J. Vis. Commun. Image R. 80 (2021) 103325

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Full length article

Adversarial steganography based on sparse cover enhancement☆,☆☆

Chuan Qin, Weiming Zhang *, Xiaoyi Dong, Hongyue Zha, Nenghai Yu

School of Information Science and Technology, University of Science and Technology of China, Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences, China

ARTICLE INFO

Keywords:
Steganography
Adversarial example
Deep neural network

ABSTRACT

CNN (Convolutional Neural Network) steganalyzers achieve enormous improvements in detecting stego images. However, they are easily deceived by adversarial steganography, which combines adversarial attack and steganography. Currently, there are two kinds of adversarial steganography, function separation and cover enhancement. ADV-EMB (**ADV**ersarial **EMB**edding) is a typical function separation method. It forces the steganographic modifications along side the gradient directions of the target CNN steganalyzer on partial image elements. It results in relatively low deceiving success rate against the target model. ADS (**AD**versarial **S**teganography) is the first adversarial steganography, which is based on cover enhancement. It introduces much distortions, so it can be easily detected by non-target steganalyzers. To overcome such defects of the previous works, in this paper, we propose a novel cover enhancement method, denoted as SPS-ENH (**SP**ar**S**e **ENH**ancement). Through sparse ±1 adversarial perturbations, we effectively compress the distortions caused in cover enhancement. In addition, a re-trying scheme is introduced to further reduce the distortion scale. Extensive experiments show that the proposed method outperforms the previous works in the average classification error rates under non-target steganalyzers and deceiving success rates against target CNN models. When combining with the min–max strategy, the proposed method converges in less iterations and provides higher security level than ADV-EMB.

1. Introduction

Steganography is a technique of covert communication [1–4]. It aims to embed the secret messages into the cover while arousing minimal suspicions of the detector. Content-adaptive steganography [5] is considered the most secured steganographic approach. It formulates the steganography problem as source coding with fidelity constraint. Two tasks under the optimization problem of it: (1) defining modification cost of each cover element, (2) designing steganographic codes that embed the secret messages into the cover while minimizing the costs defined before. Several approaches have realized near-optimal coding performance, i.e. syndrome-trellis codes (STC) [6] and steganographic polar codes (SPC) [7].

Defining the cost function has become a research hotspot recently. In spatial domain, HUGO [8] defines modification costs by referring to SPAM feature changes between cover and stego images. WOW (Wavelet Obtained Weights) [9] and UNIWARD (UNIversal Wavelet Relative Distortion) [10] defines costs based on the residual map obtained by a bank of directional filters. Li et al. proposed *the spreading rule* in HILL [11]. Aside from aforementioned heuristic-defined cost functions, several model-driven cost functions have been proposed, such as MG [12], MVG [13], and MiPOD [14]. In JPEG domain, Holub et al. extended S-UNIWARD to JPEG domain and side-informed domain. Considering the intra- and inter-block dependency, Guo et al. proposed UED (Uniform Embedding Distortion metric) [15] and improved it to UERD (Uniform Embedding Revised Distortion metric) [16]. Recently, inspired by the success of GANs (Generative Adversarial Networks) in computer vision area, several GANs-based cost functions were developed, e.g. ASDL-GAN [17], UT-6HPF-GAN [18] and SPAR-RL [19]. The security of steganography increases significantly with more subtle cost functions being proposed.

Along side with the development of steganography, extensive works have been proposed in steganalysis [20]. The researchers mostly focused on designing handcrafted features that capture the steganographic distortions in high-dimensional spaces. The diversity

☆ This work was supported in part by the Natural Science Foundation of China under Grant U1636201 and 61572452, by Anhui Initiative in Quantum Information Technologies under Grant AHY150400.

☆☆ This paper has been recommended for acceptance by Zicheng Liu.

* Corresponding author.
E-mail addresses: qc94@mail.ustc.edu.cn (C. Qin), zhangwm@ustc.edu.cn (W. Zhang), dlight@mail.ustc.edu.cn (X. Dong), zhahongyue@163.com (H. Zha), ynh@ustc.edu.cn (N. Yu).

https://doi.org/10.1016/j.jvcir.2021.103325
Received 30 June 2021; Received in revised form 14 September 2021; Accepted 25 September 2021
Available online 2 October 2021

of steganographic methods and source images drives the steganalyzer to assemble the sub-models generated from various high pass filters. SRM [21], TLBP [22], GFR [23] and DCTR [24] are all constructed based on such logic. To cooperate with high-dimensional features, sophisticated machine learning tools are adopted, such as FLD (Fisher Linear Discriminant) ensemble classifier [25].

Inspired by enormous progress of CNNs (Convolutional Neural Networks) in computer vision area [26–30], various studies were made on CNN-based steganalyzer. Qian et al. [31] and Xu et al. did some early works in designing CNN-based steganalyzers. YeNet [32] is the first one that outperforms handcrafted feature-based steganalyzers. More recently, Boroumand et al. [33] proposed SRNet for both the spatial and JPEG domain. Some other advanced CNN structures were also adopted and effectively improve the detection accuracy of CNN steganalyzers, such as co-variance pooling [34,35], depth-wise separable convolution [36,37], Siamese structure [38,39]. CNN-based steganalyzers now significantly outperform handcrafted feature-based ones.

Despite the success of CNNs in many areas, they are found to be vulnerable against imperceptible and subtly crafted adversarial perturbations [40–43]. Such perturbations could be added on any samples and deceive CNNs to output incorrect results. Vast number of researches are made on the generation of adversarial perturbations. The discovery of adversarial perturbations motivates the study of steganography. But, directly adding adversarial perturbations on stego images would destroy the message extraction. Zhang et al.'s work called ADS (**AD**versarial **S**teganography) [44] is the first one that successfully combines adversarial perturbations and steganography. They utilize FGSM (Fast Gradient Sign Method) to enhance cover images until they are still classified as the cover by the target CNN steganalyzer even after the secret messages embedded. Tang et al. proposed ADV-EMB (**ADV**ersarial **EMB**edding) [45], which accomplishes the adversarial attack by synchronizing the modification directions with the gradient signs. To reduce the excessive modifications caused by imbalanced plus and minus cost maps, ADV-EMB controls the amount of the cost-adjusted elements through a iteration process. Similar to the iteration process of ADS, ADV-EMB also requires embedding the secret messages to evaluate whether the adversarial attack is accomplished. We call these two as the basic adversarial steganography, since there are several works inspired by them or directly adopted them as tools to generate more secured stego images under the detection of non-target steganalyzers. One typical example is min–max strategy [46]. It was built upon a game between the steganographer and the steganalyst. The most secured stego images are selected as the ones that obtains highest cover class probabilistic outputs of the steganalyzers in the previous iterations. Meanwhile, the optimal steganalyzer of the current iteration is trained with the cover images and the corresponding most secured stego images. The candidate set of the most secured stego images are updated by generating adversarial stego images targeting the steganalyzer in the last iteration. Hence, an adversarial steganographic method that can deceive the CNN steganalyzer and improve the security from the basic method is the basics of min–max strategy.

However, both ADS and ADV-EMB have defects in different perspectives. During the cover enhancement, ADS introduces a large amount of distortions. It significantly reduces its security under non-targeted steganalyzers. ADV-EMB effectively controls the excessive distortions by forcing part of modifications along the gradient signs. While the cover image is randomly divided into two groups. This operation introduces two defects: (1) Relatively low deceiving success rates against target CNN steganalyzers will be observed. The amount of the steganographic modifications that can be adjusted for adversarial attack is limited. Sometimes, they are not enough to "erase" the steganographic trace. One can observe that the deceiving success rate of ADV-EMB is lower than ADS and SPS-ENH, and it drops when the relative payload decreases. (2) Neglecting the absolute values of gradients would introduce excessive modifications. The modification on the element with larger absolute gradient values would bring about more significant changes on the predictions. Yet this is not considered by ADV-EMB.

