ITC 2/55
Information Technology and Control
Vol. 55 / No. 2/ 2026
pp. 447-468
DOI 10.5755/j01.itc.55.2.43026

Face Privacy Protection and Anonymization Model Integrating Multi-Condition Collaborative Control GANs

Received 2025/10/10 | Accepted after revision 2026/01/09

**HOW TO CITE:** Zhang, Y., Song, H. (2026) Face Privacy Protection and Anonymization Model Integrating Multi-Condition Collaborative Control GANs. *Information Technology and Control*, 55(2), 447-468. https://doi.org/10.5755/j01.itc.55.2.43026

# Face Privacy Protection and Anonymization Model Integrating Multi-Condition Collaborative Control GANs

**Yuzhen Zhang***
Science & Technology Department, Wuxi Vocational Institute of Arts &Technology, Wuxi 214206, China

**Houfei Song**
Teaching Affairs Department, Wuxi Vocational Institute of Arts &Technology, Wuxi 214206, China

**Corresponding author:** zhangyzhen06@163.com

The widespread deployment of face recognition and visual analytics has made balancing privacy protection and image usability a critical challenge. Existing anonymization methods often rely on blurring or occlusion, which suppress identity but distort key non-identity attributes. To address this issue, this study proposes a reversible face anonymization framework based on multi-conditional collaborative control (RFAMCC), integrating multi-level feature-driven anonymization, identity-space deflection, context fusion, and a steganography–recovery mechanism. Experimental evaluations demonstrated that RFAMCC preserved semantic consistency under original, compressed, and cropped conditions. It achieved correlation scores of 0.312, 0.308, and 0.301, as well as face retrieval accuracies of 83.7%, 82.4%, and 80.9%, respectively. In privacy-sensitive evaluations, gender and age inference attacks attained low success rates of 21.4% and 23.1%, indicating strong resistance to attribute leakage. Overall, RFAMCC effectively balances anonymization strength and reversibility, providing a practical solution for privacy-preserving and legally compliant facial data sharing.

KEYWORDS: Multi-condition; Collaborative control; GAN; Facial privacy protection; Anonymization; Feature steganography

## 1. Overview

In recent years, with the widespread application of facial recognition technology in public safety, financial payments, and intelligent interaction, facial images have emerged as a crucial carrier of biometric data, with their value and associated risks becoming increasingly prominent [1]. However, facial images not only contain core features for identity verification but also carry sensitive attributes such as gender, age, and expressions. If improperly collected or misused, they can easily lead to serious issues like privacy breaches and identity theft [16]. Particularly against the backdrop of widespread surveillance and social media proliferation, protecting facial privacy has become a critical issue demanding urgent solutions [10, 12]. Therefore, constructing anonymization models that balance privacy security and image usability has become a significant research direction in the fields of facial recognition and computer vision [7]. In response to this trend, scholars have proposed various approaches from perspectives such as feature decoupling, conditional generation, and adversarial optimization. He et al. proposed a face privacy protection method based on a diffusion model, unifying anonymization and visual identity information hiding within the same framework. This method generated conditional embeddings through multi-scale image inversion and set different energy functions during the inference process to achieve two types of tasks [4]. Osorio-Roig et al. addressed the lack of adversarial attack validation in existing privacy-enhancing facial recognition systems by proposing an attack method based on soft biometric attributes. This approach compared intercepted facial representations with an attacker database and inferred unknown attributes based on maximum similarity [17]. Wang et al. proposed an identity concealer to address privacy violations caused by unintentional observation of facial images in databases. This method used virtual face generation and appearance transformation modules to alter appearances while preserving parsing maps and identity features, balancing diversity and background protection [24]. Ghani et al. proposed a privacy-aware framework integrating generative adversarial networks (GANs), blockchain, and distributed computing to address privacy concerns arising from the widespread application of facial recognition in the digital age. This framework balances identity verification with data security and enabled scalable facial analysis through advanced tools [2]. Sun et al. proposed a mask-based adversarial attack method addressing the rapid development of mask-wearing facial recognition algorithms and associated privacy leakage risks during the pandemic. This approach utilized projection networks to generate adaptive masks that degraded recognition capabilities under rotation and lighting variations [20].

Subsequently, researchers explored practical challenges such as surveillance footage, social sharing, and data recovery, proposing more targeted and scalable frameworks for facial privacy protection. Sivalakshmi et al. introduced a novel deep learning model [19]. This approach comprised two stages, facial detection and privacy protection, achieving optimal detection performance through a hybrid convolutional neural network combined with an enhanced Hawk algorithm while maintaining service quality control [19]. Xiao et al. proposed a protection scheme based on multi-feature fusion tensors to address privacy concerns arising from the misuse of social media data for training facial recognition models. This approach integrated the correlation between deep features and artificial features, enhancing the robustness and portability of disguised images [25]. Morris et al. proposed an image privacy protection system, which used location aware multi-party access control models to achieve real-time privacy protection for large-scale users on social networks. When sensitive locations were detected, the system automatically replaced faces with synthetic faces [15]. Zhong et al. proposed a personalized privacy protection method called "one person, one mask". This approach generated personalized universal masks by optimizing the orientation of training samples in feature subspaces, and was validated across different network architectures and datasets [30]. Kim et al. proposed a reversible facial de-identification method based on format-preserving encryption to address the risks of identity theft and deepfakes associated with facial images stored across various platforms. This approach combined

a deep neural network-based face swapping model to separate identity and attribute information while supporting authorized restoration [9].

Subsequently, researchers proposed more adaptive privacy protection schemes from the perspectives of steganographic communication, network security, and generative model anonymization. Zhou et al. proposed a generative steganography method based on latent vector optimization for consumer electronics scenarios [31]. This method adaptively allocated embedding capacity in the latent space of the streaming generative model. This significantly expanded the amount of hidden information and improves steganography quality. Thus, it makes up for the capacity and visual quality shortcomings of traditional GIS [31]. Yang et al. proposed an adaptive steganography algorithm that combined hybrid machine learning [26]. This algorithm achieved differentiated embedding through foreground-background partitioning, encrypts secret information, and improved embedding capacity while maintaining high perceptual quality. The peak signal-to-noise ratio (PSNR) exceeded 52 dB [26]. In the field of anonymization, Shaheryar et al. proposed a biconditional diffusion model, which preserved task attributes such as pose and expression by simultaneously modeling identity and non-identity features, and maintained higher de-identification stability and temporal consistency under extreme poses [18]. You et al. further proposed a diffusion-based recoverable anonymization framework (DIFP), which used a conditional diffusion generator and latent variable optimization to generate high-resolution encrypted faces and achieves authorization recovery through diffusion denoising. It achieved excellent performance in both privacy protection and reversibility [28].

As summarized in the existing research, although face anonymization has made some progress in identity removal and visual quality, several key issues remain unresolved. First, most methods rely on single-scale or single-module feature perturbations. This can weaken the representation of identity while damaging task-related attributes, such as expression, pose, and lighting. This leads to a decline in the application value of anonymized images. Second, mainstream research currently focuses on irreversible anonymization. Although reversible schemes have made progress in format preservation and diffusion reconstruction, they still have limitations in cross-scene consistency, controllable recovery capability, and stability of embedded information. Third, although recent diffusion model anonymization methods and format-preserving reversible methods have improved the quality and controllability of synthesis, they have not yet been deeply integrated into the GAN anonymization framework. This is due to the lack of a unified mechanism that can simultaneously consider privacy protection, reversible recovery, and multi-condition semantic control. This study proposes a reversible face anonymization model based on multi-condition collaborative control (RFAMCC) to address the research gaps and achieve a balance between privacy protection and identity recovery. The innovation of this research lies in the organic integration of reversible feature steganography with the GAN anonymization framework. A more consistent anonymization-recovery process is constructed through a joint design of identity perturbation, attribute preservation, and controllable recovery. RFAMCC employs multi-level, feature-driven mechanisms, identity space shifting, context fusion, and steganography-recovery coupling to suppress identity features while preserving non-identity attributes, such as facial expressions and poses, to the greatest extent possible. This achieves a controllable balance between privacy protection and image usability. Compared to existing diffusion-based anonymization methods and format-preserving reversible techniques, the overall framework further enhances reversibility, multi-condition control capabilities, and cross-scene consistency.

The main contributions of this study are as follows: (1) A new framework for reversible anonymization: A reversible anonymization model based on a multi-condition collaborative control system is constructed. This model organically combines GAN anonymization generation and a feature steganography mechanism. This combination overcomes the problem of limited reversibility. (2) Multi-scale feature collaborative driving: The multi-level feature driving and identity space deflection module is designed to maximize preservation of non-identity attributes, such as expressions and poses, while ensuring the identity remains unrecognizable. This achieves a balance between privacy protection and image usability. (3) Dual-channel steganography and recovery mechanism: A unified process that combines deep steganography and reversible

reconstruction is proposed. This process enables the model to recover identities under authorized conditions while maintaining anonymity. (4) Cross-scene robust anonymization capability: The consistency of the model is verified under different lighting, pose, expression and scene conditions, proving its generalization and robustness in multi-scene tasks.

The structure of this study is as follows: Section 1 introduces the relevant research and technical background. Section 2 presents the overall framework and core methods of RFAMCC. Section 3 presents the experimental design and results analysis. Section 4 is the discussion. Section 5 summarizes the entire study and looks forward to future research directions.

## 2. Methodology

This study first proposes multi-level feature-driven face anonymization (MLFA). By employing semantic decomposition and multi-scale feature extraction, it achieves equilibrium between weakening identity characteristics and preserving usable attributes. Second, by introducing mechanisms such as ISD, context fusion module (CFM), and FSR, the RFAMCC framework is further constructed to achieve a unified approach to anonymization and reversible reconstruction during the generation process.

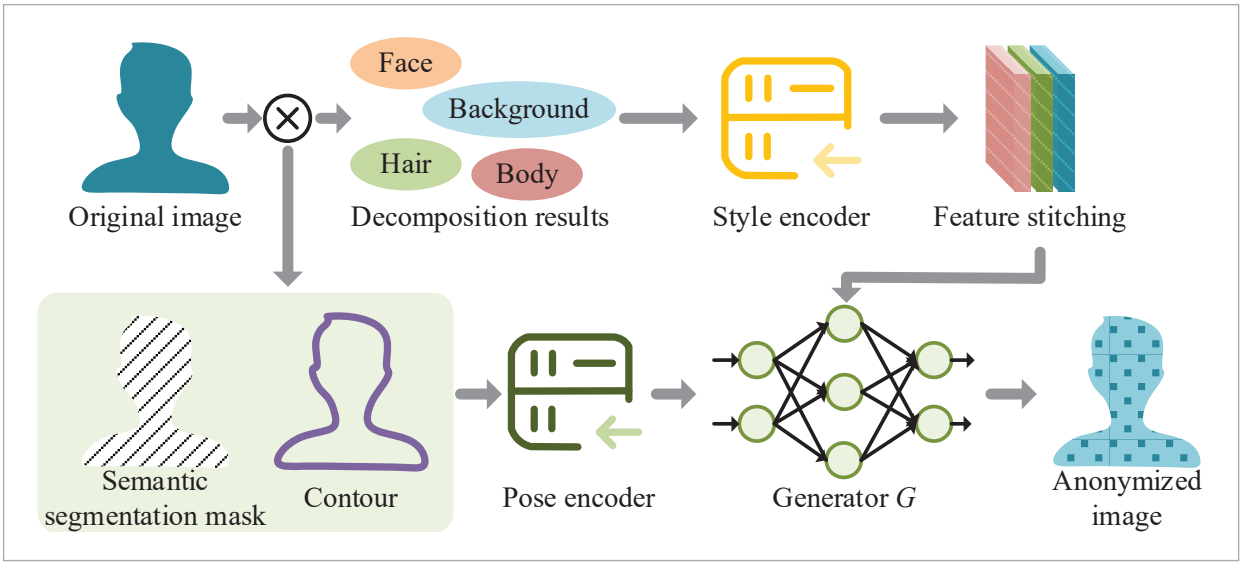
### 2.1. MLFD's Face Privacy Protection and Anonymization Mechanism

Given GAN's strengths in image generation and feature modeling, this study introduces MLFD-based facial anonymization mechanisms on top of GAN, proposing MLFA. The overall framework is illustrated in Figure 1.

