2025 International Conference on Intelligent Computing
July 26-29, Ningbo, China
https://www.ic-icc.cn/2025/index.php

Adaptive and High-security Image Steganography via Adversarial Embedding

Jibin Zheng, Li Ma, Wenyin Yang(✉), Fen Liu, Jihui Li

School of Computer Science and Artificial Intelligence,
Foshan University, No.33 Guangyun Road, Foshan, 528225, Guangdong, China
cswyyang@fosu.edu.cn

Abstract. The existing adversarial embedding methods, which based on distortion cost function and syndrome trellis codes (STCs), achieve adversarial embedding by manually adjusting the embedding cost. To address the limitation of hand-crafted embedding cost, we propose an end-to-end adversarial image steganography method, which automatically achieves adversarial embedding by using the gradients from the steganalytic network. We utilize gradients of the stego image to generate the adversarial embedding mask, then integrate it wtih the loss function to guide the secret messages embedded into the specific security-enhanced regions. Comparing with several state-of-the-art steganography methods, extensive experimental results demonstrate that our method significantly improves the security performance against convolutional neural network (CNN)-based steganalyzers and re-trained steganalyzers. For example, when against steganalyzers, the security improvement in terms of detection accuracy of our method achieves 30.68% higher than the SOTA steganography methods at 0.4 bpp (bit per pixel).

Keywords: Image steganography, adversarial sample, invertible neural network.

1 Introduction

Image steganography [1-4] is an invisible communication method that hides secret messages in the pixels or frequency coefficients. Currently, the most remarkable image steganography methods utilize the structure of "distortion cost function + STCs". The critical issue of these methods is designing embedding costs for embedding units. For instance, in the JPEG domain, UED [5] uniformly extends the embedding modification to the histogram bin corresponding to all possible DCT coefficient values, thereby reducing the overall variation of statistical characteristics. Work [6] uses the relative embedding changes of wavelet decomposition coefficients in the decompressed spatial domain to define the distortion cost, thereby obtaining the JPEG image steganography distortion cost function J-UNIWARD. With the development of neural networks, steganography methods based on deep learning have shown strong performance. JS-GAN [7] applies generative adversarial networks to JPEG adaptive steganography for the first time and verifies the feasibility of using GAN in the design of JPEG steganography

cost functions. It can automatically learn distortion cost values based on the features of the steganalyzer and compensate for the security vulnerabilities of steganography based on the detection features. In the spatial domain, there are several typical steganography methods, such as WOW [8], HILL [9], and UT-GAN [10]. On the contrary of steganography, steganalysis [11] is used to distinguish stego images from cover images.

Inspired by the idea that adversarial samples [12-14] can effectively deceive deep neural network-based classifiers, Researchers have introduced adversarial samples into image steganography to deceive CNN-based steganalyzer, thereby achieving the goal of improving security. Adversarial embedding (ADV-EMB) [15, 16] randomly divides image elements into common and adjustable groups according to a certain proportion. Min-max [17] can be seen as the iterative version of ADV-EMB. Adversarial enhancing method (AEN) [18] fully utilizes the gradients of cover and multiple stegos to guide the modification of embedding costs. In addition, an adversarial embedding method [19] suitable for both the spatial and JPEG domains is proposed. This method first generates some candidate stegos randomly and then selects the stego that can deceive the target steganalyzer.

The above methods have not incorporated adversarial embedding into network training, and the embedding and extracting of secret messages rely entirely on syndrome trellis codes. In this paper, we propose an end-to-end image steganography method based on automatic adversarial embedding without using the "distortion cost function + STCs" framework. The proposed method can automatically learn how to embed secret messages and maintain the stego image with visual and statistical imperceptibility. The contributions are as follows:

1. We first propose an end-to-end adversarial image steganography method, which automatically achieves adversarial embedding by using the gradients from the pre-trained CNN-based steganalyzer.
2. We propose a mask loss to guide the secret messages embedded into the specific regions. This design guide the stego signal along the opposite direction of the signs of gradient that can decrease the classification loss of the steganalyzer, enabling the steganalyzer to distinguish the stego image as a cover image.
3. Extensive experimental results demonstrate that our method significantly improves security compared with state-of-the-art steganography methods. For example, at 0.2 bpp (bit per pixel) and 0.4 bpp (bit per pixel), the detection accuracies against SRNet of our method are respectively about 49.40% and 46.16% lower than the state-of-the-art adversarial embedding steganography methods.

