Journal of Information Security and Applications 68 (2022) 103252

Contents lists available at ScienceDirect

Journal of Information Security and Applications

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

Robustness enhancement against adversarial steganography via steganalyzer outputs☆

Chuan Qin [a], Weiming Zhang [a,*], Hang Zhou [b], Jiayang Liu [c], Yuan He [d], Nenghai Yu [a]

[a] School of Cyber Science and Technology, University of Science and Technology of China, Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, China
[b] Simon Fraser University, Canada
[c] National University of Singapore, Singapore
[d] Alibaba Security, Alibaba Group, China

ARTICLE INFO

Keywords:
Steganalysis
Robust framework
Adversarial steganography
Probabilistic outputs

ABSTRACT

Recently, CNN (convolutional neural network) steganalyzers have significantly outperformed handcrafted features in detecting steganography. However, adversarial steganography has challenged the applications of them in the real world. Adversarial steganography can easily deceive the target CNN steganalyzer while sending secret messages. In this paper, a general framework is proposed. It that can improve the robustness of CNN steganalyzers against adversarial steganography while keeping detecting cover and conventional stego images accurately. Specifically, a rough filter that filters adversarial stego images out of the input data is set. It exploits the differences between cover and adversarial stego images on probabilistic outputs of the target CNN steganalyzer and a handcrafted steganalyzer. Extensive experiments show that the proposed framework can significantly improve the robustness of CNN steganalyzers. In the real-world scenario where cover, conventional stego and adversarial stego images are mixed, the robustness enhanced CNN steganalyzers can achieve the optimal overall performance.

1. Introduction

Image steganography [1–5] is the science and art of covert communication that embeds secret messages into cover images with minimal distortions [2,6–9]. Currently, the most successful steganographic approaches are based on the minimal distortion model [10,11], which formulates the steganography problem as source coding with a fidelity constraint. Under the framework of the minimal distortion model, there are two tasks: (1) defining the costs of modifying the elements of a cover image and (2) designing a practical embedding methodology while minimizing the arbitrary cost defined previously. Since syndrome-trellis codes (STCs) [10,11] perform near the maximum theoretical bound at the second task, steganography research mostly focuses on the design of the cost function, such as WOW [12], UNIWARD [13], HILL [14], MiPOD [15], UERD [16] and J-MiPOD [17] etc.

With the development of steganography, many steganalysis methods [18–24] have been proposed. Steganalysis is an image binary classification task that aims to classify cover and stego images. As general image classification tasks, steganalysis has evolved from handcrafted features combined with traditional machine learning models (called handcrafted steganalyzers) to CNNs. The most successful handcrafted features are the spatial rich model (SRM) [25] and the Gabor filter rich (GFR) [19], in the spatial domain and the JPEG domain respectively. The best performing traditional classifier model is the ensemble classifier (EC) [26], a random forest-based machine learning model. And since YeNet [27], CNN steganalyzers [27–37] have significantly outperformed handcrafted ones. Currently, SRNet [31], CovNet [38], and SiaStegNet [35] are considered to be the most successful CNN steganalyzers.

Whereas CNNs have provided breakthroughs in various areas, they have been found to be vulnerable to adversarial attacks [39–43]. Generally, the adversarial attack is a technique that deceives CNNs into outputting incorrect results by adding elaborately designed small adversarial perturbations to original images. Images generated by this technique are called adversarial examples. Similarly, adversarial steganography [44–50] has been proposed. It can communicate secret messages and deceive target CNN steganalyzers at the same time. Currently,

☆ This work was supported in part by the Natural Science Foundation of China under Grant 6200233462072421, and 62121002, Anhui Science Foundation of China under Grant 2008085QF296, and by Anhui Initiative in Quantum Information Technologies under Grant AHY150400.

* Corresponding author.
E-mail addresses: qc94@mail.ustc.edu.cn (C. Qin), zhangwm@ustc.edu.cn (W. Zhang), zhouhang2991@gmail.com (H. Zhou), ljyljy@nus.edu.sg (J. Liu), heyuan.hy@alibaba-inc.com (Y. He), ynh@ustc.edu.cn (N. Yu).

https://doi.org/10.1016/j.jisa.2022.103252

Available online 29 June 2022

the state-of-the-art way to defend against adversarial steganography is retraining [40,47–49,51], i.e., augmenting the training set with adversarial stego images. However, it is found that adversarial steganography can still deceive retrained CNN steganalyzers. Thus, CNN steganalyzers will get stuck in the "arms race" with adversarial steganography [48,49]. Even worse, the detection accuracy on common samples (cover and conventional stego images) of retrained CNN steganalyzers will drop [47–49].

In this paper, a robustness enhancement framework for CNN steganalyzers against adversarial steganography is proposed. Different from retraining, it avoids CNN steganalyzers stuck in the "arms race" with adversarial steganography. Though adversarial steganography can easily fool CNN steganalyzers, it has a small impact on handcrafted steganalyzers. Hence, the images labeled as cover by the target CNN steganalyzer and as stego by a handcrafted steganalyzer may contain a substantial number of adversarial stego images. Also, adversarial stego images obtain substantially distinguished probabilistic outputs, which is caused by minimizing adversarial perturbations. By exploiting such characteristics, adversarial stego images are filtered from the input and labeled them by a specific classifier. The proposed scheme is robust against adversarial steganography while maintaining higher detection accuracies on cover and conventional stego images than handcrafted steganalyzers.

Our framework is evaluated based on the area under the curve (AUC) in the real-world scenario with a mixture of adversarial images and conventional stego images. The experimental results show that the robustness enhanced CNN steganalyzers substantially outperforms the previous works. The contributions of the proposed framework are summarized as follows.

- Previously, the arms race like retraining was the only way to defend against adversarial steganography. This paper proposed a framework that avoids CNN steganalyzers stuck in the arms race like retraining.
- By utilizing the robustness gap between CNN steganalyzers and handcrafted steganalyzers, the proposed framework filters adversarial stego images out of the input stream.
- The extensive experiments prove that the robustness enhanced CNN steganalyzers obtain superior comprehensive detection ability in the real-world scenario.

The rest of this paper is organized as follows. In Section 2, two retraining strategies are briefly reviewed. The proposed framework is detailed in Section 3. Extensive evaluations and comparisons are carried out in Section 4. The paper is concluded in Section 5.

2. Previous works

Previously, the only effective way to defend against adversarial steganography was retraining. Tang et al. [47] and Bernard et al. [48,49] discussed two retraining methods respectively. In this section, these two methods are briefly introduced.

2.1. Tang et al.'s method

In each round of Tang et al.'s [47] setting, the steganographer takes the first step to attack the CNN steganalyzer, then the steganalyzer augments the training set with some adversarial stego images in the current round and gets retrained.

In a three-round experiment, the success rates of ADV-EMB are 99.39%, 97.37% and 95.23%, respectively. It indicates retraining with current-round adversarial stego images provides little robustness against adversarial steganography, even if the generation method keeps the same. Moreover, the retrained CNN steganalyzer suffers from accuracy drop of common samples (cover and conventional stego images). The decreases are 2.57% and 4.31% in the second and third rounds.

2.2. Min–max retraining

Bernard et al. [48,49] described a more complex scenario than Tang et al. [47] Instead of augmenting the training set with randomly selected adversarial stego images, the steganalyzer only adds the adversarial stego images that obtain the highest detectability. The detectability is measured by the probabilistic outputs of the steganalyzers in previous rounds.

In Bernard et al.'s experiment, the success rates of ADV-EMB are mostly higher than 90% across the tested target CNN steganalyzers and payloads. Also, the average detection accuracy of CNN steganalyzers on conventional stego images and adversarial stego images of previous rounds continuously drops as the "arms race" goes.

3. The proposed robustness enhancement framework

In this section, the proposed robustness enhancement framework is detailed. Specifically, the motivation (Section 3.1) is first introduced. The overall architecture (Section 3.2) is presented. Two key constitutions of the proposed framework, the rough filter and the specific classifier, are detailed in Sections 3.3 and 3.4. The theoretical analysis about the probabilistic outputs of steganalyzers on adversarial stego images is elaborated in Section 3.3.4 and Appendix.

3.1. Motivation

Though detecting adversarial steganography is quite a challenge, traces they still leave in the outputs of CNN steganalyzers and handcrafted steganalyzers. It indicates a chance to enhance CNN steganalyzers' robustness.

Calculations in CNN are mostly derivable. The gradient maps are accessible, allowing the attacker to modify clean images to deceive CNN models. While key calculations in extracting handcrafted features are underivable, such as generating co-occurrence matrices or histograms. It prevents the attacker to craft adversarial stego images against them. Therefore, the target CNN steganalyzer cannot correctly classify adversarial stego images, but handcrafted models are almost immune to adversarial steganography. Then the images that the handcrafted model labels as stego and the CNN labels as cover may contain a significant number of adversarial stego images.