As shown in Figure 1, MLFA first performs semantic parsing on the input facial image to obtain multi-region representations including the face, background, hair, and body. These are then combined with contour and pose priors to form structural constraints. Subsequently, a multi-scale encoder extracts hierarchical sensitive features from each region and applies anonymization transformations to the subspaces carrying identity information. The anonymized regional features are concatenated along the channel dimension to form a conditional set. Together with the geometric representations output by the pose encoder, this set serves as input to the generator. This enables the reconstruction of images that maintain visual consistency while rendering the identity unrecognizable. The resulting anonymized image preserves non-identity attributes such as facial expressions and poses while reducing recognizability, thereby achieving privacy protection.

Specifically, the original facial image is first segmented into distinct regions, including the face, back-

**Figure 1**
Overall framework of the MLFA model.



ground, hair, and body, via semantic parsing. A multi-scale feature encoder then extracts multi-level feature representations. These features are concatenated into a unified feature vector, as shown in Equation (1) [29].

$$s = concat(s_{face}, s_{bg}, s_{hair}, s_{body}). \quad (1)$$

In Equation (1), $s_{face}$, $s_{bg}$, $s_{hair}$, and $s_{body}$ denote the latent feature vectors for the face, background, hair, and body regions, respectively. Concurrently, the semantic mask image $x_m$, along with the face landmarks and contour map $x_{land}$, are fed into the pose encoder $E_{pose}(\cdot)$ to extract geometric and structural information, resulting in the pose latent vector as shown in Equation (2).

$$z_{pose} = E_{pose}(x_m, x_{land}). \quad (2)$$

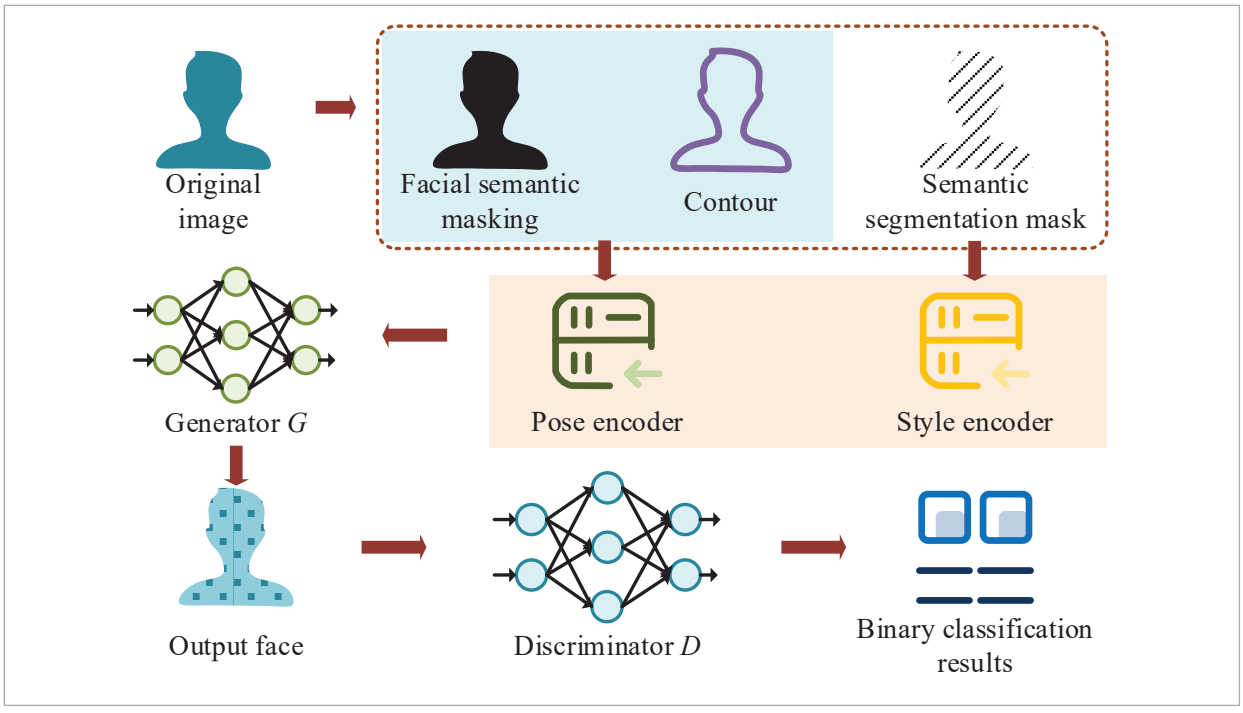
In Equation (2), $x_m$ provides the semantic distribution of the primary facial region, while $x_{land}$ supplies local geometric constraints. Their combination ensures that anonymized images retain reasonable spatial relationships during generation. Subsequently, the anonymized regional feature embeddings $s$ are input alongside the pose encoding vector $z_{pose}$ into the generator $G(\cdot)$ to synthesize new anonymized facial images, as shown in Equation (3).

$$x_{rec} = G(z_{pose}, s). \quad (3)$$

In Equation (3), $G(\cdot)$ denotes the generator, and $x_{rec}$ represents the final reconstructed anonymous image. Finally, the generated image $x_{rec}$ is fed into the discriminator $D(\cdot)$, which assesses the authenticity distribution of the image. Therefore, the processing flow of the face synthesis network is illustrated in Figure 2.

In Figure 2, the entire anonymization process consists of components $E_{style}$, $E_{pose}$, $G$, and $D$. First, the semantic segmentation map and contour annotation map are input into the pose encoder $E_{pose}$ to extract pose-related structural priors. Simultaneously, the original image undergoes multi-scale feature encoding by the multi-scale feature encoder $E_{style}$ to obtain hierarchical feature representations across different regions. Subsequently, the generator $G$ fuses the pose vector with the multi-scale features to output

**Figure 2**
The operation process of the anonymous face generation network.



the anonymized reconstructed image $x_{rec}$. Finally, the discriminator $D$ distinguishes between genuine and synthetic input images, continuously driving the generator through adversarial training to produce more realistic anonymized images.

Therefore, the model can extract multidimensional representations of facial regions through a multi-scale feature encoder. However, these features often contain identity-sensitive information, and using them directly in the generation process may still compromise personal privacy. Consequently, further anonymization of sensitive features is required to weaken or remove identity-related patterns [3, 6]. Specifically, it is assumed that the input image yields a feature set $\{f_1, f_2,...,f_n\}$ after passing through the multi-scale feature encoder. Among these, each $f_i$ represents a feature vector at a distinct scale. To achieve anonymization, these features must first be clustered, making features within the same category more similar and less distinguishable in terms of individual differences. The k-means method accomplishes this by minimizing the distance between samples and their respective cluster centers. Its optimization objective is given by Equation (4) [8, 13].

$$\min \sum_{i=1}^{n} \left\| f_i - c_{j(i)} \right\|^2 . \tag{4}$$

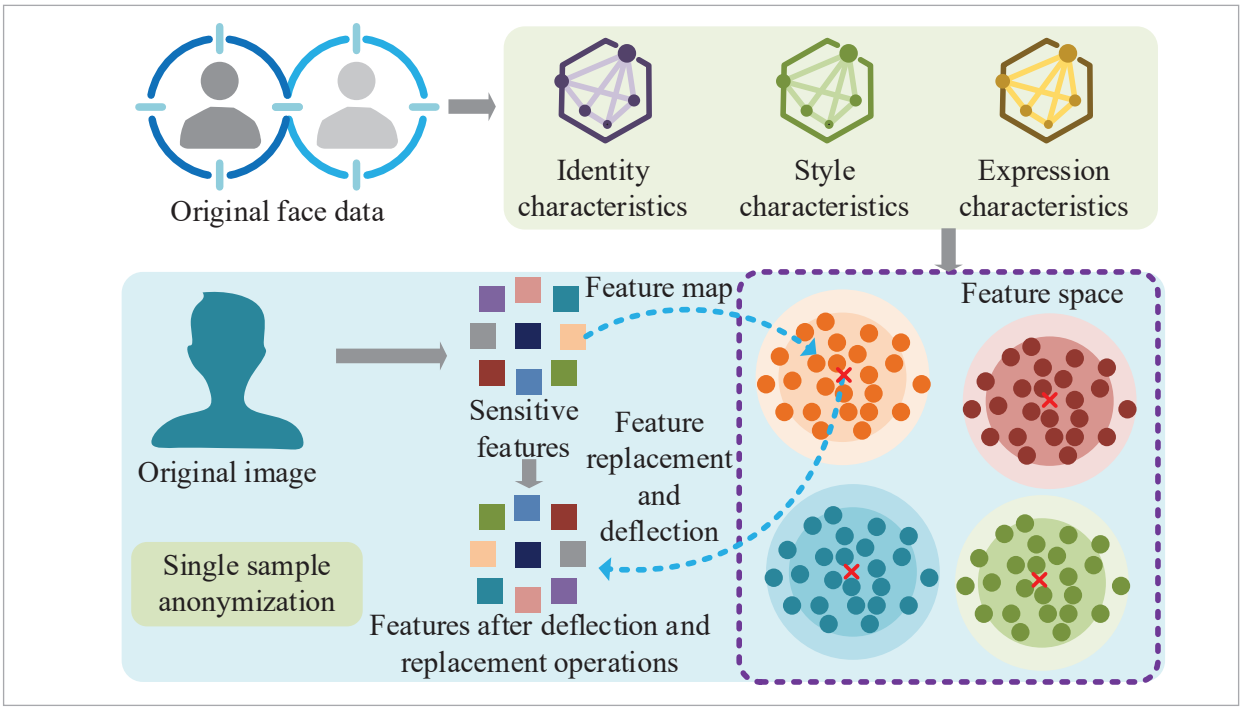
In Equation (4), $c_{j(i)}$ denotes the centroid vector of the cluster to which feature $f_i$ belongs. This objective ensures that sensitive features within the same cluster tend to aggregate, thereby blurring individual differences. After completing the clustering, each input feature is replaced with its corresponding cluster centroid, achieving an anonymized mapping as shown in Equation (5) [5, 21].

$$f_i^{anon} = c_{j(i)}, \quad i = 1,2,...,n . \tag{5}$$

In Equation (5), $f_i^{anon}$ denotes the anonymized feature representation. This substitution process ensures that features within the same cluster are unified, thereby reducing the distinguishability of original identity features. To illustrate this process more intuitively, the clustering and substitution workflow for identity-sensitive features is shown in Figure 3.

**Figure 3**

Identity-sensitive feature clustering and replacement process.



As shown in Figure 3, the original facial image undergoes encoding to extract multi-scale sensitive features, which are then classified using a clustering algorithm. Each category's features are represented by its cluster center. Finally, during the anonymization stage, the sensitive features of the input image are replaced with cluster centers that correspond to their respective categories. This completes the masking of identity-sensitive information.

