Journal of Advanced Digital Communications
https://www.sciltp.com/journals/jadc

Article

DR-DPGAN: Dual Prior GAN for Image Privacy Based on Dynamic Reversibility

Tingting Lin [1], Tao Wang [2,*] and Zhigao Zheng [3]

[1] Central China Normal University Wollongong Joint Institute, Central China Normal University, Wuhan 430079, China
[2] Hubei Key Laboratory of Digital Education, Faculty of Artificial Intelligence in Education, Central China Normal University, Wuhan 430079, China
[3] School of Computer Science, Wuhan University, Wuhan 430072, China
* Correspondence: tmac@ccnu.edu.cn

How To Cite: Lin, T.; Wang, T.; Zheng, Z. DR-DPGAN: Dual Prior GAN for Image Privacy Based on Dynamic Reversibility. *Journal of Advanced Digital Communications* **2025**, *2*(1), 1. https://doi.org/10.53491/jadc.2025.10001.

Received: 25 April 2025
Revised: 12 June 2025
Accepted: 14 June 2025
Published: 23 June 2025

Abstract: With the development of image acquisition technology, the volume of image data has surged, highlighting the contradiction between data publication and privacy protection. Generative Adversarial Networks (GANs) offer a solution to find a balance between the development of image data and privacy security. However, the unidirectional image generation of GANs fails to satisfy the reversible requirements of privacy-sensitive images. To address this limitation, this study proposes an image privacy protection method based on a dual-prior GAN with dynamic reversibility, called DR-DPGAN. This method uses StyleGAN2 latent space editing to make targeted modifications to image identity features, allowing modification and reconstruction of features. To achieve image privacy protection, a fake identity generator composed of two-layer multi-layer perceptrons is designed. By combining identity-related guidance information, it precisely controls the generation of fake features to avoid excessive or insufficient modification. Meanwhile, three-dimensional prior constraints are introduced to extract geometric feature vectors, maximizing the retention of original non-identity attribute features and ensuring the usability of images in downstream tasks. To ensure reversible image restoration, this paper converts the original identity attribute information into binary vectors through a binary encoding mapping network, generating reversible encrypted features to ensure precise restoration of original identity features. In addition, four loss functions are used jointly to optimize the network to balance the quality of the generation. To verify the reversibility and effectiveness of the proposed method, comprehensive experimental tests are conducted on two different datasets. The experimental results demonstrate the effectiveness of this method in image anonymization and reversible restoration.

Keywords: image privacy; generative adversarial networks; reversibility

1. Introduction

In the digital era, widespread image sharing has raised critical privacy concerns as images often contain sensitive personal information [1]. While Generative Adversarial Networks (GANs) have shown promise in privacy protection through techniques like noise addition [2] and blurring [3], these irreversible methods face significant limitations in scenarios requiring original image recovery.

The irreversible nature of current GAN-based approaches poses particular challenges in fields like criminal investigations. When processing crime scene photos to protect bystander privacy, irreversible alterations may permanently erase forensically valuable details, potentially hindering case resolution.

Current reversible solutions face dual limitations: Conventional reversible data hiding (RDH) techniques [4–6] exhibit inadequate performance on complex images with constrained embedding capacity, while encoder-decoder

anonymization frameworks [7] often degrade biometric fidelity or inadvertently modify non-target attributes. These shortcomings stem from two unresolved challenges: (1) precise control over identity attribute manipulation, and (2) guaranteed reversibility between privacy protection and feature reconstruction.

In order to overcome the deficiencies of existing methods, this paper proposes a dual prior GAN for image privacy based on dynamic reversibility, named DR-DPGAN, which introduces a dynamic reversibility mechanism, enabling the result of privacy protection to be reversibly operated as needed, thus while protecting image privacy, it retains the possibility of restoring the original image. This method uses the collaborative design of StyleGAN2 latent space editing, 3D attribute modeling, and reversible information embedding to provide a more flexible and secure solution for image privacy protection.

Specifically, this paper uses StyleGAN2 latent space editing to perform directional modification of the identity features of the image, generates a high-fidelity de-identified image, and embeds the encrypted features into the image. In the reversible restoration stage, the original identity features are accurately restored through the inverse decoder. Combined with the geometric coefficients extracted by the 3D prior constraint module, the StyleGAN2 generator performs inverse feature mapping, and finally outputs the reconstructed image.

The main contributions of this paper are as follows.

- We generate synthetic identity features through randomized noise while incorporating identity-related guidance information to precisely control feature generation direction. This approach effectively prevents both excessive and insufficient modifications. By integrating with StyleGAN2's latent space editing technology, we achieve directional modification of facial identity features while maintaining high-fidelity de-identified image generation.
- Our method extracts geometric feature vectors from source images to maximally preserve non-identity attributes. This constraint mechanism prevents topological distortion in generated images, thereby enhancing their usability for downstream computer vision tasks.
- We combine binary-encoded mapping networks with StyleGAN2 for identity protection. The method converts original identity attributes into binary vectors to generate reversibly encrypted features, which are then embedded into images via StyleGAN2. Authorized users can precisely reconstruct original images by extracting and decoding the embedded features.
- Extensive experiments on real-world image datasets demonstrate that our DR-DPGAN framework maintains exceptional data utility while achieving perfect reversible reconstruction of identity features, outperforming existing de-identification methods.

The remainder of this paper is structured as follows: Section 2 provides a comprehensive review of related works in the field. Section 3 presents a detailed description of our proposed image privacy protection method, which is based on the Dynamic Reversible Dual-Prior Generative Adversarial Network (DR-DPGAN). Section 4 conducts extensive experiments to evaluate and analyze the performance of the proposed method. Finally, Section 5 concludes the paper by summarizing our contributions and discussing potential future directions.

2. Related Work

In today's digital age, image data is widely used in numerous fields such as healthcare, security, and social media, and the importance of image privacy protection has become increasingly prominent. In the early days, reversible image privacy protection technologies mainly revolved around encryption and digital watermarking techniques [8–10].

Scholars have proposed the use of Reversible Data Hiding (RDH) technology [5, 11] to embed encryption keys or privacy metadata into images. Specifically, high-capacity embedding is achieved through difference expansion or prediction error coding, while ensuring that the original image can be recovered without loss with the help of the key [12]. In addition, Encrypted Image Processing (EIP) technology also plays an important role. It protects sensitive areas through block encryption [13] or homomorphic encryption [14], and authorized users can recover the original pixels by decrypting. These methods have been verified to be feasible in the fields of medical imaging [15] and military communications. However, they face bottleneck problems such as high computational complexity and the difficulty of balancing robustness and reversibility.

In order to address the above challenges, some schemes have explored alternative approaches. For example, FIT [16] and RAPP [17] have adopted the identity vector encryption strategy, that is, the encrypted features are input into the generation network to achieve anonymization. Although these schemes can accurately recover the original data under password verification, the key management system has become a new security vulnerability and is vulnerable to attacks.