Summarizing the defects of ADS and ADV-EMB above, one can conclude that current adversarial steganographic methods cannot obtain a balance between high deceiving success rate against a target CNN steganalyzer and the **holistic security** that is measured by non-target steganalyzers when the steganalyzers are trained with the steganographic method. In this paper, we propose a sparse cover enhancement method, which obtains high deceiving success rate and high holistic security. To achieve satisfactory deceiving success rate, we take the cover enhancement framework to combine the adversarial perturbations with steganographic embedding. To reduce the large distortions introduced when enhancing the cover images, we propose a novel sparse adversarial perturbation method. By restricting the perturbation amplitude, it can easily minimize the L_1 distances. The distortions it introduces are significantly smaller than the previous methods against steganalysis models. Since the steganographic modifications are random, we repeat the steganographic embedding with new random seeds several times when the prediction probabilities are large enough. Such mechanism effectively reduce the amount of adversarial perturbations.

The contributions of this paper are summarized as follows:

1. We propose a novel adversarial steganographic method, which obtains comparable holistic security as ADV-EMB while higher deceiving success rates against target CNN steganalyzers.
2. A novel sparse adversarial perturbations is introduced to deceive CNN steganalyzers. It is specifically designed for steganography task and introduces much smaller distortions than the previous works.
3. In the similar cover enhancement framework, we introduce a restarting and repeating mechanism. It effectively reduces the adversarial perturbations.

The rest of this paper is organized as follows. In Section 2, several typical adversarial perturbation methods and adversarial steganographic methods are briefly reviewed. We detail the proposed method in Section 3. In Section 3, the proposed method is detailed. Extensive experiments are carried out in Section 4. The paper is concluded in Section 5.

2. Related work

2.1. Adversarial perturbations

Let $\boldsymbol{x}$ be a input image, with y_{true} being the ground-truth label. In steganalysis, $y_{\text{true}} \in \{0, 1\}$, where 0 and 1 represent cover and stego class respectively. $P(x, \theta)$ represents a CNN model with hyper-parameters θ. The loss of the CNN model with y set as the ground-truth label is denoted as $L(P(\boldsymbol{x}, \theta), y)$.

2.1.1. FGSM

FGSM (**F**ast **G**radient **S**ign **M**ethod) is proposed by Goodfellow et al. [40], which generate adversarial perturbations by multiplying gradient maps with a scalar. With θ being fixed, the gradient map of $L(P(\boldsymbol{x}, \theta), y_{\text{true}})$ with reference to the input image $\boldsymbol{x}$ is denoted as $\boldsymbol{\eta}$:

$$\boldsymbol{\eta} = \frac{\partial L(P(\boldsymbol{x}, \theta), y_{\text{true}})}{\partial \boldsymbol{x}}. \tag{1}$$

By multiplying the gradient map $\boldsymbol{\eta}$ with a proper scalar ϵ, the adversarial perturbations $\boldsymbol{\delta} = \epsilon \cdot \boldsymbol{\eta}$ would alter the predicted label of the target model.

2.1.2. DDN (Decoupled Direction and Norm)

Rony et al. [43] proposed to project the perturbations onto a L_2-sphere of the clean image. Then the norm will be changed based on whether the adversarial attack is successful or not. DDN introduces significantly smaller L_2 distortions and requires fewer iterations than the previous works.

2.2. Adversarial steganography

2.2.1. ADS (*ADversarial Steganography*)

ADS [44] is the first adversarial steganography. It utilizes FGSM to iteratively enhance the original cover image until the predicted label would remain cover after the secret messages embedded. In each iteration the modification is generated by:

$$\boldsymbol{\delta}_i = \epsilon \cdot \frac{\partial L(P(\boldsymbol{z}_{i-1}, \theta), 1)}{\partial \boldsymbol{z}_{i-1}}, \tag{2}$$

where $\boldsymbol{z}_{i-1}$ is the stego image generated on the enhanced cover image of the last iteration. Assuming the final iteration is n, the enhanced cover image is:

$$\boldsymbol{c}' = \boldsymbol{c} + \sum_{i=1}^{n} \boldsymbol{\delta}_i. \tag{3}$$

2.2.2. ADV-EMB (*ADVersarial EMBedding*)

ADV-EMB [45] generates adversarial stego images by forcing the embedding cost fit the gradient sign. It divides the elements of the cover image into two disjoint groups, common group and adjustable group. The embedding costs in common group are defined by the base cost function, such as UNIWARD [10], HILL [11], UERD [16], and etc. The embedding costs in adjustable group are adjusted based on the gradient map of the stego image generated in the last iteration:

$$q_{i,j}^{+} = \begin{cases} \rho_{i,j}^{+}/\alpha, & \text{if } \eta_{i,j} < 0, \\ \rho_{i,j}^{+}, & \text{if } \eta_{i,j} = 0, \\ \rho_{i,j}^{+} \cdot \alpha, & \text{if } \eta_{i,j} > 0, \end{cases} \tag{4}$$

$$q_{i,j}^{-} = \begin{cases} \rho_{i,j}^{-}/\alpha, & \text{if } \eta_{i,j} > 0, \\ \rho_{i,j}^{-}, & \text{if } \eta_{i,j} = 0, \\ \rho_{i,j}^{-} \cdot \alpha, & \text{if } \eta_{i,j} < 0, \end{cases} \tag{5}$$

where the gradient value, base cost value and adjusted cost value at the element with position index i, j are denoted as $\eta_{i,j}$, $\rho_{i,j}$ and $q_{i,j}$.

3. The sparse adversarial enhancement scheme

3.1. The gradient distribution of CNN steganalyzers

Given an input image $\boldsymbol{x}$, the gradient map of the CNN steganalyzer is denoted as $\boldsymbol{\eta} = \frac{\partial L(P(\boldsymbol{x},\theta), y_{\text{true}})}{\partial \boldsymbol{x}}$. It indicates how to modify the input image could change the prediction. Both ADV-EMB [45] and ADS [44] utilize the gradient map to deceive CNN steganalyzers. We plot the gradient maps and their histograms of several typical stego images generated by S-UNIWARD [10] from BOSSBase 1.01 [47] in Fig. 1.

It can be observed from the histogram that the standard deviations of the gradient distributions are high. Most elements are with low gradient values. The modifications on them could not effectively change the predictions. Meanwhile, it can be observed from the second row, which is the 3-D gradient value maps, the gradient values peak at some elements. Modifying them could effectively help deceive the target CNN steganalyzer. Interestingly, these pixel points with high gradient values are sometimes located in smooth area, making them difficult to be exploited by cost adjustment methods, such as ADV-EMB. This phenomenon may be due to the randomness of steganographic modifications, i.e. though steganographic modifications cluster in textured area where the modification costs are set low, the elements with high modification costs could still be changed.

Using cost adjustment method as ADV-EMB cannot precisely control where modifications are made. Thus, we adopt cover enhancement as the general framework of the proposed method.

Table 1
The average proportion and steganographic modifications of common group and adjustable group of ADV-EMB.

Payload (bpp)	Group	Proportion	Modification
0.1	Common	79.90%	773.47
	Adjustable	20.10%	219.35
0.2	Common	73.76%	1599.87
	Adjustable	26.23%	655.87
0.3	Common	73.43%	2572.71
	Adjustable	26.56%	1094.53
0.4	Common	73.61%	3651.31
	Adjustable	26.38%	1564.69
0.5	Common	75.74%	4949.23
	Adjustable	24.25%	1925.11

3.2. Formulating a L_0 and L_∞ constraint optimization problem

In cover enhancement, the steganographer enhances cover images before embedding secret messages. As analyzed before, the defects of ADS, i.e., low holistic security, are caused by large distortions introduced during enhancing cover images. In this section, we compress such distortions by formulating adversarial attack as a L_0 and L_∞ constraint optimization problem.

