

## ORIGINAL RESEARCH

# Hiding image into image with hybrid attention mechanism based on GANs

Yuling Zhu<sup>1,2</sup>  | Yunyun Dong<sup>1,2,3</sup>  | Bingbing Song<sup>2,3</sup> | Shaowen Yao<sup>1,2</sup>

<sup>1</sup>School of Software, Yunnan University, Kunming, China

<sup>2</sup>Engineering Research Center of Cyberspace, Yunnan University, Kunming, China

<sup>3</sup>School of Information Science and Engineering, Yunnan University, Kunming, China

## Correspondence

Yunyun Dong, School of Information Science and Engineering, Yunnan University, Kunming 650000, China.

Email: dongyy929@ynu.edu.cn

## Funding information

Science and Technology Plan in Key Fields of Yunnan Province, Grant/Award Number: 202202AD080002; Fundamental Research Funds for the Central Universities, Grant/Award Number: 2042022kf0021; Youth Project for Basic Research of Yunnan Province Science and Technology Department, Grant/Award Number: 202301AU070194

## Abstract

Image steganography is the art of concealing secret information within images to prevent detection. In deep-learning-based image steganography, a common practice is to fuse the secret image with the cover image to directly generate the stego image. However, not all features are equally critical for data hiding, and some insignificant ones may lead to the appearance of residual artifacts in the stego image. In this article, a novel network architecture for image steganography with hybrid attention mechanism based on generative adversarial network is introduced. This model consists of three subnetworks: a generator for generate stego images, an extractor for extracting the secret images, and a discriminator to simulate the detection process, which aids the generator in producing more realistic stego images. A specific hybrid attention mechanism (HAM) module is designed that effectively fuses information across channel and spatial domains, facilitating adaptive feature refinement within deep image representations. The experimental results suggest that the HAM module not only enhances the image quality during both the steganography and extraction processes but also improves the model's undetectability. Stego images are mixed with varying levels of noise in the training process, which can further improve robustness. Finally, it is verified that the model outperforms current steganography approaches on three datasets and exhibits good undetectability.

## 1 | INTRODUCTION

Steganography is a technique of covert communication that involves concealing secret information within a carrier. This carrier can take various forms, such as text, image, audio, and video. Among these media, image steganography involves concealing secret information within images and is a commonly used steganography method. In image steganography, the sender conceals secret messages within cover images to generate stego images, which the intended receiver retrieves and decodes. This technique ensures that only the sender and receiver are aware of the hidden information, keeping it secure and inaccessible even if intercepted by unauthorized users.

The least significant bit (LSB) method and LSB matching [1, 2] are among the most traditional spatial domain-based techniques in steganography. In a typical LSB algorithm, pixel values of the cover image and secret messages are represented in binary form. The stego image generation process involves replacing the least significant bits of the cover image with the most sig-

nificant bits of secret information. However, the LSB method causes noticeable alterations in the statistical characteristics of the cover image, making it difficult to meet the requirements of covert communication. In contrast, traditional content-adaptive steganography algorithms take a more sophisticated approach. These methods design hand-crafted distortion functions to select embedding locations within the image based on various criteria, such as textural complexity. Examples of these methods include highly undetectable steganography (HUGO), wavelet obtained weights (WOW), spatial universal wavelet relative distortion (S-UNIWARD), and high-pass, low-pass, and low-pass (HILL) [3–6]. Traditional steganographic methods can only hide a small amount of information, prompting the exploration of new steganographic techniques.

In recent years, reversible data hiding has emerged as a critical area of research due to its wide-ranging applications in digital security and communication. Jana et al. implemented various reversible data hiding methods [7–11]. Each of these methods utilized different techniques such as weighted matrices,

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2024 The Author(s). *IET Image Processing* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

hamming codes, image interpolation etc., to achieve the goal of hiding data within images and fully recovering the original image when necessary. Debasis et al. proposed a dual-image-based reversible data hiding scheme utilizing three pixel value differences and difference expansion [12]. Meikap et al. utilized generalized the directional pixel value overlapping (DPVO) technique for reversible data hiding [13]. Chowdhuri et al. proposed two different data hiding methods. One method involves embedding data into dual-color images in a reversible manner using a weighted matrix [14]. This technique utilizes the weighted matrix to determine embedding positions within the images and modify pixel values accordingly. The other method is a weighted matrix based steganographic scheme for highly compressed color image through discrete cosine transform (DCT) to maintain a good balance between payload and imperceptibility [15]. In this scheme, the color images undergo high compression, followed by the embedding of data into the compressed images using weighted matrices. Pal et al. introduced three different digital image watermarking schemes. The first method used weighted matrices to determine embedding positions within the image and embeds the watermark into the image while ensuring data reversibility [16]. The second method utilized local binary pattern (LBP) and dual images for image authentication and tamper detection [17]. The third method combined local binary pattern, Lagrange interpolation polynomial, and weighted matrix techniques [18]. This approach enables the embedding of a watermark into the color image while preserving both data reversibility and security. Mukherjee et al. utilized difference expansion for achieving high-capacity reversible data hiding [19]. Biswapati et al. employed weighted matrices and image interpolation for reversible data hiding [20]. Singh et al. achieved robust reversible data hiding through super pixel segmentation, Arnold transform, discrete cosine transform, and cellular automata (CA) [21]. The above methods provide diverse technical solutions in the field of reversible data hiding. However, they also have limitations, such as potential capacity constraints when dealing with large amounts of data and significant loss of image details, resulting in lower image quality.

Generative adversarial networks (GAN) [22] have offered new inspiration for image steganography. GANs typically consist of two adversarial subnetworks: the generator aims to generate data distribution approximate to the real data distribution, while the discriminator tries to distinguish between real and generated samples. The basic principle of GANs is to leverage adversarial learning to enhance the respective performance of these two subnetworks. The ultimate goal of data hiding is to make hidden data imperceptible to the detector, and this necessitates an adversarial relationship in data hiding applications. This confrontation naturally resembles the relationship between steganography and steganalysis, implying the rationality of using GAN for developing steganographic algorithms. Yu et al. [23] integrated the evaluation metrics of secure data hiding with the latest GAN principle and proposed a novel end-to-end framework. It utilized a spatial attention model to generate a mask identifying less sensitive regions in the cover images for concealing secret information. However, the convolution

layer's edge detection and texture extraction capabilities limit the effectiveness of the spatial attention model. Similarly, Cong et al. [24] introduced a new steganography without embedding based on the attention-GAN model. It utilized the self-attention mechanism to capture internal correlations between image patches. However, since the embedding operation takes into account holistic image information, it makes minimal direct contributions to steganography. Furthermore, the study in [25] introduced a novel end-to-end network architecture for image steganography, incorporating channel attention mechanisms based on generative adversarial networks. However, using dimensionality reduction for channel attention has a negative impact on prediction, and capturing dependencies across all channels is inefficient and unnecessary.