2 Related Works

2.1 Adversarial Sample

Image adversarial samples refer to samples formed by adding carefully constructed and subtle disturbances that are difficult for humans to discern through their senses to the input image. These adversarial samples cause the model to produce incorrect clas-

sification results with high confidence. Adversarial examples have evolved into effective techniques for attacking both traditional machine learning models and deep learning models, achieving impressive performance in computer vision, information hiding [20], and other fields [21]. Next, we will introduce two classical adversarial attack algorithms.

The fast gradient sign method (FGSM) [22] is a typical gradient-based adversarial attack method. This method adds perturbations along the opposite direction of the gradient to rapidly increase the loss function, leading to the model producing incorrect classification results. Carlini & Wagner (C&W) attacks [23] is a classic optimization-based adversarial attack method. This method customizes different objective functions and selects the optimal objective through experimental data to achieve adversarial attacks. C&W attack is characterized by small adversarial perturbations, high adversarial strength, and difficult to defend. However, a drawback is that it is time-consuming.

2.2 Steganalysis Methods

As the effectiveness of steganographic methods continues to improve, steganalysis methods are also being adjusted to enhance detection accuracy. As a comprehensive measure of detection accuracy and computational complexity, spatial rich model (SRM) [24] is the most commonly used hand-designed high-dimensional steganographic feature set in spatial images, while discrete cosine transform residual (DCTR) [25] and gabor filtering residual (GFR) are the most commonly used in JPEG images.

With the advancement of neural networks, several deep learning-based steganalysis networks with high security have been proposed. J-XuNet [26] is a 20-layer CNN-based steganalysis network. The preprocessing layer of this network uses high-pass filters from DCTR to suppress image content. The residuals obtained through high-pass filtering pass through an absolute value layer and then a truncation layer. All downsampling layers are processed using 3×3 convolutional layers with a stride of 2. SRNet [27] applies to both spatial and JPEG images. This network extracts residuals through an unpooled front-end network, replacing the prefiltering layer. ReLU activation functions are used to reduce computational complexity. DengNet [28] is a CNN-based fast and effective steganalysis network. It first introduces global covariance pooling into steganalysis, utilizing the second-order statistics of high-level features to enhance the detection performance of the steganalysis network.

3 Proposed Method

3.1 Model Architecture

In this paper, we propose an end-to-end adversarial image steganography framework to enhance the security of the image steganography. The framework of the proposed method is displayed in Fig. 1. Cover image (denoted as $\mathbf{C}$) refers to the spatial image or the quantized DCT coefficients of JPEG image. The size of cover image is $\omega \times h$ and the number of channel is c. The corresponding stego image with the same size of $\mathbf{C}$ is

characterized as **S**. In addition, we use **M** and **M′** to depict the original secret message and the revealed secret message, respectively. Our proposed method utilizes the invertible neural network (INN) [29-31] to realize the embedding and extracting of secret messages. The cover image and secret message are input into the forward process of INN to generate the stego image. The backward process of INN can obtain the revealed secret message and revealed cover image from the stego image.

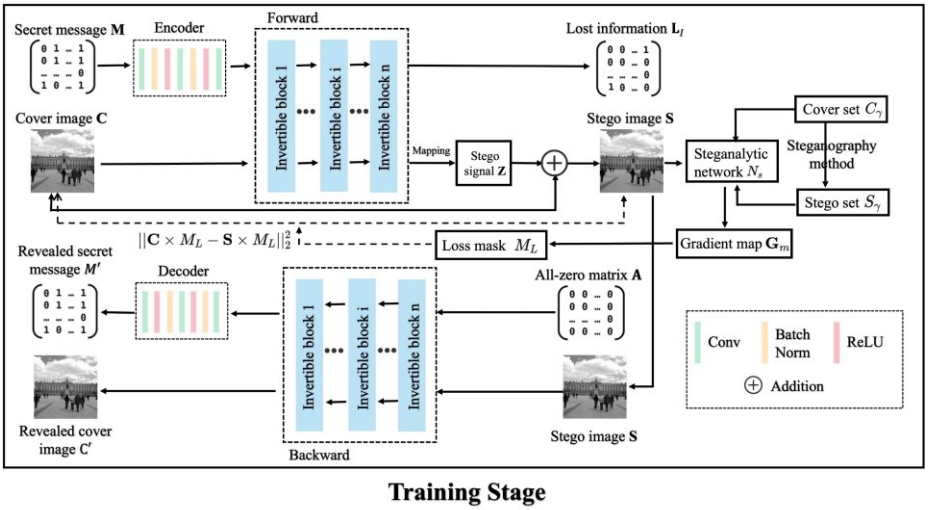

Fig. 1:Overview of the proposed architecture.

