







ARES: On Adversarial Robustness Enhancement for Image Steganographic Cost Learning

Qi Cui , Zhili Zhou , Senior Member, IEEE, Ruohan Meng , Shaowei Wang , Member, IEEE, Hongyang Yan , and Q. M. Jonathan Wu , Senior Member, IEEE

Abstract—Taking the steganalytic discriminators as the adversaries, the existing Generative Adversarial Networks (GAN)-based steganographic approaches learn the implicit cost functions to measure the embedding distortion for steganography. However, the steganalytic discriminators in these approaches are trained by the stego-samples with insufficient diversity, and their network structures offer very limited representational capacity. As a result, these steganalytic discriminators will not exhibit robustness to various steganographic patterns, which causes learning suboptimal cost functions, thus compromising the anti-steganalysis capability. To address this issue, we propose a novel GAN-based steganographic approach, in which the Diversified Inverse-Adversarial Training (DIAT) strategy and the Steganalytic Feature Attention (SteFA) structure are designed to train a robust steganalytic discriminator. Specifically, the DIAT strategy provides the steganalytic discriminator with an expanded feature space by generating diversified adversarial stego-samples; the SteFA structure enables the steganalytic discriminator to capture more various steganalytic features by employing the channel-attention mechanism on higher-order statistics. Consequently, the steganalytic discriminator can build a more precise decision boundary to make it more robust, which facilitates learning a superior steganographic cost function. Extensive experiments demonstrate that the proposed steganographic approach achieves promising anti-steganalysis capability over the state-of-the-arts under the same embedding payloads.

Index Terms—Steganography, GAN, Adversarial training, Cost learning.

Manuscript received 20 April 2022; revised 25 May 2023 and 31 October 2023; accepted 29 December 2023. Date of publication 12 January 2024; date of current version 10 April 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62372125, Grant 61972205, and Grant U1936218, in part by the Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2023B1515020041, in part by the China Scholarship Council under Grant 202008320533 and Grant 202109040027, and in part by the Guangzhou Research Foundation under Grant 202201010194 and Grant 202201020139. The Associate Editor coordinating the review of this manuscript and approving it for publication was Dr. Sen-Ching Samson Cheung. (Corresponding authors: Zhili Zhou; Ruohan Meng.)

Qi Cui and Ruohan Meng are with the Engineering Research Center of Digital Forensics, Ministry of Education, the School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China, and also with the School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore 639798 (e-mail: qi.cui@ntu.edu.sg; ruohan.meng@ntu.edu.sg).

Zhili Zhou, Shaowei Wang, and Hongyang Yan are with the Institute of Artificial Intelligence, Guangzhou University, Guangzhou 510006, China (e-mail: zhou_zhili@163.com; wangsw@gzhu.edu.cn; hyang_yan@gzhu.edu.cn).

Q. M. Jonathan Wu is with the Department of Electrical and Computer Engineering, University of Windsor, Windsor N9B 3P4, Canada (e-mail: jwu@uwindsor.ca).

Digital Object Identifier 10.1109/TMM.2024.3353543

I. INTRODUCTION

AS an important application in the field of covert communication, steganography embeds secret messages in multimedia imperceptibly and then enables the transmission in untrusted channels [1]. The research of image steganography started from the least significant bit (LSB) embedding, which was highly vulnerable to statistical analysis. To overcome this issue, the concept of embedding cost was empirically defined for evaluating the statistical anomalies [2], [3]. By using the coding algorithms, e.g., Syndrome-Trellis Code (STC) [4] and Steganographic Polar Code (SPC) [5], the practical embedding cost almost met the theoretical minimum bounds of the given cost function. Some other empirical approaches defined the cost functions by assigning lower embedding cost to textured regions [6], [7], [8]. However, the performance of empirically defined cost functions is limited by the human knowledge of proper regions for embedding, which will compromise anti-steganalysis capability.

Generative Adversarial Network (GAN) [9] consists of a pair of adversaries, i.e., a generator and a discriminator, which originally aims to generate “realistic-looking images” via noises obeying a prior continuous distribution, e.g., Gaussian distribution. The relationship between the generator and the discriminator coincides with that between steganographer and steganalyst. Generally, the embedding-probability generator learned the embedding probability maps of the cover-samples, serving as the cost function; the steganalytic discriminator learned the decision boundary between the cover-samples and the generated stego-samples, guiding the optimization [10], [11], [12]. Upon iteratively learning the generator and the discriminator, GAN-based steganographic approaches can define more precise embedding costs than empirical ones [13], [14], [15]. However, they still suffer the following issues. 1) The feature space of the stego-samples for training steganalytic discriminator is narrowed, as the stego-samples are derived from the same steganographic pattern, i.e., the current generator. In addition, considering that the cover-samples are drawn from the non-continuous distribution, the stego-samples inherit this sparsity. 2) The existing steganalytic discriminators generally have insufficient parameters and inadequately consider the feature interdependencies in feature extraction. Consequently, the structures of these steganalytic discriminators show very limited representational capacity, which makes them hard to capture more effective steganalytic features. The above can be respectively regarded as the external and internal issues of the existing steganalytic

TABLE I
NOTATION AND EXPLANATION

Notation	Explanation
x	Cover-image
y	Stego-image
p	Embedding-probability map
ρ	Embedding cost map
m	Modification map
δ	Adversarial perturbation
S	Solution space of the adversarial image
ϕ	Modification space of ± 1 embedding steganography
ψ	Solution space of the adversarial stego-image
ξ	Embedding payload
\mathcal{G}_p	Embedding-probability generator
\mathcal{G}_{adv}	Adversarial-sample generator
\mathcal{D}_s	Steganalytic discriminator
\mathcal{Q}	Encoder network
\mathcal{B}	Decision boundary of the steganalytic discriminator

discriminators. As a result, those steganalytic discriminators will build imprecise decision boundaries during the training process. In this case, even with a small progress of the generator, the stego-samples of the new steganographic pattern can deceive the non-robust steganalytic discriminator. This will cause the GAN-based steganographic approaches to learn a suboptimal cost function, thus compromising the anti-steganalysis capability. Therefore, it is reasonable and feasible for the steganalytic discriminator to learn more steganographic patterns to enhance its robustness.

Recently, training classifiers with mixing adversarial samples, i.e., adversarial training (AT) [16], [17], [18], exhibited the effectiveness of enhancing model robustness. Hence, the above external issue can be intuitively addressed by employing AT, where the steganalytic discriminator will be trained with mixed adversarial stego-samples. However, the feature space of these adversarial stego-samples will also be narrowed if one employs explicit adversarial attack methods in AT. To this end, the Diversified Inverse-AT (DIAT) strategy is proposed by further expanding the distribution of the adversarial stego-samples. To address the internal issue, a novel network structure of the steganalytic discriminator is designed, which employs the channel-attention mechanism [19] on the higher-order statistics. Owing to such a well-designed structure, the steganalytic discriminator can capture more various and effective steganalytic features. Consequently, the above techniques will tackle the external and internal issues of the existing steganalytic discriminators, which contributes to the steganalytic discriminator establishing a more precise decision boundary to enhance its robustness. In such a manner, the robust steganalytic discriminator will strengthen the embedding-probability generator, thereby enhancing the anti-steganalysis capability. The main contributions of this paper are summarized as follows:

- 1) A novel GAN-based steganographic cost learning approach called ARES is proposed. ARES constructs a robust steganalytic discriminator that forces the embedding-probability generator to become a superior cost function. Consequently, the stego-images with higher security performance can be obtained.
- 2) A Diversified Inverse-Adversarial Training (DIAT) strategy is designed for the external issue. Conducting AT

based on the Inverse Manner, and providing the steganalytic discriminator with diversified adversarial stego-samples, DIAT greatly expands the feature space.

- 3) A Steganalytic Feature Attention (SteFA) network structure is designed for the internal issue. SteFA employs the channel-attention mechanism on the higher-order statistics, which enable the steganalytic discriminator to capture more effective steganalytic features automatically.
- 4) Diversified steganalytic features generated in DIAT can be extracted more efficiently by SteFA. Thus, it enables fewer cover-samples to support a more diverse features-set in the training and improves the performance of the learned steganographic cost function, which is proven by experimental results in Section IV.

The rest of this paper is organized as follows. In Section II, the related works of empirical steganography, GAN-based steganography, and adversarial attack-based steganography are introduced. In Section III, we give the key observations and the hypothesis in the aspect of the robustness of the steganalyzers and then elaborate on the ARES approach. The experimental results and analysis are given in Section IV, and the conclusion is drawn in Section V. For convenience, the notations are summarized in Table I.

II. RELATED WORKS

A. Empirical Steganography

Early steganographic algorithms were commonly easy to detect by steganalysis tools. Focusing on this security issue, Fridrich and Pevny et al. [2], [3] proposed the concept of adaptive steganography to minimize the embedding impact. However, these approaches could not preserve the statistical characteristics well, causing unexpected distortion. To reduce the distortion, some empirical approaches employed hand-crafted features to design the cost functions. [6], [7], [8]. Generally, they tended to find the textured regions to modify. WOW [6] employed directional filters to build the cost function, which assigned low costs to the textured regions in all directions. Evaluating the distortion by the wavelet transformations, UNWARD [7] provided a universal steganographic cost function for the cover-images in arbitrary domains. In HILL [8], the cost function was defined by a group of filters containing a High-Pass Filter (HPF) and two Low-Pass Filters (LPFs).