The adversarial attack is often regarded as a L_p-norm constraint optimization problem.

$$\min_{\delta} \|\boldsymbol{\delta}\|_p \quad \text{s.t. } \arg\max P(\boldsymbol{x}+\boldsymbol{\delta}, \theta) \neq y_{\text{true}} \quad \text{and } 0 \leq \boldsymbol{x}+\boldsymbol{\delta} \leq M, \tag{6}$$

where M denotes the upper bound of image element value and $\boldsymbol{\delta}$ denotes the modification matrix. L_0-norm leads to sparse perturbations, which meets the requirement of improving holistic security. Solving L_0-constraint optimization problem is NP-hard to solve. Fortunately, CNN steganalyzers are easily deceived by ± 1 adversarial perturbations. Thus, we add L_∞ constraint and resort to a greedy algorithm to find a local optimal solution. Before introducing the detailed algorithm, we analyze the vulnerability of CNN steganalyzers against ± 1 adversarial perturbations.

The spatial steganographic modifications are rather small and exist in highly textured areas, which can be considered as high frequency signals. In CNN steganalyzers, stacking unpooled layers, pre-processing the input images with a HPF (High Pass Filter) and etc. are taken to preserve the high frequency signals. Intuitively, the perturbations of small amplitude would effectively influence the prediction of CNN steganalyzers.

Specifically, a phenomenon from ADV-EMB inspired us. The modifications in the adjustable group are alongside with the gradient directions. One can consider that the modifications in the adjustable group "erase" the steganographic trace in the common group. In Table 1, we exhibit the proportions and the average modification numbers in each group under relative payloads from 0.1 bpp (bit per pixel) to 0.5 bpp. One can see that the ± 1 adversarial perturbations with the half of the common steganographic modification number can deceive the CNN steganalyzer.

Hence, we restrict the modification amplitude to 1 and formulate the optimization problem we are solving as follows:

$$\min_{\delta} \|\boldsymbol{\delta}\|_0 \quad \text{s.t. } \arg\max P(\boldsymbol{x}+\boldsymbol{\delta}, \theta) \neq y_{\text{true}} \quad \text{and } \|\boldsymbol{\delta}\|_\infty = 1 \text{ and } 0 \leq \boldsymbol{x}+\boldsymbol{\delta} \leq M, \tag{7}$$

Now solving such problem becomes much easier. We aim to minimize the number of ± 1 modifications. The detailed algorithm is discussed in the next section.

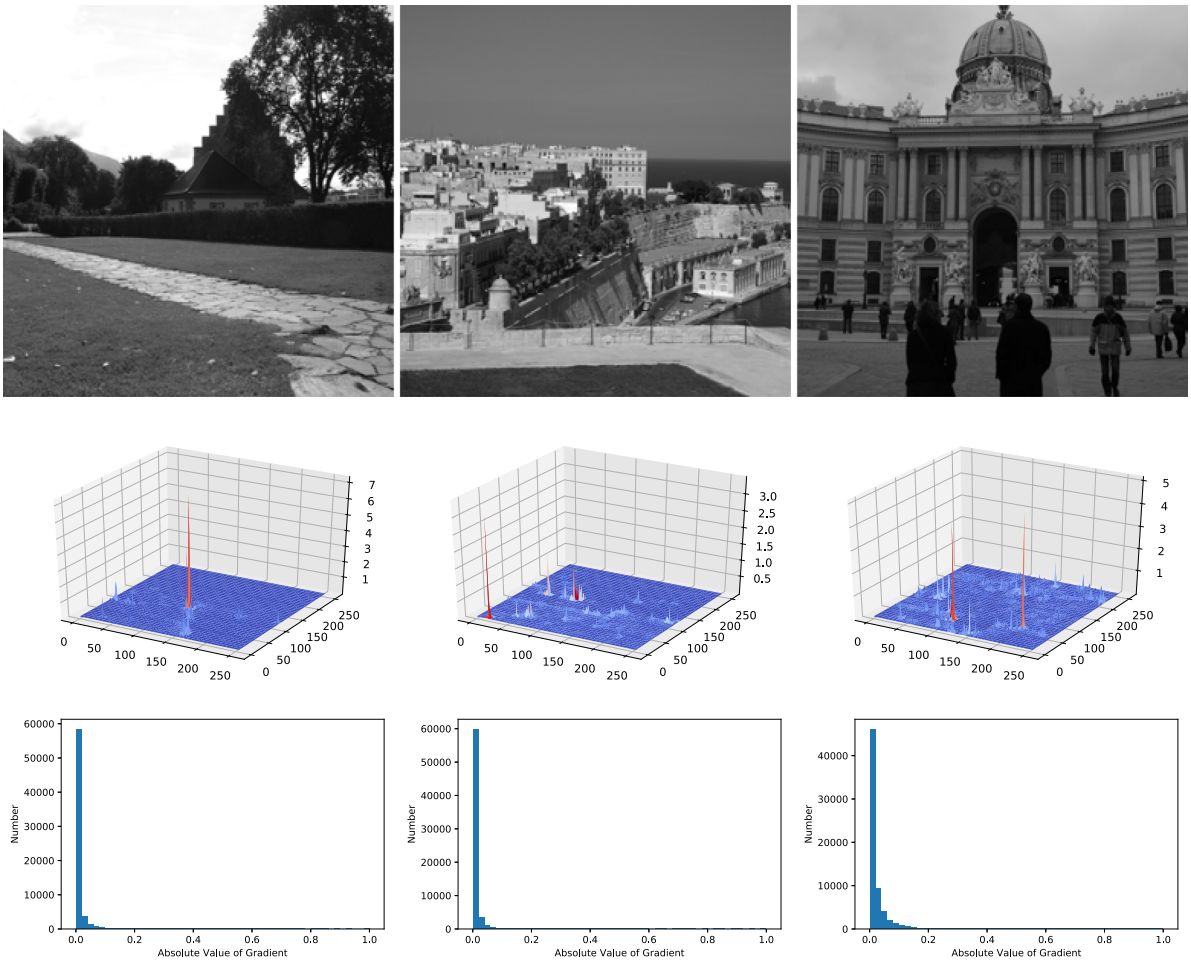

Fig. 1. The first row exhibits the stego images. The second row exhibits the corresponding gradient maps. The third row exhibits the histograms of the gradients.

3.3. Generating adversarial stego images

We propose to iteratively enhance the cover image to generate adversarial stego images. The general process is shown in Fig. 2. We simultaneously modify the original stego image with the same adversarial perturbations and calculate the gradient map with reference to it in each iteration:

$$\boldsymbol{\eta}_i = \frac{\partial L(P(\boldsymbol{s}_i^t, \theta), 1)}{\partial \boldsymbol{s}_i^t}, \tag{8}$$

where s_i^t represents the modified stego image in the current iteration. Specifically, in the first iteration, s_i^t is the original stego image. The intuition behind such operation is the perturbations that lower the loss $L(P(\boldsymbol{s}_i^t, \theta), 0)$, i.e., making the original stego image more and more "cover", could "erase" the steganographic modification traces on a slightly modified cover image generated by the same method.

In each iteration, we enhance the image elements that are with top-k gradient values and not modified before. Specifically, we leverage a mask $\boldsymbol{m}$ to control whether elements can be enhanced. If the element $x_{p,q}$ was modified before, the corresponding flag $m_{p,q}$ will be set 1.

$$\boldsymbol{\eta}_i' = \boldsymbol{\eta}_i \cdot (1 - \boldsymbol{m}). \tag{9}$$

$$d_1, d_2, \ldots, d_k = \text{argtop}_k(|\boldsymbol{\eta}_i'|). \tag{10}$$

$$\boldsymbol{m}_{d_1, d_2, \ldots, d_k} = \mathbf{1}. \tag{11}$$

Formally, we enhance the cover image in each iteration as follows:

$$\boldsymbol{e}_{i+1} = \text{clip}_0^M(\boldsymbol{e}_i + \frac{\boldsymbol{\eta}_i'}{|\boldsymbol{\eta}_i'|}), \tag{12}$$

where the enhanced cover image in ith iteration is denoted as $\boldsymbol{e}_i$. The steganographic modifications will be totally different if the LSB or 2LSB of image elements are changed. So, the modified stego image $\boldsymbol{s}_i'$ being predicted cover is not the terminal condition of the iteration. It will continue until the stego image that generated from the enhanced cover image is predicted cover.

