Pattern Recognition 155 (2024) 110691

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

# Robust image hiding network with Frequency and Spatial Attentions

Xiaobin Zeng [a], Bingwen Feng [a,*], Zhihua Xia [a], Zecheng Peng [a], Tiewei Qin [a], Wei Lu [b]

[a] College of Cyber Security, Jinan University, Guangzhou 510632, China
[b] School of Computer Science and Engineering, Guangdong Province Key Laboratory of Information Security Technology, Ministry of Education Key Laboratory of Machine Intelligence and Advanced Computing, Sun Yat-sen University, Guangzhou 510006, China

ARTICLE INFO

*Keywords:*
Convert Image Communication (CIC)
Image hiding
Robustness
JPEG compression
Attention mask
Frequency

ABSTRACT

Convert Image Communication (CIC) is a promising technology to protect the privacy of images. Recently, the emergence of robust CIC resistant to JPEG compression has gained due to the widespread use of JPEG compression in image communication. This paper introduces a Robust image hiding network with Frequency and Spatial Attentions (RFSA) to implement robust CIC. RFSA can hide an image within another image with high robust. It incorporates multiple image attentions corresponding to imperceptibility, recovered image quality, and resistance to JPEG compression, which ensure that secret images are hidden within regions that cause little distortion and can well withstand JPEG compression. Additionally, two encoders, that is, a frequency encoder and a spatial encoder, are mixed to adaptively embed secret images across both frequency and spatial domains. Experimental results demonstrate that the proposed scheme not only maintains high image quality and capacity but also exhibits exceptional resistance to JPEG compression compared to other state-of-the-art image hiding methods. The average Peak Signal-to-Noise Ratio (PSNR) of the recovered image remains at 24.96 dB even under JPEG compression with a quality factor of 55.

## 1. Introduction

The privacy of images has garnered significant attention in recent years. Convert Image Communication (CIC) is a promising technology to protect the privacy of images. It enables the communication confidentiality of an image by discreetly concealing it within other multimedia data, usually another public image [1,2]. Only the authorized recipient can correctly recover it.

CIC can be achieved by traditional steganographic techniques, which effectively hide secret data within the given cover images with limited distortion [3–5]. Embedding algorithms such as pixel matching [3], Syndrome-Trellis Codes (STC) [4], Steganographic Polar Codes (SPC) [5] have been manually designed to this end. However, these algorithms impose strict limitations on the embedding capacity, typically up to 1 bit per pixel. With the rapid development of deep learning, an increasing number of learning-based CIC schemes have emerged. The pioneer work suggested by Baluja [6] employed a deep network model to achieve color image hiding. After that, Yu et al. [7] proposed an Attention Based Data Hiding framework (ABDH), which introduced attention modules to find inconspicuous areas for secret information hiding. Liu et al. [8] utilizes a joint compression autoencoder to address the image mapping in potential space for high-quality image concealment. Invertible networks have been frequently used recently to achieve high visual quality of stego and recovered images [9,10]. While the aforementioned methods exhibit impressive imperceptibility and capacity, they fail to account for the pervasive presence of JPEG compression noise in the transmission channel, leading to the fragility of CIC.

To enable the CIC in lossy compression channels, a scheme must be able to withstand the impact of JPEG compression. A series of learning-based data hiding schemes have been proposed for this purpose. Ahmadi et al. [11] devised a differentiable approximation algorithm for JPEG and trained it with selected attack types with assigned probabilities. Zhang et al. [12] developed a pseudo-differentiable method for JPEG compression to address the non-differentiability problem. Jia et al. [13] proposed a method to enhance robustness by selecting one noise layer from multiple noise environments for small batch images. However, they have a small capacity and can only hide secret information at the bit level. By integrating self-supervised learning and adversarial training, Zheng et al. [14] proposed a Composition-Aware Image Steganography (CAIS), which achieved the generation of steganographic images with increased capacity and enhanced robustness. Ying et al. [15] enhanced the JPEG compression simulator based on [13] and improved robustness through a progressive recovery strategy. Luo et al. [16] employed visual perception loss and two-stage

* Corresponding author.
*E-mail addresses:* dannis@stu2022.jnu.edu.cn (X. Zeng), bingwenfeng@jnu.edu.cn (B. Feng), xiazhihua@jnu.edu.cn (Z. Xia), pzc202134511007@stu2021.jnu.edu.cn (Z. Peng), qtiewei@stu2021.jnu.edu.cn (T. Qin), luwei3@mail.sysu.edu.cn (W. Lu).

https://doi.org/10.1016/j.patcog.2024.110691
Received 7 February 2024; Received in revised form 6 May 2024; Accepted 13 June 2024
Available online 20 June 2024

training to reduce noise interference in invertible networks. However, many of them are learned in the spatial domain, which may be sensitive to the low-frequency sub-band and some salient subbands, and thus limits their robustness [17,18].

On the other hand, traditional hiding methods frequently opt to embed data in more resilient areas to enhance robustness. Many of them embed message bits in the transform domain to balance imperceptibility and robustness [19–21]. Zhang et al. [19] created resilient regions by utilizing Discrete Cosine Transform (DCT). Sun et al. [20] also embed message bits in the DCT domain. Further, a CNN based method is suggested to select a resilient domain. Singh et al. [21] used a chaotic kbest gravitational search algorithm in DCT and Singular Value Decomposition (SVD) domain. Although these methods have limited capacity, the concept of embedding in the frequency domain is promising. We can impose the data hiding network to embed secret images in the frequency domain to enhance robustness against JPEG compression. Additionally, attention allows the network to focus on important features [22]. This concept is similar to embedding region selection. Various attentions have been developed. Zhu et al. [23] proposed grouping-aggregation and hybrid coding for channel attention to reduce information loss caused by convolutional dimensionality reduction. Obeso et al. [24] focused on external visual saliency attention as new complementary to help network optimization. They have been successfully applied it on image classification, object detection, etc. Therefore, We expect to design different attentions tailored to different CIC requirements.

This paper presents a robust CIC framework with frequency and spatial attentions (RFSA). RFSA integrates frequency and spatial data hiding, and incorporates multiple attentions to effectively extract robust features across diverse domains. This strategy ensures that the encoder can adaptively conceal the secret image. Specifically, it contains a frequency encoder module and a spatial encoder module. Additionally, an attention generation module is introduced to generate distinct attention masks for the encoders. The contributions of this paper can be summarized as follows.

- The proposed scheme extends the embedding domain to the frequency domain, leveraging the intrinsic distortion on DCT coefficients caused by JPEG compression to enhance the corresponding robustness.
- Multiple attention masks are generated based on JPEG compression influence and the human visual system. These masks guide the network to achieve a good balance among imperceptibility, recovered image quality, and resistance to JPEG compression.
- Extensive experimental results demonstrate that the stego images generated by our scheme present high visual quality and exhibit strong resilience to JPEG compression.

The rest of this paper is organized as follows. In Section 2, we review robust hiding methods at both the message and image levels. In Section 3, we describe the proposed method and loss function in detail. Section 4 shows the experimental setup and verifies the effectiveness of the proposed method. The conclusion is presented in Section 5.

## 2. Related work

In recent years, significant advancements have been achieved in learning-based robust covert communication. These previous methods can be categorized broadly into two groups depending on the payload of covert communication.

### 2.1. Robust convert communication of binary message

To achieve accurate message recovery in a lossy channel, several robust data hiding schemes for binary messages have been proposed. One effective method to improve the robustness is to insert a distortion layer during the training process. The distortion layer was introduced earlier by Zhu et al. [25]. Zhang et al. [12] further designed it to make the network compatible with non-micro distortions. Liu et al. [26] utilizes a multilayer perceptron for redundant mapping of the message and combines it with a distortion layer to fine-tune the encoder. Bui et al. [27] combined multiple noises and a single noise in a specific ratio to form a distortion layer. Furthermore, methods that leverage attention mechanisms and frequency coefficients to enhance robustness have progressively emerged. Tan et al. [28] and Fang et al. [29] underwent training on attentional mechanisms to enhance the resilience of images containing secrets. Lan et al. [30] employed invertible networks to embed confidential information into DCT coefficients, effectively enhancing both robustness and security.