Furthermore, adversarial steganography aims to trick the CNN steganalyzer into labeling adversarial stego images as cover with minimal perturbations. Subject to an adversarial stego image being labeled as cover, the minimal adversarial perturbations will be made when the probabilistic output of predicting as cover class is just larger than that of predicting as stego class. Meanwhile, for cover images, the probabilistic outputs of predicting as cover class are mostly much larger than that of predicting as stego class. Therefore, it is reasonable to assume that cover, conventional stego and adversarial stego images are all distinct in the two-dimensional feature space constructed by the probabilistic output of predicting as cover class of CNN and handcrafted models.

The classification problem with a mixture of cover, conventional stego, and adversarial stego images may be decomposed into the classification of cover and conventional stego images and the classification of cover and adversarial stego images. This allows us to let CNN steganalyzers and handcrafted steganalyzers exploit their advantages to solve the tricky task of steganalysis in the presence of adversarial stego images.

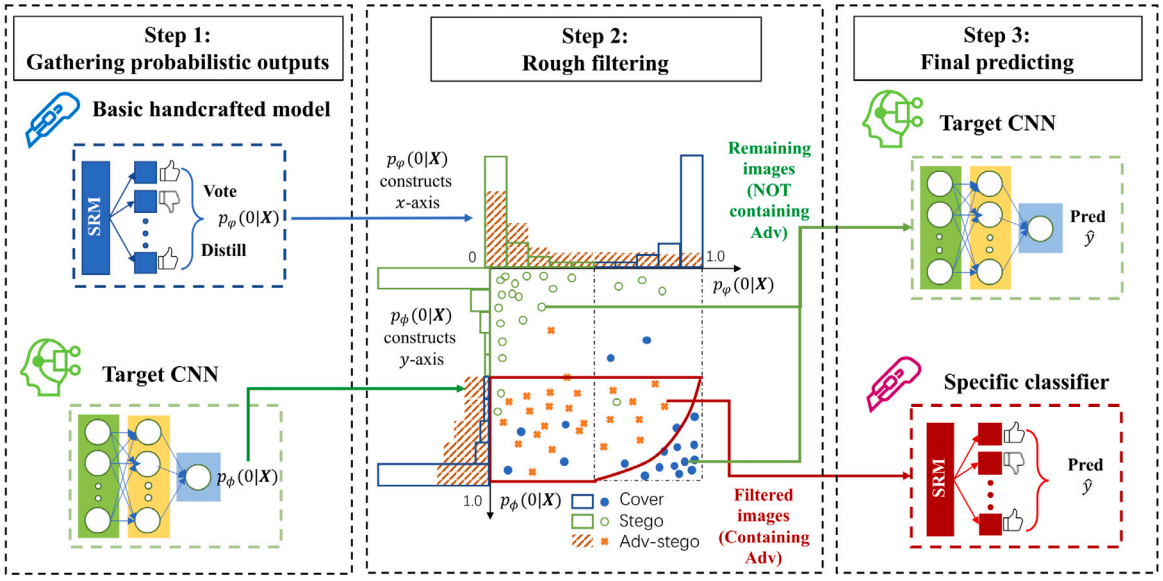


Fig. 1. The architecture the proposed robustness enhancement framework. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.2. The architecture of the proposed framework

Fig. 1 exhibits the architecture of the proposed framework. It consists of a basic handcrafted steganalyzer (in the left column), a rough filter (the middle column), and a specific classifier (in the right column), and a slot for any CNN steganalyzer to fit in (the "Target CNN" in the figure). The classification process of the proposed framework is detailed as follows.

The steganalysis task is accomplished by the proposed framework in three steps. In the first step, the probabilistic outputs of predicting as cover class ($\hat{y} = 0$) of ϕ and φ are collected, i.e., $p_\phi(0|\boldsymbol{X})$ and $p_\varphi(0|\boldsymbol{X})$, where ϕ is the target CNN steganalyzer and φ is a handcrafted steganalyzer trained by conventional stego images. In the second step, the rough filter is utilized to divide the input data into two groups, i.e., *the filtered images* and *the remaining images*. $p_\phi(0|\boldsymbol{X})$ and $p_\varphi(0|\boldsymbol{X})$ construct a two-dimensional space, in which adversarial stego images stand out from the others. Thus the rough filter (the red line in the center of Fig. 1) can divide the input images into the filtered images and the remaining images. In the third step, all the inputs are labeled. The filtered images, which contain a large number of adversarial stego images, are classified by a handcrafted model trained with cover and adversarial stego images. This model is called *the specific classifier*. The remaining images, which contain almost no adversarial stego images, are classified by the CNN steganalyzer ϕ. It is worth noting that the specific classifier is required to be robust against the adversarial steganography, since deceiving multiple CNN steganalyzers is implemented by Zhang et al. [44]. The detailed discussion about the risk of a non-robust specific classifier would be exhibited in Section 4.5.

3.3. The rough filter

The rough filter could be expressed as the boundary in the feature space of $p_\phi(0|\boldsymbol{X})$ and $p_\varphi(0|\boldsymbol{X})$, as shown in Fig. 2. This filter consists of two parts: (1) a label filter that utilizes the label outputs of the CNN steganalyzer and the handcrafted feature-based steganalyzer. (2) A probabilistic filter that filters the adversarial stego images that could not be identified by labels.

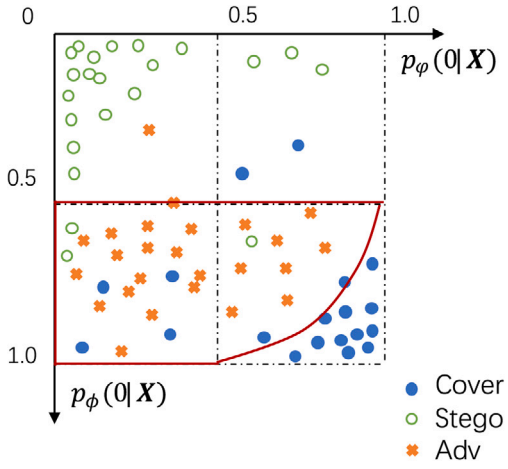


Fig. 2. The rough filter in the two-dimensional feature space (the boundary drawn in red). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

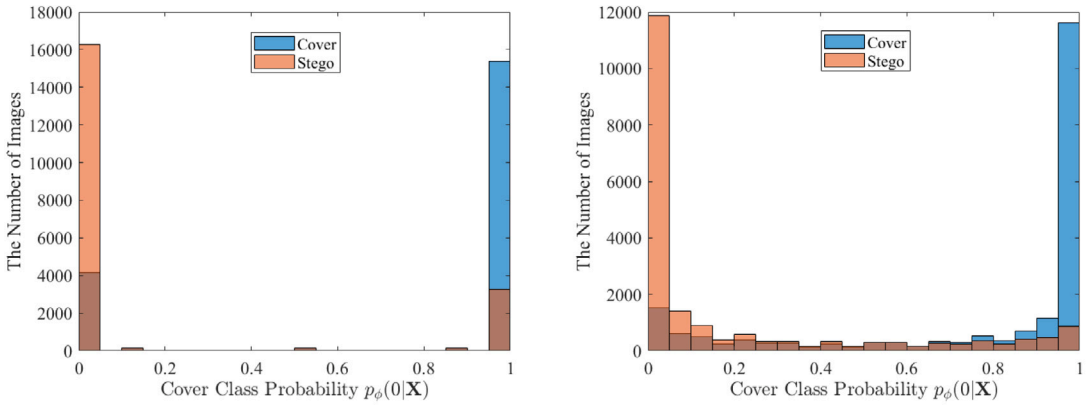
3.3.1. The label filter

Briefly speaking, the label filter filters the images being predicted as cover by the basic handcrafted steganalyzer while being predicted as stego by the target CNN steganalyzer, i.e., $\{\boldsymbol{X}|\hat{y}_\phi(\boldsymbol{X}) = 0, \hat{y}_\varphi(\boldsymbol{X}) = 1\}$.

More detailed than introduced in Section 3.1, 0.4 bpp (bit per pixel) in BOSSBase 1.01 and BOWS2 is taken for instance. ADV-EMB deceives the target SRNet with a 95.00% success rate while only achieving 33.42% missed detection on an S-UNIWARD trained SRM + EC. The missed detection rate gap between two steganalyzers motivates the design of the label filter. Combining the label outputs $\hat{y}_\phi$ and $\hat{y}_\varphi$, plenty of adversarial stego with images $\hat{y}_\phi = 0$ and $\hat{y}_\varphi = 1$ could be filtered out. The specific number of the images filtered out is discussed in Section 4.4.