To further reduce the enhancement modifications, we introduce a repeat embedding scheme. When the probabilistic output $p(\boldsymbol{s}_i|0, \theta)$ is larger than a threshold τ, we scramble the secret message and re-generate steganographic modifications until the adversarial stego image $\boldsymbol{s}_i^j$ is predicted cover or the repetition reaches the maximum. The complete generation process is shown in Algorithm 1.

4. Experiments

4.1. Setups

Datasets. The datasets in this paper, we select BOSSBase 1.01 [47] and BOWS2 [48]. Each contains 10 000 grayscale image of 512 × 512. To match the settings of most CNN steganalyzers, we resize the original images using **imresize**() of MatLab to resize the images to 256 × 256.

Steganalyzers. The state-of-the-art CNN steganalyzers SRNet [33] is adopted as the target model in this paper. It is trained with 14 000 randomly selected cover images and the corresponding stego images

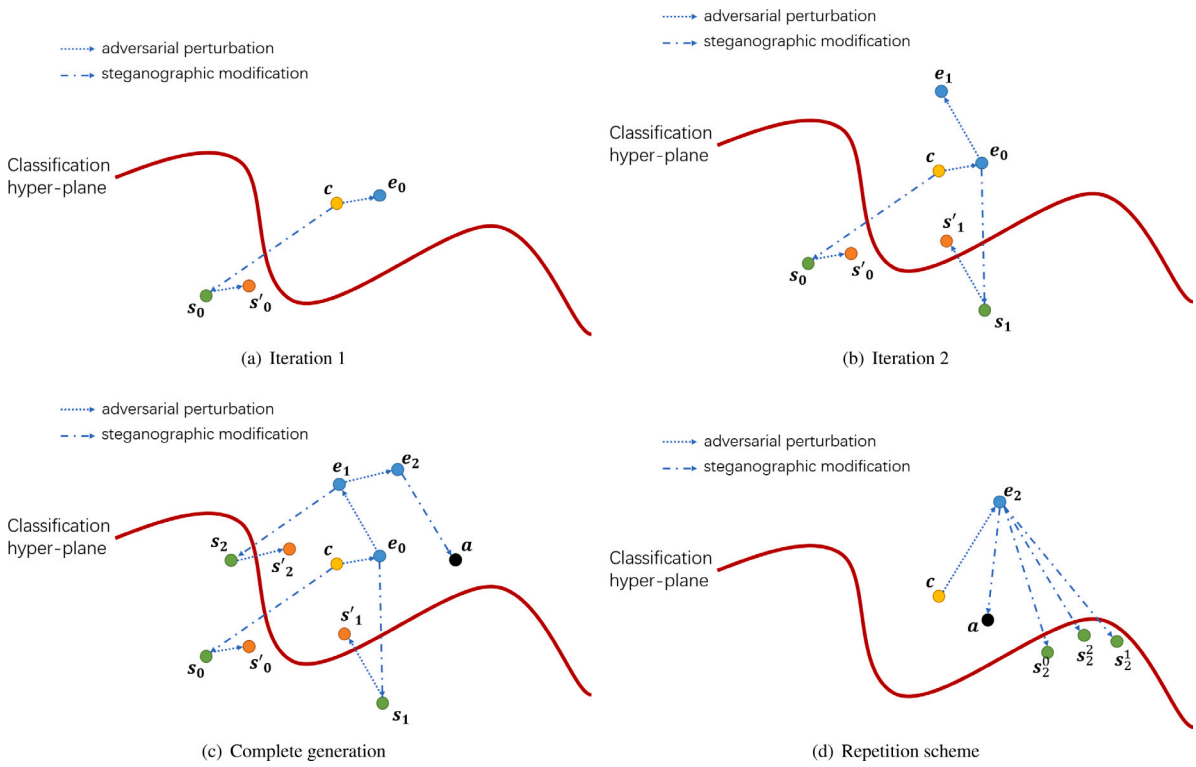

Fig. 2. The process of SPS-ENH. c, s_i, s'_i, e_i and a represent the cover image, stego image, the modified stego image, the enhance cover image and the adversarial stego image respectively. In each iteration, the gradient value is calculated with reference to the stego image towards the cover class. We enhance the top-k high gradient value elements. Since the steganographic modifications would be totally different if the LSB or 2LSB are changed, whether s'_i is predicted as cover does not matter. (c) represents the repetition scheme. Sometimes the stego image generated from the enhanced cover image is close to the classification hyper-plane (the probabilistic output is close enough to 0.5). We scramble the secret messages to produce a legal adversarial stego image.

Algorithm 1 SPS-ENH (**SP**ar**S**e **ENH**ancement)

Input: The input image $\boldsymbol{x}$, target model $P(\boldsymbol{x}, \theta)$, ground-truth label y_{true}.
Parameter: Max iteration t, modification number k, modification map $\boldsymbol{m}$, secret messages $\boldsymbol{msg}$.
Output: Adversarial example $\boldsymbol{x}^{adv}$.

Initialization: $\boldsymbol{e}_0 \leftarrow \boldsymbol{c}$, $\boldsymbol{m} \leftarrow \boldsymbol{0}$, $i \leftarrow 0$, $k \leftarrow 1$ $\rho \leftarrow \text{CostDef}(\boldsymbol{c})$, $\boldsymbol{s}_0 \leftarrow \text{StegEmb}(\boldsymbol{e}_0, \rho, \boldsymbol{msg})$, $\boldsymbol{s}_0^t \leftarrow$;
while $i < t$ and $p(\boldsymbol{s}_i|0,\theta) < 0.5$ **do**
  $\boldsymbol{\eta}_i \leftarrow \frac{\partial L(P(\boldsymbol{s}_i^t,\theta), y_{\text{true}})}{\partial \boldsymbol{s}_i^t}$
  $\boldsymbol{\eta}'_i \leftarrow \boldsymbol{\eta}_i \cdot (1 - \boldsymbol{m})$
  $d_1, d_2, ..., d_k \leftarrow \text{argtop}_k(|\boldsymbol{\eta}'_i|)$
  $\boldsymbol{m}_{d_1,d_2,...,d_k} \leftarrow \boldsymbol{1}$
  $\boldsymbol{e}_{i+1} \leftarrow \text{clip}_0^M(\boldsymbol{e}_i + \frac{\boldsymbol{\eta}'_i}{|\boldsymbol{\eta}'_i|})$
  $\boldsymbol{s}_{i+1}^t \leftarrow \text{clip}_0^M(\boldsymbol{s}_i^t + \frac{\boldsymbol{\eta}'_i}{|\boldsymbol{\eta}'_i|})$
  $\boldsymbol{s}_{i+1} \leftarrow \text{StegEmb}(\boldsymbol{e}_{i+1}, \rho, \boldsymbol{msg})$
  if $\tau < p(\boldsymbol{s}_i|0,\theta) < 0.5$ **then**
    $\boldsymbol{s}_{i+1} \leftarrow \text{StegEmb}(\boldsymbol{e}_{i+1}, \rho, \text{Scramble}(\boldsymbol{msg}))$
  $i \leftarrow i + 1$
  $k \leftarrow k + 1$
if $p(\boldsymbol{s}_i|0,\theta) > 0.5$ **then**
  $\boldsymbol{x}^{adv} \leftarrow \boldsymbol{s}_i$
else
  $\boldsymbol{x}^{adv} \leftarrow \boldsymbol{s}_0$

from BOSSBase 1.01 and BOWS2. To fully evaluate the holistic security of stego images, we include SRM+EC (the ensemble classifier trained with the Spatial Rich Model) and YeNet [32] as reference models. The steganalyzers are trained with cover images and their corresponding stego images. We set SRM+EC by default to minimize the total classification error rate while a lower false alarm rate P_{FA} is prior $P_{\text{E}} = \min_{P_{\text{FA}}} \frac{1}{2}(P_{\text{FA}} + P_{\text{MD}})$, where P_{FA} and P_{MD} are false alarm rate and missed detection rate respectively.

