFace, FaceNet512, and VGG-Face models achieving 99.18%, 99.99%, and 99.05% success rates, respectively. For the restoration of de-identified images, the success rates were equally high, with ArcFace at 99.99%, FaceNet512 at 99.81%, and VGG-Face at 99.99%.

c: VGGFace2-HQ DATASET

The results from the VGGFace2-HQ dataset showed slightly lower success rates compared to the LFW and FFHQ datasets, but remained high overall. The ArcFace model achieved a de-identification success rate of 96.87%, FaceNet512 achieved 99.92%, and VGG-Face achieved 97.91%. For the restoration of de-identified images, ArcFace recorded 98.96%, FaceNet512 achieved 97.53%, and VGG-Face recorded 99.08%.

d: CelebA-HQ DATASET

FaceShifter produced strong results on the CelebA-HQ dataset, with de-identification success rates of 96.53% (ArcFace), 98.22% (FaceNet512), and 97.02% (VGG-Face), and restoration success rates of 100.0% (ArcFace), 99.73% (FaceNet512), and 99.99% (VGG-Face). SimSwap achieved de-identification success rates of 97.84% (ArcFace), 99.95% (FaceNet512), and 92.51% (VGG-Face), with restoration success rates of 99.91% (ArcFace), 98.51% (FaceNet512), and 99.98% (VGG-Face).

These results confirm that the proposed method, when combined with the FaceShifter model, provides robust de-identification and reliable restoration across

TABLE 2. Quantitative evaluation of image quality for de-identified and restored outputs using (a) FaceShifter and (b) SimSwap across four datasets (LFW, FFHQ, VGGFace2-HQ, and CelebA-HQ). Metrics include SSIM, FID, LPIPS, PSNR, and BRISQUE, where ↑ indicates higher is better and ↓ indicates lower is better. BRISQUE values are reported as both the difference from the original and the absolute score (difference / absolute).

(a) FaceShifter

Dataset	Metric	De-Identified	Restored
LFW	SSIM ↑	0.9486	0.9657
	FID ↓	1.0787	0.6668
	LPIPS ↓	0.0240	0.0187
	PSNR ↑	30.4010	34.3245
	BRISQUE ↓	-0.2347 / 31.7490	+1.8156 / 33.7993
FFHQ	SSIM ↑	0.8978	0.9056
	FID ↓	6.7071	4.7227
	LPIPS ↓	0.0519	0.0565
	PSNR ↑	25.7045	28.6898
	BRISQUE ↓	+3.5324 / 6.6942	+9.7293 / 12.8911
VGGFace2-HQ	SSIM ↑	0.8721	0.8472
	FID ↓	6.2875	3.8112
	LPIPS ↓	0.0651	0.0851
	PSNR ↑	23.9561	24.6397
	BRISQUE ↓	-0.1860 / 14.2166	+4.6634 / 19.0660
CelebA-HQ	SSIM ↑	0.9399	0.9543
	FID ↓	2.1414	2.2271
	LPIPS ↓	0.0277	0.0307
	PSNR ↑	28.8664	31.8741
	BRISQUE ↓	+0.4876 / 15.1296	+3.7176 / 18.3596

(b) SimSwap

Dataset	Metric	De-Identified	Restored
LFW	SSIM ↑	0.9514	0.9579
	FID ↓	0.8298	0.6452
	LPIPS ↓	0.0208	0.0170
	PSNR ↑	31.9974	33.5091
	BRISQUE ↓	+1.3454 / 33.3291	+1.6945 / 33.6782
FFHQ	SSIM ↑	0.8341	0.8190
	FID ↓	19.7097	15.8504
	LPIPS ↓	0.0797	0.0773
	PSNR ↑	26.1322	26.0516
	BRISQUE ↓	+11.8354 / 14.9972	+12.6911 / 15.8529
VGGFace2-HQ	SSIM ↑	0.7312	0.6719
	FID ↓	19.9047	16.7163
	LPIPS ↓	0.1309	0.1285
	PSNR ↑	22.3970	21.0113
	BRISQUE ↓	+10.1827 / 24.5853	+10.9031 / 25.3057
CelebA-HQ	SSIM ↑	0.9219	0.9248
	FID ↓	3.8333	2.8213
	LPIPS ↓	0.0293	0.0291
	PSNR ↑	29.4248	30.1624
	BRISQUE ↓	+0.8797 / 15.5217	+2.0768 / 16.7188

various datasets. Notably, the near-perfect restoration success across all models demonstrates that the identity information encrypted via format-preserving encryption can be accurately recovered without compromising visual quality or data utility.

