Feature Fusion Based Adversarial Example Detection against Second-Round Adversarial Attacks

Chuan Qin, Yuefeng Chen, Kejiang Chen, Xiaoyi Dong, Weiming Zhang, Xiaofeng Mao, Yuan He, Nenghai Yu

***Abstract*—Convolutional Neural Networks (CNNs) achieve remarkable performances in various areas. However, adversarial examples threaten their security. They are designed to mislead CNNs to output incorrect results. Many methods are proposed to detect adversarial examples. Unfortunately, most detection-based defense methods are vulnerable to second-round adversarial attacks, which can simultaneously deceive the base model and the detector. To resist such second-round adversarial attacks, handcrafted steganalysis features are introduced to detect adversarial examples, while such a method receives low accuracy at detecting sparse perturbations. In this paper, we propose to combine handcrafted features with deep features via a fusion scheme to increase the detection accuracy and defend against second-round adversarial attacks. To avoid deep features being overwhelmed by high-dimensional handcrafted features, we propose an expansion-then-reduction process to compress the dimensionality of handcrafted features. Experimental results show that the proposed model outperforms the state-of-the-art adversarial example detection methods and remains robust under various second-round adversarial attacks.**

***Impact Statement*—Currently, deep learning systems have important applications and outstanding performance in various areas, e.g., recognizing traffic objects in autopilot systems, classifying images for online search engines. But, their outputs will be wrong if inputs are added imperceptible malicious perturbations, called adversarial perturbations. For instance, the deep learning autopilot systems will ignore an adversarially perturbed "STOP" sign and keep moving. The detector proposed in this paper identifies various types of adversarial perturbations with averaging more than 95% accuracy. The adaptive perturbations against the detector only achieve about 10% success rate. With such detection accuracy and robustness, the detector can effectively protect deep learning systems from being attacked. For instance, it can help autopilot systems identify malicious objects. For online search engines, it can help to detect sensitive images that are hidden by adversarial perturbations.**

***Index Terms*—Adversarial examples, Detection, Steganalysis, Information hiding, Second-round adversarial attacks**

I. Introduction

RECENTLY, deep neural networks make breakthroughs in various fields, such as image classification [1], [2], [3], object detection [4], [5], [6]. However, its applications are threatened by adversarial examples [7]. These images are intentionally perturbed to induce the target CNN to output incorrect results. The perturbations are often small and visually invisible to humans.

Most of the adversarial attack methods reduce the perturbations by minimizing the l_p distance between adversarial examples and original images. Plenty of l_2 and l_∞-based methods [8], [9], [10], [11], [12], [13], [14] have been proposed since they are believed to own lower visibility and easy to optimize. Even though the amplitude of perturbations is large, l_0-based methods only require modifying a few pixels. There are also several works focusing on it [15], [16], [17], [12].

As adversarial examples become the major security concern of CNNs, some methods were proposed to defend them by enhancing the robustness of the target model, such as adversarial training [18] and gradient masking [19], [20], or pre-processing the input images [21], [22], [23], [24]. They are called robustness-based defenses. While these methods improved the robustness against adversarial examples, the classification accuracy on clean images is reduced. Besides, some robustness-based defenses require modifying target models. For this reason, the research community turned to detection-based defenses recently.

Some detection-based methods focused on anomalies of the features from hidden layers, such as via the use of PCA analysis [25], [26] or LID (Local Intrinsic Dimensionality) [27]. Some other methods [28], [29], [30], [31], [32] detected adversarial examples by analyzing the probabilistic outputs of the target CNN. Grosse et al., Gong et al. and Metzen et al. [33], [34], [35] proposed to detect adversarial examples by constructing another CNN detector. Unfortunately, most of the aforementioned detecting methods were proved to only work on datasets with small image size or be vulnerable to the second-round adversarial attacks [36], which can fool the base model and the detector simultaneously by merging them as a $N+1$ classification model.

Inspired by the explanation of Ian Goodfellow [8], *adversarial examples can be viewed as a sort of "accidental steganography"*[37], [38], [39], [40], Liu et al. [41] have proposed to detect adversarial examples from the viewpoint of steganalysis[42], [43], [44], which is the countermeasure of steganography. The same as adversarial attacks, steganography minimizes the quantity and amplitude of perturbations.

The corresponding authors: Weiming Zhang and Kejiang Chen. E-mail: {zhangwm, chenkj}@ustc.edu.cn

C. Qin, K. Chen, X. Dong, W. Zhang, N. Yu are with School of Information Science and Technology, University of Science and Technology of China, Key Laboratory of Electromagnetic Space Information, Chinese Academy of Sciences.

Y. Chen, X. Mao, Y. He are with Alibaba Group.

This work was supported in part by the Natural Science Foundation of China under Grant 62002334, 62072421, and 62121002, Anhui Science Foundation of China under Grant 2008085QF296, and by Anhui Initiative in Quantum Information Technologies under Grant AHY150400.

Furthermore, the perturbations are forced to happen in the regions where the image textures are complex to further conceal the secret messages. To counter steganography, steganalysis calculates the residual maps of various high-pass filters to capture the distortions caused by artificial operations on image textures. Following the logic of steganalysis, Liu et al. adopted SRM (Spatial Rich Model) [43] to detect adversarial examples. Their method outperforms the previous ones on ImageNet [45]. Moreover, due to various underivable operations in SRM feature extraction, Liu et al.'s [41] method is robust against second-round adversarial attacks.

However, to counter the "curse of dimensionality" caused by profuse filters, SRM clamps the elements' value in residual maps. It severely limits its perception of large perturbations. On the other hand, the effectiveness of various filters decreases in detecting sparse perturbations. These factors inevitably make SRM receive relatively low accuracies on detecting l_0-based adversarial examples. Hence, we turn to the recent advances of CNN steganalysis [46], [47], [48], which suggest that CNN detectors can outperform SRM, in adversarial example detection. Meanwhile, the last layer features of CNN models are often of low dimension. However, under the second-round adversarial attacks [36], CNN detectors provide little robustness. Fortunately, the advantages of SRM detectors and CNN detectors are complementary, i.e., the former is robust against second-round adversarial attacks while the latter own higher detection accuracies. Hence, in this paper, a fusion model is proposed to combine these complementary advantages.

In this paper, we propose a detection-based defense. It utilizes an ensemble classifier-based fusion scheme to combine deep features and optimized SRM features. It enables them to provide their predictions to the final results. They become the complement of each other. Specifically, the deep features complement the information loss of optimized SRM features caused by residual truncation, and optimized SRM features would make the correct predictions when the deep features are bypassed. Since SRM is a high-dimensional handcrafted feature, it will overwhelm deep features if they are direct concatenated. Hence, we propose an expansion-then-reduction process to compress SRM's dimensionality and improve its detection ability. Specifically, we first expand dimensions by adding channel correlations and enlarging truncation threshold on residual maps. Then we adopt Fisher score [49], [50] to scissor redundant features from the expanded SRM. The final optimized SRM is with only 2048 dimension, which is less than $1/10$ of the original. Extensive experiments show that the proposed detector outperforms the previous works in detecting adversarial examples in datasets with either small or large image sizes. Furthermore, the proposed detector remains robust under various second-round attacks, including Carlini et al.'s method [36], [12], BPDA [51] and adaptive query-based attack [14].