Base Cost Functions. To make fair comparisons with conventional steganography and the previous works (ADS and ADV-EMB), we keep consistent the base cost function of adversarial steganographic methods and the target CNN steganalyzers are trained with. Specifically, in this paper, we utilize a widely-used cost functions S-UNIWARD [10].

4.2. Holistic security

Steganography endeavours to increase its holistic security. Specifically, it is measured when the steganalyst is fully aware of the steganographic method and deploy a specific steganalyzer. In this section, the average classification error rate P_E of cover images and S-UNIWARD (baseline), ADS, ADV-EMB, SPS-ENH are compared under the detection of CNN models and handcrafted-feature models.

From Table 2, one can observe that SPS-ENH obtains higher P_E than ADS and ADV-EMB, especially under the detection of handcrafted-features. ADS sacrifices its holistic to deceive CNN steganalyzers. ADV-EMB and SPS-ENH could increase the holistic, while SPS-ENH obtains superior performances. The largest improvement is made under relative payload 0.2 bpp. The P_E of SRM+EC increases with 0.0099.

For NonADV-ENH, though it does not aim to deceive CNN models, by enhancing the cover images with the perturbations that can "erase" the steganographic modifications on the original cover, it improves the holistic of base steganographic method. Under the detection of SRM+EC, the most significant improvement 0.0255 is made under relative payload 0.3 bpp. Under the detection of SRNet and YeNet, the most significant improvements 0.0770 and 0.0683 are made under payload 0.2 and 0.3 respectively. The performance of SPS-ENH is better than

Table 2
The performance (P_E) comparison between the proposed scheme and previous works. The target model is set as **SRNet**.

Payload (bpp)	Algorithm	SRM+EC	SRNet	YeNet
0.1	Baseline	0.4364 (±0.0038)	0.3247	0.3637
	ADS	0.3743 (±0.0023)	0.2437	0.2846
	ADV-EMB	0.4433 (±0.0027)	**0.3960**	0.4082
	SPS-ENH	**0.4519** (±0.0037)	0.3909	**0.4142**
0.2	Baseline	0.3579 (±0.0030)	0.2013	0.2648
	ADS	0.3094 (±0.0036)	0.1538	0.1934
	ADV-EMB	0.3711 (±0.0038)	**0.2830**	0.2920
	SPS-ENH	**0.3810** (±0.0029)	0.2783	**0.3303**
0.3	Baseline	0.2894 (±0.0036)	0.1325	0.1623
	ADS	0.2438 (±0.0035)	0.1204	0.1547
	ADV-EMB	0.3143 (±0.0025)	0.1997	**0.2485**
	SPS-ENH	**0.3149** (±0.0041)	**0.2016**	0.2306
0.4	Baseline	0.2315 (±0.0027)	0.0866	0.1135
	ADS	0.1954 (±0.0048)	0.0826	0.0993
	ADV-EMB	0.2462 (±0.0040)	0.1297	0.1629
	SPS-ENH	**0.2541** (±0.0029)	**0.1406**	**0.1731**
0.5	Baseline	0.1834 (±0.0021)	0.0604	0.0746
	ADS	0.1439 (±0.0035)	0.0710	0.0847
	ADV-EMB	0.1992 (±0.0024)	**0.1024**	**0.1410**
	SPS-ENH	**0.2021** (±0.0032)	0.0970	0.1133

ADV-EMB under the detection of handcrafted feature-based SRM+EC, and comparable with it under the detection of CNN steganalyzers. But, it is found SPS-ENH can achieve higher security when combining with min–max [46] strategy than ADV-EMB. Such experiment is shown in Section 4.4.

4.3. Deceiving target CNN steganalyzers

In this section, we evaluate the performance of SPS-ENH in deceiving target CNN steganalyzers when the steganographer grasps the complete settings and hyper-paramters of the target model. The same as Tang et al. [45], we set steganalyzers with two types: (1) adversary-unaware, the steganalyzers are trained with conventional stego images. (2) adversary-aware, the steganalyzers enhance their robustness with adversarial training, where the adversary-aware steganalyzers utilize the weights of adversary-unaware steganalyzers for initialization and replace the stego images in the training set with the adversarial stego ones.

4.3.1. Adversary-unaware steganalyzers

In Table 3, It can be observed that the success rate of ADV-EMB is lower than ADS and SPS-ENH, and it drops with the relative payload. Specifically, the success rates of ADV-EMB are lower than 90% when the relative payloads are less than 0.3 bpp. It is mainly because that the pixels with high gradient values are not necessarily locate in textured areas. Such phenomenon can be observed in Fig. 5. Though ADV-EMB forces the modification directions along side with the gradient signs in the adjustable group. Critical modifications at the pixels with high gradient values could be skipped. Even with the whole cost map being modified, there are still probability that the modifications are not sufficient to alter the prediction. Such phenomenon would be more severe when the relative payload drops. Since the quantity of modifications would significantly reduced with the relative payload. It would be harder for ADV-EMB to achieve adversarial attacks by only guide the modification directions.

4.3.2. Adversary-aware steganalyzers

In Table 4, such adversarial training strategy brought little robustness against ADS and SPS-ENH. ADV-EMB receives clear drops on deceiving success rate across all the tested payloads. The largest drop of ADV-EMB is 28.55% under relative payload 0.4 bpp. Meanwhile, The largest success rate decrease of SPS-ENH is only 0.03% under relative payload 0.1 bpp. SPS-ENH exhibits superior robustness against adversarial trained CNN steganalyzers.

Table 3
The success rate of ADS, ADV-EMB and SPS-ENH in the white-box scenario against a **adversary-unaware** steganalyzer under the relative payload (bpp) from 0.1 to 0.5. The adversarial steganographic method is abbreviated as "ADV" in the table.

Target model	ADV	0.1	0.2	0.3	0.4	0.5
SRNet	ADS	100.00%	99.99%	99.98%	100.00%	99.86%
	ADV-EMB	88.12%	88.78%	89.92%	95.00%	94.78%
	SPS-ENH	100.00%	99.96%	99.68%	99.66%	99.36%

Table 4
The success rate of ADS, ADV-EMB and SPS-ENH in the white-box scenario against a **adversary-aware** steganalyzer under the relative payload (bpp) from 0.1 to 0.5.

Target model	Adversarial steganography	0.1	0.2	0.3	0.4	0.5
SRNet	ADS	98.99%	96.59%	99.13%	97.08%	97.06%
	ADV-EMB	76.39%	78.14%	70.71%	68.53%	72.36%
	SPS-ENH	99.97%	99.97%	99.99%	99.85%	99.96%

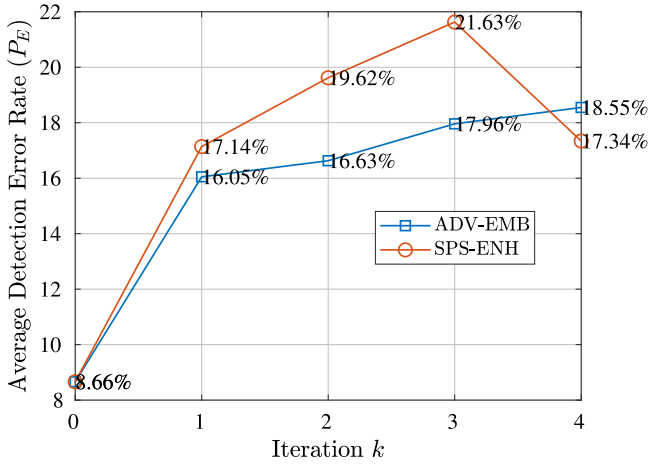

Fig. 3. The holistic (P_E) in each iteration. (Round 0 is consists of all stego images generated via base cost function).

