SteganoDDPM: A high-quality image steganography self-learning method using diffusion model

Mengnan Qu, Yuhao Jin, Guanghua Zhang
Hebei University of Science and Technology
qumengnan@quanfita.cn, jyh8888@88.com, zhanggh@hebust.edu.cn

***Abstract*—Image steganography has become a focal point of interest for researchers due to its capacity for the covert transmission of sensitive data. Traditional diffusion models often struggle with image steganography tasks involving paired data, as their core principle of gradually removing noise is not directly suited for maintaining the correspondence between carrier and secret information. To address this challenge, this paper conducts an in-depth analysis of the principles behind diffusion models and proposes a novel framework for an image steganography diffusion model. The study begins by mathematically representing the steganography tasks of paired images, introducing two optimization objectives: minimizing the secrecy leakage function and embedding distortion function. Subsequently, it identifies three key issues that need to be addressed in paired image steganography tasks and, through specific constraint mechanisms and optimization strategies, enables the diffusion model to effectively handle paired data. This enhances the quality of the generated stego-images and resolves issues such as image clarity. Finally, on public datasets like CelebA, the proposed model is compared with existing generation model-based image steganography techniques, analyzing its implementation effects and performance parameters. Experimental results indicate that, compared to current technologies, the model framework proposed in this study not only improves image quality but also achieves significant enhancements in multiple performance metrics, including the imperceptibility and anti-detection capabilities of the images. Specifically, the PSNR of its stego-images reaches 93.14dB, and the extracted images' PSNR reaches 91.23dB, an approximate improvement of 30% over existing technologies; the attack success rate is reduced to 2.4×10^{-38}. These experimental outcomes validate the efficacy and superiority of the method in image steganography tasks.**

I. Introduction

In the digital age, information security has become a focal point of societal concern. With the widespread adoption of the internet and the explosive growth of digital media, traditional encryption technologies can no longer meet the ever-increasing security demands. Information hiding techniques [10], particularly image steganography, as a significant branch within the realm of information security, offer new solutions for the covert transmission of sensitive data. The essence of image steganography lies in embedding secret information into digital images, rendering it visually imperceptible, thereby facilitating the clandestine transfer of information.

As deep learning technology rapidly evolves, image processing methods based on generative models have demonstrated exceptional performance across various domains. Deep learning models such as Generative Adversarial Networks (GANs) [7], [12], [28]and Variational Autoencoders (VAEs) [8], [31] have achieved remarkable success in image generation [19], editing [4], and restoration [27]. These models are capable of learning complex data distributions and generating high-quality images, paving new possibilities for image steganography.

However, despite the tremendous success of deep learning in image processing, its application in the field of image steganography still faces numerous challenges. Primarily, generative models must possess a high degree of flexibility and robustness to accommodate different types of images and varying levels of concealment requirements. Furthermore, the quality of the generated images must be sufficiently high to avoid arousing suspicion among attackers. Additionally, effectively balancing the trade-off between stealthiness and image quality remains an important research question.

Among them, diffusion models [3], [6], [21], [22] have attracted considerable attention for their excellent performance in generating high-resolution and high-quality images.However, despite the significant achievements of diffusion models in the field of image generation, they are typically not considered suitable for carrier-secret paired data image steganography tasks. The primary reason is that the core principle of diffusion models is to gradually remove noise to generate images, which is not compatible with traditional steganography techniques.

This study aims to challenge this conventional view by conducting an in-depth analysis of the workings of diffusion models and exploring how to combine the diffusion process with steganography techniques tailored to the unique requirements of image steganography tasks. We propose a novel framework that not only retains the ability of diffusion models to generate high-quality images but also effectively embeds information into images for the purpose of steganography. By breaking down the steganography task into stages, we meticulously designed each component of the model to ensure efficient information embedding and extraction without compromising image quality.

The significance of this study lies in two aspects: firstly, it expands the application scope of diffusion models, providing a new perspective and method for image steganography; secondly, by combining generative models with steganography techniques, our approach is expected to enhance the security and imperceptibility of steganography, holding significant practical value for scenarios requiring highly confidential communication. Furthermore, the findings of this research offer new insights and potential research directions for the future integration of deep learning with information hiding technologies, contributing to the advancement and innovation in related fields.

II. Related Works

In recent years, with the rapid development of deep learning technologies, image steganography methods based on deep neural networks have gradually emerged, bringing new breakthroughs to traditional steganography techniques. Gyojin et al. [5] proposed a deep cross-modal steganography framework that uses implicit neural representations (INRs) to hide various formats of secret data within cover images. Mengnan et al. [17] introduced an image steganography and extraction scheme based on implicit symmetric generative adversarial networks, utilizing two sets of generative adversarial networks to form a zero-sum game relationship and reducing the risk of steganographic data leakage by optimizing the loss function. Guobiao et al. [11] suggested disguising the steganography network as a steganography deep neural network model performing ordinary machine learning tasks. During the model camouflage process, they selected and adjusted a subset of filters in the secret DNN model to preserve its functionality on secret tasks, while the remaining filters were reactivated according to partial optimization strategies, disguising the secret DNN model as an implicit DNN model. Despite the significant progress made in the field of deep image steganography by the above research, issues such as image quality loss and unclear image boundaries still occur during the steganography process.