To address the aforementioned challenges, we propose a hybrid attention mechanism (HAM) based end-to-end network architecture for image steganography with GAN. Different from these methods in [23–25]. Our approach integrates the hybrid attention mechanism (HAM) model, which combines both channel attention modules and spatial attention modules. The channel attention module embeds the secret information by fusing it with multi-channel feature maps in the feature map. As the number of meaningful features varies across channels, channel attention allocates varying levels of attention to each channel, preventing irrelevant features from compromising stego quality. Simultaneously, the spatial attention contributes to extracting crucial information from different positions in space, enabling more effective concealment of secret information in the image and reducing the risk of detection. Compared with the previous work, this article makes the following novel contributions:

- We propose a novel end-to-end network architecture for image steganography with an HAM based on GANs.
- We introduce a novel HAM module to enhance the quality of generated stego images and improve the accuracy of secret information recovery.
- Our network architecture learns to resist various attacks (dropout, crop, compression etc.).
- We optimize the number and location of HAM additions to ensure model reliability while enhancing important features.
- Extensive experiment results show that the proposed method achieve excellent performances on image hiding, compared to other methods.

## 2 | RELATED WORK

Steganography is typically understood as the practice of concealing information within other media, and these methods can be categorized into three types.

### 2.1 | Traditional steganography algorithm

LSB [1] is the most traditional spatial domain based method in steganography. In contrast, traditional content adaptive

steganography algorithms design the hand-crafted distortion functions which are used for selecting the embedding localization of the image. For example, WOW [4] embeds information into the cover image according to the textural complexity of regions, while HUGO [3] defines a distortion function domain by assigning costs to pixels according to the effect of embedding information within a pixel. Additionally, HILL [6] introduced a new cost function for spatial image steganography. These methods are more robust and undetectable than LSB, but they can only hide bit-level information.

## 2.2 | Deep learning based steganography

Hayes et al. proposed the HayesGAN [26], utilized coding-decoding networks to hide text information within images. However, during real-world applications, images are susceptible to various noise attacks, making it challenging for this models to extract secret information reliably. Zhu et al. proposed the HiDDeN image steganography framework [27], this model introduced a noise layer between the coding network and the decoding network to simulate potential noise attacks and improve the model's resistance to noise interference. However, due to the network structure design, there is still significant room for improvement in the model's embedding capacity. Baluja et al. [28] introduced a neural network-based steganographic model for the novel task of embedding color images into color images. However, this model fails to meet the demand for high-capacity information hiding. Weng et al. [29] successfully implemented video steganography through temporal residual modeling. Additionally, Fu et al. proposed a HISGAN [30], incorporating a coding structure similar to U-Net and integrating GAN concepts. Zhang et al. [31] leveraged the fact that the Y channel of the YUV color space lacks color information and embedded grayscale images into the Y channel to mitigate color distortion in stego images. Subsequently, Chen et al. [32] chose the B channel as the steganography channel in the RGB color space, presented a new steganography framework. However, the stego images generated by these methods tend to exhibit reduced quality and susceptibility to detection by steganographic analyzers such as XuNet [33], YeNet [34], SRNet [35], and others. A novel method of coverless information hiding was proposed in [36], involving the construction and training of an improved Wasserstein GAN model using both disguised and secret images. Furthermore, Chen et al. proposed a novel coverless information hiding method named StarGAN [37], based on image selection, to overcome the limitations of existing coverless information hiding techniques, which are constrained by the size of the image database. Additionally, in [38], a coverless image steganographic scheme with high capacity was proposed, enabling the concealment of a color secret image into a color stego image of the same size, without the need for cover images. These methods typically feature strong hiding security by avoiding cover modification operations. However, these coverless methods tend to have lower hiding capacities.

## 2.3 | Steganography based on GANs with attention mechanism

Yu et al. [23] proposed a method that introduces an spatial attention model to generate an attention mask, which helps create a higher-quality target image. However, the effectiveness of the spatial attention model is limited due to the inherent edge detection and texture extraction capabilities of the convolution layer. In [24], the self-attention mechanism is employed to capture internal correlations among image patches. However, as the embedding operation considers holistic image information, it makes minimal direct contributions to steganography. Furthermore, Tan et al. [25] proposed a novel end-to-end network architecture for image steganography, integrating channel attention mechanisms based on GANs. However, the use of dimensionality reduction to obtain channel attention has a detrimental effect on channel attention prediction. Zhang et al. introduced the adaptive frequency-domain channel attention network (AFcaNet) [39], which utilized a fine-grained manner of assigning weights to fully exploit the frequency features in each channel. Nevertheless, this method does not significantly enhance steganographic capacity. Lu et al. introduced the deep adaptive hiding network (DAH-Net) [40], which is designed to systematically extract and combine essential secret and cover information across different frequency levels and network layers. The method proposes a module called the Adaptive Frequency-Domain Channel Attention Network (AFcaNet), which makes full use of the frequency features in each channel by finely assigning weights. However, the method doesn't greatly improve the quality of stego images. To the best of our knowledge, there is no work to explore hybrid attention mechanism in image steganography task.

## 3 | PROPOSED METHOD

In this section, we will first provide an overview of the model architecture and basic ideas. Following that, we will provide a detailed of each component of our model, and illustrate the loss functions.

### 3.1 | The proposed network structure

In the proposed framework, the process unfolds in several stages. First, the secret image ( $I_S$ ) and cover image ( $I_C$ ) serve as input to the generator. Subsequently, the generator generates a stego image ( $I_{C'}$ ) containing the secret image information. The discriminator is responsible for scoring the input cover image and stego image. Finally, the extractor aims to decode the secret image ( $I_S'$ ) from the stego image. Trainable parameters  $\theta_G$ ,  $\theta_E$ , and  $\theta_D$  correspond to the generator, extractor, and discriminator, respectively. The generator's output is denoted as  $G(\theta_G; I_C, I_S)$ , produced using the cover image ( $I_C$ ) and secret image ( $I_S$ ) as inputs. Similarly,  $E(\theta_E; I_{C'})$  represents the output of the extractor, with  $I_{C'}$  as its input. Furthermore,

$D(\theta_D; I_C, I_{C'})$  signifies the output of the discriminator, taking both  $I_C$  (from the real dataset) and  $I_{C'}$  (generated stego images) as input.