4.4. Under min–max strategy

Bernard et al. [46] proposed the min–max strategy, in which they utilized ADV-EMB to attack the CNN steganalyzers in each iteration. As SPS-ENH is a more securer adversarial steganographic method, one could anticipate that it can achieve higher holistic security than ADV-EMB when cooperating with the min–max strategy. In Fig. 3, we compare the holistic security of SPS-ENH-min–max with ADV-EMB-min–max under the detection of SRNet. We run 4 iterations to collect statistics. It can be observed that SPS-ENH converges faster than ADV-EMB and achieves superior security performances under the detection of the optimal steganalyzer. The average detection error rate on SPS-ENH-min–max peaks at the third iteration with 21.63%, while the holistic security of ADV-EMB-min–max converges at the fourth iteration with 18.55%. SPS-ENH obtains superior performance when combining with the min–max strategy.

Moreover, min–max strategy could be utilized by the steganalyst to enhance the steganalyzers' robustness against adversarial steganography. We compare the deceiving success rate of ADV-EMB and SPS-ENH in each iteration in Fig. 4. The success rate of ADV-EMB against min–max retrained models drops significantly. It steadies at about 60% lastly, while the success rate of SPS-ENH even increases a little with the iteration. Hence, ADV-EMB can be defended via min–max retraining. But such strategy does not work against SPS-ENH.

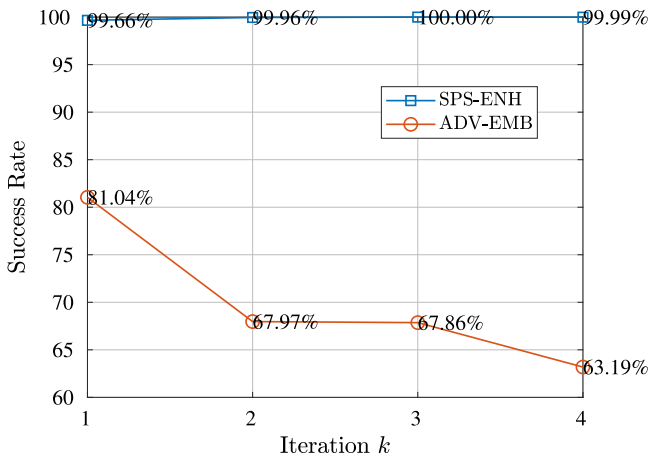

Fig. 4. The attacking success rate of SPS-ENH and ADV-EMB in each iteration (starting from round 1).

Table 5
The comparison of distortions among FGSM, DDN and HiD-PeT when targeting CNN-based steganalyzer SRNet.

	FGSM	DDN	HiD-PeT
L_0	65 241.80	24 610.98	**52.71**
L_∞	4.0	1.0	**1.0**
L_2	1017.62	16.20	**6.20**

4.5. Ablation study

4.5.1. Distortions caused by the adversarial enhancement

The sparse enhancement is one of the key elements in guaranteeing the holistic security of SPS-ENH. The adversarial perturbation of SPS-ENH is denoted as HiD-PeT (**HiD**den **P**er**T**urbation). In this section, we compare the distortions caused by the proposed method, HiD-PeT, with adversarial attacks in computer vision area, e.g. FGSM (L_∞-based, utilized in ADS), DDN (L_2-based, introducing the least L_2 distortions). To make full-scale comparison, we compare L_0, L_2 and L_∞ distortions of them. SRNet is trained using S-UNIWARD under the payload of 0.4 bpp. Note that the goal of the adversarial enhancement and other adversarial attacks is set to alter the predictions of the target steganalyzer on stego images regardless whether the secret message can be extracted.

L_0 distortions could represent the sparsity of perturbations. Through the greedy search process, HiD-PeT could alter the prediction of target CNN-based steganalyzer with average 52.71 pixels being modified. Since we restrict the modifications to only ±1, HiD-PeT obtains the least L_∞ than all the other adversarial perturbations. With the least L_0 and L_∞ distortions, L_2 distortions introduced by HiD-PeT would be less than other methods. The results are shown in Table 5.

Moreover, to exhibit the visual performance of HiD-PeT, we display the modification maps of it on three randomly selected images from BOSSBase 1.01 in Fig. 5. The original stego images, the steganographic modifications on them, FGSM perturbations, DDN perturbations and the HiD-PeT perturbations are listed in columns respectively. The modified pixels are colored in white, while the unmodified ones are in black. FGSM modifies most of the pixels, except the ones in the most smooth areas, such as the pedestrian in black in the first image and the white car at the corner in the third image. The modification map of DDN is much sparser than that of FGSM, but there are still some modifications made in smooth areas. It can be clearly observed that the number of HiD-PeT modifications is so small. Only a few pixels are modified. Furthermore, there are only a few pixels in the gradient map are highlighted, which means only a few pixels make remarkable impact on the predictions while the modification on other pixels are insignificant. Based on the statistics and the visualization above, one can conclude that the proposed HiD-PeT introduces much less distortions than the previous works for adversarial steganography.

Table 6
The average classification error rates of SRM on adversarial modified stego images.

Payload (bpp)	Average classification error rate (P_E)
0.1	0.4899 (±0.0028)
0.2	0.4811 (±0.0026)
0.3	0.4806 (±0.0036)
0.4	0.4608 (±0.0045)
0.5	0.4355 (±0.0056)

Table 7
The average modification numbers and holistic security of different modification amplitudes.

Amplitude	±1	±2	±4
Average modifications	390.00	344.82	3542.71
Holistic security P_E	0.2409 (±0.0041)	0.2142 (±0.0032)	0.1251 (±0.0036)

4.5.2. Holistic security of the adversarial enhancement

From the perspective of steganography, low L_0, L_2 and L_∞ distortions do not necessarily mean high holistic security. Since the modification positions also influence it.

To further exhibit the holistic security of the proposed cover enhancement, HiD-PeT, we try to detect it using SRM+EC. In Table 6, it is observed that the adversarial modified stego images remains undetectable for the steganalyzers. The adversarial modified stego images are predicted by the target model as stego and cover. About 50% training accuracy means the steganalyzer is not capable to differentiate the samples. Hence, the security of HiD-PeT targeting CNN steganalyzers is quite high, which guarantees the performance of SPS-ENH.

4.5.3. The amplitude of adversarial perturbations

The terminal condition of SPS-ENH is the adversarial stego images can deceive the target CNN steganalyzer. But, when conducting the following ablation study, changing the adversarial perturbation amplitudes or the calculation gradient maps in each iteration will consume significantly more time. Thus, for the convenience, we change the terminal condition of standard SPS-ENH to $p(0|\boldsymbol{s}_j^t) > \tau'$. Such modified version of SPS-ENH is denoted as NonADV-ENH (**Non-ADV**ersarial **ENH**ancement). The later discussed hyper-parameters in NonADV-ENH are the same and play the same roles as in SPS-ENH. Without loss of generality, we substitute SPS-ENH with NonADV-ENH to discuss the impact of adversarial perturbation amplitude in this section and the calculation of gradient maps in the next section.