On the journey of exploring digital steganography, some researchers have also noticed the remarkable effects of diffusion models in the field of image generation. Ping Wei et al. [25] proposed a steganalysis diffusion model that utilizes non-Markov chains and fast sampling techniques to achieve efficient stego image generation. It constructs an ordinary differential equation (ODE) based on the transition probabilities of the generation process in steganalysis diffusion, and uses an approximate solver of the ODE - Euler iteration formula to interconvert steganographic data and stego images, enabling the use of irreversible but more expressive network structures to achieve model reversibility. Yinyin Peng et al. [16] exploited the probability distribution between intermediate states in the reverse process of the diffusion model and the generated image, hiding secret messages within the generated images through message sampling, following the same probability distribution as normal generation. Although the above achievements applied diffusion models to image steganography tasks, their form of steganography was carrier-less, with the carrier image being randomly generated during the image generation process, rather than the traditional sense of carrier-secret paired data image steganography tasks.

III. SteganoDDPM Method

This study firstly provided a description of the carrier-secret paired data image steganography task, identifying two optimization objective functions. Secondly, it briefly introduced the prior knowledge required by the method. Lastly, it analyzed three issues that need to be addressed when applying diffusion models to image steganography tasks and resolved these issues through practical solutions, resulting in a complete training and prediction workflow for the image steganography denoising diffusion model.

A. Representation for Image Steganography

In the carrier-secret paired data image steganography task, there are two inputs: the carrier image A and the secret image B to be hidden. A function S is defined to perform the steganography process, where the output of $S(A, B)$ is the new image A^* that hides the secret image B. The goal is for the new image A^* to be as close as possible to the original image A, meaning that the smaller the difference $\|A - A^*\|$, the better the steganography effect. Therefore, a minimization function for secret leakage is defined to optimize the steganography process, as equation (1).

$$\min \|A - S(A, B)\| \tag{1}$$

Upon completion of the steganography task, the hidden secret image B is extracted from the stego result A^*. A function F is defined, where its input is the stego image A^*, and its output is the extracted secret image B^*. Theoretically, this process aims for the extracted secret image B^* to be as close as possible to the original secret image B, meaning that the smaller the difference $\|B - B^*\|$, the better the extraction result. Therefore, a minimization function for embedding distortion is defined to optimize the extraction process, as equation (2).

$$\min \|B - F(S(A, B))\| \tag{2}$$

This problem can also be described as a game where the distance between A^* and B^* and A and B will determine the winner. If A^* is closer to A, A gets a payoff of 1; if B^* is closer to B, B gets a payoff of 1. A uses method S to make A^* gradually approach A, and B uses method M to make B^* gradually approach B. In addition, there is a correlation between A^* and B^*, that is, when B^* is closer to B, the distance between A^* and B^* is farther; when A^* is closer to A, the distance between B^* and A^* is closer.

Participant A can use method S to gradually approach A^*, but since there is a correlation between A^* and B^*, A needs to consider whether B^* may also be approached. Since when B^* is closer to B, A^* is farther away from D, so A should avoid method S getting too close to A^* to prevent B^* from getting too close to B and causing A^* to become too far away from B^* .

Similarly, participant B can use method M to gradually approach B^*, but since there is a correlation between A^* and B^*, B needs to consider whether A^* may also be approached. Since when A^* is closer to A, B^* is closer to A^*, so B should use method M to approach B^* as much as possible, and ensure that B^* does not get too close to B, so that A^* and B^* Keep the distance within a reasonable range.

Suppose B and B wish to maximize their payoff, where the payoff depends on the distance between A^* and B^* from them. Let the distance between A and A^*be d_{AC}, the distance between A^* and B^* be $d_{A^*B^*}$, and the distance between B and B^* be d_{BB^*}. Then the income of A and B can be expressed as: $P_A = f(d_{AA^*})$ and $P_B = f(d_{BB^*})$ respectively.

Let S and M denote the policy functions of A and B respectively, then it can be expressed as:

$$d_{AA^*}(t+1) = S(d_{AA^*}(t), d_{BB^*}(t)) \tag{3}$$

$$d_{BB^*}(t+1) = M(d_{BB^*}(t), d_{AA^*}(t)) \tag{4}$$

where t represents the time step. Both S and M are functions of two distances, representing the change in distance in the next time step. Let F be a function of the distance between A and A^*, and G be a function of the distance between A^* and B^*, then:

$$F(d_{AA^*}) = d_{AA^*} \tag{5}$$

$$G(d_{A^*B^*}) = -d_{A^*B^*} \tag{6}$$

So the problem is transformed into a problem of minimizing F and maximizing G.

B. DDPM

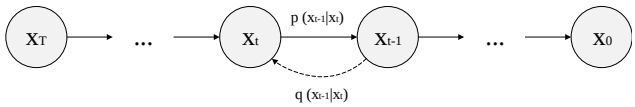

Fig. 1. Schematic diagram of the forward noise addition and reverse denoising process of the noise reduction diffusion model.

This study commences with the Denoising Diffusion Probabilistic Model (DDPM) [3], [6], [21], [22], and this article elucidates the fundamental principles of DDM in accordance with the literature [14], [15].

