Robust Image Steganography: Hiding Messages in Frequency Coefficients

Yuhang Lan[1], Fei Shang[1], Jianhua Yang[2], Xiangui Kang[1]*, Enping Li[3]

[1]Guangdong Key Laboratory of Information Security Technology, School of Computer Science and Engineering, Sun Yat-Sen University, Guangzhou, China
[2]Guangdong Polytechnic Normal University, Guangzhou, China
[3]Computer Science Department, Bridgewater State University, Massachusetts, USA
{lanyh5, shangf5}@mail2.sysu.edu.cn, yangjh86@gpnu.edu.cn, isskxg@mail.sysu.edu.cn, eli@bridgew.edu

Abstract

Steganography is a technique that hides secret messages into a public multimedia object without raising suspicion from third parties. However, most existing works cannot provide good robustness against lossy JPEG compression while maintaining a relatively large embedding capacity. This paper presents an end-to-end robust steganography system based on the invertible neural network (INN). Instead of hiding in the spatial domain, our method directly hides secret messages into the discrete cosine transform (DCT) coefficients of the cover image, which significantly improves the robustness and anti-steganalysis security. A mutual information loss is first proposed to constrain the flow of information in INN. Besides, a two-way fusion module (TWFM) is implemented, utilizing spatial and DCT domain features as auxiliary information to facilitate message extraction. These two designs aid in recovering secret messages from the DCT coefficients losslessly. Experimental results demonstrate that our method yields significantly lower error rates than other existing hiding methods. For example, our method achieves reliable extraction with 0 error rate for 1 bit per pixel (bpp) embedding payload; and under the JPEG compression with quality factor $QF = 10$, the error rate of our method is about 22% lower than the state-of-the-art robust image hiding methods, which demonstrates remarkable robustness against JPEG compression.

Introduction

A safe and robust image steganography system attempts to hide the secret messages within the cover image, and the generated stego image is more inclined to evade malicious distortion and detection. Figure 1 shows the universal pipeline of robust image steganography. In real-world applications, the stego image inevitably encounters distortions during transmission on social networks, such as JPEG compression, which dramatically complicates extracting messages. As modifying the different pixel positions of the image has different embedding effects, traditional steganography algorithms combine the Syndrome Trellis Codes (Filler, Judas, and Fridrich 2011) and the additive distortion cost functions (Holub, Fridrich, and Denemark 2014; Li et al. 2014; Pevnỳ, Filler, and Bas 2010) to minimize the steganographic framework overall distortion. However, to evade statistical steganalysis detection, these traditional methods only have a low embedding payload. Encouraged by the representational power of convolutional networks, learning-based steganography methods achieve higher embedding capacity. Work (Hayes and Danezis 2017) proposes the first generative adversarial network (GAN) based steganography framework by simultaneously training both a steganographic generator and a steganalyzer. SteganoGAN (Zhang et al. 2019a) proposes a flexible image steganography method that expands the binary secret information into a three-dimensional tensor, significantly increasing the embedding capacity. ABDH (Yu 2020) utilizes a spatial attention mechanism to improve the visual quality of generated stego images, which embeds the secret messages into locations that are not sensitive to human vision. CHAT-GAN (Tan et al. 2021) designs a channel attention module to promote the network to concentrate on the critical channel features. However, most of these methods are vulnerable to malicious JPEG compression when transmitting stego images on lossy channels. That is, a slight disturbance usually results in a poor secret message revealing. This substantial drop in performance makes them not applicable in practical scenarios.

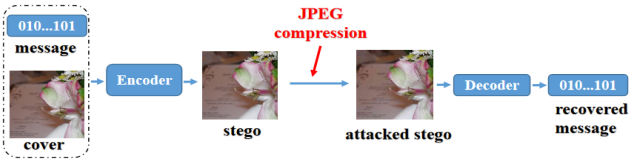

Figure 1: The universal pipeline of robust image steganography. Our model has robustness against JPEG compression with different quality factors.

This paper proposes an end-to-end robust steganography system in which the concealing and revealing of secret messages are realized by INN's forward and backward processes. Instead of choosing spatial image pixels, our method utilize the DCT coefficients for message embedding. As the quantization and rounding operations in JPEG compression directly act on the DCT coefficients, the modification of DCT coefficients by JPEG compression is intuitive and easier to model than in spatial domain or wavelet domain. However, one DCT coefficient of a block is calculated by all pixel values of the block in the spatial domain, which indicates that

*Corresponding author: Xiangui Kang

modifications in pixel values may significantly affect corresponding DCT coefficients. This complex statistical characteristic of DCT coefficients incurs the difficulty of robust steganography in the DCT domain. Our proposed method efficiently achieves JPEG robustness under any quality factors, and it meets low error rate requirements for the robust steganography algorithm. The main contributions are summarized as follows:

- We introduce an end-to-end robust steganography system based on INN, which directly hides/extracts secret messages in the DCT coefficients. Our method ensures artifact imperceptibility of the generated stego image and robustness against JPEG compression.
- We propose a mutual information loss in the concealing process and devise a two-way fusion module (TWFM) used in the revealing process. These two designs minimize the lost information and utilize the reserved crucial features to reveal the secret messages as much as possible.
- Experimental results demonstrate the superior robustness of our steganography system in lossless or distorted conditions. Specifically, our method can reliably achieve a 0 error for message extraction at an embedding payload of 1 bpp. Moreover, even with an aggressive JPEG compression of a very low quality factor, such as $QF = 10$, the error rate does not exceed 3%.