With the technological breakthroughs in GANs and reversible neural networks, new dynamic reversible privacy

protection frameworks have emerged. For instance, the literature [18] designed a reversible perturbation generator that injects an imperceptible noise layer into the image through adversarial training, and only authorized users can recover the original image with the aid of the inverse network. The literature [19] proposed a domain migration model based on Cycle Generative Adversarial Network (CycleGAN), which implicitly maps sensitive images to the privacy protection domain, such as cartoonization, while retaining the reversible decoding path. RiDDLE [20] innovatively integrates the latent space mapping ability of StyleGAN2 [21] and realizes the generation and inverse recovery of anonymous images through a hider. During the same period, Privacy-Net [7] proposed a hierarchical reversible transformation module. Although these methods have demonstrated the ability to resist inverse attacks in natural image scenarios, they have the drawback of poor visual natural effects.

The current reversible image privacy protection technologies still have many problems. From the perspective of image content, for complex images with rich textures and details, existing methods are difficult to achieve perfect reversible recovery, and there are certain errors. From the perspective of data processing scale, when dealing with large-scale image data, the efficiency of existing algorithms is insufficient and cannot meet the needs of practical applications.

Looking at traditional GAN desensitization methods, they mainly rely on irreversible identity perturbations to achieve anonymization. Once processed, the original identity information will be permanently changed. Although significant progress has been made in existing reversible protection technologies, the anonymization and restoration effects of most methods are still unsatisfactory. Taking biometric data processing as an example, the topological structure is extremely vulnerable to damage during the anonymization process, resulting in situations such as distortion of data features and deviation of key feature points. The existing anonymization technologies mainly include the identity tampering model based on the basic encoder-decoder architecture and the latent space editing method based on the inverse mapping technology of the generative adversary network. However, the former is likely to cause texture detail degradation when processing biometric data and cannot accurately restore the feature information; the latter is often accompanied by unexpected changes in identity-independent attributes, such as the environmental conditions during data collection and the presentation angle of features, which affects the availability and accuracy of the data.

In conclusion, how to precisely manipulate identity information while protecting privacy and enable the image to be accurately restored to its original identity state has become a crucial challenge that urgently needs to be addressed in the field of reversible image privacy protection.

3. Methodology: DR-DPGAN

In this section, we elaborate the design of DR-DPGAN, a dual prior GAN for image privacy based on dynamic reversibility, which aims to overcome the irreversibility of GAN in privacy protection.

3.1. Framework of DR-DPGAN

As shown in Figure 1. The DR-DPGAN consists the fake identity generator model, 3D prior constraints model, as well as identity manipulation and anonymization model. The fake identity generator is a network consisting of two layers of Multi-Layer Perceptrons (MLP), aiming to guide the generation of false identity features from random noise. Identity manipulation and anonymization model consists of two parts: the hiding of identity information and the latent space editing of StyleGAN2. The 3D prior constraints play a major role in extracting the geometric feature vectors from images. The Hiding of Identity Information is responsible for password extraction and contains an independent convolutional neural network encoder, which can generate reversible encrypted features. The latent space editing of StyleGAN2 contains the StyleGAN generator G. It is a feature generation module based on the W^+ latent space and has the ability of adaptive feature fusion, enabling the manipulation and editing of images as well as information embedding.

The workflow of the framework is divided into two stages: anonymization and reversible recovery.

In the anonymization stage, when the real image $original$ is input, the 3D prior constraints model first extracts the geometric coefficients and performs tensor concatenation with the virtual identity embedding $F_{id}^{'}$ to generate a geometrically aware identity encoding F_{concat}. The F_{id} is converted to a binary vector and encrypted as $Password$. Then, the StyleGAN2 generator maps the encoding to the W^+ space. At the same time, the original image goes through an encoder to extract the identity-independent feature $F_G^{'}$. Subsequently, the E_{Ste} model performs feature fusion on $Password$ and $F_G^{'}$, and finally reconstructs and generates the anonymized image $anonymization$, which retains the original geometric features but replaces the identity features.

In the reversible recovery stage, given the anonymized image $anonymization$ and the real identity F_{id}, the 3D prior constraints module extracts the geometric coefficients and concatenates them with the real identity embedding

F_{concat}. Then, the same E_{Ste} architecture is used to reversely fuse the real identity features, and the StyleGAN generator performs the reverse feature mapping. Finally, the reconstructed image $recovery$ is output, which can precisely restore the features of the original image.

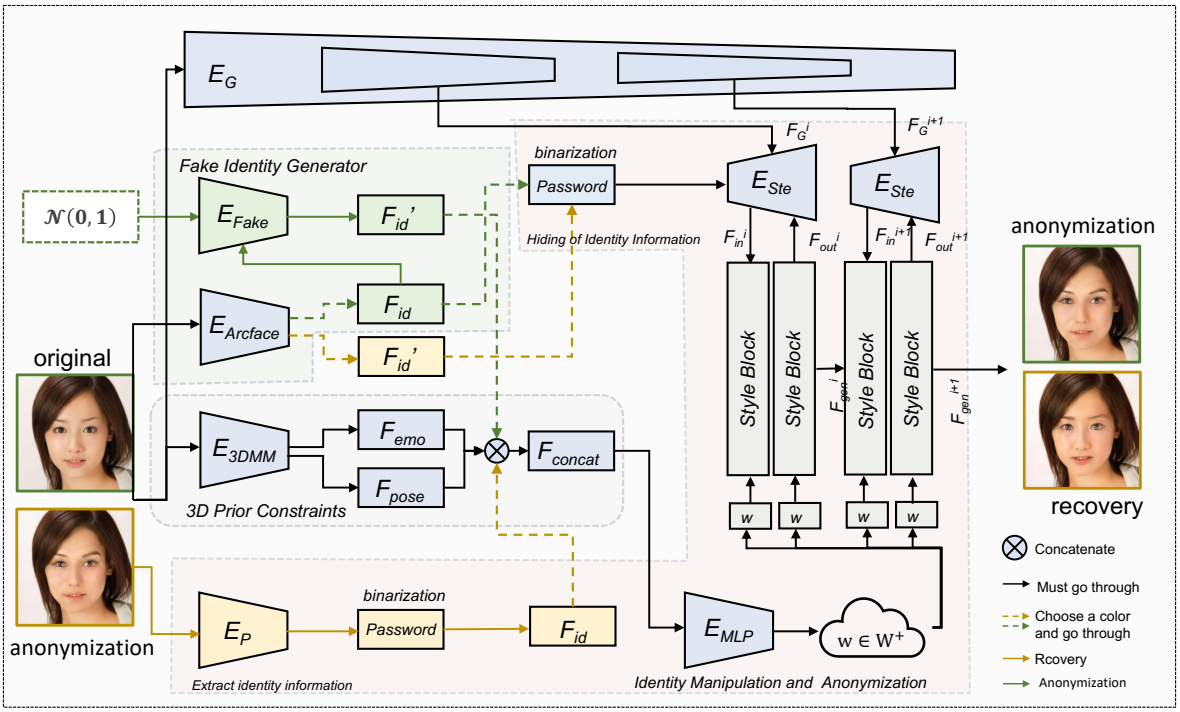


Figure 1. Framework of DR-DPGAN.