3.3.2. Generating probabilistic outputs of handcrafted steganalyzers

As presented above, the proposed framework requires the probabilistic outputs of the CNN steganalyzer ϕ and the handcrafted steganalyzer φ. CNN steganalyzers generate probabilistic outputs via softmax



(a) Softmax conversion (b) Softmax function with temperature

Fig. 3. The comparison of SRM + EC probabilistic outputs histograms, (a) before and (b) after the conversion of using softmax function with temperature.

function normalizing class scores (also called logits):

$$p_\phi(i|\boldsymbol{X}) = \frac{e^{z_\phi(i|\boldsymbol{X})}}{\sum_j e^{z_\phi(j|\boldsymbol{X})}}, i, j \in \{0, 1\}, \tag{1}$$

where $z_\phi(i|\boldsymbol{X})$ represents the logits of class i. However, for the widely used handcrafted steganalyzer, ensemble classifier, which consists of a series of base learners and adopts majority voting to decide predicted labels, it does not output class probabilities $p_\varphi(0|\boldsymbol{X})$ and $p_\varphi(1|\boldsymbol{X})$.

The most straightforward way to generate $p_\varphi(i|\boldsymbol{X})$ is to normalize the votes into probabilities by using softmax function:

$$p_\varphi(i|\boldsymbol{X}) = \frac{e^{z_\varphi(i|\boldsymbol{X})}}{\sum_j e^{z_\varphi(j|\boldsymbol{X})}}, i, j \in \{0, 1\}, \tag{2}$$

where $z_\phi(i|\boldsymbol{X})$ represents the vote for class i. But, the probabilities generated in this way cluster tightly near 0 and 1, as shown in Fig. 3-(a). In other words, the adversarial stego images that are predicted as cover will "hide" in the cover cluster. It creates difficulties in filtering them from the input.

To solve this problem, the softmax function with temperature [52] is adopted:

$$p_\varphi(i|\boldsymbol{X}) = \frac{e^{z_\varphi(i|\boldsymbol{X})/T}}{\sum_j e^{z_\varphi(j|\boldsymbol{X})/T}}, i, j \in \{0, 1\}, \tag{3}$$

where T is the temperature. Higher values of T can better scatter the images across $[0, 1]$, as shown Fig. 3-(b). The temperature value is set as $T = 16$ in this paper.

3.3.3. The probabilistic filter

There are some adversarial stego images with both $\hat{y}_\phi = 0$ and $\hat{y}_\varphi = 0$. These adversarial stego images would generate low $p_\phi(0|\boldsymbol{X})$ or $p_\varphi(0|\boldsymbol{X})$.

The payload with 0.4 bpp in BOSSBase 1.01 and BOWS2 is taken for instance. The target CNN model ϕ is SRNet, and the basic handcrafted steganalyzer φ is SRM + EC. Both are trained with cover and S-UNIWARD image pairs. The histograms of the probabilistic outputs $[p_\phi(0|\boldsymbol{X}), p_\varphi(0|\boldsymbol{X}])$ of the target model on cover, ADV-EMB and ADS are exhibited in Fig. 4.

A two-dimensional feature space $[p_\phi(0|\boldsymbol{X}), p_\varphi(0|\boldsymbol{X})]$ is constructed, as shown in Fig. 2. To draw the boundary between adversarial stego images and cover images, a Gaussian kernal SVM as $p_\phi(0|\boldsymbol{X})$ is utilized, and $[p_\phi(0|\boldsymbol{X}), p_\varphi(0|\boldsymbol{X})]$ is taken as the input. The SVM will output the labels that indicate whether to filter.

In summary, for the input, two rules are taken to filter the adversarial stego images: (1) images with labels $\hat{y}_\phi = 0$ and $\hat{y}_\varphi = 1$. (2) The images labeled as adversarial stego by the Gaussian SVM filter from the images with $\hat{y}_\phi = 0$ and $\hat{y}_\varphi = 0$. The filtered images are sent to be classified by the specific classifier, while the remaining images are classified by the CNN steganalyzer ϕ. The framework is shown in Fig. 1.

3.3.4. Explaining the trade-off between high probabilistic outputs and adversarial perturbations

In this section, the theoretical analysis about why adversarial steganography will obtain lower $p_\phi(0|\boldsymbol{X})$ is presented.

First, adversarial perturbations and the probabilistic outputs of the target CNN model $p_\phi(0|\boldsymbol{X})$ are bridged.

For convenience, the gradient map towards the cover class of the input image $\boldsymbol{X}$ is denoted as

$$\boldsymbol{\eta}(\phi, \boldsymbol{X}, 0) = \frac{\partial L_\phi(\boldsymbol{X}, 0)}{\partial \boldsymbol{X}}. \tag{4}$$

Specifically, for ADS [44], which embeds the secret messages on the cover images that are iteratively enhanced by FGSM [53] perturbations, the increase of the probabilistic output $\Delta p_\phi(0|\boldsymbol{X})$ could be estimated by the adversarial perturbations as

$$\Delta p_\phi(0|\boldsymbol{X}) = \sum_{i=1}^{n} \epsilon \ln 2 \int_{\boldsymbol{S}}^{\boldsymbol{S}+\boldsymbol{\Delta}_i} 2^{p_\phi(0|\boldsymbol{Z}_i)} \boldsymbol{\eta}(\phi, \boldsymbol{Z}_i, 0)\, \mathrm{d}\, \boldsymbol{S}, \tag{5}$$

where ϵ denotes the scalar of adversarial perturbations in each iteration of ADS. The deduction process of the above equation is detailed in Appendix. With larger ϵ and more iterations n, the accumulated adversarial perturbations $\boldsymbol{\Delta} = \sum_{i=1}^{n} \epsilon \cdot \boldsymbol{\eta}(\phi, \boldsymbol{Z}_i, 0)$ grow, and the increase of probabilistic output also increases.

ADV-EMB [47] forces the steganographic modification directions to be the same as the gradient directions. Since steganographic modifications are ± 1, the steganographic modifications in adjustable groups could be considered adversarial perturbations with a negative amplitude:

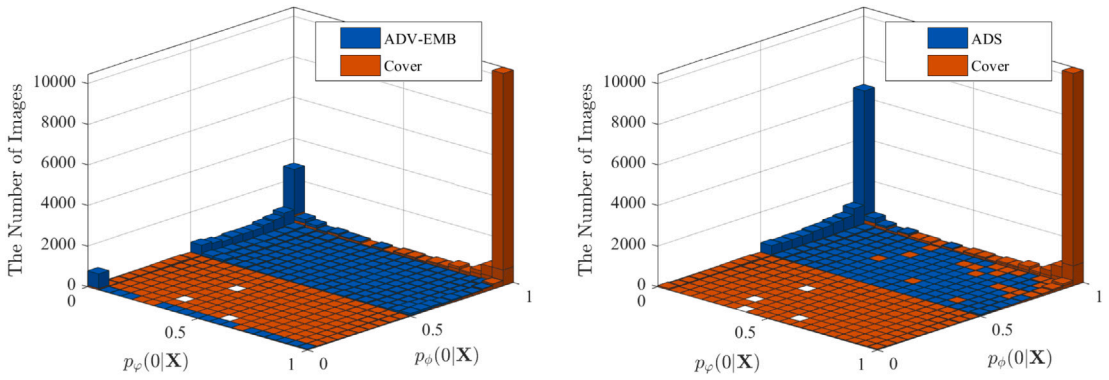
$$\Delta_{x,y} = -\frac{\eta_{x,y}(\phi, \boldsymbol{X}, 0)}{\left|\eta_{x,y}(\phi, \boldsymbol{X}, 0)\right|}, \tag{6}$$

where x, y represent the position of the element.

According to Gibbs constructions [54] in steganography, the steganographic modifications will increase if the modification probabilities of $+1$ and -1 are imbalanced. Hence, with larger adjustable groups, the modification rate of ADV-EMB increases. Still, more image elements in the adjustable group will produce a higher probabilistic output $p_\phi(0|\boldsymbol{X})$ as more adversarial perturbations are generated. A simple experiment in BOSSBase 1.01 [55] and BOWS2 [56] is conducted to show that a larger amount of adjustable group elements will generate a higher $p_\phi(0|\boldsymbol{X})$ while introduce more modifications. The results are shown in Table 1.

Hence, for both ADS and ADV-EMB, the following conclusion can be drawn: the probabilistic outputs of the target CNN steganalyzer $p_\phi(0|\boldsymbol{X})$ are positively correlated with the quantity of adversarial perturbations.