Related Work

Robust Image Hiding

Robust image steganography methods have been developed to meet the application requirements of covert communication in lossy channels, in which the model is resistant to distortions. HiDDeN (Zhu et al. 2018) proposes an end-to-end framework incorporating typical digital attacks into the encoder-decoder structure to simulate distortions. ReDMark (Ahmadi et al. 2020) presents a deep diffusion framework to improve the robustness. However, these two simple network structures provide poor universality to different distortions. Work (Wengrowski and Dana 2019) focuses on particular robustness in the light field messaging (LFM): encoded image displayed on a screen and captured by a camera. StegaStamp (Tancik, Mildenhall, and Ng 2020) uses a set of differentiable image augmentations to simulate the print-shooting process and hides hyperlinks in a physical photograph. To model the non-differentiable distortion, work (Liu et al. 2019) designs a two-stage separable framework, which jointly trains the encoder and decoder without noise in the first stage, and then presents noise attacks in the second stage but restricts the loss to only propagate back through the decoder. Later, work (Jia, Fang, and Zhang 2021) randomly chooses one of simulated JPEG, real JPEG, and noise-free layer as the noise layer for each mini-batch to enhance the robustness against JPEG compression. Besides, work (Zhang et al. 2021) uses a forward attack simulation layer to make the pipeline compatible with non-differentiable distortion. As for adaptive steganography algorithms in the frequency domain, works (Zhang et al. 2019b; Zhu et al. 2021) hide secret messages into compressed DCT coefficients. These works achieve promising robustness and security with small embedding capacity.

Invertible Neural Network

The affine coupling layer (Kingma and Dhariwal 2018) is the fundamental building block of the invertible neural network (INN) (Dinh, Krueger, and Bengio 2014). The encoding and decoding process shares the same parameters, making the model lightweight. Since the reversible network is information lossless in theory, it can preserve details of the input as much as possible. Due to these exceptional properties, many works with invertible architecture gain more satisfactory performance than traditional autoencoder frameworks, especially for image-related tasks with inherent invertibility. Work (Van der Ouderaa and Worrall 2019) applies a reversible network as the central workhorse to the image-to-image translation task, which reduces memory consumption and generates images with high fidelity. For the image rescaling task, work (Xiao et al. 2020) utilizes INN to discover a bijective mapping between high-resolution and low-resolution images, which alleviates the ill-posed problem of image upscaling reconstruction.

The image steganography composed of concealing and revealing processes can be considered inverse problems. Work (Lu et al. 2021) utilizes the INN to conduct high-capacity spatial image steganography, in which multiple secret images can be hidden into one cover image. HiNet (Jing et al. 2021) attempts to hide the secret image in the wavelet domain using INN for high invisibility. It presents a low-frequency loss to confine hiding messages into high-frequency wavelet subbands. Work (Xu et al. 2022) achieves a certain degree of robustness under various distortions using an invertible structure. Despite managing a large number of bits, all of the above INN methods incur high error rates for the recovered messages. So applying them straightly for the scenario with zero tolerance of even a single incorrectly recovered bit is impractical.

Proposed Method

Overview of the Framework

The traditional neural network utilizes two independent encoder and decoder to model the mappings among secret message $\mathbf{M}$, cover image $\mathbf{C}$, and stego image $\mathbf{S}$, which can be denoted as: $(\mathbf{M}, \mathbf{C}) \rightarrow \mathbf{S}$ and $\mathbf{S} \rightarrow (\mathbf{M}, \mathbf{C})$, respectively. This separate encoder-decoder structure will result in inaccurate bijective mapping and may accumulate the error of one mapping into the other. In our proposed method, we attempt to find an invertible and bijective network that can hide and extract the secret message simultaneously, denoted as: $(\mathbf{M}, \mathbf{C}) \leftrightarrow \mathbf{S}$. In Figure 2, the framework of our method consists of several components: INN, JPEG compression layer, discriminator, and two-way fusion module. We employ the affine coupling layers to structure the concealing and revealing blocks. There are 12 such reversible blocks for both the concealing and revealing process of the INN. These two processes share parameters during training, which drive the model more effectively than the traditional encoder-decoder