3.2. Fake Identity Generator

In the process of generating fake identity IDs, we introduce identity-related guidance information to enable more precise control over the direction and intensity of fake identity ID generation, thereby preventing two extreme scenarios during the creation of pseudo-identity identifiers. The first scenario is over-modification, where excessive alteration of identity features causes the modified images to lose the essential connection with either the original or target identity. The second scenario is inadequate modification, where insufficient changes to identity features fail to achieve the anonymization objective effectively, allowing the original identity to remain easily recognizable.

As shown in Algorithm 1, the pre-trained recognition model $E_{Arcface}$ is employed to extract real identity IDs from images. Simultaneously, a two-layer MLP E_{Fake} guides normally distributed random noise vectors to generate virtual identity IDs. In this MLP architecture, the first layer implements non-linear feature transformation using $ReLU$ activation, while the second layer projects features into a dimension aligned with ArcFace's feature space. Through cosine similarity constraints, the virtual identity features $F_{id}^{'}$ maintain a controllable distance from the real identity features F_{id} during generation.

Through statistical analysis of cosine similarity between different identity pairs in large-scale face datasets, it can be observed that the similarity between different identities mostly falls within the range $[0.1, 0.4]$, while the similarity of the same identity is predominantly distributed in $[0.6, 1.0]$. Therefore, we select $[s_{min}, s_{max}] = [0.1, 0.4]$ to ensure that the similarity between the virtual identity and the original identity is significantly lower than the same-identity threshold, while remaining higher than random noise similarity, thereby balancing the naturalness and anonymity of the virtual identity.

3.3. 3D Prior Constraints

In traditional image synthesis tasks using Generative Adversarial Networks, due to the lack of explicit three-dimensional geometric constraints, the generated image results often suffer from geometric inaccuracies. For example, there are issues such as asymmetric distortion of objects in the images and distortion of the topological structure. These problems limit the semantic fidelity of non-critical features in the images (such as the local morphology and placement angles of objects).

Algorithm 1 Fake identity generator

Input: Original image X_{ori}, noise d, target similarity range $[s_{\min}, s_{\max}]$, initial cosine similarity $s = 0$, number of generator iterations T_g

Output: Virtual identity feature vector $F_{id}^{'}$

$F_{id} \leftarrow E_{Arcface}(X_{ori})$

Sample noise vector $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}) \in \mathbb{R}^d$

for $t_1 \leftarrow 1$ **to** T_g **do**

  $h \leftarrow ReLU(W_1 z + b_1)$

  $F_{id}^{'} \leftarrow W_2 h + b_2$

  $F_{id}^{'} \leftarrow \frac{F_{id}^{'}}{\|F_{id}^{'}\|_2}$

  Compute cosine similarity: $s \leftarrow \frac{F_{id}^{\top} F_{id}^{'}}{\|F_{id}\|_2 \|F_{id}^{'}\|_2}$

  Construct loss: $\mathcal{L} \leftarrow \max(0, s - s_{\max}) + \max(0, s_{\min} - s)$

  Update parameters: $\theta_{\text{MLP}} \leftarrow \theta_{\text{MLP}} - \eta \nabla_\theta \mathcal{L}$

  if $s \in [s_{\min}, s_{\max}]$ **then**

    End of training

return $F_{id}^{'}$

To address this issue, this study introduces an optimization framework based on the Parametric 3D Morphable Model. By conducting decoupled modeling of the morphological variation coefficients and rigid placement parameters of objects in the images, a differentiable three-dimensional object prior constraint is constructed. This approach can provide synthetic data support with explicit semantic decoupling for downstream tasks such as object recognition.

In the specific operation process, with the help of a pre-trained 3D object reconstruction model, predictions are made on the input image, thus obtaining the object morphological variation parameter F_{emo} and the placement posture parameter F_{pos}. These parameters will serve as the 3D object prior conditions for generating anonymized images.

Subsequently, the virtual identity embedding value $F_{id}^{'}$ is combined with these parameters, and a 3D prior fusion identity embedding value F_{concat} is obtained through a concatenation operation. The formula can be found in Equation (1).

$$F_{concat} = \begin{cases} Concat(F_{id}^{'}, F_{emo}, F_{pos}) & \text{if } t = 1 \\ Concat(F_{id}, F_{emo}, F_{pos}) & \text{else} \end{cases} \tag{1}$$

In the anonymization process at $t = 1$ and in the recovery process at $t = 0$.

3.4. Identity Manipulation and Anonymization

Identity manipulation and anonymization aim to protect identity characteristics and achieve reversible recovery of images. Specifically, the process primarily includes three core stages: the hiding of identity information, the extraction of identity information, and the latent space editing of StyleGAN2. Firstly, in the identity information hiding phase, the original identity information is converted into a binary format and embedded into the image using steganography, ensuring the security of the identity information. Secondly, in the identity information extraction phase, a multi-layer convolutional network is employed as a password extractor to precisely retrieve the embedded information from the steganographic image and restore it to floating-point format identity information. Finally, leveraging the latent space editing technique of StyleGAN2, a nonlinear mapping network is constructed to project a synthetic identity feature vector incorporating 3D morphable priors into the extended latent space of StyleGAN2, enabling identity attribute editing of the generated images. Throughout this process, feature fusion technology is utilized to adaptively merge identity-independent features of the generated image with intermediate features, while the original image's identity information is retained in the form of password embedding. This achieves high-quality reconstruction of the image and reversible recovery of identity information.

3.4.1. Hiding of Identity Information

In the context of today's critical information security requirements, steganography is skillfully integrated to achieve reversible recovery of image information while ensuring the security of identity information.

First, preprocessing of identity information is required. Specifically, the identity id F_{id} involved is converted

from its original format into a binary format $Password$. This conversion process can be simply expressed as:

$$Password = BinaryCodeMapping(F_{id}) \tag{2}$$

Figure 2 illustrates the feature-to-binary mapping process in the Hiding of Identity Information module. The core component is a binary code mapping network constructed using fully connected (FC) layers and non-linear activation functions. The network architecture consists of three FC layers and a Sigmoid activation layer. Additionally, a Dropout strategy with a probability of 0.5 is applied to FC layers to prevent network overfitting. This strategy randomly selects neurons to participate in training iterations during training, masking non-selected neurons to reduce complex co-adaptive relationships between neurons. The Sigmoid activation layer is placed after the FC layers to generate identity features with approximate high entropy. Specific implementation details can be found in Figure 3.

