2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

DDSR: An Identity Preservation Framework for Facial Privacy Protection

Jingxian Zhou[1] , Wanying Zhao[2(✉)] , Shuang Wang[1] , and Ping Chen[3]

[1] Information Security Evaluation Center,Civil Aviation University of China,Tianjin 300300，China
[2] College of.Computer Science and Technology,Civil Aviation University of China, Tianjin 300300,China
Wanying778850@163.com
[3] Travelsky Technology Limited, Beijing 101318

Abstract. With the increasing risk of facial privacy leakage during image transmission, achieving both recognizability and privacy protection remains a major challenge. This paper proposes a novel diffusion-based framework for facial image privacy, termed Diffusion-Driven Steganography and Recovery (DDSR). DDSR utilizes a prompt-guided dual-phase diffusion strategy: during the steganography phase, identity prompts guide latent perturbation, and reverse diffusion under unrelated prompts generates semantically irrelevant stego images; during the recovery phase, the model re-encodes the stego image and reconstructs the original face under the guidance of the identity prompt. To enhance semantic alignment, we introduce a lightweight Prompt Consistency Regularization (PCR), which aligns recovered images and prompts in CLIP semantic space during training. This regularization improves prompt controllability without adding inference overhead. DDSR is compatible with low-resolution data and small-scale training, and does not rely on high-resolution inputs or large datasets. Extensive experiments demonstrate that DDSR achieves up to 98% face recognition accuracy on recovered images and outperforms prior methods by over 30% in resisting recognition attacks. Furthermore, DDSR provides improved robustness under image degradation while maintaining high visual quality and identity fidelity.

Keywords: Deep learning, facial privacy protection, diffusion models, image steganography.

1 Introduction

With the rapid development of computer vision and deep learning, facial recognition has been extensively adopted in identity verification, social media, and intelligent surveillance[1]. As a result, facial images have become highly sensitive personal data. However, most existing face recognition systems still transmit raw images or extracted features without sufficient encryption or obfuscation, leaving them vulnerable to interception and unauthorized recognition during transmission. Therefore, a critical

challenge is how to enhance privacy protection during image transmission without degrading recognition performance.

Existing face privacy protection methods can be broadly classified into two categories. The first is anonymization, which prevents identity recognition by modifying facial features through blurring, masking, or substitution, as seen in methods like Fawkes and Low-Key. The second is visual information hiding, which attempts to conceal identity features through adversarial perturbations, encryption, or embedding, while maintaining the image's natural appearance. However, these approaches are fundamentally contradictory: anonymization seeks to disable recognition, whereas information hiding relies on preserving it. As such, most current solutions require a trade-off between privacy and utility. A unified and controllable privacy-preserving framework that achieves concealment, naturalness, and identity recoverability remains a significant challenge[2].

To address this conflict, recent research has turned to image steganography[3], which aims to embed sensitive information into images in a way that is imperceptible to unauthorized receivers. In particular, coverless steganography[4], which generates visually unrelated images rather than embedding content into an existing one, has shown potential in improving concealment and transmission security. However, traditional steganography methods are mainly designed for bit-level message hiding and lack semantic-level representation capabilities, making them unsuitable for reversible face recovery. Even with generative models like GANs[5] and encoder-decoder models[6], existing methods still suffer from poor semantic controllability, limited realism, and weak robustness.

Recently, diffusion models have emerged as a promising alternative for image synthesis and privacy protection, due to their powerful generative capacity, symmetric noise modeling, and strong semantic guidance. Their inherent ability to progressively add and remove noise, guided by text prompts, offers new possibilities for building controllable and reversible privacy-preserving frameworks.

In this paper, we propose DDSR (Diffusion-Driven Steganography and Recovery), a novel diffusion-based framework for secure facial image transmission. DDSR leverages prompt-guided diffusion to convert identity images into semantically unrelated stego images, preventing recognition during transmission. At the receiver end, the original image can be faithfully reconstructed using only the correct identity prompt.DDSR offers key advantages: Security, by restricting recovery to authorized prompts; Reversibility, by preserving identity consistency; Robustness, under low resolution and lossy channels; and Controllability, through natural, style-flexible stego images. Extensive experiments on public face datasets demonstrate that DDSR outperforms existing methods in both image fidelity and privacy protection.

1. We propose DDSR, a reversible framework for facial privacy protection. It transforms facial images into semantically unrelated stego images and enables high-fidelity identity recovery via prompt-guided diffusion.
2. We develop a dual-phase diffusion strategy for controllable identity transformation. By combining DDIM inversion and prompt switching, DDSR supports both effective identity obfuscation and precise restoration.

3. 王4We introduce Prompt Consistency Regularization (PCR) to enhance semantic alignment. PCR aligns image-text embeddings in the CLIP space during training, enabling prompt-controllable and semantically faithful recovery.

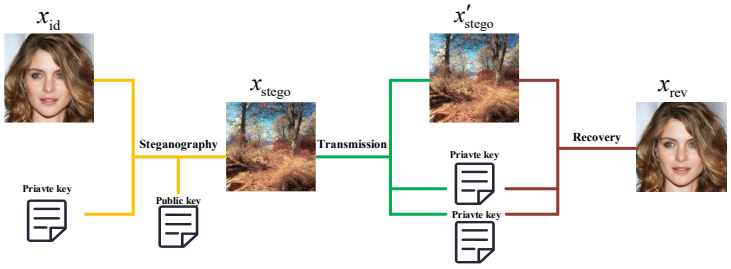

Fig. 1. The definition and composition of DDSR.