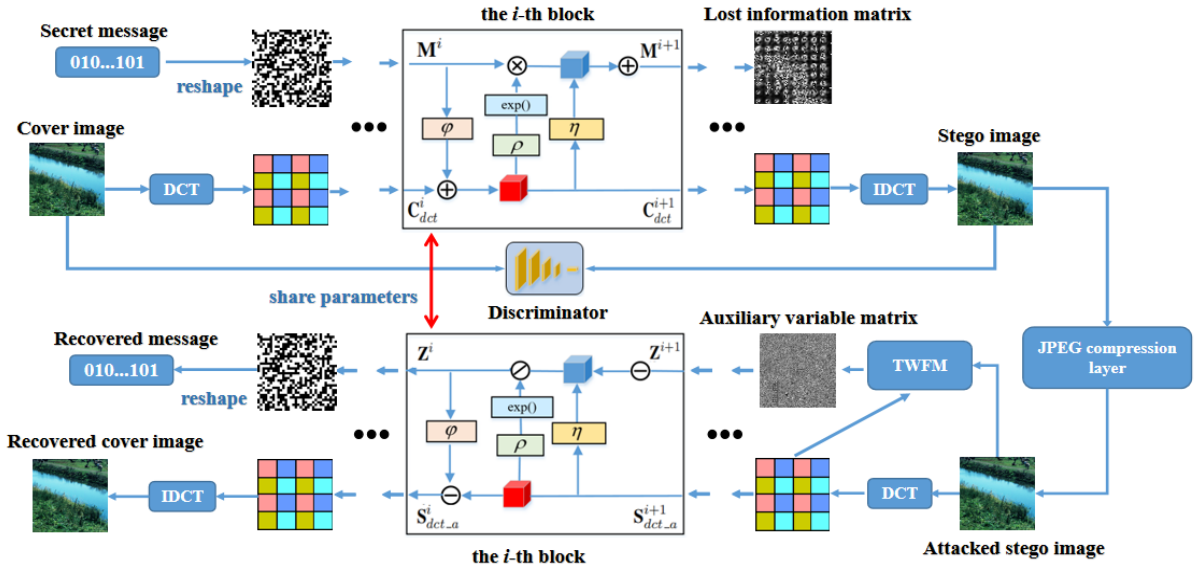

Figure 2: Overview of the proposed steganography architecture.

architecture. The JPEG compression layer is a simulator to simulate the non-differentiable JPEG compression in a differentiable and approximate form, which is then introduced to the end-to-end network to alleviate the distortion influence and efficiently improve the robustness. The adversarial discriminator is utilized to distinguish whether the generated stego image is similar to the original cover image, which takes advantage of adversarial learning to improve the visual fidelity and steganographic security of the generated stego image. The TWFM is designed delicately to combine spatial and DCT domain features fully to recover the messages.

Invertible Concealing/Revealing Process

We denote the cover image with c channels and $h \times w$ size as $\mathbf{C} \in \{0, ..., 255\}^{c \times h \times w}$, where $c = 3$, $h = 256$, and $w = 256$ in our work. The secret message is denoted as $m \in \{0, 1\}^l$ of length l. To ensure that the two inputs of INN are the same size, we manage the secret message input as a three-dimensional volume by padding each bit with the same values into 8×8 small block if the payload is less than 1 bpp. The shaped secret message is represented as $\mathbf{M} \in \{0, 1\}^{c' \times h \times w}$, where c' is the number of channel. In the forward concealing process, the cover image is converted by DCT transform to form a three-dimensional DCT coefficient cube, the same size as the spatial image. Then we utilize concealing blocks to embed secret messages $\mathbf{M}$ into DCT coefficients of cover image $\mathbf{C}_{dct}$. Considering the i-th concealing block in Figure 2, we input $\mathbf{C}^i_{dct}$ and $\mathbf{M}^i$, and output $\mathbf{C}^{i+1}_{dct}$ and $\mathbf{M}^{i+1}$, which can be formulated as:

$$\begin{aligned} \mathbf{C}^{i+1}_{dct} &= \mathbf{C}^i_{dct} + \varphi(\mathbf{M}^i) \\ \mathbf{M}^{i+1} &= \mathbf{M}^i \odot \exp(\rho(\mathbf{C}^{i+1}_{dct})) + \eta(\mathbf{C}^{i+1}_{dct})), \end{aligned} \tag{1}$$

where $\odot$ indicates the Hadamard product operation, and $\varphi(\cdot)$, $\rho(\cdot)$, and $\eta(\cdot)$ denote the arbitrary functions but do not require to be invertible. Here, we employ the 5-layer Denseblock (Wang et al. 2018) to represent these three functions for satisfactory performance in image processing. Note that after a series of concealing blocks, the model outputs the DCT coefficients of stego image $\mathbf{S}_{dct}$ and the lost information matrix $\mathbf{L}$. To obtain the corresponding stego image $\mathbf{S}$, the IDCT module receives the frequency features $\mathbf{S}_{dct}$ and transforms them back to the spatial domain.

In the backward revealing process, the DCT module receives attacked stego image $\mathbf{S}_a$ and converts it into DCT features $\mathbf{S}_{dct\_a}$. Since only the stego image can be transmitted on the communication channel, we hold the attacked stego image $\mathbf{S}_a$ at the start of the revealing process. Therefore, to substitute the lost information matrix $\mathbf{L}$ and refine essential information for the revealing, the TWFM incorporates the attacked stego image $\mathbf{S}_a$ and its frequency coefficients $\mathbf{S}_{dct\_a}$ to generate the auxiliary variable matrix $\mathbf{Z}$. The detailed structure of TWFM will be introduced next. It is noted that the concealing blocks and the revealing blocks are almost identical, except that the information flowing direction is opposite. For the i-th revealing block, the inputs are $\mathbf{S}^{i+1}_{dct\_a}$ and $\mathbf{Z}^{i+1}$, and obtains $\mathbf{S}^i_{dct\_a}$ and $\mathbf{Z}^i$ as the outputs according to the following formula:

$$\begin{aligned} \mathbf{Z}^i &= (\mathbf{Z}^{i+1} - \eta(\mathbf{S}^{i+1}_{dct\_a})) \oslash \exp(\rho(\mathbf{S}^{i+1}_{dct\_a})) \\ \mathbf{S}^i_{dct\_a} &= \mathbf{S}^{i+1}_{dct\_a} - \varphi(\mathbf{Z}^i), \end{aligned} \tag{2}$$

where $\oslash$ denotes the matrix division operation. The last revealing block outputs the recovered DCT coefficients of cover image $\mathbf{C}'_{dct}$ and the recovered message $\mathbf{M}'$. Eventually, $\mathbf{C}'_{dct}$ goes through the IDCT module to transform into the recovered cover image $\mathbf{C}'$, and $\mathbf{M}'$ reshapes and maps back to binary messages $\mathbf{m}'$.