With the dimensionality of the extracted features growing in nowadays steganalysis [20], [21], it is hard for these steganographic approaches to keep desirable security performance. Moreover, the development of deep learning encouraged many advanced CNN-based steganalysis algorithms [22], [23], [24], [25], [26], [27], [28]. Hence, it is critical to consider improving steganographic security with the guidance of powerful steganalyzers.

B. GAN-Based Steganography

Motivated by the concept of GAN, many GAN-based steganographic approaches have been proposed to enhance the anti-steganalysis capability of steganography. In the schemes of directly generating cover/stego-images, such as SGAN [29]

and Stego-WGAN [30], the main idea is to learn a distribution that exhibits the inherent ability to resist steganalyzers. Besides, SGSR-GAN [31], S-CycleGAN [32], and CEOA-GAN [33] enhanced the textures of the cover-images based on GAN. However, it is difficult for these approaches to find a good trade-off between steganographic security and image quality during training. Instead of generating secure cover-images/regions, GAN-based automatic steganographic cost learning approaches generated probability maps to indicate the embedding cost [10], [11], [12]. ASDL-GAN [10] adopted XuNet [22] as the steganalytic discriminator, and simulated the staircase ± 1 embedding function by a differentiable network. To accelerate the convergence, UT-GAN [11] employed a double-tanh function to simulate the embedding function. Wu-GAN [12] improved the UT-GAN by augmenting the channels of the generator. UMC-GAN [15] adopted a linear-clipped embedding simulator to alleviate the gradient vanishing problem of the double-tanh function.

Generally, the steganalytic discriminators in these algorithms were trained by stego-samples with insufficient diversity in feature space, and they did not have a well-designed structure to extract more effective steganalytic features.

C. Adversarial Attack-Based Steganography

CNN classifiers are vulnerable to malicious adversarial perturbations [34], [35], and the same happens with CNN-based steganalyzers. Based on this characteristic, Zhang et al. [36] added the adversarial perturbations to the cover-images before embedding messages, aiming to cheat a specific CNN-based steganalyzer. However, the magnitude of the added adversarial perturbations is larger than that of the steganographic modifications, which may raise suspicion with other detectors. Tang et al. [37] proposed an adversarial embedding scheme called ADV-EMB, based on asymmetric embedding [38]. According to the directions of the gradients, ADV-EMB attacked the steganalyzer by adjusting the probabilities of ± 1 modifications of a group of image elements. Qin et al. [39] argued that ADV-EMB ignored preserving image features since the changeable image elements were randomly selected, and proposed ensuring the same modification directions of the adjacent image elements. Ma et al. [40] built saliency maps for assessing the attack level using the absolute values of the gradients and then implemented the attack in a non-iterative way to improve efficiency. Bernard et al. [41] selected the least detectable adversarial stego-images to attack the strongest steganalyzer based on a min-max criterion. However, there is virtually no chance to access the target steganalyzer to conduct attacks in the practical scenario.

III. THE PROPOSED ARES APPROACH

A. Preliminaries

1) *GAN-Based Steganographic Cost Learning*: We denote x and y as the cover-image and its corresponding stego-image with the size of $H \times W$, and $x_{i,j}$ and $y_{i,j}$ as the (i, j) -th elements in the two images, respectively. Thus, the embedding cost of changing $x_{i,j}$ to $y_{i,j}$ can be denoted as $\rho_{i,j}$. According to

the existing approaches [10], [11], a GAN-based approach can learn the embedding cost implicitly. Specifically, the generator \mathcal{G}_p will learn the embedding probability map $p = \mathcal{G}_p(x)$. Afterwards, each embedding probability $p_{i,j}$ can be transformed to the embedding cost $\rho_{i,j}$ [42]:

$$\rho_{i,j} = \ln(1/(p_{i,j} - 2)). \quad (1)$$

Given an expected embedding payload ξ , the practical embedding payload can be evaluated by calculating:

$$\xi' = \sum_{i=1}^H \sum_{j=1}^W (-p_{i,j}^{+1} \log_2 p_{i,j}^{+1} - p_{i,j}^{-1} \log_2 p_{i,j}^{-1} - p_{i,j}^0 \log_2 p_{i,j}^0). \quad (2)$$

The practical embedding-capacity will be kept as ξ by minimizing:

$$\mathcal{L}_{emb} = |\xi' - \xi|. \quad (3)$$

Note $p_{i,j}^{+1} = p_{i,j}^{-1} = p_{i,j}/2$ since it needs to be subject to the condition of symmetric embedding, and $p_{i,j}^0 = 1 - p_{i,j}$. In addition, we use the double-tanh function designed in UT-GAN [11] to make the staircase embedding function differentiable, which is defined as:

$$m_{i,j} = -0.5 \times \tanh(\beta(p_{i,j} - 2 \times r_{i,j})) + 0.5 \times \tanh(\beta(p_{i,j} - 2 \times (1 - r_{i,j}))), \quad (4)$$

$$\tanh(x) = (e^x - e^{-x}) / (e^x + e^{-x}), \quad (5)$$

where β controls the extent of the slope, $r_{i,j}$ is the element of the random noise map $r \sim U(0, 1)$. Then the corresponding stego-image y can be obtained by adding m to x . Next, x and y will be fed into the steganalytic discriminator \mathcal{D}_s for calculating the discriminative loss:

$$\mathcal{L}_{dis} = -[\log(\mathcal{D}_s(x)) + \log(1 - \mathcal{D}_s(y))], \quad (6)$$

where l_i denotes the cover label or stego label.

2) *Adversarial Attack and Adversarial Training*: Given an image I_l with the ground truth label l , the adversarial attack algorithm calculates the adversarial perturbation δ by the following optimization function.

$$\delta := \operatorname{argmax}_{\delta \in S} \mathcal{L}(f_\theta(I_l + \delta), l), \quad (7)$$

where $S = \{\delta : \|\delta\|_u \leq \epsilon\}$, $u \in (0, \infty)$, $f_\theta(\cdot)$ is the trained target classifiers with parameter θ , and $\mathcal{L}(\cdot, \cdot)$ is the classification loss function. Then an AT strategy can be formulated as:

$$\min_{\theta} \frac{1}{n} \sum_{i=1}^n \max_{\delta \in S} \mathcal{L}(f_\theta(I_{l_i} + \delta), l). \quad (8)$$

The purpose of the AT strategy is solving the minimax problem to defend against potential attacks.

TABLE II
PERFORMANCES (IN %) OF THE STEGANALYZERS ON DETECTING DIFFERENT STEGANOGRAPHIC MODIFICATIONS

	Test	WOW	HILL	UT-GAN
Yedroudj-Net	18.85	26.01(↑7.16)	29.53(↑10.68)	35.52(↑16.67)
SRNet	13.53	16.59(↑3.06)	21.96(↑8.43)	29.57(↑16.04)

The bold values indicate a more robust steganalyzer.

TABLE III
PERFORMANCES (IN %) OF THE STEGANALYZERS ON DETECTING DIFFERENT STEGANOGRAPHIC-LIKE MODIFICATIONS

	Test	Avg32	Avg16	Gauss	Lap4
Yedroudj-Net	18.85	33.17(↑14.32)	35.12(↑16.27)	21.59(↑2.74)	27.35(↑8.50)
SRNet	13.53	21.98(↑8.45)	21.62(↑8.09)	14.85(↑1.32)	19.46(↑5.93)

The bold values indicate a more robust steganalyzer.

B. Our Motivations

Existing research indicated that the steganalysis algorithms are sensitive to steganographic algorithms of different patterns [43]. Meanwhile, it was revealed that CNN models are vulnerable to high-frequency components, e.g., filter-guided noises [44], [45], [46]. Furthermore, besides adversarial noises, the CNN-based classifiers can also be attacked by some semi-random noises [47], [48]. Considering the steganographic ± 1 modifications can be regarded as a type of “payload-constrained” noise, we also wonder whether the performance of the CNN-based steganalyzers degrades in detecting the stego-samples within ± 1 embedding from other steganographic patterns, especially unknown patterns. Therefore, the two following preliminary studies were conducted.

Initially, two CNN-based steganalyzers, Yedroudj-Net [24] and SRNet [25] were trained with the stego-samples produced by S-UNIWARD [7] under the payload of 0.4 bpp (bit per pixel). Note that Yedroudj-Net uses fixed HPFs to extract high-frequency features, while SRNet adopts the learnable structures to extract high-frequency features. Consequently, we obtained Yedroudj-Net and SRNet with the test error rates of 18.85% and 13.53% on S-UNIWARD, respectively. In the first preliminary study, we used the above two steganalyzers to distinguish the cover-samples and their corresponding stego-samples produced by WOW [3], HILL [8], and UT-GAN [11] with 0.4 bpp. The specific results are shown in Table II. In the second preliminary study, we empirically used several common image filters as cost functions to construct several types of steganographic-like modifications with 0.4 bpp, aiming to further observe the performance of the trained steganalyzers on more unknown steganographic patterns. The examples of the modification maps are shown in Fig. 1, where Avg32 and Avg16 denote the average filters with the kernel sizes of 32 and 16, and then be subtracted by 255, respectively; Gauss denotes the default Gaussian filter with a kernel size of 3; Lap4 denotes the Laplace filter with a kernel size of 4. The specific results are shown in Table III. From Tables II and III, we can obtain the two following observations, respectively.

