Ren and Wu *Cybersecurity (2024)7:73*
https://doi.org/10.1186/s42400-024-00299-5

Cybersecurity

RESEARCH

Open Access

A Robust joint coverless image steganography scheme based on two independent modules

Chang Ren[1,2] and Bin Wu[1,2*]

Abstract

With the development of deep learning technology, great progress has been made in the field of coverless steganography based on deep learning technology, including some selection-based steganography methods that use deep learning technology and all generation-based steganography methods, however both of which have their limitations. The former is difficult to meet actual communication requirements in terms of communication capacity and completeness due to the limit of the algorithm. Due to the irreversibility of the process of generating secret images from message codeword, the recovery accuracy of the latter is very poor. To this end, this paper designs a robust joint coverless image steganography scheme called Joint Coverless Image Steganography (JoCS). Firstly, this paper proposes the Semantic Factorization Fitting module (SeFF) and the Transform Domain Steganography module (TrDS). The former adds the secret message to the input vector of the low resolution layer in the StyleGAN generator network, which establishs a mapping rule between message codeword and the coarse feature of the generated image, and then the extractor is used to fit the above mapping rule, which has excellent robustness and completeness; the latter encodes the main content area of the image based on the encoder in VQGAN, and then adds secret message to the latent vector of the encoded image, which achieves the steganography in the latent domain of the image. Secondly, we demonstrate the independence between two modules and the advantages of connecting two modules. By using the image generated in the SeFF module as the cover image in the TrDS module, secondary steganography of a single image is achieved, based on which we design the JoCS scheme. The results show that our scheme breaks through the communication capacity limit in the selection-based coverless methods while guaranteeing 100% completeness, excellent image quality and outstanding robustness against various image attacks. Moreover, our scheme exhibits strong security against detection by multiple steganalysis tools and excellent practicality in practical communication. Finally, this paper also discusses the following three points as further elaboration of the scheme: (1) the advantages of the mapping rule in the SeFF module (2) the verification of the independence between the two modules (3) the flexibility of the joint steganography scheme.

Keywords Selection-based and generation-based coverless steganography, Generative adversarial network, Steganalysis, Joint steganography

Introduction

Image steganography refers to a technique of transmitting secret message in a public channel by embedding the secret message into the redundant information in the cover image. It can be divided into cover-modified steganography and coverless steganography according to different embedding principles.

*Correspondence:
Bin Wu
wubin@iie.ac.cn
[1] Key Laboratory of Cyberspace Security Defense, Institute of Information Engineering, CAS, Beijing, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China

Cover-modified steganography methods usually embed secret message into the cover image by modifying pixel values or coefficients in the transform domain (Mielikainen 2006; Liao et al. 2019; Su et al. 2020; Tao et al. 2018). However, such embedding methods often cause an inevitable distortion to the cover image, which is difficult to resist detection by steganalysis tools (Pevnỳ et al. 2009). Thus, coverless steganography method (Zhou et al. 2015) has been proposed.

The coverless steganography methods can be divided into two categories: selection-based coverless steganography and geneartion-based coverless steganography. The former achieves the transmission of secret message by establishing a mapping rule between secret message and image features, which can either be manually designed or extracted through deep learning networks. And then the sender can select the corresponding image to represent the secret message based on the above mapping rule. The latter directly generates image from the encoded secret message based on generator model. Such generator models are usually Generative Adversarial Networks (Goodfellow et al. 2020) or Diffusion models (Ho et al. 2020). Since coverless steganography methods do not modify the cover image, they can effectively resist detection by steganalysis tools and ensure the security of the communication process.

However, there are some problems with the coverless steganography methods. Due to the limit of algorithm, the selection-based coverless steganography methods generally cannot support the high-capacity steganography in a single image, and the completeness of mapping rule are also unsatisfactory. As for generation-based coverless steganography methods, receiver generally is not able to correctly recover all secret message due to the irreversibility of the image generation process. Also, the robustness against image attacks is even worse.

Thus, enhancing the communication capacity of selection-based coverless steganography methods and improving the recovery accuracy of generation-based coverless steganography methods are urgent problems in current research. Therefore, this paper proposes two modules to solve the above problems: the SeFF module(selection-based) and the TrDS module(generation-based). The steganography schemes in both modules are independent of each other and have very good robustness. We use the images generated in the SeFF module as the cover image of the TrDS module, which connects the two modules, on the basis of which we finally propose a robust joint coverless image steganography scheme, JoCS, where 'joint' refers to the integration of two sub-modules presented in this paper. The main contributions of this paper are as follows.

(1) We found that the mapping rules between the input vector of the low resolution layer and the coarse features of the generated image in the StyleGAN network are very robust, which is suitable for selection-based coverless steganography, so we proposed the SeFF module. By adding message codeword to the input vector which will be inputted to the low resolution layer of the StyleGAN (Karras et al. 2019) generation network, we establish a complete mapping rule between secret message and the coarse features of the generated image, which will be learned and fitted by the extractor network, and then each image can represent a message codeword. We also analyze the feature points distribution between the original image and the attacked image, which shows that the above mapping rule has good robustness.
(2) We found that VQGAN (Esser et al. 2021) can encode and reconstruct images without disrupting the coarse features of the image. Based on above characteristics, we proposed the TrDS module to achieve steganography in the latent domain of the image. We add the secret message to the latent vector of the main content area of the image, which enables the TrDS module to possess excellent resistance to image geometric attacks. In addition, the attack is added to the training process of the extractor model to further improve the robustness of the steganography scheme.
(3) We have proven the independence between the above two modules through theoretical analysis and experimental verification. And then we design a joint coverless steganography scheme to combine the two modules. By conducting extensive experiments on different datasets, it is proved that the joint coverless steganography scheme proposed in this paper has advanced completeness, robustness, security and communication capacity.

Related work

Selection-based steganography method

In selection-based coverless steganography methods, secret messages are represented by images selected by a specific mapping rule. This scheme was first proposed by Zhou et al. (2015), uses a hash algorithm to calculate the hash value of the image block, then generates a sequence to represent the secret message of the image. Subsequently, Zheng et al. (2017) proposed using the orientation information of the SIFT feature point to calculate the hash value of the image and establish the mapping rule between the hash value of the image and the secret message, so that the sender can communicate with the receiver according to the mapping rule. Yuan

et al. (2017) proposed a coverless steganography scheme based on SIFT and Bag, firstly clustering the image with Bag of feature, then extracting the SIFT of the image to obtain the corresponding hash sequence. Thus, a mapping rule between the hash sequence and the secret message is established. Luo et al. (2020) extracted objects in images based on Faster RCNN and established a mapping rule between image objects and secret message. However, these schemes perform poorly in terms of completeness of mapping rules, as there are cases where the secret message cannot match the corresponding image.

Wang and Wu (2020) proposed a method called CIHDN to learn the mapping rule between secret message and image with over-fitting neural network in deep learning, which achieves 100% completeness of the mapping rule. It means that all secret messages codeword can be matched to at least one image in CIHDN's image database, avoiding a risk of communication failure. Wu and Xue (2024) established a mapping rule between face images and secret messages, using facial image identity information as a pivot to connect the images with the secret message, further enhancing the robustness of the scheme on the basis of ensuring that the scheme has 100% completeness.

Generation-based steganography method

With the rapid development of the Generative Adversarial Network (GAN), more and more researchers began to apply GAN to steganography. Hu et al. (2018) proposed to generate secret images by mapping secret messages to noise vectors as the input part of DCGAN. Then we used an extractor based on a convolutional neural network to recover secret messages. You et al. (2022) proposed CIS NET, which directly synthesizes 32*32 images based on secret messages. By adding noise modules to the extractor network, the robustness of the scheme is improved. Liu et al. (2022) proposed the Image Dis-Entanglement Autoencoder for Steganography (IDEAS), a scheme that decomposes an image into two representations, structural and textural, and trains the StyleGAN generator and structural vector extractor to hide and extract secret messages. Peng et al. (2022) proposed an image steganography framework based on a generative adversarial network and a gradient descent approximation.

Yu et al. (2023) introduced diffusion models into the image steganography field for the first time, which proposed a new steganography framework applying the robustness of diffusion models to noisy data and the ability to transform between two images without training. Peng et al. (2023) proposed StegaDDPM, a method that utilizes the probability distribution between the intermediate state and the generated image in the reverse process of the diffusion model. In this approach, secret information is hidden into the generated image through message sampling, ensuring that the probability distribution remains consistent with normal generation. Wei et al. (2023) designed an invertible diffusion model called StegoDiffusion, which utilizes a non-Markov chain and fast sampling techniques to achieve efficient stego image generation. Kim et al. (2023) proposed a generation-based steganography method based on the diffusion model, in which Diffusion-Stego projects the secret message into the latent noise of the diffusion model and generates the stego image through an iterative denoising process.

Although most of the generation-based coverless steganography schemes have large communication capacity, their recovery accuracy and robustness against image attacks are very poor, which makes it impossible to ensure consistent information between the sender and receiver. In addition, the stability of the generator is also critical, as whether high-quality images can be generated for transmission determines the security of the communication process.

Progressive generative model StyleGAN

StyleGAN is a style-based GAN generator architecture proposed by Karras et al. (2019). StyleGAN adopts a progressive generation method. By separately modifying the input vectors corresponding to each resolution layer of the generative network, the visual features expressed in that layer can be controlled. Taking the StyleGAN model trained on the FFHQ dataset as an example, as shown in Fig. 1, each resolution layer has two input vectors (e.g., S1 and S2 for the 4*4 layer). The inputs at the 4*4 to 8*8 resolution layers influence image features such as pose, general hairstyle, and facial shape. The inputs at the 16*16 to 32*32 resolution layers influence more fine-grained facial features, such as hair details and whether the eyes are open or closed. The inputs at the 64*64 to 1024*1024 resolution layers influence color features (eyes, hair, and skin) as well as microscopic details. This method allows the model to first learn basic features on simpler images and then gradually learn details and complex features. It helps stabilize the training process, improve the quality of generated images, and provides fine control over the generated content.

Method

This section first introduces the design of the SeFF module and the TrDS module, and then elaborates the communication process of the joint coverless image steganography scheme (JoCS), which is based on two modules.

