Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

SCFformer: a binary data hiding method against JPEG compression based on spatial channel fusion Transformer*

Xintao DUAN†‡[1,2], Chun LI[1,2], Bingxin WEI[3], Guoming WU[1,2], Chuan QIN[4], Haewoon NAM[3]

[1]*College of Computer and Information Engineering, Henan Normal University, Xinxiang 453007, China*

[2]*Key Laboratory of Artificial Intelligence, Henan Normal University, Xinxiang 453007, China*

[3]*Department of Electrical and Electronic Engineering, Hanyang University, Ansan 15588, South Korea*

[4]*Optical-Electrical and Computer Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China*

†E-mail: duanxintao@htu.edu.cn

Received Nov. 8, 2023; Revision accepted Mar. 28, 2024; Crosschecked

Abstract: To enhance information security during transmission over public channels, images are frequently employed for binary data hiding. Nonetheless, this data is vulnerable to distortion due to JPEG compression, leading to challenges in recovering the original binary data. Addressing this issue, this paper introduces a pioneering method for binary data hiding that leverages a combined spatial and channel attention Transformer, termed SCFformer, designed to withstand JPEG compression. This method employs a novel DCT quantization truncation mechanism during the hiding phase to bolster the stego image's resistance to JPEG compression, utilizing spatial and channel attention to conceal information in less perceptible areas and thereby enhancing the model's resistance to steganalysis. In the extraction phase, the DCT quantization stage mechanism minimizes secret image loss during compression, facilitating easier information retrieval. The incorporation of scalable modules adds flexibility, allowing for variable-capacity data hiding. Experimental findings validate the high security, large capacity, and flexibility of our scheme, alongside a marked improvement in binary data recovery post-JPEG compression, underscoring our method's leading-edge performance.

Unedited

Key words: Binary data hiding; Against JPEG compression; DCT quantization; SCFformer
https://doi.org/10.1631/FITEE.2300762 **CLC number:**

1 Introduction

As a method of embedding secret information in digital media, data hiding technology (Li and Zhang, 2019; Shang et al., 2023; Zhu et al., 2018) has important application value in protecting privacy and confidential communication. In the field of data hiding, images(Mallika et al., 2022), text(Jassim, 2013), video(Mou et al., 2023) and speech signals(Abood et al., 2022) can be transmitted as message carriers. In order to take advantage of the redundancy of image information, image is selected as the research content in this paper. Image has been widely used as a cover of binary information. However, in practical application scenarios, social platforms and major websites use compression operations on images due to storage and transmission costs, the most common of which is JPEG compression. After the image is compressed, the binary information cannot be completely extracted, and the non-difference of JPEG makes it more difficult to train a model to resist lossy compression. This affects the significance of the user for private transmission of information. The essence of transmission should be to ensure the integrity of information. Therefore, the data hiding algorithm

‡ Corresponding author

* This work was supported in part by the National Natural Science Foundation of China (Nos. U1904123, 62172280, and U20B2051), in part by the Key Scientific Research Projects of Colleges and Universities in Henan Province (No. 23A520006), and in part by Henan Provincial Science and Technology Research Project (No. 222102210199)

© Zhejiang University Press 2024

with practical application value should not only ensure strong anti-detection ability, but also ensure the complete extraction of embedded messages.

The algorithm proposed in Reference(Tao et al., 2018) can resist the content loss caused by JPEG compression. However, this method uses an ideal JPEG channel, which may not be applied to non-differential JPEG compression channels in practical applications. Therefore, SRJS(Lu et al., 2021) proposed a JPEG data hiding algorithm, which can resist JPEG compression with QF = 75,85 and 95, and use adaptive BCH coding to reduce the error rate of secret message extraction. However, these algorithms still cannot resist JPEG compression with QF lower than 75 in the face of high capacity, and may not meet the requirements of current social networks. RDHINN(Shang et al., 2023) directly embeds the secret message into the quantized discrete cosine transform (DCT) coefficients of the cover image and decodes the secret message from the quantized DCT coefficients of the attacked stego image. Although this method has a certain resistance to lossy JPEG compression in essence, the extraction error rate of the extracted secret message is still not ideal. FNNS(Kishore et al., 2022) achieves low error rate extraction by simulating the JPEG compression process, and the algorithm can resist JPEG compression with QF = 80. However, these methods lack effectiveness in practical applications.

Aiming at the problem that the current data hiding algorithm has a large error in extracting embedded messages after low quality factor JPEG compression and lacks effectiveness in practical applications. In this paper, a novel anti-JPEG compression binary data hiding method based on spatial channel fusion Transformer(SCFformer) is proposed, which can recover secret information with high precision after image processing. First of all, we use an efficient fusion Transformer to capture the dimensional relationship of the entire space and channel while maintaining computational efficiency. Secondly, we introduce the DCT quantization truncation mechanism to improve the anti-JPEG compression ability. In addition, we also use the scalable module to realize the flexibility of secret information hiding, so that our scheme can deal with different capacity of secret information. The specific contributions are as follows :

1. A novel anti-JPEG compression binary data hiding method based on Spatial Channel Fusion Transformer (SCFformer) is proposed, which can provide high-precision secret information recovery with low quality factor.
2. The DCT quantization truncation mechanism is proposed, which improves the anti-JPEG compression ability of the stego image, and further improves the flexibility by using the scalable module, and realizes the secret data hiding with variable capacity.
3. The efficient fusion Transformer is proposed, which improves capturing the dimensional relationships of the entire space and channels while maintaining high efficiency, thereby achieving a lower secret data extraction error rate at the same capacity.

The remainder of this article is organized as follows. The second part introduces related work; the third part introduces the Spatial Channel Fusion Transformer (SCFformer) proposed in this article; the fourth part introduces the experimental results and analysis; finally, it summarizes the work of this article and looks forward to future research directions.

Unedited

2 Relate work

2.1 Traditional anti-JPEG compression data hiding

Since the emergence of image watermarking, researchers have explored various methods to improve the anti-JPEG compression ability and hiding capacity of data hiding. In order to improve the anti-JPEG compression ability of data hiding, the trend has shifted to hiding in the transform domain, such as DCT(Singh and Bhatnagar, 2018), discrete wavelet transform (Lu et al., 2009) or a mixture of them (Singh et al., 2012; Harish et al., 2013).In order to achieve a good balance between capacity and anti-JPEG compression capability, adaptive data hiding has emerged to embed hidden data based on the local features of the carrier image (Bhinder et al., 2018; Chen et al., 2018) . In (Zhang et al., 2015), by using the relative relationship between DCT coefficients, a message embedding method with strong resistance to JPEG compression is constructed. This method can resist JPEG compression under multiple quality

factors, but it will sacrifice some anti-detection capabilities under high embedding capacity. Reference (Yu et al., 2020) proposed an anti-JPEG data hiding method combined with jitter modulation, which constructs the asymmetric cost of different modification polarity. By combining the recompression knowledge, this method can improve the anti-JPEG compression ability and security under the known channel quality factor.In (Zhang et al., 2021), by analyzing the principle of JPEG compression, the quantization rounding principle is used to construct the message embedding domain, and based on the coefficient relationship and adaptive modification, it has the ability to resist JPEG compression, Gaussian noise and so on. In addition, the method combines the embedding cost calculation function and the coding method to achieve message embedding, which achieves higher embedding efficiency. In (Zhao et al., 2019), an anti-JPEG compression method based on transmission channel matching is proposed. This method adjusts the carrier image by repeatedly inputting the carrier image into the transmission channel until the generated carrier image can resist the JPEG compression of the transmission channel. In addition, they also use adaptive BCH coding to reduce the extraction error rate. However, these methods cannot solve the problem of third-party interference in public channels faced by data hiding.