In order to achieve good robustness, the aforementioned methods usually encode the hidden messages with error correction codes, and repeatedly embed them into the cover image. This inevitably limits the embedding capacity of the methods. Despite notable robustness improvements, this strategy falls short of CIC's image-level robust hiding requirements. Nevertheless, the introduction of a distortion layer and attention mechanism offers promising avenues for enhancing robustness. In our method, we employ the attention to concentrate embedding energy in more robust regions while incorporating a distortion layer to enhance robustness performance further.

### 2.2. Robust convert communication of image message

Baluja [6] pioneered the application of deep neural networks for concealing color images. However, robust CIC remains challenging due to the deep neural networks' vulnerability to intermediate distortion. To enhance robustness, ABDH [7] fine-tunes attacked images for adversarial training. Zhang et al. [31] proposed a Deep Adaptive Hiding Network (DAH), which employed deep frequency features and added a perturbation layer to improve robustness. Luo et al. [16] employed a two-stage training approach in invertible networks to mitigate noise interference during the backpropagation process. Nevertheless, the robustness of these methods is not particularly strong. Luo et al. [32] proposed a patch-level image hiding method that uses small patch blocks as embedding locations to enhance robustness. However, the patch image size is only 1/16 of the carrier image. Shang et al. [33] employed invertible networks to conceal a secret image within quantized DCT coefficients. Cao et al. [34] realized robust CIC using frequency domain channel attention. However, the hidden secret images in them are grayscale.

It can be found that the capacity of these schemes remains inadequate. To compensate for this limited capacity, these methods resort to reducing the sizes or color channels of the hidden images. Conversely, our proposed scheme strives to embed a full-size color secret image, necessitating a deeper investigation into the embedding potential of image hiding networks. To this end, we integrate representations from both the frequency and spatial domains, aiming to achieve an optimal balance between capacity, imperceptibility, and robustness.

Both [32] and [34] utilized the Universal Deep Hiding (UDH) [35] framework to attain satisfactory imperceptibility. Compared to Dependent Deep Hiding (DDH), UDH hides messages in an unrelated manner to the cover image, resulting in stego images with high quality [35]. Unfortunately, the original UDH is not robust. In this proposed scheme, we extend the UDH framework to maintain high capacity and imperceptibility while also ensuring good robustness.

## 3. Method

The proposed scheme delves into the frequency domain to embed secret images. Previous methods, as cited in [30,33,34], have demonstrated that concentrating secret embedding within the frequency domain enhances the visual quality of stego or recovered images. However, our approach goes beyond mere image quality, aiming to achieve

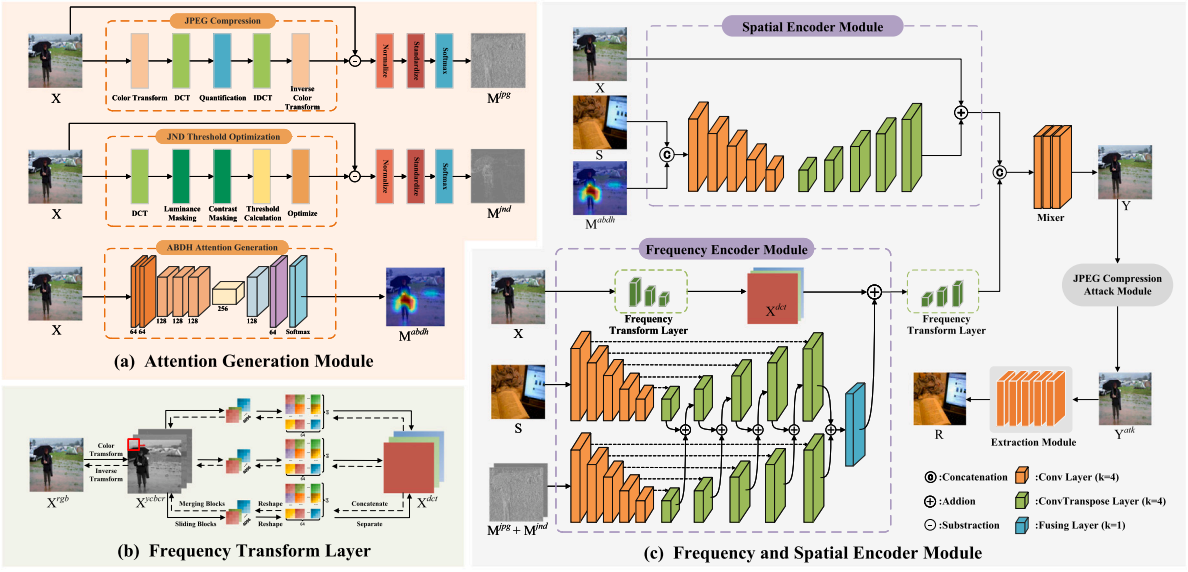



**Fig. 1.** The architecture of RFSA.

**Table 1**
Symbol definition.

| Symbols | Descriptions |
|---|---|
| $X$ | cover image in RGB model |
| $X^{ycbcr}$ | cover image in YCbCr model |
| $X^{dct}$ | DCT coefficients of $X^{ycbcr}$ |
| $M^{jpg}$ | JPEG attention mask |
| $M^{jnd}$ | JND attention mask |
| $M^{abdh}$ | ABDH attention mask |
| $S$ | secret image |
| $R$ | recovered image |
| $Y$ | stego image |
| $Y^{atk}$ | JPEG compressed image |

robustness against JPEG compression. Recognizing that the information loss in JPEG compression primarily stems from the quantization of DCT coefficients, we identify stable regions within these coefficients. It is expected that the inherent linkage between DCT domain representation and JPEG compression will facilitate a good balance between robustness and visual quality.

This section describes the proposed image hiding network for robust CIC, named RFSA. It is an end-to-end network that aims to hide a secret image within another image while achieving high imperceptibility, recovered image quality, and robustness against JPEG compression. For convenience, we summarize the symbols used in Table 1.

### 3.1. Overview

Fig. 1 displays the proposed RFSA framework, which consists of three components: the attention generation module, the frequency and spatial encoder module, and the extraction module. During the hiding process, a cover image $X$ and a secret image $S$ serve as the inputs. The attention generation module processes $X$ to generate multiple attention masks $M^{\cdot}$. Subsequently, $X$, $M^{\cdot}$, and $S$ are fed into the frequency and spatial encoder module to obtain the stego image $Y$. The extraction network can then retrieve the recovered image $R$ from $Y$ or JPEG compressed stego image $Y^{atk}$. The frequency transform layer in the frequency encoder module transforms $X$ to its frequency representation.

### 3.2. Attention generation module

Attention mechanisms can guide neural networks to focus on the most important features for a given task [22–24]. Inspired by this concept, the attention generation module generates attention masks to incorporate secret information into the feature maps of the cover with desired attributes such as robustness and imperceptibility. Since the hiding network should balance imperceptibility, recovered image quality, and resistance to JPEG compression, distinct attention masks should be generated to satisfy individual requirements. We explore three types of masks as shown in the top left of Fig. 1. They are the JPEG attention mask $M^{jpg}$, the JND attention mask $M^{jnd}$, and the ABDH attention mask $M^{abdh}$. The first two masks enhance robustness, while the third mask emphasizes imperceptibility.

#### 3.2.1. JPEG attention mask

This mask aids in enhancing the robustness against JPEG compression. It guides the encoder to embed information in the features that are less affected by JPEG compression. This mask is defined in the DCT domain, as JPEG compression is also performed in this domain.

We apply JPEG compression with quality factor $QF \in (50, 100)$ to the cover image $X$ to generate a JPEG compressed image $X^{atk}$. Subsequently, the corresponding residual image is collected by subtracting $X^{atk}$ from $X$. Residual image coefficients with larger values indicate weaker resistance to JPEG compression, and, thus, a higher risk of secret information deletion. Therefore, we should assign a small value to these sensitive areas. During the training process, the QF is randomly selected within a given range, and set to be consistent with the JPEG compression attack module.