II. Related Works

A. Spatial Rich Model

SRM [43] was utilized to detect adversarial examples in Liu et al.'s work [41]. The diversity of residual computing enables SRM to capture the distortions caused by invisible noises, such as adversarial perturbations.

Residuals Computing. The typical residual form of the linear predictor can be represented as:

$$R_{i,j} = \widehat{X}_{i,j}(\mathcal{N}_{i,j}) - cX_{i,j}, \tag{1}$$

where $R_{i,j} \in \mathbb{R}^{W\times H}$, and $c \in \mathbb{N}$ is the residual order. $\mathcal{N}_{i,j}$ represents local neighborhoods of pixel $X_{i,j}$, but $X_{i,j} \notin \mathcal{N}_{i,j}$. $\widehat{X}_{i,j}$ is a predictor defined on $\mathcal{N}_{i,j}$.

Residual Quantization and Truncation. To control the dimension and reduce sparsity, the float residual maps are truncated and quantized as the following formula:

$$R_{i,j} \leftarrow \mathrm{trunc}_T(\mathrm{round}(\frac{R_{i,j}}{q})), \tag{2}$$

where q is the quantization step, and trunc_T limits the quantized residual values to $[-T, T]$. In SRM, the default set of q is $\{1, 1.5, 2\}$ and the default value of T is set as 2.

Co-occurrence Matrices Computing. SRM computes co-occurrence matrices from the truncated and quantized residual maps as the final features. Co-occurrence matrix indexed with $\mathbf{d} = (d_1, d_2, d_3, d_4) \in \mathcal{T}_4 \triangleq \{-T, \ldots, T\}^4$ is denoted as $\mathbf{C_d}$. Given residual $\mathbf{R} = (R_{i,j})$ and four neighbouring residual samples with values d_1, d_2, d_3, d_4, the $\mathbf{d}$-th element of co-occurrence as follows.

$$\begin{aligned}\mathbf{C_d} = \frac{1}{Z}\Big|\{(R_{i,j}, R_{i,j+1}, R_{i,j+2}, R_{i,j+3})| \\ R_{i,j+k-1} = d_k, k = 1, 2, 3, 4\}\Big|,\end{aligned} \tag{3}$$

where Z is the normalization factor that guarantees $\sum_{\mathbf{d}\in\mathcal{T}_4} \mathbf{C_d} = 1$.

B. Second-Round Adversarial attacks

To evaluate the robustness of a detector, Carlini et al. [36] proposed second-round adversarial attacks. It generates adversarial examples that will simultaneously fool the base model F_{base} and the detector D. Specifically, with all the other parts kept the same, a function which combines the logits of F_{base} and D is formulated to replace the logits of F_{base} in C&W attack:

$$G(\mathbf{X})_i = \begin{cases} Z_{F_{base}}(\mathbf{X})_i, & i \le N \\ (Z_D(\mathbf{X}) + 1) \cdot \max_j Z_{F_{base}}(\mathbf{X})_j, & i = N + 1 \end{cases} \tag{4}$$

where $Z(\cdot)$ is the logits of CNN, $\mathbf{X}$ is the input image and $i \in \{i|\mathbb{N}, 1 \le i \le N + 1\}$ represents the class index. It can be noticed that $G(\cdot)$ is like a classifier on $N+1$ classes. It has two important properties: if $Z_D(\mathbf{X}) > 0$, i.e. the detector predicts $\mathbf{X}$ as adversarial example, the label output of $G(\mathbf{X})$ is $N + 1$. If $Z_D(\mathbf{X}) < 0$, i.e. the detector predicts $\mathbf{X}$ as clean image, the label output of $G(\mathbf{X})$ is $\arg\max_i (Z_F(\mathbf{X})_i)$. Hence, when the output label of $\mathbf{X}$ is neither the correct label nor $N + 1$, the second-round adversarial example is successfully generated.

III. The Proposed Method

A. Motivation

As briefly introduced in Section I, SRM steganalysis features are powerful in detecting small perturbations, including most

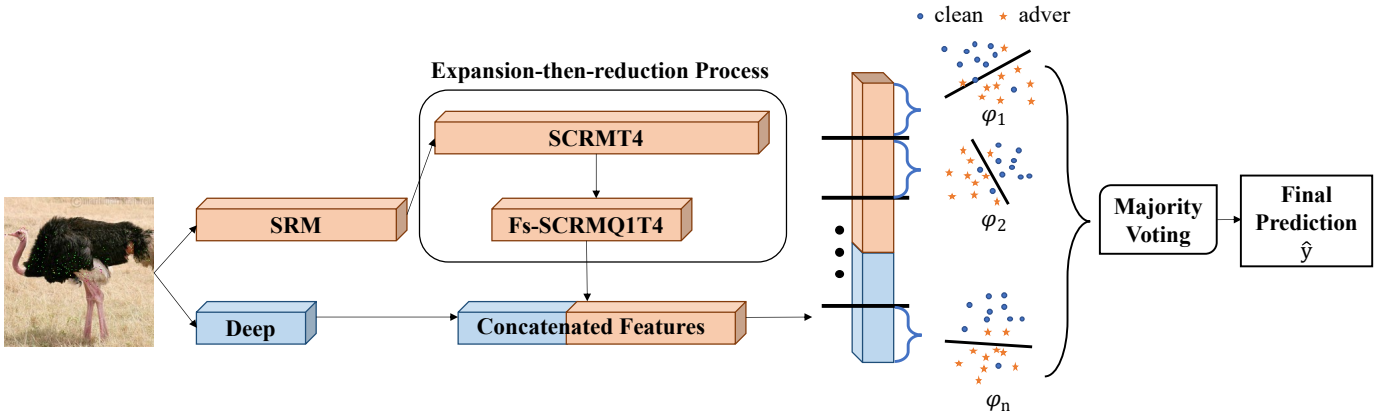


Fig. 1. The proposed fusion model. The deep features and optimized SRM features (Fs-SCRMQ1T4 in the figure) are concatenated. Each of the two kinds of features is fed separately to an ensemble classifier. Top-$k\% \times N_h$ and top-$(1-k\%) \times N_d$ base classifiers φ_i are considered in the majority voting for final predictions.

kinds of adversarial perturbations. However, the perception of SRM features on the perturbations of large amplitude is quite limited. Since steganographic modifications are always with ± 1. To detect such small perturbations, SRM assembles many diverse high-pass filters, which enables it to capture small artificial perturbations in various kinds of texture patterns. However, a large number of filters also produce high dimension. To compress dimension, SRM clamps the element value of residual maps to a small range. This operation makes little impact on detecting steganography due to its small amplitude. For l_∞-based and most l_2-based perturbations, SRM can still detect them accurately because the number of modifications is still quite large. However, for l_0-based adversarial perturbations, which are designed to perturb as few pixels as it can be, the high diversity of filters only produces redundancy. On the other hand, SRM is unable to perceive large perturbations. It is not surprising that relatively low detection accuracies are witnessed.

Since both steganographic perturbations and adversarial perturbations can be regarded as high-frequency noises, the ideas of designing CNN steganalysis models can be applied to detect adversarial perturbations. Furthermore, CNNs are free from the limitations of truncation. Higher detection accuracies can be obtained. However, as Carlini et al. [36] indicated, CNN detectors are vulnerable to the second-round adversarial attack.