Introducing more adversarial perturbations will expose adversarial stego images to the detection of handcrafted steganalyzers. Specifically,



(a) ADV-EMB (b) ADS

Fig. 4. The bi-histogram of $p_\phi(0|\boldsymbol{X})$ and $p_\varphi(0|\boldsymbol{X})$ of cover and (a) ADV-EMB or (b) ADS. The target CNN steganalyzer is SRNet trained on cover and S-UNIWARD image pairs from BOSSBase 1.01 and BOWS2.

Table 1
The mean modification number, $p_\phi(0|\boldsymbol{X})$ and the average detection error rate under SRM+EC of ADV-EMB with different β under relative payload 0.4 bpp.

β	0.1	0.3	0.5	0.7	0.9
Average L_0 distance	2478.37	2528.61	2628.86	2733.42	2832.49
Average $p_\phi(0|\boldsymbol{X})$	0.1481	0.4873	0.7465	0.8550	0.8950

ADS introduces larger perturbations than ADV-EMB. The average L_2 distortions per image caused by ADS and ADV-EMB and the detected rate on ADS and ADV-EMB of the basic handcrafted steganalyzer φ are compared. The target CNN steganalyzer is SRNet trained on cover and S-UNIWARD image pairs. The results are shown in Table 2.

From Table 2, one can observe the distortions caused by ADS are clearly larger than those caused by ADV-EMB. Hence, according to Fig. 4 and Table 2, ADS is easily detected by the basic handcrafted feature-based steganalyzer φ. In Fig. 4, ADS is harder to distinguish from cover images than ADV-EMB.

From the analysis above, adversarial steganography either introduces more adversarial perturbations, which will lead to detection by handcrafted steganalyzers, or subtly controls the perturbations, which leads to identification by the probabilistic filter.

3.4. The specific classifier

The specific classifier is set to accurately classify the filtered images, which are mainly the adversarial stego images and the cover images. In this section, we introduce the structure of the specific classifier.

The best-performing structure of a steganalyzer is CNN. However, it has been proven that CNN steganalyzers are vulnerable to adversarial steganography. Based on the experiments from the paper of Tang et al. [47], the retrained CNN steganalyzers are still vulnerable to adversarial steganography targeting at the updated model. Furthermore, ADS has a version that targets multiple models [44]. Referring to adversarial attacks in computer vision, which is similar to adversarial steganography, Carlini et al. [57] have also shown that an extra CNN detection model would provide little improvement in robustness.

There are underivable operations in the process of handcrafted feature extraction [25], such as calculating co-occurrence matrices or histograms, and residual map truncations. These operations make generating adversarial stego images against handcrafted steganalyzers much harder than the equivalent process for CNN steganalyzers. Furthermore, attacking two steganalyzers with totally different structures at the same time is even harder. SRM [25] and GFR [19] incorporated with the ensemble classifiers [26] are the state-of-the-art handcrafted models. For these reasons, SRM + EC and GFR + EC are utilized as the specific classifiers in the spatial domain and JPEG domain, respectively. Moreover, the specific classifier is trained with the same cover images as ϕ and φ and the corresponding adversarial stego images targeting ϕ, as the steganalyzer would obtain a higher detection ability with paired training. During the testing period, the specific classifier only processes the filtered images.

Note that setting the specific classifier is different from the retraining strategy. As mentioned before, retrained CNN steganalyzers are still vulnerable to adversarial steganography. Besides, retraining requires seeing every type of adversarial steganography, while the specific classifier only requires seeing one type of it. Specifically, in this paper, the specific classifier is only trained with ADV-EMB [47]. On the other hand, the proposed scheme is robust against updated adversarial steganography. The specific classifier is set to reduce the misclassification of cover images.

4. Numerical evaluations and comparisons

To evaluate the performance of the proposed robustness enhancement framework, the following experiments are conducted.

1. As stressed in this paper, adversarial stego images are now mixed with cover and conventional stego images in real-world data streams. The AUCs (area under the curve) are taken to evaluate the performance of the enhanced CNN steganalyzers and the previous works. It will be reported in Section 4.2.
2. Before the proposed framework, retraining was the only effective way to defend against adversarial steganography. The detection accuracy on the adversarial steganography targeting the current CNN steganalyzer is taken as the robustness measurement. The robustness of the enhanced CNN steganalyzers and retrained ones are compared in Section 4.3.
3. In the proposed framework, the rough filter and the specific classifier are two key constitutions. The former filters adversarial stego images out of the input. The latter guarantees lower false alarm rates. In Sections 4.4 and 4.5, they are analyzed respectively.
4. To better exhibit how the proposed framework works, the classification process of several example images are visualized in Section 4.6.

4.1. Settings

1. **Image Sets:** In this paper, two widely-used image datasets BOSSBase 1.01 [55] and BOWS2 [56] are adopted. Each dataset contains 10,000 grayscale images of size 512 × 512. To train

Table 2
The comparison of the average L_2 distortions, $p_\phi(0|\boldsymbol{X})$ and the detection rate under φ between ADV-EMB and ADS.

Payload	L_2		$p_\phi(0\|\boldsymbol{X})$		Detected rate (%)	
	ADV-EMB [47]	ADS [44]	ADV-EMB [47]	ADS [44]	ADV-EMB [47]	ADS [44]
0.1	22.12	**66.82**	0.6141	**0.7150**	56.98	**60.86**
0.2	33.31	**87.41**	0.6705	**0.7548**	61.61	**67.63**
0.3	42.95	**85.75**	0.7151	**0.8163**	62.19	**73.68**
0.4	50.90	**85.17**	0.7802	**0.8163**	67.55	**78.61**
0.5	58.70	**95.31**	0.8302	**0.8680**	77.82	**83.46**

CNN steganalyzers, the images are resized to 256 × 256 by the MATLAB function **imresize()** with the default settings. For the JPEG domain experiments, the images of size 256 × 256 are compressed into JPEG format with the quality factor of 75. 14 000 images are randomly selected as the training set. Another random 1000 images from the validation set and the other 5000 images from the testing set.

2. **Steganographic Methods:** For adversarial steganographic methods, two state-of-the-art methods, namely, ADV-EMB [47] and ADS [44], are selected. For conventional steganographic methods, S-UNIWARD [13] and J-UNIWARD [13] are adopted in the spatial and the JPEG domain, respectively. They are also the base cost functions utilized for ADV-EMB and ADS. All conventional and adversarial stego images are generated using the optimal embedding simulator.
3. **Steganalyzers:** Three state-of-the-art CNN steganalyzers, namely, SRNet [31] and SiaStegNet [35], are selected. SRNet is utilized in both the spatial and JPEG domains, while SiaStegNet is only used in the spatial domain due to its design. CNN steganalyzers are trained with default settings introduced in their papers. The target CNN steganalyzers ϕ are trained with cover and conventional stego image pairs.
   For handcrafted steganalyzers, ensemble classifier [26] is trained with two state-of-the-art feature sets SRM [25] and GFR [19] in the spatial and JPEG domains, respectively. The retrained SRM + EC and GFR + EC are trained with the ratio of cover: stego: adversarial stego = 2:1:1.
   In the proposed scheme, the basic handcrafted steganalyzers φ are trained with cover and conventional stego image pairs. Please note that we set the basic handcrafted steganalyzers trained with conventional stego images with a fixed relative payload 0.5 bpp (or bpnzAC) to compress the false alarm rate. The specific classifiers of the proposed scheme are trained with cover and the adversarial stego images targeting CNN steganalyzers ϕ. Since ensemble classifiers optimize their parameters through cross-validation, 15 000 image pairs are utilized to train ensemble classifiers.

4.2. Performance comparisons in the real-world scenario

In real-world steganalysis, the input images include cover, conventional stego and adversarial stego images. To comprehensively compare the robustness enhanced CNN steganalyzers with the previous works, metrics that address all accuracies on cover, stego and adversarial stego images should be adopted. Therefore, in this subsection, the AUC (area under the curve) is adopted as the metric. AUC is a classic evaluation metric of binary classification tasks. It reflects the comprehensive performance of models under different classification thresholds. Since steganalysis is a binary classification problem, the ratio of class 0 (cover) and class 1 (stego) is kept 1 : 1. The ratio of conventional and adversarial stego images is 1 : 1. Thus, the testing data consist of all cover, conventional stego and adversarial stego images with proportions 2 : 1 : 1. Note that the adversarial stego images are all targeting the tested CNN steganalyzers. The results are shown in Fig. 5.