To address the dimensional inconsistency of the difference values and guarantee the rationality of the attention features, we normalize the residual image collection through standardization and normalization. Subsequently, a softmax function is utilized to transform the normalized residuals into a JPEG attention mask $M^{jpg} \in [0, 1]$. This process can be formed as:

$$M^{jpg} = \sigma\Big(S(\mathcal{N}\left(X - X^{atk}\right))\Big) \tag{1}$$

where $\sigma$ is the sigmoid function multiplied by a constant factor, and $S$ and $\mathcal{N}$ represent the standardization and normalization processes, respectively.

#### 3.2.2. JND attention mask

This mask further balances imperceptibility and embedding strength. Just Noticeable Distortion (JND) defines the maximum level of image distortion that the human eye cannot perceive, reflecting the tolerance of the human visual system to changes in images. It has been widely used in traditional robust watermarking [36]. It is expected that the designed JND attention mask $M^{jnd}$ can transfer this characteristic to the hiding network.

The mask is also defined in the DCT domain for the same reason as the JPEG attention mask. To maintain consistency in dimensions and sizes between the mask and DCT coefficients, we use the 4 × 4 sized JND model introduced by Watson [36] as the backbone. Precisely, visibility threshold, luminance, and contrast masking factor are calculated using cover image $X^{dct}$ to obtain the JND coefficients. Subsequently, through residual calculations, standardization, normalization, and other computations, JND attention mask $M^{jnd}$ is generated. This process can be formed as:

$$X^{jnd} = t \cdot f^{lum} \cdot f^{con} \cdot X^{dct} \tag{2}$$

$$M^{jnd} = \sigma\Big(S(\mathcal{N}\big(X - X^{jnd}\big))\Big) \tag{3}$$

Where $f^{lum}$ is the modulation factor from the luminance masking, $f^{con}$ is another modulation factor from the contrast masking, and $t$ refers to the threshold. $f^{lum}$, $f^{con}$ and $t$ are calculated as follow:

$$t_{i,j} = \left(\sum_k \left|\frac{c_{i,j} - R[c_{i,j,k}/q_{i,j}] \cdot q_{i,j}}{f^{con}_{i,j,k}}\right|^4\right)^{1/4} \tag{4}$$

$$f^{lum}_{i,j,k} = \frac{1}{2} \cdot q_{ij} \cdot (c_{0,0,k}/\bar{c}_{0,0})^{0.649} \tag{5}$$

$$f^{con}_{i,j,k} = \frac{1}{2} Max\left[f^{lum}_{i,j,k}, \left|c_{i,j,k}\right|^{w_{i,j}} \cdot {f^{lum}_{i,j,k}}^{1-w_{i,j}}\right] \tag{6}$$

$$w_{i,j} = \begin{cases} 0 & i = 0,\ j = 0 \\ 0.7 & otherwise \end{cases} \tag{7}$$

where $R$ denotes the rounding, $c_{i,j,k}$ is the component of the $k$th DCT block, $\bar{c}_{0,0}$ is the DC coefficient corresponding and $q_{i,j}$ is the quantization matrix.

#### 3.2.3. ABDH attention mask

This mask improves imperceptibility by using visual attention that simulates the human attention mechanism. We employ the ABDH attention mask $M^{abdh}$ suggested by [7]. It implicitly focuses on specific areas of the image, aligning the feature weights of the intermediate network with human attention emphasis. To generate the ABDH attention mask $M^{abdh}$, we utilize the first two layers of ResNet50 [37] as the backbone network and input the cover image $X$ as the input.

### 3.3. Frequency and spatial encoder module

The structure of this module is shown on the right side of Fig. 1, where the orange and green blocks represent the convolutional and transpositional convolutional layers with the kernel size of 4 × 4, respectively, and the fusion layer is a convolutional layer with the kernel size of 1 × 1. The frequency and spatial encoder module is composed of three sub-networks: frequency encoder, spatial encoder, and mixer. Both the frequency encoder and the spatial encoder have a similar structure with UDH. The frequency encoder hides the secret image $S$ within the DCT coefficients, while the spatial encoder hides $S$ within the spatial pixels. The mixer is composed of three convolution layers with the kernel size of 4 × 4. It adaptively integrates frequency and spatial embedding results.

The convolutional structure is hierarchical and can extract features at various levels, perceiving detailed information in shallow layers and structural information in deeper layers. Our objective is to dynamically extract the essential frequency information from different depths of the network. As a result, the frequency encoder employs a parallel U-Net [38] structure as the network backbone. This approach dynamically fuses essential information from $S$ and attention masks $M^{\cdot}$, leveraging the design of Gradual Depth Extraction from [31].

In the frequency encoder, secret image $S$, together with generated $M^{jpg}$ and $M^{jnd}$, are used as inputs and fed into the network. During the encoding stage, the frequency feature information of the mask and secret image is extracted from different depths through the parallel network structure. The feature information is cross-shared during the decoding stage, progressively fusing details and structural features from different scales of the parallel network.

In the spatial encoder, our focus is on capturing the spatial position information of the secret image $S$. To avoid multiple convolution computations on the cover image $X$ and prevent the loss of high-frequency details, we draw inspiration from [35] and maintain the utilization of the U-Net as the network backbone. The input image and $M^{abdh}$ are directly combined and fed into the spatial encoder to extract and fuse the required information. Multiple convolution and linear modules are applied to the final output in the mixer.

### 3.4. Extraction module

This module recovers the secret image $R$ from stego image $Y$ or JPEG compressed stego image $Y^{atk}$. It simply superposes 5 Conv-InstanceNorm-ReLU blocks, each utilizing 4 × 4 convolution operations and kernel size.

### 3.5. Frequency transform layer

The frequency transform layer in the frequency encoder transforms the cover image $X$ to the frequency representation, producing DCT coefficient tensor of the same size as that of $X$. Specifically, $X$ is first transformed to the YCbCr color space $X^{ycbcr} \in \mathbb{R}^{(3,h,w)}$, where $h$ and $w$ are the height and width of the image. We then perform block-DCT using a 4 × 4 sliding window, obtaining the DCT frequency coefficients $X^{dct} \in \mathbb{R}^{(3,\frac{h}{4},\frac{w}{4},4,4)}$. Empirically we find that the convolution receptive field of the subsequent encoder should stride over different DCT channels so that it can map the secret information in the frequency domain well. As a result, we reshape it to form $X^{dct} \in \mathbb{R}^{(3,\frac{H}{4}\times 4,\frac{W}{4}\times 4)}$, as shown in the left-lower of Fig. 1.

### 3.6. Loss functions

The overall objective loss function $\mathcal{L}_{total}$ consists of four components: the hiding loss $\mathcal{L}_h$ and the recovery loss $\mathcal{L}_r$ to ensure the hiding performance, the perceptual loss $\mathcal{L}_p$ and the frequency loss $\mathcal{L}_f$ to enhance stego image quality.

*Hiding loss.* The hiding process should produce a stego image that is indistinguishable from the cover image. Toward this goal, the hiding loss is defined as:

$$\mathcal{L}_h(\theta) = \sum_{n=1}^{N} \ell_h\big(X, Y\big) \tag{8}$$

where $\theta$ represents the network parameters, $N$ is the number of training samples, and $\ell_h$ is the difference between the cover image $X$ and the stego image $Y$, which can be $L1$ or $L2$ norm.

*Recovery loss.* The extraction process should recover the secret image from the generated stego image. Consequently, the recovery loss is defined as:

$$\mathcal{L}_r(\theta) = \sum_{n=1}^{N} \ell_r\big(S, R\big) \tag{9}$$

Similar to $\ell_h$, $\ell_r$ measures the difference between the recovered image $R$ and the ground-truth secret image $S$.