Though truncation hinders SRM's perception of large perturbations, suchlike underivable operations in SRM provide the robustness against second-round adversarial attacks. On the other hand, the deep features from the penultimate layer of CNN-based steganalysis models could perceive large and sparse perturbations. Hence, in this paper, we propose to fuse the optimized SRM features and deep features to improve the detecting abilities against sparse perturbations while maintaining robustness against second-round adversarial attacks.

B. The Fusion Model

To construct a fusion model that obtains robustness and high detection ability, both optimized SRM features and deep features should contribute to the final predictions. In this paper, we propose to fuse them by random combination and a majority voting structure. As shown in Figure 1, the optimized SRM features (Fs-SCRMQ1T4) and deep features are fed to several base classifiers, then the predictions are considered comprehensively through a majority voting structure. Note that the base classifiers φ_i are either trained with optimized SRM features or deep features. Specifically, the number of base classifiers and the dimensions of sub-spaces of each base classifiers are optimized via the process of the ensemble classifier in steganalysis [52]. We adopt the top-$k\% \times N_h (k \in [0, 100])$ accurate base classifiers from optimized SRM feature space and top-$(1-k\%) \times N_d$ accurate base classifiers from deep feature space to construct the voting panel, where N_h and N_d represent the number of base classifiers trained with the optimized SRM features and the deep features.

The fusion structure has two key properties: 1) the random combination and voting mechanism provide extra robustness against second-round adversarial attacks. 2) It combines the optimized SRM features and deep features through base classifiers and majority voting, which can improve the detection ability of the scheme. Since both optimized SRM and deep features are effective independently, they can be utilized when concatenated through proper classifiers. In the proposed scheme, we adopt FLD (Fisher Linear Discriminator) as the base classifier.

When the deep features are bypassed, the optimized SRM features are the ones that detect the adversarial perturbations. Hence, it is important to improve the detection ability of them. Moreover, typical handcrafted features such as SRM [43] are of high dimension. Direct concatenation will make deep features overwhelmed and introduce excessive computations. Hence, in this paper, we propose an expansion-then-reduction process to improve the detection ability of SRM features also significantly reduce their dimensions.

C. Dimension Expansion and Reduction

Adding Cross-Channel Features. SRM is designed for grayscale images [43]. Liu et al. [41] consider the color images as three independent grayscale images without utilizing the dependencies across color channels. To enhance the ability to

capture the cross-channel distortions, CRM [53] (Color Rich Model) is added to the SRM features in this work.

For three channels of a true-color image $\mathbf{X}$, let $\mathbf{R}^{(r)} = (R_{i,j}^{(r)})$, $\mathbf{R}^{(g)} = (R_{i,j}^{(g)})$ and $\mathbf{R}^{(b)} = (R_{i,j}^{(b)})$ be the residual maps of three color channels, which are truncated and quantized as Eq. 2. Similar to Eq. 3, the cross-channel co-occurrence matrices are formed from the triplets $(R_{i,j}^{(r)}, R_{i,j}^{(g)}, R_{i,j}^{(b)})$ as follows:

$$\begin{aligned}\mathbf{C_d} =& \frac{1}{Z}\Big|\{(R_{i,j}^{(r)}, R_{i,j}^{(g)}, R_{i,j}^{(b)})| \\ & R_{i,j}^{(r)} = d_1, R_{i,j}^{(g)} = d_2, R_{i,j}^{(b)} = d_3\}\Big|,\end{aligned} \tag{5}$$

where $\mathbf{d} = (d_1, d_2, d_3)$, and Z guarantees $\sum_{\mathbf{d}\in\mathcal{T}_3} \mathbf{C_d} = 1$. To reduce the feature redundancy, the residuals of CRM are quantized with single quantization step $q = 1$. Lastly, the same as the original SRM, CRM features are symmetrized. The final 5404-D features are obtained.

Expanding the Range of Residual Maps. Steganographic perturbations are minimal, which are always ± 1 [54]. It allows SRM to truncate the residual maps with a small threshold value $T = 2$ while not damaging the detection accuracy. However, adversarial perturbations often own larger amplitudes. The truncation would hinder the SRM from capturing the distortion caused by large adversarial perturbations on image textures.

Based on the observation above, we amplify the truncation threshold T to 4. Since each co-occurrence matrix has $(2T+1)^4$ elements, further amplification would consume excessive computation and storage resources in the feature extraction process. Since we add CRM features and modify the truncation value T, the fully expanded features are denoted as SCRMT4, which is 389295-D.

Dimension Reduction. We take two steps to reduce the dimension: 1) imposing a single quantization step, 2) selecting features with Fisher scores.

SRM is the combination of features with three quantization steps $\{1, 1.5, 2\}$. Inspired by CRM, we take a single quantization step with $q = 1$ to reduce the dimension to 135169-D. We denote these 135169-D features as SCRMq1T4.

However, 135169-D is still too large. To further reduce dimension, we utilize Fisher score to rank features and keep the first D_{opt} ones. Given the clean image set $\mathcal{C}$ and the corresponding adversarial example set $\mathcal{A}$, we denote the feature matrices of all the images in the datasets as $\mathbf{F}_c$ and $\mathbf{F}_a$. For the d-th feature $\mathbf{F}_c^d$ and $\mathbf{F}_a^d$, the Fisher score $Fs(d)$ is:

$$Fs(d) = \frac{\left(\mu_c^d - \mu_a^d\right)^2}{\left(\sigma_c^d\right)^2 + \left(\sigma_a^d\right)^2}, \tag{6}$$

where μ_c^d and μ_a^d represent the mean values of d-th feature, and σ_c^d and σ_a^d represent the standard deviations.

Similar to the Fisher score of single feature, Lu et al. [50] proposed the Fisher score of feature matrices:

$$Fs = \frac{\Omega\left(\boldsymbol{\mu}_c, \boldsymbol{\mu}_a\right)^2}{\frac{1}{N}\sum_{n=1}^{N}\Omega\left(\mathbf{F}_c^{(n)}, \boldsymbol{\mu}_c\right)^2 + \frac{1}{N}\sum_{n=1}^{N}\Omega\left(\mathbf{F}_a^{(n)}, \boldsymbol{\mu}_a\right)^2}, \tag{7}$$

where $\Omega(\cdot)$ represents the Euclidean distance between two vectors, and $\boldsymbol{\mu}_c$, $\boldsymbol{\mu}_a$ represent the vectors of mean values of all the features, and $\mathbf{F}_c^{(n)}$, $\mathbf{F}_a^{(n)}$ represent the feature vectors of the n-th sample from $\mathcal{C}$ and $\mathcal{A}$.

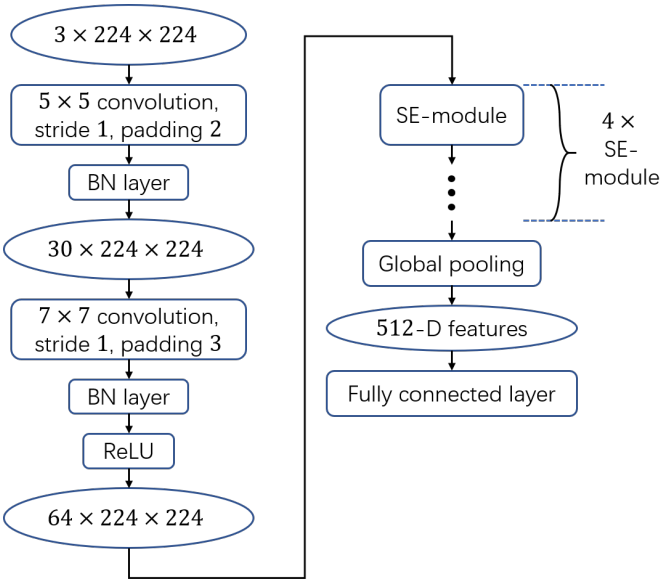


Fig. 2. The proposed CNN detector. We delete the max-pooling layer after the second convolutional layer and reduce the stride to 1 in the early stages.