3.2 Adversarial Embedding

Pre-training the steganalytic network. We obtain a cover set C_γ from the BOSSBase ver.1.01 and BOWS2 datasets. Then, we utilize UNIWARD to calculate its corresponding embedding cost and generate a stego set S_γ with the given embedding capacity. Finally, we train the cover set and stego set to obtain a pre-trained steganalytic network N_s (also named target steganalyzer). We use XuNet [32, 26] and SRNet [27] as the steganalytic network N_s. It is worth noting that the steganalytic network is only pre-trained once and will not be updated during the process of embedding and extracting secret messages. All gradient calculations involved in the proposed method are implemented through this pre-trained steganalytic network.

Adversarial embedding. We generate the stego image using the forward process of INN. Then, we input the stego image into the steganalytic network and use the back-propagation algorithm to calculate its gradient map G_M. The specific calculation process is as follows:

$$\begin{aligned} G_M &= \nabla_s L(s,t;N_s), \\ L(s,t;N_s) &= -t\log\big(N_s(s)\big) - (1-t)\log\big(1-N_s(s)\big). \end{aligned} \tag{1}$$

where s denote a stego image. $L(s,t;N_s)$ represents the binary cross entropy loss between the targeted label t and the output of steganalytic network $N_s(x)$ ($0 \leq$

$N_s(x) \leq 1$). For targeted label t, we define $t = 0$ as the cover image while $t = 1$ as the stego image. The purpose of adversarial image steganography is to enable the target steganalyzer N_s to distinguish the stego image s as the target label ($t = 0$). From the property of the fast gradient sign method (FGSM), we should modify the input stego image s along the opposite direction of the signs of gradient map G_M, which can decrease the loss $L(s, t; N_s)$.

In order to introduce the idea of adversarial samples mentioned above into image steganography, the existing adversarial image steganography methods utilize the gradient signs of the cover and stego images to update the original symmetric embedding cost to the asymmetric embedding cost as follows:

$$\begin{cases} \rho_{i,j}^{+1} > \rho_{i,j}^{-1}, if \nabla x_{i,j} L(x, t, N_s) > 0; \\ \rho_{i,j}^{+1} = \rho_{i,j}^{-1}, if \nabla x_{i,j} L(x, t, N_s) = 0; \\ \rho_{i,j}^{+1} < \rho_{i,j}^{-1}, if \nabla x_{i,j} L(x, t, N_s) < 0. \end{cases} \tag{2}$$

where x denotes cover or stego image, t denotes the targeted label.

After obtaining the asymmetric embedding cost, the existing adversarial image steganography methods use STC to embed the given secret message into the cover image to obtain the final stego image. From this process, we can see that existing adversarial image steganography methods achieve adversarial embedding by adjusting the embedding cost rather than incorporating the idea of adversarial embedding into network training. Unlike these methods, our proposed method achieves adversarial embedding automatically by integrating the loss function with gradients from the pre-trained steganalytic network. The detailed process is as follows:

Firstly, the output (value range is [0,255]) of the forward process of INN is mapped into the stego signal **Z** (value range is [-1,1]). Adding the cover image **C** and the stego signal **Z** yields stego image **S**. Then, we feed the stego image **S** into the steganalytic network N_s to obtain the gradient map G_M according to Formula (1). Next, we utilize the sign of stego signal $Z(i,j)$ and gradients of the stego image $G_M(i,j)$ at coordinate (i,j) to generate the embedding mask.

$$M_L(i,j) = \begin{cases} \delta, if \quad sign\big(Z(i,j)\big) \times G_M(i,j) \geq 0 \\ -\delta, if \quad sign\big(Z(i,j)\big) \times G_M(i,j) < 0 \end{cases} \tag{3}$$

where the mask factor $\delta > 1$ denotes the adversarial magnitude. The value of δ will be discussed in the next section. Finally, we design the following mask loss $\mathcal{L}_m$ according to the embedding mask:

$$\mathcal{L}_m = ||C \times M_L - S \times M_L||_2^2 \tag{4}$$

This mask loss can guide the secret messages embedded in specific security-enhanced regions. For example, when the sign of the stego signal is consistent with gradients of the stego image, adversarial magnitude δ will increase the mask loss $\mathcal{L}_m$. In other words, the position (i,j) that satisfies $sign\big(Z(i,j)\big) \times G_M(i,j) \geq 0$ will reduce the embedding probability. The secret message is iteratively embedded into the cover image along the opposite direction of the signs of the gradient graph, minimizing the cross-entropy loss of the pre-trained steganalytic network N_s as much as possible. Therefore, the stego image will be determined as the cover image by the pre-trained steganalytic network as much as possible.