*Perceptual loss.* We introduced perceptual loss in the hiding and recovery processes to restrict the visual perception gap between the cover/stego and secret/recovered image pairs. It utilizes the $L2$ distance of intermediate features of a pre-trained VGG16-Network [39]. The perceptual loss is written as:

$$\mathcal{L}_p(\theta) = \left|\left|\Phi_3(X) - \Phi_3(Y)\right|\right|_2^2 + \left|\left|\Phi_3(S) - \Phi_3(R)\right|\right|_2^2 \tag{10}$$

Where $\Phi_i$ represents the $i$th layer of the pre-trained VGG16 network.

*Frequency loss.* We extend the perceptual loss to the frequency domain by asking for the similarity between DCT coefficients of cover and stego images. The frequency loss is calculated by:

$$\mathcal{L}_f(\theta) = \sum_{n=1}^{N} \ell_f\left(X^{dct}, Y^{dct}\right) \tag{11}$$

where $\ell_f$ measures the difference in DCT domain.

*Total loss function.* The final loss function of the entire network is formed as

$$\mathcal{L}_{total} = \lambda_1\mathcal{L}_h + \lambda_2\mathcal{L}_c + \lambda_3\mathcal{L}_p + \lambda_3\mathcal{L}_f \tag{12}$$

Where $\lambda_1, \lambda_2, \lambda_3$, and $\lambda_4$ are the weight factors to balance the loss term.

## 4. Experimental results

### 4.1. Experimental setting

*Implementation details.* We train and test our network on the COCO dataset, randomly selecting 5000 pairs of cover/secret images as the training set and 1000 pairs of cover/secret images as the test set. The proposed RFSA is implemented using PyTorch and accelerated using Nvidia RTX 3090 GPU. We employ the Adam optimizer with $\beta = (0.5, 0.999)$ and a learning rate starting from 0.0001. The total number of training epochs is set to 200. The input images are uniformly cropped to a size of $256 \times 256$, and the batch size is set to 5. The weight factors used in Eq. (12) are experimentally set as $\lambda_1 = \lambda_2 = 1$, $\lambda_3 = 0.75$, and $\lambda_4 = 1.25$.

*Benchmarks.* To validate the effectiveness of our scheme, we compared it with existing CIC schemes that hide color images into another image of equal size, including ABDH [7], UDH [35], CAIS [14], and DAH [31]. For a fair comparison, we retrained and evaluated these schemes using the same dataset as ours, and the parameters followed the default settings mentioned in those references.

*Evaluation metrics.* Four metrics are used to measure the image quality for cover/stego and secret/recovered image pairs: Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM), which are commonly used objective evaluation metrics, as well as Average Pixel Difference (APD) calculated using $L1$ norm, and Learned Perceptual Image Patch Similarity (LPIPS) [40]. Higher values of PSNR/SSIM indicate better image quality, while lower values of APD/LPIPS indicate better image quality. In the following experimental results, -C represents the metrics on the cover/stego image pairs (e.g., APD-C, PSNR-C), and -S represents the metrics on the secret/recovered image pairs (e.g., APD-S, PSNR-S).

### 4.2. Performance without distortion

We test the proposed scheme on the clean cover images in the test set. Fig. 2 provides a visual comparison of our RFSA with four other schemes: ABDH, UDH, CAIS, and DAH. We zoom in on the stego and recovered images to observe their intricacies and compare them with the others. It can be observed that, even in regions with intricate textures, the stego and recovered images generated by our scheme exhibit remarkable visual similarity to the originals, preserving their fine details. Additionally, we examine the embedding distortion by displaying a magnified version of the cover image's difference with the stego image. As shown in Fig. 2, the contrast reveals that the difference obtained by the compared schemes reflects the details of cover images, indicating a detail loss on their stego images. In contrast, the difference obtained by our scheme contains little visible information. The magnified difference between secret and recovered images also suggests that the proposed scheme can provide better recovered image quality. Table 2 further quantitatively compares our RFSA with the other schemes using image quality metrics. The results demonstrate that our scheme outperforms the others in terms of hiding and recovering performance. We achieved improvements of 3.23 dB/4.24 dB in PSNR and 0.002/0.002 in SSIM compared to the second-best results. Similar enhancements are also observed in APD and LPIPS.

### 4.3. Performance under distortion

JPEG compression is one of the primary causes of information loss during image communication. In our experiments, we evaluate the robustness of RFSA against JPEG compression with various QFs. Fig. 3 depicts the visual effects of the recovered images obtained by our scheme under JPEG compression with different QFs. Table 3 reports the performance averaged over the test images. It can be observed that, even under the JPEG compression with $QF = 55$, our scheme still maintains high-quality recovered images. It can achieve an averaged PSNR-S of 24.96 dB and SSIM-S of 0.812 in this case.

We further conduct a comparative analysis of the embedding and extraction performance between RFSA and compared schemes under JPEG compression. It should be noted that the compared schemes are not originally equipped with robust training against JPEG compression. In view of this, we first remove the JPEG compression layer from our scheme for a fair comparison. Fig. 4 visually compares our scheme with others under JPEG compression with $QF = 85$. The results indicate that these compared schemes suffer from significant information loss of secret images. On the contrary, our scheme provides superior recovered image quality. Table 4 presents a quantitative comparison of these schemes, confirming the robustness of our scheme to JPEG compression.

Then we introduce JPEG compression distortion layers into all these schemes. This enabled these schemes to be trained robustly against JPEG compression. With the robust training, we obtained fine-tuned ABDH+, UDH+, CAIS+, DAH+, and our scheme RFSA+. Fig. 5 presents the visual comparison of RFSA+ with the other fine-tuned schemes under JPEG compression with $QF = 85$. It can be observed that the compared ones still exhibit visible color distortions, blurring, and blocky artifacts on the recovered images. Contrarily, our scheme is more effective in reconstructing image details and smooth regions, resulting in images that closely resemble the original secret image. We further conduct a comparison of the quality of our scheme and other fine-tuned schemes under JPEG compression with different QFs, as listed in Table 5. It suggests that our scheme outperforms the others under all JPEG compression distortions. As a result, the network structure in our scheme can provide strong robustness to JPEG compression.

However, due to the lack of robust design against other channel disturbances, RFSA is not effective in the face of other distortions. For example, when subjected to Gaussian noise with a standard deviation of $\sigma = 2$, the PSNR value of the recovered image reaches only 21.57 dB, while the SSIM metric drops to 0.816.

### 4.4. Ablation study

*Effect of modules and losses.* Here, we primarily delve into the efficacy of the proposed spatial encoder module, frequency encoder module, frequency loss, and perceptual loss. We juxtapose them with the baseline


**Fig. 2.** Visual comparison of RFSA without JPEG compression distortion with ABDH [7], UDH [35], CAIS [14], and DAH [31]. To account for the difference between the original image and the generated image, the pixel-level residuals are magnified by a factor of 10.

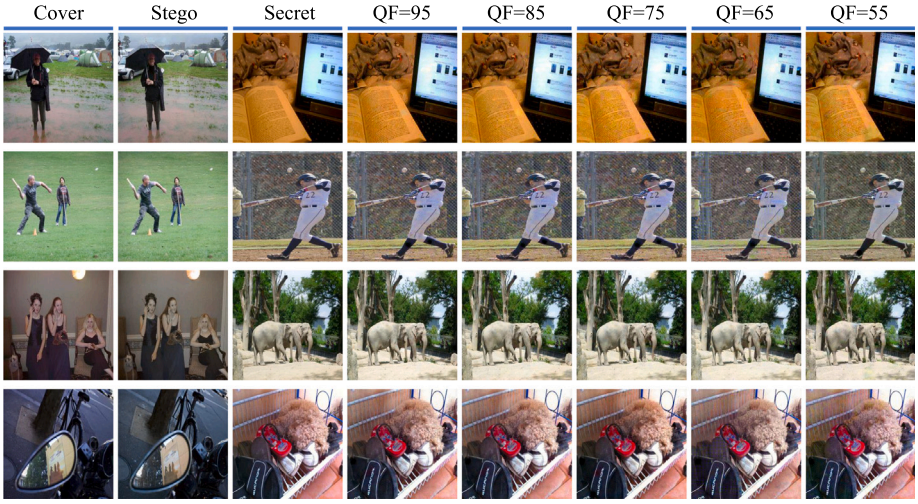



**Fig. 3.** The visual effect of recovered images by our scheme under JPEG compression with different QFs.

**Table 2**
Comparison with other schemes without JPEG compression distortion. The best and second-best results are marked in bold and underlined, respectively.