Based on the Fisher score of single-dimension features and feature matrices, the best separability will be obtained if the feature with the highest Fisher score is always selected until the Fisher score of feature matrices reaches its maximum. Based on this logic, the feature selection process is presented as follows:

1) For each single-dimension feature pair $\mathbf{F}_c^d$ and $\mathbf{F}_a^d$ ($d \in [1, D]$), the Fisher score is calculated using Eq. 6.
2) All the features are sorted in descending order of Fisher score. The rearranged feature matrices are $\overline{\mathbf{F}}_c$ and $\overline{\mathbf{F}}_a$.
3) We calculate the Fisher score of the first D'-dimension features using Eq. 7.
4) The maximum Fisher score is obtained with the first D_{opt}-dimension features.

D. CNN Detector Design

Previous CNN steganalysis models suggested that pooling operations are like average filtering [47], which will erase the subtle perturbations, such as steganographic modifications and adversarial perturbations. A similar effect is also brought by the strides which are larger than 1. Moreover, the design of CNN steganalysis models fell behind the most progressive computer vision areas. Stacking unpooled modules in early stages [47] creates CNN models with a large number of parameters. Hence, we cancel the pooling operations and reduce the stride to 1 in the early stage of typical CNN structures. The backbone of the model we adopt is SE-ResNet18 [2], [55]. The structure of the CNN detector is shown in Figure 2.

IV. Experiments

A. Setup

1) Datasets and base models: Two widely studied datasets ImageNet [45] and CIFAR-10 [56] are utilized for performance evaluation in this paper. VGG16 [3] and ResNet18

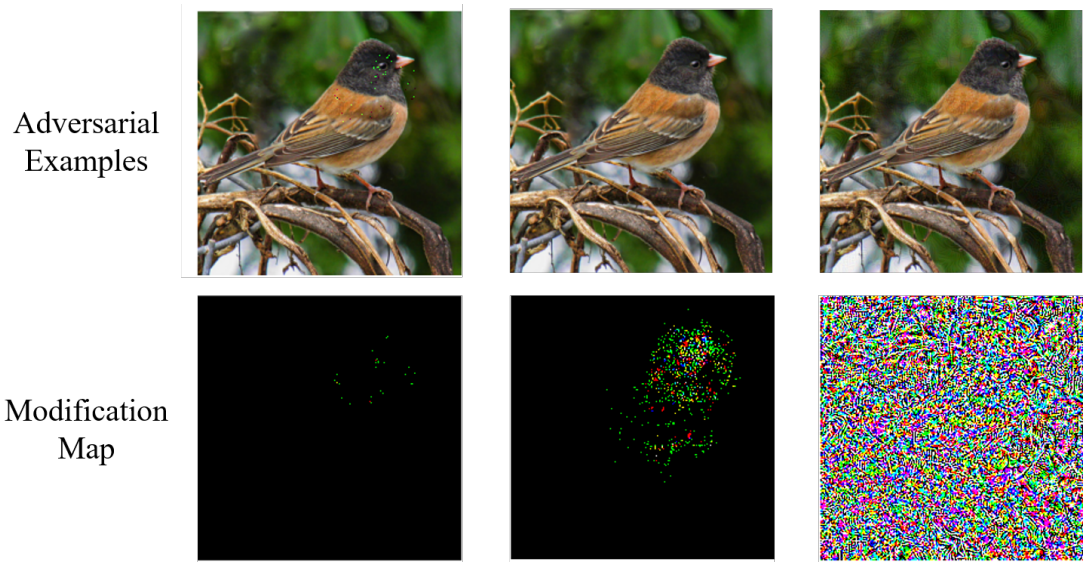


Fig. 3. The adversarial examples and modification maps of l_0, l_2 and l_∞-based perturbations. We select SparseFool, DDN and PGD ($\epsilon = 2$) as instance.

(or ResNet34) [2] are adopted as the base model. We adopt ResNet18 in CIFAR-10 and ResNet34 in ImageNet. For each of the two datasets, the target model is pretrained on the designed training set. The designed test datasets are used to train and test the detector. For ImageNet, $20,000$ images from the testing set are randomly selected and evenly divided into two disjoint sets. One is used for training the detector, while the other is used for performance evaluation. For CIFAR-10, from the testing set of the target model, which contains $10,000$ images, $1,000$ images are selected for performance evaluation while the other is utilized for training the detector.

2) Adversarial examples: The proposed fusion detector relays differentiate textures of natural images and adversarial examples to classify them, thus various l_p constrained adversarial examples are adopted to evaluate the detection ability of the proposed detector. For l_0-based methods, we select CornerSearch [17], C&W-l_0 [12] and SparseFool [16]. Since the generation of CornerSearch and C&W-l_0 on ImageNet is extremely time-consuming, they are only used on CIFAR-10. For l_2-based methods, we select DeepFool [11] C&W-l_2 [12] and DDN [13]. For l_∞-based methods, we select PGD [10] and BIM [9]. Carlini et al. [36] proposed a method against such type of detectors as the ones of Grosse et al. [33], Gong et al. [34] or Metzen et al. [35] and the proposed model. It has been proved to be effective against the previous detectors. Hence we adopt it to evaluate the robustness of the proposed model against second-round attacks, which is also called adaptive attacks.

3) Training process and hyper-parameters: The training of the fusion detector consists of two stages. In the first one, we train the CNN detector using Adamax optimizer [57], which utilized later for deep feature extraction. On ImageNet, the batch size is set as 32. We train the CNN detector with learning rate 1e-3, 1e-4 and 1e-5 for $200,000$, $50,000$ and $50,000$ iterations. On CIFAR-10, the batch size is set as 256. The CNN detector is trained with learning rate 1e-3, 1e-4 and 1e-5 for $15,000$, $5,000$ and $5,000$ iterations respectively. In the second stage, the weights of the CNN detector is fixed. The outputs of the penultimate layer are taken as the deep features of the fusion detector. Two ensemble classifiers [52] are trained independently with the optimized SRM features and deep features. The training hyper-parameters are set as the default [52]. Each FLD (Fisher Linear Discriminant) is trained with a random subset of the complete feature cluster. The dimension of the subsets and the number of FLD (base learners) are optimized via cross validation. Lastly, since adversarial example detection is a binary classification task, the fusion detector is trained with natural image and the corresponding adversarial example pairs.

B. Optimal Dimensionality of Optimized SRM Features

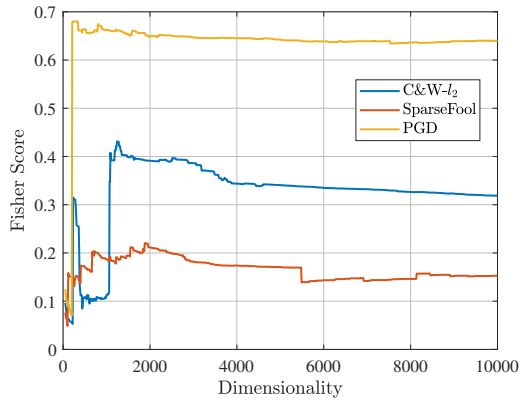
The dimension optimization of optimized SRM features is shown in this section. Our goal is to minimize the dimension while keeping the detection ability.