It is evident that the robustness enhanced CNN steganalyzers outperform the original CNN steganalyzers and the handcrafted ones. The largest gap is 0.0541 when the target model SRNet is trained with S-UNIWARD under payload 0.4 bpp. Furthermore, to visually exhibit the AUC comparison, in Fig. 6, we draw the receiver operating characteristic (ROC) curves of SRNet, SRM + EC and the proposed scheme. The target model is SRNet trained with S-UNIWARD under a relative payload of 0.4 bpp.

Table 3
The quantity of each type of the images in the filtered images.

Payload (bpp)	0.1	0.2	0.3	0.4	0.5
Cover	1972	1315	1262	1044	1129
Stego	1695	832	668	352	384
Adversarial stego	3490	3211	3476	3835	4373
Total	7157	5358	5406	5231	5886

4.3. Comparing robustness with retraining

Before the proposed framework, retraining is the only way to defend against adversarial steganography. Bernard et al. [48,49] discussed a min–max retraining strategy for CNN steganalyzers. In this section, the robustness of the min–max retraining and the proposed framework are compared. The robustness is evaluated by the detection accuracy of the adversarial stego images that target it.

The min–max strategy involves 8 and 6 iterations in the JPEG domain and the spatial domain, respectively. The model structure we adopt is SRNet, and the conventional steganographic methods are S-UNIWARD and J-UNIWARD. The tested relative payload is 0.4 bpp (or bpnzAC).

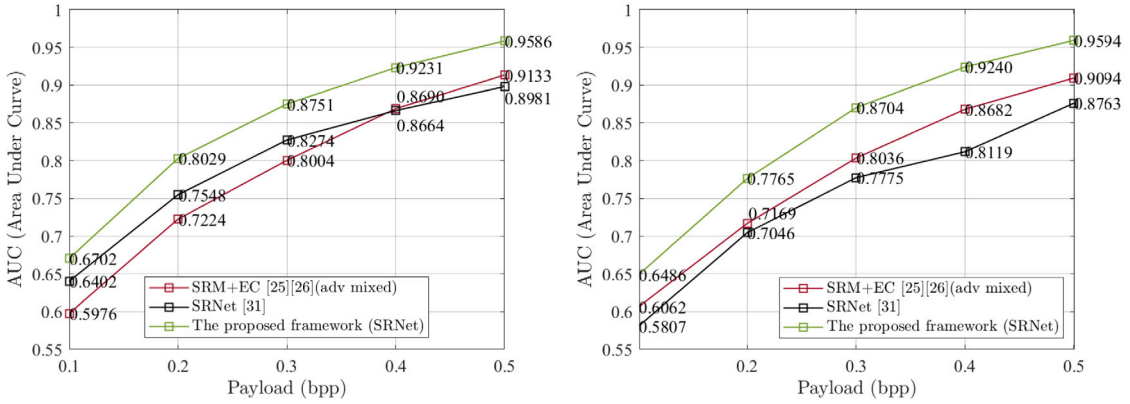
According to the results in Fig. 7, little improvement is brought by the min–max strategy. Even in the best case of retraining, i.e. in the 6th iteration and the spatial domain, the detection accuracy on ADV-EMB is almost 30% lower than the proposed framework. Moreover, the min–max retrained CNN steganalyzer is still vulnerable to unknown adversarial steganographic methods. In the last iteration of the spatial domain, the ADV-EMB retrained CNN steganalyzer can only detect 0.32% ADS [44] stego images targeting it. While the proposed framework, which is also not trained with any ADS stego images, detects 71.32% of the ADS stego images targeting ϕ. Hence, the robustness brought by the proposed scheme is clearly superior.

4.4. The effect of the rough filter

The rough filter is set to filter out adversarial stego images from the input. Thus, the constitutions of the filtered images are displayed in this section. Note that the total number of images belonging to each type (cover, conventional stego, adversarial stego) is 5000.

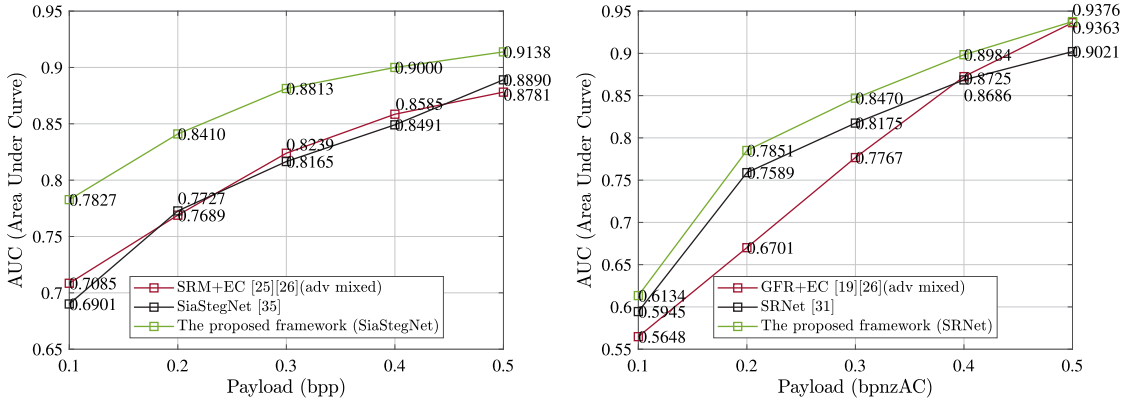
The rough filter consists two parts: (1) the label filter that filters images with $\hat{y}_\phi = 0$ and $\hat{y}_\varphi = 1$ and (2) the probabilistic filter that filters the images with lower $p_\phi(0|\boldsymbol{X})$ and $p_\varphi(0|\boldsymbol{X})$.

The experiments are conducted as SRNet is the target steganalyzer of the adversarial steganographer, which utilizes ADV-EMB to generate adversarial stego images in the spatial domain. The constituents of the images filtered by the rough filters are shown in Table 3. Taking relative payload 0.5 bpp for instance, the filtered images consist of



(a) Domain: spatial, target model: SRNet, adversarial steganography: ADV-EMB

(b) Domain: spatial, target model: SRNet, adversarial steganography: ADS



(c) Domain: spatial, target model: SiaStegNet, adversarial steganography: ADV-EMB

(d) Domain: JPEG, target model: SRNet, adversarial steganography: ADV-EMB

Fig. 5. The comparisons on AUC (Area Under Curve) of SRNet, retrained SRM + EC (or GFR + EC) and the proposed scheme in the spatial domain (a,b,c) and the JPEG domain (d).

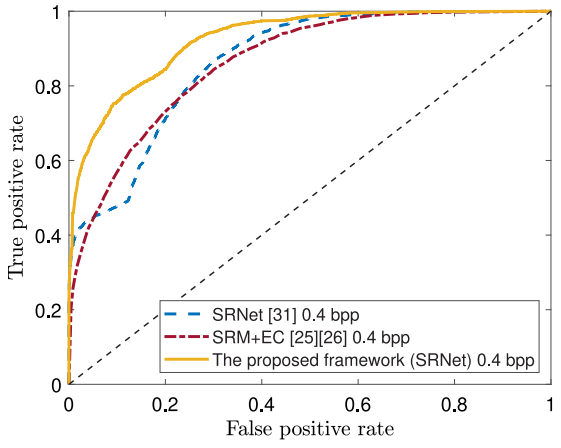


Fig. 6. The ROC curves of SRM + EC (mix trained), SRNet and the proposed scheme.

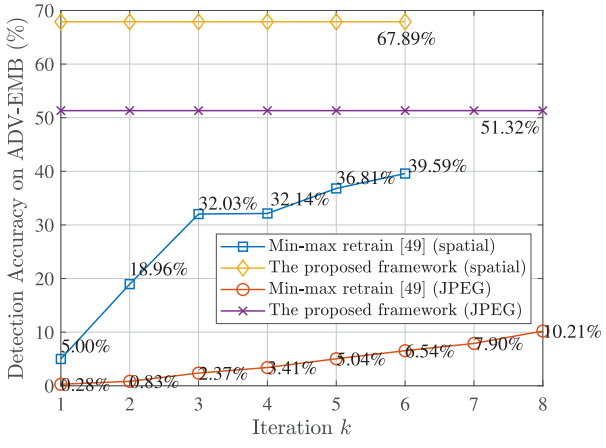


Fig. 7. The detection accuracy of the steganalyzer in kth iteration on the ADV-EMB targeting it. It represents the robustness of the steganalyzers.

1129 cover images, 4373 adversarial stego images and 384 conventional stego images. This result proves the rough filter's effectiveness, which guarantees that most of the adversarial stego images would not be classified by the target CNN steganalyzers.

4.5. The effect of the specific classifier

The specific classifier aims to classify the filtered images, which consist of mostly adversarial stego images and some cover images. First, the impact on the detection accuracy of cover/conventional

Table 4
The detection accuracy of the schemes with and without the specific classifier. We take ADV-EMB and the SRNet trained with cover and S-UNIWARD image pairs as the example. The specific classifier is abbreviated as SC.