2.2 Anti-JPEG compression data hiding based on deep learning

Traditional data hiding methods usually require practitioners to have a good understanding of a wide range of related professional knowledge, and methods based on deep learning automate the entire process, which greatly promotes its widespread use. In addition to its success in a wide range of applications, deep learning has also shown a trend of success in data hiding(Isac and Santhi, 2011). Reference (Shang et al., 2023) has shown the possibility of hiding another complete image in one image. In Reference(Hayes and Danezis, 2017), a method of hiding binary data by adversarial learning was proposed. The above deep learning-based methods do not take into account its anti-interference ability, and the hidden data can be easily destroyed by some common lossy compression methods such as JPEG. In order to improve the resistance to JPEG compression, Reference (Zhu et al., 2018) proposed to add a ' noise layer ' to simulate JPEG during the training process to enhance the encoded image. In order to overcome the problem that real JPEG is non-differentiable, there are two approximate solutions, namely JPEG-Mask and JPEG-Drop, to differential JPEG compression to remove high-frequency content in the image. Although JPEG-Mask and JPEG-Drop also remove high-frequency content like real JPEG, their behavior is still far from real JPEG, resulting in overfitting and poor performance when tested under real JPEG compression. In order to minimize the gap between approximate and ' real ' JPEG, a new method was proposed recently in Reference (Li et al., 2023) to carefully simulate the important steps in ' real ' JPEG. A similar method is also proposed in Reference (Bui et al., 2023) to generate adversarial samples against JPEG compression. This paper also uses a similar method to combat JPEG compression.

Unedited

3 The Proposed Method

3.1 Net Framework

We introduce a non-convolutional U-Net-like Spatial Channel Fusion Transformer (SCFformer) data hiding scheme, as shown in Fig 1. In the process of hiding, given the cover image $X_{co} \in \{H \times W \times C\}$, the secret information $X_{se} \in \left\{\frac{H}{4} \times \frac{W}{4} \times n\right\}$. The spatial dimension of the image is $H \times W$, the number of channels is C, the spatial dimension of the secret information is $\frac{H}{4} \times \frac{W}{4}$, and the number of channels is n. First, n Xse are upsampled and merged with the DCT quantized truncated Xco in the dimension of the channel. Then we use the Patch Merging module (Liu et al., 2021; Huang et al., 2023)(to obtain 4×4 tokens of size from the merged results). Finally, the tokens ($x_n \times d$) are passed through the encoder module, which has two stacked encoder blocks, each of which consists of a fusion Transformer layer and a Patch Merging layer. In the merging process, 2×2 tokens are merged to reduce the spatial dimension and double the channel dimension. It enables the network to obtain multi-scale representation in a hierarchical manner. In the decoder, the tokens in each block are expanded twice again. Then, the output of each patch Merging layer is fused with the skip connection of the parallel encoder layer using SCCA.

3.2 Efficient Fusion Transformer

The traditional self-attention mechanism generates a redundant context matrix, and Eq 1 is a more efficient process of calculating the self-attention mechanism.

$$E(Q,K,V) = \rho_{\mathbf{q}}(Q)(\rho_{\mathbf{k}}(K)^T V) \tag{1}$$

Where andare the normalization functions of Q and K, respectively. Driven by the data hiding task, the goal of attention is to embed the secret data into the cover image in the process of data hiding, while minimizing the impact and distortion of the stego image. The attention mechanism can help the network pay more attention to the unimportant parts of the carrier when embedding secret data, so as to reduce the impact on the stego image as much as possible, and improve the concealment and security of embedded secret data. We consider that different types of attention mechanisms can be used to achieve the above goals. In order to capture cross-channel information more effectively, we use the transposed attention mechanism. Eq 2 shows the process of transposed attention mechanism.

$$\begin{aligned} T(Q,K,V) &= V \cdot C_T(K,Q) \\ C_T(K,Q) &= \mathrm{Softmax}(\frac{K^T Q}{\tau}) \end{aligned} \tag{2}$$

After $QandK$ are transposed, the attention weight must be calculated based on the cross-covariance matrix. is a global context vector that refers to the transposition of attention. Before calculating the attention weight, the temperature parameter t is introduced to offset the $L2-norm$ scaling from Q and K. this operation will increase the stability of the training process, but reduce a certain degree of freedom. The literature on attention mechanism(Guo et al., 2022) shows that the combination of spatial attention and channel attention can improve the ability to capture more contextual features. Therefore, we construct an efficient fusion Transformer, which combines transpose (channel) attention and efficient (spatial) attention. Its structure is shown in Fig 2 It includes an efficient attention block, followed by an addition and normalization block, and a transposed attention block, and finally an addition and normalization block, which is expressed by the formula as shown in Eq 3. The structure can capture more unnoticed image regions more efficiently, thereby efficiently hiding information. Here, $Tblock$ denotes the transposed attention block and $Eblock$ denotes the spatial attention block. Q_1, K_1, and V_1 are calculated based on the input X, and Q_2, K_2, and V_2 are calculated from the input of the transposed at-

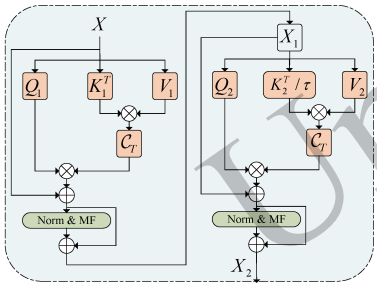

Fig. 2 Efficient fusion Transformer module. It includes an efficient spatial attention block and a channel attention block, followed by a specification and MF.

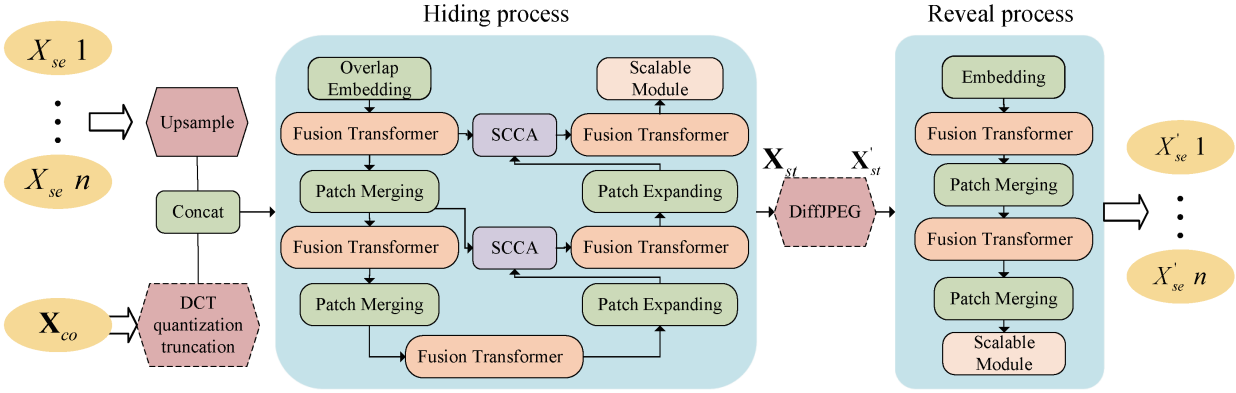

Fig. 1 The overall framework of SCFformer data hiding scheme mainly includes two parts : hiding process and extraction process.