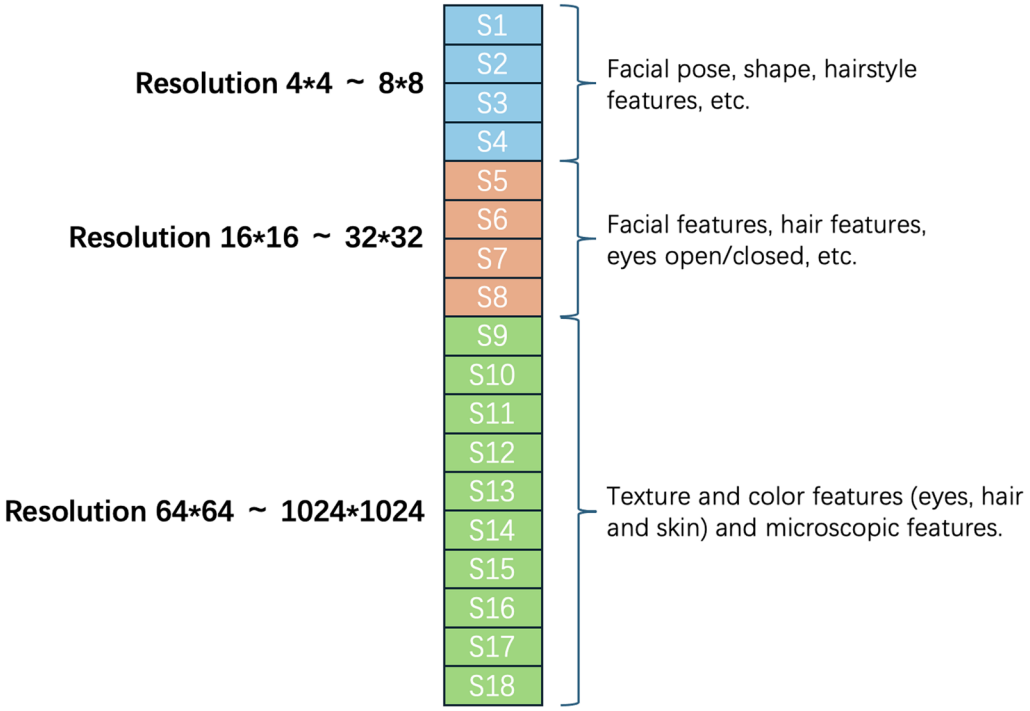

Fig. 1 The features controlled by input vectors at different resolution layers

Semantic factorization fitting module

Most of the existing selection-based coverless methods are as below, first design a mapping rule between the secret message and the image, and then search for an eligible database in an image dataset as a mapping cover set for the secret message. However, there are some problems in the above design idea:

(1) The message codes mapped by the image database do not always contain all the secret message to be transmitted, and there may be cases where the secret message does not match any of the images in the image database.
(2) Natural image datasets are limited in size, and as the communication capacity of the scheme increases, the required database become larger and larger. It becomes increasingly difficult to find a eligible image database.
(3) The mapping rules of most schemes have poor robustness. When the secret images attacked by various image algorithms, the receiver's recovery accuracy for the secret message will be very poor.

In order to solve the above problems, we propose a module in this paper inspired by the principle of the StyleGAN generator model (Fig. 2). In this module, since the mapping rule is established between the vectors and corresponding features controlled by them (the input vectors of the low resolution layer of the StyleGAN generation network correspond to the coarse features of the generated image), we name the module after the Semantic Factorization Fitting Module (SeFF). The framework of the SeFF module is shown in Fig. 3, where w is the input vector of the StyleGAN generator network, with dimensions of (1, 14, 512), w_{1-14} are sub-vectors of w, with dimensions of (1, 512), x_t represents a partial segment of vector w_1, with dimensions of (1, 32) and mean of 0, m_t is the bit value of the secret message, w' is the secret vector after added the secret message, with dimensions of (1, 14, 512).

In the SeFF module, the secret message M is first added to the input vectors of the low resolution layer of the StyleGAN generation network (the input vector of the generation network is w, and we only add the secret message to w_1), then the mapping rule between the input vectors and the generated images is established based on the StyleGAN generator, and then the input vectors are extracted from the secret image I_S based on the extractor E_x designed by us, finally the secret message M' is recovered.

We add the secret message to the vector w_1 which will be input to the 4*4 resolution layer of the generation network, the vector w_1 follows the standard normal distribution with vector dimension (1, 512), and

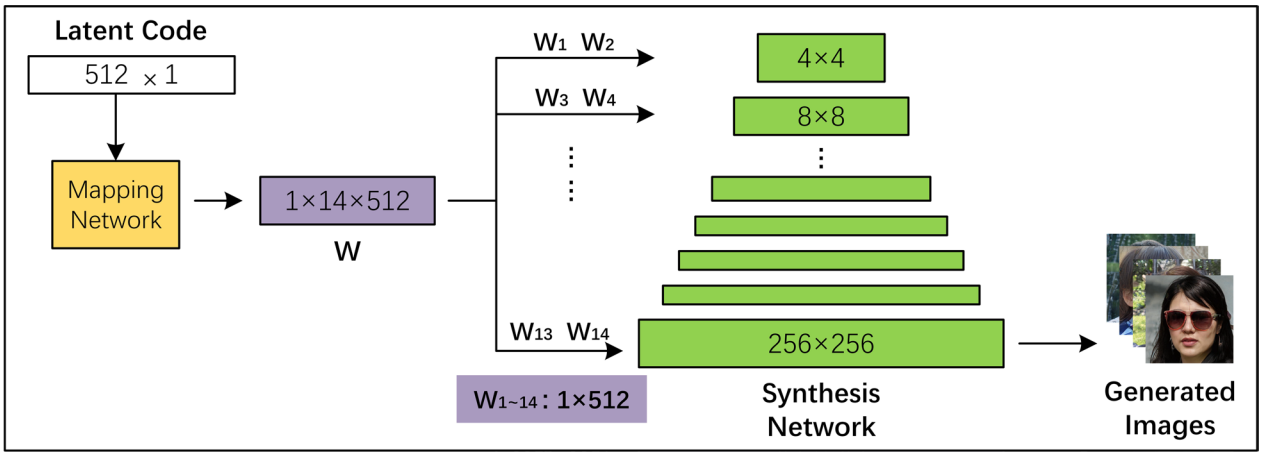

Fig. 2 Random latent code transforms into vector w through the mapping network, whose dimension depends on the resolution of the generated image. When the resolution of the generated image is 256*256, the dimension of w is (1,14,512). Synthsis Network generates images progressively layer by layer, where each resolution layer contains two convolution blocks. The input vector w_x to each convolution block is from w, where the dimension of w_x is (1,512) and $0 < x \leq 14$. (For example, the input of the first convolution block of the 4*4 resolution layer is w_1)

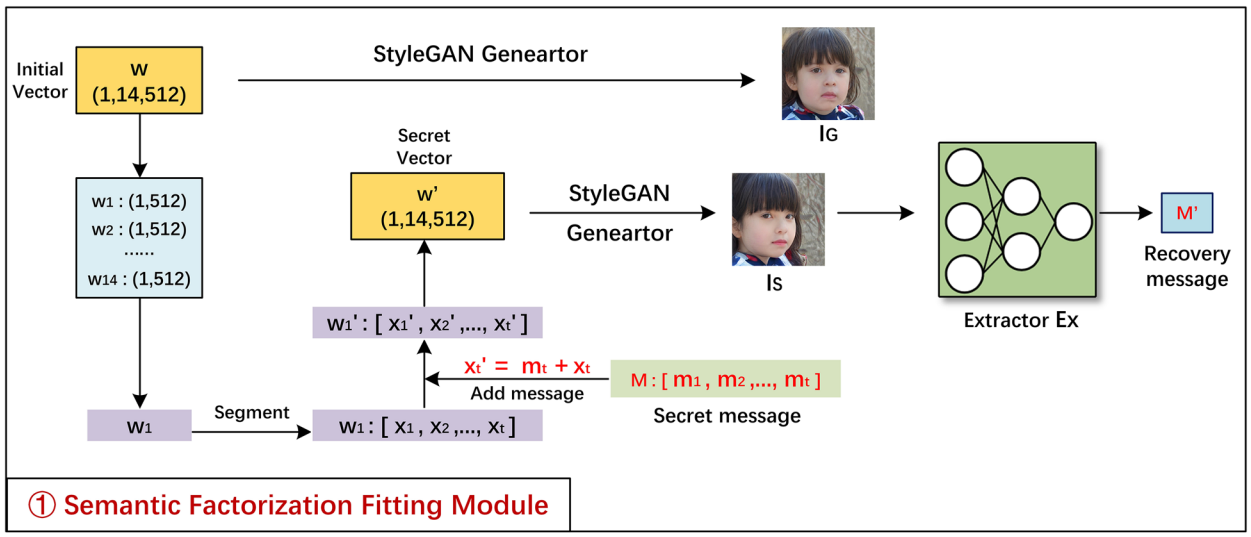

Fig. 3 Semantic factorization fitting module

we divide it into 16 segments $w_1 = [x_1, x_2, x_3, ..., x_{16}]$, the dimension of each vector segment x_t is (1, 32), and the secret message M consists of a 16-bit binary string $M = [m_1, m_2, m_3, ..., m_{16}]$, where $m_t = 0$ *or* 1 $(0 < t \leq 16)$.

After that, we add a secret message in each vector segment x_t of w_1. As shown in formula (1), the mean value of x_t is always 0 and depending on the value of m_t added, the mean value of each segment x'_t is close to 0 or 1. Therefore, the receiver can determine whether m'_t is 1 or 0 based on whether the mean of x'_t extracted from the stego image is greater than or less than 0.5.

$$\begin{cases} x'_t = x_t + m_t & (0 < t \leq 16) \\ w'_1 = [x'_1, x'_2, x'_3, ..., x'_{16}] \end{cases} \tag{1}$$

The vector with the secret message added is w'_1, and the secret image I_S is then synthesized by w', thus establishing a mapping rule between the secret message and the generated image. In addition, since the secret message hiding process only modifies the input vector of the low resolution layer in the generation network, the generated image quality will not be affected. As shown in Fig. 3, the image I_G generated by the initial vector w and the image

I_S generated by the secret vector w' only differ in the pose of the face, which will not alert third parties.

$$\begin{cases} (mean(x'_t) < 0.5,\ m'_t = 0(0 < t \leq 16) \\ (mean(x'_t) \geq 0.5,\ m'_t = 1(0 < t \leq 16) \end{cases} \tag{2}$$

We then design the extractor E_x to extract the vector w'_1 from the generated images I_S. Since the vector w_1 follows the standard normal distribution, as shown in formula (2), the mean value of each vector segment x'_t represents one-bit of the secret message, and the secret message recovered from I_S is $M' = [m'_1, m'_2, m'_3, ..., m'_{16}]$. In order to improve the recovery accuracy of secret message, we directly train the extractor network E_x to the overfitting (example: when transmitting a secret message of 8 bit, we add 00000000−11111111 in sequence to 256 initial vectors and generate the corresponding images. Each image can represent a message code and then fit its mapping rule with E_x), so as to guarantee that the extractor can fit the mapping rule between the vector w_1 and the generated image. Thus, secret messages are accurately recovered from the vector w'_1 that be extracted from I_S.

When using the SeFF module in this paper, the sender can directly generate the images containing the mapping rule without having to search for an eligible image database to assign the code for it, effectively solving the image dataset shortage. In addition, the above mapping rule has very good completeness, which can ensure that each message codeword can be mapped to an image. When extracting secret messages from generated images, we adopt the mean value of the vector segment x'_t as the extraction benchmark of the secret message, even if some values in the recovered vector segment x'_t are deviated, it can be mitigated by calculating the mean of the entire vector segment, which improves the robustness of the SeFF module.