The difficulty of optimizing the dimension stems from the variety of adversarial perturbations. As shown in Figure 3, l_0-based adversarial attacks tend to modify a few pixels with large modification values. l_∞-based adversarial attacks often modify massive pixels with small values. l_2-based adversarial perturbations often maintain a better balance between the number of modified pixels and the modification values. Meanwhile, the adversarial perturbations with the same distortion measurement share similar modification patterns, so we adopt PGD ($\epsilon = 8$), C&W-l_2 and SparseFool as representations to optimize the dimension.

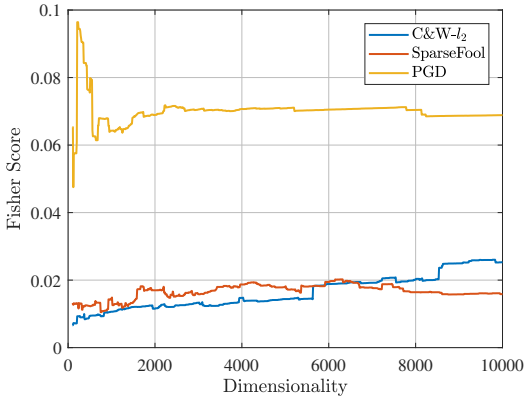
The changes of Fisher scores with dimension are shown in Figure 4. The Fisher scores fluctuate acutely when the dimension is low. The detection accuracy of optimized SRM features with low dimension on C&W-l_2 in CIFAR-10 and PGD ($\epsilon = 8$) in ImageNet is far from optimal. Hence, the range of valid dimension start from 1000 in ImageNet.

The Fisher scores of feature matrices reach their peaks when the dimension is rather small. $D_{opt}^{l_0} = 1790$, $D_{opt}^{l_2} = 1150$,

$D_{opt}^{l_\infty} = 130$. The universally optimal dimension as the maximum among all the detectors is $max\{D_{opt}^{l_0}, D_{opt}^{l_2}, D_{opt}^{l_\infty}\} = 1790$. To guarantee the top-$k\%$ optimized SRM features could cover all the optimal dimensions of various adversarial perturbations, we heuristically set the dimension of optimized SRM features in CIFAR-10 as 2048. Furthermore, in ImageNet, the complexity of image texture requires higher dimension to capture the distortion caused by adversarial perturbations. As the guidance of the Fisher score of feature matrices, the optimal dimension of l_0 and l_2 based method on ImageNet are 6060-D and 9720-D. Following the same logic, we set the dimension in ImageNet as 10000, which is slightly larger than 9720.



(a) CIFAR-10



(b) ImageNet

Fig. 4. (a) and (b) are the Fisher score of feature matrices of SparseFool, C&W-l_2 and PGD ($\epsilon = 8$) in CIFAR-10 and ImageNet respectively.

C. Detection Accuracy Evaluation

In this section, we compare the detection accuracy of the fusion model with the previous works on CIFAR-10 and ImageNet. SRM [41], FS (Feature Squeezing) [32], LID [27] and MDA [30] are adopted for comparison.

The testing set consists of the adversarial examples and their corresponding clean images. The average detection accuracies of different methods are evaluated for comparison. The results are shown in Table I and Table II. It is evident that no matter on CIFAR-10 or ImageNet, the proposed fusion model outperforms the previous works. In CIFAR-10, the fusion model outperforms the previous works with the largest gap of 13.00%. In ImageNet, the largest gap is 18.00%. Compared with SRM, the detection accuracy on sparse perturbed adversarial examples (l_0-based methods and DDN) is significantly improved with an average of 13.38% on CIFAR-10 and 20.39% on ImageNet. One can conclude the proposed scheme outperforms the previous works in detecting oblivious adversarial attacks.

D. Robustness against Second-Round Adversarial Attacks

In this section, the robustness of the fusion model against the second-round adversarial attacks [36] is evaluated. The robustness is evaluated based on the knowledge of the attacker about the detector, white-box (perfect-knowledge) attack scenario and gray-box (limited-knowledge) attack scenario are discussed in this section. The zero-knowledge scenario, where the attacker is not aware of the detector, has already been discussed in Section IV-C.

1) White-box Attack Scenario: When the attacker is fully aware of the defense and parameters, the second-round adversarial attacks can target the detector D and the base model F_{base} by using Eq. 4. The robustness of models is evaluated from three aspects: 1) **Success rate**. A second-round adversarial attack would be considered successful if it can deceive both F_{base} and D. The higher success rate means lower robustness against second-round attacks. 2) **Average l_2 distortion**. The larger distortion makes second-round adversarial examples more easily exposed to another detector or human observations. The higher average l_2 distortion means higher robustness. 3) **The average attempt time**. Excessive attempt time makes real-time attacks harder to implement and creates more opportunities for the defender to fix the bug or find the attacker. Sometimes, massive attempt time may stop some attackers who lack computing resources. More attempt time also means higher robustness of a detector.

To deceive the proposed detector, which contains a majority voting scheme, the attacker mainly has two ways: 1) using a fully connected layer to simulate the classifier; 2) taking the voting results as the logits of the detector. We evaluate the robustness of the fusion model under the attack of the two kinds of attacks. Specifically, the success rate is calculated as N_{succ}/N_{total}, where N_{succ} is the number of adversarial examples that successfully deceive both the detector D and the base model F_{base} and N_{total} is the number of all the images that the attacker tries to craft adversarial examples. The average l_2 distortion is calculated between the adversarial examples, which deceive the substituted detector and the corresponding clean images. Since the white-box second-round adversarial attacks are extremely time-consuming, we only conduct experiments on CIFAR-10. The results are shown in Table III.

Meanwhile, we exhibit the success rate and other statistics of the second-round adversarial attacks against bare CNN detectors and the fusion models, which utilize the fully connected layer as the classifiers in Table IV and Table V. It can be observed that the bare CNN detector is quite vulnerable against the second-round adversarial attacks. When the base

TABLE I
AVERAGE DETECTION ACCURACY (%) COMPARISON IN CIFAR-10.