An adversarial loss is introduced within the generative adversarial framework. The discriminator $D(\cdot)$ distinguishes between real and anonymized images, while the generator continuously optimizes to enhance the realism of generated images, as shown in Equation (6) [23, 27].

$$L_{adv} = E_x[\log D(x)] + E_{x_{rec}}[\log(1 - D(x_{rec}))]. \quad (6)$$

To further weaken identity information, the study proposes a contrastive identity suppression loss. This loss uses the center of the anonymized cluster as a positive anchor sample, while treating the original identity vector and other sample identity vectors as negative samples. This forces anonymized images to move away from their original identities and toward the neutral center in the identity space, as shown in Equation (7) [11, 14].

$$L_{CIS} = -\log \frac{\exp(sim(v,a)/\tau)}{\exp(sim(v,a)/\tau) + \exp(sim(v,u)/\tau) + \sum_{k=1}^{K}(sim(v,n_k)/\tau)}. \quad (7)$$

In Equation (7), $v$ represents the identity features of the anonymized image. $u$ denotes the original identity features. $a$ signifies the anonymized cluster center. $\{n_k\}$ indicates the random negative sample. $\tau$ represents the temperature coefficient. Additionally, it is necessary to ensure the stability of the anonymized image's attributes and pose. To this end, an attribute preservation constraint is introduced, as shown in Equation (8).

$$L_{att} = \left\| \Phi_{att}(x_{rec}) - \Phi_{att}(x) \right\|_1. \quad (8)$$

In Equation (8), $\Phi_{att}(\cdot)$ denotes the non-identity feature encoder, ensuring that the generated image matches the input image in non-identity features such as facial expressions and poses. Integrating the preceding sections, the generator's optimization objective is defined by Equation (9).

$$L_G = \lambda_{adv} L_{adv} + \lambda_{CIS} L_{CIS} + \lambda_{att} L_{att}. \quad (9)$$

In Equation (9), $\lambda_{adv}$, $\lambda_{CIS}$, and $\lambda_{att}$ are the weight parameters representing adversarial loss, identity suppression loss, and attribute retention loss, respectively. These parameters are used to balance the relative importance of different objectives during training.

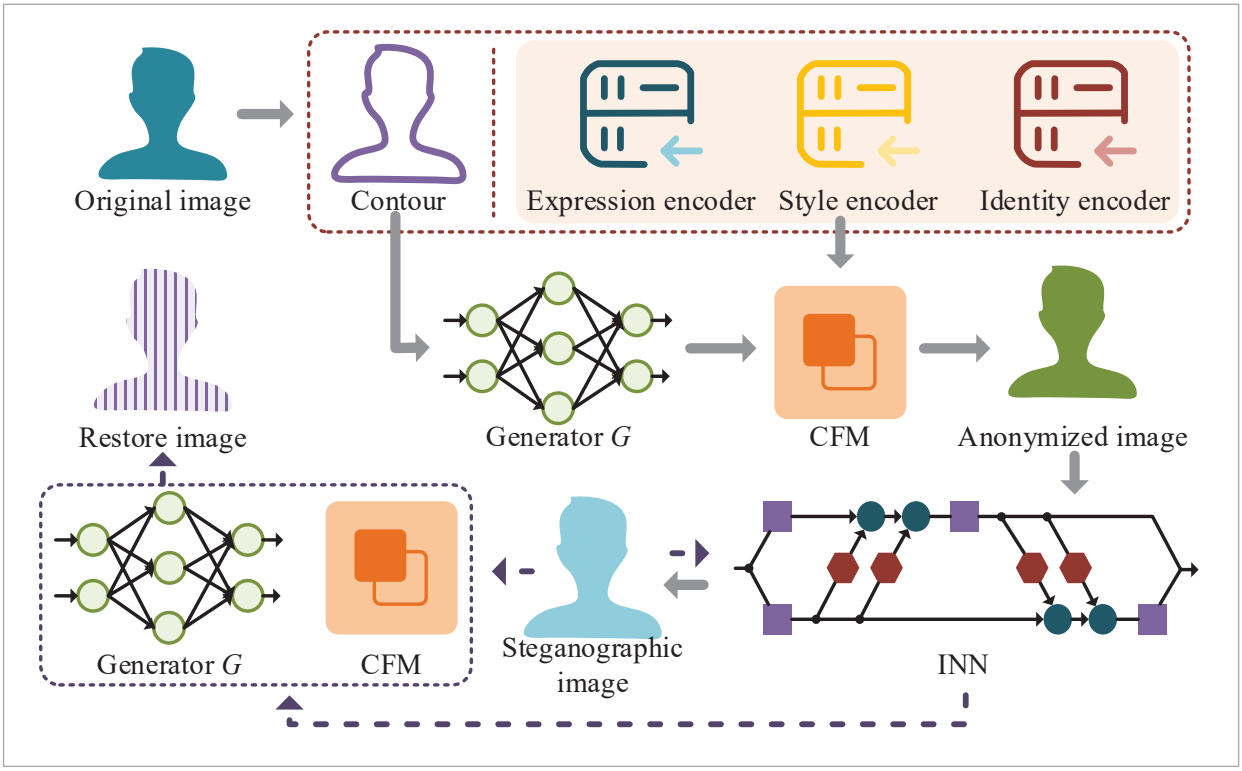
## 2.2. Reversible Face Privacy Protection and Anonymization Model for Multi-Condition Collaborative Control

To further enhance the model's flexibility and application value, and to endow the anonymization process with reversible recovery capabilities to strengthen privacy protection, this study introduces a multi-condition collaborative control mechanism based on MLFAd, constructing a reversible face anonymization model RFAMCC. Its overall architecture is shown in Figure 4.

As shown in Figure 4, after inputting the original face, semantic mask, and contour constraints, the multi-condition encoder extracts identity vector $z_{id}$, style vector $z_{style}$, and expression vector $z_{exp}$. The ISD module deflects $z_{id}$ on a controlled identity manifold to obtain the de-associated anonymous identity code $z_{anu}$, ensuring sufficient separation from the original identity while aligning with the neutral reference. The remaining condition vectors remain stable to preserve available attributes. Subsequently, the generator synthesizes anonymous faces under ($z_{anu}$, $z_{style}$, $z_{exp}$) 's multi-condition collaborative control. CFM performs geometric alignment and edge blending between the generated facial region and the original image background, ensuring scene consistency and detail continuity. To achieve reversibility, the reversible steganography mechanism embeds key recovery information into the anonymized result in an imperceptible manner, creating a steganographic image. In authorized scenarios, the same steganography network decodes the recovery information, which is then fed alongside ($z_{anu}$, $z_{style}$, $z_{exp}$) into the generator to reconstruct a high-fidelity original identity image.

The CFM module takes the generator-synthesized facial region and the original background as inputs,

**Figure 4**
RFAMCC overall framework.




utilizing a predicted fusion mask to achieve geometric alignment and edge transitions. This approach reduces artifacts and enhances visual coherence. The fusion process can be formalized as Equation (10).

$$x_{rec} = (1-\gamma)(x \otimes (1-G_M)) + \gamma(x_{gen} \otimes G_M)\,. \quad (10)$$

In Equation (10), $x$ denotes the original image. $x_{gen}$ represents the face region output by the generator. $G_M$ is the mask predicted by the generator. $\gamma$ is the fusion weight.

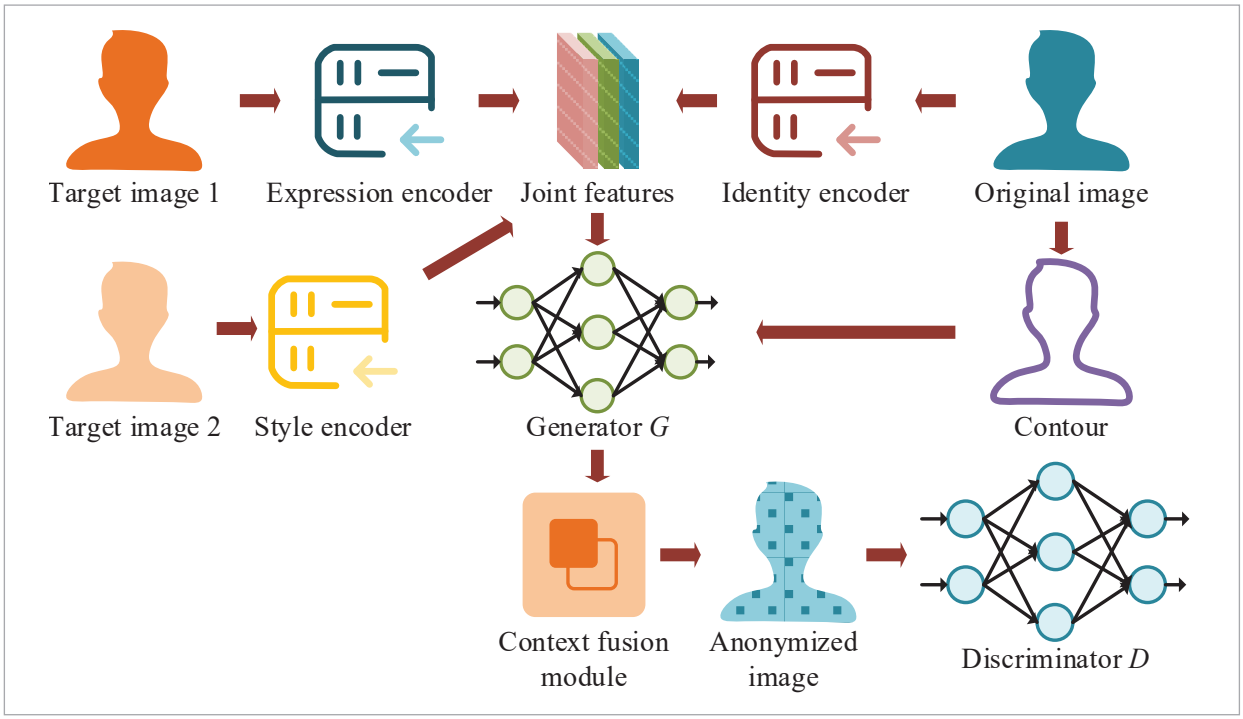
Furthermore, multi-condition collaborative control serves as one of the core components of RFAMCC, enabling the unification of identity anonymization and reversible restoration within the same framework. The overall workflow of multi-condition control learning is illustrated in Figure 5.

As shown in Figure 5, the original image first undergoes identity encoding by encoder $E_{id}$ to extract an identity vector, which is then anonymized via the ISD mechanism to generate an anonymous identity representation. Concurrently, the target image's expression and style information are extracted by $E_{exp}$ and $E_{style}$ respectively, ensuring the generated result inherits non-identity attributes. Subsequently, these conditional features are input alongside pose information into Generator $G$ to produce an anonymized face. The generated result is then combined with the original background via CFM to ensure overall scene consistency. Finally, the discriminator $D$ evaluates the authenticity of the generated image, thereby guiding the generator to continuously improve the visual quality and naturalness of the anonymized image. This joint modeling can be formalized as Equation (11).

$$x_{rec} = G(z_{anu}, z_{id}, z_{style}, z_{exp}, z_{pose}) \quad (11)$$

In Equation (11), $x_{rec}$ represents the anonymization result. Through this approach, the generator ensures the coordinated consistency between anonymization and multi-attribute preservation under multiple conditional constraints.