Unedited

tention block. MF represents Mix-FFN feedforward neural network(Huang et al., 2023) which by Eq 4.

$$MF(X) = FC(GELU(DW(FC(X))) \quad (4)$$

FC is a fully connected layer. GELU refers to GELU activation (Gaussian error linear units), DW refers to deep convolution.

3.3 Skip Connection Cross Attention

The SCCA module is shown in Fig 3. Unlike simply connecting the characteristics of the encoder and decoder layers, we cross-participate in them to retain the underlying characteristics more effectively. The module can effectively provide spatial information for each encoder, so that fine-grained detail information can be restored when a stego-image is generated. In Eq 5, skip connection cross attention (SCCA) applies efficient attention, but it is not simple to use the same input for K, Q and V. Q is output by the encoder layer of X_2, and the input of K and V is the output of the decoder layer X_1. In order to fuse these two features, X_1 needs to use the linear layer scaling to the same dimension as X_2 for embedding. The role behind using X_2 as Q input is to perform multi-level representation in efficient spatial attention blocks.

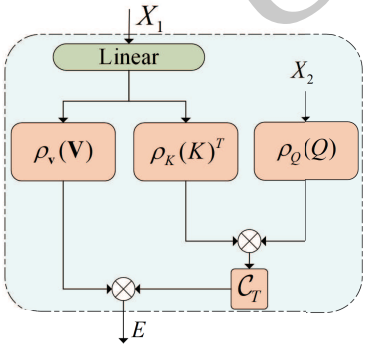

Fig. 3 The SCCA module contains a linear layer and an efficient attention module.

$$\begin{aligned} K, V &= Proj(FC(X_1)), Q = Proj(X_2) \\ E &= \rho_v(V)\rho_k(K^T)Q \end{aligned} \quad (5)$$

3.4 DCT Quantization Truncation mechanism

In JPEG compression, there is a special mechanism called recompression mechanism. In the implementation of some JPEG data hiding schemes (such as OutGuess and F5), additional JPEG compression may be performed before data embedding, and then in the social network, the picture will be re-compressed and the DCT coefficient of the Stego image will be changed, resulting in adaptive data hiding failure. Although some methods can hide in the undiscarded position of DCT coefficients, when the carrier image is uploaded to the online social network, using a smaller QF for JPEG compression will still cause the data hiding scheme to fail. We found that we can effectively deal with the problems caused by JPEG compression by quantifying an image by DCT quantization, generating the corresponding DCT quantization coefficients, and then performing quantization truncation according to the degree of interference, that is, the JPEG quality compression factor in the disturbed scene, and then performing inverse DCT transformation, and finally performing data embedding. This mechanism is called DCT quantization truncation mechanism. Under this mechanism, we can select a set of specific DCT quantization tables and use them to convert the original image into DCT coefficients. Then, we can embed these coefficients without worrying about the changes of DCT coefficients that may be caused by JPEG compression. After embedding the data, we can restore the DCT coefficients to the image as needed. This mechanism can improve the anti-JPEG compression ability of the data hiding scheme, and can realize the secure embedding and extraction of

$$\begin{aligned} E_{block}(X, Q_1, K_1, V_1) &= E(Q_1, K_1, V_1) + X \\ MF_1(E_{block}) &= MF(LN(E_{block})) \\ T_{block}(E_{block}, Q_2, K_2, V_2) &= T(MF_1(E_{block}) + E_{block}) + MF_1(E_{block}) \\ MF_2(T_{block}) &= MF(LN(T_{block})) \\ FT(T_{block}) &= MF_2(T_{block}) + T_{block} \end{aligned} \quad (3)$$

Unedited

data without affecting the image quality. This mechanism has been widely used in various JPEG data hiding schemes and proved to be a very effective solution.

3.5 Key Control Mechanism

This paper introduces a new key control scheme, which can be used to further encrypt the secret information when the image is hidden, and to recover the secret information under certain key conditions. The design structure of this scheme is shown in Fig 4. In this scheme, given a cover image, it first passes through a layer of fusion Transformer layer. Next, a specific key is generated by a key encoder, which consists of a fully connected (FC) layer. Then, by fusing the input key at the end of the Transformer layer, a condition vector of the 2Crpm channel is generated, which is divided into two modulation vectors $\alpha, \beta \in \mathbb{R}^{Crpm\times 1\times 1}$ in the channel dimension. Here, CRPM is the number of feature channels for each fused Transformer layer in RPM. Finally, the output feature is $R = Frpm \times \alpha + \beta$. During the training process, we constrain the output and key and secret information during extraction. During the training process, the scheme constrains the output, key and secret information during extraction. In this way, it can ensure that the scheme has higher security and accuracy when hiding and restoring secret information.

3.6 Scalable Module

In this way, it can ensure that the scheme has higher security and accuracy when hiding and restoring secret information. It is located before the output layer. Our scalable design, shown in Fig 5, can be considered as a special convolution layer whose kernel size can be changed according to the input.

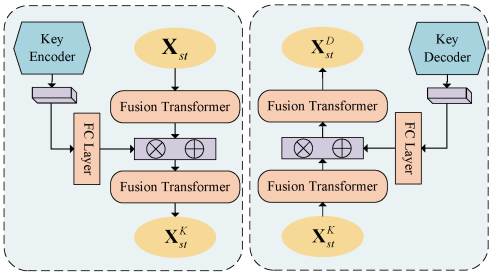

Fig. 4 The key control mechanism consists of two parts, namely key generation and information extraction

All convolution kernels of different dimensions come from the same base kernel and share parameters. For a given input feature, we intercept a convolution kernel with a size of k ÃŮ k to match the input dimension, and then convolve. We implement this operation by training all convolution kernels.