Transform domain steganography module

Since the communication efficiency of the SeFF module is limited by the size of the dataset (transmitting a secret message of n bits requires the extractor to fit the mapping rule between 2^n pairs of images and codewords, as n increases, the efficiency of the scheme becomes lower and lower), whether we can find a method to conduct secondary steganography in the secret image without affecting the mapping rule of the SeFF module is the key to solving the communication efficiency problem. Secondary steganography refers to the process where, after generating a stego image with mapping rule based on the SeFF module, an additional steganography operation is performed on the same image to hide another secret message. This allows a single image to independently hide two separate secret messages. In this paper, since the modification of pixel-level feature will disrupt the mapping rule in SeFF, the secondary steganography of the image should be performed in the latent space of the image, and it should satisfy the following conditions:

(1) The process of secondary steganography should not affect the mapping rule established in the SeFF module.
(2) The secondary steganography scheme should have excellent robustness while improving the communication capacity.

Inspired by Bui et al. (2023), this paper proposes the Transform Domain Steganography Module (TrDS), the TrDS module achieves a reversible transformation of the image from the spatial domain to the latent domain with the Encoder and Decoder models in the VQGAN(Esser et al. 2021) image compression model.

As shown in Fig. 4, firstly, object detection is performed on the cover image with Fast R-CNN (Girshick 2015), and the main detected object is cropped to obtain the main content area and the edge part of the cover image. Next, the VQGAN-Encoder encodes the main content area into a vector Z_0, while the CNN-Encoder encodes the secret message S into a vector δ of the same size as Z_0. The VQGAN-Decoder then recovers $Z_0 + \delta$ to the reconstruction image, which is further spliced with the remaining edge part of the cover image to obtain the stego image. Finally, the secret message S' is recovered by a decoder designed based on Resnet50.

In the TrDS module, the secret message is added to the latent vector of the image after encoded by the CNN-Encoder, then the image is reconstructed by the VQGAN-Decoder. In order to make the reconstructed image as similar as possible to the original image, the loss of image reconstruction and the loss of perception are added to the training process, so that the encoded secret message will not affect the reconstruction process of the image. As a result, there is no significant difference in the coarse features between the images before and after reconstruction, which means that the steganography in the TrDS module does not affect the mapping rule in the SeFF module (this will be further verified later), which satisfies the first condition of secondary steganography.

To ensure the robustness of the steganography in the TrDS module, we add an attack module to the training process of the Resnet50 decoder. When the recovery accuracy of the decoder reaches more than 90%, we perform image attacks, such as adding random noise, filtering, cropping, and JPEG compression, on training data for data augmentation, which makes the Resnet50 decoder more robust. Moreover, since the secret message is hidden in the main content area of the stego

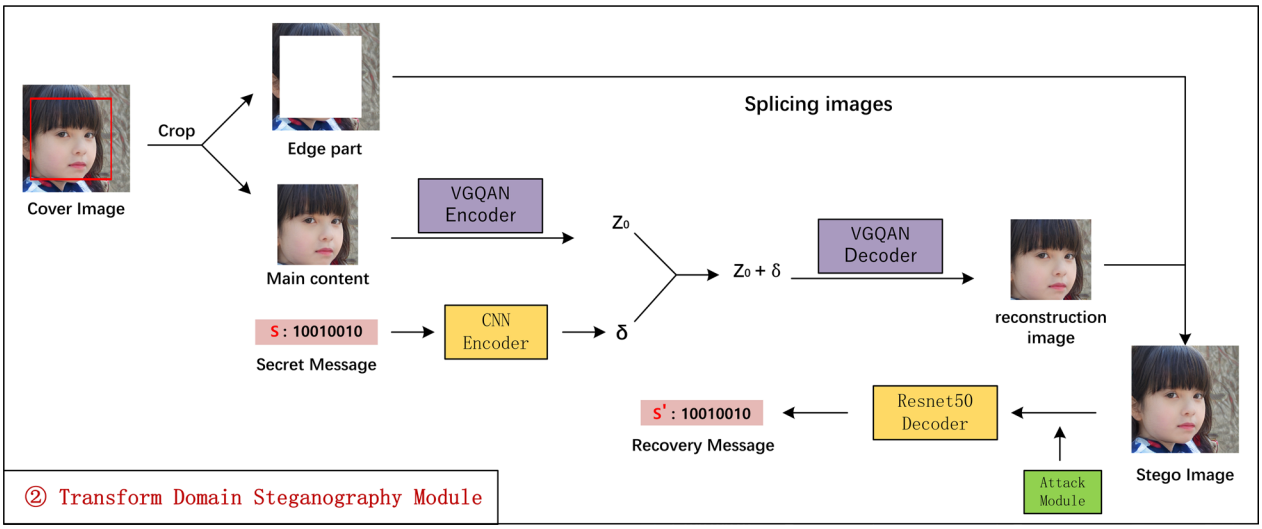

Fig. 4 Transform domain steganography module

image, and this area tends not to be drastically cropped in the transmission process, the Resnet50 decoder's resistance to geometric attacks is also excellent. In conclusion, the TrDS module also satisfies the second condition of secondary steganography.

The communication process of joint steganography scheme

Although we can use the two modules mentioned above to independently hide two different messages, this approach is highly inefficient for practical communication. For example, in the case of the SeFF module, the communication capacity of a single image is only 16 bits, which means that transmitting a secret message often requires a large number of images. To improve communication efficiency, we aim to design a steganography scheme that can use the two aforementioned modules to transmit a continuous secret message within a single image.

Since the steganography in the SeFF module and the steganography in the TrDS module are independent and the two modules both have excellent robustness, we combine the above two modules by using the image generated in the SeFF module as the cover image of the TrDS module. Then, we design a joint coverless image steganography scheme based on above two modules, which we refer to as JoCS in the subsequent sections. Here, the term 'joint' indicates that the JoCS scheme is constructed by connecting two independent modules. Figure 5 shows the communication process between the sender and the receiver in the joint steganography scheme:

(a) The sender, Alice, shares the trained extractor E_x, the Resnet50 decoder and other necessary information with the receiver, Bob through a secure channel. So far the negotiation process between the two parties is finished. Thereafter, Bob can recover the secret message in the received image with the E_x and Resnet50 decoder without exchanging extra information with Alice.

(b) As shown in Fig. 5, in sending state, Alice needs to transform the complete message into a binary string M_s and divide it into several message fragments M_{sk} of length $n+N$. Each segment of M_{sk} is composed of the corresponding a_k and b_k, where the size of a_k is n, referring to the communication capacity of the SeFF module, and the size of b_k is N, referring to the communication capacity of the TrDS module.

(c) First, Alice searches for the vector with codeword added n-bit code a_k in the vector base of the SeFF module, and inputs it into StyleGAN generator to generate image I_x which represents the codeword a_k. After that Alice encodes the image I_x to obtain the latent vector of I_x with VQGAN-Encoder in TrDS module. Next Alice encodes the remaining N -bit code b_k and adds it to the latent vector of I_x. Finally, the VQGAN-Decoder reconstructs the secret image I'_x from the latent vector embedded with the second secret message b_k.

(d) By repeating the operation in step (c), the sender Alice converts all message segments $[M_{s1}, M_{s2}, ..., M_{sk},]$ into the corresponding stego images. Then, Alice sends the image sequence through the social network, during which the

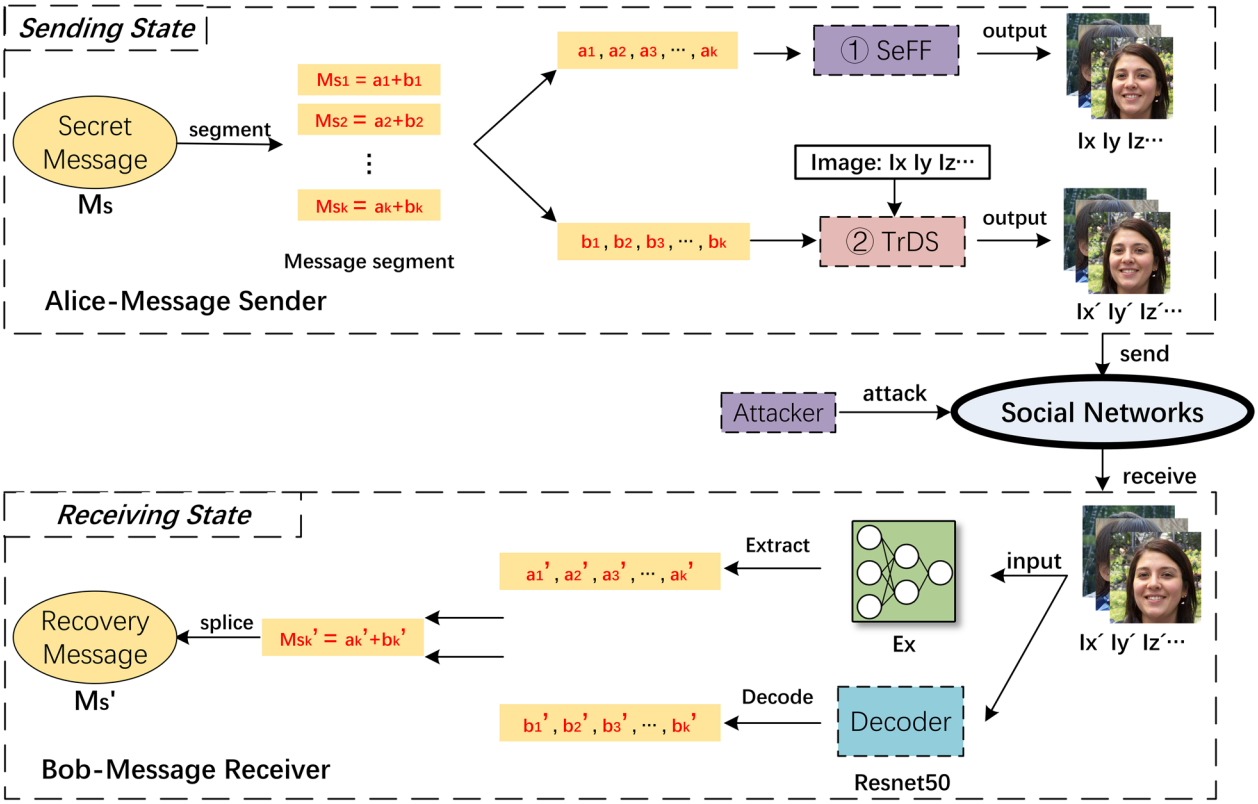

Fig. 5 The communication process of the joint steganography scheme

images may be detected by steganalysis tools or attacked by third-party attackers in the channel.

(e) In receiving state, Bob receives the image sequence, uses the extractor E_x to recover the message code a'_k, uses the Resnet50 decoder to recover the message code b'_k, then splices them to obtain the message segment M'_{sk}. After performing the above operations on all the received images, Bob obtained the complete secret message segment $[M'_{s1}, M'_{s2}, ..., M'_{sk}]$. Finally, Bob correctly splices all message segments M'_{sk} in order, then obtains the complete message M'_s, so far the communication between the two parties is finished.

Experiment

Experimental setup

The experiment of the SeFF module is conducted in Ubuntu 20.04 LTS, Pytorch 1.9.0, 1080Ti (12GB). After adding the secret message (16-bit binary representation of 0-65535) into the initial vector, we generate 65,536 secret images of 256*256 with the StyleGAN generator models respectively trained on six datasets, CELEBA, CHURCH, FFHQ, BEDROOM, CAT, HORSE and then train the extractor E_x to the over-fitting state (error bits of the secret messages recovered from the training set is 0).