The steganographic modifications are mostly restricted as ±1. It motivates us to enhance the cover images with ±1. We compare the average modification numbers and holistic of NonADV-ENH with different enhancement amplitudes. The target model and the base cost function we adopt SRNet and S-UNIWARD. The results in Table 7 show that the minimal amplitude ±1 is most effective and provides the optimal holistic security.

It can be noted that larger modification amplitudes significantly reduce the holistic security of SPS-ENH. Though ±2 spends less modifications to alter the predictions, its holistic security is much lower than ±1, even lower than the baseline (0.2315 (±0.0027)). Meanwhile, the modification number of ±4 is almost 9 times more than that of ±1. Not surprisingly, its holistic security is also much lower than the standard SPS-ENH and the baseline. Hence, using ±1 to enhance cover images could compress the modification number and improve the holistic security of adversarial stego images.

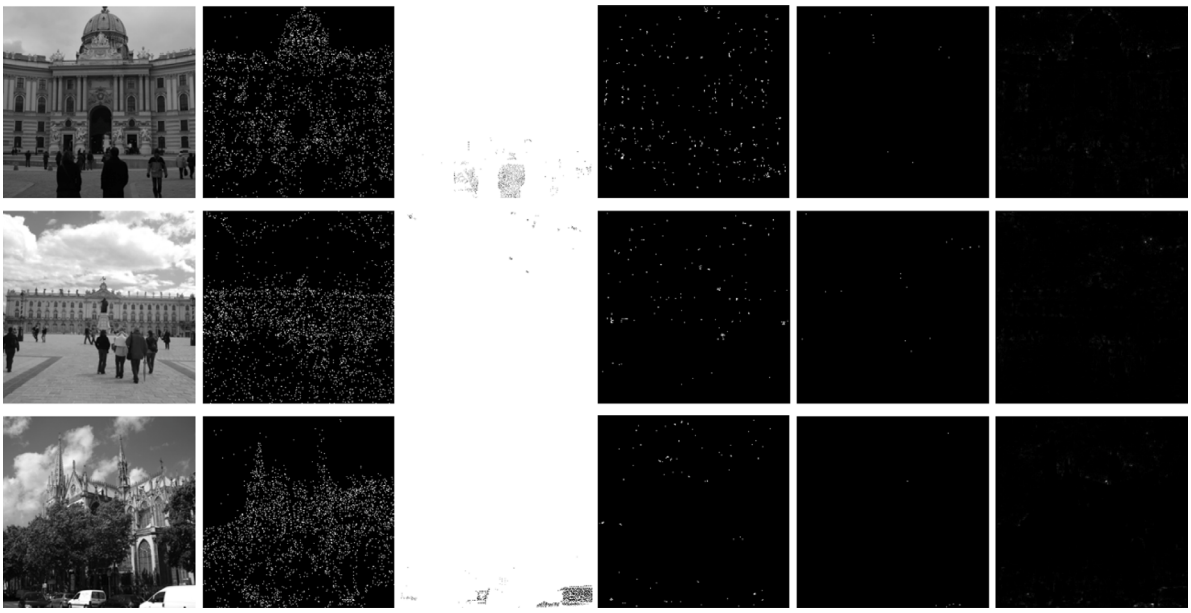

Fig. 5. The exhibition of the original stego images (the first column), steganographic modifications (the second column), FGSM perturbations (the third column), DDN perturbations (the fourth column), HiD-PeT modifications (the fifth column) and the gradient map (the sixth column). If a pixel is modified, it will be painted white. Otherwise, it will be in black.

Table 8
The average modification numbers and holistic security comparison between using the gradients of s_i^t and s_i.

SPS-ENH	Average modifications	390.00
	Holistic security (P_E)	0.2409 (±0.0041)
Ref to s_i	Average modifications	551.27
	Holistic security (P_E)	0.2283 (±0.0037)

4.5.4. The calculation of gradient maps

Different from ADS and ADV-EMB, SPS-ENH leverages the gradient maps with reference to adversarially modified stego images s_i^t. In this section, we analyze the modifications and holistic security of the proposed scheme and if the gradient maps is generated as $\eta_i \leftarrow \frac{\partial L(P(s_i,\theta),y_{\text{true}})}{\partial s_i}$. We take SRNet as the target model, S-UNIWARD as the base cost function. The detailed statistics is shown in Table 8.

It can be observed from Table 8 that utilizing the gradients of s_i^t produces much less perturbations to deceive the target model. As a result, the holistic security of the proposed scheme is higher than using the gradients of s_i. The reason behind the performance gap is the steganographic modifications on the modified stego images in each iterations are totally different. The gradients that is effective to "erase" the steganographic trace of the last iteration could be useless for the new steganographic modifications. Hence, the accumulated adversarial perturbations are often useless or even could be against the effective directions. While the adversarial perturbations that drive the original stego images further away from the classification hyper-plane, i.e., gain high probabilistic outputs, could be transferable sometime for the final steganographic modifications.

4.5.5. The terminal threshold of NonADV-ENH

The terminal threshold of NonADV-ENH is the parameter that influences the holistic security of NonADV-ENH. A higher threshold introduces more perturbations but could potentially brings about more transferability, which could improve the holistic security of NonADV-ENH. Hence, in this section, we analyze the specific influence of the terminal threshold τ' on the holistic security, which is measured by SRM+EC. Without loss of generality, we take relative payload 0.4 bpp and target model SRNet as an example to exhibit the influence of τ'. The results are shown in Table 9.

Table 9
The holistic security (P_E) and average modification numbers of NonADV-ENH with different terminal thresholds *tau'*.

Threshold τ'	SRM+EC	Average modifications
$\tau' = 0.5$	0.2376 (±0.0035)	48.49
$\tau' = 0.6$	0.2389 (±0.0041)	52.59
$\tau' = 0.7$	0.2404 (±0.0034)	57.68
$\tau' = 0.8$	0.2387 (±0.0039)	65.07
$\tau' = 0.9$	0.2409 (±0.0041)	78.58
$\tau' = 0.99$	0.2354 (±0.0035)	131.26

The average classification error rate fluctuates from 0.2354 to 0.2409. It peaks when $\tau' = 0.9$. It basically increases with τ' until $\tau' = 0.99$. Since $\tau' = 0.99$ is quite high, the modification number significantly increases, almost doubled than $\tau' = 0.9$. Thus, the transferability brought by the increase of τ' is overwhelmed by the excessive perturbations.

5. Conclusions

In this paper, we propose a novel adversarial steganography, SPS-ENH. Through a novel sparse steganalysis-specific perturbations, we successfully use sparse ±1 perturbations to enhance the holistic security from the base cost function and deceive target CNN steganalyzers. Compared with the state-of-the-art adversarial attack, DDN (Decoupled Direction and Norm), the proposed cover enhancement method achieves significant less distortions to deceive the target CNN steganalyzers. SPS-ENH achieves comparable holistic security as the state-of-the-art method, ADV-EMB, even better under the evaluation of traditional steganalyzers. Meanwhile, the deceiving success rates of SPS-ENH are higher than these of ADV-EMB.

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.

References

[1] J. Fridrich, Steganography in Digital Media: Principles, Algorithms, and Applications, Cambridge University Press, 2009, http://dx.doi.org/10.1017/CBO9781139192903.

[2] B. Li, J. He, J. Huang, Y.Q. Shi, A survey on image steganography and steganalysis, J. Inf. Hiding Multimedia Signal Process. 2 (2) (2011) 142–172.

[3] S. Ma, X. Zhao, Steganalytic feature based adversarial embedding for adaptive JPEG steganography, J. Vis. Commun. Image Represent. 76 (2021) 103066, http://dx.doi.org/10.1016/j.jvcir.2021.103066.

[4] B. Qi, C. Yang, L. Tan, X. Luo, F. Liu, A novel haze image steganography method via cover-source switching, J. Vis. Commun. Image Represent. 70 (2020) 102814, http://dx.doi.org/10.1016/j.jvcir.2020.102814.