		FS [32]	LID [27]	MDA [30]	SRM [41]	FD
CornerSearch	ResNet18	67.90%	63.70%	69.30%	84.70%	**93.15%**
	VGG16	62.92%	63.20%	70.20%	85.35%	**95.90%**
CW-l_0	ResNet18	62.10%	65.20%	64.80%	88.00%	**90.40%**
	VGG16	63.02%	62.20%	63.60%	89.00%	**91.20%**
SarpseFool	ResNet18	80.10%	79.40%	77.30%	89.20%	**98.95%**
	VGG16	87.37%	83.70%	88.50%	88.60%	**98.35%**
DeepFool	ResNet18	94.80%	99.60%	99.70%	100.00%	**100.00%**
	VGG16	92.60%	99.40%	99.50%	99.75%	**99.85%**
CW-l_2	ResNet18	82.00%	65.50%	69.60%	89.65%	**98.35%**
	VGG16	89.97%	73.60%	71.50%	86.05%	**97.50%**
DDN	ResNet18	72.60%	67.40%	69.60%	81.00%	**90.10%**
	VGG16	**90.88%**	88.50%	86.50%	71.95%	87.65%
PGD($\epsilon = 2$)	ResNet18	76.30%	68.75%	73.00%	94.30%	**99.90%**
	VGG16	62.69%	69.70%	74.20%	92.40%	**99.75%**
PGD($\epsilon = 4$)	ResNet18	81.80%	69.50%	78.20%	98.25%	**99.90%**
	VGG16	73.99%	73.30%	76.00%	97.70%	**99.95%**
PGD($\epsilon = 8$)	ResNet18	82.50%	70.82%	78.22%	99.75%	**100.00%**
	VGG16	78.36%	73.10%	73.80%	99.80%	**99.95%**
BIM($\epsilon = 2$)	ResNet18	74.50%	69.00%	74.10%	94.90%	**98.95%**
	VGG16	66.85%	69.80%	74.30%	88.55%	**99.90%**
BIM($\epsilon = 4$)	ResNet18	81.10%	69.50%	78.00%	97.20%	**99.90%**
	VGG16	74.44%	71.20%	75.00%	95.50%	**99.90%**
BIM($\epsilon = 8$)	ResNet18	79.70%	73.90%	76.20%	98.45%	**99.85%**
	VGG16	76.10%	71.60%	74.20%	97.75%	**99.70%**

TABLE II
AVERAGE DETECTION ACCURACY (%) COMPARISON IN IMAGENET.

		FS [32]	LID [27]	MDA [30]	SRM [41]	FD
SarpseFool	ResNet34	59.97%	66.01%	71.25%	82.75%	**96.16%**
	VGG16	62.32%	79.73%	72.40%	82.43%	**96.91%**
DeepFool	ResNet34	52.24%	72.96%	76.23%	95.03%	**98.46%**
	VGG16	54.21%	73.14%	75.61%	93.66%	**98.59%**
CW-l_2	ResNet34	91.08%	73.14%	75.61%	88.88%	**98.09%**
	VGG16	94.24%	73.87%	78.94%	91.74%	**98.62%**
DDN	ResNet34	67.72%	61.90%	64.61%	64.52%	**83.14%**
	VGG16	70.00%	60.51%	62.29%	64.24%	**81.42%**
PGD($\epsilon = 2$)	ResNet34	95.61%	90.14%	99.02%	97.67%	**99.58%**
	VGG16	99.30%	82.33%	82.33%	90.14%	**99.74%**
PGD($\epsilon = 4$)	ResNet34	99.59%	94.86%	99.32%	99.37%	**99.93%**
	VGG16	99.76%	99.37%	99.93%	99.69%	**99.92%**
PGD($\epsilon = 8$)	ResNet34	99.86%	99.93%	99.95%	99.87%	**99.98%**
	VGG16	99.95%	99.04%	99.93%	99.86%	**99.99%**
BIM($\epsilon = 2$)	ResNet34	97.12%	84.46%	88.86%	95.33%	**99.07%**
	VGG16	98.22%	86.67%	92.99%	97.60%	**99.34%**
BIM($\epsilon = 4$)	ResNet34	99.45%	94.79%	96.89%	97.70%	**99.71%**
	VGG16	99.79%	96.03%	99.69%	99.14%	**99.76%**
BIM($\epsilon = 8$)	ResNet34	**99.86%**	99.04%	99.43%	98.80%	99.79%
	VGG16	99.86%	92.28%	**100.00%**	99.63%	99.93%

model is ResNet18, the success rate is almost 100%. The average distortion is also lower than attacking VGG16. It may be because that the detector and the base model have a similar structure. The fully connected layer improves the robustness against the second-round adversarial attacks. The success rate drops about 50%. But the proposed fusion model obtains better robustness than the fully connected layer. It detects almost all the second-round adversarial examples against it. Moreover, attacking a fusion model consumes much more time than attacking a bare CNN detector. It takes such a long time that it could stop some attackers and create time for the detector to update the detector's hyper-parameters.

The white-box attack scenario is the worst case of the detector. In this scenario, the proposed scheme exhibits effective robustness against the second-round attacks. It is more robust than bare CNN detectors and the fully connected fusion models.

2) Gray-box Attack Scenario: We examined whether the second-round adversarial examples targeting the CNN detector can be transferred to deceive the fusion model. We assume the attacker obtains the perfect knowledge of the CNN detector, which is merged into the fusion model D_{FM}. Firstly, we evaluate the transferring attack success rate of the second-round adversarial examples towards the detector that only replace the fully connected layer with the ensemble classifier. It can be observed from Table VI that replacing the classifier hardly improves the robustness against second-round adversarial attacks. Secondly, the proposed detector remains quite robust against such transferring attack. The success rates of the

TABLE III
THE SUCCESS RATE, CAUSED l_2 DISTORTION AND ATTEMPTING TIME (SECOND PER IMAGE) OF THE WHITE-BOX SECOND-ROUND ADVERSARIAL ATTACKS AGAINST **THE PROPOSED FUSION MODEL**.

	Success Rate		l_2 distortion		Attempt time (sec/image)	
	VGG	ResNet	VGG	ResNet	VGG	ResNet
Substitute with FC layer	1.45%	2.06%	0.9381	0.9067	276.71	274.75
Take votes as logits	0.00%	0.00%	-	-	259.2	291.5

TABLE IV
THE SUCCESS RATE, CAUSED l_2 DISTORTION AND ATTEMPTING TIME (SECOND PER IMAGE) OF THE WHITE-BOX SECOND-ROUND ADVERSARIAL ATTACKS AGAINST **THE BARE CNN DETECTOR**.

Success Rate		l_2 distortion		Attempt time (sec/image)	
VGG	ResNet	VGG	ResNet	VGG	ResNet
83.96%	99.84%	0.5945	0.4344	18.26	18.09

TABLE V
THE SUCCESS RATE, CAUSED l_2 DISTORTION AND ATTEMPTING TIME (SECOND PER IMAGE) OF THE WHITE-BOX SECOND-ROUND ADVERSARIAL ATTACKS AGAINST **THE FC (FULLY CONNECTED) CLASSIFIER**.

Success Rate		l_2 distortion		Attempt time (sec/image)	
VGG	ResNet	VGG	ResNet	VGG	ResNet
46.19%	48.94%	0.9381	0.9067	276.71	274.75

transferring attack against it are only 1.67% and 1.65%.

TABLE VI
THE SUCCESS RATE OF THE SECOND-ROUND TRANSFERRING ATTACKS FROM THE CNN DETECTOR TOWARDS 1) THE MODEL THAT ONLY REPLACE THE FULLY CONNECTED LAYER WITH ENSEMBLE CLASSIFIER AND 2) THE PROPOSED DETECTOR.

	VGG	ResNet
Replace FC with EC	99.90%	93.50%
The proposed detector	1.67%	1.65%

Based on the results above, the second-round adversarial attacks targeting the deep feature part could hardly be transferred to the fusion model. It proves one of the properties we want in the fusion model, the optimized SRM features still function when the deep features are bypassed, is achieved.

3) Against Backward Pass Differentiable Approximation: Obfuscated gradients [51] (also called gradient masking [58], [59]) are a group of robustness-based defenses. They add or substitute underivable calculations in CNNs to stop the attacker from obtaining valid gradients. Papernot et al. [58] and Athalye et al. [51] proposed transferring attack and BPDA (Back Pass Differentiable Approximations) to circumvent obfuscated gradients respectively. Specifically, transferring attack generate adversarial examples against a local model and transfer to the target one. BPDA leverages differentiable models to generate valid gradients to conduct effective attacks.