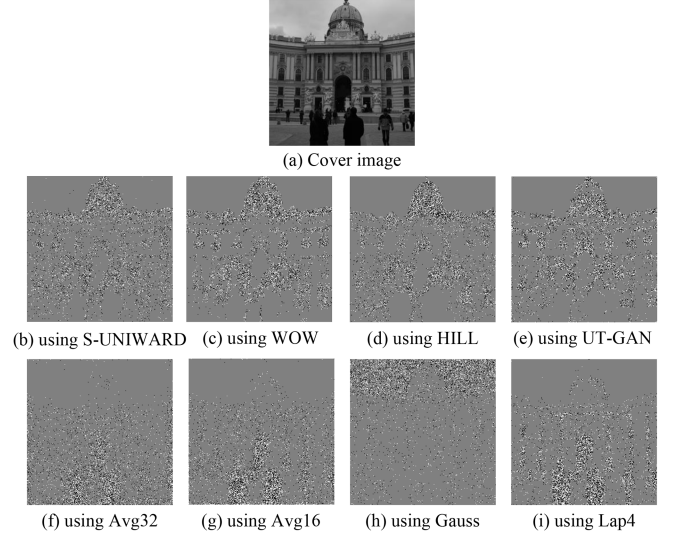


Fig. 1. Examples of the enhanced modification maps using different steganographic approaches with 0.4 bpp. The white pixels and the black pixels represent +1 and -1 modifications, respectively.

Observation 1: A CNN-based steganalyzer trained under one steganographic pattern degraded the performance on detecting the stego-images from unknown steganographic patterns.

Observation 2: A CNN-based steganalyzer using limited HPFs structures performed worse than the one using learnable structure, on unknown steganographic patterns.

These observations imply that the steganalyzer will show poor robustness when trained with a single steganographic pattern and/or on limited HPFs structures. Also, such a steganalytic discriminator may become a non-robust one in the GAN-based cost learning. Consequently, it will mislead the generator to be sub-optimal and thus compromise the anti-steganalysis capability. According to the above observation and analysis, we propose the following hypothesis.

Hypothesis: In GAN-based cost learning, enhancing the robustness of the steganalytic discriminator benefits in improving the security performance of the learned steganographic cost function.

Explanation: Suppose the training reaches a local Nash equilibrium at time t , where the embedding-probability generator \mathcal{G}_p^t is learned under the supervision of the steganalytic discriminator \mathcal{D}_s^t with decision boundary \mathcal{B}^t . As illustrated in Fig. 2(a), the stego-samples will show limited diversity from y_1^{t-2} to y_1^t , as they are all generated based on \mathcal{G}_p^t . As a result, an imprecise decision boundary \mathcal{B}^t would be built. In this case, the optimal steganographic distortion v of y_1^t is not desired, even y_1^t can confuse the steganalytic discriminator. Whereas, as illustrated in Fig. 2(b), once enabling the steganalytic discriminator to learn adversarial stego-samples y_2^{t-2} and y_2^{t-1} that tend to cause misclassification, the steganalytic discriminator will be pushed to build a precise decision boundary. On this basis, if we further increase the diversity of the adversarial stego-samples as shown in Fig. 2(c), the steganalytic discriminator will be forced to establish a more precise decision boundary. Consequently,

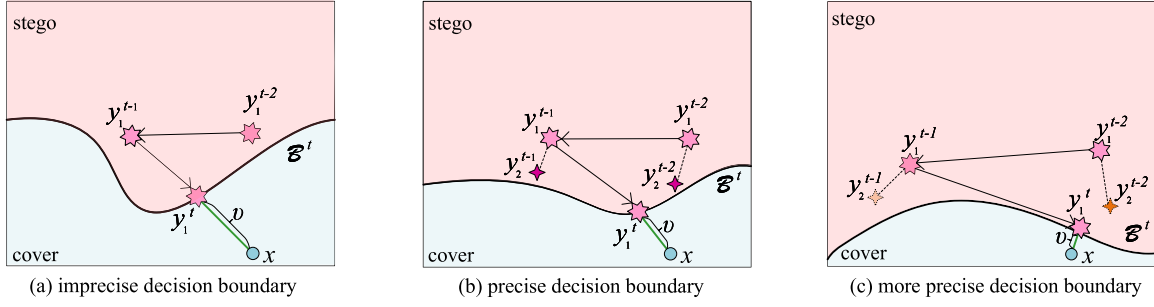


Fig. 2. Conceptual illustration of a robust steganalytic discriminator benefits in improving the security performance of the learned cost function, by showing the decision boundaries. From (a) to (c), the decision boundary becomes more precise, thus the steganalytic discriminator becomes more robust. \mathcal{B}^t denotes the decision bounds of the steganalytic discriminators. x denotes a cover-sample. y_1^{t-2} to y_1^t denote the corresponding stego-samples produced by \mathcal{G}_p^{t-2} to \mathcal{G}_p^t . y_2^{t-2} and y_2^{t-1} in (b) and (c) denote the corresponding adversarial stego-samples with limited diversity and sufficient diversity, respectively. v denotes the corresponding practical steganographic distortion, respectively.

the steganalytic discriminator \mathcal{D}_s^t can be more robust. Under such a strict condition, the optimal steganographic distortion v will be reduced effectively, which helps to enhance the security performance.

C. Inverse Manner for Conducting AT on Steganalytic Discriminator

The AT strategy is adopted for the external issue since AT calculates the most harmful samples, i.e., adversarial stego-samples, to correct the decision boundary of the steganalytic discriminator efficiently. In ± 1 embedding steganography, the modification space of $m_{i,j}$ a pixel is restricted to the discrete set $\phi = \{-1, 0, +1\}$ in an 8-bit grayscale image. To avoid introducing redundant information, we set the same modification space for the adversarial perturbation in our AT task, i.e., $\psi = \{\delta : \|\delta\|_\infty = 1\}$. The amount of the perturbed pixels should be consistent with the embedding payload by minimizing:

$$\mathcal{L}_{num} = \left| \sum_{i=1}^H \sum_{j=1}^W \frac{|\delta_{i,j}|}{H \times W} - \xi \right|. \quad (9)$$

Next, the reasons why conventional AT cannot be performed on steganalytic discriminators are given as follows. On the one hand, it is unreasonable to conduct AT on the cover-class by computing the adversarial perturbation and then adding it to the cover-samples. That is because the resulting adversarial samples will essentially become stego-samples. On the other hand, it is also infeasible to conduct AT on the stego-class by computing the adversarial perturbation and adding it to the stego-samples. The reason is the addition of such adversarial perturbation will cause the resulting adversarial samples to not obey the condition ψ . Therefore, the Inverse Manner is designed, where AT is conducted on the stego-class by calculating the adversarial perturbation and then adding the adversarial perturbation to the cover-samples. This is exactly the inverse form of the conventional defined AT, i.e., (8), and can be defined as the following problem.

$$\min_{\mathcal{D}_S} \frac{1}{n} \sum_{k=1}^n \max_{\delta \in \psi} \mathcal{L}_{\mathcal{D}}(\mathcal{D}_S(x_k + \delta), 0). \quad (10)$$

It first calculates the adversarial perturbation δ , which will make the adversarial sample $(x_k + \delta)$ to be classified as the stego-label 0 with a higher probability. Then the steganalytic discriminator \mathcal{D}_S will be trained for minimizing the classification loss $\mathcal{L}_{\mathcal{D}}$ with the adversarial sample by optimizing its parameters. We denote such AT strategy as Inverse Adversarial Training (IAT).

D. Diversified Inverse-Adversarial Training (DIAT) Strategy

In IAT, the adversarial samples generated by a single attack algorithm will cluster in sample space and thus lose the diversity [49], [50], [51]. Inspired by constructing distributed adversarial samples [49], [50], in which an entropy term of the perturbation δ is added to increase the support of $p(\delta)$, Diversified IAT (DIAT) strategy is proposed. It can further expand the sample distribution based on (10):

$$\min_{\mathcal{D}_S} \frac{1}{n} \sum_{k=1}^n \max_{p(\delta) \in \mathcal{P}_\psi} \{ \mathbb{E}[\mathcal{L}_{\mathcal{D}}(\mathcal{D}_S(x_k + \delta), 0)] + \lambda \mathcal{H}[p(\delta|x_k)] \}, \quad (11)$$

where the last term denotes the entropy of $p(\delta|x_k)$. To better approximate the distribution $p(\delta|x_k)$, an auxiliary generator \mathcal{G}_{adv} is leveraged such that $\delta = \mathcal{G}_{adv}(x_k; z)$. The diversity of a distribution will enhance once its density increases. Following the existing work [50], this can be achieved by maximizing the variational lower bound $\mathbb{E}_{z \sim p(z)}[\log q(z|\delta)]$ of the perturbation entropy, as detailed in (12).

$$\mathcal{H}[p(\delta|x)] \geq \mathbb{E}_{z \sim p(z)}[\log q(z|\delta)] + c. \quad (12)$$