3.3 Loss Function

The objective loss function of our method includes mask loss $\mathcal{L}_m$, generating loss $\mathcal{L}_g$, extracting loss $\mathcal{L}_e$, and alignment loss $\mathcal{L}_a$. The training objective is to minimize the following:

$$\mathcal{L}_t = \lambda_m \mathcal{L}_m + \lambda_g \mathcal{L}_g + \lambda_e \mathcal{L}_e + \lambda_a \mathcal{L}_a. \tag{5}$$

where $\lambda_m = 3$, $\lambda_g = 1$, $\lambda_e = 15$, and $\lambda_a = 5$ are the weight factors to adjust loss terms. We have already introduced mask loss $\mathcal{L}_m$. Next, we will provide a detailed introduction to the remaining three losses.

Generating Loss $\mathcal{L}_g$. The generating loss $\mathcal{L}_g$ ensures the generated stego image resembles as close as possible to the original cover image. The typical image quality evaluation indicators structural similarity index (SSIM) and mean square error (MSE) are employed to measure the difference between the generated stego image and original cover image depicted as:

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

Extracting Loss $\mathcal{L}_e$. The extracting Loss $\mathcal{L}_e$ indicates the difference between the revealed and original secret messages. The low extracting loss means obtaining high extraction accuracy during the process of the message recovery, which can be calculated as:

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

Alignment Loss $\mathcal{L}_a$. The process of embedding secret messages inevitably causes damage to the information in the cover image. In addition, to ensure the high visual quality of the stego image, it is difficult to entirely embed large amounts of secret messages into the cover image. So the lost information matrix $\boldsymbol{L_I}$ consists of these two information losses. In this paper, we propose the following alignment loss $\mathcal{L}_a$, which minimizes the difference between the lost information matrix $\boldsymbol{L_I}$ and the matrix $\boldsymbol{I_0}$ to embed as many secret messages as possible into the stego image while minimizing the information loss of the entire network.

$$\mathcal{L}_a = \left|\left|\boldsymbol{L}_I - \boldsymbol{I_0}\right|\right|_2^2 \tag{8}$$

where $\boldsymbol{I_0}$ is an all-zero matrix of the same size as $\boldsymbol{L_I}$. Because after network training, lost information matrix $\boldsymbol{L_I}$ hardly contains useful information about the process of embedding secret messages. We input an all-zero matrix $\boldsymbol{A}$ and the stego image $\boldsymbol{S}$ during the process of extracting secret messages to recover the original secret message.

4 Experimental Results

4.1 Experimental Settings

Datasets. We use 20,000 grayscale images from the BOSSBase ver.1.01 and BOWS2 datasets as cover images. To maintain consistency with the baseline methods, we resize all cover images to a size of 256×256 using the "imresize" function in Matlab.

Evaluation Metrics. We adopt the peak signal-to-noise ratio (PSNR) and structural similarity index measure (SSIM) to measure the image quality of the generated stego images. Moreover, we use the extraction accuracy to evaluate the recovery performance of our method. We utilize the receiver operating characteristic (ROC) curve generated by the statistical steganalysis tool StegExpose [33], together with detection accuracy to measure the security of our method. The larger value of PSNR, SSIM, and extraction accuracy indicate higher performance. The smaller value of detection accuracy indicate higher performance.

Table 1. Detection accuracy (%) evaluated on four steganalyzers in spatial domain. "Without" denotes the proposed method does not use adversarial embedding. "Ours-Xu" and "Ours-SR" represent our proposed method adopt XuNet and SRNet as the steganalytic network, respectively.