3.7 Loss function and training process

In order to understand the influence of different scenarios on the data hiding effect, we use three model variants to calculate the loss function respectively. For convenience, three different variants are called $Variation\#A$, $Variation\#B$,$Variation\#C$ respectively. In order to distinguish different variants, we set different security levels. $Variation\#A$ can be used in scenarios where the user requires higher security without considering the computational cost, $Variation\#B$ can be used in scenarios where the user requires certain security and considers the computational cost, and $Variation\#C$ can be used in scenarios where the user requires certain security and considers the image quality. In our model, the loss function is used to constrain two parts, namely hiding and recovering. Hiding is to hide multiple secret information in the carrier image. The secret image Xst generated by should not detect the secret information and be as similar to the carrier image as possible. Therefore, we constrain the stego image X_{st} to be the same as the carrier image X_{co} : In $Variation\#A$, we use the DCT quantization truncation mechanism to calculate the hidden loss, the formula is shown in Eq 6.

$$\mathcal{L}_h = \frac{1}{\mathrm{n}} \sum_{i=1}^{\mathrm{n}} \| DiffJPEG(\mathbf{X}_{st}, QF) - \mathrm{QT}(\mathbf{X}_{co}, QF) \|_2^2 \tag{6}$$

In $Variation\#B$, we abandon the DCT quantization truncation mechanism to calculate the hidden

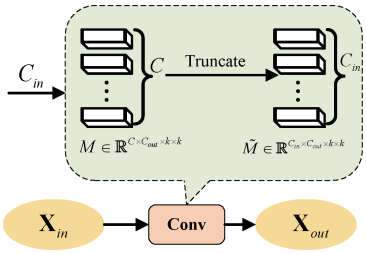

Fig. 5 Scalable Module

Unedited

loss, and the formula is shown in Eq 7.

$$\mathcal{L}_h = \frac{1}{\mathrm{n}} \sum_{i=1}^{\mathrm{n}} \left\| \mathbf{X}_{co}^i - \mathbf{X}_{st}^i \right\|_2^2 \tag{7}$$

In $Variation\#C$, based on the DCT quantization truncation mechanism, we perform the inverse process of image degradation on the generated carrier image, trying to restore the image degradation caused by JPEG compression, and calculate the hidden loss. The formula is shown in Eq 8.

$$\mathcal{L}_h = \frac{1}{\mathrm{n}} \sum_{i=1}^{\mathrm{n}} \left\| DiffJPEG(\mathbf{X}_{st}^i, JCF) - \mathbf{X}_{co}^i \right\|_2^2 \tag{8}$$

Algorithm 1: The training process

Data: The secret message $X_{se}^1, X_{se}^2, X_{se}^n$, the cover image X_{co}

Result: Trained networks $Hiding()$,$Reveal()$, stego image X_{st} and rccovered secret message$X_{se}^{'}$.

for *all layers* **do**
  Initialization;
end
for *each of epoch* **do**
  for *each of iteration* **do**
    $X_{se} = Concat(X_{se}^1, X_{se}^2, X_{se}^n)$;
    $X_{co}^{'} = QT(X_{co})$;
    $X_{st} = Hiding(X_{se}, X_{co}^{'})$;
    $Loss = MSE(X_{st}, X_{co}^{'})$;
    $X_{st}^{'} = DiffJPEG(X_{st})$;
    $X_{se}^1, X_{se}^2, X_{se}^n = Reveal(X_{st}^{'})$;
    Update network parameters alternately;
    $\theta_t \leftarrow \theta_t - \eta \times \frac{\partial(\lambda_c L_c + \lambda_r L_r - \lambda_d L_d)}{\partial \theta_t}$;
  end
  Shuffle the training set.
end

where $\mathbf{X}_{co}^i$ is the cover image, the stego image, the image degradation process, and the JCF represents the JPEG compression factor. In the recovery process, our goal is to recover the secret information from the stego image to be the same as the original secret information, then the corresponding loss function is shown in formula 9.

$$\mathcal{L}_r = \frac{1}{\mathrm{n}} \sum_{i=1,j=1}^{\mathrm{n}} \left\| \hat{\mathbf{X}}_{se@\mathrm{j}}^i - \mathbf{X}_{\mathrm{se@j}}^i \right\|_2^2 \tag{9}$$

For the overall optimization model, we calculate the total loss by minimizing the hidden loss and the extraction loss, as shown in Formula 10.

$$\mathcal{L} = \mathcal{L}_h + \lambda \mathcal{L}_r \tag{10}$$

where λ is a hyperparameter that can be traded between hiding and recovering. We set $\lambda = 0.5$ to balance these two losses. The whole training process of this method is shown in Algorithm 1.

4 Experiments

This chapter mainly introduces the experimental situation of our work. We first give an overview of the implementation details, then briefly introduce the evaluation indicators used, and then perform capacity analysis, visual quality analysis, anti-JPEG compression capability analysis, and security analysis. Finally, ablation experiments were carried out to verify the effectiveness of the proposed method.

4.1 Experiment Setting

In this work, we use 30,000 images selected from ImageNet as the training set of cover images to train our model. The spatial resolution is 224 ÃŮ 224. We use AdamW optimizer(Zhang et al., 2021), $\beta 1 = 0.9$, $\beta 2 = 0.5$. We set the batch size to 16.The weight attenuation factor is set to 1x10-12. Using an initial learning rate of 1x10-4, the hidden network will be reduced by half every 5 iterations without decreasing the loss, and the recovery network will be reduced by half every 10 iterations without decreasing the loss. The total number of iterations is set to 300. On a NVIDIA Tesla V100 GPU, the training process can be completed within 5 days.

4.2 Parameter Evaluation

In this paper, the error-sensitive peak signal-to-noise ratio (PSNR) (Hore and Ziou, 2010) and structural similarity index measurement (SSIM)(Ye et al., 2017b)are used as indicators to measure the data hiding effect. MSE represents the mean square error C' between the carrier image C and the Stego image. PSNR index has been widely used and recognized in the field of computer vision. Its calculation source is the difference between the pixels at the same position in the two images, which can objectively reflect the image quality after compression or other processing. Since the image often deviates during transmission or processing, the higher the PSNR

Unedited

value, the smaller the image change after processing. In this paper, PSNR is selected for measurement because it can accurately measure the difference between the model coverage image and the Stego image. In this paper, the embedding of information in the steganography process is regarded as the destruction of image quality, so PSNR can be used to evaluate the distortion degree of the secret information embedded in the stego image. Pe is a commonly used measure to evaluate the undetectability of stego images, which is defined as formula 11.

$$Pe = \frac{1}{2}\left(P_{FA} + P_{MD}\right) \tag{11}$$

P_FA and P_MD were false positive rate and missed detection rate. The value range is [0,1], and the optimal value is 0.5. Data hiding capacity can be obtained by Payload by calculating the number of bits hidden per pixel. The commonly used calculation formula 12.

$$Payload = \frac{m}{W \times H}(bpp) \tag{12}$$

where m represents the total number of embedded binary data, w and h represent the width and height of the image embedded in the data, respectively. bpp represents the number of bits per pixel(bpp).

$$Errorrate = \sum_{i=1}^{n}\left(\frac{|x_i - x_i'|}{m}\right) \tag{13}$$

The error rate can be calculated by Formula 13, where x represents the secret information to be sent, and represents the secret message to be sent. After our algorithm is embedded into the carrier image, the carrier image is compressed and interfered, and the secret message is re-extracted. Where i denotes the i-th secret message.