The Two-Way Fusion Module (TWFM)

When hiding secret messages into the cover image, high-capacity embedding will inevitably damage the carrier. Besides, to avoid the visual distortion of the generated stego image, it is difficult to embed the secret message into the carrier entirely. Thus, the above two information losses constitute the lost information matrix $\mathbf{L}$. Work (Jing et al. 2021) randomly samples from a Gaussian distribution to make up the auxiliary variable matrix $\mathbf{Z}$. Then every sampled $\mathbf{Z}$ is trained to ensure that the INN can recover the secret message. However, this approach does not fully consider that the $\mathbf{L}$ contains valid information from input features, which should be utilized as much as possible in the revealing process. That is, discarding the $\mathbf{L}$ will result in accuracy degradation for the secret message recovery.

Motivated by the self-attention mechanism (Vaswani et al. 2017) that maps the input into several different patches to capture the internal correlation of the features, TWFM aims at effectively extracting and fusing spatial domain and frequency domain features to improve the accuracy of secret message recovery. The architecture of TWFM is shown in Figure 3. The inputs are the attacked stego image $\mathbf{S}_a$ and its frequency features $\mathbf{S}_{dct\_a}$. The attention weight map assigns the weights between 0 and 1 to the corresponding features in the spatial and frequency domains through an element-wise product. It selects meaningful features from the $\mathbf{S}_a$ and $\mathbf{S}_{dct\_a}$ while suppressing some irrelevant details. The overall process can be depicted as follows:

$$\begin{gathered}\mathbf{W} = \sigma[\delta(f_c(\mathbf{S}_a), f_c(\mathbf{S}_{dct\_a})) \otimes \delta(f_c(\mathbf{S}_{dct\_a}), f_c(\mathbf{S}_a))], \\ \mathbf{Z} = f_c(\mathbf{W} \otimes \mathbf{S}_a + \mathbf{W} \otimes \mathbf{S}_{dct\_a})\end{gathered} \tag{3}$$

where $\otimes$ is an element-wise product operation, and $\sigma(\cdot)$, $\delta(\cdot)$, and $f_c(\cdot)$ denote the softmax function, the concat operation, and the convolution respectively. The attention weight map $\mathbf{W}$ indicates the importance of the regional information obtained from the input feature matrix, where more attention needs to be paid to the high-weight elements. Note that TWFM is plug-and-play, that is, it can be trained end-to-end jointly with the INN.

Design of the JPEG Compression Layer

The JPEG compression process contains a non-differentiable step: the rounding function is a piecewise step function that truncates the gradient propagation. Thus, JPEG is not appropriate for direct end-to-end training optimization. To solve this problem, we adopt a smooth rounding function $R(x)$ in work (Shin and Song 2017) to simulate the rounding step, which can be formulated as:

$$R(x) = [x] + (x - [x])^3, \tag{4}$$

where $[x]$ is the rounding operation on x, and $R(x)$ represents the quantized and simulated rounded DCT coefficient. The simulated rounding function $R(x)$ is approximately continuous, which indicates that the derivative is nonzero. Through this way, the non-differentiable JPEG compression can be simulated in a differentiable form to keep gradient propagation in the training process.

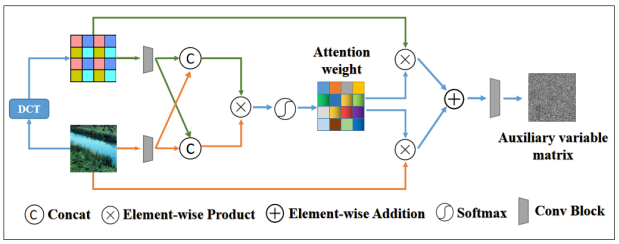

Figure 3: The architecture of the two-way fusion module.

Design of the Discriminator

The discriminator evaluates the difference between cover and stego images and provides feedback on generator performance, which further stimulates the generated instances to be closer to the data from the original class. In our proposed method, the discriminator module is composed of 6 groups. From group 1 to group 5, each group consists of a convolutional layer (kernel size = 3, stride = 2, padding = 1), a BN layer, and a LeakyReLU activation function. Group 6 contains a global average pooling (GAP) and a linear layer to output the classification probability. Benefited from the basic principle of adversarial learning, the generated stego images have higher visual fidelity and steganographic security against statistical detection.

Loss Function

To generate stego images with high fidelity and recover messages with low error, the overall optimization object is:

$$\mathcal{L}_{total} = \lambda_c \mathcal{L}_c + \lambda_r \mathcal{L}_r - \lambda_d \mathcal{L}_d + \lambda_m \mathcal{L}_m. \tag{5}$$

where λ_c, λ_r, λ_d and λ_m are the weight factors to balance different loss terms.