The experiment environment for the TrDS module is the same as that for the SeFF module. We used 80,000 images from the MIRFlickR dataset as the training set and 800 images as the test set. The main content area of all the above images was selected by Fast R-CNN and then cropped to 192*192 before training. After adding the secret message (32-bit binary codeword) and reconstructing, we spliced them to 256*256 with the remaining edge part and used it as the input of the training process. The above trained model is used on the six datasets generated by the SeFF module for the second steganography.

In this paper, six coverless steganography methods are selected as contrast experiments, including SIFT-HASH (Zheng et al. 2017), PIXEL (Zhou et al. 2015), CIHDN (Wang and Wu 2020), IDEAS (Liu et al. 2022), MDI (Xue and Wang 2021) and RoSteAls (Bui et al. 2023). The evaluation metrics for the experiments are completeness, communication capacity, security, and robustness.

Quality of generated images

In order to better apply the steganography scheme proposed in this paper to practical communication processes, the quality of the images generated in the scheme

is evaluated. This paper employs Fréchet Inception Distance(Heusel et al. 2017) (FID) and Image Quality Assessment(Wu et al. 2023) (IQA) to indicate the stability of the entire generated data set and the quality of individual images, respectively.

The FID metric is commonly used to evaluate the quality of images generated by Generative Adversarial Networks (GANs). The FID metric is calculated by comparing the feature distributions of generated images with those of real images, which combines both the realism and diversity of generated images. Generally speaking, a lower FID score indicates that the generated images are closer to the real images and of better quality. However, the FID score cannot guarantee that each generated image will be free from distortion. The FID score relies on features extracted by a pre-trained Inception network. If the generated images are similar to the real images in these features, the FID score will be lower, but this does not necessarily reflect the actual visual quality of all the images. Therefore, in this paper, the IQA metric proposed in Heusel(Wu et al. 2023) is adopted to evaluate the quality of individual images. In Heusel(Wu et al. 2023), the authors classify image quality into five levels: excellent, good, fair, poor, and bad, corresponding to scores ranging from 5 to 1. A higher composite score indicates better image quality, which can very well reflect the visual quality of each generated image.

In our experiments, we used six StyleGAN generator models to generate six categories image datasets: CELEBA, CHURCH, FFHQ, BEDROOM, HORSE, and CAT, each containing 4096 different images. And it should be emphasized that our steganography scheme has been successfully implemented on all six datasets. We calculated the FID score, the average IQA score and DR (Distortion Rate) in each dataset, and the calculation formula for DR is as follows (3). Furthermore, we also analyzed the distribution of IQA scores for images in the six datasets, as shown in Fig. 6.

$$DR = \frac{N_{distortion}}{N_{number}} \cdot 100\% \tag{3}$$

where N_{number} represents the total number of images in each image dataset, while $N_{distortion}$ represents the number of distorted images. Furthermore, distorted images are defined as all images with an IQA score less than T. The value of T is determined through multiple experiments and manual review, which varies for different types of datasets. Table 1 shows the corresponding T values for different datasets.

From Table 1, we observe that the FID values for all six datasets are relatively low, and the FID scores of the CELEBA, FFHQ, and CAT datasets are around 10, while those of the CHURCH, BEDROOM, and HORSE datasets are around 5, which indicates that the StyleGAN generator models are stable, and the distribution of generated images closely resembles that of real images. However, based on the validation results of the paper(Jayasumana et al. 2023), it is evident that there is a discrepancy between the FID scores and the assessments of the human evaluators, indicating that the FID metric does not accurately capture the level of distortion in images. Our experiments have also confirmed this observation, since despite the low FID scores of the generated datasets, there are still instances of image distortion.

Table 1 Image quality data in different datasets

DataType	FID(<50)	IQA	DR(%)	T
CELEBA	9.46	3.96	0	2.7
CHURCH	4.87	3.05	0.24	2.1
FFHQ	10.16	3.98	0	2.8
BEDROOM	5.40	3.48	0.17	2.1
HORSE	5.08	2.90	62.03	3.0
CAT	11.12	3.26	34.03	3.0

To prevent the occurrence of distortion in individual images, we further conducted IQA analysis experiments. From Table 1 and Fig. 6, it can be observed that in datasets such as CELEBA and FFHQ, which are facial datasets, the average IQA of images is close to 4, with the majority of images having IQA distribution between 3 and 4. This indicates that the images in these datasets have high image visual quality, and the image distortion rates in the two datasets are both 0. In the CHURCH and BEDROOM datasets, the average IQA of the images is approximately 3 and 3.5, respectively, with most images having an IQA score between 2 and 4. Although there are some images with lower IQA scores in these two datasets, through multiple experimental tests and manual reviews, we have found that when the IQA score of the images in these datasets is greater than 2.1, distortion is generally not observed. Therefore, the distortion rate in these two datasets is relatively low. In animal datasets like CAT and HORSE, we have found that when the value of T is not less than 3, it can ensure that the generated images will not be distorted. However, a large proportion of images in these two data sets have IQA scores below 3, especially in the HORSE dataset, where the proportion exceeds 50%. Consequently, this leads to a higher distortion rate in the images of these two datasets.

Thus, we recommend using CELEBA, FFHQ, CHURCH, and BEDROOM datasets in our scheme. Additionally, to address the issue of distorted images in datasets like CAT and HORSE, we introduced a filtering operation during the image generation process. Specifically, we first calculate the IQA score of the generated

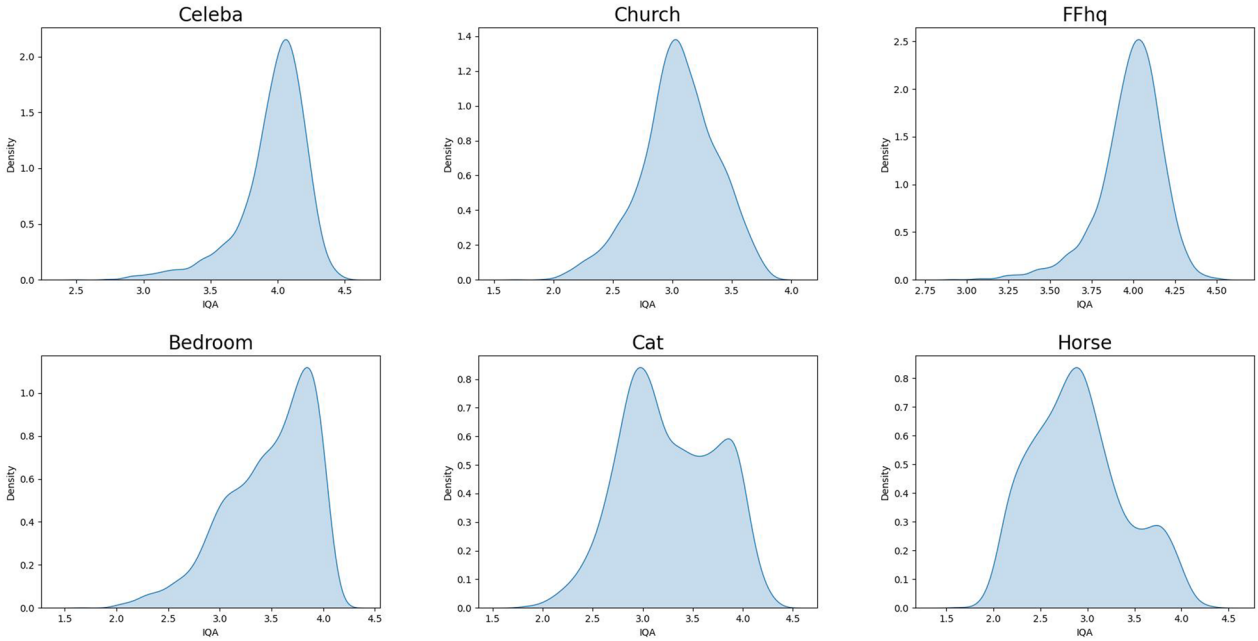

Fig. 6 The distribution of image IQA in different datasets

images, and only when the score is not less than 3 do we consider the generated images for extractor training E_x. This ensures that the secret images generated in the SeFF module will not be distorted. Even with generator models like HORSE and CAT presented in this paper, we are able to apply them to our steganography scheme while ensuring the quality of transmitted images. And we have also uploaded samples of the generated steganography images to github (RenChang 2024).

Communication capacity

This paper uses the length of message that can be transmitted by each image as the communication capacity. It is called absolute capacity AC in the literature (Hu et al. 2018). The larger the AC, the higher the communication capacity. In selection-based coverless steganography methods, AC is related to the mapping rule. Due to the limit of the algorithm, AC is generally limited to 18 bits/image, while in generation-based coverless steganography methods AC is generally above 100 bits/image (Table 2).

The capacity of SIFT-HASH, PIXEL, and CIHDN is limited to 18 bits due to the algorithm limit of the selection-based coverless steganography methods. IDEAS and RoSteAls, generation-based coverless steganography methods, have higher communication capacity. The communication capacity of MDI is only 11 bits/image due to the limit in categories of its facial attribute. Our joint steganography scheme combines the communication capacity of the two modules, breaking through the limit in selection-based steganography methods and achieving higher communication capacity.

Table 2 Communication capacity of schemes

Scheme	AC(bit(s)/image)
SIFT-HASH(Zheng et al. 2017)	18
PIXEL(Zhou et al. 2015)	8
CIHDN(Wang and Wu 2020)	8
IDEAS(Liu et al. 2022)	256
MDI(Xue and Wang 2021)	11
RoSteAls(Bui et al. 2023)	56
JoCS	48

Completeness

The coverless steganography methods design a mapping rule between the secret message and the image, based on which the sender sends the secret image indexed to specific secret message. Therefore, it must be ensured that there is at least one image corresponding to each message. If some secret messages cannot be matched to any image, there is a risk of communication failure. Therefore, excellent completeness of the coverless steganography methods is a prerequisite to ensure that the sender and the receiver can communicate normally.

This paper uses the coverage rate (CR) proposed in CIHDN (Wang and Wu 2020) as a measure of completeness. The coverage rate CR is defined as:

$$CR = \frac{N_{codeword}}{N_{number}} \cdot 100\% \tag{4}$$

where N_{number} is the number of cover images in the image database, $N_{codeword}$ represents the number of messages with duplicates removed that the image database can map.

A scheme with 100% coverage means that all message fragments are able to find the corresponding images in the cover image database. From Table 3 we can see that CIHDN, MDI, and the scheme proposed in this paper are able to achieve 100% coverage, and all of them are able to fully utilize the image database to map the messages and communicate stably under the corresponding AC. CIHDN achieves high coverage based on the mapping rule of the code-table between secret message and corresponding images. While MDI selects several main attributes of the human face as the basis for establishing mapping rule, which achieves a high degree of completeness at the expense of the communication capacity of the scheme. Our scheme utilizes the one-to-one correspondence between the input vectors and the images generated in GAN to achieve 100% coverage.

Although the SIFT-HASH scheme is able to reach 18bits/image, its coverage is only 38.41%, that is, the number of messages with duplicates removed only accounts for 38.41% of the number of cover images, as a result of which some segments cannot be mapped, which greatly increases the uncertainty of the communication process. The coverage of the PIXEL scheme is only 28.52%, since it shares a similar shortage in methods with SIFT-HASH. Due to the fact that the mapping rules established by the above two schemes are uncontrollable, the features corresponding to some of the message segments will always be missing from the image database they build. Whereas IDEAS and RoSteAls directly generate secret images by secret messages, there is no process of using a secret message to match a secret image in image database, so their coverage generally cannot be calculated.