2 Revated works

2.1 Privacy protection in facial image data

Several methods for facial privacy protection have been proposed, primarily implemented through visual obfuscation, image generation, and adversarial perturbations. Mrityunjay et al.[7] introduced a system that reduces the recognizability of faces through filtering processes. Newton's K-Same method [8] computes an average facial image via rotation and cropping operations to obscure the original identity. TIP-IM [9] ensures privacy by overlaying adversarial identity masks on the face. With the advancement of deep learning technologies, researchers in the field of facial privacy protection have increasingly focused on Generative Adversarial Networks (GANs).Adv-Makeup [10] achieves adversarial protection through makeup, while AMT-GAN [11] generates adversarial facial images with makeup applied. CLIP2Protect [12] optimizes adversarial latent encoding by incorporating user-defined text prompts. The goal of these methods is to induce a loss of facial feature information, rendering the original facial data non-reusable. Haoxuan et all. [13] proposed a facial privacy enhancement method based on latent encoding optimization and facial masking, which maintains the usability of facial recognition while protecting privacy. Yang Yang et all. [14] presented a facial privacy protection method based on an Invertible Mask Network (IMN), capable of generating privacy-protected facial images and restoring the original facial information when necessary. However, due to the unnatural appearance of obscured faces, they are more easily detected by attackers. These methods fail to transmit facial images without raising suspicion. To address these issues, we draw on techniques from image steganography.

2.2 Image steganography

Image steganography [3] differs from traditional encryption techniques in that it aims to create an information carrier by embedding secret data within a host medium, thereby keeping the message concealed during transmission. Conventional image steganography typically involves designating a cover image into which secret information is embedded, making subtle modifications to render it difficult to detect. To achieve a more covert method of information hiding, coverless steganography [4] has emerged as a technique that avoids explicit alterations to the carrier image, thus reducing the risk of detection associated with traditional steganography. In recent years, Generative Adversarial Networks (GANs) have been integrated into coverless steganography, using adversarial training to generate images that make information hiding more discreet and harder to detect [15].However, coverless steganography still faces several challenges, such as limited steganographic capacity, difficulties in maintaining image quality, low robustness, and poor controllability. To address these issues, recent advances in generative models offer promising solutions.

2.3 Diffusion models

Diffusion models [16] are deep generative frameworks that progressively generate high-quality samples from random noise through a reverse denoising process, conditioned on specific inputs. Starting from Gaussian noise, the model iteratively reconstructs data that approximates the target distribution. Owing to their remarkable performance, diffusion models have been widely adopted in image generation, restoration, and translation tasks. For instance, SSIE-For example, SSIE-Diffusion [17] introduces a personalized conditional generation model to produce theme-relevant content, while, DINO-ViT [18] proposes a multi-reference translation approach for high-quality style transfer.Compared to traditional steganography techniques, diffusion models not only enable conditional image synthesis with better privacy preservation, but also exhibit superior robustness and controllability. These properties make them well-suited for tasks involving both image hiding and high-fidelity recovery.

3 Method

3.1 Preliminaries

Before introducing the DDSR framework, we first briefly review diffusion models and their latent-space extensions to establish the methodological foundation. Diffusion models [16] are a class of probabilistic generative models that aim to reconstruct the real data distribution by progressively denoising samples drawn from Gaussian noise. This process is typically formulated as a fixed-length Markov chain, where the forward process gradually adds noise to the original data until it becomes nearly pure noise, while the reverse process learns how to iteratively recover the original samples from noisy inputs.

In this work, we adopt the Stable Diffusion Model (SDM) [21] as the generative backbone, which performs diffusion and reverse diffusion in the latent space of images. Specifically, the input image x_{id} is first encoded into a latent representation z_0 using a pre-trained Variational Autoencoder (VAE). Noise is then added and removed within the latent space, and the image is finally reconstructed through the decoder. The forward noise addition process is defined as:

$$z_t = \sqrt{\alpha_t} \cdot z_0 + \sqrt{1-\alpha_t} \cdot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

where α_t is a time-dependent weighting coefficient.

For the reverse process, we adopt the deterministic sampling strategy proposed in Denoising Diffusion Implicit Models (DDIM), which eliminates the randomness in traditional diffusion and improves the stability of recovery. The update rule is defined as:

$$z_{t-1} = \sqrt{\alpha_{t-1}} \cdot \left(\frac{z_t}{\sqrt{\alpha_t}} - \frac{\sqrt{1-\alpha_t}}{\sqrt{\alpha_t}} \cdot \epsilon_\theta(z_t, t, C) \right) + \sqrt{1-\alpha_{t-1}} \cdot \epsilon_\theta(z_t, t, C) \tag{2}$$

where C denotes the embedding of the conditional prompt, and ϵ_θ is the learned denoising network.

In addition, to enable reversible transformation between images and latent noise, we incorporate DDIM Inversion, which maps an image backward into a noisy latent variable z_T at a chosen timestep. This serves as the basis for subsequent identity restoration in our framework.