Concealing loss $\mathcal{L}_c$. Concealing loss $\mathcal{L}_c$ indicates the difference between cover image $\mathbf{C}$ and stego image $\mathbf{S}$. The concealing loss needs to be managed at a decent level to obtain good imperceptibility. We employ mean square error (MSE) to measure the difference between them, which can be defined as:

$$\mathcal{L}_c = MSE(\mathbf{C}, \mathbf{S}) = \frac{1}{c \times h \times w} \|\mathbf{C} - \mathbf{S}\|_2^2, \tag{6}$$

Revealing loss $\mathcal{L}_r$. Revealing loss $\mathcal{L}_r$ states the difference between the recovered message $\mathbf{M}'$ and the original secret message $\mathbf{M}$. A low revealing loss is desired to obtain a high accuracy in message recovery, which can be expressed as:

$$\mathcal{L}_r = MSE(\mathbf{M}, \mathbf{M}') = \frac{1}{c' \times h \times w} \|\mathbf{M} - \mathbf{M}'\|_2^2, \tag{7}$$

Discrimination loss $\mathcal{L}_d$. The discrimination loss is adopted to enhance the visual fidelity of the generated stego image and the anti-steganalysis ability of the network, where as described in the following cross-entropy loss function:

$$\mathcal{L}_d = \frac{1}{N} \sum_i - \left[y_i \cdot \log\left(p_i\right) + \left(1 - y_i\right) \cdot \log\left(1 - p_i\right) \right], \tag{8}$$

where y_i refers to the ground truth label (the cover as 0 and stego as 1). The p_i represents the classification probability of the discriminator.

Methods	Cover/Stego PSNR(dB)/SSIM	Secret/Recovery BER(%)
HiDDeN	36.61/0.922	24.72
Hayes	35.54/0.914	26.29
SteganoGAN	40.47/0.971	1.43
ABDH	42.38/0.988	1.95
HiNet	47.43/0.993	0.47
Ours	**48.41/0.996**	**0**

Table 1: Comparison between our proposed method and other steganography methods with 1 bpp embedding payload.

Mutual information loss $\mathcal{L}_m$. Mutual information reflects the correlation between two variables: the amount of information contained in one variable about the other variable. In Figure 2, the forward concealing process has the secret message **M** as the input and the lost information matrix **L** as the output. Ideally, when the mutual information between them is close to 0, it can be assumed that the distribution of **L** is independent of the distribution of the input **M**. In this case, we can discard **L** without rendering information loss for revealing. Thus, to preserve valid information in embedding, a mutual information loss $\mathcal{L}_m$ is proposed to constrain the direction of information flow. It can be defined as follows:

$$\mathcal{L}_m = H(\mathbf{M}) + H(\mathbf{L}) - H(\mathbf{M}, \mathbf{L}), \tag{9}$$

where $H(\mathbf{M})$ and $H(\mathbf{L})$ represent information entropy, and $H(\mathbf{M}, \mathbf{L})$ represents the joint entropy of the **M** and **L**. The information entropy and joint entropy are calculated as:

$$H(\mathbf{M}) = -\sum_{i=0}^{N-1} P_i \log P_i, \tag{10}$$

$$H(\mathbf{M}, \mathbf{L}) = -\sum_{i,j} P_{ML}(i,j) \cdot \log P_{ML}(i,j), \tag{11}$$

where N is the number of distinct pixel values in the matrix, P_i denotes the probability that the pixel with value i appears in the matrix. $P_{ML}(i,j)$ is the probability that the pixel at the same position has a value of i in matrix **M** and a value of j in matrix **L**.

Experimental Results

Datasets and implemental details. The dataset for training and testing is MSCOCO (Lin et al. 2014). We randomly select 5000 images as cover images for training and 1000 images for testing, respectively. The cover images used for training or testing above do not overlap and are all cropped at resolution 256 $\times$ 256 using the center-cropping strategy with MATLAB. The Adam optimizer (Kingma and Ba 2014) with standard parameters is adopted to optimize our network. The initial learning rate is set as 0.0001 and batch size as 2 to adapt our devices. The whole training process includes 120 epochs. The whole framework is implemented by PyTorch and executed on NVIDIA GeForce RTX 2080 Ti. At the end of the training, the model has already converged sufficiently. The weight factors λ_c, λ_r, λ_d and λ_m are set to 1.0, 15.0, 3.0, and 5.0, respectively.

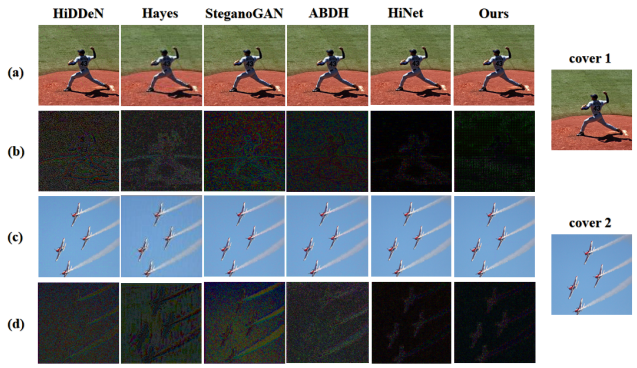

Figure 4: Visual comparison of our proposed method and other steganography methods: (a)/(c) the generated stego images, (b)/(d) the differences between the cover and stego images (magnified 10 times).

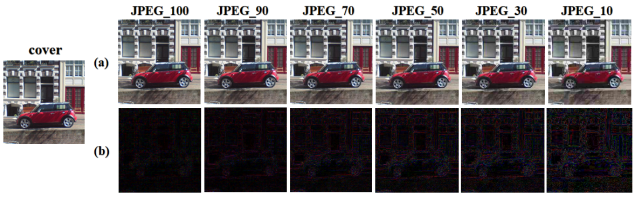

Figure 5: The visual effect of our proposed method under JPEG compression with different quality factors: (a) the generated stego images, (b) the difference between the cover and compressed stego images (magnified 5 times).

Evaluation metrics. We evaluate the performance from the following several aspects. The peak signal-to-noise ratio (PSNR) (Welstead 1999) and structural similarity index measure (SSIM) (Wang et al. 2004) are adopted as the image quality assessment. The bit error rate (BER) is utilized to measure the robustness, and a lower BER means the recovered message m' is closer to the original message m. Since the recovered message m' are floating numbers, we map the bits in m' greater than 0.5 to 1, and the rest to 0. As for steganographic security, the detection error indicates that the steganalysis network is unable to distinguish cover/stego images yielded by a specific steganography method.