[5] J. Fridrich, T. Filler, Practical methods for minimizing embedding impact in steganography, in: Security, Steganography, and Watermarking of Multimedia Contents IX, Vol. 6505, International Society for Optics and Photonics, 2007, 650502.

[6] T. Filler, J. Judas, J. Fridrich, Minimizing additive distortion in steganography using syndrome-trellis codes, IEEE Trans. Inf. Forensics Secur. 6 (3) (2011) 920–935.

[7] W. Li, W. Zhang, L. Li, H. Zhou, N. Yu, Designing near-optimal steganographic codes in practice based on polar codes, IEEE Trans. Commun. (2020).

[8] T. Pevnỳ, T. Filler, P. Bas, Using high-dimensional image models to perform highly undetectable steganography, in: International Workshop on Information Hiding, Springer, 2010, pp. 161–177.

[9] V. Holub, J. Fridrich, Designing steganographic distortion using directional filters, in: 2012 IEEE International Workshop on Information Forensics and Security (WIFS), IEEE, 2012, pp. 234–239.

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

[11] B. Li, M. Wang, J. Huang, X. Li, A new cost function for spatial image steganography, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 4206–4210.

[12] J. Fridrich, J. Kodovskỳ, Multivariate Gaussian model for designing additive distortion for steganography, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2013, pp. 2949–2953.

[13] V. Sedighi, J. Fridrich, R. Cogranne, Content-adaptive pentary steganography using the multivariate generalized Gaussian cover model, in: Media Watermarking, Security, and Forensics 2015, Vol. 9409, International Society for Optics and Photonics, 2015, p. 94090H.

[14] V. Sedighi, R. Cogranne, J. Fridrich, Content-adaptive steganography by minimizing statistical detectability, IEEE Trans. Inf. Forensics Secur. 11 (2) (2015) 221–234.

[15] L. Guo, J. Ni, Y.Q. Shi, Uniform embedding for efficient JPEG steganography, IEEE Trans. Inf. Forensics Secur. 9 (5) (2014) 814–825.

[16] L. Guo, J. Ni, W. Su, C. Tang, Y.-Q. Shi, Using statistical image model for JPEG steganography: uniform embedding revisited, IEEE Trans. Inf. Forensics Secur. 10 (12) (2015) 2669–2680.

[17] W. Tang, S. Tan, B. Li, J. Huang, Automatic steganographic distortion learning using a generative adversarial network, IEEE Signal Process. Lett. 24 (10) (2017) 1547–1551.

[18] J. Yang, D. Ruan, J. Huang, X. Kang, Y. Shi, An embedding cost learning framework using GAN, IEEE Trans. Inf. Forensics Secur. 15 (2020) 839–851.

[19] W. Tang, B. Li, M. Barni, J. Li, J. Huang, An automatic cost learning framework for image steganography using deep reinforcement learning, IEEE Trans. Inf. Forensics Secur. 16 (2020) 952–967.

[20] J. Chen, W. Lu, Y. Fang, X. Liu, Y. Yeung, Y. Xue, Binary image steganalysis based on local texture pattern, J. Vis. Commun. Image Represent. 55 (2018) 149–156, http://dx.doi.org/10.1016/j.jvcir.2018.06.004.

[21] J. Fridrich, J. Kodovsky, Rich models for steganalysis of digital images, IEEE Trans. Inf. Forensics Secur. 7 (3) (2012) 868–882.

[22] B. Li, Z. Li, S. Zhou, S. Tan, X. Zhang, New steganalytic features for spatial image steganography based on derivative filters and threshold LBP operator, IEEE Trans. Inf. Forensics Secur. 13 (5) (2017) 1242–1257.

[23] X. Song, F. Liu, C. Yang, X. Luo, Y. Zhang, Steganalysis of adaptive JPEG steganography using 2D gabor filters, 2015, pp. 15–23.

[24] V. Holub, J. Fridrich, Low-complexity features for JPEG steganalysis using undecimated DCT, IEEE Trans. Inf. Forensics Secur. 10 (2) (2015) 219–228.

[25] R. Cogranne, T. Denemark, J. Fridrich, Theoretical model of the FLD ensemble classifier based on hypothesis testing theory, 2014, pp. 167–172.

[26] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, 2012, pp. 1097–1105.

[27] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, 2014, arXiv preprint arXiv:1409.1556.

[28] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, 2016, pp. 770–778.

[29] R. Girshick, Fast R-CNN, 2015, pp. 1440–1448.

[30] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: towards real-time object detection with region proposal networks, Vol. 2015, 2015, pp. 91–99.

[31] Y.T. Qian, J. Dong, W. Wang, T. Tan, Deep learning for steganalysis via convolutional neural networks, Electron. Imaging 9409 (2015).

[32] J. Ye, J. Ni, Y. Yi, Deep learning hierarchical representations for image steganalysis, IEEE Trans. Inf. Forensics Secur. 12 (11) (2017) 2545–2557.

[33] M. Boroumand, M. Chen, J. Fridrich, Deep residual network for steganalysis of digital images, IEEE Trans. Inf. Forensics Secur. 14 (5) (2019) 1181–1193.

[34] P. Li, J. Xie, Q. Wang, W. Zuo, Is second-order information helpful for large-scale visual recognition?, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 2070–2078.

[35] X. Deng, B. Chen, W. Luo, D. Luo, Fast and effective global covariance pooling network for image steganalysis, in: Proceedings of the ACM Workshop on Information Hiding and Multimedia Security, 2019, pp. 230–234.

[36] C. Szegedy, S. Ioffe, V. Vanhoucke, A. Alemi, Inception-v4, inception-resnet and the impact of residual connections on learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 31.

[37] R. Zhang, F. Zhu, J. Liu, G. Liu, Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis, IEEE Trans. Inf. Forensics Secur. 15 (2019) 1138–1150.

[38] S. Chopra, R. Hadsell, Y. LeCun, Learning a similarity metric discriminatively, with application to face verification, in: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), Vol. 1, IEEE, 2005, pp. 539–546.

[39] W. You, H. Zhang, X. Zhao, A siamese CNN for image steganalysis, IEEE Trans. Inf. Forensics Secur. 16 (2020) 291–306.

[40] I. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, ArXiv: Mach. Learn. (2014).

[41] A. Kurakin, I. Goodfellow, S. Bengio, Adversarial examples in the physical world, 2016, arXiv preprint arXiv:1607.02533.

[42] N. Carlini, D. Wagner, Towards evaluating the robustness of neural networks, 2017, pp. 39–57.

[43] J. Rony, L.G. Hafemann, L.S. Oliveira, I.B. Ayed, R. Sabourin, E. Granger, Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses, 2019, pp. 4322–4330.

[44] Y. Zhang, W. Zhang, K. Chen, J. Liu, Y. Liu, N. Yu, Adversarial examples against deep neural network based steganalysis, 2018, pp. 67–72.

[45] W. Tang, S. Tan, B. Li, J. Huang, Automatic steganographic distortion learning using a generative adversarial network, IEEE Signal Process. Lett. 24 (10) (2017) 1547–1551.

[46] S. Bernard, P. Bas, J. Klein, T. Pevny, Explicit optimization of min max steganographic game, IEEE Trans. Inf. Forensics Secur. 16 (2021) 812–823, http://dx.doi.org/10.1109/TIFS.2020.3021913, Conference Name: IEEE Transactions on Information Forensics and Security.

[47] P. Bas, T. Filler, T. Pevnỳ, " break our steganographic system": the ins and outs of organizing BOSS, in: International Workshop on Information Hiding, Springer, 2011, pp. 59–70.

[48] P. Bas, T. Furon, BOWS-2 Contest (Break Our Watermarking System), (Organized between the 17th of July 2007 and the 17th of April 2008) URL http://bows2.ec-lille.fr/, 0000.