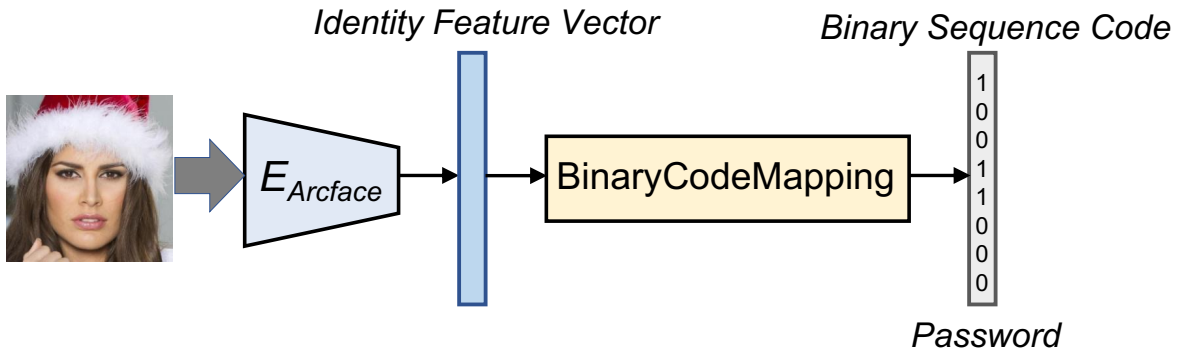


Figure 2. Flowchart of feature-to-binary mapping.

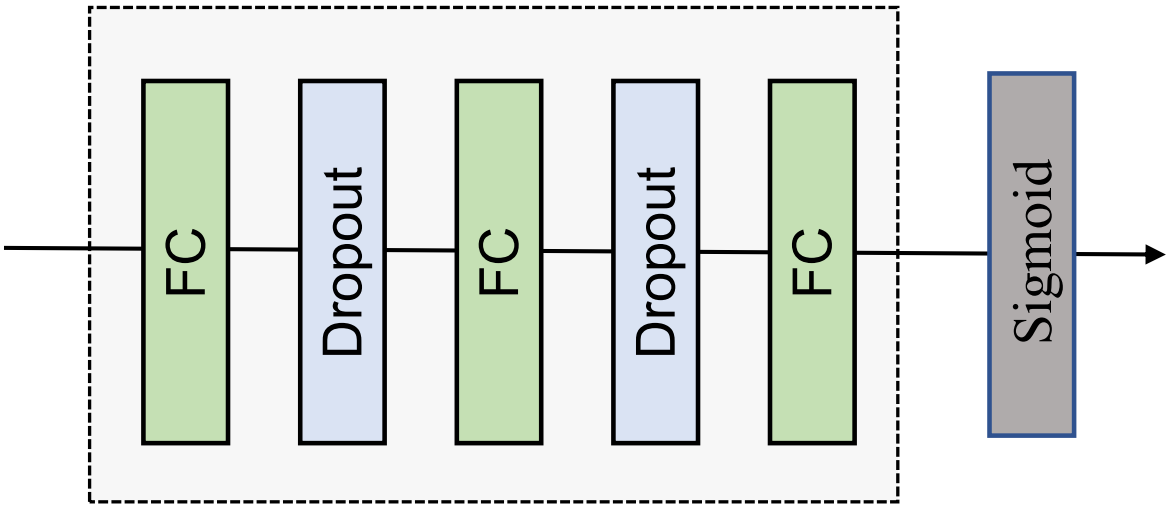


Figure 3. The detail of binary code mapping.

After non-linear projection matrix operations, the generated identity feature vector F_{id} (floating-point data with element values distributed between 0 and 1) requires a binary quantization operation to obtain the mapped binary sequence code B'. A dynamic quantization threshold $\bar{F} = (\sum_{i=1}^{l} F_i)/l$ (where F_i represents the i-th element of F_{id}, and l is the length of F_{id}) is used to minimize quantization errors. Finally, elements B'_i (for $1 \leq i \leq l$) in the mapped binary sequence code B' are defined as:

$$\mathbf{B'} = [B'_1, \ldots, B'_i, \ldots, B'_l] = [q(F_1), \ldots, q(F_i), \ldots, q(F_l)] \tag{3}$$

where the binary quantization function $q(F_i)$ follows the rule:

$$q(F_i) = \begin{cases} 1, & \text{if } F_i \geq \bar{F} \\ 0, & \text{otherwise} \end{cases} \quad \text{for } 1 \leq i \leq l \tag{4}$$

The necessity of this operation lies in the fact that directly embedding original identity features into images often

leaves detectable traces. Such traces, once identified by detection tools, may lead to identity information decryption and pose significant security risks. In contrast, the binary format offers distinct advantages. Through specific algorithms and advanced techniques, it can be embedded into images in a more concealed manner. Compared to raw identity features, this binary representation is more abstract and difficult to directly interpret, thereby effectively reducing detection risks and significantly enhancing information concealment security.

3.4.2. Extract Identity Information

After the steganography-embedded image has been utilized, a series of meticulously designed steps are executed for image recovery. A multi-layer convolutional network is employed as a password extractor E_P, which precisely extracts the information embedded in the image. Following successful extraction, further processing is performed to convert the retrieved information back into corresponding floating-point identity information F_{id}. The obtained floating-point identity information F_{id} is then fused with features extracted through 3DMM. Through sophisticated processing, the fused information is projected into the W^+ space of StyleGAN2 to enable image manipulation. Through this complex yet well-organized operational sequence, the original image is ultimately successfully reconstructed, accomplishing reversible recovery of image information.

3.4.3. Latent Space Editing of StyleGAN2

Based on the research findings from literature [20], the latent space W^+ of StyleGAN2 can achieve facial identity attribute editing through latent variable disentanglement. Building upon this, we construct a nonlinear mapping network E_{MLP} to project a synthetic identity feature vector $F_{\text{concat}} \in \mathbb{R}^d$ (incorporating 3D morphable priors) into the extended latent space W^+ of the pre-trained StyleGAN2. This mapping process can be formally expressed as:

$$w_i = E_{MLP}(F_{\text{concat}}), \quad i \in \{1, 2, 3\} \tag{5}$$

Here, i corresponds to the level. These latent codes are like instructions that control the image generation of StyleGAN2. Different levels of w_i can control the generated images at different levels of detail granularity. The structure of latent space editing is in the Figure 4.

Each prior block in StyleGAN2 contains two style convolution blocks. Taking the i-th GAN block as an example, the first style convolution block receives the input feature F^i_{gen} and the latent code w, and then generates the intermediate feature F^i_{out}. In this process, since the latent code w is generated from F_{concat}, the generated intermediate feature F^i_{out} changes the identity information while retaining the geometric features.

In order to make the generated images more realistic and retain the features that are irrelevant to the identity, feature fusion is required. The encoder E_G is used to extract multi-scale features F^i_G from the original image. These features contain information in the image that is irrelevant to the identity, such as certain elements in the image and the background of the image. Then, the intermediate features F^i_{out} are adaptively fused with the multi-scale features F^i_G. During the fusion process, the identity information F_{id} of the original image is also embedded in the form of password embedding. Specifically, first, F_{id} is converted into a binary vector and encrypted into $Password$, and then $conv_{1\times1}(Password)$ is obtained through convolution. Each E_{Ste} module takes F^i_{hide}, F^i_{out}, and F^i_G as inputs. Generate an adaptive mask is

$$M_i = conv_{1\times1}(F^i_{hide}, F^i_G, F^i_{out}) \tag{6}$$

through a 1×1 convolution and a Sigmoid operation. This mask can adaptively capture the pixels irrelevant to the identity and guide the feature fusion. The fusion process is:

$$F^i_{in} = M_i \cdot E_G(X_{ori}, i) + (1 - M_i) \cdot F^i_{out} + conv_{1\times1}(Password) \tag{7}$$

Subsequently, the fused feature F^i_{out} is fed into the subsequent convolutional layers of the generator G, which utilizes it for the generation of features.