4.3 Capacity analysis

Table 1 Capacity analysis different N values

N	1	2	3	4	5	6	7	8
Payload(bpp)	0.0625	0.125	0.1875	0.25	0.3125	0.375	0.4375	0.5

Effective data hiding algorithms need to consider three key factors : visual quality, security and payload capacity. Payload capacity refers to the amount of data that can be hidden under the premise of maintaining the visual quality of the image. If the payload capacity is larger, the visual quality of the image will become lower, and vice versa. Therefore, a well-designed data hiding algorithm should have a high payload capacity while maintaining the visual quality of the image. The values shown in the Table 1 are calculated according to the Bpp formula, where N is the capacity criterion of our method.

4.4 Visual Quality Analysis

Visual quality is an important metric for data hiding algorithms and an important consideration for user selection algorithms. Firstly, when N = 3, the subjective visual quality of our algorithm under different QF is shown in Fig 6. It can be seen from the graph that under different quality factors, the stego-image and the stego-image after interference still maintain visual similarity with the original image, and the residual is basically invisible. From the histogram of the residual graph, it can be seen that the residual pixels are basically concentrated in the left region to ensure the concealment of the message. Therefore, in practical applications, objective assessment is widely used. In general, we use peak signal-to-noise ratio (PSNR) and structural similarity index

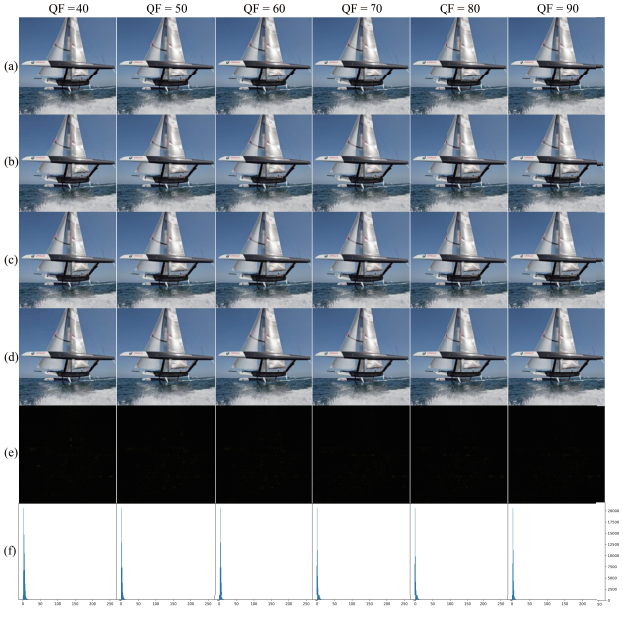

Fig. 6 When N = 3, the subjective visual quality under different QF, (a) cover image, (b) The quantized truncated cover image (c) Stego image (d) Compressed stego image (e) The residual of the Stego image and the quantized truncated cover image. (f) The histogram of the residual between the Stego image and the quantized truncated cover image.

Unedited

(SSIM) to evaluate the impact of data hiding on the image, so as to obtain the visual quality of the image. Generally speaking, the higher the PSNR and SSIM, the smaller the impact of the data hiding process on the image, and the better the visual quality of the image. We draw the trend of SSIM and PSNR as the capacity increases under different N values and different QF values. It can be seen from Fig 7,8 that as the capacity increases, the image quality continues to decline, but it is still within the acceptable range. Under the same capacity, the higher the QF value, the better the visual performance.

Table 2) shows that our scheme has a lower information extraction error rate than FNNS, and the visual quality is also higher than other schemes under the same quality factor. For RDHINN, our scheme also performs better under the same quality factor, and achieves a low error rate extraction with a lower quality factor.

Table 2 Visual quality comparison

QF	Method	Secret type	PSNR	SSIM	Error rate
65	RDHINN	image	33.02	0.906	–
	EFDR	image	31.47	0.913	-
	FNNS	binary	28.35	0.852	**0.000067**
	Our method	binary	**38.12**	**0.969**	0.000665
75	RDHINN	image	33.88	0.929	-
	EFDR	image	32.16	0.931	-
	FNNS	binary	28.56	0.857	**0.000065**
	Our method	binary	**38.85**	**0.976**	0.000721
85	RDHINN	image	34.77	0.951	-
	EFDR	image	33.64	0.927	-
	FNNS	binary	28.80	0.865	**0.000059**
	Our method	binary	**39.21**	**0.988**	0.000654
90	RDHINN	image	36.23	0.961	-
	EFDR	image	34.25	0.925	-
	FNNS	binary	29.10	0.874	**0.000043**
	Our method	binary	**39.4**	**0.982**	0.000512

4.5 Analysis of anti-JPEG compression ability

In the case of different capacity. Fig 9 shows the error rate performance under different QFs at different capacities. For the convenience of representation, we logarithmically process the error rate in the figure to clearly show its changes. It can be seen from the graph that the extraction performance of the fusion Transformer basically exceeds the traditional Unet by an order of magnitude. And in the case of high capacity and low quality factor, it can still maintain an error rate close to 0. Table3 shows that the traditional STC-based methods J-UNIWARD and UED can indeed recover messages under different

Unedited

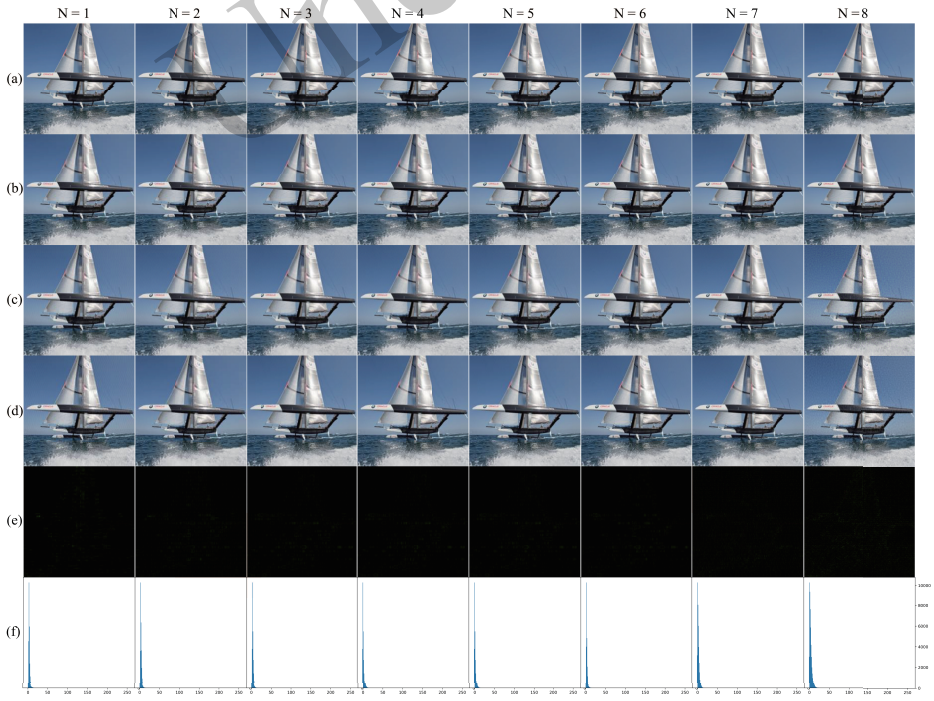

Fig. 7 When QF = 75, the subjective visual quality under different N, (a) cover image, (b) The quantized truncated cover image (c) Stego image (d) Compressed stego image (e) The residual of the Stego image and the quantized truncated cover image. (f) The histogram of the residual between the Stego image and the quantized truncated cover image.