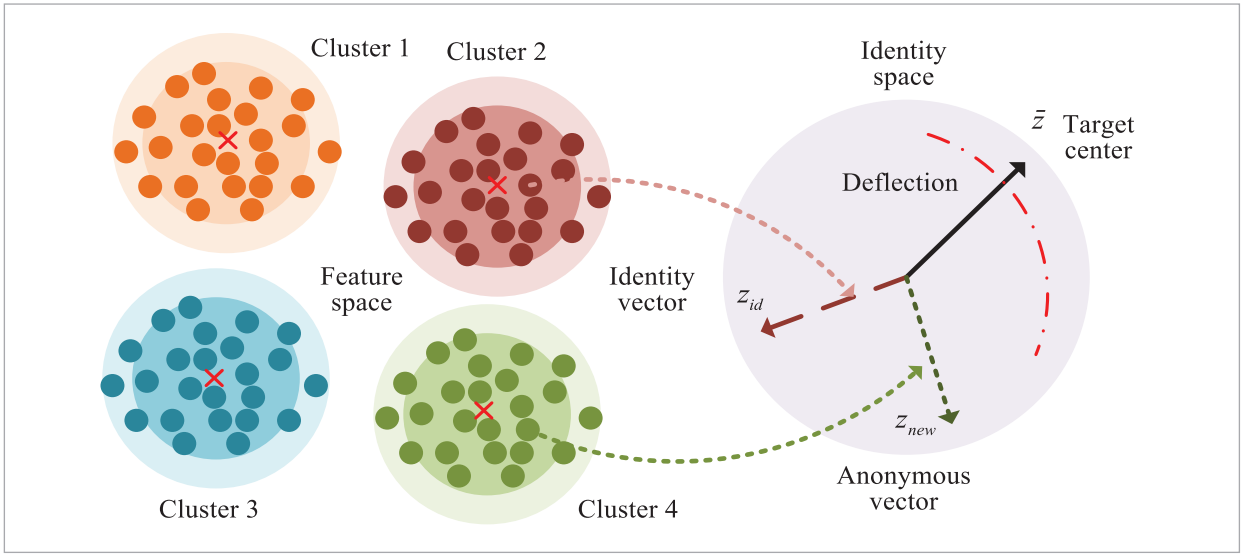
**Figure 5**
Overall flow chart of multi-condition collaborative control learning.




Furthermore, RFAMCC achieves identity obfuscation and unidentification by shifting the identity vector in the feature space from the original identity center to the anonymity cluster center, thus achieving privacy protection. The operation flow of the identity space shifting mechanism is shown in Figure 6.

As shown in Figure 6, the original identity feature vector $z_{id}$ first undergoes a clustering process to iden-

**Figure 6**
Schematic diagram of the de-identified identity feature reconstruction mechanism.



tify several candidate cluster centers $\{\bar{z}_1, \bar{z}_2, \ldots, \bar{z}_m\}$ in the feature space. Next, the Euclidean distances between $z_{id}$ and these cluster centers are calculated, and the cluster center $\bar{z}$ with the greatest distance is selected as the target direction for anonymization. Subsequently, the model combines the original vector with this target center through linear interpolation to obtain new identity features, as shown in Equation (12).

$$z_{new} = (1-\alpha)\cdot z_{id} + \alpha \cdot \bar{z}\,. \tag{12}$$

In Equation (12), $z_{new}$ denotes the anonymized identity feature. $\bar{z}$ represents the cluster center farthest from $z_{id}$. $\alpha \in [0,1]$ is the adjustment coefficient controlling the magnitude of feature offset. When $\alpha$ is small, the anonymized feature remains closer to the original identity. As $\alpha$ increases, the anonymized feature gradually approaches the anonymization cluster, thereby achieving stronger privacy protection.

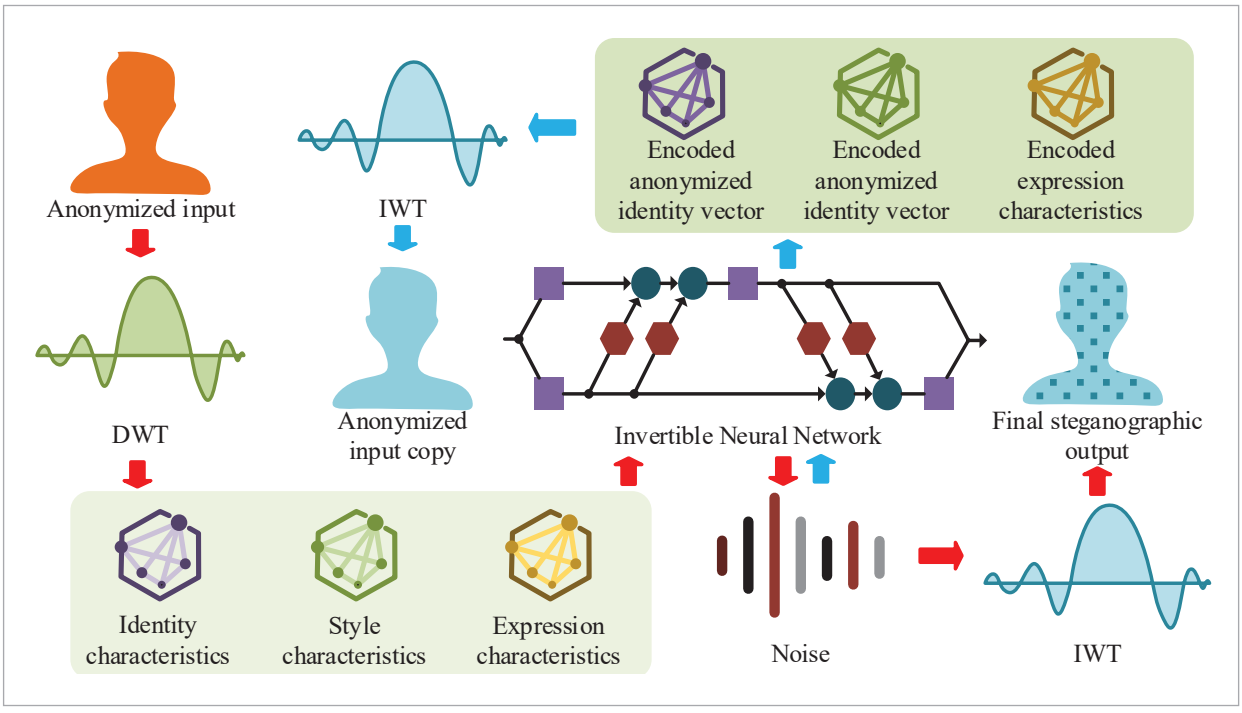
To ensure that the model has the ability to reversibly recover the original identity image from the anonymized result when authorized, RFAMCC introduces a mechanism based on feature steganography and recovery, the overall process of which is shown in Figure 7.

As shown in Figure 7, the original image is first decomposed into high-frequency and low-frequency components by discrete wavelet transform (DWT). Subsequently, the anonymized components $z_{id}$, $z_{exp}$, and $z_{style}$, along with the noise-encoded signal $\{z_{noise1}, z_{noise2}, z_{noise3}\}$, are input into the spatial subspace $INN(\cdot)$ to complete the information embedding. The $IWT$ then reconstructs the processed results back into the spatial domain, yielding an anonymized image containing the stashed information. This process can be formalized as Equation (13).

$$y = IWT(INN(DWT(x), z_{id}, z_{style}, z_{exp}, z_{noise1}, z_{noise2}, z_{noise3}))\,. \tag{13}$$

In Equation (13), $z_{exp}$ represents facial expression features, used to maintain natural consistency in facial expressions. $z_{style}$ denotes style features, ensuring coherence in texture and style of the generated results. $y$ indicates the anonymized result

**Figure 7**
Overall framework of FSR.



containing steganographic information. $INN(\cdot)$ is a reversible neural network. $DWT$ and $IWT$ denote the forward wavelet transform (FWT) and inverse wavelet transform (IWT), respectively, used to switch between the frequency domain and spatial domain. During forward steganography, the input vector undergoes progressive encoding into steganographic features through nonlinear mapping $\phi(\cdot)$, weight function $\rho(\cdot)$, and offset function $\eta(\cdot)$. During reverse recovery, these mapping functions are invoked in reverse order, ensuring a one-to-one correspondence between input and output. The computational rules for forward and reverse operations are detailed in Equation (14).

$$\begin{cases} y^{i+1} = y^i + \phi(x_i), \quad x^{i+1} = x^i \otimes \rho(y^{i+1}) + \eta(y^{i+1}) \\ x^i = x^{i+1} - \eta(y^{i+1}) / \eta(y^{i+1}), \quad y^i = y^{i+1} - \phi(x_i), \end{cases} . \tag{14}$$

In Equation (14), $x^i$ and $y^i$ denote the inputs to the reversible block at the layer. $\phi$, $\rho$, and $\eta$ represent learnable nonlinear transformation functions. $\otimes$ denotes element-wise operations. During model training, the study unifies anonymization effectiveness, image authenticity, attribute preservation, and reversible restoration into a single joint loss function. The overall objective function is expressed as Equation (15).

$$L = \lambda_{adv} L_{adv} + \lambda_{id} L_{id} + \lambda_{rec} L_{rec} . \tag{15}$$

In Equation (15), the adversarial loss $L_{adv}$ is used to enhance the realism of generated images, and its form is given by Equation (16).

$$L_{adv} = E_x[\log D(x)] + E_{x'}[\log(1 - D(x_{rec}))] . \tag{16}$$

In Equation (16), $x_{rec}$ denotes the generation of anonymous images. The identity separation loss $L_{id}$ is employed to reduce the similarity between the generated results and the original identity features, thereby achieving privacy protection, as shown in Equation (16).

$$L_{id} = 1 - \frac{z_{id} \cdot z_{anu}}{\|z_{id}\| \|z_{anu}\|} . \tag{17}$$

In Equation (17), $z_{anu}$ represents the anonymized identity vector after deflection. The reconstruction loss $L_{rec}$ constrains the effectiveness of the reversible recovery stage, requiring the reconstructed image $x'$ to be close to the original image $x$ in the pixel space, as shown in Equation (18).

$$L_{rec} = \|x - \hat{x}\|_1 . \tag{18}$$

By combining these three factors through weighted integration, the model achieves a balance during training between anonymization, realism, and reversible restoration.

To clearly describe how RFAMCC operates, the full anonymization–recovery pipeline can be summarized as follows:

- **Input decomposition:** The input face image and its semantic map are processed by the expression, style, and identity encoders to obtain the latent codes $z_{id}$, $z_{exp}$, and $z_{style}$.
- **Identity-space deflection:** The identity code $z_{id}$ is shifted toward a neutral cluster center to produce the anonymized identity representation $z_{anu}$.
- **Multi-condition control generation:** The generator takes ($z_{anu}$, $z_{style}$, and $z_{exp}$) and synthesizes the anonymized face.
- **Context fusion:** The CFM module geometrically aligns and seamlessly blends the synthesized face region with the original background, producing the final anonymized image.
- **Steganographic embedding:** The reversible information necessary for identity restoration is embedded in the anonymized image using the INN-based feature steganography module.
- **Reversible restoration (authorized scenario):** The steganographic image is passed through the inverse INN to extract the hidden features and, together with ($z_{style}$, $z_{exp}$), reconstruct the original identity face.