| Method | APD-C↓ | PSNR-C↑ | SSIM-C↑ | LPIPS-C↓ | APD-S↓ | PSNR-S↑ | SSIM-S↑ | LPIPS-S↓ |
|---|---|---|---|---|---|---|---|---|
| ABDH [7] | 3.71 | 34.56 | 0.949 | 0.0108 | 4.82 | 31.84 | 0.933 | 0.0267 |
| UDH [35] | 2.35 | 39.13 | 0.985 | **0.0001** | 3.56 | 35.0 | 0.976 | 0.0136 |
| CAIS [14] | 2.61 | 38.12 | 0.981 | 0.0005 | 3.85 | 34.20 | 0.972 | 0.0055 |
| DAH [31] | 2.05 | 39.84 | 0.988 | 0.0009 | 2.71 | 37.19 | 0.987 | 0.0024 |
| **RFSA (Ours)** | **1.06** | **43.07** | **0.996** | **0.0001** | **1.73** | **41.43** | **0.989** | **0.0016** |

**Table 3**
Embedding and extraction performance under JPEG with different QFs.

| QFs | APD-C↓ | PSNR-C↑ | SSIM-C↑ | LPIPS-C↓ | APD-S↓ | PSNR-S↑ | SSIM-S↑ | LPIPS-S↓ |
|---|---|---|---|---|---|---|---|---|
| QF=95 | 3.22 | 35.22 | 0.985 | 0.0308 | 7.79 | 27.29 | 0.884 | 0.1469 |
| QF=85 | 3.59 | 34.47 | 0.981 | 0.0351 | 8.12 | 26.89 | 0.871 | 0.1506 |
| QF=75 | 3.77 | 34.19 | 0.972 | 0.0389 | 8.66 | 26.27 | 0.857 | 0.1655 |
| QF=65 | 3.85 | 33.50 | 0.957 | 0.0407 | 8.83 | 25.43 | 0.836 | 0.1828 |
| QF=55 | 4.06 | 32.86 | 0.943 | 0.0465 | 9.04 | 24.96 | 0.812 | 0.2033 |

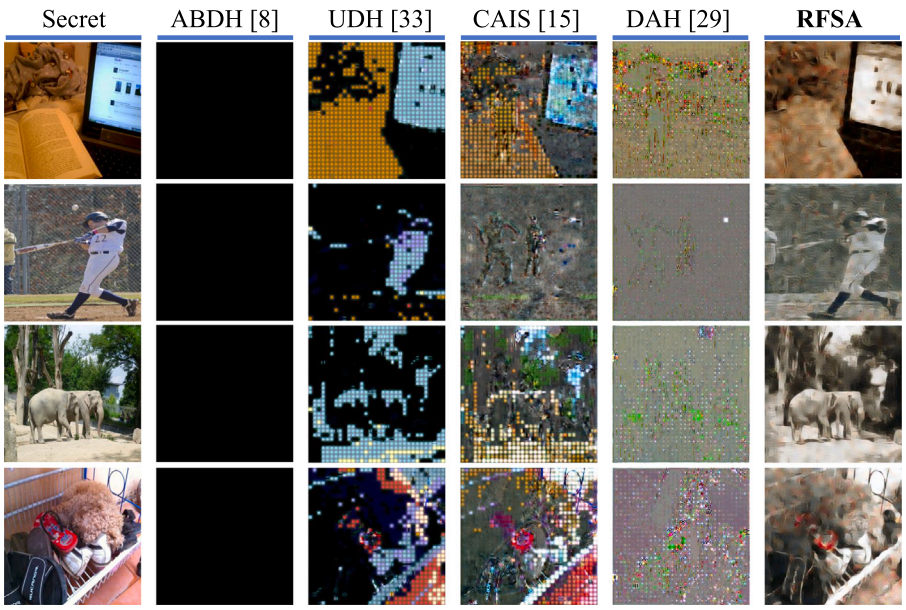



**Fig. 4.** Visual comparison of secret images recovered by our scheme and other schemes under JPEG compression with $QF = 85$.

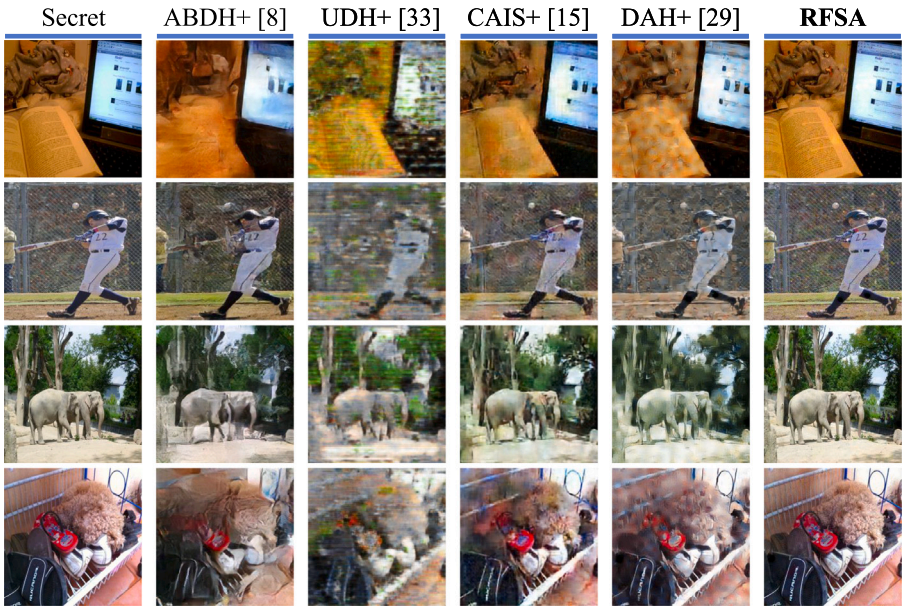
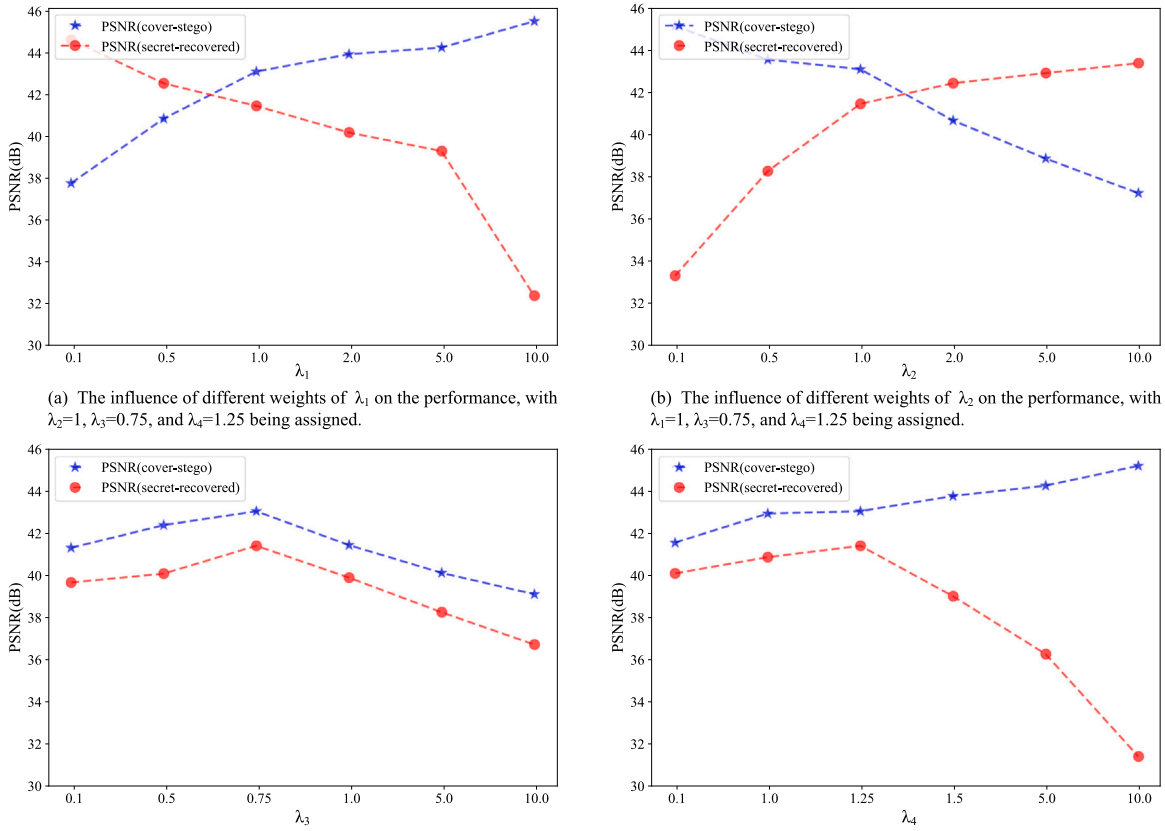