In the last section, we try to transfer second-round attack from the CNN detector to the proposed fusion model. The success rates are 1.67% and 1.65%. In this section, we take the CNN detector, which is the most similar model to the optimized SRM, as the differentiable approximation to back propagate valid gradients. We take PGD as the base model. The experiment is conducted on CIFAR-10 as VGG16 being the target model. With 2000 tested images, the success rate is **12.25%**. Though it is higher than transferring attack, the success rate is still quite low. Hence, the proposed detector remains robust under the second-round BPDA attack.

4) Against Adaptive Query-based Attacks: Unlike CNNs, the optimized SRM generates no gradients, so query-based attacks [14], [60] that only requires the predictions of target models can be utilized to conduct a second-round attack against the proposed detector.

Chen et al. [14] proposed a query efficient attack, HSJA (HopSkipJump Attack). In this section, we combine Carlini et al.'s strategy, which considers the detector and the target model as a $N+1$ classifier, and HSJA to conduct a second-round attack against the proposed detector.

The same as the last section, the experiment is conducted on CIFAR-10 and take VGG16 as the target model. Since query-based attacks are time-consuming, we evaluate 200 images. For targeted attacks, the success rate is 13%. For untargeted attacks, the success rate is 55%. As for the CNN detector, the success rate of targeted attacks is 16%, and that of untargeted attacks is 100%. Hence, the proposed detector keeps robust under the second-round query-based attacks [14], [60]. And compared with the CNN detector, the optimized SRM provides such robustness.

E. Ablation Study

1) Channel-wise Correlations: In [41], Liu et al. utilized SRM for color images in a naïve way. They concatenated the three channels to extract SRM features. The correlations among the RGB channels are ignored. To improve the detection ability of SRM features, we add CRM to exploit channel-wise correlations. This section compares the performance between SRMQ1T2 and the SCRMQ1T2 to exhibit the benefits we gain from utilizing the correlations among the channels. For l_0, l_2 and l_∞-based perturbations we take SparseFool, BIM and C&W-l_2 as instances. The results are shown in Table VII. It is evident that the detection accuracies of SCRMQ1T4 are significantly improved from SRM. Especially in detecting sparse l_2-based adversarial perturbations, i.e. DDN, including channel-wise correlations into the detection feature bank improves the detection accuracies by 12.70%, 6.30%, 13.73%, 13.15% in CIFAR-10 and ImageNet. But the improvements on sparse l_0-based perturbations are not that significant. One could infer that the channel-wise correlations are not the key element to better detect sparse perturbations.

2) Reducing the Volume of the Quantization Step Set: The quantization steps of the original SRM consist of 1, 1.5 and 2. In the optimized SRM in this paper, we adopt the quantization step $q = 1$. Thus, the dimensionality of SRMQ1 is reduced to $1/3$ of the original. We compare the average

TABLE VII
THE AVERAGE DETECTION ACCURACIES OF SRMQ1T2 AND SCRMQ1T2 ON SPARSEFOOL, DDN AND BIM ($\epsilon = 2$) IN CIFAR-10 AND IMAGENET.

		SparseFool		DDN		BIM ($\epsilon = 2$)	
		VGG	ResNet	VGG	ResNet	VGG	ResNet
CIFAR-10	SRM	88.60%	89.20%	71.95%	81.00%	88.55%	94.90%
	SCRMQ1T2	**92.55%**	**93.45%**	**84.65%**	**87.30%**	**97.95%**	**98.10%**
ImageNet	SRM	82.43%	82.75%	63.23%	64.52%	97.36%	94.84%
	SCRMQ1T2	**84.19%**	**85.67%**	**76.96%**	**77.67%**	**99.05%**	**98.57%**

classification accuracy of SRM and SRMQ1 on CIFAR-10 and ImageNet. The results are shown in Table VIII. It can be observed that reducing the volume of the quantization step set causes the detection accuracy to drop a little. For some adversarial perturbations, the detection accuracies are even higher than the original. Since the dimension reduction is significant, we adopt $q = 1$ as the only quantization step in the optimized SRM.

3) Expanding The Clamping Threshold: As mentioned in Section III-C, the rich model features are designed for detecting steganographic modifications, which are limited to ± 1. Clamping the residual maps could reduce redundancy for steganalysis, but it would hinder the feature bank from perceiving the sparse but large perturbations. Hence, we propose to expand the clamping threshold T to 4 in order to detect l_0-based adversarial perturbations. In this section, we verify the effect of expanding the clamping threshold. We compare the detection accuracy of SCRMQ1T2 and SCRMQ1T4 on several typical l_0-based adversarial perturbations, i.e., SparseFool, CornerSearch and C&W-l_0. The results are shown in Table IX.

The improvements are not that significant. It is mainly because of the redundancy introduced by expanding the threshold. Combining with the Fisher score based feature reduction, the improvements brought by expanding the threshold is evident. The detailed statistics and analysis are shown in the next section.

4) Dimension Reduction: The Fisher score guides the dimension reduction of the proposed scheme. It not only reduces the complexity of training classifiers but also improves the detection accuracy on l_0-based adversarial perturbations. We evaluate the performances of the full SCRMQ1T4 feature bank and the Fisher reduced SCRMQ1T4 feature bank. The detection accuracies are shown in Table X and Table XI. In Table XI, we specifically exhibit the detection accuracies on l_0-based perturbations in CIFAR-10.

Though increasing the clamping threshold enables the detector to perceive large amplitude perturbations, much redundancy is introduced. Hence, scissoring such redundant features help to reduce the training complexity and improve the prediction accuracy. But the situations in CIFAR-10 and ImageNet are different. In ImageNet, if the dimension is reduced to 2048, the detection accuracies on l_2-based perturbations drop severely. It is because there is a gap between the complexity of the texture of datasets with different resolutions. The images in CIFAR-10 are with 32×32 while the images in ImageNet are much larger and contain much more diversified and complex textures. To detect the images in ImageNet requires features with higher dimension.

5) Unpooled SE-ResNet: The key element in the CNN detector is the unpooled layer in the early stage. To exhibit the improvements brought by it, we compare the detection accuracies of SE-ResNet18 [55], [2] and the proposed CNN detector in Table XII. For l_0, l_2 and l_∞-based perturbations we take SparseFool, DDN and BIM ($\epsilon = 2$) as instances. The training parameters of the max-pooling SE-ResNet18 are the same as the proposed unpooled one.

It is clear that the proposed network outperforms the conventional SE-ResNet18 in detecting adversarial perturbations. In CIFAR-10, only when detecting SparseFool that the accuracy of the conventional SE-ResNet18 is comparable with the unpooled SE-ResNet18. In ImageNet, the model even could not converge when detecting DDN. It is mainly because the l_0-based adversarial perturbations are of large amplitude. They could survive pooling operations, which can be considered as a kind of low-pass filtering. While smaller perturbations, i.e. l_2 and l_∞-based ones, are mostly erased. To exhibit the effect caused by removing the pooling operations in shallow layers. We display the feature maps extracted by the proposed unpooled CNN model and ordinary SE-ResNet18 in Figure 5. The adversarial examples are DDN in the ImageNet. It can be observed that the features extracted by the ordinary SE-ResNet18 are quite flat. One can infer that the effect is caused by the max-pooling in the first layer of SE-ResNet18.