## 3. Results and Analysis

### 3.1. RFAMCC Basic Performance Testing

The experimental setup is built using Python 3.10 and the PyTorch framework. The runtime environment comprises an Intel Core i7-12700K processor,

32GB of RAM, and the Ubuntu 22.04 operating system. Model training and inference are performed on an NVIDIA RTX 3090 GPU. The experiment adopts two publicly available facial datasets: CelebA-HQ (a high-resolution facial image dataset containing over 30,000 diverse facial samples covering attributes such as gender, expression, and pose) and LFW (a widely used benchmark dataset for facial verification and recognition, comprising over 13,000 unconstrained facial images). The dataset is divided into 70% for training, 10% for validation, and 20% for testing. All images are resized to 256×256 pixels and enhanced through random cropping, horizontal flipping, and color jittering. The comparison methods encompass three representative advanced techniques: diffusion-based face privacy protection (Diff-Privacy), seeing is not believing: an identity hider (IH), and one person one mask (OPOM). To ensure comparability of results, all are evaluated using the same pre-trained recognizer and data partition.

To ensure reproducibility, complete training hyperparameters and dataset partitioning are provided. All experiments use the Adam optimizer with a learning rate of $1\times10^{-4}$ (linearly decaying to $1\times10^{-8}$), a batch size of 16, and 120 training epochs. The loss weights are set to $\lambda_{adv}$=1.0, $\lambda_{CIS}$=0.7, and $\lambda_{att}$=0.5. To analyze the performance of the proposed RFAMCC model on key parameters, parameter sensitivity experiments are conducted first. The effects of identity deviation coefficient $\alpha$, number of cluster centers $k$, reversible reconstruction weight $\lambda_{rec}$, and context fusion weight $\gamma$ on privacy and perception quality are examined separately to determine default parameter configurations for subsequent scenario experiments. Results are shown in Table 1. The ablation study aims to quantify the contribution of each core module (MLFD, ISD, CFM, and FSR) to privacy suppression and utility preservation. TPR@FPR=1e−4 is used as the privacy metric, Attr-Acc reflects attribute consistency, and the number of parameters (Params) indicates the model's structural complexity.

**Table 1**
Parameter sensitivity experiment (L16 orthogonal combination).

| Run | $\alpha$ | $k$ | $\lambda_{rec}$ | $\gamma$ | EER (%) ↓ | LPIPS ↓ |
|---|---|---|---|---|---|---|
| 1 | 0.3 | 16 | 0.3 | 0.3 | 8.35±0.18 | 0.280±0.006 |
| 2 | 0.3 | 32 | 0.5 | 0.5 | 8.15±0.16 | 0.255±0.005 |
| 3 | 0.3 | 64 | 0.7 | 0.7 | 8.25±0.17 | 0.251±0.006 |
| 4 | 0.3 | 96 | 1.0 | 0.9 | 8.75±0.19 | 0.267±0.006 |
| 5 | 0.5 | 16 | 0.5 | 0.7 | 6.80±0.15 | 0.275±0.005 |
| 6 | 0.5 | 32 | 0.7 | 0.9 | 6.55±0.14 | 0.250±0.006 |
| 7 | 0.5 | 64 | 1.0 | 0.3 | 6.45±0.14 | 0.223±0.005 |
| 8 | 0.5 | 96 | 0.3 | 0.5 | 6.50±0.13 | 0.265±0.006 |
| 9 | 0.7 | 16 | 0.7 | 0.5 | 5.75±0.12 | 0.258±0.005 |
| 10 | 0.7 | 32 | 1.0 | 0.7 | 5.50±0.12 | 0.230±0.005 |
| 11 | 0.7 | 64 | 0.3 | 0.9 | 5.15±0.11 | 0.255±0.006 |
| 12 | 0.7 | 96 | 0.5 | 0.3 | 5.50±0.12 | 0.250±0.005 |
| 13 | 0.9 | 16 | 1.0 | 0.9 | 5.25±0.11 | 0.307±0.007 |
| 14 | 0.9 | 32 | 0.3 | 0.3 | 4.35±0.10 | 0.285±0.006 |
| 15 | 0.9 | 64 | 0.5 | 0.5 | 4.50±0.10 | 0.283±0.006 |
| 16 | 0.9 | 96 | 0.7 | 0.7 | 5.00±0.11 | 0.298±0.006 |

As shown in Table 1, with increasing identity shift strength $\alpha$, the equal error rate (EER) gradually decreases, indicating more pronounced anonymization effects. However, the learned perceptual image patch similarity (LPIPS) slightly increases, suggesting that image naturalness is somewhat affected. The number of cluster centers $k$ within the range of 32 to 64 effectively balances privacy and visual quality. The reversible reconstruction weight $\lambda_{rec}$ between 0.5 and 0.7 optimally balances restoration accuracy and robustness. Meanwhile, the context fusion weight $\gamma$ within the 0.5-0.7 range effectively preserves background coherence. After comprehensive evaluation, the final experimental parameter combination adopted is: identity shift strength $\alpha = 0.7$, number of cluster centers $k = 32$, reversible reconstruction weight $\lambda_{rec} = 0.5$, and context fusion weight $\gamma = 0.5$.

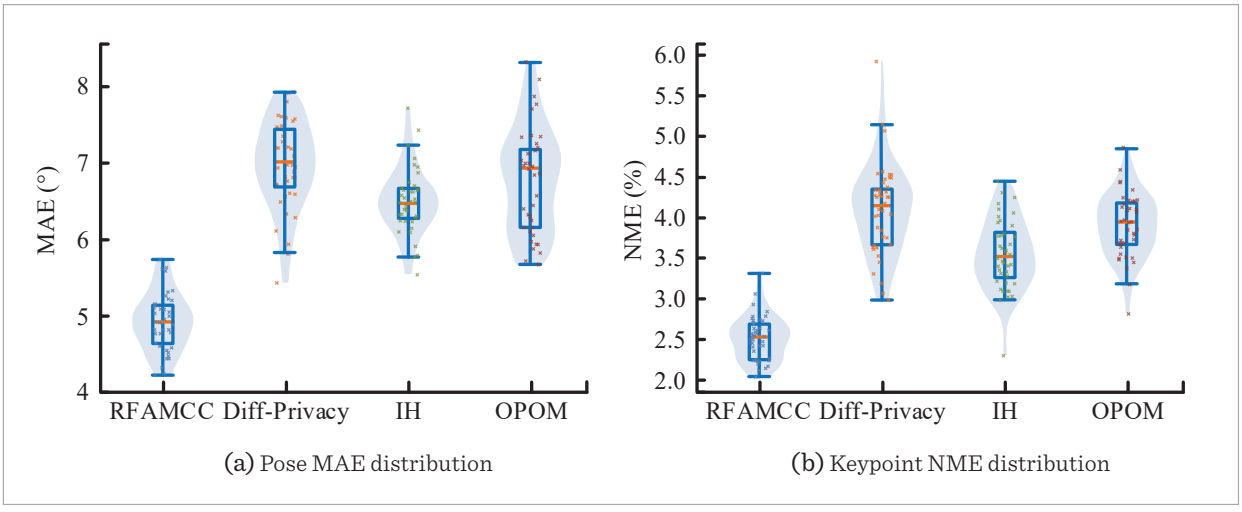
Subsequently, this experiment examines geometric stability by evaluating pose consistency (MAE) and landmark accuracy (NME), aiming to assess robustness under structural variations before and after anonymization. The experimental results are shown in Figure 8.

Figure 8(a) shows that RFAMCC exhibits the most concentrated distribution in terms of mean absolute error (MAE) for head pose. Its average value is approximately 4.9°, outperforming Diff-Privacy (7.0°), IH (6.5°), and OPOM (6.9°). In Figure 8(b), RFAMCC exhibits a mean normalized mean error (NME) of approximately 2.5%, outperforming Diff-Privacy (4.2%), IH (3.5%), and OPOM (3.9%). This demonstrates its superiority in preserving facial expressions and local geometric details. Through multi-condition collaborative control, RFAMCC achieves joint modeling of identity deflection, context fusion, and feature steganography, enhancing identity concealment while constraining geometric and textural consistency during the generation phase. In contrast, Diff-Privacy, though utilizing a diffusion model, lacks explicit constraints on pose preservation and is prone to deviation at large angles. IH employs latent space substitution for appearance transformation, but keypoint regions remain difficult to fully align. OPOM primarily focuses on identity feature perturbation, exhibiting slightly weaker performance in local geometric consistency.

The performance of different methods in reversible restoration and identity reconstruction is then further evaluated. The experiment is designed to progressively increase the steganography embedding rate to 0%, 10%, 20%, 30%, 40%, and 50%, and to calculate the PSNR and identity reconstruction similarity (ID-Sim, cosine) of the facial region in the restored image. PSNR is used to measure restoration fidelity, and ID-Sim is used to measure identity consistency. All results in Figure 9 are expressed as the mean±standard deviation of five random runs, with error bars representing the standard deviation.

**Figure 8**

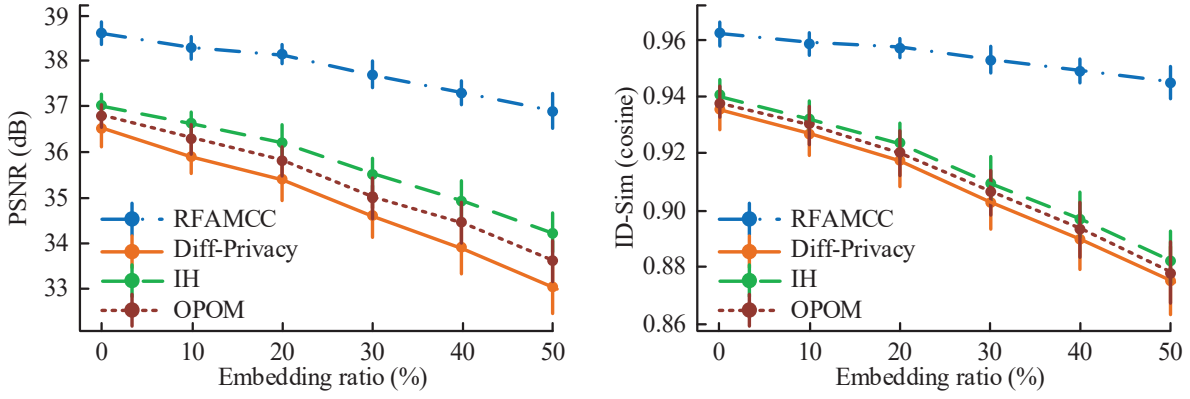
Results of pose and structure robustness.




(a) Pose MAE distribution

(b) Keypoint NME distribution

**Figure 9**
Results of reversible recovery quality and identity restoration.




(a) PSNR under different embedding ratios

(b) ID-Sim under different embedding ratios

In Figure 9(a), when the steganographic embedding rate reaches 50%, RFAMCC achieves a PSNR of 37.0±0.7 dB, outperforming Diff-Privacy (33.1±0.9 dB), IH (34.0±0.8 dB), and OPOM (33.6±0.78 dB). This demonstrates its ability to maintain high image quality even at high embedding intensities. In Figure 9(b), under identical conditions, RFAMCC achieves an ID-Sim of 0.948±0.005, while Diff-Privacy, IH, and OPOM yield values of 0.884±0.009, 0.889±0.007, and 0.887±0.008, respectively. This demonstrates RFAMCC's superiority in preserving identity consistency. The identity feature deflection and feature steganography recovery mechanisms introduced by RFAMCC strike a balance between anonymization and reversible reconstruction under multi-condition collaborative control. In contrast, competing methods exhibit declining performance with increasing embedding rates when lacking reversible constraints or sufficient steganographic robustness. A paired t-test shows that the improvements of RFAMCC over Diff-Privacy, IH, and OPOM are statistically significant ($p<0.05$).