**Fig. 5.** Visual comparison of secret images recovered by our scheme and other fine-tuned schemes under JPEG compression with $QF = 85$.

**Table 4**
Comparison with other schemes under different JPEG compression conditions. Value1/value2 represents the scores calculated from cover/stego and secret/recovered image pairs using the corresponding index.

| QFs | ABDH [7] | | UDH [35] | | CAIS [14] | | DAH [31] | | RFSA(Ours) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR-C/-S↑ | SSIM-C/-S↑ | PSNR-C/-S↑ | SSIM-C/-S↑ | PSNR-C/-S↑ | SSIM-C/-S↑ | PSNR-C/-S↑ | SSIM-C/-S↑ | PSNR-C/-S↑ | SSIM-C/-S↑ |
| QF=85 | 34.19/7.42 | 0.949/0.552 | 38.47/9.35 | 0.984/0.486 | 37.60/13.28 | 0.980/0.625 | 39.72/10.23 | 0.988/0.563 | **42.55/17.56** | **0.994/0.758** |
| QF=75 | 34.19/5.89 | 0.949/0.529 | 38.47/8.04 | 0.984/0.460 | 37.60/11.24 | 0.980/0.581 | 39.72/9.63 | 0.988/0.552 | **42.55/15.66** | **0.994/0.704** |
| QF=65 | 34.19/5.82 | 0.949/0.526 | 38.47/5.78 | 0.984/0.424 | 37.60/8.76 | 0.980/0.552 | 39.72/8.18 | 0.988/0.531 | **42.55/13.64** | **0.994/0.639** |
| Average | 34.19/6.38 | 0.949/0.536 | 38.47/7.72 | 0.984/0.456 | 37.60/11.09 | 0.980/0.586 | 39.72/9.35 | 0.988/0.549 | **42.55/15.62** | **0.994/0.700** |

network. The comparative results are summarized in Table 6, where Spatial means the spatial encoder module and Frequency means the frequency encoder module, and $\mathcal{L}_f$ and $\mathcal{L}_p$ represent the frequency loss and the perceptual Loss, respectively. It can be observed that the quality of the stego and recovered images produced by the frequency encoder module is demonstrably superior to that produced by the spatial encoder module when used independently. The best results are achieved by combining the frequency domain and spatial domain. Further, the best hiding and recovery performance is achieved when the frequency and spatial encoder module is utilized alongside frequency and perceptual losses as the training constraints. As a result, they play a pivotal role in determining the final outcomes.




(a) The influence of different weights of $\lambda_1$ on the performance, with $\lambda_2$=1, $\lambda_3$=0.75, and $\lambda_4$=1.25 being assigned.

(b) The influence of different weights of $\lambda_2$ on the performance, with $\lambda_1$=1, $\lambda_3$=0.75, and $\lambda_4$=1.25 being assigned.

(c) The influence of different weights of $\lambda_3$ on the performance, with $\lambda_1$=1, $\lambda_2$=1, and $\lambda_4$=1.25 being assigned.

(d) The influence of different weights of $\lambda_4$ on the performance, with $\lambda_1$=1, $\lambda_2$=1, and $\lambda_3$=0.75 being assigned.

**Fig. 6.** Comparison of image quality ablation under different weight factors.

**Table 5**
Comparison with other fine-tuned schemes under different JPEG compression conditions.

| QFs | ABDH+ [7] | | UDH+ [35] | | CAIS+ [14] | | DAH+ [31] | | RFSA+ (Ours) | |
|---|---|---|---|---|---|---|---|---|---|---|
| | PSNR-C/-S↑ | SSIM-C/-S↑ | PSNR-C/-S↑ | SSIM-C/-S↑ | PSNR-C/-S↑ | SSIM-C/-S↑ | PSNR-C/-S↑ | SSIM-C/-S↑ | PSNR-C/-S↑ | SSIM-C/-S↑ |
| QF=85 | 33.41/19.46 | 0.918/0.668 | 29.75/19.29 | 0.853/0.747 | 29.28/21.39 | 0.907/0.826 | 31.92/23.75 | 0.952/0.776 | **34.47/26.89** | **0.981/0.871** |
| QF=75 | 32.45/18.55 | 0.909/0.653 | 29.47/18.76 | 0.838/0.722 | 28.64/20.86 | 0.881/0.817 | 31.26/23.12 | 0.945/0.768 | **34.19/26.27** | **0.972/0.857** |
| QF=65 | 31.86/16.61 | 0.900/0.639 | 28.93/17.57 | 0.829/0.670 | 27.73/19.43 | 0.870/0.788 | 30.93/22.56 | 0.938/0.738 | **33.50/25.43** | **0.957/0.836** |
| Average | 32.57/18.20 | 0.909/0.653 | 29.38/18.54 | 0.840/0.713 | 28.55/20.56 | 0.886/0.810 | 31.37/23.14 | 0.945/0.760 | **34.05/26.19** | **0.970/0.854** |

**Table 6**
Ablation study for different modules.

| | Spatial | Frequency | $\mathcal{L}_f$ | $\mathcal{L}_p$ | APD-C/-S↓ | PSNR-C/-S↑ | SSIM-C/-S↑ | LPIPS-C/-S↓ |
|---|---|---|---|---|---|---|---|---|
| Clean | ✗ | ✓ | ✗ | ✗ | 2.14/3.94 | 39.47/38.13 | 0.985/0.971 | 0.0027/0.0038 |
| | ✓ | ✗ | ✗ | ✗ | 2.46/3.14 | 38.18/36.26 | 0.979/0.950 | 0.0042/0.0066 |
| | ✓ | ✓ | ✗ | ✗ | 1.95/2.90 | 40.72/39.31 | 0.990/0.974 | 0.0017/0.0024 |
| | ✓ | ✓ | ✓ | ✗ | 1.58/2.32 | 41.63/40.20 | 0.993/0.981 | 0.0011/0.0020 |
| | ✓ | ✓ | ✓ | ✓ | **1.06/1.73** | **43.07/41.43** | **0.996/0.989** | **0.0001/0.0016** |
| QF=85 | ✗ | ✓ | ✗ | ✗ | 4.72/16.07 | 30.32/20.45 | 0.923/0.784 | 0.1065/0.2982 |
| | ✓ | ✗ | ✗ | ✗ | 5.44/23.73 | 28.93/17.75 | 0.881/0.729 | 0.1616/0.3247 |
| | ✓ | ✓ | ✗ | ✗ | 3.83/11.47 | 32.18/24.68 | 0.952/0.808 | 0.0729/0.2361 |
| | ✓ | ✓ | ✓ | ✗ | 3.66/9.84 | 33.37/25.22 | 0.975/0.826 | 0.0462/0.1798 |
| | ✓ | ✓ | ✓ | ✓ | **3.59/8.12** | **34.47/26.89** | **0.985/0.884** | **0.0351/0.1506** |