Final, the output F^{i+1}_{gen} is obtained by performing an Adaptive Instance Normalization (AdaIn) operation on the input feature F^i_{in}, where the parameters for the AdaIn operation are w_i, and then performing a 3×3 convolution operation. Its mathematical expression is:

$$F^{i+1}_{gen} = conv_{3\times3}(F^i_{in}, AdaIn(w_i)) \tag{8}$$

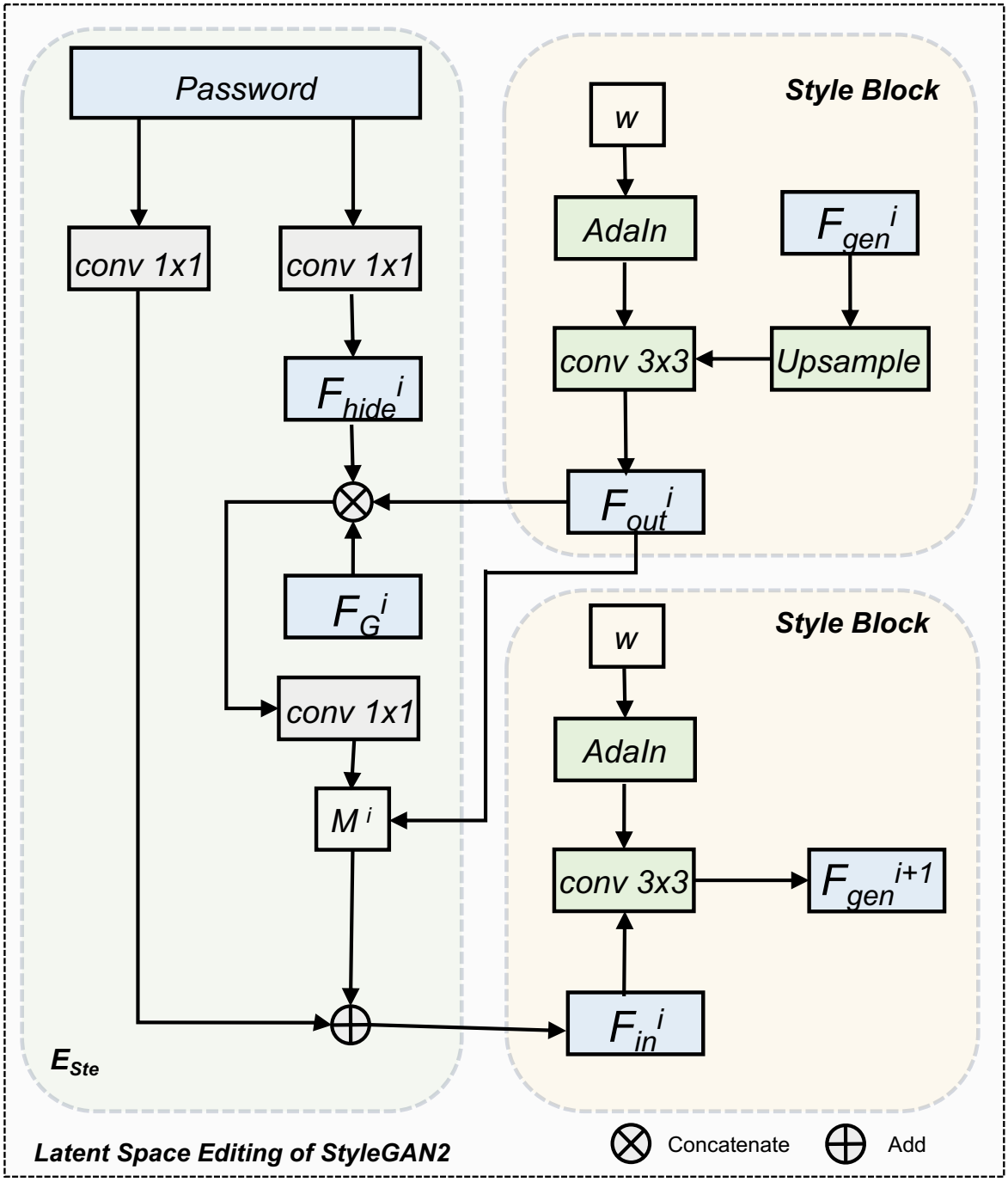


Figure 4. Structure of latent space editing.

3.5. Loss Function

To balance the generation quality, identity difference, and non-identity feature consistency, a joint optimization method using multiple loss functions is proposed. The original image is represented by X_{ori}, and the generated image is represented by X_{gen}.

3.5.1. Adversarial Loss

According to StyleGAN2, an adversarial loss is also introduced to enhance the realism of the generated images X_{gen}:

$$\mathcal{L}_{adv}(X_{gen}) = \mathbb{E}_{X_{gen}} \ln(1 + e^{-C(X_{gen})}) \tag{9}$$

Here, C represents the discriminator, and its architecture is similar to that of the discriminator in StyleGAN2.

3.5.2. Discriminative Loss

By minimizing the discriminative loss, the discriminator's judgment of the carrier image approaches 1, and the judgment result of the stego-image approaches 0, so as to achieve a state of correct classification.

$$\mathcal{L}_{dic} = (1 - Critic(X_{ori}))^2 + (Critic(X_{gen}) - 0)^2 \tag{10}$$

Among them, $Critic(X_{ori})$ and $Critic(X_{gen})$ respectively represent the discriminator's judgment results of the two images.

3.5.3. Identity Loss

The purposes in the anonymization stage and the recovery stage are different.

- Anonymization. The objective is to enhance the disparity between the synthesized face and the genuine identity to the greatest extent possible, and simultaneously, to reduce the disparity between the synthesized face and the false identity to the lowest degree. So the formula is

$$\begin{aligned} &\mathcal{L}_{id}^{ano}(X_{gen}, F_{id}, F_{id}^{'}) = 1 - \cos(F_{id}', E_G(X_{gen})) \\ &+ \max(0, \cos(F_{id}, E_G(X_{gen}))) \end{aligned} \tag{11}$$

  where F_{id} represents the original ID embedding of the input image X, and $E_G(\cdot)$ represents the identity embedding extracted by the pre-trained ArcFace. This loss achieves its goal through two components: first, $1 - \cos(F_{id}', E_G(X_{gen}))$ drives the identity of the anonymized image to closely align with the dummy identity, and second, $\max(0, \cos(F_{id}, E_G(X_{gen})))$ penalizes any similarity between the original identity and the anonymized image. This ensures that the anonymized image effectively dissociates from the original identity while binding to the dummy identity, thus safeguarding privacy and maintaining the rationality and controllability of the anonymization result.
- Recovery. The aim is to reduce the discrepancy between the recovered image and the actual identity to the smallest possible extent. So the formula is

$$\mathcal{L}_{id}^{rec}(X_{ori}, F_{id}) = 1 - \cos(F_{id}, E_G(\hat{X_{ori}})) \tag{12}$$

3.5.4. Information Hiding Loss

The information hiding loss $\mathcal{L}_{info}$ is determined by the binary cross-entropy loss function $CrossEntropy$. First, a small Gaussian noise ϵ (used to enhance the stability of information embedding) is added to the output image X_{gen}. Then, the processed image is input into the password extractor $E_P(\cdot)$ to obtain the extracted password. Finally, this password and the target value V are substituted into $CrossEntropy$ for calculation.