3.2 Overview of the DDSR Framework.

To achieve privacy-preserving and reversible face image transmission, we propose a diffusion-based steganographic framework, DDSR (Diffusion-Driven Steganography and Recovery). The core idea is to transform identity images into semantically unrelated stego images through prompt-guided diffusion, and enable identity recovery on the receiver side via prompt-based inversion. As shown in **Fig. 1**, DDSR comprises three stages: Steganography, Transmission, and Recovery.

In the Steganography Phase, the sender encodes the input facial image x_{id} into a latent representation z_0 using a VAE. Under the guidance of the identity prompt $\text{prompt}_{\text{id}}$ (serving as a private key), the model performs DDIM inversion [19] to diffuse z_0 into an intermediate noisy state z_t, embedding identity information into the diffusion trajectory. Then, switching to a disguise prompt $\text{prompt}_{\text{stego}}$ (serving as a public key), the model reverses the diffusion process to generate a new latent variable z_0', which is decoded into the stego image x_{stego}. The result is visually unrelated to the original and semantically aligned with the disguise prompt (e.g., a landscape), making it suitable for open transmission.

In the Transmission Phase, the stego image x_{stego} is transmitted through public channels. During real-world transmission, it may undergo slight degradation due to compression or noise, resulting in a modified version x'_{stego}.Without the correct

identity prompt, it resists both human recognition and feature extraction by existing facial recognition models, thereby mitigating risks of privacy leakage and identity attacks.

In the Recovery Phase, the receiver provides x'_{stego} and the identity prompt $\text{prompt}_{\text{id}}$. The model re-encodes x'_{stego} into z'_0, performs DDIM inversion under $\text{prompt}_{\text{stego}}$ to obtain z'_t, and then applies reverse diffusion guided by $\text{prompt}_{\text{id}}$ to reconstruct the latent $\hat{z}_0$, which is finally decoded into the recovered image x_{rev}.

Here, $\text{prompt}_{\text{id}}$ can be considered a private key—it must be kept confidential to enable successful identity recovery. In contrast, $\text{prompt}_{\text{stego}}$ functions as a public key that relates to the stego image's appearance and can often be inferred from the image itself.

3.3 Steganography based on diffusion models.

This method is built upon a prompt-guided diffusion model, consisting of two main stages: a steganographic process and a recovery process. In our implementation, the two stages are respectively guided by an identity prompt and a disguise prompt, enabling the hiding and reconstruction of facial information through controllable diffusion trajectories. The complete workflow is illustrated in **Fig. 2**. The following sections provide a detailed explanation of both stages.

Steganography Phase: Prompt-Guided Identity Obfuscation. To enable identity hiding during image transmission, we propose a prompt-guided steganographic process in latent space. This phase aims to convert a facial image into a semantically irrelevant stego image, such that the identity is visually obfuscated while retaining semantic recoverability, as shown in **Fig. 3**. This stage consists of three key steps: image encoding, prompt-guided DDIM inversion, and denoising sampling with prompt switching

Image Encoding. The input facial image $x_{id} \in \mathbb{R}^{3\times H\times W}$ is first encoded into the latent space using the pre-trained VAE encoder $E(\cdot)$ from the stable diffusion model:

$$z_0 = E(x_{\text{id}}) \tag{3}$$

where $z_0 \in \mathbb{R}^{4\times\frac{H}{8}\times\frac{W}{8}}$ serves as the starting point of the subsequent diffusion process.

DDIM Inversion: Prompt-Guided Diffusion under $\text{prompt}_{\text{id}}$.Under the semantic guidance of an identity prompt $\text{prompt}_{\text{id}}$ (e.g., "a smiling man"), we apply the DDIM inversion technique to embed the latent variable z_0 into the diffusion trajectory, generating an intermediate state z_t. The process is defined as:

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\cdot z_t + \left(\sqrt{1-\alpha_{t+1}} - \sqrt{\frac{\alpha_{t+1}}{\alpha_t}}\cdot\sqrt{1-\alpha_t}\right)\cdot \epsilon_\theta\left(z_t, t, \text{prompt}_{\text{id}}\right) \tag{4}$$

where α_t is the noise attenuation coefficient at timestep t, and $\epsilon_\theta(z_t, t, \text{prompt}_{\text{id}})$ is the noise predicted by the U-Net under the condition of prompt $\text{prompt}_{\text{id}}$. This step is iteratively applied up to a predefined timestep T, resulting in the intermediate state z_T. Semantically, this process encodes identity-relevant information into a noise trajectory.

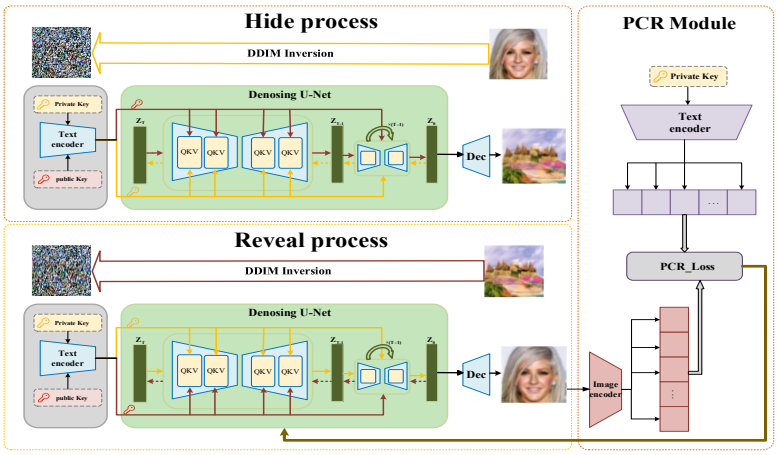

Fig. 2. Overview of the DDSR Framework.