Performances Analysis

Steganographic imperceptibility. From the perspective of quantitative results, Table 1 shows the significant superiority of our method compared with HiDDeN (Zhu et al. 2018), Hayes (Hayes and Danezis 2017), SteganoGAN (Zhang et al. 2019a), ABDH (Yu 2020), and HiNet (Jing et al. 2021). The framework of ABDH and HiNet are mainly designed to hide a color image in one carrier image. To make it available at capacity with 1 bpp, we finetuned the networks for hiding message bits. For cover/stego image pairs, the average PSNR and SSIM for our generated stego image are distinctly higher than the comparison methods. Notably, for secret/recovery pairs, our method achieves 100% recovery accuracy, which can be applied to the scenario to extract the

Metrics	Identity	JPEG_100	JPEG_90	JPEG_70	JPEG_50	JPEG_30	JPEG_10
BER(%)	0	0.19	0.31	0.66	0.92	1.47	2.92
PSNR(dB)/SSIM	48.41/0.996	45.54/0.992	44.13/0.985	41.39/0.977	37.08/0.967	36.27/0.955	34.58/0.948

Table 2: The average metric values of our method for cover/stego pairs and secret/recovery pairs with the payload of 1 bpp. The QFs chosen for training and testing are the same.

Methods	Payload (bit)	Identity	JPEG_90	JPEG_70	JPEG_50	JPEG_30	JPEG_10
HiDDeN	120	37.41	38.18	38.83	38.11	44.58	49.01
TSR	30	9.29	15.12	29.21	34.64	39.88	46.59
MBRS	256	**0**	**0.00063**	**0.0098**	1.35	8.58	33.17
Ours	**256×4**	0.0022	0.0072	0.081	**0.74**	**3.47**	**10.93**

Table 3: Comparison for the BER of secret message restoration. We chose JPEG compression with quality factor 50 in training, then tested with different quality factors. The PSNR value of the stego image is adjusted to about 33.5 dB for fair.

Methods	Detection Error(%)		
	XuNet	SRNet	WISERNet
HiDDeN	0.22	0	2.55
Hayes	0.04	0	1.13
SteganoGAN	2.35	0	1.67
ABDH	10.11	2.86	4.89
Ours	**19.27**	**5.14**	**9.37**

Table 4: Comparison on detection error using three steganalysis networks with the embedding payload of 1 bpp.

hidden message without error. Figure 4 compares the stego images of our method with the other four methods. The difference between cover and stego images shows the extent of damage to the cover image after embedding secret messages. We can see that the stego images generated by our proposed method differ slightly from the cover images and have no obvious text-copying artifacts or color distortion. In contrast, the method HiDDeN, Hayes, and SteganoGAN have apparent differences between cover and stego images, especially in complex texture and edge regions. The above comparison demonstrated the effectiveness of our proposed method in generating stego images with pleasing visual fidelity and high reconstruction quality.

Robustness against JPEG compression. We analyze the robustness of our proposed method with the embedding payload of 1 bpp. We train the model using the simulated JPEG compression with different quality factors (QFs) and testing with the same QFs. Figure 5 shows the generated stego images, and the difference between cover and stego images after JPEG compression. The stego images yielded by our method maintain high visible fidelity without block-blurring artifacts, and the difference between cover and stego images is nearly invisible. The quantitative results are displayed in Table 2. The QFs chosen for training and testing are the same, which means the testing is under a known channel. Our method can reach satisfactory reconstruction results under different QFs. In particular, our model achieves the BER of less than 3% even subjected to a high-intensity JPEG compression with a quality factor of 10, which indicates that our method has strong resistance against severe JPEG compression distortion. Due to such a low BER for extracting messages, error correction codes can be utilized to achieve 100% extraction accuracy in practical applications.

To better illustrate the robustness of our method, we compare the performance or our method with several advanced robust image hiding methods: HiDDeN (Zhu et al. 2018), TSR (Liu et al. 2019), and MBRS (Jia, Fang, and Zhang 2021). All the methods apply JPEG compression with $QF = 50$ to the training process, while different QFs vary from 10 to 90 for testing. That is, the testing is under an unknown channel. Note that there is a trade-off relationship between the message embedding and extracting. To make the comparative experiment fair, we maintain the PSNR of the stego images generated by the embedding process as 33.5 dB, then compare the BER of the recovered secret message. In addition, the embedding payload of our method has reduced to 256×4 bits to achieve better robust performance. As seen in Table 3, although the embedding payload of our method is 4 times that of MBRS, our method provides much better BER performance, especially when JPEG compression QF is decreased to 50 or even lower. As the intensity of JPEG compression increases, the superiority of our method over MBRS becomes more pronounced. The above experimental results demonstrated that our method provides outstanding robustness for aggressive JPEG compression as low as $QF = 10$, no matter the transmission channel is known or unknown. Rather than holding reconstruction capability at a specific compression factor, the robustness of our method is more stable and general in practical applications.