In our context, q represents a diagonal Gaussian distribution, parameterized by a compact encoder network \mathcal{Q} , which provides its mean and standard deviation, and c is a minor constant. Consequently, \mathcal{G}_{adv} generate the perturbation $\delta = \mathcal{G}_{adv}(x_k; z)$, with z being sampled from a uniform distribution $U(-1, 1)$. This leads to a reformulation of the DIAT objective from (11) as:

$$\min_{\mathcal{D}_S} \max_{\mathcal{G}_{adv}, \mathcal{Q}} \frac{1}{n} \sum_{k=1}^n \{ \mathbb{E}[\mathcal{L}_{\mathcal{D}}(\mathcal{D}_S(x_k + \mathcal{G}_{adv}(x_k; z)), 0)] + \lambda \log q_{\mathcal{Q}}(z|\mathcal{G}_{adv}(x_k; z)) \}, \quad (13)$$

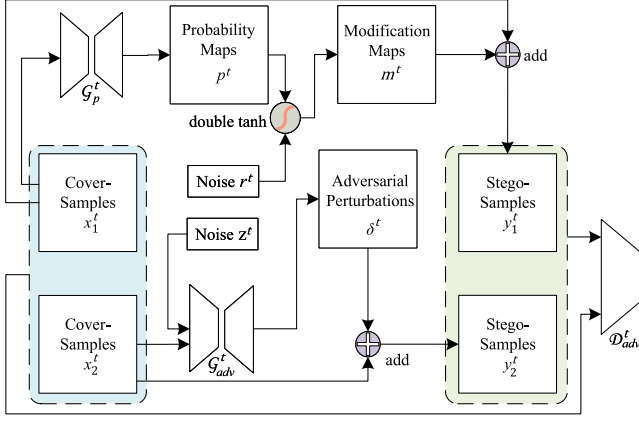


Fig. 3. Training details of the proposed DIAT strategy.

where λ controls the weight of entropy term *w.r.t.* the adversarial perturbations. By applying the stochastic gradient descent on the parameters of \mathcal{D}_S , and stochastic gradient ascent on that of \mathcal{G}_{adv} and \mathcal{Q} , the optimization can be achieved. In addition, to train the steganalytic discriminator with the samples from both \mathcal{G}_{adv} and \mathcal{G}_p , a batch of cover-samples are equally divided into two sub-batches to fed in \mathcal{G}_{adv} and \mathcal{G}_p simultaneously during training. Then the objective of the steganalytic discriminator is represented by:

$$\min_{\mathcal{D}_S} \max_{\mathcal{G}_{adv}, \mathcal{Q}} \left\{ \frac{1}{n} \sum_{k=1}^n \mathbb{E}_{p(z), p(x)} [\mathcal{L}_{\mathcal{D}} (\mathcal{D}_S ((x_k + \mathcal{G}_{adv}(x_k; z); y_k), 0) + \mathcal{L}_{\mathcal{D}} (\mathcal{D}_S (x_k), 1) + \lambda \log q_{\mathcal{Q}} (z | \mathcal{G}_{adv}(x_k; z)) + \mathcal{L}_{dis}] \right\}. \quad (14)$$

By solving the above function, the distribution of the adversarial stego-samples can be expanded effectively.

In the implementation of the DIAT strategy, as shown in Fig. 3, the cover-sample x_1^t in one of the sub-batches is fed in \mathcal{G}_p^t to get the corresponding embedding probability maps p^t . Then p^t is sent to the embedding function (4), together with the random noise r^t . Then the function outputs modification maps m^t . The other sub-batch x_2^t with z^t are fed in \mathcal{G}_{adv}^t to output adversarial perturbations δ^t , which aims to attack \mathcal{D}_S^{t-1} . Then m^t and δ^t are added to x_1^t and x_2^t , respectively, to obtain two sub-batches of stego-labeled samples y_1^t and y_2^t . Finally, y_1^t and y_2^t are merged and then sent to the steganalytic discriminator \mathcal{D}_S^t . The detailed progress of the proposed DIAT is shown in Algorithm 1.

E. Steganalytic Feature Attention (SteFA) Structure

The proposed DIAT provides more diverse stego-samples for the steganalytic discriminator. However, one still cannot build a robust steganalytic discriminator if the structure of the steganalytic discriminator does not support exploiting various steganographic patterns. Thus, SteFA is designed to capture more effective steganalytic features and analyze the complex feature

Algorithm 1: Diversified Inverse-Adversarial Training Strategy.

Input: Embedding-probability generator \mathcal{G}_p^{t-1} ; adversarial-sample generator \mathcal{G}_{adv}^{t-1} ; steganalytic discriminator \mathcal{D}_S^{t-1} ; sub-batch of cover-images x_1^t and x_2^t ; noise r^t and z^t ; the number of maximization phases T .

Output: Learned embedding-probability generator \mathcal{G}_p^T .

- 1: **While** $t < T$:
- 2: update $\mathcal{G}_p^t \leftarrow \mathcal{G}_p^{t-1}$ by minimizing (3) and maximizing (6).
- 3: update $\mathcal{G}_{adv}^t \leftarrow \mathcal{G}_{adv}^{t-1}$ by solving the inner maximization of (14).
- 4: $p^t = \mathcal{G}_p^t(x_1^t)$.
- 5: $m^t = \text{double-tanh}(p^t, r^t)$.
- 6: $y_1^t = x_1^t + m^t$.
- 7: $\delta^t = \mathcal{G}_{adv}^t(x_2^t; z^t)$.
- 8: $y_2^t = x_2^t + \delta^t$.
- 9: $\mathcal{D}_S^{t-1}(\text{stack}(x_1^t, x_2^t), \text{stack}(y_1^t, y_2^t))$.
- 10: update $\mathcal{D}_S^t \leftarrow \mathcal{D}_S^{t-1}$ by solving the outer minimization of (14).
- 11: $t = t + 1$.
- 12: **End**

interdependencies automatically. SteFA enhances the perception of steganalytic features by employing more learnable HPFs, channel-attention mechanism, and global covariance pooling. The detailed structure is shown in Fig. 4.

More learnable HPFs: Recall Observation 2 that the CNN-based steganalyzer performs better when using learnable structures to capture high-frequency features. However, directly using the noise residual structure of SRNet may have a negative influence on convergence, since its random initialization discards any prior knowledge. To enable the steganalytic discriminator to extract steganalytic features from multiple perspectives while maintaining effectiveness, two groups of parallel HPFs are set in its first layer. One group contains 30 fixed high-pass kernels from SRM, while the other group contains 30 learnable high-pass kernels initialed by the fixed ones, to adaptively learn complementary steganalytic features. As the parameters for extracting high-frequency features increase, this design encourages the networks to capture more sufficient high-frequency patterns while keeping the feature extraction of the widely used traditional high-frequency features.

Channel-attention mechanism: To facilitate the learnable HPFs to exploit more various and significant steganographic features, we establish cross-channel interdependencies between the aggregated features maps. Specifically, a channel-attention module [19] is constructed inside the backbone, as shown in Fig. 4. It can make the networks build inherent interdependencies among the feature channels so that the networks pay more attention to the significant features adaptively. The channel-attention module consists of three main steps, i.e., squeeze, excitation, and scale [19]. The first step is extracting the features from a

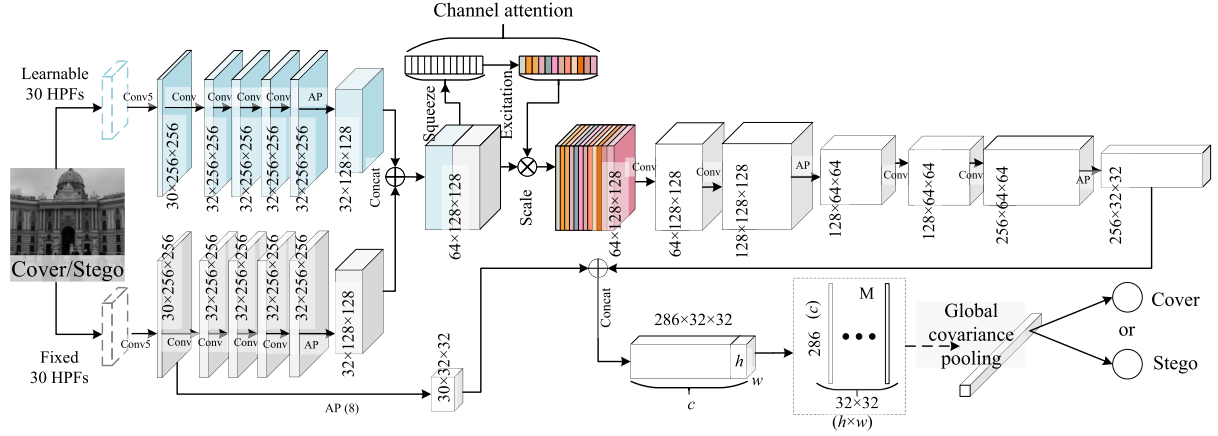


Fig. 4. Details of the proposed SteFA structure. Conv5/Conv denotes the convolution operation with kernel size of 5/3. AP denotes the average pooling. Concat denotes the channel concatenation.

channel to a descriptor by global average pooling, which can be formulated as

$$Fea_c^1 = \mathcal{F}_{sq}(Fea_c^0) = \frac{1}{h \times w} \sum_{i=1}^h \sum_{j=1}^w Fea_{c,i,j}^0, \quad (15)$$

where $Fea_c^0 = \{fea_1^0, \dots, fea_c^0\}$ is a group of feature maps with the size of $h \times w \times c$. Next, the excitation operation \mathcal{F}_{ex} will adaptively capture inherent interdependencies by assigning weights for the feature channels:

$$Fea_c^2 = \mathcal{F}_{ex}(Fea_c^1) = \text{Sigmoid}(\text{lin}_2(\text{Relu}(\text{lin}_1(Fea_c^1)))) , \quad (16)$$

where lin_1 and lin_2 are two linear layers. Then, channel-wise multiplication is used to transform the channel-weights into the original feature maps:

$$Fea_c^3 = \mathcal{F}_{scale}(Fea_c^0, Fea_c^2) = Fea_c^0 \cdot Fea_c^2. \quad (17)$$

The channel-attention mechanism promotes the networks to exploit the connections of the steganographic feature channels, which integrate the fixed HPFs and the learnable HPFs implicitly. In the second step, the output of the squeeze operation is attached to the weights to learn the non-linear interactions of the channels. Next, the original channels are updated by multiplying them with the learned weights of the excitation. As a result, it can stimulate the networks to be aware of more significant steganalytic features flexibly.