**Table 7**
Ablation study for the effect of different attention masks.

| Configuration | Clean | | QF=85 | | QF=75 | | QF=65 | |
|---|---|---|---|---|---|---|---|---|
| | PSNR-C/-S↑ | SSIM-C/-S↑ | PSNR-C/-S↑ | SSIM-C/-S↑ | PSNR-C/-S↑ | SSIM-C/-S↑ | PSNR-C/-S↑ | SSIM-C/-S↑ |
| Without $M$ | 40.91/37.47 | 0.982/0.974 | 31.89/22.81 | 0.938/0.774 | 31.28/22.19 | 0.936/0.757 | 30.65/21.87 | 0.914/0.749 |
| Only $M^{jpg}$ | 42.02/40.04 | 0.988/0.980 | 32.49/24.51 | 0.956/0.867 | 32.27/24.26 | 0.948/0.825 | 32.04/23.84 | 0.943/0.791 |
| Only $M^{jnd}$ | 41.78/39.02 | 0.986/0.981 | 32.05/23.95 | 0.943/0.795 | 31.77/23.56 | 0.941/0.786 | 31.12/23.02 | 0.929/0.763 |
| Only $M^{abdh}$ | 42.53/40.90 | 0.990/0.984 | 33.38/24.23 | 0.970/0.810 | 33.01/23.98 | 0.965/0.785 | 32.65/23.53 | 0.956/0.770 |
| Fully $M$ | **43.07/41.43** | **0.996/0.989** | **34.47/26.89** | **0.981/0.871** | **34.19/26.27** | **0.972/0.857** | **33.50/25.43** | **0.957/0.812** |

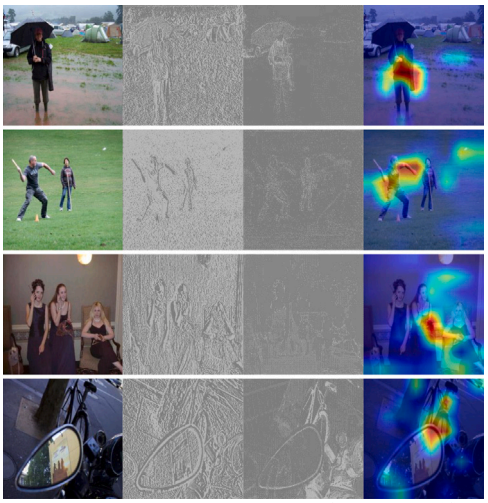



**Fig. 7.** Examples of the attention masks. The quality factor for JPEG is set to $QF = 85$.

*Effect of weight factors.* We investigate the impact of various weight factors in Eq. (12) on image quality. $\lambda_1$, $\lambda_2$, $\lambda_3$, and $\lambda_4$ correspond to the loss weights of the hiding process, recovery process, visual perception, and frequency coefficient similarity, respectively. Fig. 6(a) demonstrates that increasing $\lambda_1$ improves the visual quality of the stego image, but significantly decreases the quality of the recovered image. Conversely, Fig. 6(b) illustrates the opposite effect of $\lambda_2$. Moreover, Fig. 6(c) and (d) reveal that excessively large or small values of $\lambda_3$ and $\lambda_4$ have a substantial impact on image quality. The optimal balance of these weights relies on the specific requirements of the application. This paper takes into account the image quality of both stego and recovered images. Taking $\lambda_1$ as an example, search progressively within the range $[0.1, 10]$ while keeping the other weights constant. Therefore, we choose $\lambda_1 = 1$ to obtain the best trade-off. Finally, we choose $\lambda_1 = 1$, $\lambda_2 = 1$, $\lambda_3 = 0.75$, and $\lambda_4 = 1.25$.

*Effect of image attention masks.* Ablation experiments are conducted to investigate the impact of attention selection in the attention feature extraction module. Fig. 7 illustrates some examples of attention masks for a cover image, including the cover image itself, as well as the JPEG, JND, and ABDH attention masks, from left to right. We compare the effect of these attentions by varying the input of the frequency and spatial encoder module. As shown in Table 7, inputting JPEG and JND attention masks can effectively improve the quality of recovered images, and enhance robustness against JPEG compression. On the other hand, ABDH attention primarily help ensure the quality of stego images. The best hiding and recovery performance is achieved when all attentions are utilized simultaneously.

## 5. Conclusion

Convert transmitting a secret image can be a challenging task, especially when dealing with a carrier image of the same size and the presence of interference of JPEG compression in the channel. In this work, we propose an end-to-end image hiding network called RFSA for robust covert image communication over the Internet. This scheme incorporates an attention generation module to balance imperceptibility, recovered image quality, and resistance to JPEG compression. Especially, it outputs three attention masks: the JPEG attention mask that enhances robustness against compression, the JND attention mask that balances imperceptibility and embedding strength, and the ABDH attention mask that improves imperceptibility. Furthermore, both frequency and spatial domains are considered in the encoder module. By using multiple attention masks to guide the embedding of the secret image in both frequency and spatial encoders, the proposed technique can select regions that reduce distortion. This provides effective protection of the structural and perceptual quality of the image and better resists JPEG compression while ensuring the imperceptibility of the proposed method. In comparison to existing state-of-the-art methods, our method exhibits superior hiding and recovery performance without channel distortion. The PSNR values of the stego and recovered image can achieve 43.07 dB and 41.43 dB, respectively. The visual quality of the recovered image can also reach 26.19 dB on average in JPEG lossy channels. The high PSNR value and the satisfactory visual quality of the image indicate that the proposed method is effective in controlling the distortion and loss of the image. A limitation of this method lies in the manual design and selection of injected attentions, which diminishes the effectiveness of the proposed RFSA when confronted with alternative channel disturbances. Furthermore, the absence of robust design against disturbances from other channels results in suboptimal resistance against such disturbances. In the future, we will focus on automatically learning the attentions tailored to other lossy channels, such as the rounding effects in the lossy compression. Additionally, enhancing the payload capacity of robust CIC through multiple secret image hiding remains a task for improvement.

## CRediT authorship contribution statement

**Xiaobin Zeng:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation, Conceptualization. **Bingwen Feng:** Writing – review & editing, Writing – original draft, Methodology, Funding acquisition, Conceptualization. **Zhihua Xia:** Writing – review & editing, Supervision, Funding acquisition. **Zecheng Peng:** Validation, Software, Data curation. **Tiewei Qin:** Validation, Software, Data curation. **Wei Lu:** Writing – review & editing, Supervision, Methodology.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

Data will be made available on request.

## Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 61802145, 61932010, 62261160653, 62102101), Natural Science Foundation of Guangdong Province, China (Grant No. 2023A1515011348, 2019B010137005), the Fundamental Research Funds for the Central Universities, China.

## References

[1] C. Kumar, A.K. Singh, P. Kumar, A recent survey on image watermarking techniques and its application in e-governance, Multimedia Tools Appl. 77 (2018) 3597–3622.
[2] N. Subramanian, O. Elharrouss, S. Al-Maadeed, A. Bouridane, Image steganography: A review of the recent advances, IEEE Access 9 (2021) 23409–23423.
[3] B. Feng, W. Lu, W. Sun, Novel steganographic method based on generalized k-distance n-dimensional pixel matching, Multimedia Tools Appl. 74 (21) (2015) 9623–9646.
[4] T. Filler, J. Judas, J. Fridrich, Minimizing additive distortion in steganography using syndrome-trellis codes, IEEE Trans. Inf. Forensics Secur. 6 (3) (2011) 920–935.
[5] W. Li, W. Zhang, L. Li, H. Zhou, N. Yu, Designing near-optimal steganographic codes in practice based on polar codes, IEEE Trans. Commun. 68 (7) (2020) 3948–3962.