Denoising Sampling with Prompt Switching. After obtaining the intermediate state z_T, we switch the prompt from $\text{prompt}_{\text{id}}$ to a semantically irrelevant disguise prompt $\text{prompt}_{\text{stego}}$ (e.g., "a landscape"), and perform DDIM reverse sampling to reconstruct the latent representation $\hat{z}_0$. Each update step is given by:

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \cdot z_t + \left(\sqrt{1-\alpha_{t-1}} - \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \cdot \sqrt{1-\alpha_t} \right) \cdot \epsilon_\theta \left(z_t, t, \text{prompt}_{\text{stego}} \right) \quad (5)$$

where $\epsilon_\theta \left(z_t, t, \text{prompt}_{\text{stego}} \right)$ denotes the noise prediction under the disguise prompt $\text{prompt}_{\text{stego}}$. This step is applied iteratively from $t = T$ down to $t = 0$, resulting in the stego latent variable $\hat{z}_0$.Finally, $\hat{z}_0$ is decoded by the VAE decoder $G(\cdot)$ to obtain the stego image x_{stego}:

$$x_{stego} = G(\hat{z}_0) \quad (6)$$

The resulting image x_{stego} is a visually natural image (e.g., a landscape) that bears no apparent similarity to the original face image. However, the identity-relevant information remains implicitly embedded in the latent representation, enabling reversibility.

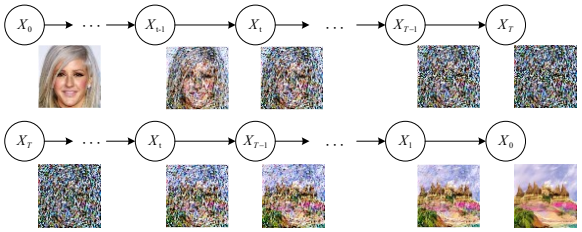

Fig. 3. Noise addition and denoising process performed by the sender.

Recovery Phase: Identity Restoration from the Stego Image. Given the identity prompt $\text{prompt}_{\text{id}}$, the goal of the recovery phase is to reconstruct a high-fidelity facial image x_{rev} from the stego image x'_{stego}. This process is structurally symmetric to the steganography phase, as illustrated in **Fig. 4**, and consists of three main steps: stego image encoding, DDIM inversion under prompt $\text{prompt}_{\text{stego}}$, and reverse sampling guided by $\text{prompt}_{\text{id}}$.

Stego Image Encoding. First, the stego image x'_{stego} is encoded into its latent representation using the VAE encoder:

$$\hat{z}_0 = E\left(x'_{stego}\right) \tag{7}$$

where $\hat{z}_0$ serves as the initial latent variable for the recovery process.

*DDIM Inversion.*We apply DDIM Inversion to $\hat{z}_0$ using the same disguise prompt prompt_{stego} that was used in the steganography phase, embedding the latent into a diffusion trajectory to obtain an intermediate state $\hat{z}_T$. The update step is given by:

$$\hat{z}_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} \cdot \hat{z}_t + \left(\sqrt{1-\alpha_{t+1}} - \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} \cdot \sqrt{1-\alpha_t} \right) \cdot \epsilon_\theta \left(\hat{z}_t, t, \text{prompt}_{stego} \right) \tag{8}$$

This process is iteratively applied up to $t = T$, resulting in the intermediate diffusion state $\hat{z}_T$ under the guidance of prompt_{stego}.

Prompt Switching: Reverse DDIM Sampling with $\text{prompt}_{\text{id}}$.Next, we switch the prompt back to the original identity prompt $\text{prompt}_{\text{id}}$ and perform reverse DDIM sampling starting from $\hat{z}_T$. Each update is defined as:

$$\tilde{z}_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \cdot \tilde{z}_t + \left(\sqrt{1-\alpha_{t-1}} - \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} \cdot \sqrt{1-\alpha_t} \right) \cdot \epsilon_\theta \left(\tilde{z}_t, t, \text{prompt}_{\text{id}} \right) \tag{9}$$

This process is applied iteratively from $t = T$ to $t = 0$, yielding the recovered latent variable $\tilde{z}_0$, which is then decoded to reconstruct the final image:

$$x_{rev} = G(\tilde{z}_0) \tag{10}$$

The recovered image x_{rev} contains identity features that are highly consistent with the original facial image, making it suitable for downstream tasks such as face recognition or verification.

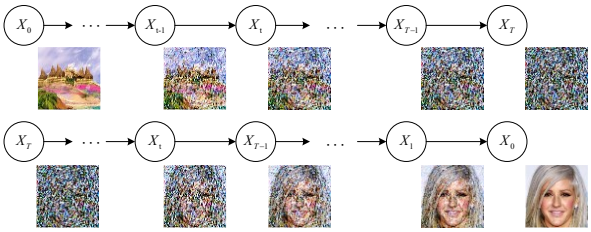

Fig. 4. Noise addition and denoising process performed by the receiver.