2) DE-IDENTIFICATION AND RESTORATION (SimSwap)

a: LFW DATASET

SimSwap achieved de-identification success rates of 99.04%, 99.95%, and 91.91% with ArcFace, FaceNet512, and VGG-Face, respectively. Restoration results remained high across all models, reaching 99.98% for both ArcFace and VGG-Face, and 99.56% for FaceNet512.

b: FFHQ DATASET

The model continued to perform well on FFHQ, achieving de-identification rates of 99.10% (ArcFace), 99.95% (FaceNet512), and 90.96% (VGG-Face). Restoration success was also high, with all models achieving above 97%.

c: VGGFace2-HQ DATASET

De-identification rates were 98.41%, 99.98%, and 92.75% for ArcFace, FaceNet512, and VGG-Face, respectively. Restoration success remained strong, with ArcFace at 99.85%, FaceNet512 at 97.71%, and VGG-Face at 99.90%.

d: CelebA-HQ DATASET

Finally, on CelebA-HQ, SimSwap achieved de-id success rates of 97.84%, 99.95%, and 92.51%. Restoration success was again strong, with ArcFace reaching 99.91%, FaceNet512 at 98.51%, and VGG-Face at 99.98%.

While SimSwap showed slightly lower de-identification rates with the VGG-Face model, particularly on LFW and FFHQ, the overall restoration performance remained excellent. These results suggest that although SimSwap offers slightly less aggressive anonymization in some cases, it effectively maintains high restoration fidelity, making it a viable alternative for reversible facial de-identification.

3) IMAGE QUALITY EVALUATION (FACESHIFTER)

a: LFW DATASET

The de-identified images achieved an SSIM of 0.9486, indicating strong structural similarity with the original images. The FID score of 1.0787 and LPIPS of 0.0240 suggest minimal perceptual deviation. PSNR reached 30.4010, while the BRISQUE score was 31.7490, only slightly deviating from the original, with a BRISQUE gap of -0.2347. The restored images exhibited even higher fidelity, with an SSIM of 0.9657, FID of 0.6668, LPIPS of 0.0187, and PSNR of 34.3245. The BRISQUE difference remained low at $+1.8156$, with an overall score of 33.7993.

b: FFHQ DATASET

The de-identified images showed an SSIM of 0.8978 and a FID of 6.7071, with LPIPS and PSNR values of 0.0519 and 25.7045, respectively. BRISQUE was 6.6942, with a difference of $+3.5324$ from the original. For restored images, SSIM increased to 0.9056, FID improved to 4.7227, and PSNR rose to 28.6898, demonstrating improved visual quality. However, the BRISQUE score increased to 12.8911, with a total difference of $+9.7293$, reflecting a slight degradation in perceptual quality.

c: VGGFace2-HQ DATASET

De-identified images recorded SSIM and FID scores of 0.8721 and 6.2875, respectively, with LPIPS of 0.0651 and PSNR of 23.9561. The BRISQUE score was 14.2166, with a deviation of -0.1860 from the original. Restored images maintained high fidelity, with SSIM of 0.8472, FID of 3.8112, and PSNR of 24.6397. Although LPIPS slightly increased