As depicted in Figure 1, the process from x_0 to x_T represents the forward noise addition process. The diffusion commences with a clean data point x_0 and subsequently adds noise to it in sequence. The data point x_0 follows the distribution $q(x_0)$. For each time step t within the interval $[1, T]$, the relationship between x_t and its preceding state $x_(t-1)$ is defined as follows:

$$x_t = \sqrt{1-\beta_t}x_{t-1} + \sqrt{\beta_t}\epsilon \tag{7}$$

In the equation, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ denotes the noise sampled from a Gaussian distribution, where β_t is a fixed constant that increases with t. Owing to the additive property of normal distributions, the noisy data x_t at a specified time step t is expressed as:

$$x_t = \gamma_t x_0 + \sigma_t \epsilon \tag{8}$$

Herein, γ_t and σ_t respectively define the scaling factors for the signal and the noise, satisfying the relationship $\gamma_t^2 + \sigma_t^2 = 1$ as cited in [14], [15]. By setting $\alpha_t = 1 - \beta_t$, then

$$\gamma_t = \sqrt{\prod_{i=1}^{t} \alpha_i} \tag{9}$$

The denoising diffusion model learns how to remove noise during the backward denoising process. Conditioned on the time step t, the model employs Bayes' theorem to compute the posterior probability:

$$p(x_{t-1}|x_t, x_0) = \frac{p(x_t|x_{t-1}, x_0)p(x_{t-1}|x_0)}{p(x_t|x_0)} \tag{10}$$

From equation (4), we can derive

$$p(x_{t-1}|x_t) = \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\gamma_t}\epsilon\right), \frac{\beta_t\gamma_{t-1}}{\gamma_t^2}\right) \tag{11}$$

Contrarily to the original DAE [23] which predicts a clean input, modern DDPMs often predict the noise ϵ The loss function for this formulation is minimized as follows:

$$\|\epsilon - \mathrm{net}(x_t)\|^2 \tag{12}$$

In the equation, $\mathrm{net}(x_t)$ represents the output of the neural network. Given a noise schedule conditioned on the time step t, the network is trained across multiple levels of noise. During generation, the trained model is iteratively applied until the clean signal x_0 is reached. As depicted in Figure 2, this illustrates the classical DAE adding and predicting noise in the image space.

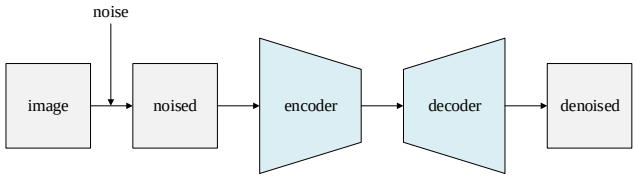

Fig. 2. Schematic diagram of the classic DAE model for adding and predicting noise in image space.

C. Steganographic Denoising DAE Model

In the context of image steganography, the key lies in maintaining the visual quality and original characteristics of the carrier image while covertly encoding information into the image. This often requires specific algorithms to make subtle adjustments to the pixel values, colors, or other features of the image, thereby concealing data imperceptibly. Unlike image steganography, the goal of DDPM models is to learn data distributions and generate high-quality new images through a process of gradual noise addition and removal. Therefore, although DDPM excels in image generation, it is not directly applicable to image steganography tasks as it is not specially designed for the fine control and stealth required for such tasks. This study proposes a steganographic denoising diffusion model based on the DAE model structure, as shown in Figure 3.

Integrating DDPM models into image steganography tasks presents three main challenges: First, the issue of image input; existing DDPM models typically handle multiple inputs by encoding information into the encoding results of the DAE encoder, thereby affecting its output. However, the output is uncertain because DDPM can only ensure the generation of images that conform to the target distribution, not necessarily their closeness to the carrier image. Second, adapting the image steganography process to the training framework of the DDPM model is necessary. The image steganography process involves a multi-step learning procedure, including the primary steps of embedding and extraction. In deep learning-based image steganography, embedding and extraction are usually divided into two different models that are jointly or separately trained. Their training and validation processes constitute a complete image processing workflow, whereas

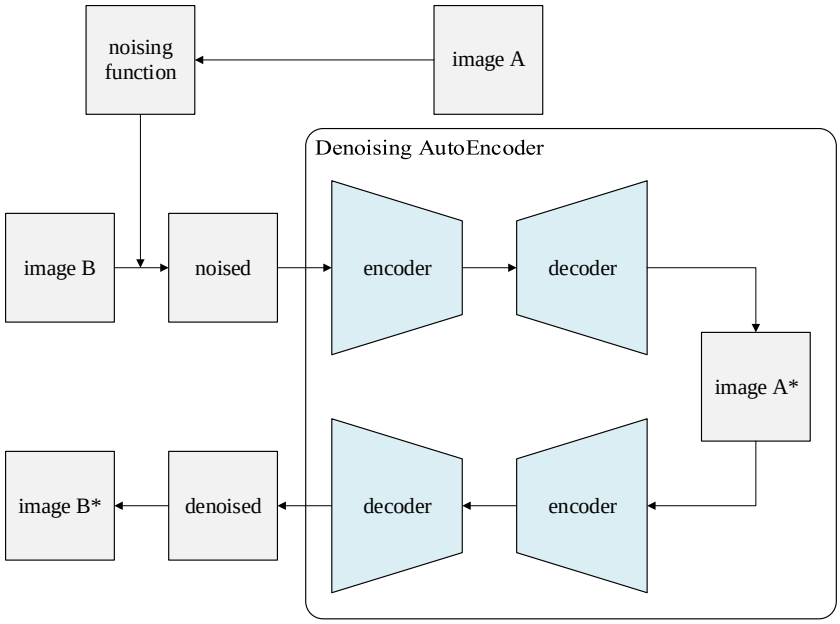

Fig. 3. SteganoDAE involves attaching the carrier image A in the form of noise to the secret image B, entering the entire denoising diffusion process. The primary process remains adding noise ϵ to image B and using DAE to learn how to remove noise ϵ, ultimately resulting in image B^*.