The formula can be rewritten using mathematical symbols as follows:

$$Y = X_{gen} + \epsilon \tag{13}$$

$$Password = E_P(Y) \tag{14}$$

$$\mathcal{L}_{info} = CrossEntropy(Password, V) \tag{15}$$

3.5.5. Total Loss

To achieve joint optimization across generation quality, identity preservation, and information hiding, we formulate the total loss as a weighted combination of individual loss components. The objective is to minimize the total loss during training:

$$\mathcal{L}_{total}^{ano} = \lambda_{adv}\mathcal{L}_{adv} + \lambda_{dic}\mathcal{L}_{dic} + \lambda_{id}\mathcal{L}_{id}^{process} + \lambda_{info}\mathcal{L}_{info} \tag{16}$$

$$process = \begin{cases} ano & \text{if it is the anonymization process,} \\ rec & \text{if it is the recovery process.} \end{cases}$$

where λ_{adv} means weight for adversarial loss, λ_{dic} means weight for discriminative loss, λ_{id} means weight for identity loss, λ_{info} means weight for information hiding loss. The training process minimizes the total loss with

respect to generator parameters θ_G and discriminator parameters θ_D:

$$\min_{\theta_G,\theta_D} \mathcal{L}_{total} \tag{17}$$

4. Experiments and Analysis

4.1. Experimental Setting

4.1.1. Datasets

In this experiment, two data sets are used. The first is CelebA-HQ, the high-definition version of the CelebA dataset. It contains 30,000 face images and also includes information such as facial landmarks and facial attributes. The second is the CASIA-WebFace dataset, which has 494,414 face images from 10,575 different identities. In this study, it is referred to as CASIA for short. It is diverse, covering face images of different ages, genders, ethnicities, and expressions, as well as challenging scenarios such as lighting and pose variations.

4.1.2. Baseline and Parameters

In this experiment, RiDDLE [20] and FIT [16] are selected as comparison algorithms. Both are open-source methods that support reversible image anonymization.

Training with the adam optimizer, the StyleGAN2 blocks have a learning rate of 0.0001, the other trainable parts score 0.001, and the batch size is 8. For the loss functions, we set (λ_{adv}, λ_{dic}, λ_{id}, λ_{info}) = (1.0, 1.0, 1.0, 5.0).

4.2. Evaluation Metrics

- Cosine Similarity: It quantifies the similarity between two vectors in a multidimensional space. This measure captures the directional alignment of the vectors rather than their magnitude, providing an assessment of their relative orientation.
- Mean Absolute Error (MAE): It serves the purpose of assessing the mean discrepancy between the predicted and the actual values. The smaller the value, the more accurate the prediction will be.
- Peak Signal-to-Noise Ratio (PSNR): It is a metric to measure the quality distortion of reconstructed images between the maximum possible power of a signal and the average power of noise, where a higher value indicates less distortion and better quality.
- Bit Error Rate (BER). It represents the ratio of the quantity of error bits to the overall number of bits that have been transmitted. The lower the value, the more reliable the data transmission will be.
- Float Mean Squared Error (FMSE): It calculates the mean value of the squares of the errors between the predicted and the true values of the floating-point numbers. A smaller value indicates better performance.

4.3. Experimental Results

4.3.1. Generated Image Quality

As shown in the generated images in Figure 5, our algorithm demonstrates excellent performance. Whether it is dealing with the large number of high-resolution and diverse face images in the CASIA dataset or the high-definition face images in the CelebA-HQ dataset, it can ensure good generation quality. In the anonymization stage, our algorithm can accurately process the data, effectively concealing sensitive information and ensuring privacy. In the recovery stage, the algorithm can efficiently restore the image features, retaining the key information of the images to the greatest extent possible. This enables the generated images to maintain high visual quality and usability while having privacy protection features.

As can be seen in Figures 6 and 7, different types of random noise are applied to specific original images to generate anonymized images. The results generated exhibit a high degree of diversity. Notably, although there are various changes in the images, the non-identity features such as contours, hairstyles, and backgrounds remain highly consistent. This fully demonstrates that this algorithm has excellent control capabilities during the processing and also effectively proves that the applied prior knowledge has achieved the expected results.

Figure 5. The generation quality of our algorithm under two datasets.

Figure 6. The generation results of our algorithm for the CelebA-HQ dataset under different random noises.

Figure 7. The generation results of our algorithm for the CASIA dataset under different random noises.

4.3.2. Usability

In the process of studying the GAN-based generation results, the accuracy of face recognition classification is a key indicator, which can intuitively reflect whether there is a one-to-one correspondence between the identity information in the images generated by the GAN and the original identity information. To explore this relationship in depth, experiments were conducted on the CASIA dataset. Specifically, three different algorithms were used to generate images, and then these generated images were used to train the FaceNet classifiers, and their classification accuracies were evaluated. First, the generative model was used to create data for privacy protection, and then the classifier is trained on these data. After the training was completed, the publicly available test set was used to test the trained model. It is worth noting that for anonymized images, the lower the classification accuracy, the better the de-identification effect, that is, the identity information in the images has been successfully eliminated; while for recovery images, the higher the classification accuracy, the better the recovery effect, which means that the identity information in the original images has been restored as much as possible.

As shown in Table 1, in the anonymization tasks for the FaceNet classifier on the CASIA dataset, different algorithms exhibit varying performances. Among them, the accuracy rate of the anonymized images generated by our algorithm is only 6.85% under the FaceNet classifier, which is the lowest among the three algorithms. This clearly indicates that our algorithm performs excellently in hiding the original identity information of images, and the generated images are more difficult to be recognized by the classifier. In contrast, the accuracy rate of the FIT algorithm is relatively high, reaching 9.63% under the FaceNet classifier, which means that its anonymization effect is relatively weak, and the generated images are more likely to leak the original identity information.

Table 1. Evaluation of classification accuracy of images generated by different algorithms on the CASIA dataset for FaceNet classifier.