Table 3 Coverage of coverless steganography schemes

Scheme	AC(bit(s)/image)	Number of images	Coverage
SIFT-HASH(Zheng et al. 2017)	18	262144	38.41%
PIXEL(Zhou et al. 2015)	8	256	28.52%
CIHDN(Wang and Wu 2020)	8	256	100%
IDEAS(Liu et al. 2022)	256	–	–
MDI(Xue and Wang 2021)	11	2048	100%
RoSteAls(Bui et al. 2023)	56	–	–
JoCS	48	65536	100%

It is worth mentioning that our steganography scheme is not limited by the size of dataset. Given that the image dataset generated by the SeFF module achieves 100% coverage, we can expand the scale of the dataset by using multiple different images to represent the same message codeword. This approach also increases the diversity of images transmitted during the communication process, avoiding the alertness of third parties caused by the frequent use of certain images.

Security

In the JoCS scheme of this paper, the process of constructing mapping rules in the SeFF module is essentially a random semantic editing of the image. Typically, image editing in GANs is performed by linear manipulation (addition, subtraction, multiplication, scaling) along specific semantically related directions in the latent vector. In SeFF, secret messages are added in the latent vector using the above method, and the direction of semantic editing is random depending on the added message codeword. As a result, the stego image containing the mapping rules shows no regular changes in statistical, residual, or other features, making it difficult for steganalysis tools to successfully detect the SeFF module; the TrDS module modifies the image vectors in the latent domain and adds image reconstruction loss and image perceptual loss during the model training process. As a result, the stego image also has excellent security. Therefore, the JoCS scheme based on the above two modules has excellent security. In order to verify the above view, we conducted the following experiment.

In this paper, we choose three steganalysis methods based on hand-crafted features, SPAM686(Pevnỳ et al. 2009), SRM(Fridrich and Kodovsky 2012), SRMQ1(Fridrich and Kodovsky 2012) and four steganalysis methods based on deep learning, SRNet (Boroumand et al. 2018), YeNet (Ye et al. 2017), XuNet (Xu et al. 2016) and YedroudjNet (Yedroudj et al. 2018), to detect our JoCS scheme, and we randomly generate 10,000 cover images and the corresponding stego images. For the steganalysis methods based on hand-crafted features, cover images were randomly paired with stego images and split into training and testing sets in a 1:1 ratio. For the steganalysis methods based on deep learning, the images were divided into training, validation, and testing sets in a 5:1:4 ratio.

As shown in Table 4: the detection results of the above seven steganalysis tools based on the testing sets

Table 4 Detection rate under steganalysis tools

Method	Detection rate(%)
SPAM686(Pevnỳ et al. 2009)	49.88
SRM(Fridrich and Kodovsky 2012)	50.09
SRMQ1(Fridrich and Kodovsky 2012)	49.63
SRNet(Boroumand et al. 2018)	50.13
YeNet(Ye et al. 2017)	50.00
XuNet(Xu et al. 2016)	50.25
YedroudjNet(Yedroudj et al. 2018)	50.54

are all around 50%, which means that the steganalysis tools almost randomly determine whether the image is a cover or stego. In addition, the loss of the four steganalysis models based on deep learning show no decrease after 100 epoch of training, indicating that they could not learn effective features to make judgments. This is because the current mainstream steganalysis tools primarily use features such as statistical, residual, transform, and other features extracted from images as classification criteria, which achieved good results when facing spatial domain and transform domain steganography methods (such as S-UNIWARD(Holub et al. 2014), HUGO(Pevnỳ et al. 2010)). However, in our scheme, the steganography operation in the SeFF module is merely random semantic edits of the image, and the steganography operation in the TrDS module is performed in the latent space of the image. Moreover, two types of loss functions are added during the image reconstruction process to ensure the quality of image reconstruction. Therefore, the aforementioned steganalysis tools are unable to detect the steganography operations in the JoCS scheme proposed in this paper, which also confirms that the proposed scheme has good security.

Robusteness

The main application scenario of coverless steganography methods is the social network, which is a lossy channel, and it is difficult to maintain the consistency of images sent and received, which will also affect the extraction process of the secret message during the transmission process. Therefore, the robustness of coverless steganography methods determines whether the receiver can correctly extract the secret message, which in turn affects the whole communication process. Thus, the robustness of coverless steganography methods is a critical indicator.

In order to evaluate and compare the performance of different coverless steganography schemes in terms of robustness, we use several common image attack algorithms to process secret images, and evaluate the robustness performance of each scheme in different datasets.

This paper uses bit accuracy (BA) as the measure of robustness, and secret messages are composed of binary codewords. The definition of BA is as follows:

$$BA = \frac{\sum_{i=0}^{n-1} sim(m_i, m_i')}{n \cdot l} \cdot 100\% \tag{5}$$

where i represents the image sequence number, n represents the number of images that need to be decoded, $sim()$ represents the number of equal values in the same position of the two inputs, that is, the number of *True* in $[m_i = m_i']$, m_i represents the message fragment sent by the sender, m_i' represents the message fragment recovered by the receiver, and l represents the absolute capacity AC of the scheme.

In addition to common image attack algorithms in real communication processes, we also tested the steganography schemes using two adversarial attack algorithms, FGSM (Goodfellow et al. 2014) and PGD (Madry et al. 2017), which has not been conducted in lots of steganography schemes. The image attack algorithms and parameters used in this paper are as follows.

(1) Color Histogram.
(2) Edge cropping, including coefficients of 0.05, 0.1, and 0.2.
(3) Center cropping, including coefficients of 0.2 and 0.5.
(4) Rotation, including coefficients of 30 and 50.
(5) Scaling, including coefficients of 0.3, 0.5 and 1.5.
(6) Gaussian noise, with a mean of 0 and variances of 0.001, 0.005, and 0.01.
(7) Salt and pepper noise, with a mean of 0 and variances of 0.001, 0.005, and 0.01.
(8) Mean filtering, filtering window size of 3×3, 5×5, 7×7.
(9) Median filtering, filtering window size of 3×3, 5×5, 7×7.
(10) Gaussian filtering, filtering window size of 3×3, 5×5, 7×7.
(11) JPEG compression, with compression coefficients of 0.3, 0.5, 0.7.
(12) FGSM (Fast Gradient Sign Method)
(13) PGD (Projected Gradient Descent)

Tables 5 and 6 show the BA of the SeFF module, the TrDS module, and the JoCS scheme together with other 6 coverless steganography schemes under several common image attacks on two datasets, and the results in two tables are the average of multiple experiments. Additionally, we conducted the same experiments

Table 5 Under CELEBA datasets, compared BA with the coverless steganography methods

Data		CELEBA										
Attack type	**ratio**	**SIFT-HASH**	**PIXEL**	**CIHDN**	**SeFF**	**IDEAS**	**MDI**	**RoSteAls**	**TrDS-**	**TrDS**	**TrDS+**	**JoCS**
		Zheng et al. (2017)	**Zhou et al. (2015)**	**Wang and Wu (2020)**		**Liu et al. (2022)**	**Xue and Wang (2021)**	**Bui et al. (2023)**				
Color Histogram	none	68.88	95.86	94.33	77.47	91.28	91.08	95.14	95.36	99.97	99.26	92.47
Edge Cropping	0.05	81.24	93.88	99.07	100.00	92.56	–	91.93	96.97	99.99	99.64	99.99
	0.1	79.67	87.16	97.26	95.87	90.71	–	88.03	96.97	99.99	99.65	98.62
	0.2	74.17	73.89	92.62	62.72	82.76	–	81.23	96.40	99.98	98.77	89.16
Center Cropping	0.2	98.13	91.93	99.07	75.43	97.97	–	98.57	96.49	99.99	99.54	93.95
	0.5	78.74	65.12	68.02	54.06	86.95	–	96.67	91.33	99.20	96.06	83.86
Rotation	30	53.77	77.70	76.48	50.39	64.23	90.13	53.58	55.78	51.02	65.27	50.81
Scaling	0.3	64.39	99.94	96.32	100.00	82.41	80.33	77.37	71.04	99.92	97.45	99.95
	0.5	72.62	99.97	98.74	100.00	90.14	89.92	85.27	91.07	99.99	99.54	99.99
	1.5	89.25	99.98	99.12	100.00	98.52	92.14	98.63	96.79	99.99	99.63	99.99
Gaussian noise	0.001	98.17	100.00	100.00	100.00	99.17	92.66	98.74	96.99	99.99	99.60	99.99
	0.005	98.17	100.00	100.00	100.00	99.17	92.03	98.74	96.99	99.99	99.60	99.99
	0.01	98.17	100.00	100.00	100.00	99.17	91.57	98.74	96.99	99.99	99.60	99.99
Salt noise	0.001	94.06	99.93	100.00	100.00	99.13	92.21	98.50	96.42	99.99	99.54	99.99
	0.005	84.05	99.91	100.00	100.00	99.08	92.03	97.68	94.14	99.99	99.29	99.99
	0.01	72.79	99.82	99.46	100.00	97.80	90.89	96.43	91.96	99.99	98.81	99.99
Mean filtering	3x3	71.88	99.96	98.88	100.00	99.82	91.84	98.42	95.90	99.99	99.62	99.99
	5x5	65.04	99.96	93.45	100.00	99.49	90.77	97.98	93.77	99.99	99.57	99.99
	7x7	60.43	99.92	87.45	100.00	98.65	88.91	97.25	92.45	99.99	99.53	99.99
Median filtering	3x3	76.28	99.86	99.75	100.00	98.78	92.36	98.51	96.11	99.99	99.62	99.99
	5x5	68.35	99.75	96.29	100.00	96.22	90.89	98.10	94.34	99.99	99.60	99.99
	7x7	63.29	99.65	95.54	100.00	92.12	89.25	97.49	93.01	99.99	99.53	99.99
Gaussian filtering	3x3	74.09	99.96	99.65	100.00	99.86	92.16	98.53	96.27	99.99	99.63	99.99
	5x5	69.63	99.96	97.61	100.00	98.14	90.67	98.35	95.65	99.99	99.61	99.99
	7x7	65.65	99.96	93.85	100.00	97.57	88.62	98.15	94.68	99.99	99.59	99.99
JPEG Compress	0.3	77.30	99.92	100.00	100.00	64.00	–	85.29	80.73	99.91	98.25	99.93
	0.5	80.44	99.93	100.00	100.00	79.43	–	95.14	93.11	99.97	98.98	99.97
	0.7	83.05	99.98	100.00	100.00	86.95	–	97.54	96.34	99.99	99.47	99.99
FGSM(Goodfellow et al. 2014)	none	–	–	100.00	100.00	54.63	50.84	49.88	50.06	51.17	99.57	99.71
PGD(Madry et al. 2017)	none	–	–	100.00	100.00	52.75	52.12	53.74	49.42	50.72	96.28	97.52

Table 6 Under CHURCH datasets, compared BA with the coverless steganography methods