To quantify the contribution of each core mechanism to model performance, mechanism ablation

**Table 2**
Ablation experiment results.

| Method setting | TPR@FPR=1e-4 (%) ↓ | Attr-Acc (%) ↑ | Params (M) ↓ |
|---|---|---|---|
| RFAMCC (Full) | 7.8 | 92.3 | 52.1 |
| -MLFD (w/o multi-level feature-driven) | 10.9 | 89.7 | 47.8 |
| -ISD (w/o identity space deflection) | 18.9 | 92.1 | 49.6 |
| -CFM (w/o CFM) | 11.6 | 90.2 | 50.4 |
| -FSR (w/o FSR) | 9.8 | 91.0 | 48.9 |
| -MLFD-ISD | 23.5 | 88.9 | 45.3 |
| -MLFD-CFM | 15.2 | 87.8 | 46.1 |
| -ISD-FSR | 20.1 | 90.5 | 46.7 |
| -CFM-FSR | 12.9 | 89.6 | 46.3 |

experiments are designed. The four modules, MLFD, ISD, context-fused CFM, and FSR, are removed individually, with partial combination ablations also conducted. Changes in privacy metrics (true positive rate at false positive rate=1e-4, TPR@FPR=1e-4), attribute accuracy (Attr-Acc), and model parameter count are examined. The experimental results are shown in Table 2.

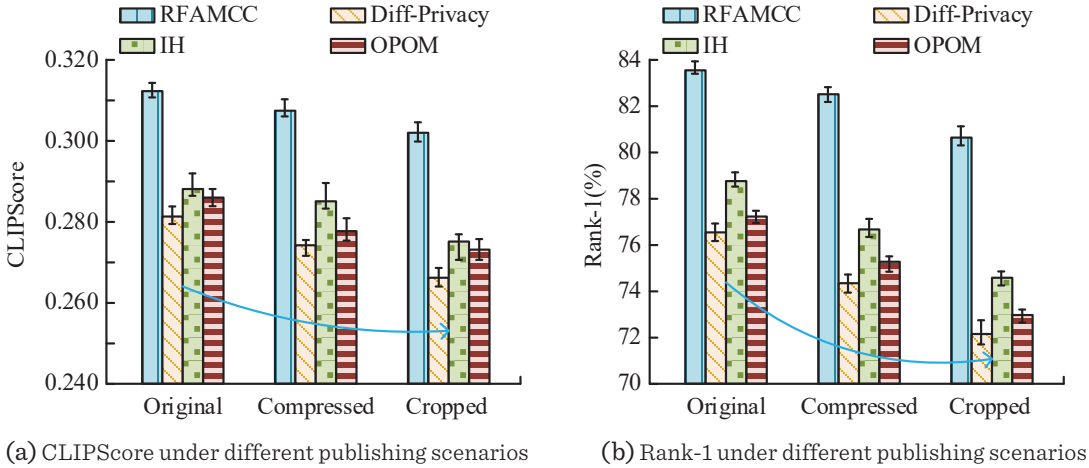
As shown in Table 2, the complete RFAMCC achieves the best balance between privacy protection and attribute preservation (TPR@FPR=1e−4 is 7.8%, Attr-Acc is 92.3%). Removing the ISD module increases the TPR to 18.9%. This indicates that identity space deflection plays a key role in weakening the recognizability of deep identities. When this mechanism is absent, the anonymous identity vector cannot sufficiently deviate from the original identity center, leading to a sharp increase in recognition risk. Decreases in Attr-Acc of 89.7% and 90.2% are observed when MLFD and CFM are removed, respectively. This indicates that multi-layer feature modeling and context fusion are crucial for maintaining consistency between expression, pose, and background. The absence of MLFD leads to single-scale bias in structure and texture, while the absence of CFM produces artifacts at the fusion boundary, thus affecting attribute preservation. Although removing FSR has a smaller impact on TPR (9.8%), it slightly reduces attribute stability. This reflects the fact that its main role is reversible recovery rather than privacy suppression.

## 3.2. RFAMCC Practical Application Test Results

To validate the adaptability and practical applicability of the proposed model in real-world scenarios, experiments will be conducted across four key application dimensions: social media publishing, surveillance scenarios, cross-domain generalization, and attack defense. These diverse scenarios not only address the preservation of image quality and semantic representation but also evaluate privacy robustness and identity reversibility under conditions involving multi-source data, complex environments, and adversarial threats. First, in real-world social media scenarios, users often seek to protect their personal identity privacy when posting images while simultaneously demanding high-quality semantic expression and visual consistency in their content. Therefore, the experiment simulates the posting environment using a facial data subset. It employs a cross-modal semantic matching model to compute the semantic relevance score (contrastive language-image pretraining score, CLIPScore), while simultaneously testing the retrieval accuracy (rank-1 accuracy, Rank-1) of anonymized images based on a standard face retrieval framework. CLIPScore measures semantic consistency and uses Rank-1 accuracy to measure retrieval utility. The experimental results are shown in Figure 10.

Figure 10(a) shows that under the CLIPScore metric, RFAMCC achieves scores of 0.312, 0.308, and

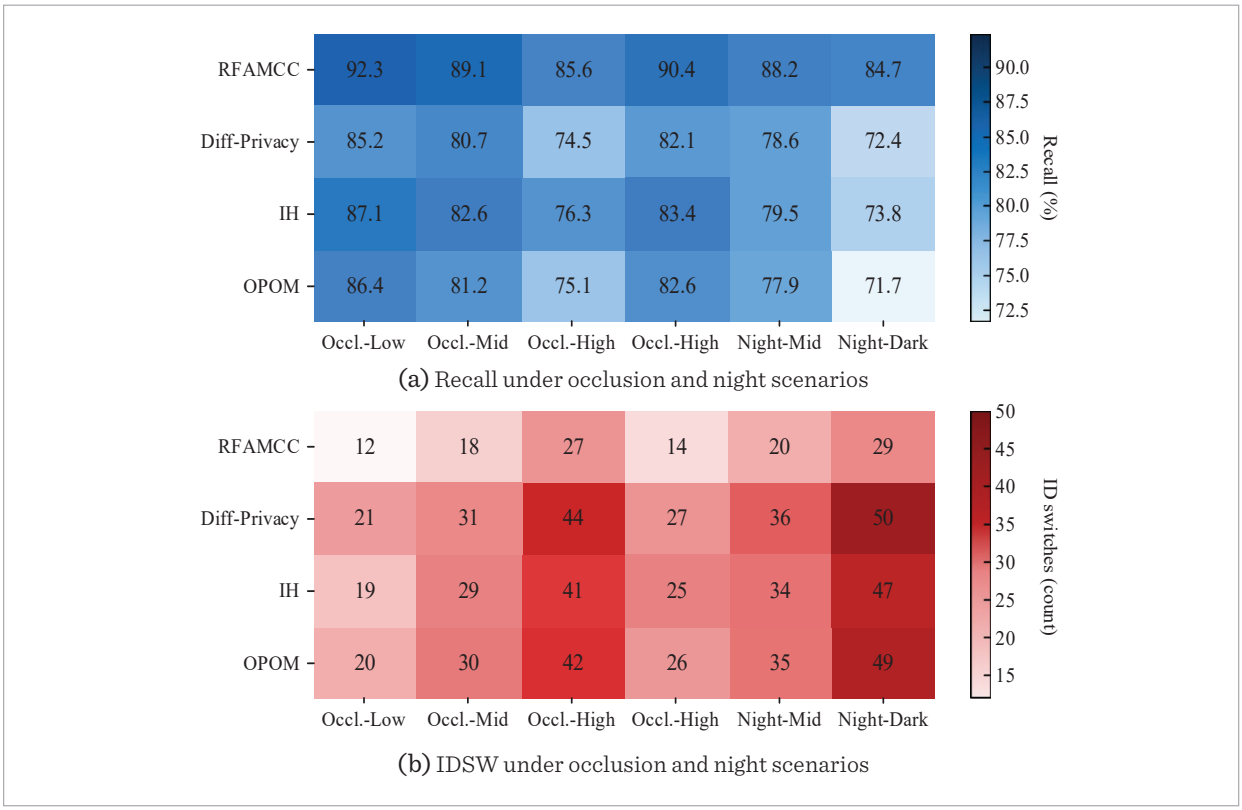
**Figure 10**
Results in social media publishing scenario.




(a) CLIPScore under different publishing scenarios

(b) Rank-1 under different publishing scenarios

**Figure 11**
Performance in monitoring scenarios (occlusion and night conditions).




0.301 in the original, compressed, and cropped scenarios, respectively. In contrast, Diff-Privacy scores only 0.281, 0.274, and 0.266. IH scores 0.289, 0.283, and 0.275. OPOM scored 0.286, 0.279, and 0.272. Figure 10(b) shows the Rank-1 performance for face retrieval. RFAMCC achieves 83.7%, 82.4%, and 80.9% across the three scenarios, outperforming Diff-Privacy's 76.2%, 74.5%, and 72.1%. The results demonstrate that RFAMCC significantly outperforms the comparison methods in preserving image semantic consistency and maintaining identity retrieval usability. The comparison methods often sacrifice semantic integrity or retrieval accuracy under strong anonymization constraints.

In complex surveillance environments, facial detection and tracking performance often deteriorates due to occlusions or insufficient lighting. To validate the proposed model's robustness in such scenarios, experiments evaluate the performance of various methods across different occlusion levels (low, medium, high) and nighttime lighting conditions (bright, medium, dim) by comparing metrics including detection recall and identity switches (IDSW). Experimental results are shown in Figure 11.

In Figure 11(a), RFAMCC achieves a Recall of 85.6% under high occlusion conditions and 84.7% in dim nighttime scenes, outperforming Diff-Privacy (74.5%, 72.4%), IH (76.3%, 73.8%), and OPOM (75.1%, 71.7%) across all scenarios. RFAMCC maintains high detection performance even under complex lighting and severe occlusion. Figure 11(b) shows that RFAMCC achieves an IDSW of 29 in dim nighttime scenes, outperforming Diff-Privacy (50), IH (47), and OPOM (49) while demonstrating superiority in maintaining tracking chains and suppressing identity switching. While Diff-Privacy maintains acceptable recall under mild occlusions, its performance degrades significantly under strong occlusions and low-light conditions. IH and OPOM introduce geometric structural bias during identity perturbation, leading to reduced tracking stability.

To evaluate the adaptability of the proposed model across heterogeneous domains, two transfer directions are set: CelebA→VGGFace2 and VGGFace2→CelebA. After training is completed in the source domain, the model is directly tested on the target domain without additional fine-tuning. Considering that compressed distortion may occur in facial images during practical applications, experiments are conducted under both Clean (original resolution without compression) and Compressed (JPEG Q=50) conditions. Two distribution consistency metrics, Fréchet inception distance (FID) and kernel inception distance (KID), are adopted as evaluation criteria. Lower values indicate that the generated anonymized images are closer to the target domain distribution. Experimental results are shown in Table 3.