JPEG compression, but their error rates are high. The proposed JPEG steganography scheme can effectively learn the inverse transform of the compression channel, so that the proposed scheme can learn the JPEG channel compression process under different QF values and provide anti-JPEG compression performance.

4.6 Generalization performance analysis

In order to show the generalization performance of the model, we analyze the generalization performance of the three model variants when QF = 75, N = 3. To evaluate the performance of the model on different data sets. We can use multiple different data sets to test the generalization performance of the model to ensure that the model not only performs well on the training set, but also performs well on unseen data. If the model performs well on multiple data sets, then we can think that its generalization performance is good. After training, we used ImageNet1K, CelebA, DIV2K and MS-coco datasets to evaluate the three model variants, respectively. We use several evaluation metrics to evaluate the generalization performance of the model, including loss value, error rate, PSNR and SSIM. According to the experimental results in Table 4, our model shows good generalization performance on different data sets. The highest value of PSNR reaches more than 40, and the highest even reaches 45.77. At the same time, the average value is about 38, and the average value is up to 42.91. Although the performance of $Varation\#B$ is not as good as $Varation\#A$, it is also acceptable because it does not use the DCT quantization truncation mechanism. The performance of $Varation\#C$ is extremely poor because it is trying to restore irreversible image degradation, but it is still within the acceptable range.

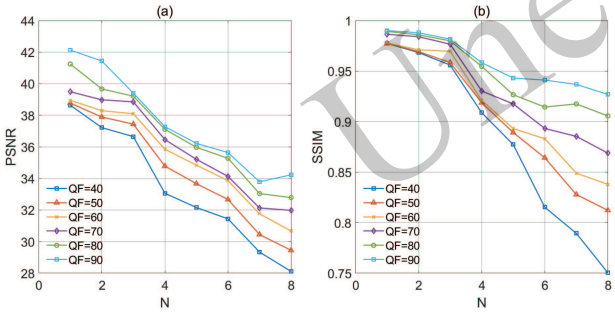

Fig. 8 (a) PSNR mean. (b) SSIM mean. N represents the capacity, and the ordinate represents the PSNR value or SSIM value. Different color lines represent different QF values.

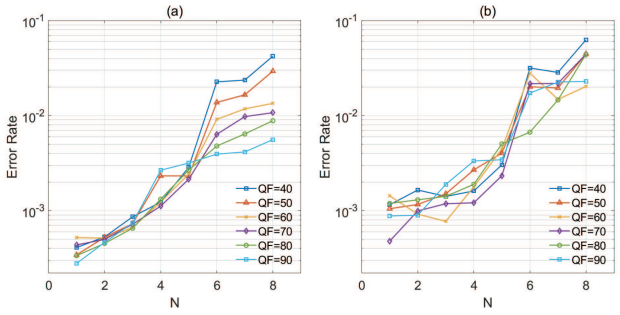

Fig. 9 Error rate performance under different N (a) using SCFformer error rate performance, (b) using Unet error rate performance.

Table 3 Comparison of the error rate with the advanced anti-JPEG compression method. The capacity of this method is represented by N value, and the other methods are represented by Payload.

	Payload(bpp)/N	0.05/1	0.1/2	0.2/4	0.3/6
QF=90	UED2(Guo et al., 2012)	0.3876	0.3451	0.3134	0.3003
	UED2-JCRISBE(Zhao et al., 2019)	0.1207	0.1157	0.1023	0.0889
	UED2-SRJS (Lu et al., 2021)	0.1141	0.1024	0.0979	0.0757
	J-UNIWARD (Holub et al., 2014)	0.3507	0.3314	0.3146	0.2943
	J-UNIWARD- JCRISBE	0.1102	0.0842	0.0705	0.0564
	J-UNIWARD- SRJS	0.0679	0.0565	0.0453	0.0392
	Our method	**0.0002**	**0.0004**	**0.0026**	**0.0039**
QF=85	UED2	0.3786	0.3452	0.3161	0.2961
	UED2-JCRISBE	0.1196	0.1088	0.0957	0.0848
	UED2- SRJS	0.1191	0.1053	0.0897	0.0767
	J-UNIWARD	0.3561	0.3316	0.3121	0.2976
	J-UNIWARD- JCRISBE	0.1121	0.0864	0.0735	0.0601
	J-UNIWARD- SRJS	0.0635	0.0532	0.0439	0.0385
	Our method	**0.0003**	**0.0004**	**0.0013**	**0.0047**
QF=75	UED2	0.3823	0.3668	0.3418	0.3182
	UED2-JCRISBE	0.1174	0.1057	0.0976	0.0831
	UED2- SRJS	0.1102	0.1042	0.0913	0.0784
	J-UNIWARD	0.3527	0.3372	03197	0.3013
	J-UNIWARD- JCRISBE	0.1089	0.0811	0.0651	0.0519
	J-UNIWARD- SRJS	0.0615	0.0517	0.0439	0.0391
	Our method	**0.0004**	**0.0005**	**0.0011**	**0.0063**

Table 4 Generalization performance results in four data sets

DataSet	Type	MSE	Error rate	PSNR		SSIM	
				Max	Average	Max	Average
ImageNet1K	Varation#A	0.000180	0.000721	45.77	38.85	0.990	0.976
	Varation#B	0.000195	0.000972	43.29	38.12	0.986	0.967
	Varation#C	0.000572	0.001216	41.33	33.61	0.978	0.928
CelebA	Varation#A	**0.000064**	**0.000139**	**46.06**	**42.91**	0.990	**0.980**
	Varation#B	0.000090	0.000140	43.68	41.12	0.983	0.968
	Varation#C	0.000198	0.000420	42.33	37.98	0.974	0.954
Div2K	Varation#A	0.000222	0.000801	43.91	38.40	0.988	0.978
	Varation#B	0.000241	0.000725	41.48	36.97	0.988	0.971
	Varation#C	0.000741	0.001615	38.14	32.09	0.973	0.933
MS-coco	Varation#A	0.000172	0.000801	45.58	39.35	**0.991**	0.977
	Varation#B	0.000194	0.000537	42.98	38. 23	0.987	0.967
	Varation#C	0.000540	0.001154	41.23	33.64	0.931	0.977

4.7 Security analysis

In order to verify the security performance of the algorithm, we use StegExpose(Boehm, 2014), YeNet(Ye et al., 2017a), XuNet(Xu et al., 2016), SRNet(Boroumand et al., 2019) steganalysis algorithms to evaluate the security performance of each quality factor using the original cover image and Stego image

Unedited

or Stego image after the compressed channel. Firstly, we calculate the average error rate Pe value of our method and compare it with the advanced model, as shown in Table 5 In addition, we also compared the AUC with the advanced model, as shown in Fig 10. The results show that our method has good safety performance compared with advanced models.

In addition, since the algorithms used by the S-UNIWARD and J-UNIWARD algorithms are basically the same and their performance is basically the same, repeated comparisons will not be made.