[6] S. Baluja, Hiding images in plain sight: Deep steganography, Adv. Neural Inf. Process. Syst. 30 (2017) 2066–2076.
[7] C. Yu, Attention based data hiding with generative adversarial networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 34, (01) 2020, pp. 1120–1128.
[8] X. Liu, Z. Ma, Z. Chen, F. Li, M. Jiang, G. Schaefer, H. Fang, Hiding multiple images into a single image via joint compressive autoencoders, Pattern Recognit. 131 (2022) 108842.
[9] S.-P. Lu, R. Wang, T. Zhong, P.L. Rosin, Large-capacity image steganography based on invertible neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 10816–10825.
[10] J. Jing, X. Deng, M. Xu, J. Wang, Z. Guan, HiNet: deep image hiding by invertible network, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 4733–4742.
[11] M. Ahmadi, A. Norouzi, N. Karimi, S. Samavi, A. Emami, ReDMark: Framework for residual diffusion watermarking based on deep networks, Expert Syst. Appl. 146 (2020) 113157.
[12] C. Zhang, A. Karjauv, P. Benz, I.S. Kweon, Towards robust deep hiding under non-differentiable distortions for practical blind watermarking, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 5158–5166.
[13] Z. Jia, H. Fang, W. Zhang, 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, 2021, pp. 41–49.
[14] Z. Zheng, Y. Hu, Y. Bin, X. Xu, Y. Yang, H.T. Shen, Composition-aware image steganography through adversarial self-generated supervision, IEEE Trans. Neural Netw. Learn. Syst. 34 (11) (2023) 9451–9465.
[15] Q. Ying, H. Zhou, X. Zeng, H. Xu, Z. Qian, X. Zhang, Hiding images into images with real-world robustness, in: 2022 IEEE International Conference on Image Processing, ICIP, IEEE, 2022, pp. 111–115.
[16] Y. Luo, T. Zhou, F. Liu, Z. Cai, IRWArt: Levering watermarking performance for protecting high-quality artwork images, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 2340–2348.
[17] J. Lu, J. Ni, W. Su, H. Xie, Wavelet-based CNN for robust and high-capacity image watermarking, in: 2022 IEEE International Conference on Multimedia and Expo, ICME, IEEE, 2022, pp. 1–6.
[18] M. Khayatkhoei, A. Elgammal, Spatial frequency bias in convolutional generative adversarial networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, (7) 2022, pp. 7152–7159.
[19] Y. Zhang, X. Luo, Y. Guo, C. Qin, F. Liu, Multiple robustness enhancements for image adaptive steganography in lossy channels, IEEE Trans. Circuits Syst. Video Technol. 30 (8) (2019) 2750–2764.
[20] W. Sun, J. Zhou, Y. Li, M. Cheung, J. She, Robust high-capacity watermarking over online social network shared images, IEEE Trans. Circuits Syst. Video Technol. 31 (3) (2020) 1208–1221.
[21] R. Singh, A. Ashok, An optimized robust watermarking technique using CKGSA in frequency domain, J. Inf. Secur. Appl. 58 (2021) 102734.
[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017) 6000—-6010.
[23] M. Zhu, W. Min, J. Han, Q. Han, S. Cui, Improved channel attention methods via hierarchical pooling and reducing information loss, Pattern Recognit. 148 (2024) 110148.
[24] A.M. Obeso, J. Benois-Pineau, M.S.G. Vázquez, A.Á.R. Acosta, Visual vs internal attention mechanisms in deep neural networks for image classification and object detection, Pattern Recognit. 123 (2022) 108411.
[25] J. Zhu, R. Kaplan, J. Johnson, L. Fei-Fei, Hidden: Hiding data with deep networks, in: Proceedings of the European Conference on Computer Vision, ECCV, 2018, pp. 657–672.
[26] G. Liu, Y. Si, Z. Qian, X. Zhang, S. Li, W. Peng, WRAP: Watermarking approach robust against film-coating upon printed photographs, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, pp. 7274–7282.
[27] T. Bui, S. Agarwal, N. Yu, J. Collomosse, Rosteals: Robust steganography using autoencoder latent space, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 933–942.
[28] J. Tan, X. Liao, J. Liu, Y. Cao, H. Jiang, Channel attention image steganography with generative adversarial networks, IEEE Trans. Netw. Sci. Eng. 9 (2) (2021) 888–903.
[29] H. Fang, D. Chen, F. Wang, Z. Ma, H. Liu, W. Zhou, W. Zhang, N. Yu, TERA: Screen-to-camera image code with transparency, efficiency, robustness and adaptability, IEEE Trans. Multimed. 24 (2021) 955–967.
[30] Y. Lan, F. Shang, J. Yang, X. Kang, E. Li, Robust image steganography: hiding messages in frequency coefficients, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, (12) 2023, pp. 14955–14963.
[31] L. Zhang, Y. Lu, J. Li, F. Chen, G. Lu, D. Zhang, Deep adaptive hiding network for image hiding using attentive frequency extraction and gradual depth extraction, Neural Comput. Appl. (2023) 1–19.
[32] Y. Luo, T. Zhou, S. Cui, Y. Ye, F. Liu, Z. Cai, Fixing the double agent vulnerability of deep watermarking: A patch-level solution against artwork plagiarism, IEEE Trans. Circuits Syst. Video Technol. (2023) 1.
[33] F. Shang, Y. Lan, J. Yang, E. Li, X. Kang, Robust data hiding for JPEG images with invertible neural network, Neural Netw. 163 (2023) 219–232.
[34] F. Cao, T. Wang, D. Guo, H. Yao, J. Li, C. Qin, Universal screen-shooting robust image watermarking with channel-attention in DCT domain, Expert Syst. Appl. (2023) 122062.
[35] C. Zhang, P. Benz, A. Karjauv, G. Sun, I.S. Kweon, Udh: Universal deep hiding for steganography, watermarking, and light field messaging, Adv. Neural Inf. Process. Syst. 33 (2020) 10223–10234.
[36] A.B. Watson, DCTune: A technique for visual optimization of DCT quantization matrices for individual images, in: Sid International Symposium Digest of Technical Papers, Vol. 24, Citeseer, 1993, p. 946.
[37] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[38] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.
[39] J. Johnson, A. Alahi, L. Fei-Fei, Perceptual losses for real-time style transfer and super-resolution, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11-14, 2016, Proceedings, Part II 14, Springer, 2016, pp. 694–711.
[40] H.R. Sheikh, A.C. Bovik, Image information and visual quality, IEEE Trans. Image Process. 15 (2) (2006) 430–444.

**Xiaobin Zeng** is currently studying for a master's degree in the College of Cyber Security at Jinan University. His research interests include multimedia security and artificial intelligence.

**Bingwen Feng** received the M.S. degree in Software Engineering and the Ph.D. degree in Computer Science from Sun Yat-sen University, China in 2008 and 2014, respectively. He is currently an associate researcher with the College of Cyber Security, Jinan University, Guangzhou, China. His research interests include software security and multimedia security.

**Zhihua Xia** received the B.S. degree from Hunan City University, China, in 2006, and the Ph.D. degree in computer science and technology from Hunan University, China, in 2011. He is currently a professor with the College of Cyber Security, Jinan University, Guangzhou, China. His research interests include digital forensic and encrypted image processing.

**Zecheng Peng** is currently studying for a master's degree in the College of Cyber Security at Jinan University. His research interests include multimedia security and artificial intelligence.

**Tiewei Qin** is currently studying for a master's degree in the College of Cyber Security at Jinan University. His research interests include multimedia security and artificial intelligence.

**Wei Lu** received the B.S. degree in automation from Northeast University, China, in 2002, and the M.S. and Ph.D. degrees in computer science from Shanghai Jiao Tong University, China, in 2005 and 2007, respectively. He was a research assistant at Hong Kong Polytechnic University from 2006 to 2007. He is currently a professor with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. His research interests include multimedia forensics and security, data hiding and watermarking, privacy protection. He has published more than 100 papers in security conferences and journals, such as TIFS, TDSC, TCSVT, TPAMI, TNNLS, TCYB, etc. He is an associate editor for the Signal Processing and the Journal of Visual Communication and Image Representation.