### 3.1.1 | Generator

The network structure of the stego image generative model generator includes nine convolution layers (kernel size = 3, stride = 1, pad = 1), two convolution layers (k = 4, s = 2, p = 1), another convolution layer (k = 1, s = 1, p = 0), a deconvolution layer (k = 4, s = 2, p = 1), and two HAM modules. Typically, within the generator's network structure, each convolution and deconvolution layer follows with an instance normalization layer and a ReLU layer. However, the final layer of the generator network differs as it employs a Tanh activation layer. This setup ensures that the generator's output values are appropriately normalized and fall within the desired range, thereby aiding in the generation of high-quality stego images.

### 3.1.2 | Extractor

The network structure of the extractor model includes eight convolution layers (kernel size = 3, stride = 1, pad = 1), and two HAM modules. Each convolution layer follows with an instance normalization layer and a ReLU activation layer, except that the last layer, which employs a Tanh activation layer.

### 3.1.3 | Discriminator

The network structure of the discriminator model includes three convolution layers (kernel size = 3, stride = 1, pad = 1), a average pooling layer and a linear layer. The primary objective of the discriminator is to assign higher scores to stego images and lower scores to cover images. Moreover, to improve the convergence performance, we use Adam optimizer. It is computationally efficient and has little memory requirements. The hyper-parameters of Adam optimizer are:  $\beta_1 = 0.5$ ,  $\beta_2 = 0.999$ . The base learning rate is 0.0001.

## 3.2 | The proposed HAM

The attention mechanism in deep learning enables the network to learn and focus on important features while disregarding irrelevant ones. In image steganography, it is crucial to suppress unimportant information in the secret image to prevent artifacts in the generated stego image while also identifying suitable hiding positions within the cover image for the secret information. To address these requirements, we propose a new HAM module that effectively implements image steganography. The structure of the hybrid attention module and the computation process of each attention map is depicted in the lower part of

Figure 1. Given an intermediate feature map  $F \in R^{C \times H \times W}$  as input, the HAM sequentially infers a 1D channel attention map  $M_c \in R^{C \times 1 \times 1}$  and a 2D spatial attention map  $M_s \in R^{1 \times H \times W}$ . The overall attention process can be summarized as,

$$\begin{aligned} F' &= M_c(F) \otimes F, \\ F'' &= M_s(F') \otimes F', \end{aligned} \quad (1)$$

where  $\otimes$  denotes element-wise multiplication, and  $F''$  is the final refined output. The following describes the details of each attention module.

### 3.2.1 | Channel attention module

We first aggregate the spatial information of a feature map by using the average pooling and max-pooling operations, generating two different spatial context descriptors,  $F_{avg}^c$  and  $F_{max}^c$ , which denote average-pooled features and max-pooled features, respectively. Both descriptors are then forwarded to a 1D convolution to produce our channel attention map  $M_c \in R^{C \times 1 \times 1}$ . In short, the channel attention is computed as,

$$\begin{aligned} M_c(F) &= \sigma(C1D_k(AvgPool(F)) + C1D_k(MaxPool(F))) \\ &= \sigma(C1D_k(F_{avg}^c) + C1D_k(F_{max}^c)), \end{aligned} \quad (2)$$

where  $C1D$  indicates a 1D convolution with  $k = 3$ , and  $\sigma$  denotes the sigmoid function.

### 3.2.2 | Spatial attention module

We generate a spatial attention map by utilizing the inter spatial relationship of features. Differently from the channel attention, the spatial attention focuses on 'where' is an informative part, which is complementary to the channel attention. To compute the spatial attention, we first apply the average-pooling and max-pooling operations along the channel axis and then concatenate them to generate an efficient feature descriptor. We aggregate the channel information of a feature map using two pooling operations, generating two 2D maps,  $F_{avg}^s \in R^{1 \times H \times W}$  and  $F_{max}^s \in R^{1 \times H \times W}$ , which denotes average-pooled features and max-pooled features across the channel, respectively. These features are then concatenated and convolved by a standard convolution layer, producing our 2D spatial attention map. In short, the spatial attention is computed as,

$$\begin{aligned} M_s(F) &= \sigma(f^{(7 \times 7)}([AvgPool(F); MaxPool(F)])) \\ &= \sigma(f^{(7 \times 7)}([F_{avg}^s; F_{max}^s])), \end{aligned} \quad (3)$$

where  $\sigma$  denotes the sigmoid function, and  $f^{7 \times 7}$  represents a convolution operation with the filter size of  $7 \times 7$ .