Global covariance pooling: As the features flow into deeper layers, the high-frequency information extracted from the HPFs will be suppressed. Particularly, the average pooling operation will suppress the extraction of high-frequency features. To address these limitations and enhance the sensitivity of the steganalytic discriminator to more high-frequency variations, global covariance pooling [52] is adopted to extract second-order characteristics. As indicated in fine-grained classification tasks and some steganalysis algorithms [53], [54], global covariance pooling allows the network to have more representative and discriminatory power. As shown in Fig. 4, for the output feature maps $Fea^{sl} = \{fea_1, \dots, fea_c\}$ with the size of $h \times w \times c$ from the

second-last layer, it first converts Fea^{sl} to a feature matrix M with $h \times w$ feature vectors of c -dimension. Then one can compute the covariance matrix by

$$\Sigma = M\bar{M}^T, \quad (18)$$

where $\bar{M} = \frac{1}{h \times w}(M - \frac{1}{h \times w}\mathbf{1})$, in which I and $\mathbf{1}$ are the identity matrix and the matrix of all ones. Then the eigenvalue decomposition (EIG) is performed for the obtained Σ by

$$\Sigma = U\Lambda U^T, \quad (19)$$

where U is an orthogonal matrix and $\Lambda = \{\kappa_1, \dots, \kappa_c\}$ is a diagonal matrix. Next, we follow [52] to find the square root of the eigenvalues for global covariance pooling.

This operation helps increase the awareness of the useful high-frequency features while converging fast. Cooperated with the channel-attention mechanism, it can make the networks more sensitive to capture the feature variance of steganographic modifications, which will improve the representational capacity of the steganalytic discriminator significantly.

F. Network Structures of Embedding-Probability Generator, Adversarial-Sample Generator, and the Encoder

Based on UT-GAN [11], we design a new structure for our embedding-probability generator \mathcal{G}_p , as shown in Table IV. Considering the generation of the embedding-probability map can be regarded as an image translation task, the Instance Normalization [55] is used to substitute the Batch Normalization [56] in UT-GAN. For a given cover-sample, \mathcal{G}_p will generate the embedding-probability map that is highly dependent on the cover-sample itself rather than the other samples in the batch. In the upsampling process, bilinear operation with convolution and padding is adopted to replace the transpose convolution in U-Net structure, which can avoid the unexpected artifacts caused by the excessive zero paddings. In addition, the concatenation operations are also employed to make better use of the features from different levels. The specific concatenations are shown in

TABLE IV
NETWORK ARCHITECTURE OF \mathcal{G}_p

Group	Concat	Type	Output size
Input: x			$1 \times 256 \times 256$
Group 1		Conv+IN+LReLU	$16 \times 128 \times 128$
Group 2		Conv+IN+LReLU	$32 \times 64 \times 64$
Group 3		Conv+IN+LReLU	$64 \times 32 \times 32$
Group 4		Conv+IN+LReLU	$128 \times 16 \times 16$
Group 5		Conv+IN+LReLU	$128 \times 8 \times 8$
Group 6		Conv+IN+LReLU	$128 \times 4 \times 4$
Group 7		Conv+IN+LReLU	$128 \times 2 \times 2$
Group 8		Conv+IN+LReLU	$128 \times 1 \times 1$
Group 9	Group 7	BiL+(LReLU+Pad+Conv+IN) $\times 2$	$256 \times 2 \times 2$
Group 10	Group 6	BiL+(LReLU+Pad+Conv+IN) $\times 2$	$256 \times 4 \times 4$
Group 11	Group 5	BiL+(LReLU+Pad+Conv+IN) $\times 2$	$256 \times 8 \times 8$
Group 12	Group 4	BiL+(LReLU+Pad+Conv+IN) $\times 2$	$256 \times 16 \times 16$
Group 13	Group 3	BiL+(LReLU+Pad+Conv+IN) $\times 2$	$128 \times 32 \times 32$
Group 14	Group 2	BiL+(LReLU+Pad+Conv+IN) $\times 2$	$64 \times 64 \times 64$
Group 15	Group 1	BiL+(LReLU+Pad+Conv+IN) $\times 2$	$32 \times 128 \times 128$
Group 16		BiL+(LReLU+Pad+Conv+IN) $\times 2$	$1 \times 256 \times 256$
Output: p		(Sigmoid-0.5)+ReLU	$1 \times 256 \times 256$

Conv, IN, BiL, LReLU, Concat, and Pad denotes linear, 3×3 convolution, instance normalization, Bilinear upsampling (2), LeakyReLU (0.1), channel concatenation, and zero pad (1), respectively.

TABLE V
NETWORK ARCHITECTURE OF \mathcal{G}_{adv} AND \mathcal{Q}

\mathcal{G}_{adv}	Output size	\mathcal{Q}	Output size
Input: x, z		Input	
$f1 = 2 \times \text{Linear}(z)$	1024	Conv+BN+ReLU	$32 \times 5 \times 5$
$f2 = \text{view}(f1)$	$1 \times 32 \times 32$	Conv+BN+ReLU	$64 \times 4 \times 4$
$f3 = \text{Concat}(x, f2)$	$2 \times 32 \times 32$	Conv+BN+ReLU	$128 \times 4 \times 4$
$3 \times \text{ResBlk}(f3)$	$128 \times 32 \times 32$	Conv+BN+ReLU	$256 \times 4 \times 4$
Conv	256×256	AP	$256 \times 1 \times 1$
		Linear	256

$f1, f2$, and $f3$ denote the feature maps. Linear, Conv, BN, AP, Concat, and ResBlk denotes linear, 2×2 convolution, batch normalization, average pooling, channel concatenation, and residual block, respectively.

Table IV. Eventually, subtracting 0.5 after the Sigmoid function ensures that the output probability value will be bounded to (0, 0.5).

Following [50], a simple image-to-image structure of the adversarial-sample generator \mathcal{G}_{adv} was adopted, which is equipped with 3 residual blocks. The encoder \mathcal{Q} uses a four-layer convolutional group and is equipped with an average pooling structure. The detailed architecture of \mathcal{G}_{adv} , and the encoder \mathcal{Q} are detailed in Table V.

IV. EXPERIMENTS

A. Settings

1) *Datasets*: The following two datasets are used in the experiments.

BOSSBase256: This dataset is used to train the proposed approach and the contrastive GAN-based steganographic approaches. It contains 10,000 grayscale images with the size of 256×256 by resizing the original images in the BOSSBase v1.01 [57] dataset using bilinear interpolation.

ALASKA256: This dataset is obtained by randomly selecting 50,000 uncompressed grayscale images with the size of 256×256 from the ALASKA #2 [58] dataset. It was then randomly separated into two parts with 40,000 images and 10,000 images separately. The first part with its corresponding stego-images is used to train the steganalyzers, and the second part with

its corresponding stego-images is used to test the performance of different steganographic approaches.

2) *Contrastive Steganographic Approaches*: Four contrastive steganographic approaches are tested in the experiments.

UT-GAN [11]: We used the trained model with the TensorFlow version of UT-6HPF-GAN as the network structure, which contains a U-Net [59]-based generator with 6 HPFs and a XuNet-based steganalytic discriminator. The corresponding stego-sample sets for training the steganalyzers and evaluating the security performance are respectively generated by the trained model.

S-UNIWARD [7]: It is the spatial version of the UNIWARD steganographic approach. It employs multiple direction filters to calculate steganographic costs, by catching steganographic changes in any direction.

WOW [3]: It defines the steganographic cost function by computing the residuals of the image in each direction.

HILL [8]: It uses a HPF and two LPFs to assign a lower steganographic cost to regions with higher texture complexity in a more concentrated manner.

3) *Steganalyzers*: There were five steganalyzers used to test the security performance of the steganographic approaches, including four CNN-based methods and an empirical method. The steganalyzers were introduced as follows. **SRNet [25]**: It mainly uses two types of shortcut structures to capture the noise residuals. There are no HPFs in the network structure.

Ye-Net [23]: It employs 30 HPFs and uses a new activation function called Truncated Linear Unit (TLU) to capture high-frequency features, and utilizes a selection channel to boost the detection performance.

Yedroudj-Net [24]: It employs 30 HPFs and uses the TLU in Ye-Net to capture high-frequency features. Note that we utilize the covariance pooling [52] version [54].

Zhu-Net [60]: It utilizes separable convolution to capture the channel correlation of residuals, and employs spatial pyramid pooling to enhance the representation ability.

SRM [20]: This empirical steganalyzer extracts steganalytic rich features by handcrafted HPFs. The classification process is equipped with an FLD-based ensemble classifier [61].