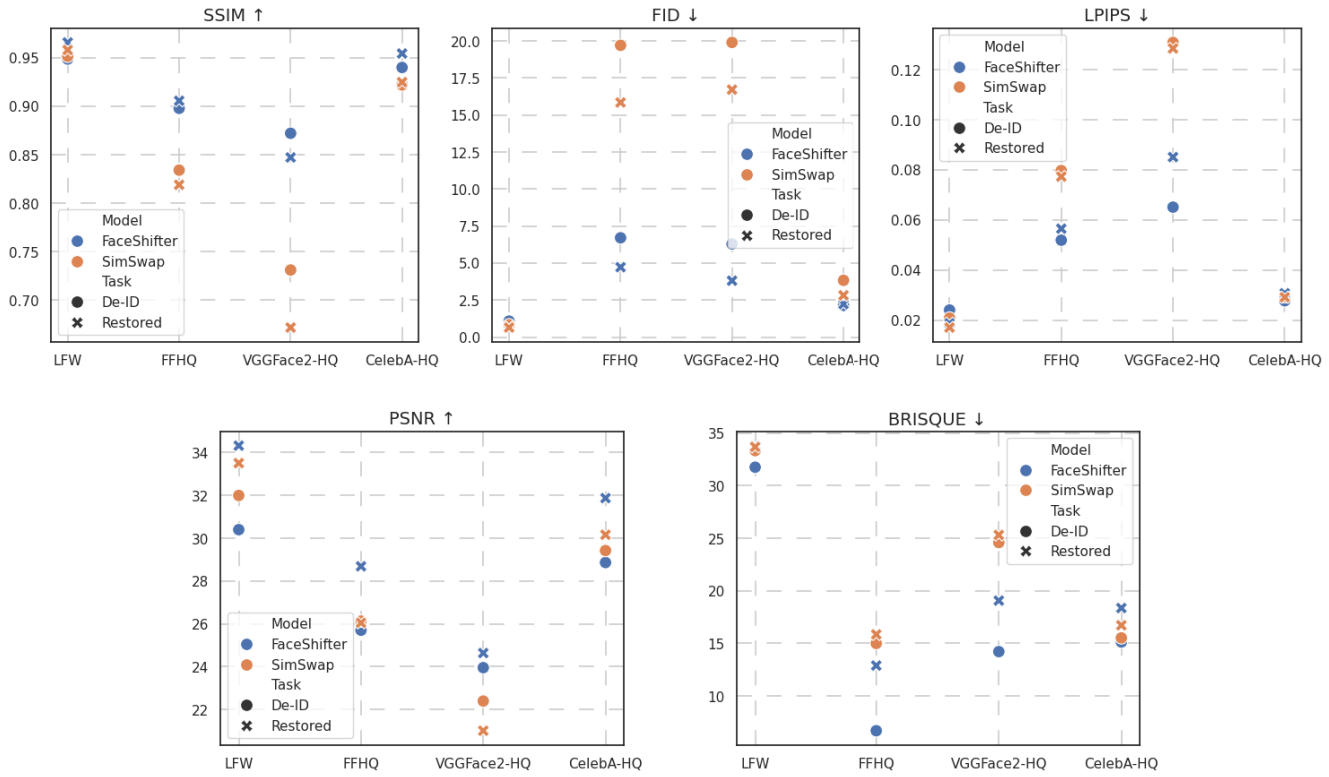


FIGURE 5. Visualization of image quality metrics from Table 2, comparing SSIM, FID, LPIPS, PSNR, and BRISQUE scores for de-identified (circle markers) and restored (cross markers) images produced by FaceShifter and SimSwap across four datasets. Lower FID, LPIPS, and BRISQUE indicate better quality, whereas higher SSIM and PSNR indicate better fidelity.

to 0.0851, the BRISQUE score rose to 19.0660, with a difference of +4.6634.

d: CelebA-HQ DATASET

This dataset yielded the highest SSIM among all cases, with de-identified images at 0.9399 and restored images at 0.9543. The FID scores were low—2.1414 for de-identified and 2.2271 for restored images—indicating high visual similarity. LPIPS was consistently low (0.0277 and 0.0307), and PSNR values were 28.8664 and 31.8741, respectively. The BRISQUE scores were also favorable: 15.1296 (de-identified) and 18.3596 (restored), with moderate differences of +0.4876 and +3.7176.

Overall, these results confirm that the proposed method, when applied with FaceShifter, preserves both perceptual and structural fidelity during the de-identification and restoration processes. The restored images closely resemble the original inputs across diverse datasets, validating the effectiveness of reversible privacy-preserving face processing.

4) IMAGE QUALITY EVALUATION (SIMSWAP)

a: LFW DATASET

The SimSwap model achieved an SSIM of 0.9514 and a FID of 0.8298 for the de-identified images, demonstrating strong visual similarity to the originals. LPIPS was very low at 0.0208, and PSNR reached 31.9974. The BRISQUE score was 33.3291, with a difference of +1.3454 compared to the original. Restoration further improved image quality, yielding an SSIM of 0.9579 and a reduced FID of 0.6452. LPIPS decreased to 0.0170, PSNR increased to 33.5091, and the BRISQUE difference was +1.6945, with a restored image score of 33.6782.