Security analysis. The steganography security is usually evaluated by measuring the detection error of a certain steganalyzer to distinguish the images generated by a steganographic algorithm. Current research (Yang et al. 2019) has demonstrated that the CNN-based steganalyzer is able to reduce the detection error dramatically compared with the conventional steganalyzer. To verify the security, we adopt three advanced steganalysis networks XuNet (Xu, Wu, and Shi 2016), SRNet (Boroumand, Chen, and Fridrich 2018), and WISERNet (Zeng et al. 2019) to assess the steganographic security for different steganography methods with

Discrete cosine transform	$\mathcal{L}_d$ loss	$\mathcal{L}_m$ loss + TWFM	Cover/Stego PSNR(dB)/SSIM	Secret/Recovered BER(%)	Detection Error(%)
✗	✗	✗	36.75/0.961	1.4	4.81
✓	✗	✗	43.29/0.975	0.23	10.36
✓	✓	✗	48.13/0.988	0.39	19.01
✓	✓	✓	48.41/0.996	0	19.27

Table 5: Ablation study on discrete cosine transform, discrimination loss $\mathcal{L}_d$, mutual information loss $\mathcal{L}_m$ and TWFM.

Figure 6: Visual results of ablation study on DCT. The differences are between the cover and compressed stego images with $QF = 50$.

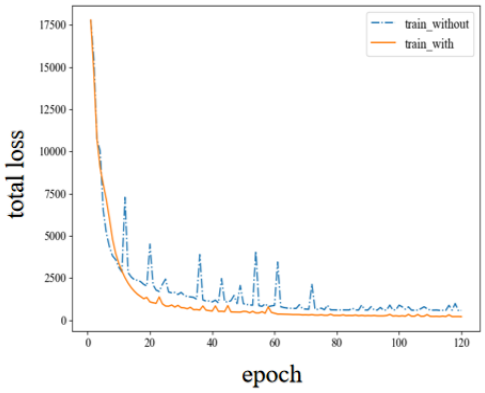

Figure 7: The total loss curve of ablation study on $\mathcal{L}_m$ loss and TWFM. Note that when the training process reaches to 120 epochs, two models have converged.

the embedding payload of 1 bpp. We yield 5,000 cover/stego image pairs for each method to re-train the steganalysis networks. Table 4 presents the detection error using three steganalysis networks for different methods. Our proposed method increases the detection error by 9.16%, 2.28%, and 4.48% compared to the second best result, respectively. The considerably higher detection error than other methods against three steganalysis networks demonstrates that our method has relatively better steganographic security.

Ablation Study

Effectiveness of discrete cosine transform. To conduct this ablation, we embedded secret messages into spatial domain and DCT domain respectively. In Figure 6, under the JPEG compression with $QF = 50$, the generated stego using spatial domain (without using DCT) shows visible block artifacts. The difference between the cover and stego image contains some apparent regions with color distortion. In contrast, the difference between the cover and stego image using DCT domain is nearly invisible, which means embedding into frequency domain coefficients will better adapt distortion and maintain the detailed content of the cover image. From the first and second rows in Table 5, the PSNR and SSIM with DCT transform increase by 6.54 dB and 0.014, respectively, and the detection error increases from 4.81% to 10.36%. This ablation further demonstrates that the DCT transform is inherently robust to JPEG compression and plays an indispensable role in the robust framework.

Effectiveness of $\mathcal{L}_d$. The discriminator provides progressive feedback to the generator, enabling it to enhance the imperceptibility and security of the generated stego images. As shown in the second and third rows in Table 5, with the $\mathcal{L}_d$ loss, the visual quality of the stego image can be improved by 4.84 dB in terms of PSNR, by 0.013 in terms of SSIM. In addition, the steganographic security of our method is enhanced, for which the detection error is increased by 8.65%.

Effectiveness of $\mathcal{L}_m$ and TWFM. The $\mathcal{L}_m$ can guide the information flow in the INN to flow more to the branch that yields the stego image, thereby preserving the valid information for the revealing process. The TWFM makes full use of this information to facilitate the recovery of secret messages. The training loss curve with and without $\mathcal{L}_m$ and TWFM has shown in Figure 7. It can be seen that these two designs can accelerate the convergence of the network in training and decline the total loss of the entire network to a certain extent. From the third and fourth rows in Table 5, the BER in secret/recovered pairs decreases, which also highlights the contribution of $\mathcal{L}_m$ loss and TWFM.

Conclusion

This paper presents a robust end-to-end image steganography system based on INN, which is resistant to lossy JPEG compression. Unlike other works with INN, we are the first to consider from the viewpoint of reducing the valid information loss in the forward process and utilizing it in the recovery process by designing mutual information loss and TWFM module. Experiments demonstrate that our method can yield high-fidelity stego images and achieve 0 error for message extraction with 1 bpp embedding capacity. In addition, our method has robustness against JPEG compression under the QFs of the transmission channel are known or unknown, which shows the generalization performance. Moreover, compared with other advanced steganography methods, our method achieves the best security performance under three deep learning-based steganalysis network detections.

Acknowledgements

This work was supported partly by the NSFC under Grant 62072484 and 62102462, partly by the Guangdong Basic and Applied Basic Research Foundation under Grant 2022A1515010108, and partly by the Research Project of Guangdong Polytechnic Normal University under Grant 2022SDKYA027.

References

Ahmadi, M.; Norouzi, A.; Karimi, N.; Samavi, S.; and Emami, A. 2020. ReDMark: Framework for residual diffusion watermarking based on deep networks. *Expert Systems with Applications*, 146: 113157.

Boroumand, M.; Chen, M.; and Fridrich, J. 2018. Deep residual network for steganalysis of digital images. *IEEE Transactions on Information Forensics and Security*, 14(5): 1181–1193.