3.4 Prompt Consistency Regularization.

Although DDSR achieves controllable and identity-preserving image recovery through prompt switching, the diffusion model itself does not explicitly guarantee semantic consistency between the recovered image and the identity prompt p_1. To address this issue, we introduce a lightweight, unsupervised auxiliary module called Prompt Consistency Regularization (PCR), which is used during training to enhance the semantic alignment of the recovered images.

Specifically, we utilize a CLIP model [20] to extract the image embedding $f(x_{rev})$ from the recovered image x_{rev}, and the text embedding $\phi(\text{prompt}_{\text{id}})$ from the identity prompt $\text{prompt}_{\text{id}}$. We then compute the cosine similarity loss between them, defined as:

$$\mathcal{L}_{\text{PCR}} = 1 - \frac{\langle f(x_{rev}), \phi(\text{prompt}_{\text{id}}) \rangle}{\| f(x_{rev}) \| \cdot \| \phi(\text{prompt}_{\text{id}}) \|} \tag{11}$$

where $f(x_{rev})$ denotes the image features extracted by the CLIP image encoder, and $\phi(\text{prompt}_{\text{id}})$ denotes the semantic embedding of the identity prompt extracted by the CLIP text encoder. The denominator normalizes both embeddings to ensure a valid similarity measure.

This loss is only applied during training, encouraging the model to generate images that are more semantically aligned with the identity prompt. Importantly, it does not introduce any additional computation during inference. Through training-time guidance, the model implicitly learns this semantic constraint, enabling it to generate prompt-consistent recovered images at test time without requiring CLIP or loss computation.

4 Experiment

4.1 Experimental Setting

Experimental Setup. In our experiments, we use Stable Diffusion v1.5[21] as the training model and employ the deterministic DDIM sampling algorithm. Both the forward and reverse processes consist of 60 steps. All experiments were conducted on a GeForce RTX 3090 GPU.

Data Preparation. To quantitatively and qualitatively evaluate our approach, we conducted experiments using CelebA-HQ [22] , LADN [23], and LFW [24]. We selected a subset of 1,000 images from CelebA-HQ [22]. To assess the robustness of our method, we randomly selected 500 images from the public dataset for experimentation.

4.2 Reversibility

Fig. 5 presents the experimental results of face image hiding and recovery. From left to right, the images correspond to the secret image x_{id}, container image x_{stego}, and recovered image x_{rev}, where each secret image is paired with a corresponding container image. In our experiments, the container images are all landscape images unrelated to faces, effectively concealing facial information in a visually imperceptible manner. As a result, the human eye cannot detect any facial features embedded within the container image. Moreover, adversaries find it challenging to infer the presence of facial privacy information from the transmitted container images, significantly reducing the risk of information interception and misuse during transmission.

Fig. 5. Visualization of the secret and recovered images generated by our method.

To further validate the reversibility of the proposed method, we analyze the similarity between the hidden images and the recovered images to determine whether the recovered images retain characteristics similar to the original images in face recognition tasks. The detailed comparison results are presented in **Fig. 6**.

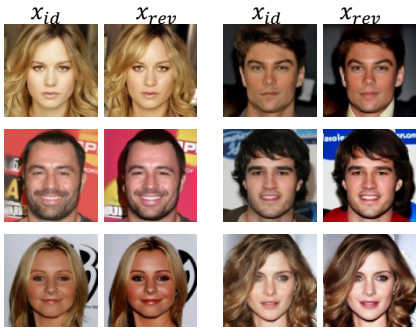

Fig. 6. Visualization of restored images compared with originals using our method.

In the experiments, four classical face recognition models—IRSE50 [25], IR152 [26], FaceNet [27], and MobileFace [28], —and three commercial face recognition APIs, namely Tencent Cloud, Face++, and Aliyun, were selected. These models and APIs were used to recognize faces by returning a confidence score between 0 and 100, with a higher score indicating stronger similarity. It is important to note that the training data and model parameters for these proprietary face recognition models are undisclosed, which more realistically simulates real-world application scenarios, thus enhancing the credibility of the experimental results.**Fig. 7** presents the evaluation results on the CelebA-HQ [22] and LADN [23] datasets.The experimental results demonstrate that the similarity confidence achieved by the proposed method exceeds 80% in most cases, reaching over 98% in some instances. This fully validates the reversibility of the proposed method, i.e., the recovered images maintain a high level of recognizability in face recognition tasks.

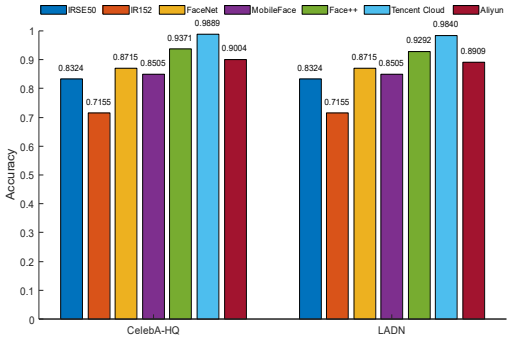

Fig. 7. Average confidence scores (higher is better) returned by commonly used face recognition models and real-world face verification APIs for evasion attacks.