the DDPM model's training process is not a complete image processing workflow, as the model only learns the noising process between adjacent states. Third, there is the design problem of the objective function. Deep learning-based image steganography typically uses a distance function between the model's generated results and the carrier and stego images as the objective function, whereas the DDPM model's objective function uses the distance between the prediction result and the noise ϵ.

1) Input to SteganoDAE Model: This method of noise addition breaks the conventional notion in deep-learning-based image steganography, which traditionally takes the carrier image as the primary input and embeds the secret image upon it. When the carrier image A is embedded into the secret image B as noise, the entire denoising process enhances the model's ability to restore hidden information. This implies that even in the face of disturbances such as image compression and transmission distortion, the embedded information is more likely to be fully recovered. Therefore, this study defines a noise function $\theta(y)$ for transforming the carrier image into noise:

$$\theta(y) = \epsilon y \tag{13}$$

where ϵ is sampled from a Gaussian distribution $\mathcal{N}(0, \mathbf{I})$, and y represents the input image. At time instances $t \in [1, T]$, x_t and x_{t-1} satisfy the following relationship:

$$x_t = \sqrt{1-\beta_t} x_{t-1} + \sqrt{\beta_t}\theta(y) \tag{14}$$

Then, at the specified time step t, the noisy data x_t is represented as:

$$x_t = \gamma_t x_0 + \sigma_t \theta(y) \tag{15}$$

This strategy allows for a more flexible approach to information embedding, preserving as much detail of the secret image as possible, and offers insights for exploring richer methods of information embedding in subsequent studies.

2) Training and Prediction of SteganoDDPM: The steganographic denoising DAE model consists of two serial autoencoders, where one encoder is responsible for recovering the carrier image A from embedded noise, with information from the secret image B remaining in its intermediate result image A^*. Subsequently, another encoder is tasked with recovering the secret image B from the intermediate result image A^* using the residual information of the secret image B, resulting in the image B^*.

During the training process, the data flow is opposite to that of the prediction process, as shown in Figure 4. Initially, as formula (13) indicates, the noisy secret image B is fed into AutoencoderB to generate the intermediate result image A^*. Following this, the intermediate result image A^* is input into AutoencoderA to produce a denoised image C, making it similar to the unnoised image B.

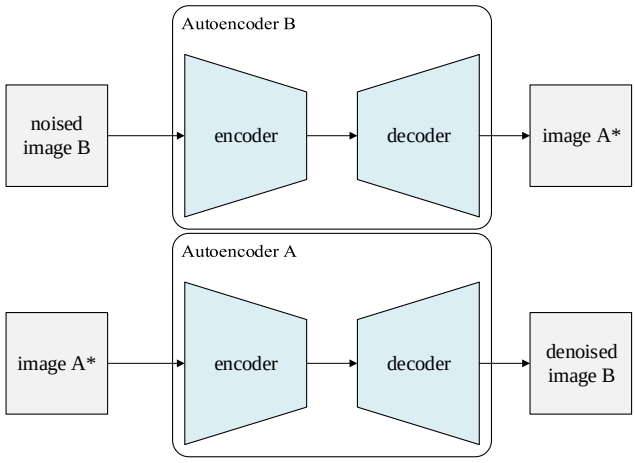

Fig. 4. Schematic diagram of the classic DAE model for adding and predicting noise in image space.

The prediction process follows Algorithm 1, taking a noiseless secret image B and carrier image A as input, starting a reverse cycle traversal at moments $t \in [1, T]$. In each cycle, first, the secret image B is predicted by AutoencoderA as the carrier image A, then noise from the carrier image A is added to the secret image B through formula (13), and finally, the secret image B is updated. The image obtained after the traversal ends is the intermediate result image A^*. Similarly, following the above method in a reverse cycle traversal $t \in [1, T]$, the intermediate result image A^* is input into AutoencoderB to predict the secret image B, and after adding its noise through formula (13), it is updated. The image obtained after the traversal ends is the extracted secret image B^*.