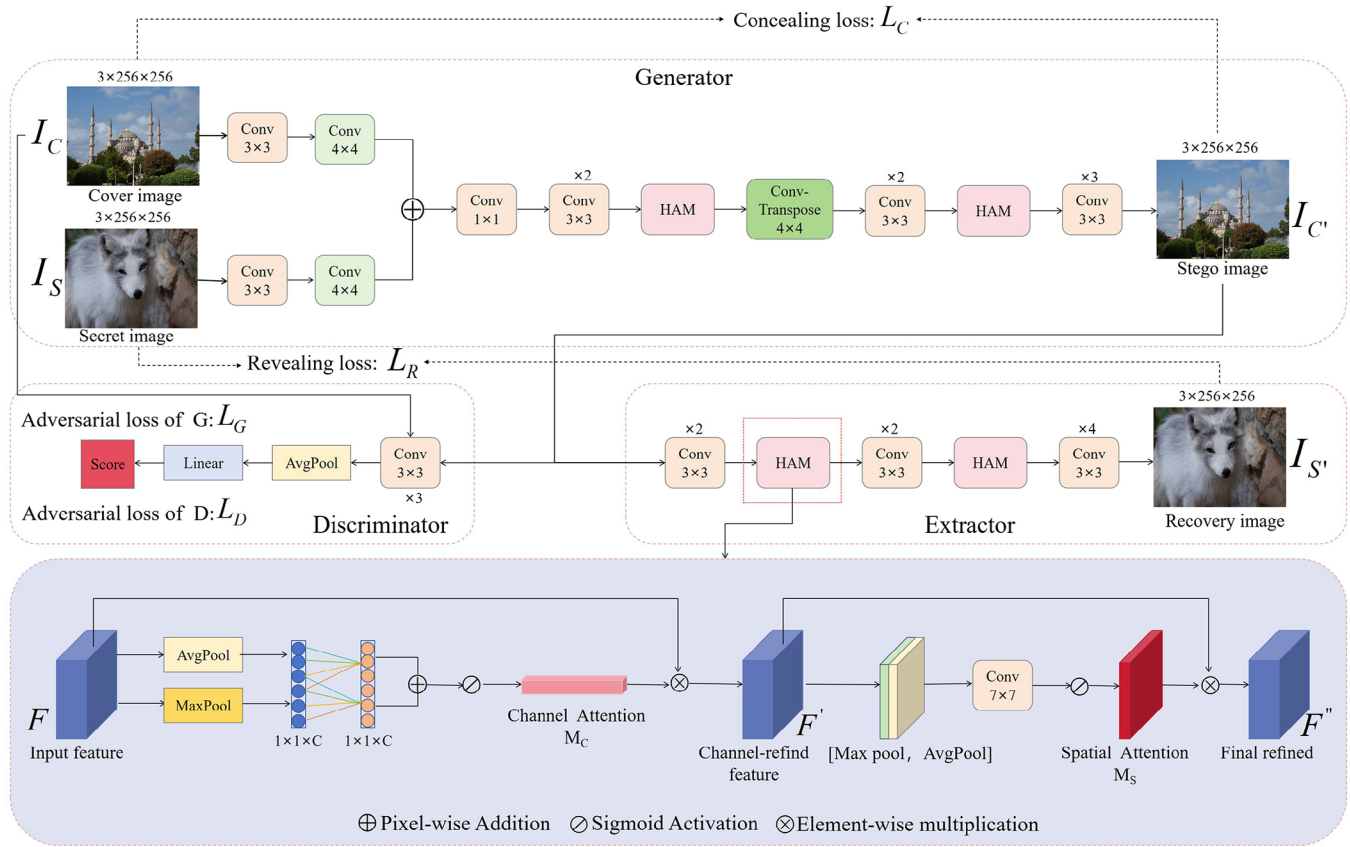


FIGURE 1 The overall architecture of the proposed method.

### 3.3 | Loss function

The total loss function consists of the concealing loss, revealing loss, and adversarial loss. Next, we introduce these losses in detail.

**Concealing loss.** The purpose of the generator is to hide secret image into cover image to generate a stego image. The cover image and stego image generated by the generator should be as similar as possible to ensure that the hidden secret image is invisible. To achieve this goal, we define the concealing loss  $L_C$  as follows,

$$L_C = \frac{1}{CHW} \|I_C - I_{C'}\|_2^2, \quad (4)$$

where  $I_C$  and  $I_{C'}$  represent the cover image and stego image, respectively, and the shape of cover image and stego image is  $R^{C \times H \times W}$ .

**Revealing loss.** The reconstructed secret image extracted by the extractor should be the same as the secret image. For this purpose, the concealing loss  $L_R$  is defined as follows,

$$L_R = \frac{1}{CHW} \|I_S - I_{S'}\|_2^2, \quad (5)$$

where  $I_S$  and  $I_{S'}$  represent the secret image and extracted secret image, respectively, and the shape of the cover image and stego image is  $R^{C \times H \times W}$ .

**Adversarial loss.** In general, generating of an image is a process of entropy reduction, and information recovery is a task of keeping entropy unchanged. Competition arises between these two types of tasks, necessitating a compromise. The discriminant network needs to distinguish as much as possible whether the input data are cover images or stego images.  $L_G$  represents the effectiveness of the discriminator in detecting stego images, while the discriminator incurs a classification loss, denoted as  $L_D$ , based on its predictions. Therefore, we optimize the following losses for the generator and discriminator,

$$L_D(I_C, I_{C'}) = \log(1 - D(I_C)) + \log(D(I_{C'})), \quad (6)$$

$$L_G(I_{C'}) = \log(1 - D(I_{C'})), \quad (7)$$

where  $I_C$  and  $I_{C'}$  represent the cover image and stego image, respectively.

**Total loss.** The total loss function  $L_{total}$  is a weighted sum of concealing loss  $L_C$ , revealing loss  $L_R$ , and adversarial loss  $L_G$ , as follow,

$$L_{total} = \lambda_1 L_C + L_R + \lambda_2 L_G, \quad (8)$$

where  $\lambda_1$  and  $\lambda_2$  are weights for balancing different loss terms.

## 4 | EXPERIMENTS

### 4.1 | Experimental settings

#### 4.1.1 | Datasets and settings

For our experiments, we use the following datasets: ImageNet [41] (15,000 training images, 5,000 validation images), COCO [42] (15,000 training images, 5,000 validation images), and DIV2K [43] (900 training images, 100 validation images). Additionally, the testing datasets include 200 images from DIV2K, 200 images from ImageNet, and 200 images from the COCO dataset. All images are resized to  $256 \times 256$  and trained the models with Adam optimization algorithm. The generated images are  $256 \times 256$ . Due to the common use of color images in real life, we train our models with RGB cover and secret images, both sized at  $3 \times 256 \times 256$  pixels. We have set the initial learning rate to 0.0001, the batch size to 16, and the parameters  $\lambda_1$  and  $\lambda_2$  to 0.7 and 0.001, respectively. We use PyTorch running in GPU mode to calculate the gradient. During the training phase, the model runs on NVIDIA 4090 GPU and 18GB of memory.

#### 4.1.2 | Benchmarks

To verify the effectiveness of our model, we compare it with several state-of-the-art (SOTA) image hiding methods, including HiDDeN [27], Weng et al [29], Baluja [28], and DAH-Net [40]. For fair comparison, we re-trained the models of HiDDeN [27], Weng et al. [29], Baluja [28], and DAH-Net [40] using the same training dataset as ours.

#### 4.1.3 | Evaluation metrics

We evaluate the performance of our model using four objective evaluation metrics: peak signal to noise ratio (PSNR), structural similarity index measure (SSIM), mean absolute error (MAE), and root mean square error (RMSE). In addition, we use the statistical steganalysis tool named SiaStegNet [44] and SRNet [35] to evaluate the security performance of our method.