Steganalyzer	0.2 bpp			0.4 bpp			bpp		
	Without	Ours-Xu	Ours-SR	Without	Ours-Xu	Ours-SR	Without	Ours-Xu	Ours-SR
SRM	75.69	74.44	**73.53**	84.70	76.41	**72.99**	79.56	**76.97**	78.62
XuNet	91.75	**50.50**	51.47	95.50	**50.50**	51.95	100.00	**54.41**	60.95
DengNet	98.10	**50.20**	50.85	100.00	53.54	**51.66**	100.00	61.69	**55.36**
SRNet	99.40	51.15	**50.00**	100.00	61.39	**53.84**	100.00	64.91	**59.32**

4.2 Security on Steganalyzers.

The pre-trained steganalyzers are trained by conventional stego images but cannot perceive adversarial stego images. In this subsection, we evaluate the security of our method in this case. The security improvement in the following text refers to the performance improvement compared to the "original" method. Table 1 and Table 2 show the detection accuracy evaluated on four pre-trained steganalyzers in the spatial domain, respectively. We can obtain the following five observations:

Table 2. Benchmark evaluations focus on image quality, extraction performance, and security in a white-box scenario. The term "Naive INN from work [2]" describes the original INN framework that does not incorporate adversarial hiding. Meanwhile, "Naive perturbation" refers to the process of directly adding perturbations to the stego image without utilizing the gradient of the stego image.

Capacities	Methods	Image quality		Extraction performance		Security
		PSNR(dB)	SSIM	BER(%)	R_s	Acc_d(%)
0.1 bpp	Naive INN from work [2]	**42.35**	**0.9821**	0.0731	7/4,000	99.87
	Naive perturbation	36.47	0.9364	0.0032	1,756/4,000	57.59
	Ours-SR	41.16	0.9648	**0**	**4,000/4,000**	**50.05**
0.2 bpp	Naive INN from work [2]	**39.76**	**0.9652**	0.3415	7/4,000	99.56
	Naive perturbation	34.73	0.9053	0.0013	1,492/4,000	57.03
	Ours-SR	39.84	0.9667	**0**	**4,000/4,000**	**50.12**

1. In comparison to the "naive INN from work [2]," our approach results in a slight reduction in image quality. This is due to the presence of adversarial perturbations, which function as noise and inevitably affect the image. Nonetheless, our method achieves a Bit Error Rate (BER) of 0, demonstrating its ability to extract secret messages without loss. Furthermore, our technique accurately retrieves all 4,000 test images.
2. Our proposed end-to-end adversarial image steganography method significantly improves the security against pre-trained steganalyzers compared with the original method. For example, when the steganalyzer is SRNet, the security improvements of "ours-SR" in the spatial domain for 0.2-1.0 bpp are 49.40%, 46.16%, and 40.68%, respectively.
3. Our method performs better in resisting CNN-based steganalyzers than traditional steganalyzers. At 0.2 bpp, the security improvements of "ours-SR" are 40.28%, 47.25%, and 49.40% in resisting XuNet, DengNet, and SRNet, respectively. However, the security improvement is only 2.04% when resisting SRM detection.
4. Due to the transferability of adversarial examples, our proposed method also achieves strong security in resisting the non-targeted pre-trained steganalyzers. For example, "ours-SR" adopts SRNet as the pre-trained steganalytic network, but it can effectively evade the detection of three other steganalyzers.
5. The security improvements of our method would decrease with increasing the embedding capacities in the spatial. For example, the security improvements of "ours-SR" in the spatial domain at 0.2 bpp is 49.40%, and at 1.0 bpp is 40.68%.

4.3 Comparison with state-of-the-art Methods.

In order to verify the progressiveness of our method, we present the comparative experimental results and analysis of our proposed method and various state-of-the-art image steganography methods in this subsection.

Comparison with adversarial embedding steganography methods in low capacity. We compare the security of our method with UNIWARD, ADV-EMB, and Work [19] at 0.4 bpp. UNIWARD is recognized as the most secure traditional steganography method. ADV-EMB is a classic steganography method based on adversarial embedding. Work [19] is currently recognized as the most advanced steganography method based on adversarial embedding. In order to show the fairness of the comparison experiment, both our method and Work [19] use UNIWARD to train the steganalyzers. In addition, we compared the experimental results obtained by the steganography method with the best security enhancement and the most advanced targeted steganalyzer in the original paper of Work [19]. Table 3 show the detection accuracy of our method and these three comparison methods with the following observations:

Table 3. Benchmark evaluations focus on image quality, extraction performance, and security in a white-box scenario. The term "Naive INN from work [2]" describes the original INN framework that does not incorporate adversarial hiding. Meanwhile, "Naive perturbation" refers to the process of directly adding perturbations to the stego image without utilizing the gradient of the stego image.