Algorithm 1 Prediction of SteganoDDPM

Require: Cover Image A; Secret Image B; Time List T; Noise Scheduler β

$x \leftarrow A$
$y \leftarrow B$
$n \leftarrow length(T)$
$\alpha \leftarrow 1 - \beta$
$\hat{\alpha} \leftarrow cumprod(\alpha)$
$i \leftarrow n$
while $i > 0$ **do**
  $p_x \leftarrow AutoencoderA(y)$
  if $i \neq 1$ **then**
    $n_x \leftarrow \frac{\epsilon x}{n}$
  else
    $n_x \leftarrow 0$
  end if
  $x \leftarrow \frac{1}{\sqrt{\alpha}}\left(x - \frac{1-\alpha}{\sqrt{1-\hat{\alpha}}}p_x\right) + \sqrt{\beta}n_x$
  $i \leftarrow i - 1$
end while
$i \leftarrow n$
while $i > 0$ **do**
  $p_y \leftarrow AutoencoderB(x)$
  if $i \neq 1$ **then**
    $n_y \leftarrow \frac{\epsilon x}{n}$
  else
    $n_y \leftarrow 0$
  end if
  $y \leftarrow \frac{1}{\sqrt{\alpha}}\left(x - \frac{1-\alpha}{\sqrt{1-\hat{\alpha}}}p_x\right) + \sqrt{\beta}n_y$
  $i \leftarrow i - 1$
end while
return x, y

3) Optimization Function of SteganoDDPM: As the steganographic denoising DAE model is divided into two steps—steganography and extraction in the image steganography process, with the intermediate result image A^* and the extraction result image B^* being the optimization targets, it follows from equation (12) that the minimization of the secrecy leakage function can be rewritten as:

$$l_A = \|\theta(A) - A^*\|^2 \tag{16}$$

The minimization of the embedding distortion function can be expressed as:

$$l_B = \|\theta(B) - B^*\|^2 \tag{17}$$

Therefore, the objective function of the steganographic denoising diffusion model is:

$$l_{total} = l_A + l_B \tag{18}$$

IV. Analyze and Compare

To assess the performance of the proposed image steganography diffusion model framework, this study has designed a series of experiments and conducted comparative analyses with existing technologies.

A. Experimental Design

This study's experimental design employs multiple datasets and baseline methods for an in-depth comparison of the proposed model framework's effectiveness, aiming to comprehensively evaluate the performance of the proposed model and to confirm its superiority through comparison with existing technologies.

1) Dataset Preparation: For a comprehensive evaluation of the model's performance, this study selected public datasets such as CelebA [30] and DIV2K [1] for experiments. The CelebA dataset encompasses a vast array of facial images, offering diversity and complexity that better tests the model's real-world performance. The DIV2K dataset includes 1000 images from various scenes, chosen from different categories to enhance the dataset's diversity and challenge.

2) Baseline Methods: To compare the model framework proposed in this study, existing image steganography techniques were selected as baseline methods. These methods include traditional LSB (Least Significant Bit) steganography [13], deep learning-based steganography [2], [29], and generative adversarial network-based steganography [17].

3) Evaluation Indicators: To thoroughly assess the model's performance, thisTo thoroughly assess the model's performance, this for comparative analysis [18], including peak signal-to-noise ratio (PSNR) [9], structural similarity (SSIM) [24], etc. These metrics enable the assessment of the generated stego-image quality from various perspectives, including the visual quality of the image, the degree of structural information preservation, and the similarity with the original image.

PSNR is used to measure the similarity between the original image and the stego-image, with the calculation formula given as:

$$PSNR = 10 \times \log\left(\frac{255^2}{MSE}\right) \tag{19}$$

where MSE (Mean Squared Error) denotes the mean squared error, with the calculation formula provided as:

$$MSE = \frac{1}{m \times n}\sum\sum I_A(i,j) - I_B(i,j) \tag{20}$$

where I_A and I_B represent the original image and the stego-image, respectively, while m and n denote the width and height of the image, respectively.

SSIM is used to compare the structural similarity between two images, with the calculation formula as:

$$SSIM(x,y) = [I(x,y)]^{\alpha}[c(x,y)]^{\beta}[s(x,y)]^{\gamma} \tag{21}$$

Among the metrics, $I(x,y)$, $c(x,y)$, and $s(x,y)$ represent brightness, contrast, and structural similarity, respectively, with α, β and γ being the weighting coefficients.

Additionally, this study also employed several other metrics to assess the model's security, including Attack Success Rate (ASR), stego-image quality, and Extraction Accuracy (EA).

Experiments simulated attacks using one of the most common attack methods, histogram attacks, by calculating the ratio of successful attacks to the total number of attacks to determine the attack success rate, thereby examining the model's ability to resist attacks.

When comparing attack success rates, we use the following method for quantitative estimation. Assuming that we have an original image I_0 and an image I_s with steganographic information, their size is m×n pixels, and each pixel contains three channels (ie, RGB channels), and the two images are represented as:

$$I_0 = [I_{0,1}, I_{0,2}, I_{0,3}], I_s = [I_{s,1}, I_{s,2}, I_{s,3}] \tag{22}$$

where $I_{0,i}$ and $I_{s,i}$ denote the pixel values of the original image and the steganographic image on the i-th channel, respectively. Using wavelet transform to extract the features of the image, for each channel i, a coefficient matrix C_i can be obtained, whose size is $\frac{m}{2} \times \frac{n}{2}$. Perform wavelet transform (DWT) on each channel i of the original image and the steganographic image to obtain:

$$C_{0,i}^L, C_{0,i}^H, C_{0,i}^V, C_{0,i}^D = DWT(I_{0,i}) \tag{23}$$

$$C_{s,i}^L, C_{s,i}^H, C_{s,i}^V, C_{s,i}^D = DWT(I_{s,i}) \tag{24}$$

Where $C_{0,i}^L$, $C_{s,i}^L$ is the coefficient matrix of the low frequency part, $C_{0,i}^H$, $C_{s,i}^H$ is the coefficient matrix of the horizontal high frequency part , $C_{0,i}^V$, $C_{s,i}^V$ is the coefficient matrix of the vertical high frequency part, $C_{0,i}^D$, $C_{s,i}^D$ is the diagonal high frequency part of the coefficient matrix. Here only the coefficient matrix of the low frequency part is kept:

$$C_{0,i} = C_{0,i}^L, C_{s,i} = C_{s,i}^L \tag{25}$$

Then, the coefficient matrices of the three channels are merged into a three-dimensional array by channel to obtain:

$$C_0 = [C_{0,1}, C_{0,2}, C_{0,3}] \tag{26}$$

Next, calculate the Euclidean distance d between these two coefficient arrays:

$$d = \|C_0 - C_s\| \tag{27}$$

Finally, calculate the attack success rate p, assuming that the attacker can guess the embedded information with a certain probability, then:

$$p = e^{-\frac{d^2}{2\alpha^2}} \tag{28}$$

Among them, α is a constant, representing the standard deviation of Gaussian distribution, which is used to control the scope of the attacker guessing the steganographic information. The smaller the α, the smaller the space range for the attacker to guess the steganographic information, and correspondingly, the lower the attack success rate; on the contrary, the larger the α, the higher the attack success rate. In order to facilitate the observation of the experimental results, this paper uses In the experiment, a larger value of α was selected, and the following results are the results when the value of α is 50.The results are shown in Table II.

EA is an indicator that measures the precision of a steganography algorithm in extracting information from a stego-image. Assuming that the information extracted from the stego-image by the steganography algorithm is S', and the original information is S, then the image extraction accuracy can be represented as equation (18).

$$EA = \frac{\|S' \cap S\|}{\|S\|} \tag{29}$$

where $\cap$ denotes the intersection of two sets, and $\|S'\|$ and $\|S\|$ represent the sizes of the extracted information and the original information, respectively.

B. Compare Results

Through experimental design and comparative analysis, this study has demonstrated that the proposed image steganography diffusion model framework achieves superior performance in multiple aspects. These experimental results fully showcase the applicability and effectiveness of the model framework in image steganography tasks.

1) Visual Effects Evaluation: Figure 5 demonstrates a comparison of images processed using the proposed steganographic denoising diffusion model with the original carrier image and the original secret image. It can be clearly observed from the figure that the steganography-processed image is very similar to the original carrier image visually, and compared to the original secret image, its detailed features have been effectively preserved.

This result indicates the superior performance of the proposed model in maintaining image quality, which is due to the ability of the DDPM model in generating images. The combination of DDPM and DAE can better restore the carrier image, thereby maintaining its visual quality.

2) Comparison of quality indicators: Table 1 presents a comparison of the results of the model framework proposed in this study with other existing technologies on different evaluation metrics over the CelebA dataset. According to the data in the table, it is evident that the model framework proposed in this study shows a significant improvement in the PSNR metric compared to other technologies, with an average increase of 56.56dB, while remaining roughly on par with other existing technologies in terms of the SSIM metric.

Methods	PSNR(Extract)	SSIM(Extract)
Ours	91.23	0.9433
Method [17]	34.67	0.9919
Method [2]	33.63	0.9429
Method [29]	25.97	0.9160
Method [20]	45.84	0.9880
LSB [13]	27.59	0.7962

TABLE I. IMAGE QUALITY ASSESSMENT RESULTS.

These findings suggest that the steganographic denoising diffusion model framework proposed in this study performs exceptionally well in generating the quality of stego-images. Compared to existing technologies, this model framework has the capability to produce higher-quality stego-images, offering better visual quality and structural information preservation.