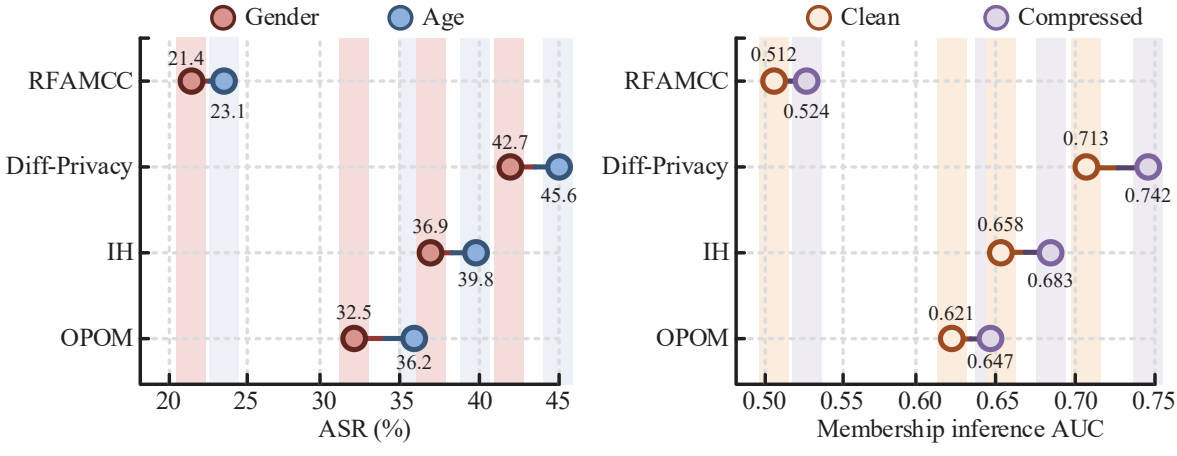
In Table 3, RFAMCC consistently achieves the best performance across cross-domain tasks. Under the VGGFace2→CelebA (Compressed) condition, RFAMCC achieves an FID of 19.0±0.6 and a KID of 0.013±0.001, surpassing Diff-Privacy's 27.2±0.9 and 0.023±0.002, IH's 23.6±0.7 and 0.019±0.002, and OPOM's 24.9±0.7 and 0.020±0.002, demonstrating its robustness under compressed degradation scenarios. Further comparison reveals that Diff-Privacy exhibits significant cross-domain distribution shifts under strong anonymization perturbations, leading to markedly elevated FID and KID metrics. While IH and OPOM demonstrate strong local texture preservation, their lack of contextual fusion and reversible information constraints results in insufficient cross-domain consistency. In contrast, RFAMCC employs a multi-condition collaborative control mechanism to simultaneously achieve identity feature deflection and semantic attribute preservation during anonymization. This enables generated results to closely align with the target domain's distribution, demonstrating the model's robust generalization capability.

**Table 3**

Cross-domain generalization results (mean±std.; ↓ lower is better).

| Transfer direction & condition | Method | FID ↓ | KID ↓ |
|---|---|---|---|
| CelebA→VGGFace2, Clean | RFAMCC | 18.7±0.6 | 0.012±0.001 |
| | Diff-Privacy | 26.4±0.8 | 0.020±0.002 |
| | IH | 23.1±0.7 | 0.017±0.001 |
| | OPOM | 24.2±0.7 | 0.018±0.001 |
| CelebA→VGGFace2, Compressed | RFAMCC | 21.1±0.7 | 0.015±0.001 |
| | Diff-Privacy | 30.3±1.0 | 0.026±0.003 |
| | IH | 26.1±0.8 | 0.021±0.002 |
| | OPOM | 27.6±0.8 | 0.022±0.002 |
| VGGFace2→CelebA, Clean | RFAMCC | 16.9±0.5 | 0.010±0.001 |
| | Diff-Privacy | 23.5±0.7 | 0.018±0.002 |
| | IH | 21.0±0.6 | 0.015±0.001 |
| | OPOM | 22.1±0.6 | 0.016±0.001 |
| VGGFace2→CelebA, Compressed | RFAMCC | 19.0±0.6 | 0.013±0.001 |
| | Diff-Privacy | 27.2±0.9 | 0.023±0.002 |
| | IH | 23.6±0.7 | 0.019±0.002 |
| | OPOM | 24.9±0.7 | 0.020±0.002 |

**Figure 12**
Attack and security evaluation results.




(a) ASR across Gender and Age

(b) Membership inference attack under different conditions

To comprehensively validate the robustness and security of the proposed model under malicious attack scenarios, two types of tests are designed: attribute attacks and membership inference attacks. Attribute attacks primarily examine the model's detectability in predicting gender and age, with the anonymized image privacy leakage risk measured by the statistical attribute attack success rate (ASR). Membership inference attacks are conducted under both clean and compressed conditions, utilizing the membership inference attack area under curve (AUC) to evaluate the model's resilience against training data leakage. Experimental results are shown in Figure 12.

In Figure 12(a), during gender prediction attacks, RFAMCC achieves an ASR of 21.4%, lower than Diff-Privacy's 32.5%, IH's 36.9%, and OPOM's 42.7%. In age prediction attacks, RFAMCC achieves an ASR of 23.1%, while Diff-Privacy, IH, and OPOM achieve 36.2%, 39.8%, and 45.6%, respectively. Figure 12(b) shows that RFAMCC achieves an AUC of only 0.512 (Clean) and 0.524 (Compressed) under member inference attacks, approaching the random guessing level (0.5), while demonstrating greater robustness against training sample inference attacks. RFAMCC's joint modeling of identity feature deflection and feature steganography recovery mechanisms weakens exploitable signals from attribute leakage. By contrast, other methods still have shortcomings in terms of steganographic embedding, feature separation, and style preservation. This makes residual privacy information more susceptible to being captured by adversarial models.

The study conducts a subjective visualization experiment on the anonymization effect on the VGGFace2 dataset (https://www.robots.ox.ac.uk/~vgg/data/vgg_face2/). VGGFace2 has higher identity diversity and wider racial distribution. It can be used to evaluate the generalization and fairness performance of the model in real population structures. Two recent representative techniques are selected for comparison: one is Mask-based adversarial anonymization [22], which generates face masks that can hide identities through adversarial perturbation. The other is DIFP [28], which uses a diffusion model to achieve a balance between privacy and realism. The results are shown in Figure 13.

Figure 13 illustrates the anonymization effects of the three methods on the VGGFace2 samples. While the mask-based method significantly destroys identity features, it relies on adversarial perturbations in local regions, which often result in blocky artifacts and texture breaks around key facial structures. This leads to a decrease in overall naturalness. The DIFP's anonymization results, which are generated based on a diffusion model, are relatively smooth. However, the diffusion resampling process weakens

**Figure 13**
Qualitative evaluation of identity suppression and visual consistency across anonymization methods on VGGFace2.


high-frequency details, resulting in slight deformations in areas such as the nose and corners of the mouth. Additionally, there is insufficient consistency in facial expressions. By contrast, RFAMCC uses multi-condition collaborative control to effectively deflect identity-related features while maintaining controllable consistency in multi-scale textures and geometric relationships. This process jointly utilizes MLFD, ISD, CFM, and FSR modules. Its anonymization results maintain stable structural coherence at facial boundaries and hairlines, as well as at local expressions. There are natural transitions to local perturbations, and the results are less prone to smooth transitions like those of the diffusion model or abrupt changes like those of adversarial perturbations. Overall, RFAMCC strikes a better balance between "de-identification strength" and "visual usability." It demonstrates higher-quality anonymization and greater practical feasibility.

## 4. Discussion

To address the shortcomings of existing facial anonymization models in terms of identity hiding, visual consistency, and cross-domain generalization, this study designed a reversible anonymization model, RFAMCC. The mask-based adversarial anonymization method proposed by Wang et al. [22] primarily used local occlusion perturbation to hide identities. However, this method easily produced visible artifacts and destroys facial expressions and local geometry. Although the diffusion-based DIFP method could generate relatively smooth anonymized images, it still degraded the original facial expressions, key point relationships, and high-frequency textures compared with the diffusion-based DIFP method [28]. Basic performance experiments showed that, with a steganalysis embedding rate of 50%, RFAMCC achieved the best reversible recovery quality, with a PSNR of 37.0 dB and an ID-Sim value that surpassed all baseline methods. This indicated stronger robustness in maintaining identity consistency reconstruction under reversible constraints. In social media post scenarios, RFAMCC achieved CLIPScores of 0.312, 0.308, and 0.301 under raw, compressed, and cropped conditions, respectively. Its face retrieval Rank-1 scores reached 83.7%, 82.4%, and 80.9%, demonstrating good semantic preservation capabilities and usability for downstream tasks after anonymization. Furthermore, the visualization experiments in Figure 13 showed the anonymization effects of RFAMCC, mask-based anonymization, and DIFP on samples with different identities.

Although the RFAMCC model was relatively complex due to its multi-module structure, the inference overhead remained within an acceptable range of approximately 41 milliseconds per frame on modern GPUs, supporting its deployment in near real-time applications. It was worth noting that CelebA-HQ and LFW had inherent biases regarding gender and race. This study addressed this issue by introducing a VGGFace2 subset for supplementary experiments. However, cross-race fairness was still an area for future research. In addition, all quantitative results were reported as mean±standard deviation from multiple random runs, thereby improving the statistical reliability of the comparisons. Slight failures were also observed under extreme lighting or severe occlusion conditions, where CFM occasionally produced minor fusion artifacts or incomplete background fusion. This indicates the model's limitations under severe perturbations. Furthermore, the results revealed an inherent trade-off between the strength of the steganalytic embedding and visual distortion. Stronger embedding can improve the fidelity of reversible recovery, but excessively high embedding rates may introduce inconsistencies in the texture of edges in challenging regions.

## 5. Summary and Future Work

To address the shortcomings of existing face anonymization methods, such as insufficient balance between identity concealment and attribute preservation, lack of reversibility and recovery capabilities, and limited cross-domain generalization and security, this study proposes the RFAMCC face anonymization model. Experimental results show that RFAMCC can provide a controllable solution for face image privacy protection and compliant applications.

Although RFAMCC demonstrates strong anonymization performance and reversible recovery, several limitations remain. First, steganographic embedding involves an inherent trade-off: higher embedding improves recovery fidelity, but it may also introduce local texture distortion. This suggests the need for more adaptive embedding strategies. Second, this study focuses on static images. Extending RFAMCC to video requires addressing temporal consistency and preventing identity drift between frames. Third, practical deployment in mobile or surveillance systems requires further optimization of latency and model size. It also requires potential integration with federated learning or blockchain-based authorization to ensure secure, auditable identity recovery. Dataset bias remains a concern. CelebA-HQ and LFW lack demographic diversity, so broader, cross-ethnic evaluations are needed. Future work may explore combining RFAMCC with diffusion-based anonymization frameworks to strengthen ethical safeguards and ensure the responsible use of reversible anonymization technology.

### Conflicts of Interest

The authors declare that they have no conflicts of interest.