Methods	SRM	XuNet	DengNet	SRNet
S-UNIWARD	77.68	82.71	89.15	96.90
ADV-EMB	78.10	78.06	84.55	86.70
Lan [2]	79.81	90.66	95.12	96.86
Liu [19]	**74.29**	76.31	82.40	84.54
Ours-SR	76.70	**61.39**	57.86	53.84

Table 4. The extraction accuracy (%), PSNR (dB), SSIM, and the detection accuracy (%) of our method and three SOTA deep steganography at 1.0 bpp embedding capacity. "Ours-SR" represent our proposed method adopt SRNet as the steganalytic network. ↑ means the larger the value, the better; while ↓ means the smaller the value, the better.

Methods	Extraction accuracy ↑	Detection by DengNet ↓	Detection by SRNet ↓
SteganoGAN	97.81	97.71	96.12
ABDH	97.44	97.82	98.53
CHAT-GAN	99.62	94.65	94.13
Lan [2]	99.75	93.27	94.86
AHDeS [34]	99.15	84.84	86.91
Ours-SR	**99.99**	**57.32**	**59.32**

1. Our method has significant security improvements in resisting CNN-based steganalyzers. Compared with Work [19] in the spatial domain, "Ours-SR" respectively improves the security by 14.49%, 24.52%, and 30.68% against XuNet, DengNet, and SRNet.
2. The security of our method will not decrease with the enhancement of steganalyzers. As is well known, DengNet and SRNet have much stronger detection capabilities than XuNet. We can observe that the three comparison methods have the best security against XuNet. When resisting DengNet and SRNet, the detection accuracy will significantly increase, which means that the security will decrease substantially. It is worth noting that our method can still achieve detection accuracy of 53.84% when resisting SRNet detection. It indicates that our method has good prospects in resisting advanced CNN-based steganalyzers.

Comparison with SOTA deep learning (DL)-based steganography methods in large capacity. Due to the limitations of STC encoding, the three comparison methods mentioned above only have good performance when the embedding capacity is less than 0.4 bpp. Recently, some deep learning-based steganography methods have shown excellent performance in large capacity. To demonstrate the superiority of our method in large capacity, we select three SOTA DL-based steganography methods (SteganoGAN, ABDH, and CHAT-GAN) as the comparison methods. Table 4 shows the

extraction accuracy, image quality (PSNR, SSIM), and security (detection accuracy) of our method and three SOTA DL-based steganography methods under 1.0 bpp embedding capacity. We can observe that our method achieves the best performance in terms of extraction accuracy and security. As shown in Table 4, "Ours-SR" achieves a 0.37% improvement in extraction accuracy compared to CHAT-GAN. When resisting the detection of DengNet and SRNet, "Ours-SR" respectively achieves 37.33% and 34.81% improvements.

4.4 Visualization results

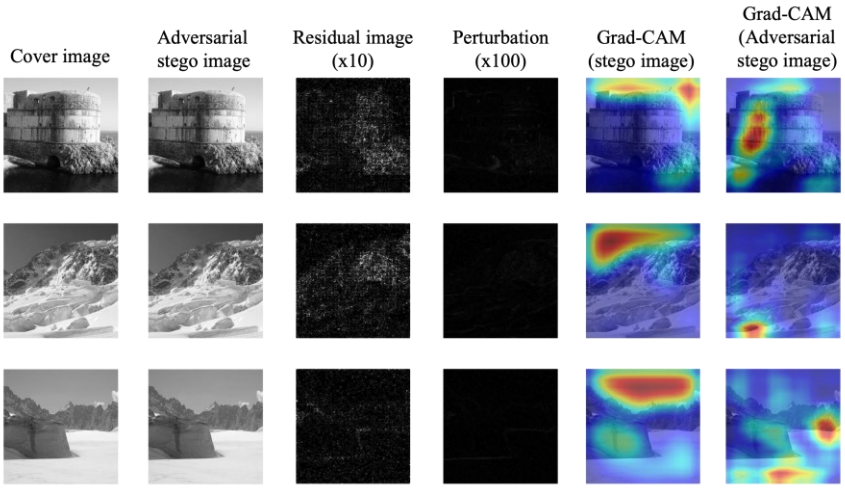

Fig. 2: The visualization results of the adversarial stego image, residual image , perturbation and Grad-CAM of our method.