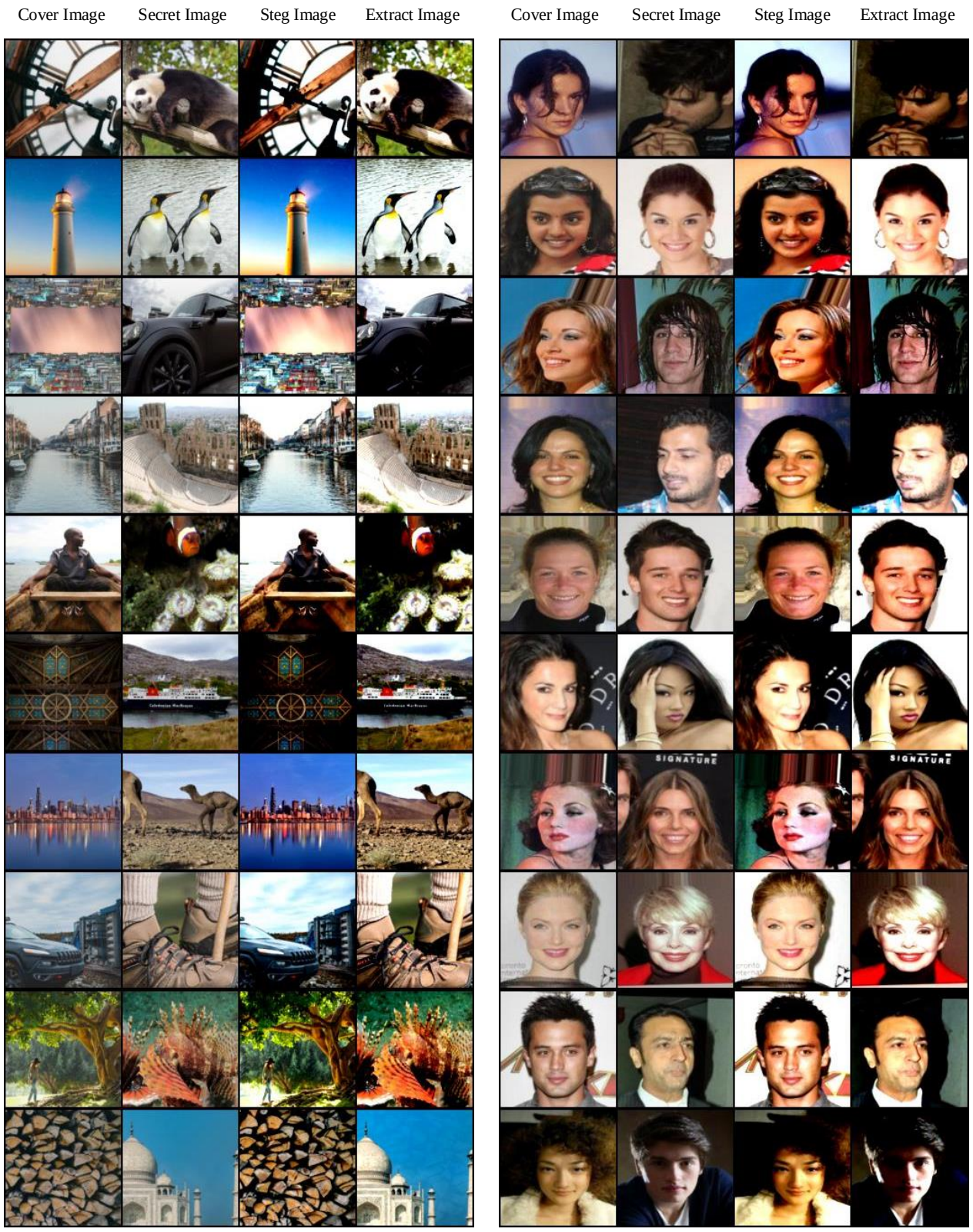

Fig. 5. Visual effects implemented by SteganoDDPM. The figure shows the effect of training on the DIV2K (left) and CelebA (right) datasets.

This outcome further substantiates the superior performance of the proposed model in generating high-quality stego-images, that is, the SteganoDAE model proposed in this study successfully achieves its goal by learning to minimize alterations to the carrier image during the steganography process. The optimization objectives of this research include minimizing secrecy leakage and embedding distortion. These two goals correspond to maintaining the closeness of the carrier image

to the original image and the closeness of the secret image to the extracted stego-image, respectively. By simultaneously optimizing these two objectives, it can be ensured that the steganography process neither significantly alters the carrier image nor compromises the accurate extraction of hidden information.

3) Steganography Robustness and Security Comparison: The comparative results are presented in Table 2, which demonstrate that the proposed method maintains high security under histogram attacks, with an attack success rate of merely $2.4 \times 10^{-38}\%$, significantly lower compared to the benchmark models; the extraction accuracy reached 62.76%, which is 16.81% higher than existing models.

Methods	ASR(%)	PSNR(Steg)	EA(%)
Ours	2.4×10^{-38}	93.14	62.76
Method [17]	2.8×10^{-17}	35.05	45.93
Method [2]	7.4×10^{-17}	34.63	44.63
Method [29]	1.3×10^{-16}	26.72	42.25
Method [20]	3.8×10^{-18}	62.34	50.00
LSB [13]	46.38	51.13	22.61

TABLE II. IMAGE SECURITY ASSESSMENT RESULTS. PART OF THE DATA COMES FROM THE CITE [26].

The method proposed in this study enhances the model's ability to recover hidden information, ensuring complete restoration of information even under interference. The low attack success rate and high extraction accuracy under histogram attacks further substantiate the model's robustness and security.

4) Border Clarity Comparison: In addressing the issue of unclear image boundaries in existing technologies, the model framework proposed by this study, through an improved diffusion model principle, is better equipped to preserve boundary information of images. As shown in Figure 6, experimental results indicate that the model framework significantly enhances boundary clarity compared to existing technologies.

Fig. 6. Comparison of border details. As shown in the figure, the boundary clarity of the steganographic and extracted images is significantly improved.

This is attributed to the model's composition of two serial autoencoders: the first autoencoder is responsible for restoring the carrier image, while the second exploits the residual secret image information to recover the secret image. This design enables the model to maintain high precision during both the steganography and extraction processes.

V. CONCLUSION

This study has conducted an in-depth exploration of the traditional notions of diffusion models in image steganography tasks and has successfully proposed an innovative image steganography diffusion model framework. Given that the core principle of diffusion models is to progressively eliminate noise, many researchers believed they were unsuitable for image steganography tasks. However, by thoroughly analyzing the principles of diffusion models and integrating the characteristics of image steganography tasks, this study has successfully challenged this conventional perspective. The proposed image steganography diffusion model framework has been validated not only theoretically but also experimentally for its applicability in image steganography tasks. To a certain extent, the framework resolves issues present in existing technologies, such as the low quality of generated stego-images and unclear image boundaries, significantly enhancing the quality of stego-images. Experimental evidence demonstrates that the model framework can produce high-quality stego-images and exhibits superior performance across multiple indicators compared to existing technologies. This research paves a new technical path for image steganography tasks and hopes that more researchers will further explore this direction in their future work.