b: FFHQ DATASET

For the FFHQ dataset, de-identified images achieved an SSIM of 0.8341, a FID of 19.7097, and a LPIPS of 0.0797, indicating slightly lower quality than LFW. PSNR was 26.1322, and the BRISQUE score was 14.9972 with a difference of +11.8354. Restoration yielded an SSIM of 0.8190 and reduced the FID to 15.8504, while LPIPS slightly decreased to 0.0773. PSNR was maintained at 26.0516, and the BRISQUE score reached 15.8529, with a gap of +12.6911.

c: VGGFace2-HQ DATASET

The de-identified images recorded an SSIM of 0.7312 and FID of 19.9047, which were the lowest among the four datasets. LPIPS was 0.1309, PSNR was 22.3970, and the BRISQUE difference was +10.1827 (score: 24.5853). The restored images improved slightly across all metrics, with SSIM of 0.6719, FID of 16.7163, LPIPS of 0.1285, and PSNR of 21.0113. The BRISQUE score increased to 25.3057,

with a difference of +10.9031, indicating moderate quality retention.

d: CelebA-HQ DATASET

The CelebA-HQ dataset yielded strong results with an SSIM of 0.9219 for de-identified images and 0.9248 for restored ones. FID values were relatively low (3.8333 and 2.8213), and LPIPS remained minimal (0.0293 and 0.0291). PSNR values were 29.4248 and 30.1624, respectively. BRISQUE scores were 15.5217 for the de-identified images and 16.7188 for the restored ones, with corresponding gaps of +0.8797 and +2.0768, indicating high visual fidelity throughout the process.

As a result, the SimSwap-based pipeline achieved favorable image quality across all datasets. Although the visual quality was slightly lower for the VGGFace2-HQ dataset, the restored images in all cases maintained high structural similarity and perceptual coherence, confirming the effectiveness of the proposed reversible de-identification framework based on SimSwap.

These experimental results verify the robustness and adaptability of the proposed reversible de-identification framework. Despite variations in dataset characteristics, both FaceShifter and SimSwap-based pipelines consistently maintained high-quality outputs across all metrics. In particular, the method preserved critical visual fidelity during both de-identification and restoration, demonstrating its practical viability for secure and privacy-preserving facial data management in real-world applications.

V. DISCUSSION

While the proposed method demonstrates strong performance across most datasets, we observed relatively lower results on the VGGFace2-HQ dataset. As shown in Table 2, this dataset exhibited a decline in SSIM and LPIPS scores compared to others, likely due to its inclusion of extreme pose variations and occlusions. These challenging facial conditions may not have been sufficiently represented during training. To improve generalization, future studies should incorporate more diverse and representative facial data, including images with motion blur, occlusions, and low-light environments.

Another practical consideration involves the computational complexity of the models used. Both FaceShifter and SimSwap rely on deep generative networks that demand significant computational resources. This could be a barrier for real-time applications, particularly on resource-constrained platforms such as mobile devices or embedded systems. Techniques such as model pruning, quantization, and the development of lightweight network architectures could be explored to address this challenge and support deployment in real-world environments.

Unlike conventional de-identification methods that manipulate embedding vectors without encryption, our approach integrates cryptographic safeguards directly into the identity representation. In non-encrypted baselines, the identity vector can remain vulnerable to inversion or model leakage, potentially enabling unauthorized identity reconstruction. By applying FPE, we ensure that even if the de-identified embedding is exposed, it cannot be reversed without the encryption key. This addresses a critical vulnerability in non-FPE models and represents a fundamental security improvement.

Furthermore, compared to traditional password-based protection schemes, FPE offers stronger resilience against brute-force and dictionary attacks while preserving the data format. However, it introduces a new dependency on secure key management. If the encryption key is compromised, all associated de-identified images become vulnerable to restoration. Future research should examine robust key storage, controlled access mechanisms, and potentially hardware-backed security modules to prevent unauthorized use.

Beyond technical safeguards, ethical and social considerations are also crucial. The ability to restore identity from anonymized data, while valuable in legal or investigative contexts, raises dual-use concerns. For instance, malicious actors gaining access to decryption keys could exploit the system. This highlights the need for governance frameworks and ethical deployment policies that go beyond algorithmic robustness.

Fairness and demographic balance must also be considered. If the training datasets lack diversity across demographic lines—such as age, gender, or ethnicity—the performance of de-identification and restoration could be uneven across population groups. To ensure equitable privacy protection, fairness-aware training strategies and the use of demographically balanced datasets are essential.