	Payload (bpp)	0.1	0.2	0.3	0.4	0.5
Cover	With SC	**52.68%**	**60.10%**	**74.48%**	**80.30%**	**90.52%**
	Without SC	34.02%	47.52%	51.46%	66.86%	71.24%
Stego	With SC	77.80%	89.12%	94.20%	96.78%	98.36%
	Without SC	**92.90%**	**95.68%**	**97.94%**	**98.48%**	**99.52%**
Adversarial stego	With SC	52.08%	53.46%	58.74%	67.89%	77.48%
	Without SC	**81.84%**	**75.28%**	**74.48%**	**82.42%**	**91.74%**

Table 5
The successful rate of ADV-EMB [47] to deceive CNN steganalyzers utilized as the specific classifier.

	YeNet [27]	retrained SRNet [31]
Success rate	75.36%	68.40%

stego/adversarial stego images is exhibited. A straightforward way to deal with the filtered images is to label them as stego. As shown in Table 3, the filtered images are mainly adversarial stego, especially under relatively high payloads. However, under relatively low payloads, the proportion of cover images significantly increases. It could be anticipated that the detection accuracy on the cover images would decrease if the filtered images are all labeled as stego. SRNet and ADV-EMB in the spatial domain are taken as an example to compare the detection accuracy as with and without the specific classifier. The results are shown in Table 4.

From Table 4, one can observe that although the detection accuracy on adversarial stego images increases, the detection accuracy on cover images decreases. Since current adversarial steganographic methods all require complete access to the target model, one can anticipate that adversarial stego images would be relatively rare in real-world data streams. So the detection accuracy of cover images is preferred. Hence, the specific classifier is required in the proposed framework to guarantee higher cover detection accuracy.

The proposed framework takes handcrafted model as the specific classifier. It is because stacking multiple CNN steganalyzers can hardly increase robustness. Thus, in this section, experiments are conducted to show the vulnerability of using CNN model as the specific classifier.

First, the previous literature has proven so. Zhang et al. [44] proposed to deceive multiple CNN steganalyzers by calculating the gradient map of the weighted loss regarding the input image. This version of ADS has a 67.3% to a success rate of deceiving YeNet [27], XuNet [29] and WuNet [58] at the same time under a relative payload of 0.4 bpp.

Second, deceiving multiple CNN steganalyzers is not discussed in ADV-EMB [47]. An intuitive method can achieve such an aim by using the average gradient map of multiple CNN steganalyzers to adjust the embedding costs.

In the experiment of this section, SRNet trained with cover and S-UNIWARD [13] is taken as the target CNN model ϕ. Two CNN steganalyzers are selected as the specific classifier: (1) YeNet with the same training set and (2) retrained SRNet. By utilizing the adaptive adversarial steganography we introduced in the last paragraph, the success rate of attack is shown in Table 5. For the first instance, where the specific classifier has a different structure from the target model, ADV-EMB can deceive them both at the same time at a rate of 75.36%. For the second instance, if the specific classifier is retrained, the adversarial steganographer can still fool both of them at a rate of 68.40%.

Hence, regardless of which adversarial steganography we select (ADV-EMB or ADS), utilizing the CNN model as the specific classifier provides little improvement in robustness. In comparison, using handcrafted feature-based steganalyzers is more secure than using CNN steganalyzers.

4.6. Visualizing classification process via some examples

To better exhibit how the proposed framework filters and classifies adversarial stego images. In this section, the classification process of several example cover/adversarial stego images is exhibited. Specifically, 2 adversarial stego images and 2 cover images are selected. The adversarial stego image (1013_a.pgm) and its corresponding cover image (1013_c.pgm) are both selected. Thus, in the rough filter part (the middle in Fig. 8), the modification maps are displayed alongside the stego images to differentiate them. The target CNN steganalyzer is SiaStegNet [35]. The relative payload is 0.4 bpp (bit per pixel).

One can take 1013_a.pgm and 1013_c.pgm as examples. First, the cover class probabilistic outputs of the input, i.e., $p_\phi(0|\boldsymbol{X})$ and $p_\varphi(0|\boldsymbol{X})$ are collected. For the adversarial stego image 1013_a.pgm, the target CNN steganalyzer ϕ predicts it with $p_\phi(0|\boldsymbol{X}) = 0.7430$, and the basic handcrafted steganalyzer φ predicts it with $p_\varphi(0|\boldsymbol{X}) = 0.0183$. For the cover image 1013_c.pgm, ϕ and φ predict it with $p_\phi(0|\boldsymbol{X}) = 1.0$ and $p_\varphi(0|\boldsymbol{X}) = 0.9661$. Second, the rough filter divides the input stream. Intuitively, every input image obtains its own locations with $(p_\phi(0|\boldsymbol{X}), p_\phi(0|\boldsymbol{X}))$. The ones located in the filtered area (circled with the red line in the middle of Fig. 8) will be divided into *the filtered images*, and ones outside are belong to *the remaining images*. 1013_a.pgm clearly locates in the filtered area. Thus, it will be sent to the specific classifier. While 1013_c.pgm stands outside the red circle. Thus, it will be labeled by ϕ then. Third, the specific classifier and the target CNN steganalyzer will label the filtered images and the remaining images respectively. 1013_a.pgm is labeled by the specific classifier as "STEGO" ($\hat{y} = 1$), and 1013_c.pgm is labeled by ϕ as "COVER" ($\hat{y} = 0$). They are either correctly classified.

For the two adversarial stego images, i.e., 1013_a.pgm and 8576.pgm, they are both misclassified by ϕ ($p_\phi(0|\boldsymbol{X}) > 0.5$). But, with the proposed framework, they are filtered and get correctly classified by the specific classifier. Meanwhile, it can be observed that the example cover images (1013_c.pgm and 1000.pgm) are located outside the filtered area and correctly labeled by ϕ. These instances show how the proposed framework filters adversarial stego images and maintains the high detection accuracy of CNN steganalyzers on cover images.

5. Conclusions

Adversarial steganography threatens the security of CNN steganalyzers. Before the proposed framework, only retraining was utilized to defend against it. However, retrained steganalyzers are still vulnerable to adversarial steganography targeting them. In this paper, a robustness enhancement framework is proposed. It filters adversarial stego images in the two-dimensional feature space constructed by the probabilistic outputs of steganalyzers. Extensive experiments show it can enhance CNN steganalyzers of different structures in different domains. Thus, in the real-world scenario, where adversarial stego images are mixed with conventional stego images, the robustness enhanced CNN steganalyzers clearly outperform the previous works. Specifically, the largest improvement is 0.0926 with SiaStegNet as the target model under relative payload 0.1 bpp. Compared with retraining, the proposed framework brings much more robustness improvements and can detect unknown adversarial steganography. Even if trained with ADV-EMB for 6 iterations, the success rate of ADS against the target model is 99.68%, while that of the enhanced model of the proposed framework is only 38.68%. Moreover, this paper also reveals the characteristic of adversarial steganography in the probabilistic outputs of steganalyzers. The tested methods (ADV-EMB [47] and ADS [44]) are filtered due to either being classified as stego by the handcrafted steganalyzer or generating lower probabilistic outputs.

Since handcrafted steganalyzers are utilized in the proposed framework, more flexible handcrafted steganalyzers for enhancing CNN steganalyzers robustness against adversarial steganography will be considered in the future. For instance, reducing the dimensionality of handcrafted features. Additionally, the specific classifier is still trained with some adversarial stego images. Designing specific classifiers never see any adversarial steganography is also part of our future work.