- (1) PSNR: PSNR is commonly used as an objective measure of image quality that quantifies the level of distortion in an image by calculating the error between corresponding pixels [45]. A higher PSNR indicates the higher quality of the stego image. The PSNR is calculated by,

$$\text{PSNR} = 10 \log_{10} \left( \frac{I_{\max}^2}{MSE} \right), \quad (9)$$

where  $MSE$  denotes the mean square error between the original image and the evaluated image,  $I_{\max}^2$  is the maximum pixel value of the image.  $I_{\max}^2$  the largest is 255 for grayscale images, for binary image  $I_{\max}^2$  the largest is 1.

- (2) SSIM: It is an image quality assessment metric used to measure the structural similarity between two images [46]. For a

pair of images, denoted as  $X$  and  $Y$ , SSIM can be defined by the mean  $\mu_X$  and  $\mu_Y$ , the variance  $\sigma_X^2$  and  $\sigma_Y^2$ , and the covariance  $\sigma_{XY}^2$  of these two images.

$$\text{SSIM}(X, Y) = \frac{(2\mu_X\mu_Y + k_1R)(2\sigma_{XY} + k_2R)}{(\mu_X^2 + \mu_Y^2 + k_1R)(\sigma_X^2 + \sigma_Y^2 + k_2R)}, \quad (10)$$

where the SSIM is usually calculated by the default setting  $k_1 = 0.01$  and  $k_2 = 0.03$ , and then by obtaining a range of values  $\text{SSIM} \in [0, 1]$ , where 1 means that the images are identical.

- (3) MAE: When considering two images denoted as  $X$  and  $Y$ , the MAE signifies the average absolute difference between the pixel values of the two images. A small MAE suggests that the disparity between the images is minimal as well. The MAE is calculated by,

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |Y_i - X_i|, \quad (11)$$

where  $n$  is the number of bits of the image.

- (4) RMSE: When provided with two images, denoted as  $X$  and  $Y$ , the RMSE is a metric that gauges the average magnitude of disparities between corresponding pixel values in the two images. As a result, a lower RMSE indicates a greater similarity between the images. The RMSE is calculated by,

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (Y_i - X_i)^2}, \quad (12)$$



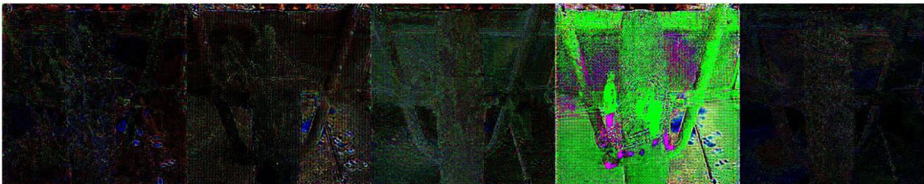
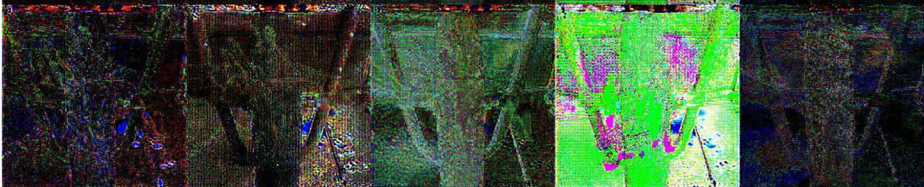


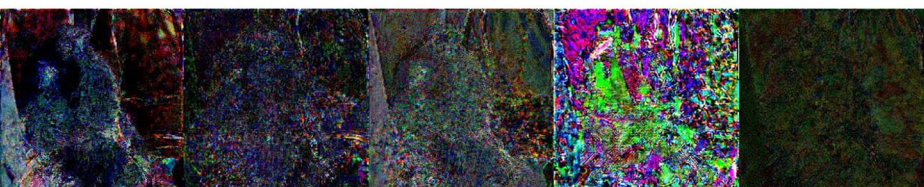
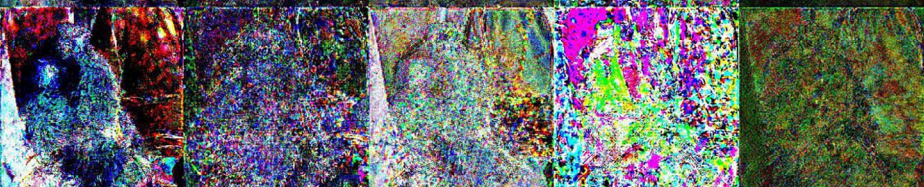
where  $n$  is the number of bits of image.

## 4.2 | Comparison against SOTA methods

### 4.2.1 | Qualitative evaluation

The experimental results, obtained by applying various training models and randomly selecting images from the ImageNet datasets, are depicted in Figure 2. In Figure 2, we present visual representations of various image components, including cover images, stego images, secret images, and recovered secret images, along with the corresponding enlarged residual images. The first row showcases the cover images, the second row depicts the stego images, and the third and fourth rows display the enlarged cover residual images at 20x and 50x magnification, respectively. The stego images are generated by combining the cover image with the secret image through the steganography process. In terms of visual effects, the generated stego images closely resemble the original cover images, maintaining a similar appearance. Even after magnifying the residual image by 50 times, we observe no significant contour information, demonstrating the security of our method in not revealing confidential information. The fifth row displays the secret images intended for concealment, while the sixth row represents the recovered



	HiDDen	Weng et al.	Baluja	DAH-Net	Ours
cover image					
stego image					
(PSNR/SSIM)	(38.13/0.96) (39.36/0.97) (40.59/0.98) (29.80/0.93) <b>(44.90/0.99)</b>				
(cover -stego)*20					
(cover-stego)*50					
secret image					
reconstructed secret image					
(PSNR/SSIM)	(35.36/0.96) (38.51/0.96) (36.24/0.97) (30.88/0.90) <b>(43.07/0.99)</b>				
(secret-re_secret)*20					
(secret-re_secret)*50					

**FIGURE 2** Visual comparisons of stego and recovery images of our method and the comparison methods HiDDen [27], Weng et al [29], Baluja [28], and DAH-Net [40]. The upper four rows show the cover image, stego image, enlarged cover residual images, while the lower four rows show the secret image, recovery secret image, and enlarged secret residual images of different methods.

**TABLE 1** Benchmark comparisons on different datasets, with the best results indicated in bold and the second bests indicated with an underline.