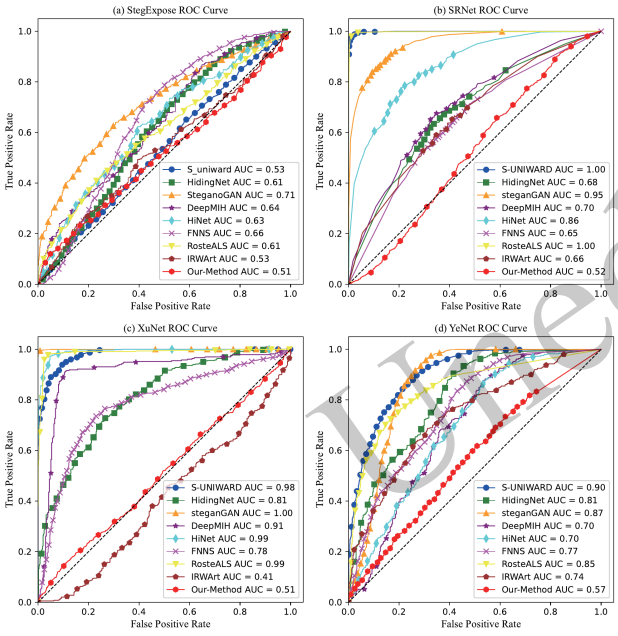

Fig. 10 (a) (b) (c) (d) represent the results under StegExpose, YeNet, XuNet, SRNet steganalysis algorithms, respectively.

Table 6 Ablation experimental results

DCTQT	EFT	KCM	PSNR	SSIM	Error rate	Loss	Pe
✓	✓	✓	35.42	0.964	0.001065	0.00022	0.41
×	✓	✓	33.12	0.877	0.004713	0.00069	0.29
✓	×	✓	34.23	0.897	0.005121	0.00054	0.31
✓	✓	×	**38.85**	**0.976**	**0.000721**	**0.00018**	**0.49**
✓	×	×	35.23	0.927	0.002126	0.00034	0.37

4.8 Ablation experiment

he experimental results of ablation experiments in Table 6, it can be seen that the DCT quantization truncation mechanism improves the visual performance of the model and can significantly reduce the error rate of message extraction. The efficient fusion Transformer can also improve the security of the visual representation set of the model. In addition, the use of key control can further enhance the security of the hidden information, but the reduction of model performance and security after use is not much, within the acceptable range.

5 Conclusions and Future Work

In this work, we propose a novel anti-JPEG compression binary data hiding method based on spatial channel fusion Transformer (SCFformer), aiming to provide a large-capacity and secure data hiding method to achieve high-precision secret information recovery with low quality factor. At the same time, a variable number of secret information hiding is realized. The experimental results show that our method can provide high-precision secret information recovery under different quality factors, and has high security, large capacity and flexibility. This makes our method have broad application prospects and can be used in scenarios such as privacy protection and secret communication. The design of

Unedited

Table 5 Security analysis was performed by Pe.

Methods	Image type	Secret type	QF_{min}	PayloadâĘŚ	$E_R(\%)\downarrow$	Pe→0.5
HidingNet(Baluja, 2020)	natural	image	100	24	92.84	0.03
UDH(Zhang et al., 2020)	natural	image	100	24	79.66	0.01
HiNet(Jing et al., 2021)	natural	image	100	24	43.16	0.02
DeepMIH(Guan et al., 2023)	natural	image	100	24	50.12	0.42
IRWArt(Luo et al., 2023)	natural	image	100	24	0.36	0.45
RDHINN	natural	image	65	8	56.77	0.43
GSN(Wei et al., 2022)	natural	binary	100	1	2.47	0.51
Diffusion-Stego(Kim et al., 2023)	natural	binary	100	1	0.18	0.42
SteganoGAN(Zhang et al., 2019)	natural	binary	100	1	0.01	0.01
FNNS(Kishore et al., 2022)	natural	binary	65	1	0.01	0.34
StegoFormer(Li et al., 2023)	natural	binary	100	0.5	≥**0**	0.46
RoSteALS(Bui et al., 2023)	natural	binary	80	3.00E-03	≥**0**	**0.50**
Our Method	natural	binary	**40**	0.5	≥**0**	0.51

SCFformer can provide some inspiration for further research on hiding a large amount of binary data in JPEG images.

6 Contributors

Xintao Duan: Conceptualization. Chun Li: Conceptualization, Data curation, Formal analysis, Investigation, Project administration, Software, Visualization, Writing review & editing. Bingxin Wei: Investigation, Visualization. Guoming Wu: Validation. Chuan Qin: Conceptualization, Writing âĂŞ review & editing. HaewoonNam:Writing original draft

7 Conflict of interest

All the authors declare that they have no conflict of interest.

8 Data availability

The data that support the findings of this study are available from the corresponding author, upon reasonable request.

Unedited

References

Abood EW, Abdullah AM, Al Sibahe MA, et al., 2022. Audio steganography with enhanced lsb method for securing encrypted text with bit cycling. *Bull Electr Eng Inf BEEI*, 11(1):185-194. https://doi.org/10.11591/eei.v11i1.3279

Baluja S, 2020. Hiding images within images. *IEEE Trans Pattern Anal Mach Intell*, 42(7):1685-1697. https://doi.org/10.1109/TPAMI.2019.2901877

Bhinder P, Singh K, Jindal N, 2018. Image-adaptive watermarking using maximum likelihood decoder for medical images. *Multimed Tools Appl*, 77:10303-10328. https://doi.org/10.1007/s11042-018-5635-z

Boehm B, 2014. Stegexpose-a tool for detecting lsb steganography. https://doi.org/10.48550/arXiv.1410.6656

Boroumand M, Chen M, Fridrich J, 2019. Deep residual network for steganalysis of digital images. *IEEE Trans Inf Foren Secur*, 14(5):1181-1193. https://doi.org/10.1109/TIFS.2018.2871749

Bui T, Agarwal S, Yu N, et al., 2023. Rosteals: Robust steganography using autoencoder latent space. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, p.933-942. https://doi.org/10.1109/CVPRW59228.2023.00100

Chen BJ, Zhou CF, Jeon B, et al., 2018. Quaternion discrete fractional random transform for color image adaptive watermarking. *Multimedia Tools and Applications*, 77(16):20809-20837. https://doi.org/10.1007/s11042-017-5511-2

Guan ZY, Jing JP, Deng X, et al., 2023. Deepmih: Deep invertible network for multiple image hiding. *IEEE Trans Pattern Anal Mach Intell*, 45(1):372-390. https://doi.org/10.1109/TPAMI.2022.3141725

Guo LJ, Ni JQ, Shi YQ, 2012. An efficient jpeg steganographic scheme using uniform embedding. Proc IEEE Int Workshop on Information Forensics and Security, p.169-174. https://doi.org/10.1109/WIFS.2012.6412644

Guo MH, Xu TX, Liu JJ, et al., 2022. Attention mechanisms in computer vision: A survey. *Comput Vis Med*, 8(3):331-368. https://doi.org/10.1007/s41095-022-0271-y

Harish NJ, Kumar BBS, Kusagur A, 2013. Hybrid robust watermarking techniques based on dwt, dct, and svd. *Int J Adv Electr Electron Eng*, 2(5):137-143.