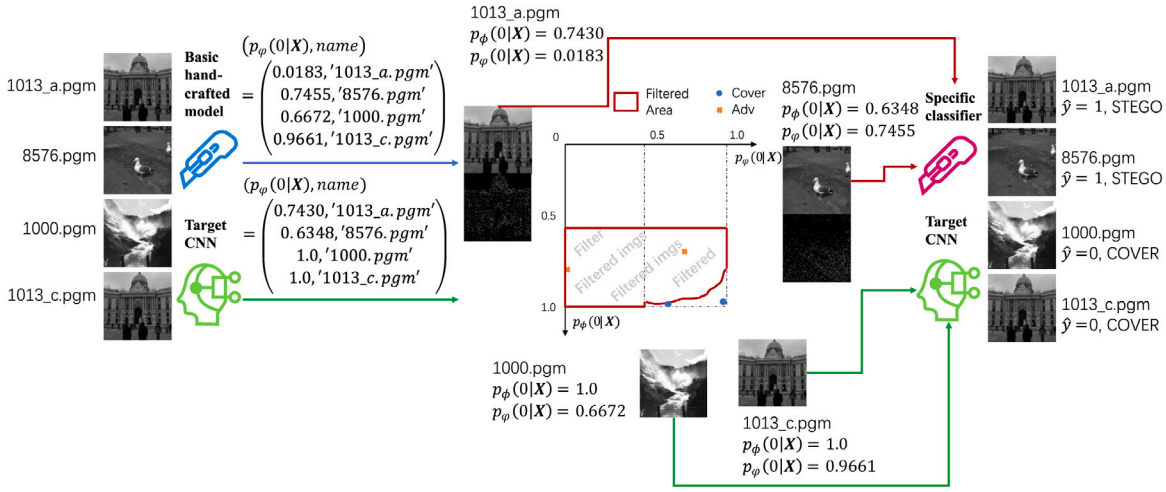


Fig. 8. The classification processes of several examples. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

CRediT authorship contribution statement

Chuan Qin: Conceptualization, Methodology, Software, Investigation, Data curation, Writing – original draft, Visualization. **Weiming Zhang:** Conceptualization, Formal analysis, Writing – review & editing, Funding acquisition. **Hang Zhou:** Software, Writing – review & editing. **Jiayang Liu:** Validation, Writing – review & editing. **Yuan He:** Writing – review & editing, Resources. **Nenghai Yu:** Resources, Funding acquisition.

Declaration of competing interest

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

Appendix. The relation between adversarial perturbations and the probabilistic outputs

In the appendix, we deduct the Eq. (5) with details. The cross-entropy loss for classification is

$$L_\phi(\boldsymbol{X}, y) = -[y \cdot \log(1 - p_\phi(0|\boldsymbol{X})) + (1 - y)\log(p_\phi(0|\boldsymbol{X}))], \tag{A.1}$$

where $p_\phi(0|\boldsymbol{X})$ denotes the probabilistic output of the CNN model ϕ predicting the input $\boldsymbol{X}$ as cover.

When the target class is cover ($y = 0$), the cross-entropy loss utilized in CNN steganalyzers is

$$L_\phi(\boldsymbol{X}, 0) = -\log(p_\phi(0|\boldsymbol{X})), \tag{A.2}$$

so the gradient map of the cross-entropy loss $L_\phi(\boldsymbol{X}, 0)$ with reference to the input image $\boldsymbol{X}$ is

$$\boldsymbol{\eta}(\phi, \boldsymbol{X}, 0) = \frac{\partial[-\log(p_\phi(0|\boldsymbol{X}))]}{\partial \boldsymbol{X}}. \tag{A.3}$$

It is well-known that the adversarial perturbations in both ADS and ADV-EMB can be expressed as $\epsilon \cdot \boldsymbol{\eta}(\phi, \boldsymbol{X}, 0)$. Since $\epsilon < 0$, increasing adversarial perturbations will reduce the cross-entropy loss $L_\phi(\boldsymbol{X}, 0) = -\log(p_\phi(0|\boldsymbol{X}))$. At the same time, it is obvious that the corresponding probabilistic output $p_\phi(0|\boldsymbol{X})$ will increase.

Specifically, for ADS, we calculate the gradient map of the probabilistic outputs with reference to the input $\boldsymbol{\vartheta}(\phi, \boldsymbol{X}, 0)$ as follows to obtain the explicit relation between the gradient map of the probabilistic outputs and the gradient map of the classification loss,

$$\begin{aligned} \boldsymbol{\vartheta}(\phi, \boldsymbol{X}, 0) &= \frac{\partial p_\phi(0|\boldsymbol{X})}{\partial \boldsymbol{X}} \\ &= \frac{\partial p_\phi(0|\boldsymbol{X})}{\partial - \log(p_\phi(0|\boldsymbol{X}))} \cdot \frac{\partial - \log(p_\phi(0|\boldsymbol{X}))}{\partial \boldsymbol{X}} \\ &= -\ln 2 \cdot 2^{p_\phi(0|\boldsymbol{X})} \cdot \boldsymbol{\eta}(\phi, \boldsymbol{X}, 0). \end{aligned} \tag{A.4}$$

As mentioned in Section 3.3.4, ADS accumulates the perturbations in each round until the enhanced cover images are predicted as cover with secret messages embedded. The accumulated distortions could be expressed as

$$\boldsymbol{\Delta} = \sum_{i=1}^{n} \epsilon \cdot \boldsymbol{\eta}(\phi, \boldsymbol{Z}_i, 0). \tag{A.5}$$

According to Eq. (A.4), the increase in probabilistic output $\Delta p_\phi(0|\boldsymbol{X})$ in ADS could be estimated by the adversarial perturbations as follows.

$$\begin{aligned} \Delta p_\phi(0|\boldsymbol{X}) &= \sum_{i=1}^{n} -\epsilon \int_{\boldsymbol{S}}^{\boldsymbol{S}+\boldsymbol{\Delta}_i} \boldsymbol{\vartheta}(\phi, \boldsymbol{Z}_i, 0)\, \mathrm{d}\, \boldsymbol{S} \\ &= \sum_{i=1}^{n} \epsilon \ln 2 \int_{\boldsymbol{S}}^{\boldsymbol{S}+\boldsymbol{\Delta}_i} 2^{p_\phi(0|\boldsymbol{Z}_i)} \boldsymbol{\eta}(\phi, \boldsymbol{Z}_i, 0)\, \mathrm{d}\, \boldsymbol{S}. \end{aligned} \tag{A.6}$$

where ϵ denotes the scalar of the adversarial perturbations in each iteration of ADS.

References

[1] Cheddad A, Condell J, Curran K, Mc Kevitt P. Digital image steganography: Survey and analysis of current methods. Signal Process 2010;90(3):727–52. http://dx.doi.org/10.1016/j.sigpro.2009.08.010, URL https://www.sciencedirect.com/science/article/pii/S0165168409003648.

[2] Li B, He J, Huang J, Shi YQ. A survey on image steganography and steganalysis. J Inf Hiding Multimedia Signal Process 2011;2(2):142–72.

[3] Ren Y, Cai S, Wang L. Secure AAC steganography scheme based on multi-view statistical distortion (SofMvD). J Inf Secur Appl 2021;59:102863. http://dx.doi.org/10.1016/j.jisa.2021.102863.

[4] Sahu AK, Swain G, Sahu M, Hemalatha J. Multi-directional block based PVD and modulus function image steganography to avoid FOBP and IEP. J Inf Secur Appl 2021;58:102808. http://dx.doi.org/10.1016/j.jisa.2021.102808.

[5] Chen B, Luo W, Zheng P, Huang J. Universal stego post-processing for enhancing image steganography. J Inf Secur Appl 2020;55:102664. http://dx.doi.org/10.1016/j.jisa.2020.102664.

[6] Fridrich J. Steganography in digital media: principles, algorithms, and applications. Cambridge University Press; 2009.

[7] Zhang X, Peng F, Long M. Robust coverless image steganography based on DCT and LDA topic classification. IEEE Trans Multimed 2018;20(12):3223–38.

[8] Lin G-S, Chang Y-T, Lie W-N. A framework of enhancing image steganography with picture quality optimization and anti-steganalysis based on simulated annealing algorithm. IEEE Trans Multimed 2010;12(5):345–57.

[9] Lin Y-T, Wang C-M, Chen W-S, Lin F-P, Lin W. A novel data hiding algorithm for high dynamic range images. IEEE Trans Multimed 2016;19(1):196–211.

[10] Filler T, Judas J, Fridrich J. Minimizing embedding impact in steganography using trellis-coded quantization. In: Media forensics and security ii, Vol. 7541. International Society for Optics and Photonics; 2010, 754105.

[11] Filler T, Fridrich J. Minimizing additive distortion functions with non-binary embedding operation in steganography. In: 2010 IEEE international workshop on information forensics and security. IEEE; 2010, p. 1–6.

[12] Holub V, Fridrich J. Designing steganographic distortion using directional filters. In: 2012 IEEE international workshop on information forensics and security (WIFS). IEEE; 2012, p. 234–9.

[13] Holub V, Fridrich J, Denemark T. Universal distortion function for steganography in an arbitrary domain. EURASIP J Inf Secur 2014;2014(1):1.

[14] 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). IEEE; 2014, p. 4206–10.

[15] Sedighi V, Cogranne R, Fridrich J. Content-adaptive steganography by minimizing statistical detectability. IEEE Trans Inf Forensics Secur 2015;11(2):221–34.

[16] Guo L, Ni J, Su W, Tang C, Shi Y-Q. Using statistical image model for JPEG steganography: uniform embedding revisited. IEEE Trans Inf Forensics Secur 2015;10(12):2669–80.