Steganography methods	ImageNet							
	Cover/stego image pair				Secret/re-secret image pair			
	PSNR $\uparrow$	SSIM $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$
HiDDen [27]	37.351	0.948	2.655	3.490	36.317	0.954	3.061	3.921
Weng et al. [29]	38.431	0.947	2.375	3.079	<u>38.511</u>	<u>0.973</u>	2.279	<u>3.048</u>
Baluja [28]	<u>38.978</u>	<u>0.975</u>	<u>2.288</u>	<u>2.952</u>	38.001	0.970	2.427	3.232
DAH-Net [40]	38.184	0.969	2.404	3.142	38.035	0.921	<u>1.819</u>	3.197
Ours	<b>43.710</b>	<b>0.990</b>	<b>1.227</b>	<b>1.668</b>	<b>42.308</b>	<b>0.990</b>	<b>1.482</b>	<b>1.960</b>
Methods	COCO							
	Cover/stego image pair				Secret/re-secret image pair			
	PSNR $\uparrow$	SSIM $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$
HiDDen [27]	37.270	0.951	2.636	3.512	37.410	0.970	2.651	3.470
Weng et al. [29]	38.106	0.959	2.419	3.188	<u>38.707</u>	0.975	<u>2.184</u>	<u>2.986</u>
Baluja [28]	<u>38.577</u>	<u>0.978</u>	<u>2.388</u>	<u>3.059</u>	38.537	<u>0.977</u>	2.246	3.057
DAH-Net [40]	37.323	0.963	2.684	3.470	38.203	0.975	2.295	3.315
Ours	<b>41.908</b>	<b>0.988</b>	<b>1.580</b>	<b>2.054</b>	<b>41.480</b>	<b>0.989</b>	<b>1.612</b>	<b>2.152</b>
Methods	DIV2K							
	Cover/stego image pair				Secret/re-secret image pair			
	PSNR $\uparrow$	SSIM $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$
HiDDen [27]	37.177	0.951	<u>2.716</u>	<u>3.547</u>	37.302	0.976	2.614	3.495
Weng et al. [29]	37.024	0.938	2.771	3.608	<u>38.434</u>	<u>0.979</u>	<u>2.079</u>	<u>3.053</u>
Baluja [28]	<u>37.361</u>	<u>0.976</u>	2.792	3.549	37.673	0.973	2.523	3.358
DAH-Net [40]	34.591	0.954	3.676	4.752	32.344	0.933	4.548	6.156
Ours	<b>40.447</b>	<b>0.989</b>	<b>1.926</b>	<b>2.444</b>	<b>41.233</b>	<b>0.988</b>	<b>1.505</b>	<b>2.035</b>

secret images. These recovered secret images are obtained from the stego image through the extraction process. In terms of visual quality, the recovered secret images recover nearly all of the semantic content present in the original secret images. This result indicates that the steganography process successfully embeds and preserves the essential information of the secret images, allowing for reliable extraction and reconstruction during the decoding phase. Moreover, the high fidelity of the reconstructed secret images demonstrates the effectiveness of the proposed method in preserving the integrity of the hidden information.

#### 4.2.2 | Quantitative evaluation

The experimental results demonstrating the superiority of our method in terms of image quality are presented in Table 1. This table compares the numerical results of our proposed network with other existing methods, namely HiDDen [27], Weng et al. [29], Baluja [28], and DAH-Net [40]. As shown in Table 1, our method outperforms other methods significantly in terms of the four metrics for both cover/stego and

secret/recovery secret pairs. Specifically, for the cover/stego image pairs, our approach outperforms the second best results by 4.732-dB on the ImageNet, by 3.331-dB on COCO, and by 3.086-dB on DIV2K in terms of PSNR. For the secret/recovery image pairs, we achieve a 3.797-dB improvement in PSNR over the second best results on ImageNet, a 2.773-dB improvement on COCO, and a 2.799-dB improvement on DIV2K. The observed enhancement in image quality can be attributed to the integration of the HAM across the spatial and channel domains within our proposed method. By employing this mechanism, our model effectively captures and utilizes relevant information from both the cover and secret images during the generation process, resulting in stego images of superior quality and ensuring better retention of concealed information. Our method compresses and distributes secret image pixel information on all available bits of the stego image, with a relative capacity of 1 byte/pixel. Stego images of superior quality typically exhibit a PSNR value of 40 dB or higher, while values below 30 dB suggest lower quality. Our method consistently achieves PSNR values on the ImageNet, COCO, and DIV2K datasets that exceed 40, reaching as high as 43.71 on ImageNet.



**TABLE 2** Effectiveness of HAM module; the second row represents our method.

HAM	Cover/stego image pair				Secret/re-secret image pair				Detection rate (%)
	PSNR $\uparrow$	SSIM $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$	PSNR $\uparrow$	SSIM $\uparrow$	MAE $\downarrow$	RMSE $\downarrow$	
$\times$	37.159	0.961	2.857	3.581	37.190	0.970	2.699	3.570	88.96
$\checkmark$	<b>42.022</b>	<b>0.989</b>	<b>1.577</b>	<b>2.055</b>	<b>41.673</b>	<b>0.989</b>	<b>1.533</b>	<b>2.049</b>	<b>59.27</b>

**TABLE 3** Performances of image hiding under different image distortions.

Method	Cropout	Dropout	JPEG-compression	Quantization	Resize	Crop 10%
HiDDen [27]	31.569	31.569	11.499	31.569	11.459	13.573
Weng et al. [29]	30.148	30.148	11.421	30.148	11.330	13.558
Baluja [28]	34.023	34.023	11.460	34.023	11.385	14.857
DAH-Net [40]	34.373	34.373	12.513	34.373	12.144	13.698
Ours	<b>37.484</b>	<b>37.482</b>	<b>16.107</b>	<b>37.482</b>	<b>16.929</b>	<b>18.261</b>

### 4.3 | Ablation study

As shown in Table 2, the HAM module plays a crucial role in enhancing the performance of our approach. Specifically, the inclusion of the HAM leads to a noteworthy 4.863-dB improvement in PSNR for the cover/stego image pair and a remarkable 4.483-dB enhancement for the secret image/recovery pair. One possible reason for this result is that our HAM module, which integrates both channel attention modules and spatial attention modules, can learn what and where to emphasize or suppress, effectively refining intermediate features. In the process of image steganography, irrelevant channel features may leave artifacts on the stego image. The channel attention mechanism suppresses irrelevant channel information to prevent artifacts on the stego image, thereby enhancing its quality. Simultaneously, finding suitable embedding positions for secret image is crucial during information hiding. Spatial attention assigns different weights to positions in the image according to their importance, facilitating better embedding of secret information and reducing steganalysis detection rates. Therefore, our attention mechanism module demonstrates excellent performance in terms of image generation quality, accuracy of secret image extraction, and resistance to steganalysis.