Algorithm	Anonymization	Recovery
RiDDLE	7.56%	92.36%
FIT	9.63%	91.02%
ours	6.85%	93.79%

Regarding the results of the recovery tasks for the FaceNet classifier on the CASIA dataset, our algorithm has an obvious advantage. After the images generated by this algorithm are recovered, the accuracy rate reaches even 93.79% under the FaceNet classifier, which is higher than those of the RiDDLE and FIT algorithms. This fully demonstrates that our algorithm can not only effectively anonymize images but also accurately restore the images to their original states when needed, enabling the classifier to accurately recognize them. However, the accuracy rate of the FIT algorithm during the recovery process is relatively low, being 91.02% under the FaceNet classifier, reflecting its relatively weak performance in image recovery.

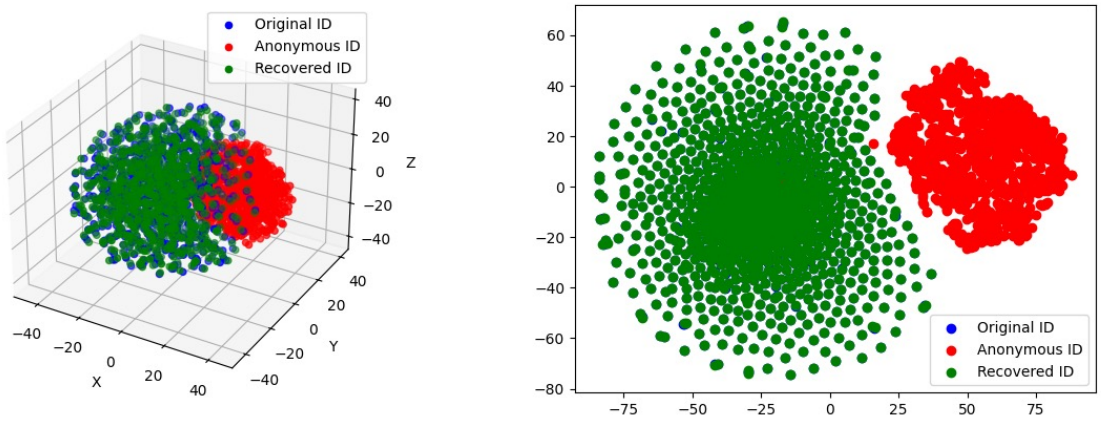
To evaluate the generated results more accurately, we will further test on the CelebA-HQ dataset. Considering that the CelebA-HQ dataset is an unstructured dataset, we will test the cosine similarity of the anonymized images, the cosine similarity of the recovered images, as well as MAE and PSNR values.

Table 2 presents the evaluation results of image quality metrics for different algorithms on the CelebA-HQ dataset. In terms of image anonymization, the cosine similarity of the ours algorithm is 0.103, which is lower than 0.159 of the RiDDLE algorithm and 0.181 of the FIT algorithm, indicating that the ours algorithm can better eliminate the identity information in the images. Regarding image recovery, the cosine similarity of the ours algorithm is 0.811, higher than 0.782 of the RiDDLE algorithm and 0.760 of the FIT algorithm. Moreover, its MAE is 0.055, lower than 0.079 of the RiDDLE algorithm and 0.070 of the FIT algorithm, and the PSNR is 23.029, higher than 20.231 of the RiDDLE algorithm and 19.215 of the FIT algorithm. Overall, the ours algorithm outperforms the RiDDLE and FIT algorithms in all aspects of image anonymization and recovery, demonstrating more excellent performance.

Figure 8 shows the visualization results of the distribution of identity characteristics, presented in three dimensions (Figure 8a) and two dimensions (Figure 8b), respectively. They involve the feature distributions of the original ID, Anonymous ID, and Recovered ID. The feature points of the original ID and the Anonymous ID are distributed at a relatively large distance, which once again proves that the anonymization process significantly alters the identity features. The feature points of the recovered ID are somewhat close to those of the original ID, indicating that the recovery algorithm can restore the original identity features to some extent. However, the two do not overlap completely, suggesting that noise has a certain impact on the results.

Table 2. Evaluation of image quality metrics on CelebA-HQ dataset.

	Anonymization	Recovery		
Algorithm	**Cosine Similarity**	**Cosine Similarity**	**MAE**	**PSNR**
RiDDLE	0.159	0.782	0.079	20.231
FIT	0.181	0.760	0.070	19.215
ours	0.103	0.811	0.055	23.029



(a) Three-dimensional (b) Two-dimensional

Figure 8. Visualization of identity feature distribution.

Finally, we conduct a rigorous verification of the accuracy of the passwords extracted from the identity. More specifically, the accuracy is gauged through the meticulous computation of the bit error rate and the float mean squared error during the process of converting binary data to floating-point data on the two datasets, namely CelebA-HQ and CASIA. The relevant results are presented in Table 3. It can be clearly seen from the evaluation results that the adopted method can accurately extract the embedded binary passwords and successfully convert them into valid identity information.

Table 3. Password accuracy evaluation results on CelebA-HQ and CASIA datasets.

Database	**Bit Error Rate**	**Float Mean Squared Error**
CelebA-HQ	0.0001	0.0010
CASIA	0.0001	0.0013

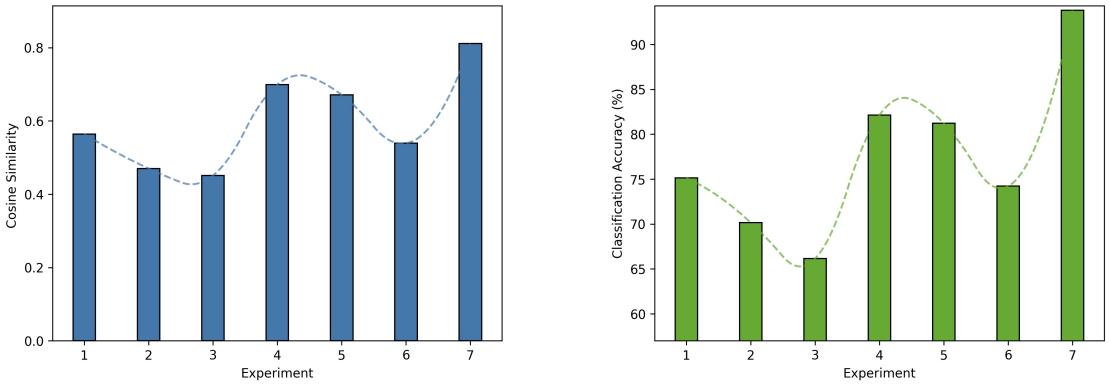
4.3.3. Ablation Experiment

To fully verify the effectiveness of the proposed optimization strategies, we combine several optimization strategies in different combination methods to construct new generative adversarial network models. The specific combination scenarios are shown in Table 4. Among them, "The control part of F_{id} in the FIG" indicates whether to retain the F_{id} part in the Fake Identity Generator module to guide the generation process of fake identities. "3DPC" means 3D Prior Constraints module. "The control part of E_G in the IMA" refers to whether to retain the E_G component in the Identity Manipulation and Anonymization module, so as to achieve effective control over the editing of identity features in the latent space. Subsequently, we select CelebA-HQ and CASIA datasets as the training data respectively, trains different GAN models constructed, and calculates their generation results. The performance of the models is evaluated through an in-depth analysis of the generation results.