Data		CHURCH									
Attack type	**ratio**	**SIFT-HASH**	**PIXEL**	**CIHDN**	**SeFF**	**IDEAS**	**RoSteAls**	**TrDS-**	**TrDS**	**TrDS+**	**JoCS**
		Zheng et al. (2017)	**Zhou et al. (2015)**	**Wang and Wu (2020)**		**Liu et al. (2022)**	**Bui et al. (2023)**				
Color Histogram	none	66.87	93.94	93.12	78.78	93.48	94.28	95.49	99.98	99.08	92.91
Edge Cropping	0.05	81.25	88.15	99.80	99.94	90.12	91.03	97.01	99.99	99.45	99.97
	0.1	80.66	78.47	98.33	94.54	88.45	87.33	97.01	99.99	99.44	98.17
	0.2	74.12	65.12	87.15	69.35	82.87	80.51	96.41	99.97	98.68	89.76
Center Cropping	0.2	97.67	89.19	100.00	82.12	98.45	98.17	96.63	99.99	99.18	94.03
	0.5	76.22	61.15	58.88	59.66	86.81	95.92	91.64	99.21	95.95	86.03
Rotation	30	51.19	58.48	73.17	51.72	67.17	58.46	55.47	49.95	65.27	50.54
Scaling	0.3	58.83	99.80	90.17	99.87	83.34	73.92	70.98	99.93	97.35	99.91
	0.5	61.59	99.92	96.93	99.98	92.56	90.12	91.30	99.99	99.29	99.99
	1.5	78.62	99.97	99.73	100.00	98.72	98.13	96.82	99.99	99.45	99.99
Gaussian noise	0.001	97.82	100.00	100.00	100.00	99.97	98.36	97.02	99.99	99.45	99.99
	0.005	97.82	100.00	100.00	100.00	99.70	98.36	97.02	99.99	99.45	99.99
	0.01	97.82	100.00	100.00	100.00	99.70	98.36	97.02	99.99	99.45	99.99
Salt noise	0.001	95.05	99.92	100.00	100.00	99.70	98.22	96.45	99.99	99.45	99.99
	0.005	85.80	99.79	99.75	100.00	99.49	97.59	94.20	99.99	99.45	99.99
	0.01	76.44	99.68	99.12	100.00	98.12	96.78	91.85	99.99	99.45	99.99
Mean filtering	3x3	60.80	99.96	92.91	100.00	99.92	98.08	95.93	99.99	99.45	99.99
	5x5	58.42	99.86	71.97	100.00	99.35	97.33	93.90	99.99	99.45	99.99
	7x7	57.67	99.78	64.79	100.00	99.20	96.49	92.57	99.99	99.45	99.99
Median filtering	3x3	62.85	99.53	97.55	100.00	99.92	98.04	95.93	99.99	99.45	99.99
	5x5	59.45	99.18	87.59	100.00	99.80	97.46	93.90	99.99	99.45	99.99
	7x7	58.58	98.77	75.73	100.00	98.68	96.74	92.57	99.99	99.45	99.99
Gaussian filtering	3x3	61.72	99.93	96.43	100.00	99.91	98.20	96.33	99.99	99.45	99.99
	5x5	60.04	99.87	88.08	100.00	99.84	97.14	95.72	99.99	99.45	99.99
	7x7	58.57	99.70	73.43	99.97	99.65	96.65	94.84	99.99	99.45	99.98
JPEG Compress	0.3	66.55	99.82	99.75	100.00	67.84	95.05	80.62	99.93	99.31	99.95
	0.5	69.30	99.92	99.75	100.00	73.69	97.13	93.27	99.98	99.45	99.99
	0.7	72.33	99.93	100.00	100.00	80.43	98.08	96.41	99.99	99.45	99.99
FGSM(Goodfellow et al. 2014)	none	–	–	100.00	100.00	51.24	48.11	51.63	53.82	99.52	99.68
PGD(Madry et al. 2017)	none	–	–	100.00	100.00	53.18	50.95	50.24	51.11	96.73	97.82

on the other four datasets (FFHQ, BEDROOM, CAT, HORSE). The results showed that the performance of the scheme in these four datasets was similar to that in CELEBA and CHURCH. Due to space constraints, these tables for the other four datasets are not included in this paper, and we have uploaded them to github (RenChang 2024).

Among them, SIFT-HASH, PIXEL, CIHDN, and SeFF are selection-based coverless steganography methods. The SIFT-HASH scheme cannot resist various image attacks in the two datasets, while PIXEL can resist all kinds of noise and filtering attacks. The experimental results of CIHDN in the two datasets are quite different, indicating that this artificially assigned mapping rule is not stable.

SeFF shows excellent robustness against all image attacks except color histogram, center cropping, and edge cropping with larger coefficients; this is because the mapping rule in the SeFF module is very robust. When the image is attacked, its coarse features can be well preserved. In addition, the receiver recovers the bit value of the secret message by calculating the mean value of the vector segment, so that even if some error occurs, the extractor can mitigate the impact by calculating the mean value of the vector segment.

IDEAS, MDI, RoSeAls, and TrDS are generation-based coverless steganography methods. Despite the high capacity of most generation-based coverless schemes, the recovery accuracy of secret messages is always low due to the irreversibility of the generation process. IDEAS and RoSteAls achieve high BA when faced with noise attacks and filtering attacks, whereas they cannot accurately recover the secret messages under cropping and JPEG compression attacks. Moreover, their performance deteriorates further when faced with adversarial attacks such as FGSM and PGD. MDI, on the other hand, is not able to recover the secret messages accurately, since image attack will affect the classifier's discrimination of the facial attributes, and geometric attacks will directly lead to the absence of their facial images, resulting in the inability to achieve the recovery of the secret message. The CHURCH dataset does not contain face images, so we did not conduct MDI experiments in this dataset.

TrDS achieves better results against various image attacks. This is because we add an attack module to attack the training data with noise, filtering, cropping, and others when training Resnet50 module, which helps the Resnet50 model achieve stronger robustness. In addition, in the steganography of the TrDS module, the secret message is added in the main content area of the image, which will probably not be cropped in communication on social networks, and thus the TrDS module excels in robustness. However, the TrDS module exhibits poor robustness against FGSM and PGD adversarial attacks. The perturbed image samples after adversarial attacks interfere with the ResNet50 model in the TrDS module. To address this issue, we introduced adversarial training during the model training process. Specifically, we add slight perturbations to the original training samples and compute the gradients with respect to the loss function. Then, we update the perturbed training samples along the direction of the gradients with a certain step size. We iterate this process 40 times on the original training samples to obtain adversarial samples for training. Finally, we use the mixed images containing adversarial samples to train the new model. We refer to the updated scheme as TrDS+. TrDS+ demonstrates good performance under FGSM and PGD adversarial attacks. Additionally, the addition of adversarial training often leads to a decrease in the overall accuracy of the model. In TrDS+, the recovery accuracy of secret messages only decreases by approximately 0.5% under various image attack algorithms, while exhibiting good robustness against adversarial attacks. This slight decrease in overall model performance is detailed in the TrDS+ column of Tables 5 and 6.

Furthermore, we conducted further analysis of the TrDS module. To examine the effectiveness of the attack module during training, we introduced a control scheme called TrDS-, which represents the scheme where the attack module is omitted during the training of the ResNet50 model. As shown in Tables 5 and 6, the robustness performance of the TrDS- scheme exhibits varying degrees of decline compared to TrDS, indicating that the attack module enhances the robustness performance of the TrDS module.

Since JoCS combines the advantages of SeFF and TrDS (in the last two rows, the data in the JoCS column are obtained using TrDS+, the rest of the rows use the TrDS module), it achieves a high communication capacity while maintaining excellent robustness. In the two datasets, the communication capacity of JoCS reaches 48bits/image, which is much higher than the other selection-based coverless steganography methods. In addition, JoCS achieves recovery accuracy of about 99.9% against most image attacks. Compared with IDEAS and RoSeAls, JoCS has an absolute advantage in terms of robustness.

Complexity

In deep learning, it is typical to consider both the computational load of the network and the number of parameters in the model. The former determines the duration of network training, while the latter dictates the amount of memory needed during training. Additionally, in the practical application process of steganography schemes, we usually need to account for the time required for the

sender to prepare cover images and embed secret messages, as well as the time required for the receiver to extract and recover secret messages. This determines whether the scheme can be used in real-time programs. Therefore, to explore what computational complexity the JoCS employs to achieve the above performance and whether JoCS can be applied to real-time programs, this paper conducts a quantitative analysis of the computational complexity of JoCS.

The analysis of the computational complexity of JoCS mainly focuses on two aspects: the training time of the extractor model and the time consumed during the steganography process. The former includes the training time of the overfitting extractor E_x in the SeFF module and the training time of the ResNet50 decoder in the TrDS module. The latter involves the time required for the sender to generate stego images, encode images, add secret messages, and reconstruct images, and the time required for the receiver to use two extractors to recover secret messages, which respectively represent the time required for the sending and receiving stages.

The data in Table 7 was measured on the 1080Ti GPU. From Table 7, it can be inferred that the training time of the extractor model module E_x is related to the length of the secret message n transmitted in the SeFF. This is because the number of images that the model needs to fit should be no less than 2^n, indicating that the training time increases exponentially with n. For example, when n is set to 8, 12, and 16, the corresponding training time is 10min, 1 h, and 36 h respectively. Additionally, according to our tests, after n exceeds 16 bits, it takes at least 48 h to train an extractor model. Although this can improve the communication capacity of the scheme, it is too inefficient. Therefore, in our scheme, we recommend setting n to 8-16 bits, which allows both sides of the communication to reasonably adjust the communication capacity based on their computational resources. On the other hand, the training time of the Resnet50 decoder in the TrDS module is relatively fixed, and it takes about 40 h to train a model.

The steganography time for the sender in Sending state and the extraction time for the receiver in Receiving state are the average results obtained from experiments using 4096 images, under the JoCS scheme, it takes about 0.11s for the sender to prepare a stego image, and it takes about 0.08s for the receiver to complete the extraction of the secret message from a received stego image.

Table 7 The average time cost at different stages

State	SeFF	TrDS	JoCS
Training	≤36h	40h	76h
Sending	0.02s	0.09s	0.11s
Receiving	0.03s	0.05s	0.08s

In practical usage scenarios, both sides of communication typically possess pre-trained generator and extractor models. Therefore, it is only necessary to consider the time spent on the sender's steganography process and the receiver's extraction process. As indicated by the experimental results, the time consumed in both the sender's steganography process and the receiver's extraction process is minimal, making it entirely acceptable for practical communication.

Further discussion

In Section "Experiment", several performances of the JoCS scheme proposed in this paper are evaluated and compared with other schemes. In this section, several experiments are conducted to further analyze the scheme proposed in this paper from multiple perspectives.

- In Sect. "The Advantages of Mapping Rule in the SeFF Module", we analyze why the StyleGAN generator model was selected as the basis to establish the mapping rule in the SeFF module.
- In Sect. "The Verification of Independence Between Two Modules", we conduct a theoretical analysis of the independence between the two modules of the JoCS scheme and design an experiment to verify it.
- In Sect. "The Flexibility of the JoCS Scheme", we explore the flexibility of the JoCS scheme.