### 4.4 | Model robustness

To evaluate the robustness of the stego images generated by our model, we apply various distortions on the stego images: Cropout ( $p = 0.3$ ), Dropout ( $p = 0.3$ ), JPEG-Compression ( $Q = 50$ ), Quantization, Resize ( $p = 0.4, q = 0.6$ ) and Crop 10% (indicating random cropping of 10% of the images). We then extract secret images from the stego images subjected to different image distortions and calculate PSNR values, as presented in Table 3. From the Table 3, it is evident that the stego images generated by our method manage to maintain good quality for secret image recovery under various distortions. This observation underscores the robustness of our approach.

**TABLE 4** The detection accuracy using different steganalysis methods.

Methods	SiaStegNet Accuracy(%)	SRNet Accuracy(%)
HiDDen [27]	81.77	89.58
Weng et al. [29]	77.60	89.06
Baluja [28]	72.92	79.17
DAH-Net [40]	87.50	81.53
Ours	<b>59.27</b>	<b>82.81</b>

### 4.5 | Security analysis

SiaStegNet [44] and SRNet [35] are two network designed for image steganalysis, with the objective of discriminating between stego images and cover images. Table 4 illustrates the detection accuracy achieved by SiaStegNet and SRNet across various hiding methods. The image hiding algorithm performs better when the detection accuracy is closer to 50% (random guess). Our method achieves a detection accuracy of 59.27% with SiaStegNet, a result close to 50%. This suggests that the stego images produced by our method are nearly indistinguishable from the cover images.

## 5 | CONCLUSION

In this paper, we propose a hybrid attention mechanism based end-to-end network architecture for image steganography with GAN. Inside the model, the generator is used to generate stego images, the extractor is used to extract secret image, and the discriminator is used to enhance model security. Our approach represents a substantial improvement in both the quality of stego images and the accuracy of secret image recovery. Specifically, we design a HAM to guide both the current hiding and extraction processes, with the aim of enhancing both the image quality and its undetectability. Extensive experimental results

demonstrate that our method can achieve high invisibility in image hiding, significantly outperforming other state-of-the-art methods in both quantitative and qualitative aspects. It also maintains a reduced detection rate under two steganalysis models, SiaStegNet and SRNet. Additionally, the mixed training dataset with noisy stego images improves the robustness of our model.

In the future, we hope to explore extending this method to other multimedia domains such as video and audio, enabling a broader range of information hiding and protection.

## AUTHOR CONTRIBUTIONS

**Yuling Zhu:** Conceptualization; methodology; writing—original draft. **Yunyun Dong:** Supervision; validation. **Bingbing Song:** Data curation. **Shaowen Yao:** Validation.

## ACKNOWLEDGEMENTS

This work is supported by the Youth Project for Basic Research of Yunnan Province Science and Technology Department (No. 202301AU070194), the Fundamental Research Funds for the Central Universities (No. 2042022kf0021), and the Science and Technology Plan in Key Fields of Yunnan Province (No. 202202AD080002).

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interest.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available in [ImageNet] at [https://cocodataset.org/\[doi\]](https://cocodataset.org/[doi]), reference number [28], [COCO] at <https://cocodataset.org/>, reference number [29], and [DIV2K] at <https://data.vision.ee.ethz.ch/cvl/DIV2K/>, reference number [30].