4) *Settings*: The Adam was used as the optimizer with a learning rate of 0.0001. β is set to 60 and 12,000 for the staircase embedding function in training and testing, respectively. λ is set to 1 for the entropy term. Note that the PGD [17] attack is adopted with 5 gradient steps in the IAT setting.

B. Ablation Studies of Conducting AT Strategies on Steganalytic Discriminators

In this part, we presented the case study of testing the performance of three variants of our ARES, i.e., with DIAT, IAT, or without AT.

1) *Comparison of the Performance of ARES-IAT and ARES-NoAT*: According to the experimental results shown in Table VI, it can be found that ARES-IAT increases the security performances with all the payloads. This verifies the effectiveness of

TABLE VI
SECURITY PERFORMANCE (IN %) OF DIFFERENT STEGANOGRAPHIC APPROACHES AGAINST DIFFERENT STEGANALYZERS

Steganalyzer	Steganographic model	0.1 bpp	0.2 bpp	0.3 bpp	0.4 bpp
SRNet	ARES-DIAT	44.90 (↑ 1.11)	40.74 (↑ 3.28)	33.17 (↑ 4.02)	30.28 (↑ 6.12)
	ARES-IAT	43.79 (↑ 4.35)	37.46 (↑ 3.44)	27.15 (↑ 2.17)	24.14 (↑ 1.65)
	ARES-noAT	39.44	34.02	24.98	22.49
Ye-Net	ARES-DIAT	46.51 (↑ 1.37)	43.69 (↑ 2.99)	37.35 (↑ 3.83)	34.06 (↑ 7.18)
	ARES-IAT	45.14 (↑ 4.36)	40.77 (↑ 4.49)	33.52 (↑ 2.02)	26.88 (↑ 1.56)
	ARES-noAT	40.78	36.28	31.50	25.32
Yedroudj-Net	ARES-DIAT	45.03 (↑ 0.84)	39.77 (↑ 1.94)	35.76 (↑ 3.75)	32.57 (↑ 7.51)
	ARES-IAT	44.19 (↑ 4.29)	37.83 (↑ 2.21)	32.01 (↑ 1.86)	25.06 (↑ 0.67)
	ARES-noAT	39.90	35.62	30.15	24.39
Zhu-Net	ARES-DIAT	41.04 (↑ 1.08)	34.86 (↑ 2.47)	29.44 (↑ 3.49)	26.30 (↑ 4.98)
	ARES-IAT	39.96 (↑ 3.33)	32.39 (↑ 1.94)	25.95 (↑ 1.53)	21.32 (↑ 1.01)
	ARES-noAT	36.63	30.45	24.42	20.31
SRM	ARES-DIAT	46.25 (↑ 1.37)	42.06 (↑ 1.59)	41.92 (↑ 1.90)	38.89 (↑ 2.42)
	ARES-IAT	44.88 (↑ 1.63)	40.47 (↑ 1.18)	40.02 (↑ 0.84)	36.47 (↑ 0.13)
	ARES-noAT	43.25	39.29	39.18	36.34

the IAT strategy. We speculate that the steganalytic discriminator in ARES-noAT cannot acquire more diversified and significant steganographic features since all the stego-samples were produced *w.r.t* the embedding-probability generator in the training. The homogeneous stego-samples can not provide sufficient diversity, though the steganalytic discriminator with SteFA is capable of extracting more features. Therefore, once providing stego-samples with greater diversity in the training of the steganalytic discriminator, it can make use of these features effectively. It proves the proposed theory that performing an AT strategy on the steganalytic discriminator can help to improve steganographic security. In addition, it can be observed in Table VI that the incremental margins (the values inside the brackets) between ARES-IAT and ARES-noAT tend to decrease as the embedding rate increases. For example, ARES-IAT yielded error rates of 43.79%, 37.46%, 27.15%, and 24.14% with 0.1 bpp to 0.4 bpp on the detection of SRNet, which gained the increments by 4.35%, 3.44%, 2.17%, and 1.65% compared with ARES-noAT. The reason behind this phenomenon is that the IAT strategy can only provide adversarial samples by the specific attack algorithm. Hence, the improvement is constrained by the limited diversity of stego-samples. Accordingly, the enhancement is gradually weakening as the modification rate (payload) increases.

2) *Comparison of the Performance of ARES-DIAT and ARES-DAT*: To verify that the proposed DIAT strategy can further improve security performance compared with the IAT strategy, we conduct the following experiment in this section.

By comparing the security performance of ARES-DIAT and ARES-IAT in Table VI, it can be found that the enhancement of DIAT is more effective than that of IAT. This verified that DIAT can provide more diverse stego-samples for the steganalytic discriminator and can further expand the steganographic feature space, which further improves the robustness of the steganalytic discriminator. Moreover, this enhancement tended to be higher as the payload increased. For example, it can be observed that the error rates of ARES-DIAT were 46.51%, 43.69%, 37.35%, and 34.06% with 0.1 to 0.4 bpp on the detection of Ye-Net. The corresponding error rates of ARES-IAT were 45.14%, 40.77%, 33.52%, and 26.88%. Consequently, the improvements were

1.37%, 2.99%, 3.83%, and 7.18%, respectively. We speculate that the increasing embedding payload allows the DIAT strategy to create more diverse adversarial samples. Thus, it pushed ARES-DIAT to build a broader awareness of the steganographic features.

C. Ablation Studies on Steganalytic Discriminator's Structures

In this section, we explored the security performance of the steganalytic discriminator when using different network structures. In detail, two variants of ARES-DIAT are involved. One adopted XuNet as the steganalytic discriminator, while the other utilized the SteFA structure. For simplicity, we denoted the above variants as ARES-Xu and ARES-SteFA. The specific experimental results are shown in Table VII. It can be observed that ARES-SteFA outperformed ARES-Xu significantly under all the embedding payloads. This phenomenon indicated that ARES-Xu could not handle the diversified stego-samples provided by DIAT. That might be due to the high diversity of the adversarial samples disturbing the feature extraction of the XuNet-based steganalytic discriminator. In this case, if the steganalytic discriminator did not have a well-designed structure to adaptively focus on more effective features, it would be lost in finding the right regions for defining the low embedding probabilities, which caused the inferior performance. This indicated that the improvements in DIAT were related to the network structures of the steganalytic discriminator. Further, we observed that the performance margins of the same steganalyzer between ARES-SteFA and ARES-Xu were not changed much among all payloads. For instance, on the detection of SRM, the performance margins were 7.08%, 8.05%, 8.12%, and 7.10%, respectively. However, the performance margins of CNN-based and conventional steganalytic discriminators were much larger at the same payload. For instance, at 0.1 bpp, the performance margins were 13.92%, 15.59%, 6.18%, 12.22%, and 7.08% on the detection of SRNet, Ye-Net, Yedroudj-Net, Zhu-Net, and SRM. This phenomenon further showed that the security performance by conducting DIAT is somehow irrelevant to the embedding payload, once the steganalytic discriminator is ill-structured. It also proved that the DIAT strategy needs to be accompanied by

TABLE VII
SECURITY PERFORMANCE (IN %) OF DIFFERENT STEGANOGRAPHIC APPROACHES AGAINST DIFFERENT STEGANALYZERS

Steganalyzer	Steganographic model	0.1 bpp	0.2 bpp	0.3 bpp	0.4 bpp
SRNet	ARES-SteFA	44.90 (↑ 13.92)	40.74 (↑ 12.09)	33.17 (↑ 13.90)	30.28 (↑ 13.26)
	ARES-Xu	30.98	28.65	19.27	17.02
Ye-Net	ARES-SteFA	46.51(↑ 15.59)	43.69(↑ 13.64)	37.35(↑ 12.67)	34.06(↑ 13.33)
	ARES-Xu	34.92	30.05	24.68	20.73
Yedroudj-Net	ARES-SteFA	45.03 (↑ 6.18)	39.77 (↑ 7.69)	35.76 (↑ 8.44)	32.57 (↑ 8.90)
	ARES-Xu	38.85	32.08	27.32	23.67
Zhu-Net	ARES-SteFA	41.04(↑ 12.22)	34.86(↑ 13.27)	29.44(↑ 11.96)	26.30(↑ 10.36)
	ARES-Xu	28.82	21.59	17.48	15.94
SRM	ARES-SteFA	46.25 (↑ 7.08)	42.06 (↑ 8.05)	41.92 (↑ 8.12)	38.89 (↑ 7.10)
	ARES-Xu	39.17	34.01	33.80	31.79