The advantages of mapping rule in the SeFF module

The robustness of selection-based coverless steganography schemes often depends on the design of the mapping rules. In CIHDN (Wang and Wu 2020), Wang et al. construct a code table that could directly assign the message codeword to the image and fit the mapping rule in the code table with a neural network. However, this artificially assigned mapping rule is unreasonable, which is reflected in the robustness of the scheme. When the transmitted image is attacked, the mapping rule will also be disrupted, as a result of which the receiver cannot recover the secret message correctly.

However, the above problems do not exist in the SeFF module proposed in this paper. The SeFF module makes full use of the progressive training characteristics of the StyleGAN generator network, which establishes a mapping rule between the input vector of the low resolution layer and the coarse features of the generated image. Even if attacked by various image algorithms, this mapping rule will not be disrupted, which makes the SeFF module very robust. To prove this we conducted the following experiments:

In the SeFF module of this paper, we randomly select 1024 initial vectors $(w_1, w_2...w_{1024})$ and divide them into 64 groups $(g_1, g_2...g_{64})$. According to Sect. "Semantic Factorization Fitting Module", we add the secret message $bin(i)$ to 16 vectors of each group g_i, where $bin()$ refers to the conversion of decimal to 6-bit binary, and i represents the serial number of the group. Then 1024 images are generated, and the 16 images in each group represent the same message codeword, for example, the 16 images in g_{19} all represent 010011. Finally, we use the extractor network to fit the above mapping rule. In CIHDN, we assign 16 different images to each of the 64 codewords$(000000 - 111111)$, and use Resnet18 to fit the mapping rule. After training the models of the two schemes, we analyze the distribution of visual feature points extracted from the original images and the attacked images.

By analyzing the distribution of the original image and the attacked image in feature space in the two schemes mentioned above, it clearly reflects the irrationality of the artificially assigned mapping rule in CIHDN, as shown in Fig. 7 (b) (c), when the images are attacked, the feature points distribution will be very chaotic. However, in (e) (f), the distance between feature points of images representing the same secret message is smaller,and the feature points of images representing different codewords are also easier to distinguish, which means our SeFF module is quite robust.

Moreover, compared to the mapping-based methods in current generation-based steganography research (e.g., MDI(Xue and Wang 2021), S-DRAGAN(Cao et al. 2020), etc.), the SeFF module shows a significant advantage in both accuracy and robustness when recovering secret messages. The current mapping-based methods of generation-based steganography primarily focus on attribute mapping of facial images, using binary attributes (e.g., smiling/not smiling, eyes open/eyes closed) as the linkage for mapping binary message codewords. However, these methods suffer from low communication capacity and low accuracy in recovering secret messages. In MDI and S-DRAGAN, the authors chose various editable attributes of facial images to construct mapping rules. However, due to the limitation of attribute types, the communication capacity of the two schemes is only 11bits/image and 14 bits/image, respectively. Even more concerning is the fact that the generator models struggle to effectively control the attributes of the generated images, making it difficult for classifier models to accurately distinguish between the two states of a given attribute. This leads to misjudgments of secret messages by the receiver, as observed in MDI and S-DRAGAN, where the accuracy of recovering secret messages is only around

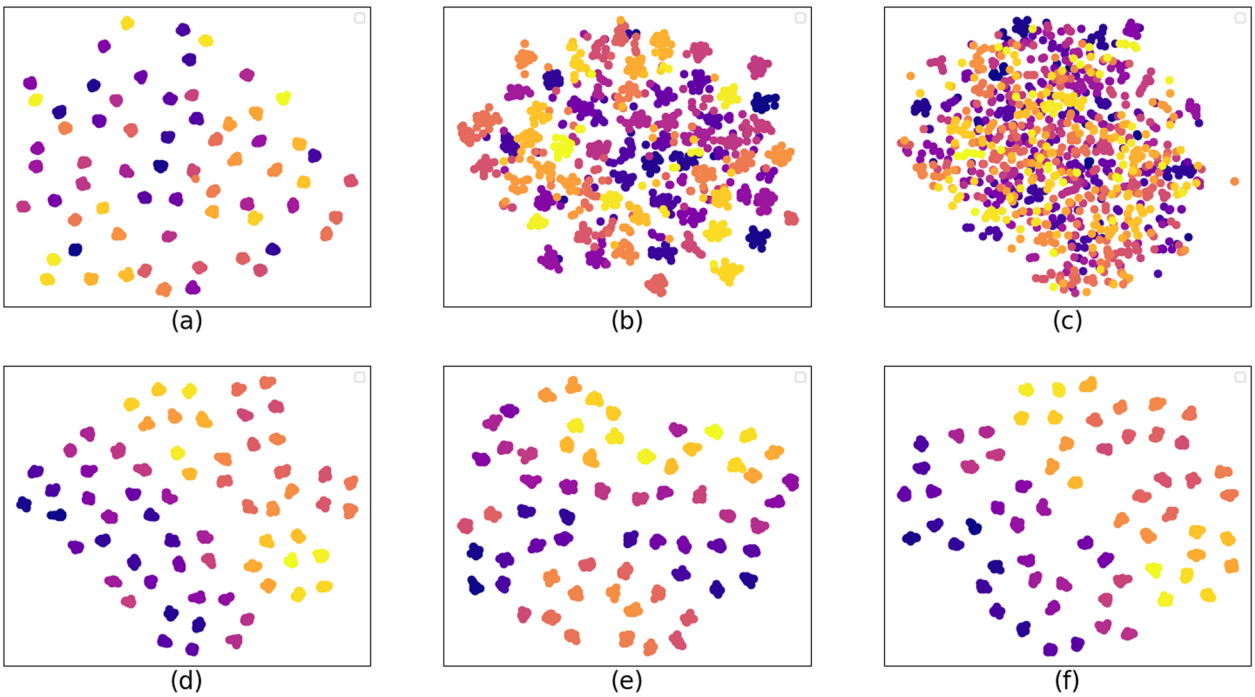

Fig. 7 **a**, **b**, **c** represent the distribution of feature points respectively extracted from the original images, JPEG Compress (0.1), and Median filtering (7x7) in the CIHDN. **d**, **e**, **f** respectively represent that in the SeFF module

90%, which is far inferior to the SeFF scheme proposed in this paper.

In summary, the SeFF module significantly enhances the robustness and accuracy of secret message recovery through a more rational design of mapping rules and more precise training of the extractor. It also effectively addresses the issues present in current mapping-based generation-based steganography methods.

The verification of independence between two modules

The JoCS scheme proposed in this paper consists of the SeFF module and the TrDS module, whose independence should be ensured, that is, the steganography between the two modules should not affect each other.

In the SeFF module, the mapping rule is established between the input vectors of the low resolution layer of the StyleGAN generator network and the coarse features of the generated image, so we need to make sure that the coarse features of the generated image are not disrupted when extracting the secret messages, which means that the steganography in the TrDS module should not disrupt the coarse features of the image. When we train the models in the TrDS module, in addition to setting the recovery loss of the secret message, we also add the image reconstruction loss and perception loss to reconstruct the image from the latent vector with the added secret message, which does not result in a significant change in the coarse features of the reconstructed image. As a result, adding the secret message to the latent vector of the image will not disrupt the mapping rule in the SeFF module, which is demonstrated through the following two experiments:

Firstly, we compare the visual features of the cover images and the stego images. Here, the cover images refer to those generated by the SeFF module, while the stego images are those reconstructed by the TrDS module. We randomly add the secret message of 32bit to the latent vector of the cover images and reconstruct stego images by VQGAN-Decoder in the TrDS module. As shown in Fig. 8, there is no significant difference between the cover images and the stego images in terms of visual features.

Secondly, we need to know whether the mapping rules in the cover images and the stego images are the same, so we extract secret messages from the cover images and stego images with the mapping rule established in the SeFF module and compare their BA on several image datasets. As shown in Table 8, $SeFF_c$ represents the BA recovering from the cover images and $SeFF_s$ represents the BA recovering from the stego images. In the six data sets, there is no significant difference between $SeFF_c$ and $SeFF_s$. As we expected, the steganography in the TrDS module will not disrupt the mapping rule in the SeFF module. In addition, we also use steganalysis tools to detect cover images and stego images. As shown in Table 9, taking the results on the CELEBA and CHRUCH datasets as examples, the detection results of $SeFF_c$ and $SeFF_s$ on both datasets are around 50%. This further illustrates that there are no features differences that would give rise to suspicion by a third party between the cover images and stego images. The results tested on the other

Table 8 The comparison between $SeFF_c$ and $SeFF_s$ on the six datasets

Data	$SeFF_c$	$SeFF_s$
CELEBA	100.00	100.00
CHURCH	100.00	99.99
FFHQ	100.00	100.00
BEDROOM	100.00	99.99
CAT	100.00	100.00
HORSE	100.00	100.00

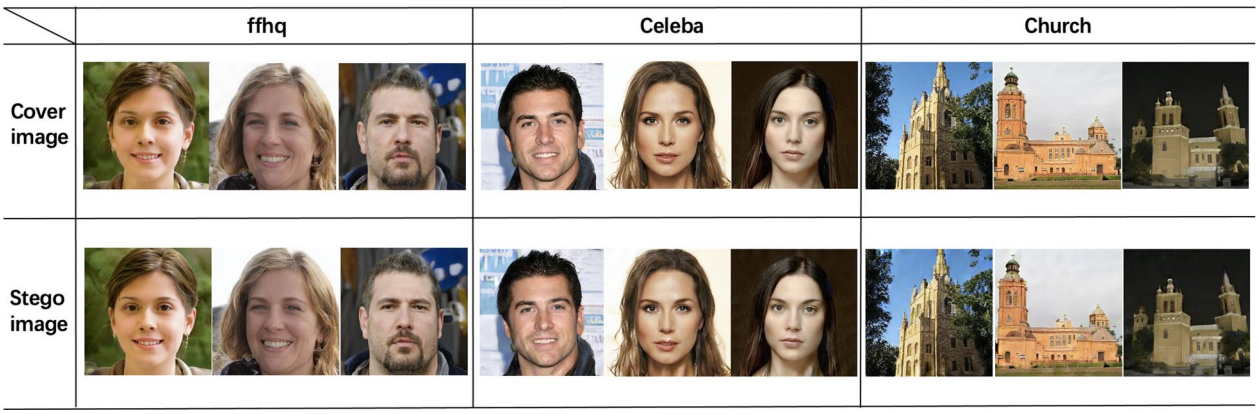

Fig. 8 Visualization comparison in cover and stego images

Table 9 Detection rates of two image sets under steganalysis tools

Data	CELEBA		CHURCH	
Detection method	$SeFF_c$	$SeFF_s$	$SeFF_c$	$SeFF_s$
SRNet(Boroumand et al. 2018)	50.12	49.98	49.92	50.01
YeNet(Ye et al. 2017)	50.02	49.99	50.00	49.97
XuNet(Xu et al. 2016)	50.24	50.20	50.11	50.13
YedroudjNet(Yedroudj et al. 2018)	50.23	50.16	50.28	50.14

four datasets are also similar to the above results. On the basis of the above experiments, the steganography in the two modules is independent of each other, and it is reasonable to design a JoCS scheme to combine them.