To quantitatively evaluate the similarity between the original and recovered images, we report the PSNR and SSIM metrics in **Table 1**. We compare our method with recent noise-based and makeup-based facial privacy preservation techniques, including TIP-IM [9], AMT-GAN [11] and CLIP2Protect [12]. As shown in the results, our method achieves the best performance in both PSNR and SSIM, outperforming all baseline methods.

Table 1. Comparison of PSNR and SSIM for Different Methods

Method	PSNR	SSIM
TIP-IM[10]	33.21	0.92
AMT-GAN[11]	19.50	0.79
CLIP2Protect[12]	19.31	0.75
ours	35.11	0.93

4.3 Security

To demonstrate the steganalysis resistance of our container images, we evaluate the robustness of the proposed method using the StegExpose [29] detector. In our experiments, we generated container images based on the CelebA-HQ [22] and LADN [23] datasets,and plotted receiver operating characteristic (ROC) curves by adjusting various thresholds. The experimental results, shown in **Fig. 8**, indicate that the ROC curves for our method are close to the diagonal of random guessing. This suggests that the method approaches the ideal detector-avoidance state and effectively conceals the original image, thereby further verifying its steganographic and security performance.

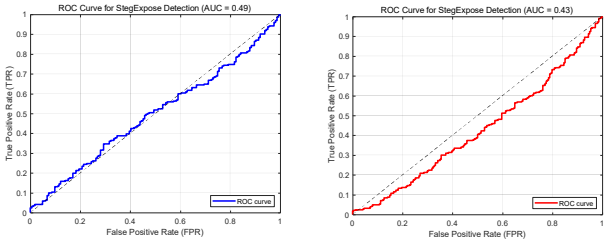

Fig. 8. ROC curves on CelebA-HQ (accuracy: 0.49, left) and LADN (accuracy: 0.43, right).

To assess the effectiveness of our generated container images in resisting recognition by black-box face recognition systems, we performed face verification experiments on the LFW dataset. The comparison methods include TIP-IM [9], Adv-Makeup [10], AMT-GAN [11] ,CLIP2Protect [12], MI-FGSM [30] and TI-DIM [31]. The results, shown in **Table 2**, provide the evasion attack values for the face recognition task in an untargeted attack scenario. Our method shows significant improvements of 9.9% and 30.9% in Rank-1 and Rank-5 evaluation metrics, respectively, compared to the

CLIP2Protect method, demonstrating its clear advantage in resisting attacks on black-box recognition systems.

Table 2. Protection success rate (PSR %) for black-box dodging on the LFW dataset.

Method	IRSE50		IR152		FaceNet		MobileFace		Average	
	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U	R1-U	R5-U
Adv-Makeup [10]	45.2	38.5	40.9	35.3	41.2	35.9	43.5	36.7	42.7	36.6
MI-FGSM [30]	70.2	42.6	58.4	41.8	59.2	34.0	68.0	47.2	63.9	41.4
TI-DIM [31]	79.0	51.2	67.4	54.0	74.4	52.0	79.2	61.6	75.0	54.7
TIP-IM [9]	81.4	52.2	71.8	54.6	76.0	49.8	82.2	63.0	77.8	54.9
AMT-GAN [11]	83.7	55.6	72.0	54.8	78.9	50.1	83.6	64.9	79.6	56.4
CLIP2Protect [12]	86.6	59.4	73.4	56.6	83.8	51.2	85.0	66.8	82.2	58.5
ours	**90.4**	**85.9**	**93.6**	**92.5**	**93.4**	**91.5**	**91.1**	**87.7**	**92.1**	**89.4**

We report the FID scores of our method in **Table 3**, which assess the naturalness of the generated container images. Adv-Makeup has the lowest FID score, as it only applies makeup to the eye region without altering the rest of the face. However, this limitation results in a poorer PSR for evading attacks. Compared to TIP-IM [9] and AMT-GAN [11], our method shows a lower FID score, suggesting that our container images exhibit stronger naturalness and can protect facial privacy in a more natural manner.

Table 3. Comparison of FID and PSR gain (relative to AdvMakeup).

Method	FID	PSR Gain	
		R1-U	R5-U
Adv-Makeup [10]	4.23	0	0
TIP-IM [9]	38.73	35.1	18.3
AMT-GAN [11]	34.44	36.9	19.8
CLIP2Protect [12]	26.62	39.5	21.9
ours	32.96	49.4	52.8

4.4 Robustness

The secret images generated by our method are primarily transmitted over public channels, which are typically lossy. Therefore, the robustness of the coverless steganography scheme directly impacts the receiver's ability to accurately extract the hidden secret information from the image. As a result, evaluating robustness becomes a crucial metric for assessing the effectiveness of the method.

To evaluate the robustness of our approach, we conduct a series of degradation experiments on the generated secret images, including the addition of Gaussian noise, JPEG compression, and pretzel noise. The specific parameter settings for each

experiment are outlined in **Table 4**. We compare our method, with recent image steganography techniques,RIIS [32], HiNet [33], Baluja [34] , ISN [35] and CRoSS [36].