REFERENCES

[1] E. Agustsson and R. Timofte, "Ntire 2017 challenge on single image super-resolution: Dataset and study," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 126–135.

[2] S. Baluja, "Hiding images in plain sight: Deep steganography," *Advances in neural information processing systems*, vol. 30, 2017.

[3] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[4] X. Feng, W. Pei, F. Li, F. Chen, D. Zhang, and G. Lu, "Generative memory-guided semantic reasoning model for image inpainting," *IEEE Transactions on Circuits and Systems for Video Technology*, p. 7432–7447, Nov 2022.

[5] G. Han, D.-J. Lee, J. Hur, J. Choi, and J. Kim, "Deep cross-modal steganography using neural representations," in *2023 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2023, pp. 1205–1209.

[6] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020.

[7] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019.

[8] D. P. Kingma and M. Welling, "An introduction to variational autoencoders," *Foundations and Trends® in Machine Learning*, p. 307–392, Jan 2019.

[9] J. Korhonen and J. You, "Peak signal-to-noise ratio revisited: Is simple beautiful?" in *2012 Fourth International Workshop on Quality of Multimedia Experience*. IEEE, 2012, pp. 37–38.

[10] K. Lamshöft, T. Neubert, C. Krätzer, C. Vielhauer, and J. Dittmann, "Information hiding in cyber physical systems: Challenges for embedding, retrieval and detection using sensor data of the swat dataset," in *Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security*, Jun 2021.

[11] G. Li, S. Li, M. Li, X. Zhang, and Z. Qian, "Steganography of steganographic networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 4, 2023, pp. 5178–5186.

[12] J. Liu, Y. Ke, Z. Zhang, Y. Lei, J. Li, M. Zhang, and X. Yang, "Recent advances of image steganography with generative adversarial networks," *IEEE Access*, p. 60575–60597, Jan 2020.

[13] J. Mielikainen, "Lsb matching revisited," *IEEE signal processing letters*, vol. 13, no. 5, pp. 285–287, 2006.

[14] A. Q. Nichol and P. Dhariwal, "Improved denoising diffusion probabilistic models," in *International conference on machine learning*. PMLR, 2021, pp. 8162–8171.

[15] W. Peebles and S. Xie, "Scalable diffusion models with transformers," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4195–4205.

[16] Y. Peng, D. Hu, Y. Wang, K. Chen, G. Pei, and W. Zhang, "Stegaddpm: Generative image steganography based on denoising diffusion probabilistic model," in *Proceedings of the 31st ACM International Conference on Multimedia*, 2023, pp. 7143–7151.

[17] J. Y. Qu Mengnan and W. Jiang, "Research on image steganography and extraction scheme based on implicit symmetric generative adversarial network," *Journal of Information Securyity Research*, vol. 9, no. 6, pp. 566–572, 2023.

[18] D. R. I. M. Setiadi, "Psnr vs ssim: imperceptibility quality assessment for image steganography," *Multimedia Tools and Applications*, p. 8423–8444, Mar 2021.

[19] T. R. Shaham, T. Dekel, and T. Michaeli, "Singan: Learning a generative model from a single natural image," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct 2019.

[20] V. Sitzmann, J. Martel, A. Bergman, D. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," *Advances in neural information processing systems*, vol. 33, pp. 7462–7473, 2020.

[21] Y. Song and S. Ermon, "Generative modeling by estimating gradients of the data distribution," *Advances in neural information processing systems*, vol. 32, 2019.

[22] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole, "Score-based generative modeling through stochastic differential equations," *arXiv preprint arXiv:2011.13456*, 2020.

[23] P. Vincent, H. Larochelle, Y. Bengio, and P.-A. Manzagol, "Extracting and composing robust features with denoising autoencoders," in *Proceedings of the 25th international conference on Machine learning*, 2008, pp. 1096–1103.

[24] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.

[25] P. Wei, Q. Zhou, Z. Wang, Z. Qian, X. Zhang, and S. Li, "Generative steganography diffusion," *arXiv preprint arXiv:2305.03472*, 2023.

[26] S. Yang, S. Song, C. D. Yoo, and J. Kim, "Flexible cross-modal steganography via implicit representations," *arXiv preprint arXiv:2312.05496*, 2023.

[27] X. Yu, Y. Qu, and M. Hong, *Underwater-GAN: Underwater Image Restoration via Conditional Generative Adversarial Network*, Jan 2019, p. 66–75.

[28] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *2017 IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[29] R. Zhang, F. Zhu, J. Liu, and G. Liu, "Depth-wise separable convolutions and multi-level pooling for an efficient spatial cnn-based steganalysis," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1138–1150, 2019.

[30] Y. Zhang, Z. Yin, Y. Li, G. Yin, J. Yan, J. Shao, and Z. Liu, "Celeba-spoof: Large-scale face anti-spoofing dataset with rich annotations," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 2020, pp. 70–85.

[31] S. Zhao, J. Song, and S. Ermon, "Infovae: Information maximizing variational autoencoders," *Cornell University - arXiv,Cornell University - arXiv*, Jun 2017.