The flexibility of the JoCS scheme

The key to successfully achieving the JoCS scheme in this paper is that the steganography in the TrDS module will not disrupt the mapping rule in the SeFF module, based on which we can speculate that on the premise of ensuring that the two modules in the scheme are independent of each other, the SeFF module can be replaced with other selection-based coverless steganography schemes, or the TrDS module can be replaced with other generation-based coverless steganography schemes. Additionally, the generative model in the SeFF module can also be flexibly replaced. With many popular generative models available, the SeFF module can be combined with them flexibly. However, since the latent space structures of different generative models may vary, it is necessary to adjust the method for constructing the mapping relationship according to the specific model. The above characteristics of the JoCS scheme greatly improve the flexibility of the scheme.

In order to verify the above speculation, we first use the Feadio scheme in Wu and Xue (2024) to replace the SeFF module in the JoCS scheme of this paper, and conduct experiments on the CELEBA dataset. Feadio is a selection-based coverless steganography scheme that establishes a mapping rule between the identity information of the face image and the message codeword, and the TrDS module will not modify the identity information of the face image. Next, we also attempted to replace the generative model in the SeFF module. We replaced StyleGAN in SeFF with StyleGAN2(Karras et al. 2020) and Stable Diffusion(Rombach et al. 2022). Then, using principles similar to those in Sect. "Semantic Factorization Fitting Module", we reconstructed the mapping rules between the codewords and the generated images. Then, we conducted experimental tests.

As shown in Table 10, the Feadio-JoCS column represents the results obtained by replacing the SeFF module in the JoCS scheme with Feadio. The SeFF-StyleGAN2 and SeFF-StableDiffusion columns represent the results obtained by replacing the generative model StyleGAN in the SeFF module with StyleGAN2 and Stable Diffusion, respectively. The original Feadio scheme and the joint Feadio-JoCS scheme have similar robustness under various image attacks, while the communication capacity of Feadio-JoCS reaches 48bits/image, which is three times that of the Feadio scheme. Additionally, as shown in the SeFF-StyleGAN2 and SeFF-StableDiffusion columns, the SeFF scheme continues to demonstrate good performance even after replacing the generative model. These results demonstrate that our JoCS scheme is highly flexible, which means that the sender can choose two independent steganography modules according to the actual communication requirements to achieve the secondary steganography on a single image.

Table 10 Experimental results after replacing the SeFF module

Data		CELEBA			
Attack type	**ratio**	**Feadio[15]**	**Feadio-JoCS**	**SeFF-StyleGAN2**	**SeFF-StableDiffusion**
EndCropping	0.05	100.00	99.99	100.00	100.00
Scaling	0.3	100.00	99.95	100.00	100.00
Gaussian noise	0.01	100.00	99.99	100.00	100.00
Salt noise	0.01	100.00	99.99	100.00	100.00
Mean filtering	7x7	100.00	99.99	100.00	100.00
Median filtering	7x7	100.00	99.99	100.00	100.00
Gaussian filtering	7x7	100.00	99.99	100.00	100.00
JPEG Compress	0.3	100.00	99.95	100.00	100.00

Conclusion

In this paper, we propose a robust joint coverless image steganography scheme, JoCS, which consists of two modules: the SeFF module and the TrDS module. The former belongs to the selection-based coverless steganography method, which uses the StyleGAN generator as a hub to connect secret message and generated image, and communicates between the sender and the receiver through an extractor based on neural networks. The latter belongs to the generation-based coverless steganography method, which uses the Encoder and Decoder in VQGAN to achieve secondary steganography in the latent vector of the encoded image. Testing on six datasets and comparing with other schemes, our JoCS scheme combines the advantages of both selection-based and generation-based coverless steganography methods, demonstrating excellent practicality, robustness, and completeness. On the basis of these, it also achieves more advanced communication capacity. In addition, we also conducted a detailed analysis of the advantages of the mapping rule in the SeFF module, the independence of the two modules in the JoCS scheme, and the flexibility of the JoCS scheme. In future research, we will study and design reversible generation-based coverless steganography methods, and explore how to further improve the communication capacity and robustness of the JoCS scheme.

Acknowledgements
Thank you to Zan Ren for her assistance with the English writing of this paper.

Author Contributions
The frst author completed the main work of the paper and drafted the manuscript. The second author reviewed the manuscript and revising the article critically. He also proofread the manuscript and corrected the grammar mistakes.

Funding
This work was supported in part by the National Natural Science Foundation of China under Grant U23B2002, and in part by the National Natural Science Foundation of China under Grant 62272007.

Availability of data and materials
The relevant data and results in the experiment have been uploaded to GitHub (RenChang 2024).

Declarations

Competing interest
Both authors declare that they have no Conflict of interest

Received: 2 July 2024 Accepted: 26 September 2024
Published online: 07 December 2024

References

Boroumand M, Chen M, Fridrich J (2018) Deep residual network for steganalysis of digital images. IEEE Trans Inf Forensics Sec 14(5):1181–1193
Bui T, Agarwal S, Yu N, Collomosse J (2023) Rosteals: Robust steganography using autoencoder latent space 933–942
Cao Y, Zhou Z, Wu QJ, Yuan C, Sun X (2020) Coverless information hiding based on the generation of anime characters. EURASIP J Image Vid Process 2020:1–15
Esser P, Rombach R, Ommer B (2021) Taming transformers for high-resolution image synthesis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12873–12883
Fridrich J, Kodovsky J (2012) Rich models for steganalysis of digital images. IEEE Trans Inf Forensics Sec 7(3):868–882
Girshick R (2015) Fast r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1440–1448
Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2020) Generative adversarial networks. Commun ACM 63(11):139–144
Goodfellow IJ, Shlens J, Szegedy C (2014) Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572
Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S (2017) Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems 30
Ho J, Jain A, Abbeel P (2020) Denoising diffusion probabilistic models. Adv Neural Inf Syst 33:6840–6851
Holub V, Fridrich J, Denemark T (2014) Universal distortion function for steganography in an arbitrary domain. EURASIP J Inf Sec 2014:1–13
Hu D, Wang L, Jiang W, Zheng S, Li B (2018) A novel image steganography method via deep convolutional generative adversarial networks. IEEE access 6:38303–38314
Jayasumana S, Ramalingam S, Veit A, Glasner D, Chakrabarti A, Kumar S (2023) Rethinking fid: Towards a better evaluation metric for image generation. arXiv preprint arXiv:2401.09603
Karras T, Laine S, Aittala M, Hellsten J, Lehtinen J, Aila T (2020) Analyzing and improving the image quality of stylegan. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8110–8119
Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4401–4410
Kim D, Shin C, Choi J, Jung D, Yoon S (2023) Diffusion-stego: Training-free diffusion generative steganography via message projection. arXiv preprint arXiv:2305.18726
Liao X, Yu Y, Li B, Li Z, Qin Z (2019) A new payload partition strategy in color image steganography. IEEE Trans Circ Syst Video Technol 30(3):685–696
Liu X, Ma Z, Ma J, Zhang J, Schaefer G, Fang H (2022) Image disentanglement autoencoder for steganography without embedding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2303–2312
Luo Y, Qin J, Xiang X, Tan Y (2020) Coverless image steganography based on multi-object recognition. IEEE Trans Circ Syst Video Technol 31(7):2779–2791
Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A (2017) Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083
Mielikainen J (2006) LSB matching revisited. IEEE Sig Process Lett 13(5):285–287
Peng F, Chen G, Long M (2022) A robust coverless steganography based on generative adversarial networks and gradient descent approximation. IEEE Trans Circ Syst Video Technol 32(9):5817–5829
Peng Y, Hu D, Wang Y, Chen K, Pei G, Zhang W (2023) Stegaddpm: Generative image steganography based on denoising diffusion probabilistic model. In: Proceedings of the 31st ACM International Conference on Multimedia, pp. 7143–7151
Pevnỳ T, Bas P, Fridrich J (2009) Steganalysis by subtractive pixel adjacency matrix. In: Proceedings of the 11th ACM Workshop on Multimedia and Security, pp. 75–84
Pevnỳ T, Filler T, Bas P (2010) Using high-dimensional image models to perform highly undetectable steganography. In: Information Hiding: 12th International Conference, IH 2010, Calgary, AB, Canada, June 28-30, 2010, Revised Selected Papers 12, pp. 161–177. Springer
Pevnỳ T, Bas P, Fridrich J (2009) Steganalysis by subtractive pixel adjacency matrix. In: Proceedings of the 11th ACM Workshop on Multimedia and Security, pp. 75–84
RenChang (2024) Supplementary experimental data. Accessed: 2024-06-12. https://github.com/rchotcocoa/JoCS

Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B (2022) High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10684–10695
Su W, Ni J, Hu X, Fridrich J (2020) Image steganography with symmetric embedding using Gaussian Markov random field model. IEEE Trans Circ Syst Video Technol 31(3):1001–1015
Tao J, Li S, Zhang X, Wang Z (2018) Towards robust image steganography. IEEE Trans Circ Syst Video Technol 29(2):594–600
Wang Y, Wu B (2020) An intelligent search method of mapping relation for coverless information hiding. J Cyber Sec 5(3):48–61
Wei P, Zhou Q, Wang Z, Qian Z, Zhang X, Li S (2023) Generative steganography diffusion. arXiv preprint arXiv:2305.03472
Wu B, Xue R (2024) A coverless image steganography method using deep learning with feature distribution optimization. J Cyber Security. https://doi.org/10.19363/J.cnki.cn10-1380/tn.2024.04.05
Wu H, Zhang Z, Zhang W, Chen C, Liao L, Li C, Gao Y, Wang A, Zhang E, Sun W, et al (2023) Q-align: Teaching lmms for visual scoring via discrete text-defined levels. arXiv preprint arXiv:2312.17090
Xu G, Wu H-Z, Shi Y-Q (2016) Structural design of convolutional neural networks for steganalysis. IEEE Sig Process Lett 23(5):708–712
Xue R, Wang Y (2021) Message drives image: A coverless image steganography framework using multi-domain image translation, 1–9. IEEE
Ye J, Ni J, Yi Y (2017) Deep learning hierarchical representations for image steganalysis. IEEE Trans Inf Forensics Sec 12(11):2545–2557
Yedroudj M, Comby F, Chaumont M (2018) Yedroudj-net: An efficient cnn for spatial steganalysis, 2092–2096. IEEE
You Z, Ying Q, Li S, Qian Z, Zhang X (2022) Image generation network for covert transmission in online social network. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 2834–2842
Yu J, Zhang X, Xu Y, Zhang J (2023) Cross: Diffusion model makes controllable, robust and secure image steganography. arXiv preprint arXiv:2305.16936
Yuan C, Xia Z, Sun X (2017) Coverless image steganography based on sift and BOF. J Int Technol 18(2):435–442
Zheng S, Wang L, Ling B, Hu D (2017) Coverless information hiding based on robust image hashing. In: Intelligent Computing Methodologies: 13th International Conference, ICIC 2017, Liverpool, UK, August 7-10, 2017, Proceedings, Part III 13, pp. 536–547. Springer
Zhou Z, Sun H, Harit R, Chen X, Sun X (2015) Coverless image steganography without embedding. In: Cloud Computing and Security: First International Conference, ICCCS 2015, Nanjing, China, August 13-15, 2015. Revised Selected Papers 1, pp. 123–132. Springer

Publisher's Note