Table 4. Parameter Settings under Different Processing Conditions

Gaussian noise			JPEG compression			Salt-and-pepper noise	
σ=10	σ=20	σ=30	Q=20	Q=40	Q=80	P=0.01	P=0.02

Our experimental results are shown in **Fig. 9**. As can be seen from the figure, our method demonstrates excellent adaptability under various levels of degradation, with the smallest performance decline compared to other methods, whose fidelity significantly decreases. Specifically, although our method did not achieve the best performance when the container image was undamaged, it exhibited superior robustness as the degradation severity increased, remaining almost unaffected. In contrast, the PSNR values of other methods dropped significantly, with HiNet and RIIS showing the most pronounced decline. Ultimately, our method achieved the best PSNR in all experiments.

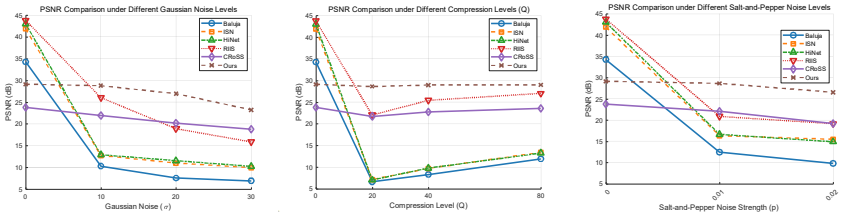

Fig. 9. PSNR (dB) of DDSR and baselines under varying degradation levels.

4.5 Ablation Study

To evaluate the effectiveness of the Prompt Consistency Regularization (PCR) module, we retrain the model with PCR removed and compare it against the full model in terms of semantic consistency and identity preservation of the recovered images. As shown in **Fig. 10**, removing PCR leads to a notable decrease in CLIP similarity between the recovered images and the identity prompts, along with a slight drop in face recognition similarity.

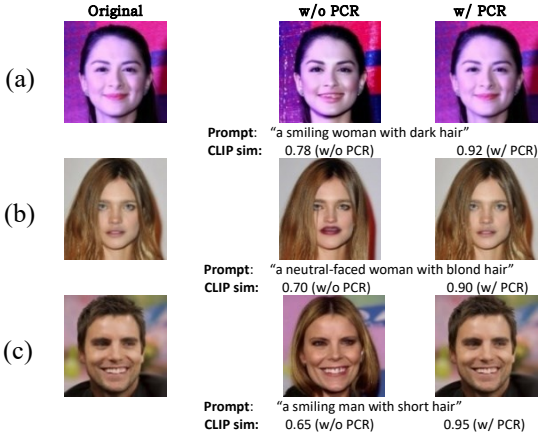

Fig. 10. Recovered image comparison with and without PCR. Rows: original, with-PCR, without-PCR.

In addition to the subjective comparison, **Table 5** reports the quantitative evaluation of our method with and without the Prompt Consistency Regularization (PCR) module. Three widely used metrics—PSNR, SSIM, and FID—are employed to measure the fidelity and quality of the recovered images. As shown in the table, incorporating PCR significantly improves image quality: PSNR increases from 28 to 35.11, SSIM rises from 0.87 to 0.93, and FID decreases from 34 to 32.96. These results indicate that PCR effectively guides the model to generate more semantically aligned images with higher fidelity to the original identity and overall visual quality. This improvement highlights the positive impact of semantic consistency supervision on identity preservation and visual realism.

Table 5. Quantitative results with and without PCR.

Method	PSNR	SSIM	FID
Ours-w/o	28	0.77	34
Ours-w	35.11	0.93	32.96

5 CONCLUSION

We propose a new method that combines a diffusion-based generative model with image steganography in order to achieve stealthy and effective protection during face image transmission while preserving the functionality of face recognition. Experimental results show that the method is capable of generating high-quality protected images, is significantly resistant to recognition attacks during transmission, and exhibits superior performance in terms of robustness. Limitations of our approach include high computational time cost in generating protected images.

Acknowledgments. This research was supported by the fundamental research funds for the central universities (3122025094).

References

1. Kortli, Y., Jridi, M., Al Falou, A., Atri, M.: Face recognition systems: A survey. Sensors **20**(2), 342 (2020).
2. Yuan, L., Chen, W., Pu, X., Zhang, Y., Li, H., Zhang, Y., Ebrahimi, T.: PRO-Face C: Privacy-preserving Recognition of Obfuscated Face via Feature Compensation. IEEE Trans. Inf. Forensics Secur. (2024).
3. Chanu, Y.J., Singh, K.M., Tuithung, T.: Image steganography and steganalysis: A survey. Int. J. Comput. Appl. (2012).