Table 4. Requirements for ablation experiments.

Experiment	The Control Part of F_{id} in the FIG	3DPC	The Control Part of E_G in the IMA
1	✓		
2		✓	
3			✓
4	✓	✓	
5	✓		✓
6		✓	✓
7	✓	✓	✓

As can be clearly seen from the experimental data in Figure 9, for both the CelebA-HQ dataset and the CASIA dataset, only when these several optimization strategies are applied simultaneously can the cosine similarity of the restored images and the accuracy of face recognition reach the optimal level. This is because this algorithm skillfully utilizes these optimization strategies to achieve precise control over the consistency of reversible images to the greatest extent, significantly reducing the deviation in the image generation process. The experimental results strongly demonstrate the effectiveness and necessity of these optimization strategies in improving the performance of the model.



(a) CelebA-HQ (b) CASIA

Figure 9. Results of ablation experiments on the recovered images under different datasets.

5. Conclusions

This paper presents DR-DPGAN (Dual Prior GAN for Image Privacy Based on Dynamic Reversibility), which employs a 3D deformable model to extract identity-irrelevant attributes while incorporating forged biometric identifiers. The framework utilizes a StyleGAN2 generator for image reconstruction and embeds authentic identity information into images, enabling password extraction from anonymized images and subsequent original image recovery. Our experimental results demonstrate the method's effectiveness in image anonymization and reversible restoration. However, potential image distortion artifacts may occur, primarily attributed to the scarcity of representative training samples in such scenarios. This limitation underscores the necessity of employing more diversified training datasets to mitigate deformation anomalies.

Author Contributions

T.L.: methodology, data curation, writing—original draft preparation; T.W.: conceptualization, methodology, supervision; Z.Z.: writing—reviewing and editing. All authors have read and agreed to the published version of the manuscript.

Funding

This study was funded by the National Natural Science Foundation of China under Grants 62477017, 62372333, 62402195, 62307019, 62277021, and 62277029, Fundamental Research Funds for the Central Universities, Central China Normal University, under Grants CCNU25ai006 and CCNU25ai004, Fundamental Research Funds for the

Central Universities, Beijing Nova Program under Grant 2023048435 and Huzhou Key Research and Development Program under Grant 2023ZD2046.

Data Availability Statement

Not applicable.

Conflicts of Interest

The authors declare no conflict of interest.

References

1. Morris, J.; Newman, S.; Palaniappan, K.; et al. Do you know you are tracked by photos that you didn't take: Large-scale location-aware multi-party image privacy protection. *IEEE Trans. Dependable Secur. Comput.* **2023**, *20*, 301–312.
2. Chunling, H.; Rui, X. Differentially private gans by adding noise to discriminator's loss. *Comput. Secur.* **2021**, *107*, 102322–102332.
3. Jindong, J.; Wafa, S.; Ali, S.; Laurent, G. Effect of face blurring on human pose estimation: Ensuring subject privacy for medical and occupational health applications. *Sensors* **2022**, *22*, 9376.
4. Kumar, A.R.; Sharma, A.K. Reversible data hiding in encrypted image using bit-plane based label-map encoding with optimal block size. *J. Inf. Secur. Appl.* **2025**, *90*, 104005.
5. Kumar, A.R.; Sharma, A.K.; Ranjan, P. High-capacity reversible data hiding in encrypted images based on difference image transfiguration. *Signal Image Video Process.* **2025**, *19*, 365.
6. Konduru, U.R.; Nagarajan, A.P.; Sai, C.V.S. An improved performance of reversible data hiding in encrypted images using decision tree algorithm. *Eng. Appl. Artif. Intell.* **2024**, *137*, 109100.
7. Kim, B.N.; Dolz, J.; Jodoin, P.M.; et al. Privacy-net: An adversarial approach for identity-obfuscated segmentation of medical images. *IEEE Trans. Med. Imaging* **2021**, *40*, 1737–1749.
8. Liao, X.; Wang, Y.; Wang, T.; et al. Famm: Facial muscle motions for detecting compressed deepfake videos over social networks. *IEEE Trans. Circuits Syst. Video Technol.* **2023**, *33*, 7236–7251.
9. Fu, L.; Liao, X.; Guo, J.; Dong, L.; Qin, Z. Waverecovery: Screen-shooting watermarking based on wavelet and recovery. *IEEE Trans. Circuits Syst. Video Technol.* **2025**, *35*, 3603–3618.
10. Li, Y.; Liao, X.; Wu, X. Screen-shooting resistant watermarking with grayscale deviation simulation. *IEEE Trans. Multimed.* **2024**, *26*, 10908–10923.
11. Yang, Y.; He, H.; Chen, F.; et al. Secure reversible data hiding in encrypted image based on 2d labeling and block classification coding. *J. Inf. Secur. Appl.* **2024**, *83*, 103771–103780.
12. Liao, X.; Yin, J.; Chen, M.; Qin, Z. Adaptive payload distribution in multiple images steganography based on image texture features. *IEEE Trans. Dependable Secur. Comput.* **2022**, *19*, 897–911.
13. Wu, H.; Cheung, Y.; Zhuang, Z.; Xu, L.; Hu, J. Lossless data hiding in encrypted images compatible with homomorphic processing. *IEEE Trans. Cybern.* **2022**, *53*, 3688–3701.
14. Xiong, J.; Chen, J.; Lin, J.; Jiao, D.; Liu, H. Enhancing privacy-preserving machine learning with self-learnable activation functions in fully homomorphic encryption. *J. Inf. Secur. Appl.* **2024**, *86*, 103887.
15. Bai, Y.; Zhao, H.; Shi, X.; Chen, L. Towards practical and privacy-preserving cnn inference service for cloud-based medical imaging analysis: A homomorphic encryption-based approach. *Comput. Methods Programs Biomed.* **2025**, *261*, 108599.
16. Gu, X.; Luo, W.; Ryoo, M.S.; et al. Password-conditioned anonymization and deanonymization with face identity transformers. In Proceedings of the of European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 727–743.
17. Zhang, Y.; Wang, T.; Zhao, R.; et al. Rapp: Reversible privacy preservation for various face attributes. *IEEE Trans. Inf. Forensics Secur.* **2023**, *18*, 3074–3087.
18. Su, Z.; Zhang, G.; Shi, Z.; et al. Message-driven generative music steganography using midi-gan. *IEEE Trans. Dependable Secur. Comput.* **2024**, *21*, 5196–5207.
19. Song, J.; Ye, J.-C. Federated cyclegan for privacy-preserving image-to-image translation. *arXiv* **2021**, arXiv:2106.09246.
20. Li, D.; Wang, W.; Zhao, K.; et al. Riddle: Reversible and diversified de-identification with latent encryptor. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 8093–8102.
21. Karras, T.; Laine, S.; Aittala, M.; et al. Analyzing and improving the image quality of stylegan. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 8107–8116.