[17] Cogranne R, Giboulot Q, Bas P. Steganography by minimizing statistical detectability: The cases of JPEG and color images. In: Proceedings of the 2020 ACM workshop on information hiding and multimedia security, 2020, p. 161–7.

[18] Lie W-N, Lin G-S. A feature-based classification technique for blind image steganalysis. IEEE Trans Multimed 2005;7(6):1007–20.

[19] Song X, Liu F, Yang C, Luo X, Zhang Y. Steganalysis of adaptive JPEG steganography using 2D gabor filters. In: Proceedings of the 3rd ACM workshop on information hiding and multimedia security. ACM; 2015, p. 15–23.

[20] Li B, Li Z, Zhou S, Tan S, Zhang X. New steganalytic features for spatial image steganography based on derivative filters and threshold LBP operator. IEEE Trans Inf Forensics Secur 2017;13(5):1242–57.

[21] Qiao T, Luo X, Wu T, Xu M, Qian Z. Adaptive steganalysis based on statistical model of quantized DCT coefficients for JPEG images. IEEE Trans Dependable Secure Comput 2019;1.

[22] Li X, Li B, Luo X, Yang B, Zhu R. Steganalysis of a PVD-based content adaptive image steganography. Signal Process 2013;93(9):2529–38. http://dx.doi.org/10.1016/j.sigpro.2013.03.029, URL https://www.sciencedirect.com/science/article/pii/S0165168413001187.

[23] Jin Z, Feng G, Ren Y, Zhang X. Feature extraction optimization of JPEG steganalysis based on residual images. Signal Process 2020;170:107455. http://dx.doi.org/10.1016/j.sigpro.2020.107455, URL https://www.sciencedirect.com/science/article/pii/S0165168420300025.

[24] Wang P, Liu F, Yang C. Towards feature representation for steganalysis of spatial steganography. Signal Process 2020;169:107422. http://dx.doi.org/10.1016/j.sigpro.2019.107422, URL https://www.sciencedirect.com/science/article/pii/S0165168419304748.

[25] Fridrich J, Kodovsky J. Rich models for steganalysis of digital images. IEEE Trans Inf Forensics Secur 2012;7(3):868–82.

[26] Kodovsky J, Fridrich J, Holub V. Ensemble classifiers for steganalysis of digital media. IEEE Trans Inf Forensics Secur 2011;7(2):432–44.

[27] Ye J, Ni J, Yi Y. Deep learning hierarchical representations for image steganalysis. IEEE Trans Inf Forensics Secur 2017;12(11):2545–57.

[28] Qian Y, Dong J, Wang W, Tan T. Learning representations for steganalysis from regularized cnn model with auxiliary tasks. In: Proceedings of the 2015 international conference on communications, signal processing, and systems. Springer; 2016, p. 629–37.

[29] Xu G, Wu H-Z, Shi Y-Q. Structural design of convolutional neural networks for steganalysis. IEEE Signal Process Lett 2016;23(5):708–12.

[30] Zeng J, Tan S, Li B, Huang J. Large-scale JPEG image steganalysis using hybrid deep-learning framework. IEEE Trans Inf Forensics Secur 2018;13(5):1200–14. http://dx.doi.org/10.1109/TIFS.2017.2779446.

[31] Boroumand M, Chen M, Fridrich J. Deep residual network for steganalysis of digital images. IEEE Trans Inf Forensics Secur 2019;14(5):1181–93.

[32] Wu S, Zhong S-h, Liu Y. A novel convolutional neural network for image steganalysis with shared normalization. IEEE Trans Multimed 2019;22(1):256–70.

[33] Li B, Wei W, Ferreira A, Tan S. ReST-Net: DIverse activation modules and parallel subnets-based CNN for spatial image steganalysis. IEEE Signal Process Lett 2018;25(5):650–4. http://dx.doi.org/10.1109/LSP.2018.2816569.

[34] Zeng J, Tan S, Liu G, Li B, Huang J. WISERNet: WIder separate-then-reunion network for steganalysis of color images. IEEE Trans Inf Forensics Secur 2019;14(10):2735–48. http://dx.doi.org/10.1109/TIFS.2019.2904413.

[35] You W, Zhang H, Zhao X. A siamese CNN for image steganalysis. IEEE Trans Inf Forensics Secur 2021;16:291–306. http://dx.doi.org/10.1109/TIFS.2020.3013204.

[36] Tan S, Wu W, Shao Z, Li Q, Li B, Huang J. CALPA-NET: channel-pruning-assisted deep residual network for steganalysis of digital images. IEEE Trans Inf Forensics Secur 2021;16:131–46. http://dx.doi.org/10.1109/TIFS.2020.3005304.

[37] Butora J, Yousfi Y, Fridrich JJ. How to pretrain for steganalysis. In: Borghys D, Bas P, Verdoliva L, Pevný T, Li B, Newman J, editors. IH&MMSec '21: ACM workshop on information hiding and multimedia security, virtual event, Belgium, June, 22-25, 2021. ACM; 2021, p. 143–8. http://dx.doi.org/10.1145/3437880.3460395.

[38] Deng X, Chen B, Luo W, Luo D. Fast and effective global covariance pooling network for image steganalysis. In: Proceedings of the ACM workshop on information hiding and multimedia security, 2019, p. 230–4.

[39] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R. Intriguing properties of neural networks. 2013, arXiv preprint arXiv:1312.6199.

[40] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. 2014, arXiv preprint arXiv:1412.6572.

[41] Papernot N, McDaniel P, Jha S, Fredrikson M, Celik ZB, Swami A. The limitations of deep learning in adversarial settings. In: 2016 IEEE european symposium on security and privacy (EuroS&P). IEEE; 2016, p. 372–87.

[42] Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: 2017 IEEE symposium on security and privacy (SP). IEEE; 2017, p. 39–57.

[43] Moosavi-Dezfooli S-M, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, p. 2574–82.

[44] Zhang Y, Zhang W, Chen K, Liu J, Liu Y, Yu N. Adversarial examples against deep neural network based steganalysis. In: Proceedings of the 6th ACM workshop on information hiding and multimedia security. ACM; 2018, p. 67–72.

[45] Li S, Ye D, Jiang S, Liu C, Niu X, Luo X. Anti-steganalysis for image on convolutional neural networks. Multimedia Tools Appl 2018;1–17.

[46] Ma S, Guan Q, Zhao X, Liu Y. Adaptive spatial steganography based on probability-controlled adversarial examples. 2018, arXiv preprint arXiv:1804.02691.

[47] Tang W, Li B, Tan S, Barni M, Huang J. CNN-Based adversarial embedding for image steganography. IEEE Trans Inf Forensics Secur 2019.

[48] Bernard S, Pevnỳ T, Bas P, Klein J. Exploiting adversarial embeddings for better steganography. In: Proceedings of the ACM workshop on information hiding and multimedia security. ACM; 2019, p. 216–21.

[49] Bernard S, Bas P, Klein J, Pevny T. Explicit optimization of min max steganographic game. IEEE Trans Inf Forensics Secur 2020;16:812–23.

[50] Mo H, Song T, Chen B, Luo W, Huang J. Enhancing JPEG steganography using iterative adversarial examples. In: 2019 IEEE international workshop on information forensics and security (WIFS). IEEE; 2019, p. 1–6.

[51] Tramèr F, Kurakin A, Papernot N, Goodfellow I, Boneh D, McDaniel P. Ensemble adversarial training: Attacks and defenses. 2017, arXiv preprint arXiv:1705.07204.

[52] Hinton G, Vinyals O, Dean J. Distilling the knowledge in a neural network. 2015, arXiv preprint arXiv:1503.02531.

[53] Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. In: Bengio Y, LeCun Y, editors. 3rd international conference on learning representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, conference track proceedings. 2015, URL http://arxiv.org/abs/1412.6572.

[54] Filler T, Fridrich J. Gibbs construction in steganography. IEEE Trans Inf Forensics Secur 2010;5(4):705–20.

[55] Bas P, Filler T, Pevnỳ T. " Break our steganographic system": the ins and outs of organizing BOSS. In: International workshop on information hiding. Springer; 2011, p. 59–70.

[56] Bas P, Teddy F. Breaking our watermarking system. (BOWS-2), 2nd ed.. URL http://bows2.ec-lille.fr.

[57] Carlini N, Wagner D. Adversarial examples are not easily detected: Bypassing ten detection methods. In: Proceedings of the 10th ACM workshop on artificial intelligence and security, 2017, p. 3–14.

[58] Wu S, Zhong S-h, Liu Y. Residual convolution network based steganalysis with adaptive content suppression. In: 2017 IEEE international conference on multimedia and expo (ICME). IEEE; 2017, p. 241–6.