4. Zhou, Z., Sun, H., Harit, R., Chen, X., Sun, X.: Coverless image steganography without embedding. In: Proc. ICCCS, pp. 1–6 (2015).
5. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV, pp. 2223–2232 (2017).
6. Qin, J., Luo, Y., Xiang, X., Tan, Y., Huang, H.: Coverless image steganography: A survey. IEEE Access **7**, 56373–56390 (2019).
7. Mrit, M., Narayanan, P.: The de-identification camera. In: NCVPRIPG, pp. 192–195 (2011).
8. Newton, E.M., Sweeney, L., Malin, B.: Preserving Privacy by De-identifying Face Images. IEEE Trans. Knowl. Data Eng. **17**(2), 232–243 (2005).
9. Yang, X., Dong, Y., Pang, T., Su, H., Zhu, J., Chen, Y., Xue, H.: Towards Face Encryption by Generating Adversarial Identity Masks. In: ICCV, pp. 3897–3907 (2021).
10. Yin, B., Wang, W., Yao, T., Guo, J., Kong, Z., Ding, S., Li, J., Liu, C.: Adv-Makeup: A New Imperceptible and Transferable Attack on Face Recognition. In: IJCAI, pp. 1234–1240 (2021).
11. Hu, S., Liu, X., Zhang, Y., Li, M., Zhang, L.Y., Jin, H., Wu, L.: Protecting Facial Privacy via Style-Robust Makeup Transfer. In: CVPR, pp. 14994–15003 (2022).
12. Shamshad, F., Naseer, M., Nandakumar, K.: CLIP2Protect: Protecting Facial Privacy Using Text-Guided Makeup via Adversarial Latent Search. In: CVPR, pp. 20595–20605 (2023).
13. Tai, H., Zhang, Y., Guo, B., Luo, G., Zhu, Y.: High Resolution Face Privacy-Enhancing Method Based on Latent Optimization with Identity-Preserving Facial Masking. In: IJCNN, pp. 1–9 (2024).
14. Yang, Y., Huang, Y., Shi, M., Chen, K., Zhang, W., Yu, N.: Invertible Mask Network for Face Privacy-Preserving. arXiv preprint arXiv:2204.08895 (2022).
15. Lu, S.P., Wang, R., Zhong, T., Rosin, P.L.: Large-Capacity Image Steganography Based on Invertible Neural Networks. In: CVPR, pp. 10811–10820 (2021).
16. Guo, L., Liu, M., Fu, K., Zhou, J.: SSIE-Diffusion: Personalized Generative Model for Subject-Specific Image Editing. In: IJCNN, pp. 1–8 (2024).
17. Xu, X., Xiao, J., Zhang, X., Jin, X., Gu, X., Xu, G.: DINO-ViT Enhanced Diffusion for Multi-Exemplar-Based Image Translation. In: IJCNN, pp. 1–9 (2024).
18. Ho, J., Jain, A., Abbeel, P.: Denoising Diffusion Probabilistic Models. In: NeurIPS (2020).
19. Song, J., Meng, C., Ermon, S.: Denoising Diffusion Implicit Models. In: ICLR (2021).
20. Radford, A., et al.: Learning Transferable Visual Models From Natural Language Supervision. In: ICML (2021).
21. Rombach, R., Blattmann, A., Lorenz, D., et al.: High-Resolution Image Synthesis with Latent Diffusion Models. In: CVPR, pp. 10684–10695 (2022).
22. Karras, T.: Progressive Growing of GANs for Improved Quality, Stability, and Variation. arXiv preprint arXiv:1710.10196 (2017).
23. Gu, Q., Wang, G., Chiu, M.T., et al.: LADN: Local Adversarial Disentangling Network for Facial Makeup and De-Makeup. In: ICCV, pp. 10481–10490 (2019).
24. Huang, G.B., Mattar, M., Berg, T., et al.: Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. In: Workshop on Real-Life Face Detection and Recognition (2008).
25. Hu, J., Shen, L., Sun, G.: Squeeze-and-Excitation Networks. In: CVPR, pp. 7132–7141 (2018).
26. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In: CVPR, pp. 4690–4699 (2019).
27. Schroff, F., Kalenichenko, D., Philbin, J.: FaceNet: A Unified Embedding for Face Recognition and Clustering. In: CVPR, pp. 815–823 (2015).
28. Chen, S., Liu, Y., Gao, X., Han, Z.: MobileFaceNets: Efficient CNNs for Accurate Real-Time Face Verification on Mobile Devices. In: Biometric Recognition, CCBR, pp. 428–438. Springer (2018).

29. Boehm, B.: StegExpose: A Tool for Detecting LSB Steganography. arXiv preprint arXiv:1410.6656 (2014).
30. Madry, A.: Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv preprint arXiv:1706.06083 (2017).
31. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting Adversarial Attacks with Momentum. In: CVPR, pp. 9185–9193 (2018).
32. Baluja, S.: Hiding Images within Images. IEEE Trans. Pattern Anal. Mach. Intell. **42**(7), 1685–1697 (2019).
33. Lu, S.P., Wang, R., Zhong, T., Rosin, P.L.: Large-Capacity Image Steganography Based on Invertible Neural Networks. In: CVPR, pp. 10816–10825 (2021).
34. Jing, J., Deng, X., Xu, M., Wang, J., Guan, Z.: HiNet: Deep Image Hiding by Invertible Network. In: ICCV, pp. 4733–4742 (2021).
35. Xu, Y., Mou, C., Hu, Y., Xie, J., Zhang, J.: Robust Invertible Image Steganography. In: CVPR, pp. 7875–7884 (2022).
36. Yu, J., Zhang, X., Xu, Y., Zhang, J.: CROSS: Diffusion Model Makes Controllable, Robust and Secure Image Steganography. In: NeurIPS **36** (2024).