TABLE VIII
SECURITY PERFORMANCE (IN %) OF DIFFERENT STEGANOGRAPHIC APPROACHES AGAINST DIFFERENT STEGANALYZERS

Steganalyzer	Steganography	0.1 bpp	0.2 bpp	0.3 bpp	0.4 bpp
SRNet	ARES-DIAT	44.90	40.74	33.17	30.28
	ARES-IAT	43.79	37.46	27.15	24.14
	ARES-noAT	39.44	34.02	24.98	22.49
	UT-GAN	40.14	32.47	24.84	21.35
	S-UNIWARD	35.38	31.69	24.62	20.50
	WOW	32.18	28.86	22.25	19.35
	HILL	37.21	32.03	24.93	21.47
Ye-Net	ARES-DIAT	46.51	43.69	37.35	34.06
	ARES-IAT	45.14	40.77	33.52	26.88
	ARES-noAT	40.78	36.28	31.50	25.32
	UT-GAN	40.36	35.49	32.57	28.12
	S-UNIWARD	39.85	32.37	28.64	26.59
	WOW	38.98	31.29	27.52	25.16
	HILL	39.97	34.78	29.68	27.93
Yedroudj-Net	ARES-DIAT	45.03	39.77	35.76	32.57
	ARES-IAT	44.19	37.83	32.01	25.06
	ARES-noAT	39.90	35.62	30.15	24.39
	UT-GAN	40.22	34.19	31.28	26.16
	S-UNIWARD	39.36	29.02	27.47	26.11
	WOW	38.75	29.04	25.41	23.46
	HILL	39.88	33.58	28.57	26.44
Zhu-Net	ARES-DIAT	41.04	34.86	29.44	26.30
	ARES-IAT	39.96	32.39	25.95	21.32
	ARES-noAT	36.63	30.45	24.42	20.31
	UT-GAN	38.72	30.64	25.33	20.47
	S-UNIWARD	29.27	26.58	22.43	18.71
	WOW	27.82	25.43	21.66	17.85
	HILL	36.78	28.40	24.62	19.48
SRM	ARES-DIAT	46.25	42.06	41.92	38.89
	ARES-IAT	44.88	40.47	40.02	36.47
	ARES-noAT	43.25	39.29	38.58	36.34
	UT-GAN	43.41	39.58	38.37	36.43
	S-UNIWARD	42.36	38.78	36.31	35.26
	WOW	40.80	36.55	36.53	34.31
	HILL	42.95	39.21	38.28	36.77

The bold values indicate the best-performing steganography method.

a well-structured steganalytic discriminator, e.g., with the proposed SteFA structure.

D. Comparisons With Other Steganographic Approaches

In this section, we conducted experiments to compare the performance of the proposed approach with several popular steganographic approaches. From the results shown in Table VIII, it can be found that the security performances of both ARES-DIAT and ARES-IAT were higher than that of UT-GAN and S-UNIWARD with all payloads, which proved the proposed hypothesis in Section III-B quantificationally. In

addition, recall that ARES-DIAT was trained by *BOSSBase256* with 10,000 images, whereas UT-GAN utilized more. Despite this, ARES-DIAT consistently outperformed UT-GAN across all payloads. These results confirmed the phenomenon that the DIAT strategy provided more diverse steganographic features and the SteFA structure could capture these features more efficiently, thus enabling the network to decrease the number of training samples. Further, both ARES-DIAT and UT-GAN were evaluated using *ALASKA256* dataset to assess their generalizability. In this case, only if their steganalytic discriminators learn more precise decision boundaries during training, their generators would show better performance when evaluated. According to the experimental results, it can also be concluded that ARES-DIAT exhibited better generalizability than UT-GAN.

For qualitatively analyzing the learned implicit cost function of different approaches, we presented a visualization of the embedding probability maps of four randomly picked cover-images by several steganographic approaches, as illustrated in Fig. 5. We first found that the GAN-based steganographic methods, i.e., the proposed ARES (ARES-DIAT) and UT-GAN, could embed the message in the complex texture regions more subtle than that of the conventional methods, i.e., HILL and S-UNIWARD. It should also be noticed that the regions with larger probabilities have a higher degree of consistency. In addition, it can be found that the ARES appeared to be more finely sorted in the regions with moderate probabilities. In other words, some regions regarded as unsuitable for embedding in UT-GAN were assigned with higher probabilities in ARES. This phenomenon occurred because we increased the diversity of steganographic samples, which helped the steganalytic discriminator learn more effective steganalytic features, especially including some subtle features. This helped to make the output probability maps with more uniformly distributed values. Eventually, it could improve security.

E. Comparisons of ARES and UT-GAN by Efficient-Net

In this part, we focus on investigating the security performance of ARES and UT-GAN by employing a more advanced classifier, i.e., Efficient-Net [62] as the steganalyzer. Following the settings in Yousfi et al. [62], we employ Efficient-Net B0 initialized with the weights pre-trained on ImageNet, and replicate the grayscale channel thrice to accommodate the input requirements. Additionally, we insert an HPF into the first layer,

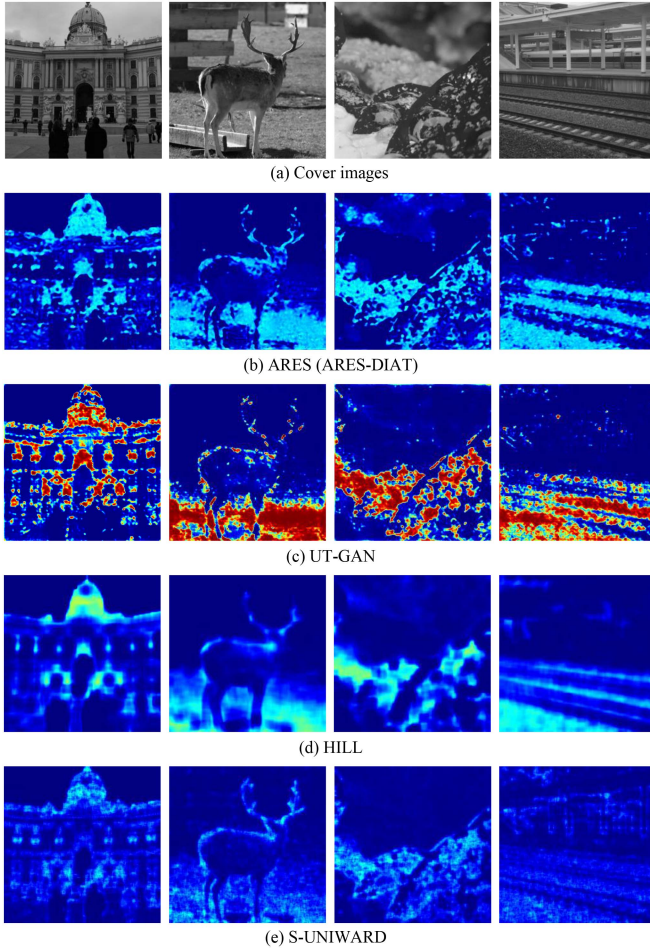


Fig. 5. Comparison of the embedding probability maps at 0.4 bpp. Best viewed in color. The bluer the color, the smaller the value. Conversely, the redder the color, the greater the value. The cover-images in the first and second columns are selected from BOSSBase, and the cover images in the third and last columns are selected from ALASKA.

TABLE IX
SECURITY PERFORMANCE (IN %) OF ARES AND UT-GAN AGAINST
EFFICIENT-NET, ON ALASKA256 DATASET

Steganography	0.1 bpp	0.2 bpp	0.3 bpp	0.4 bpp
ARES	45.32	41.49	31.27	30.03
UT-GAN	43.18	39.67	28.58	26.40

The bold values indicate the best-performing steganography method.

optimizing it for spatial steganalysis. As the results are presented in Table IX, ARES outperforms UT-GAN across all payloads, underscoring its efficacy.

V. CONCLUSION

In this paper, we observe that the robustness of the CNN-based steganalyzers is affected by the steganographic patterns and their structures. Motivated by the observations, we try to enhance the robustness of the steganalytic discriminator to improve the security of the GAN-based steganography. To this end, in the proposed ARES steganographic approach, we designed a DIAT strategy, which provides more diversified adversarial samples

for the steganalytic discriminator. Furthermore, we designed the SteFA structure of the steganalytic discriminator based on the channel-attention mechanism and covariance pooling strategy, which can extract more effective steganalytic features. Consequently, the DIAT strategy and the SteFA structure can improve the robustness of the steganalytic discriminator, significantly. Then, the steganalytic discriminator can build a more precise decision boundary, which contributes to learning a superior steganographic cost function.

In the future, considering utilizing the mutual information of perturbed pixels in adversarial attacks, we will design a more powerful AT strategy for the steganalytic discriminator to further improve the security of the GAN-based steganography.