Gradient-weighted Class Activation Mapping (Grad-CAM), also known for its heatmap visualization effects, is a technique used to visualize the internal decision-making processes of deep networks. It generates a heatmap that highlights the regions in the input image that play a crucial role in predicting a specific class by utilizing gradient information of the target class with respect to the feature maps of convolutional layers during classification tasks. Fig. 2 illustrates the visual effect of Grad-CAM. In the Grad-CAM results, the red regions represent the areas that the steganalyzer primarily focuses on. For images with concealed information that have not undergone adversarial perturbations, the red regions in Grad-CAM are mainly concentrated in texture-rich areas, which are typically key positions for the steganalyzer to extract classification features and also where secret information is embedded, making it easier for the information to be detected.

Analysis of Fig. 2 shows that the adversarial stego images generated by our method do not exhibit noticeable color distortion or text copy artifacts, making them visually indistinguishable from the cover images. The adversarial perturbations generated by this

method are primarily applied to complex texture areas. As a result, when these texture-based adversarial perturbations are added to the concealed images to produce adversarial stego images, the attention of the steganalyzer is successfully misled. The red regions in Grad-CAM shift from critical texture areas to smoother regions, and even to areas unrelated to classification. This shift indicates that our method can effectively guide the attention of the steganalyzer to regions that do not contain secret information, leading to a lack of sufficient discriminative information for correctly identifying the adversarial stego images.

5 Conclusion

This paper proposes an end-to-end adversarial image steganography method, which automatically achieves adversarial embedding by using the gradients from the pre-trained convolutional neural network (CNN)-based steganalyzer. Unlike another adversarial embedding steganography method, we introduce the idea of adversarial embedding into the automatic training of networks for the first time. We design a novel embedding mask to guide the secret messages embedded into the specific security-enhanced regions. Extensive experimental results demonstrate the superiority of our method under pre-trained and re-trained CNN-based steganalyzers in spatial domains.

The adversarial embedding proposed in this paper can be flexibly applied to other end-to-end DL-based steganography frameworks, and have good application prospects in resisting advanced CNN-based steganalyzers. However, there are several important challenges worth further overcome. For instance, while our method achieves good security against CNN-based steganalyzers in the spatial domain, it is also necessary to consider how to enhance resistance to CNN-based steganalysis in the JPEG domain. We will strive to address this issue in our future work.

Acknowledgments. This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62102098 and 62472105, the Guangdong-Foshan Joint Fund Project of Basic and Applied Basic Research Foundation of Guangdong Province under Grant 2022A1515140096, and Key Special Projects of Guangdong Provincial Department of Education under Grant 2022ZDZX1026.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. J. Tao, S. Li, X. Zhang, and Z. Wang. Towards robust image steganography. IEEE Transactions on Circuits and Systems for Video Technology, 29(2):594–600, 2018.
2. Yuhang Lan, Fei Shang, Jianhua Yang, Xiangui Kang, and Enping Li. Robust image steganography: Hiding messages in frequency coefficients. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 37, pages 14955–14963, 2023.

3. Xiao Ke, Huanqi Wu, and Wenzhong Guo. Stegformer: rebuilding the glory of autoencoder-based steganography. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 38, pages 2723–2731, 2024.
4. Kai Zeng, Kejiang Chen, Jiansong Zhang, Weiming Zhang, and Nenghai Yu. Toward secure and robust steganography for black-box generated images. IEEE Transactions on Information Forensics and Security, 19:3237–3250, 2024.
5. Linjie Guo, Jiangqun Ni, and Yun Qing Shi. Uniform embedding for efficient jpeg steganography. IEEE transactions on Information Forensics and Security, 9(5):814–825, 2014.
6. V. Holub, J. Fridrich, and T. Denemark. Universal distortion function for steganography in an arbitrary domain. EURASIP Journal on Information Security, 2014(1):1–13, 2014.
7. J. Yang, D. Ruan, X. Kang, and Y. Shi. Towards automatic embedding cost learning for jpeg steganography. In Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, pages 37–46, 2019.
8. Vojtěch Holub and Jessica Fridrich. Designing steganographic distortion using directional filters. In 2012 IEEE International workshop on information forensics and security (WIFS), pages 234–239. IEEE, 2012.
9. Bin Li, Ming Wang, Jiwu Huang, and Xiaolong Li. A new cost function for spatial image steganography. In 2014 IEEE International conference on image processing (ICIP), pages 4206–4210. IEEE, 2014.
10. J. Yang, D. Ruan, J. Huang, X. Kang, and Y. Shi. An embedding cost learning framework using gan. IEEE Transactions on Information Forensics and Security, 15(99):839–851, 2019.
11. J Zeng, S Tan, G Liu, B Li, and J Huang. Wisernet: Wider separate-then-reunion network for steganalysis of color images. IEEE Transactions on Information Forensics and Security, 14(10):2735–2748, 2019.
12. Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. IEEE Transactions on Evolutionary Computation, 23(5):828–841, 2019.
13. Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 2574–2582, 2016.
14. Zhirui Wang, Liu Yang, and Yahong Han. Robust source-free domain adaptation with anti-adversarial samples training. Neurocomputing, 614:128777, 2025.
15. W Tang, B Li, S Tan, M Barni, and J Huang. Cnn-based adversarial embedding for image steganography. IEEE Transactions on Information Forensics and Security, 14(8):2074–2087, 2019.
16. Chaoen Xiao, Sirui Peng, Lei Zhang, Jianxin Wang, Ding Ding, and Jianyi Zhang. A transformer-based adversarial network framework for steganography. Expert Systems with Applications, 269:126391, 2025.
17. Solène Bernard, Patrick Bas, John Klein, and Tomas Pevny. Explicit optimization of min max steganographic game. IEEE Transactions on Information Forensics and Security, 16:812–823, 2020.
18. Sai Ma, Xianfeng Zhao, and Yaqi Liu. Adaptive spatial steganography based on adversarial examples. Multimedia Tools and Applications, 78:32503–32522, 2019.
19. Minglin Liu, Tingting Song, Weiqi Luo, Peijia Zheng, and Jiwu Huang. Adversarial steganography embedding via stego generation and selection. IEEE Transactions on Dependable and Secure Computing, 20(3):2375–2389, 2022.