## List of Parameters

$s_{face}$: Latent feature vector of the facial region

$s_{bg}$: Latent feature vector of the background region

$s_{hair}$: Latent feature vector of the hair region

$s_{body}$: Latent feature vector of the body region

$s$: Comprehensive feature vector after concatenating features from the facial, background, hair, and body regions by channel

$x_{land}$: Contour map

$E_{pose}(\cdot)$: Pose encoder

$x_m$: Semantic mask image, used to provide the semantic distribution of the main facial regions

$x_{land}$: Facial keypoints and contour map, used to provide local geometric constraints

$z_{pose}$: Pose latent vector, representing the geometric pose information of the face

$G(\cdot)$: Generator

$x_{rec}$: Reconstructed anonymized face image

$D(\cdot)$: Discriminator

$f_i$: Feature vectors at different scales

$c_{j(i)}$: Center vector of the feature cluster

$f_i^{anon}$: Anonymized feature representation

$v$: Identity feature vector of the anonymized image

$u$: Original identity features

$k$: Number of cluster centers (range ≈ 16–96 )

$a$: Anonymous cluster center (range ≈ 0.3–0.9 )

$\{n_k\}$: Identity features of random negative samples

$\tau$: Temperature coefficient

$\Phi_{att}(\cdot)$: Non-identity feature encoder

$\lambda_{adv}$: Weight coefficient of adversarial loss

$\lambda_{rec}$: Reversible reconstruction weights (range ≈ 0.3–0.7)

$\lambda_{CIS}$: Weight coefficient of identity suppression loss

$\lambda_{att}$: Weight coefficient of attribute preservation loss

$z_{id}$: Identity vector

$z_{style}$: Style vector

$z_{exp}$: Expression vector

$z_{anu}$: Anonymous ID

$x$: Original Image

$x_{gen}$: Face Region Output by Generator

$G_M$: Mask Predicted by Generator

$\gamma$: Fusing Weights (range ≈ 0.3–0.7)

$E_{id}$: Identity Encoder

$E_{exp}$: Expression Information Provided by Target Image

$E_{style}$: Style Information Provided by Target Image
$\bar{z}$: Farthest Cluster Center
$z_{new}$: Anonymized Identity Features
$\alpha \in [0,1]$: Adjustment Coefficient
$DWT$: Forward Wavelet Transform
$IWT$: Inverse Wavelet Transform
$y$: Steganalysis Image
$\hat{x}$: Reconstructed Original Identity Image
$\phi(\cdot)$: Nonlinear Mapping
$\rho(\cdot)$: Weight Function
$\eta(\cdot)$: Offset Function
$x^i$: Input of the i-th level reversible block
$\phi, \rho, \eta$: Learnable Nonlinear Transformation Function
$L_{adv}$: Adversarial Loss (range ≈ 0.1–2.0 during training)
$L_{id}$: Identity Separation Loss (range ≈ 0.01–1.0)
$L_{rec}$: Reconstruction Loss (range ≈ 0.001–0.5)
$x'$: Reconstructed Image

## References


1. Chiphiko, B. A., Kim, H., Ali, P., Eneya, L. Forward Secrecy Attack on Privacy-Preserving Machine Authenticated Key Agreement for Internet of Things. Archives of Advanced Engineering Science, 2025, 3(1), 29-34. https://doi.org/10.47852/bonviewAAES32021937
2. Ghani, M. A. N. U., She, K., Rauf, M. A., Khan, S., Khan, J. A., Aldakheel, E. A., Khafaga, D. S. Enhancing Security and Privacy in Distributed Face Recognition Systems Through Blockchain and GAN Technologies. Computers, Materials & Continua, 2024, 79(2), 2609-2623. https://doi.org/10.32604/cmc.2024.049611
3. Golchha, R., Verma, G. K. Leveraging Quantum Computing for Synthetic Image Generation and Recognition with Generative Adversarial Networks and Convolutional Neural Networks. International Journal of Information Technology, 2024, 16(5), 3149-3162. https://doi.org/10.1007/s41870-024-01835-9
4. He, X., Zhu, M., Chen, D., Wang, N., Gao, X. Diff-Privacy: Diffusion-Based Face Privacy Protection. IEEE Transactions on Circuits and Systems for Video Technology, 2024, 34(12), 13164-13176. https://doi.org/10.1109/TCSVT.2024.3449290
5. He, Y., Ouyang, W., Li, S., Wang, L., Zhou, J., Su, W., Ren, Y. Cloud Computing Data Privacy Protection Method Based on Blockchain. International Journal of Grid and Utility Computing, 2023, 14(5), 480-492. https://doi.org/10.1504/IJGUC.2023.133436
6. Jiang, Z., Zeng, W., Zhou, X., Chen, P., Yin, S. Transferable Sparse Adversarial Attack on Modulation Recognition with Generative Networks. IEEE Communications Letters, 2024, 28(5), 999-1003. https://doi.org/10.1109/LCOMM.2024.3373222
7. Kavoliūnaitė-Ragauskienė, E. Right to Privacy and Data Protection Concerns Raised by the Development and Usage of Face Recognition Technologies in the European Union. Journal of Human Rights Practice, 2024, 16(2), 658-674. https://doi.org/10.1093/jhuman/huad065
8. Khan, W., Topham, L., Khayam, U., Ortega-Martorell, S., Heather, P., Ansell, D., Hussain, A. Person De-Identification: A Comprehensive Review of Methods, Datasets, Applications, and Ethical Aspects Along-With New Dimensions. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2024, 7(3), 293-312. https://doi.org/10.1109/TBIOM.2024.3485990
9. Kim, H., Park, S., Choi, D. Secure and Reversible Face De-Identification with Format-Preserving Encryption. IEEE Access, 2025, 13(1), 116130-116142. https://doi.org/10.1109/ACCESS.2025.3583388
10. Li, R., Gao, M., Jin, X. Recognize Me if You Can: Two-Stream Adversarial Transfer for Facial Privacy Protection Using Fine-Grained Makeup. The Visual Computer, 2025, 41(9), 6943-6954. https://doi.org/10.1007/s00371-025-04015-3
11. Liu, J., Meng, R., Sun, N., Han, G., Kwong, S. Privacy-Preserving Video Fall Detection via Chaotic Compressed Sensing and GAN-Based Feature Enhancement. IEEE MultiMedia, 2022, 29(4), 14-23. https://doi.org/10.1109/MMUL.2022.3173335
12. Lyu, T., Guo, Y., Chen, H. Understanding the Privacy Protection Disengagement Behaviour of Contactless Digital Service Users: The Roles of Privacy Fatigue and Privacy Literacy. Behaviour & Information Technology, 2024, 43(10), 2007-2023. https://doi.org/10.1080/0144929X.2023.2237603
13. Lyu, Y., Jiang, Y., He, Z., Peng, B., Liu, Y., Dong, J. 3D-Aware Adversarial Makeup Generation for Facial Privacy Protection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(11), 13438-13453. https://doi.org/10.1109/TPAMI.2023.3290175

14. Ma, C., Li, J., Ding, M., Liu, B., Wei, K., Weng, J., Poor, H. V. RDP-GAN: A Rényi-Differential Privacy Based Generative Adversarial Network. IEEE Transactions on Dependable and Secure Computing, 2023, 20(6), 4838-4852. https://doi.org/10.1109/TDSC.2022.3233580

15. Morris, J., Newman, S., Palaniappan, K., Fan, J., Lin, D. Do You Know You Are Tracked by Photos That You Didn't Take: Large-Scale Location-Aware Multi-Party Image Privacy Protection. IEEE Transactions on Dependable and Secure Computing, 2023, 20(1), 301-312. https://doi.org/10.1109/TDSC.2021.3132230

16. Osorio-Roig, D., Rathgeb, C., Drozdowski, P., Busch, C. Stable Hash Generation for Efficient Privacy-Preserving Face Identification. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2022, 4(3), 333-348. https://doi.org/10.1109/TBIOM.2021.3100639

17. Osorio-Roig, D., Rathgeb, C., Drozdowski, P., Terhörst, P., Štruc, V., Busch, C. An Attack on Facial Soft-Biometric Privacy Enhancement. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2022, 4(2), 263-275. https://doi.org/10.1109/TBIOM.2022.3172724

18. Shaheryar, M., Lee, J. T., Jung, S. K. IDDiffuse: Dual-Conditional Diffusion Model for Enhanced Facial Image Anonymization. In, Proceedings of the Asian Conference on Computer Vision, 2024, 4017-4033. https://doi.org/10.1007/978-981-96-0911-6_25

19. Sivalakshmi, M., Prasad, K. R., Bindu, C. S. Convolutional-Based Variational Autoencoders for Face Privacy Protection in Video Surveillance. Journal of Discrete Mathematical Sciences and Cryptography, 2024, 27(4), 1205-1214. https://doi.org/10.47974/JDMSC-1974

20. Sun, G., Wang, H., Bai, Y., Zheng, K., Zhang, Y., Li, X., Liu, J. PrivacyMask: Real-World Privacy Protection in Face ID Systems. Mathematical Biosciences and Engineering, 2023, 20(2), 1820-1840. https://doi.org/10.3934/mbe.2023083

21. Thapa, S., Sarkar, A. A Deep Dive into Enhancing Sharing of Naturalistic Driving Data Through Face Deidentification. The Visual Computer, 2025, 41(4), 2563-2594. https://doi.org/10.1007/s00371-024-03552-7

22. Wang, H. Improved Facial Mask-Based Adversarial Attack for Deep Face Recognition Models. In, Proceedings of the International Conference on Image Algorithms and Artificial Intelligence (ICIAAI), 2024, 714-722. https://doi.org/10.2991/978-94-6463-540-9_73

23. Wang, S., Wang, C., Dong, T., He, Y., Xiao, K. Personalized Privacy-Preserving Data Utilization Approach Powered by Distributed-GAN. Big Data Mining and Analytics, 2024, 7(4), 1098-1113. https://doi.org/10.26599/BDMA.2024.9020037

24. Wang, T., Zhang, Y., Yang, Z., Xiao, X., Zhang, H., Hua, Z. Seeing Is Not Believing: An Identity Hider for Human Vision Privacy Protection. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2025, 7(2), 170-181. https://doi.org/10.1109/TBIOM.2024.3449849

25. Xiao, Y., Han, J. J., Cen, J., Fang, Z. Tensor-Based Privacy Protection Scheme with Multifeature Fusion for Facial Recognition. IEEE Transactions on Systems, Man, and Cybernetics: Systems, 2025, 55(6), 3964-3975. https://doi.org/10.1109/TSMC.2025.3547887

26. Yang, A., Bai, Y., Xue, T., Li, Y., Li, J. A Novel Image Steganography Algorithm Based on Hybrid Machine Learning and Its Application in Cyberspace Security. Future Generation Computer Systems, 2023, 145, 293-302. https://doi.org/10.1016/j.future.2023.03.035

27. Yang, G., Lin, J., Su, Z., Li, Y. Visual Privacy Behaviour Recognition for Social Robots Based on an Improved Generative Adversarial Network. IET Computer Vision, 2024, 18(1), 110-123. https://doi.org/10.1049/cvi2.12231

28. You, X., Zhao, X., Wang, Y., Sun, W. Generation of Face Privacy-Protected Images Based on the Diffusion Model. Entropy, 2024, 26(6), 479. https://doi.org/10.3390/e26060479

29. Zhang, J., Zeng, Z. Y., Si, K. L., Ye, X. C. Entropy-Driven Differential Privacy Protection Scheme Based on Social Graphlet Attributes. Journal of Supercomputing, 2024, 80(6), 7399-7432. https://doi.org/10.1007/s11227-023-05751-w

30. Zhong, Y., Deng, W. OPOM: Customized Invisible Cloak Towards Face Privacy Protection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3), 3590-3603. https://doi.org/10.1109/TPAMI.2022.3175602

31. Zhou, Z., Bao, Z., Jiang, W., Huang, Y., Peng, Y., Shankar, A., Selvarajan, S. Latent Vector Optimization-Based Generative Image Steganography for Consumer Electronic Applications. IEEE Transactions on Consumer Electronics, 2024, 70(1), 4357-4366. https://doi.org/10.1109/TCE.2024.3354824