REFERENCES

- [1] F. Li et al., "Research on covert communication channel based on modulation of common compressed speech codec," *Neural Comput. Appl.*, vol. 34, pp. 11507–11520, 2022.
- [2] J. Fridrich and T. Filler, "Practical methods for minimizing embedding impact in steganography," *Proc. SPIE*, vol. 6505, pp. 13–27, 2007.
- [3] T. Pevný, T. Filler, and P. Bas, "Using high-dimensional image models to perform highly undetectable steganography," in *Int. Workshop Inf. Hiding*, 2010, pp. 161–177.
- [4] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," *IEEE Trans. Inf. Forensics Secur.*, vol. 6, no. 3, pp. 920–935, Sep. 2011.
- [5] W. Li, W. Zhang, L. Li, H. Zhou, and N. Yu, "Designing near-optimal steganographic codes in practice based on polar codes," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 3948–3962, Jul. 2020.
- [6] V. Holub and J. Fridrich, "Designing steganographic distortion using directional filters," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2012, pp. 234–239.
- [7] V. Holub, J. Fridrich, and T. Denemark, "Universal distortion function for steganography in an arbitrary domain," *EURASIP J. Inf. Secur.*, vol. 2014, no. 1, pp. 1–13, 2014.
- [8] B. Li, M. Wang, J. Huang, and X. Li, "A new cost function for spatial image steganography," in *Proc. IEEE Int. Conf. Image Process.*, 2014, pp. 4206–4210.
- [9] I. Goodfellow et al., "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, vol. 27, pp. 2672–2680.
- [10] W. Tang, S. Tan, B. Li, and J. Huang, "Automatic steganographic distortion learning using a generative adversarial network," *IEEE Signal Process. Lett.*, vol. 24, no. 10, pp. 1547–1551, Oct. 2017.
- [11] J. Yang, D. Ruan, J. Huang, X. Kang, and Y.-Q. Shi, "An embedding cost learning framework using GAN," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 839–851, 2020.
- [12] H. Wu, F. Li, X. Zhang, and K. Wu, "GAN-based steganography with the concatenation of multiple feature maps," in *Int. Workshop Digit. Watermarking*, 2019, pp. 3–17.
- [13] X. Mo, S. Tan, B. Li, and J. Huang, "MCTSteg: A Monte Carlo tree search-based reinforcement learning framework for universal non-additive steganography," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 4306–4320, 2021.
- [14] W. Tang, B. Li, M. Barni, J. Li, and J. Huang, "An automatic cost learning framework for image steganography using deep reinforcement learning," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 952–967, 2020.
- [15] J. Zhao and S. Wang, "A stable GAN for image steganography with multi-order feature fusion," *Neural Comput. Appl.*, vol. 34, no. 18, pp. 16073–16088, 2022.
- [16] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 274–283.
- [17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Representations*, 2018.
- [18] C. Xiao, P. Zhong, and C. Zheng, "Enhancing adversarial defense by k-winners-take-all," in *Proc. Int. Conf. Learn. Representations*, 2020.
- [19] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

- [20] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 3, pp. 868–882, Jun. 2012.
- [21] T. Denemark, V. Sedighi, V. Holub, R. Cogranne, and J. Fridrich, "Selection-channel-aware rich model for steganalysis of digital images," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2014, pp. 48–53.
- [22] G. Xu, H.-Z. Wu, and Y.-Q. Shi, "Structural design of convolutional neural networks for steganalysis," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 708–712, May 2016.
- [23] J. Ye, J. Ni, and Y. Yi, "Deep learning hierarchical representations for image steganalysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 11, pp. 2545–2557, Nov. 2017.
- [24] M. Yedroudj, F. Comby, and M. Chaumont, "Yedroudj-net: An efficient CNN for spatial steganalysis," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2018, pp. 2092–2096.
- [25] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 5, pp. 1181–1193, May 2019.
- [26] R. Zhang, F. Zhu, J. Liu, and G. Liu, "Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1138–1150, 2019.
- [27] Y. Yousfi, J. Butora, and J. Fridrich, "CNN steganalyzers leverage local embedding artifacts," in *Proc. IEEE Int. Workshop Inf. Forensics Secur.*, 2021, pp. 1–6.
- [28] J. Zhang, K. Chen, C. Qin, W. Zhang, and N.-H. Yu, "Distribution-preserving-based automatic data augmentation for deep image steganalysis," *IEEE Trans. Multimedia*, vol. 24, pp. 4538–4550, 2022.
- [29] D. Volkhonskiy, I. Nazarov, and E. Burnaev, "Steganographic generative adversarial networks," *Proc. SPIE*, vol. 11433, pp. 991–1005, 2020.
- [30] Y. Wang, K. Niu, and X. Yang, "Information hiding scheme based on generative adversarial network," *J. Comput. Appl.*, vol. 38, no. 10, 2018, Art. no. 2923.
- [31] Q. Cui et al., "Image steganography based on foreground object generation by generative adversarial networks in mobile edge computing with Internet of Things," *IEEE Access*, vol. 7, pp. 90815–90824, 2019.
- [32] R. Meng, Q. Cui, Z. Zhou, Z. Fu, and X. Sun, "A steganography algorithm based on CycleGAN for covert communication in the Internet of Things," *IEEE Access*, vol. 7, pp. 90574–90584, 2019.
- [33] R. Meng et al., "High-capacity steganography using object addition-based cover enhancement for secure communication in networks," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 2, pp. 848–862, Mar./Apr. 2022.
- [34] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Representations*, 2015.
- [35] C. Szegedy et al., "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Representations*, 2014.
- [36] Y. Zhang et al., "Adversarial examples against deep neural network based steganalysis," in *Proc. 6th ACM Workshop Inf. Hiding Multimedia Secur.*, 2018, pp. 67–72.
- [37] W. Tang, B. Li, S. Tan, M. Barni, and J. Huang, "CNN-based adversarial embedding for image steganography," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 8, pp. 2074–2087, Aug. 2019.
- [38] B. Li, M. Wang, X. Li, S. Tan, and J. Huang, "A strategy of clustering modification directions in spatial image steganography," *IEEE Trans. Inf. Forensics Secur.*, vol. 10, no. 9, pp. 1905–1917, Sep. 2015.
- [39] X. Qin, S. Tan, W. Tang, B. Li, and J. Huang, "Image steganography based on iterative adversarial perturbations onto a synchronized-directions sub-image," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2021, pp. 2705–2709.
- [40] S. Ma, X. Zhao, and Y. Liu, "Adaptive spatial steganography based on adversarial examples," *Multimedia Tools Appl.*, vol. 78, no. 22, pp. 32503–32522, 2019.
- [41] S. Bernard, P. Bas, J. Klein, and T. Pevny, "Explicit optimization of min max steganographic game," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 812–823, 2021.
- [42] V. Sedighi, R. Cogranne, and J. Fridrich, "Content-adaptive steganography by minimizing statistical detectability," *IEEE Trans. Inf. Forensics Secur.*, vol. 11, no. 2, pp. 221–234, Feb. 2016.
- [43] S. Wu, S. Zhong, and Y. Liu, "A novel convolutional neural network for image steganalysis with shared normalization," *IEEE Trans. Multimedia*, vol. 22, no. 1, pp. 256–270, Jan. 2020.
- [44] A. Agarwal, M. Vatsa, R. Singh, and N. K. Ratha, "Noise is inside me! generating adversarial perturbations with noise derived from natural filters," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 774–775.
- [45] Y. Zhou, X. Hu, J. Han, L. Wang, and S. Duan, "High frequency patterns play a key role in the generation of adversarial examples," *Neurocomputing*, vol. 459, pp. 131–141, 2021.
- [46] C. Zhang, P. Benz, A. Karjauv, and I. S. Kweon, "Universal adversarial perturbations through the lens of deep steganography: Towards a fourier perspective," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, pp. 3296–3304, 2021.
- [47] A. Fawzi, S.-M. Moosavi-Dezfooli, and P. Frossard, "Robustness of classifiers: From adversarial to random noise," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, vol. 29, pp. 131–141.
- [48] J. Y. Franceschi, A. Fawzi, and O. Fawzi, "Robustness of classifiers to uniform lp and gaussian noise supplementary material," in *Int. Conf. Artif. Intell. Statist.*, 2018, pp. 1280–1288.
- [49] T. Zheng, C. Chen, and K. Ren, "Distributionally adversarial attack," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 2253–2260.
- [50] Y. Dong, Z. Deng, T. Pang, J. Zhu, and H. Su, "Adversarial distributional training for robust deep learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, vol. 33, pp. 8270–8283.
- [51] H. Cheng et al., "Mixture of robust experts (more): A flexible defense against multiple perturbations," 2021, *arXiv: 2104.10586*.
- [52] P. Li, J. Xie, Q. Wang, and Z. Gao, "Towards faster training of global covariance pooling networks by iterative matrix square root normalization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 947–955.
- [53] X. Deng, B. Chen, W. Luo, and D. Luo, "Fast and effective global covariance pooling network for image steganalysis," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.*, 2019, pp. 230–234.
- [54] M. Yedroudj, M. Chaumont, F. Comby, A. O. Amara, and P. Bas, "Pixels-off: Data-augmentation complementary solution for deep-learning steganalysis," in *Proc. 2020 ACM Workshop Inf. Hiding Multimedia Secur.*, 2020, pp. 39–48.
- [55] D. Ulyanov, A. Vedaldi, and V. Lempitsky, "Instance normalization: The missing ingredient for fast stylization," 2016, *arXiv:1607.08022*.
- [56] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Int. Conf. Mach. Learn.*, 2015, pp. 448–456.
- [57] P. Bas, T. Filler, and T. Pevný, "Break our steganographic system": The ins and outs of organizing BOSS," in *Proc. Int. Workshop Inf. Hiding*, 2011, pp. 59–70.
- [58] R. Cogranne, Q. Giboulot, and P. Bas, "The ALASKA steganalysis challenge: A first step towards steganalysis," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.*, 2019, pp. 125–137.
- [59] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Int. Conf. Med. Image Comput. Comput.-Assist. Interv.*, 2015, pp. 234–241.
- [60] R. Zhang, F. Zhu, J. Liu, and G. Liu, "Depth-wise separable convolutions and multi-level pooling for an efficient spatial CNN-based steganalysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1138–1150, 2020.
- [61] J. Kodovsky, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," *IEEE Trans. Inf. Forensics Secur.*, vol. 7, no. 2, pp. 432–444, Apr. 2012.
- [62] Y. Yousfi, J. Butora, J. Fridrich, and C. F. Tsang, "Improving efficientnet for JPEG steganalysis," in *Proc. ACM Workshop Inf. Hiding Multimedia Secur.*, 2021, pp. 149–157.