The proposed framework is also promising for applications beyond facial images. Biometric modalities such as voice, fingerprints, and gait patterns contain sensitive personal information and can benefit from similar format-preserving, reversible anonymization. Extending this approach to multimodal privacy protection represents a compelling direction for future research, particularly in contexts requiring unified security across various data types.

From a regulatory and ethical standpoint, the proposed framework aligns well with privacy standards such as the GDPR, as it enables controlled, consent-based restoration of anonymized data. This makes it suitable for sensitive applications such as digital forensics, privacy-aware healthcare systems, and secure access control services.

Architectural differences between FaceShifter and SimSwap also suggest complementary strengths. While FaceShifter offers slightly better perceptual quality, SimSwap provides high restoration fidelity with more efficient processing. Depending on the specific application, developers may select between them based on the desired balance between visual quality and computational cost.

In summary, the proposed framework presents a secure, flexible, and practical solution for reversible facial de-identification. Continued research is necessary to enhance

its generalizability, address ethical concerns, ensure fair and inclusive performance, and support scalable deployment in real-world applications.

VI. CONCLUSION

We proposed a reversible facial de-identification framework that integrates symmetric-key format-preserving encryption (FPE) into deep neural network-based face-swapping models. This work was motivated by the increasing need to securely manage facial data in digital systems while enabling lawful restoration when required. The method supports identity removal and recovery without compromising data structure or visual fidelity.

The proposed framework is applied to two structurally different de-identification networks—FaceShifter and SimSwap—demonstrating its versatility and modularity. Notably, the approach maintains non-identity attributes such as pose and expression while ensuring that the identity information can be reconstructed when authorized. Through extensive qualitative and quantitative evaluations, we show that the framework enables high-fidelity face manipulation with minimal perceptual distortion, achieving a favorable trade-off between privacy and utility.

Beyond its technical implementation, this work highlights that reversible de-identification can be securely realized without relying on vulnerable password-based schemes. It also contributes to the broader understanding of how different network architectures affect de-identification robustness under complex visual conditions. The ability to flexibly embed FPE into different models suggests that the method can be extended to other identity-sensitive domains as well.

Future research will aim to improve restoration fidelity under highly unconstrained scenarios and expand applicability to additional real-world datasets and edge deployments where privacy guarantees are essential.

ACKNOWLEDGMENT

GPT-4o was used in the following ways: to help them iterate on LaTeX formatting; for text summarization; and as a copyediting tool.

REFERENCES

[1] A. Miotti and A. Wasil, "Combatting deepfakes: Policies to address national security threats and rights violations," 2024, *arXiv:2402.09581*.

[2] European Commission. (2018). *Eu Data Protection Rules*. Accessed: Sep. 12, 2024. [Online]. Available: https://commission.europa.eu/law/law-topic/data-protection/data-protection-euen

[3] J. Song, J. Kim, and J. Nang, "Face de-identification using convolutional neural network (CNN) models for visual-copy detection," *Appl. Sci.*, vol. 14, no. 5, p. 1771, Feb. 2024.

[4] Y. Wen, B. Liu, L. Song, J. Cao, and R. Xie, "Differential private identification protection for face images," in *Face De-Identification: Safeguarding Identities in the Digital Era*. Cham, Switzerland: Springer, 2024, pp. 75–108.

[5] 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.

[6] D. Li, W. Wang, K. Zhao, J. Dong, and T. Tan, "RiDDLE: Reversible and diversified de-identification with latent encryptor," 2023, *arXiv:2303.05171*.

[7] D. Xiao, J. Xia, M. Li, and M. Zhang, "Manipulable, reversible and diversified de-identification via face identity disentanglement," *Multimedia Tools Appl.*, vol. 83, no. 31, pp. 75653–75670, Feb. 2024.

[8] Y. Wen, B. Liu, L. Song, J. Cao, and R. Xie, "Deep motion flow guided reversible face video de-identification," in *Face De-identification: Safeguarding Identities in the Digital Era*. Cham, Switzerland: Springer, 2024, pp. 147–176.

[9] M. Bellare, T. Ristenpart, P. Rogaway, and T. Stegers, "Format-preserving encryption," in *Proc. Sel. Areas Cryptography: 16th Annu. Int. Workshop*. Alberta, ED, Canada: Springer, Jan. 2009, pp. 295–312.

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

[11] 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.