Dinh, L.; Krueger, D.; and Bengio, Y. 2014. Nice: Nonlinear independent components estimation. *arXiv preprint arXiv:1410.8516*.

Filler, T.; Judas, J.; and Fridrich, J. 2011. Minimizing Additive Distortion in Steganography Using Syndrome-Trellis Codes. *IEEE Transactions on Information Forensics and Security*, 6(3): 920–935.

Hayes, J.; and Danezis, G. 2017. Generating steganographic images via adversarial training. *Advances in neural information processing systems*, 30.

Holub, V.; Fridrich, J.; and Denemark, T. 2014. Universal distortion function for steganography in an arbitrary domain. *EURASIP Journal on Information Security*, 2014(1): 1–13.

Jia, Z.; Fang, H.; and Zhang, W. 2021. Mbrs: Enhancing robustness of dnn-based watermarking by mini-batch of real and simulated jpeg compression. In *Proceedings of the 29th ACM International Conference on Multimedia*, 41–49.

Jing, J.; Deng, X.; Xu, M.; Wang, J.; and Guan, Z. 2021. HiNet: deep image hiding by invertible network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4733–4742.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P.; and Dhariwal, P. 2018. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31.

Li, B.; Wang, M.; Huang, J.; and Li, X. 2014. A new cost function for spatial image steganography. In *Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP)*, 4206–4210. IEEE.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision*, 740–755. Springer.

Liu, Y.; Guo, M.; Zhang, J.; Zhu, Y.; and Xie, X. 2019. A novel two-stage separable deep learning framework for practical blind watermarking. In *Proceedings of the 27th ACM International Conference on Multimedia*, 1509–1517.

Lu, S.-P.; Wang, R.; Zhong, T.; and Rosin, P. L. 2021. Large-capacity image steganography based on invertible neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10816–10825.

Pevnỳ, T.; Filler, T.; and Bas, P. 2010. Using high-dimensional image models to perform highly undetectable steganography. In *Proceedings of the International workshop on information hiding*, 161–177. Springer.

Shin, R.; and Song, D. 2017. Jpeg-resistant adversarial images. In *Proceedings of the NIPS 2017 Workshop on Machine Learning and Computer Security*, volume 1, 8.

Tan, J.; Liao, X.; Liu, J.; Cao, Y.; and Jiang, H. 2021. Channel attention image steganography with generative adversarial networks. *IEEE Transactions on Network Science and Engineering*, 9(2): 888–903.

Tancik, M.; Mildenhall, B.; and Ng, R. 2020. Stegastamp: Invisible hyperlinks in physical photographs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2117–2126.

Van der Ouderaa, T. F.; and Worrall, D. E. 2019. Reversible gans for memory-efficient image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4720–4728.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, X.; Yu, K.; Wu, S.; Gu, J.; Liu, Y.; Dong, C.; Qiao, Y.; and Loy, C. C. 2018. ESRGAN: Enhanced Super-Resolution Generative Adversarial Networks. In *Proceedings of the European Conference on Computer Vision*, 63–79. Springer.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4): 600–612.

Welstead, S. T. 1999. *Fractal and wavelet image compression techniques*, volume 40. Spie Press.

Wengrowski, E.; and Dana, K. 2019. Light field messaging with deep photographic steganography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1515–1524.

Xiao, M.; Zheng, S.; Liu, C.; Wang, Y.; He, D.; Ke, G.; Bian, J.; Lin, Z.; and Liu, T.-Y. 2020. Invertible image rescaling. In *Proceedings of the European Conference on Computer Vision*, 126–144. Springer.

Xu, G.; Wu, H.-Z.; and Shi, Y. Q. 2016. Ensemble of CNNs for steganalysis: An empirical study. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, 103–107.

Xu, Y.; Mou, C.; Hu, Y.; Xie, J.; and Zhang, J. 2022. Robust Invertible Image Steganography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7875–7884.

Yang, J.; Ruan, D.; Huang, J.; Kang, X.; and Shi, Y.-Q. 2019. An embedding cost learning framework using GAN. *IEEE Transactions on Information Forensics and Security*, 15: 839–851.

Yu, C. 2020. Attention based data hiding with generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 1120–1128.

Zeng, J.; Tan, S.; Liu, G.; Li, B.; and Huang, J. 2019. WISERNet: Wider separate-then-reunion network for steganalysis of color images. *IEEE Transactions on Information Forensics and Security*, 14(10): 2735–2748.

Zhang, C.; Karjauv, A.; Benz, P.; and Kweon, I. S. 2021. Towards Robust Deep Hiding Under Non-Differentiable Distortions for Practical Blind Watermarking. In *Proceedings of the 29th ACM International Conference on Multimedia*, 5158–5166.

Zhang, K. A.; Cuesta-Infante, A.; Xu, L.; and Veeramachaneni, K. 2019a. SteganoGAN: High capacity image steganography with GANs. *arXiv preprint arXiv:1901.03892*.

Zhang, Y.; Luo, X.; Guo, Y.; Qin, C.; and Liu, F. 2019b. Multiple robustness enhancements for image adaptive steganography in lossy channels. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(8): 2750–2764.

Zhu, J.; Kaplan, R.; Johnson, J.; and Fei-Fei, L. 2018. Hidden: Hiding data with deep networks. In *Proceedings of the European conference on computer vision (ECCV)*, 657–672.

Zhu, L.; Luo, X.; Yang, C.; Zhang, Y.; and Liu, F. 2021. Invariances of JPEG-quantized DCT coefficients and their application in robust image steganography. *Signal Processing*, 183: 108015.