20. Huaxiao Mo, Tingting Song, Bolin Chen, Weiqi Luo, and Jiwu Huang. Enhancing jpeg steganography using iterative adversarial examples. In 2019 IEEE International Workshop on Information Forensics and Security (WIFS), pages 1–6. IEEE, 2019.
21. Mengte Shi, Sheng Li, Zhaoxia Yin, Xinpeng Zhang, and Zhenxing Qian. On generating jpeg adversarial images. In 2021 IEEE International Conference on Multimedia and Expo (ICME), pages 1–6. IEEE, 2021.
22. Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. 2014.
23. Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In 2017 ieee symposium on security and privacy (sp), pages 39–57. IEEE, 2017.
24. Jessica Fridrich and Jan Kodovsky. Rich models for steganalysis of digital images. IEEE Transactions on information Forensics and Security, 7(3):868–882, 2012.
25. Vojtěch Holub et al. Content Adaptive Steganography: Design and Detection. PhDthesis. Dept. of Electrical and Computer Engineering, Binghamton University, 2014.
26. H. Xu, G.and Wu and Y. Shi. Ensemble of cnns for steganalysis: An empirical study. In Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security, pages 103–107, 2016.
27. M. Boroumand, M. Chen, and J. Fridrich. Deep residual network for steganalysis of digital images. IEEE Transactions on Information Forensics and Security, 14(5):1181–1193, 2018.
28. Xiaoqing Deng, Bolin Chen, Weiqi Luo, and Da Luo. Fast and effective global covariance pooling network for image steganalysis. In Proceedings of the ACM workshop on information hiding and multimedia security, pages 230–234, 2019.
29. Fei Shang, Yuhang Lan, Jianhua Yang, Enping Li, and Xiangui Kang. Robust data hiding for jpeg images with invertible neural network. Neural Networks, 163:219–232, 2023.
30. J. Jing, X. Deng, M. Xu, J. Wang, and Z. Guan. Hinet: Deep image hiding by invertible network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 4733–4742, 2021.
31. S. Lu, R. Wang, T. Zhong, and P. Rosin. Large-capacity image steganography based on invertible neural networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 10816–10825, 2021.
32. Guanshuo Xu, Han-Zhou Wu, and Yun-Qing Shi. Structural design of convolutional neural networks for steganalysis. IEEE Signal Processing Letters, 23(5):708–712, 2016.
33. B. Boehm. Stegexpose-a tool for detecting lsb steganography. arXiv preprint arXiv:1410.6656, 2014.
34. Weixuan Tang, Haoyu Yang, Yuan Rao, Zhili Zhou, and Fei Peng. Dig a hole and fill in sand: Adversary and hiding decoupled steganography. In Proceedings of the 32nd ACM International Conference on Multimedia, MM '24, page 10440–10448, New York, NY, USA, 2024. Association for Computing Machinery.