[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] K. Mivule, "Utilizing noise addition for data privacy, an overview," 2013, *arXiv:1309.3958*.

[14] H. Liu, M. Steinebach, R. Stein, and F. Mayer, "Privacy preserving forensics for JPEG images," *Electron. Imag.*, vol. 30, no. 7, pp. 120-1–120-6, Jan. 2018.

[15] L. Fan, "Image pixelization with differential privacy," in *Proc. Data Appl. Secur. Privacy XXXII: 32nd Annu. IFIP WG 11.3 Conf.* Bergamo, Italy: Springer, Jan. 2018, pp. 148–162.

[16] C. Neustaedter, S. Greenberg, and M. Boyle, "Blur filtration fails to preserve privacy for home-based video conferencing," *ACM Trans. Computer-Human Interact.*, vol. 13, no. 1, pp. 1–36, Mar. 2006.

[17] N. Ruchaud and J.-L. Dugelay, "Automatic face anonymization in visual data: Are we really well protected?" *Electron. Imag.*, 2016.

[18] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1–27.

[19] H. Hukkelås, R. Mester, and F. Lindseth, "DeepPrivacy: A generative adversarial network for face anonymization," in *Proc. Int. Symp. Vis. Comput.* Cham, Switzerland: Springer, Jan. 2019, pp. 565–578.

[20] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Computer-Assisted Intervent. (MICCAI)*. Munich, Germany: Springer, Jan. 2015, pp. 234–241.

[21] M. Maximov, I. Elezi, and L. Leal-Taixé, "CIAGAN: Conditional identity anonymization generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 5446–5455.

[22] O. Gafni, L. Wolf, and Y. Taigman, "Live face de-identification in video," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9377–9386.

[23] Y. Liu, Q. Li, Z. Sun, and T. Tan, "A3GAN: An attribute-aware attentive generative adversarial network for face aging," *IEEE Trans. Inf. Forensics Security*, vol. 16, no. 11, pp. 2776–2790, Nov. 2021.

[24] T. Karras, T. Aila, S. Laine, and J. Lehtinen, "Progressive growing of GANs for improved quality, stability, and variation," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018.

[25] 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.

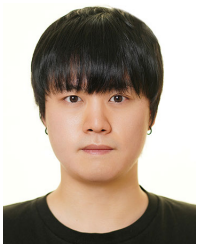
[26] J. Daemen, "Aes proposal: Rijndael," Nat. Inst. Standards Technol. (NIST), Gaithersburg, MD, USA, Tech. Rep. NIST SP 800-38G, 1999.

[27] National Institute of Standards and Technology, "Recommendation for block cipher modes of operation: Methods for format-preserving encryption," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. NIST SP 800-38G, 2016. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.800-38G.pdf.

[28] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces in'Real-Life'Images: Detection, Alignment, Recognit.*, 2008, pp. 1–10.

[29] 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.
[30] 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.
[31] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 3730–3738.
[32] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 815–823.
[33] O. Parkhi, A. Vedaldi, and A. Zisserman, "Deep face recognition," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Jan. 2015, pp. 41.1–41.12.
[34] G. Jocher, A. Chaurasia, and J. Qiu. *Ultralytics Yolov8*. Accessed: Aug. 23, 2024. [Online]. Available: https://www.ultralytics.com/yolo
[35] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.
[36] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.
[37] 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. Neural Inf. Process. Syst.*, Jan. 2017, pp. 6626–6637.
[38] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

HEEHWAN KIM received the B.S. degree in software engineering from Soongsil University, South Korea, in 2025, where he is currently pursuing the M.S. degree in AI security convergence. His research interests include information privacy protection, AI security, and generative AI.



SUNGJUNE PARK received the B.S. degree in software engineering from Soongsil University, South Korea, in 2023, where he is currently pursuing the M.S. degree in software convergence. His research interests include information privacy protection, AI security, and generative AI.

DAESEON CHOI (Member, IEEE) received the B.S. degree in computer science from Dongguk University, South Korea, in 1995, the M.S. degree in computer science from Pohang Institute of Science and Technology, South Korea, in 1997, and the Ph.D. degree in computer science from Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2009. He was a Professor with the Department of Medical Information, Kongju National University, South Korea, from September 2015 to August 2020. He is currently a Professor with the Department of Software, Soongsil University, South Korea. His research interests include identity management and information security.