## ORCID

Yuling Zhu  <https://orcid.org/0009-0004-3147-8602>

Yunyun Dong  <https://orcid.org/0009-0004-5251-9928>

## REFERENCES

- Zhang, H., Zhu, C.: Novel lsb steganography algorithm of against statistical analysis. *Comput. Eng.* 34(23), 144–146 (2008)
- Li, X., Yang, B., Cheng, D., Zeng, T.: A generalization of lsb matching. *IEEE Signal Process Lett.* 16(2), 69–72 (2009)
- Pevný, T., Filler, T., Bas, P.: Using high-dimensional image models to perform highly undetectable steganography. In: *Information Hiding: 12th International Conference, IH 2010, Revised Selected Papers 12*, pp. 161–177. Springer, Cham (2010)
- Holub, V., Fridrich, J.: Designing steganographic distortion using directional filters. In: *2012 IEEE International workshop on information forensics and security (WIFS)*, pp. 234–239. IEEE, Piscataway (2012)
- Holub, V., Fridrich, J., Denemark, T.: Universal distortion function for steganography in an arbitrary domain. *EURASIP J. Inf. Sec.* 2014, 1–13 (2014)
- Li, B., Wang, M., Huang, J., Li, X.: A new cost function for spatial image steganography. In: *2014 IEEE International conference on image processing (ICIP)*, pp. 4206–4210. IEEE, Piscataway (2014)
- Jana, B.: Dual image based reversible data hiding scheme using weighted matrix. *Int. J. Electron. Inf. Eng.* 5(1), 6–19 (2016)
- Jana, B.: High payload reversible data hiding scheme using weighted matrix. *Optik* 127(6), 3347–3358 (2016)
- Jana, B., Giri, D., Mondal, S.K.: Partial reversible data hiding scheme using (7, 4) hamming code. *Multim. Tools Appl.* 76, 21691–21706 (2017)
- Jana, B., Giri, D., Mondal, S.K.: Dual image based reversible data hiding scheme using (7, 4) hamming code. *Multim. Tools Appl.* 77, 763–785 (2018)
- Jana, B.: Reversible data hiding scheme using sub-sampled image exploiting lagrange's interpolating polynomial. *Multim. Tools Appl.* 77(7), 8805–8821 (2018)
- Debasis, G., Biswapati, J., Kumar, M.S.: Dual image based reversible data hiding scheme using three pixel value difference expansion. In: *Information Systems Design and Intelligent Applications: Proceedings of Third International Conference INDIA 2016*, vol. 2, pp. 403–412. Springer, Singapore (2016)
- Meikap, S., Jana, B.: Directional pvo for reversible data hiding scheme with image interpolation. *Multim. Tools Appl.* 77(23), 31281–31311 (2018)
- Chowdhuri, P., Jana, B.: Hiding data in dual color images reversibly via weighted matrix. *J. Inf. Sec. Appl.* 50, 102420 (2020)
- Chowdhuri, P., Jana, B., Giri, D.: Secured steganographic scheme for highly compressed color image using weighted matrix through dct. *Int. J. Comput. Appl.* 43(1), 38–49 (2021)
- Pal, P., Chowdhuri, P., Jana, B.: Weighted matrix based reversible watermarking scheme using color image. *Multim. Tools Appl.* 77, 23 073–23 098 (2018)
- Pal, P., Jana, B., Bhaumik, J.: Watermarking scheme using local binary pattern for image authentication and tamper detection through dual image. *Secur. Privacy* 2(2), e59 (2019)
- Pal, P., Chowdhuri, P., Jana, B.: A secure reversible color image watermarking scheme based on lbp, lagrange interpolation polynomial and weighted matrix. *Multim. Tools Appl.* 80(14), 21651–21678 (2021)
- Mukherjee, S., Jana, B.: A novel method for high capacity reversible data hiding scheme using difference expansion. *Int. J. Natural Comput. Res. (IJNCR)* 8(4), 13–27 (2019)
- Biswapati, J., Debasis, G., Kumar, M.S.: Weighted matrix based reversible data hiding scheme using image interpolation. In: *Computational Intelligence in Data Mining—Volume 2: Proceedings of the International Conference on CIDM*, pp. 239–248. Springer, Cham (2016)
- Singh, P.K., Jana, B., Datta, K.: Superpixel based robust reversible data hiding scheme exploiting arnold transform with dct and ca. *J. King Saud Univ.-Comp. Inf. Sci.* 34(7), 4402–4420 (2022)
- Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: *Advances in Neural Information Processing Systems*, vol. 27. MIT Press, Cambridge, MA (2014)
- Yu, C.: Attention based data hiding with generative adversarial networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, pp. 1120–1128. AAAI Press, Menlo Park, CA (2020)
- Yu, C., Hu, D., Zheng, S., Jiang, W., Li, M., Zhao, Z.-q.: An improved steganography without embedding based on attention GAN. *Peer-to-Peer Network. Appl.* 14, 1446–1457 (2021)
- Tan, J., Liao, X., Liu, J., Cao, Y., Jiang, H.: Channel attention image steganography with generative adversarial networks. *IEEE Trans. Network Sci. Eng.* 9(2), 888–903 (2021)
- Hayes, J., Danezis, G.: Generating steganographic images via adversarial training. In: *Advances in Neural Information Processing Systems*, vol. 30. MIT Press, Cambridge, MA (2017)
- Zhu, J., Kaplan, R., Johnson, J., Fei-Fei, L.: Hidden: Hiding data with deep networks. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 657–672. Springer, Berlin (2018)
- Baluja, S.: Hiding images within images. *IEEE Trans. Pattern Anal. Mach. Intell.* 42(7), 1685–1697 (2019)
- Weng, X., Li, Y., Chi, L., Mu, Y.: High-capacity convolutional video steganography with temporal residual modeling. In: *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pp. 87–95. The Association for Computing Machinery, New York (2019)
- Fu, Z., Wang, F., Cheng, X.: The secure steganography for hiding images via gan. *EURASIP J. Image Video Process.* 2020(1), 46 (2020)
- Zhang, R., Dong, S., Liu, J.: Invisible steganography via generative adversarial networks. *Multim. Tools Appl.* 78, 8559–8575 (2019)

32. Chen, B., Wang, J., Chen, Y., Jin, Z., Shim, H.J., Shi, Y.-Q.: High-capacity robust image steganography via adversarial network. *KSII Trans. Internet Inf. Syst.* 14(1), 366–381 (2020)
33. Xu, G., Wu, H.-Z., Shi, Y.-Q.: Structural design of convolutional neural networks for steganalysis. *IEEE Signal Process Lett.* 23(5), 708–712 (2016)
34. Ye, J., Ni, J., Yi, Y.: Deep learning hierarchical representations for image steganalysis. *IEEE Trans. Inf. Forensics Sec.* 12(11), 2545–2557 (2017)
35. Boroumand, M., Chen, M., Fridrich, J.: Deep residual network for steganalysis of digital images. *IEEE Trans. Inf. Forensics Sec.* 14(5), 1181–1193 (2018)
36. Duan, X., Li, B., Guo, D., Zhang, Z., Ma, Y.: A coverless steganography method based on generative adversarial network. *EURASIP J. Image Video Process.* 2020, 1–10 (2020)
37. Chen, X., Zhang, Z., Qiu, A., Xia, Z., Xiong, N.N.: Novel coverless steganography method based on image selection and StarGAN. *IEEE Trans. Network Sci. Eng.* 9(1), 219–230 (2020)
38. Li, G., Feng, B., He, M., Weng, J., Lu, W.: High-capacity coverless image steganographic scheme based on image synthesis. *Sig. Process.: Image Commun.* 111, 116894 (2023)
39. Zhang, S., Li, H., Lim, L., Lu, J., Zuo, Z.: A high-capacity steganography algorithm based on adaptive frequency channel attention networks. *Sensors* 22(20), 7844 (2022)
40. Zhang, L., Lu, Y., Li, J., Chen, F., Lu, G., Zhang, D.: Deep adaptive hiding network for image hiding using attentive frequency extraction and gradual depth extraction. *Neural Comput. Appl.* 35, 10909–10927 (2023)
41. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. *Int. J. Comput. Vision* 115, 211–252 (2015)
42. Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: *Computer Vision—ECCV 2014: 13th European Conference, Proceedings, Part V*, pp. 740–755. Springer, Berlin (2014)
43. Agustsson, E., Timofte, R.: Ntire 2017 challenge on single image super-resolution: Dataset and study. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 126–135. IEEE, Piscataway (2017)
44. You, W., Zhang, H., Zhao, X.: A siamese cnn for image steganalysis. *IEEE Trans. Inf. Forensics Secur.* 16, 291–306 (2020)
45. Rustad, S., Andono, P.N., Shidik, G.F., et al.: Digital image steganography survey and investigation (goal, assessment, method, development, and dataset). *Signal Process.* 206, 108908 (2023)
46. Agrawal, R., Ahuja, K.: Csis: Compressed sensing-based enhanced-embedding capacity image steganography scheme. *IET Image Proc.* 15(9), 1909–1925 (2021)

**How to cite this article:** Zhu, Y., Dong, Y., Song, B., Yao, S.: Hiding image into image with hybrid attention mechanism based on GANs. *IET Image Process.* 18, 2679–2689 (2024). <https://doi.org/10.1049/ipr2.13127>