Hayes J, Danezis G, 2017. Generating steganographic images via adversarial training. Proc 31st Int Conf on Neural Information Processing Systems, p.1951-1960.

Holub V, Fridrich J, Denemark T, 2014. Universal distortion function for steganography in an arbitrary domain. *EURASIP J Inf Secur*, 2014:1-13. https://doi.org/10.1186/1687-417X-2014-1

Hore A, Ziou D, 2010. Image quality metrics: Psnr vs. ssim. Proc 20th international conference on pattern recognition, p.2366-2369. https://doi.org/10.1109/ICPR.2010.579

Huang XH, Deng ZF, Li DD, et al., 2023. Missformer: An effective transformer for 2d medical image segmentation. *IEEE Trans Med Imaging*, 42(5):1484-1494. https://doi.org/10.1109/TMI.2022.3230943

Isac B, Santhi V, 2011. A study on digital image and video watermarking schemes using neural networks. *Int J Comput Appl*, 12(9):1-6. https://doi.org/10.5120/1715-2299

Jassim FA, 2013. A novel steganography algorithm for hiding text in image using five modulus method. *Int J Comput Appl*, 72(17):39-44.

Jing JP, Deng X, Xu M, et al., 2021. Hinet: deep image hiding by invertible network. Proc IEEE/CVF Int Conf on Computer Vision, p.4713-4722. https://doi.org/10.1109/ICCV48922.2021.00469

Kim D, Shin C, Choi J, et al., 2023. Diffusion-stego: Training-free diffusion generative steganography via message projection. https://doi.org/10.48550/arXiv.2305.18726

Kishore V, Chen XY, Wang Y, et al., 2022. Fixed neural network steganography: Train the images, not the network. Proc 10th Int Conf on Learning Representations, p. 1.

Li S, Zhang XP, 2019. Toward construction-based data hiding: from secrets to fingerprint images. *IEEE Trans Image Process*, 28(3):1482-1497. https://doi.org/10.1109/TIP.2018.2878290

Li ZZ, Yang XY, Shen KQ, et al., 2023. Adversarial feature hybrid framework for steganography with shifted window local loss. *Neural Netw*, 164:358-369. https://doi.org/10.1016/j.neunet.2023.05.053

Liu Z, Lin YT, Cao Y, et al., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. Proc IEEE/CVF Int Conf on Computer Vision, p.9992-10002. https://doi.org/10.1109/ICCV48922.2021.00986

Lu W, Sun W, Lu HT, 2009. Robust watermarking based on dwt and nonnegative matrix factorization. *Comput Electr Eng*, 35(1):183-188. https://doi.org/10.1016/j.compeleceng.2008.09.004

Lu W, Zhang JH, Zhao XF, et al., 2021. Secure robust jpeg steganography based on autoencoder with adaptive bch encoding. *IEEE Trans Circuits Syst Video Technol*, 31(7):2909-2922. https://doi.org/10.1109/TCSVT.2020.3027843

Luo YJ, Zhou TQ, Liu F, et al., 2023. Irwart: Levering watermarking performance for protecting high-quality artwork images. Proceedings of the ACM Web Conference 2023, p.2340-2348. https://doi.org/10.1145/3543507.3583489

Mallika, Ubhi JS, Aggarwal AK, et al., 2022. Neural style transfer for image within images and conditional gans for destylization. *J Vis Commun Image Represent*, 85:103483. https://doi.org/10.1016/j.jvcir.2022.103483

Mou C, Xu YM, Song JC, et al., 2023. Large-capacity and flexible video steganography via invertible neural network. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.22606-22615. https://doi.org/10.1109/CVPR52729.2023.02165

Shang F, Lan YH, Yang JH, et al., 2023. Robust data hiding for jpeg images with invertible neural network. *Neural Netw*, 163:219-232. https://doi.org/10.1016/j.neunet.2023.03.037

Singh SP, Bhatnagar G, 2018. A new robust watermarking system in integer dct domain. *J Vis Commun Image Represent*, 53:86-101. https://doi.org/10.1016/j.jvcir.2018.03.006

Singh SP, Rawat P, Agrawal S, 2012. A robust watermarking approach using dct-dwt. *IEEE Trans Circuits Syst Video Technol*, 2(8):300-305. https://doi.org/10.1109/TCSVT.2018.2881118

Tao J, Li S, Zhang X, et al., 2018. Towards robust image steganography. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(2):594-600.

Wei P, Li S, Zhang X, et al., 2022. Generative steganography network. Proceedings of the 30th ACM International Conference on Multimedia, p.1621-1629.

Xu GS, Wu HZ, Shi YQ, 2016. Structural design of convolutional neural networks for steganalysis. *IEEE Signal Process Lett*, 23(5):708-712. https://doi.org/10.1109/LSP.2016.2548421

Ye J, Ni JQ, Yi Y, 2017a. Deep learning hierarchical representations for image steganalysis. *IEEE Trans Inf Foren Secur*, 12(11):2545-2557. https://doi.org/10.1109/TIFS.2017.2710946

Ye YX, Shan J, Bruzzone L, et al., 2017b. Robust registration of multimodal remote sensing images based on structural similarity. *IEEE Trans Geosci Remote Sens*, 55(5):2941-2958. https://doi.org/10.1109/TGRS.2017.2656380

Yu XZ, Chen KJ, Wang YF, et al., 2020. Robust adaptive steganography based on generalized dither modulation and expanded embedding domain. *Signal Process*, 168:107343. https://doi.org/10.1016/j.sigpro.2019.107343

Zhang CN, Benz P, Karjauv A, et al., 2020. Udh: Universal deep hiding for steganography, watermarking, and light field messaging. Proc 34th Conf on Neural Information Processing Systems, p.10223-10234.

Zhang KA, Cuesta-Infante A, Xu L, et al., 2019. Steganogan: High capacity image steganography with gans. https://doi.org/10.48550/arXiv.1901.03892

Zhang Y, Luo XY, Yang CF, et al., 2015. A jpeg-compression resistant adaptive steganography based on relative relationship between dct coefficients. Proc 10th Int Conf on Availability, Reliability and Security, p.461-466. https://doi.org/10.1109/ARES.2015.53

Zhang Y, Luo XY, Wang JW, et al., 2021. Image robust adaptive steganography adapted to lossy channels in open social networks. *Inf Sci*, 564:306-326. https://doi.org/10.1016/j.ins.2021.02.058

Zhao ZZ, Guan QX, Zhang H, et al., 2019. Improving the robustness of adaptive steganographic algorithms based on transport channel matching. *IEEE Trans Inf Foren Secur*, 14(7):1843-1856. https://doi.org/10.1109/TIFS.2018.2885438

Zhu JR, Kaplan R, Johnson J, et al., 2018. Hidden: Hiding data with deep networks. Proc 15th European Conf on Computer Vision, p.682-697, https://doi.org/10.1007/978-3-030-01267-0_40

Unedited