6) Number of Base Classifiers: The number of base classifiers is one important parameter that balances the detection accuracy and the robustness of the proposed scheme. Two extreme cases are 1) all the base classifiers are fed with optimized SRM features; 2) all the base classifiers are fed with deep features. In the first case, the proposed scheme would degrade as the ensemble classifier cooperating with the dimension reduced SCRMQ1T4. In the second case, since the ensemble classifier is more complex than the fully connected layer, the prediction accuracy is slightly higher than the CNN detectors. Meanwhile, we have exhibited in Section IV-D2 that just substituting the last layer (classifier) brings little improvement on the robustness against second-round adversarial attacks. The prediction accuracy of the detector that only replaces the FC layer with the ensemble classifier on the second-round adversarial examples is less than 2.00%.

We take parameter $k\%(k \in [0, 100])$ to present the percentage of the base classifiers trained with deep features that are taken into account of the fusion model. $1-k\%$ is the percentage of the base classifiers trained with optimized SRM features are taken into account of the fusion model. The target common adversarial examples are generated via C&W-l_2. The detection accuracies on common adversarial examples and second-round

TABLE VIII
THE AVERAGE DETECTION ACCURACIES OF SRM AND SRMQ1 ON SPARSEFOOL, DDN AND BIM ($\epsilon = 2$) IN CIFAR-10 AND IMAGENET.

		SparseFool		DDN		BIM ($\epsilon = 2$)	
		VGG	ResNet	VGG	ResNet	VGG	ResNet
CIFAR-10	SRM	88.60%	89.20%	71.95%	81.00%	88.55%	94.90%
	SRMQ1	87.30%	87.70%	71.45%	81.10%	88.35%	93.90%
ImageNet	SRM	82.43%	82.75%	63.23%	64.52%	97.36%	94.84%
	SRMQ1	80.28%	80.83%	62.55%	63.54%	97.04%	94.87%

TABLE IX
THE AVERAGE DETECTION ACCURACIES OF SCRMQ1T2 AND SCRMQ1T4 ON SPARSEFOOL, CORNERSEARCH AND C&W-l_0 IN CIFAR-10 AND IMAGENET.

		SparseFool		CornerSearch		C&W-l_0	
		VGG	ResNet	VGG	ResNet	VGG	ResNet
CIFAR-10	SCRMQ1T2	86.65%	87.75%	**88.20%**	**86.65%**	**83.75%**	**81.40%**
	SCRMQ1T4	**90.55%**	**91.65%**	86.30%	85.85%	81.55%	81.05%
ImageNet	SCRMQ1T2	82.43%	82.75%	-	-	-	-
	SCRMQ1T4	**82.99%**	**84.58%**	-	-	-	-

TABLE X
THE AVERAGE DETECTION ACCURACIES OF SRMQ1T4 AND FS-SCRMQ1T4 ON SPARSEFOOL, DDN AND BIM ($\epsilon = 2$) IN CIFAR-10 AND IMAGENET.

		SparseFool		DDN		BIM ($\epsilon = 2$)	
		VGG	ResNet	VGG	ResNet	VGG	ResNet
CIFAR-10	SCRMQ1T4	90.55%	91.65%	**81.30%**	**87.15%**	96.95%	97.95%
	Fs-SCRMQ1T4	**93.60%**	**95.15%**	80.30%	84.50%	**97.80%**	**97.85%**
ImageNet	SCRMQ1T4	82.99%	84.58%	**76.25%**	**77.36%**	**98.67%**	**98.10%**
	Fs-SCRMQ1T4 (2048)	82.43%	82.75%	65.56%	65.00%	97.59%	96.13%
	Fs-SCRMQ1T4 (10,000)	**85.80%**	**86.62%**	74.19%	75.09%	98.42%	97.71%

TABLE XI
THE AVERAGE DETECTION ACCURACIES OF SCRMQ1T2 AND SCRMQ1T4 ON SPARSEFOOL, CORNERSEARCH AND C&W-l_0 IN CIFAR-10.

		SparseFool		CornerSearch		C&W-l_0	
		VGG	ResNet	VGG	ResNet	VGG	ResNet
CIFAR-10	SCRMQ1T2	86.65%	87.75%	88.20%	86.65%	83.75%	81.40%
	SCRMQ1T4	**93.60%**	**95.15%**	**89.30%**	**88.35%**	**86.20%**	**84.80%**

adversarial examples are a couple of trade-offs determined by $k\%$. We scan the range of $k\%$ with stride $s = 10\%$. The detection accuracies are shown in Figure 6. Note that we take CornerSearch as the common adversarial examples in this experiment.

It can be observed that with the increase of k, the detection accuracy on common adversarial examples slightly increases. Meanwhile, the detection accuracy on second-round adversarial examples suddenly drops when $k = 60$. To balance the detection accuracy on common adversarial examples and the robustness against the second-round adversarial attack, the optimal value of k is 50.

F. The Overhead of the Proposed Detector

The proposed detector is trained and tested with 1 Nvidia RTX 2080 Ti and 10 virtual cores of Xeon Gold 5120 and about 40 GB memory. On CIFAR-10 and ImageNet, using the hardware mentioned before, the detector takes 0.4246 sec/image and 0.4953 sec/image to output predictions, respectively.

V. CONCLUSION

Faced with the difficulties of robustness-based defenses against adversarial examples, the research community recently turned to adversarial example detection. SRM outperformed the previous works by analyzing the image textures subtly. However, in this paper, we discover the limitations of SRM on detecting sparse perturbations. The proposed model obtains robustness and high detection accuracy by combining optimized SRM features and deep features through a linear transformation. Furthermore, to reduce the dimension of handcrafted features while maintaining the detecting accuracies, an expansion-then-reduction process is introduced. Under the second-round adversarial attacks, if the attacker obtains perfect knowledge about the detector, the optimized SRM features and the fusion model still create difficulties for the attacker. The detector remains robust under various kinds of second-round attacks, including BPDA and query-based attack. If the attacker only has access to the deep features, the optimized SRM features can identify most second-round evasive adversarial examples. For the attacker who is unaware of the detector, the proposed fusion model outperforms the previous methods.

REFERENCES

[1] A. Krizhevsky, "One weird trick for parallelizing convolutional neural networks," arXiv preprint arXiv:1404.5997, 2014.

TABLE XII
THE AVERAGE DETECTION ACCURACIES OF SE-RESNET18 AND THE PROPOSED CNN DETECTOR ON SPARSEFOOL, DDN AND PGD ($\epsilon = 2$) IN CIFAR-10 AND IMAGENET.

		SparseFool		DDN		BIM ($\epsilon = 2$)	
		VGG	ResNet	VGG	ResNet	VGG	ResNet
CIFAR-10	SE-ResNet18	99.05%	98.95%	67.65%	80.30%	73.65%	86.50%
	Unpooled CNN	**99.75%**	**99.80%**	**89.70%**	**91.50%**	**98.60%**	**98.95%**
ImageNet	SE-ResNet18	99.46%	99.51%	50.10%	50.02%	84.86%	97.99%
	Unpooled CNN	**99.68%**	**99.81%**	**82.88%**	**84.36%**	**99.39%**	**99.34%**

Fig. 5. From left to right: the input DDN adversarial example, the proposed unpooled CNN feature map and the ordinary SE-ResNet18 feature map.

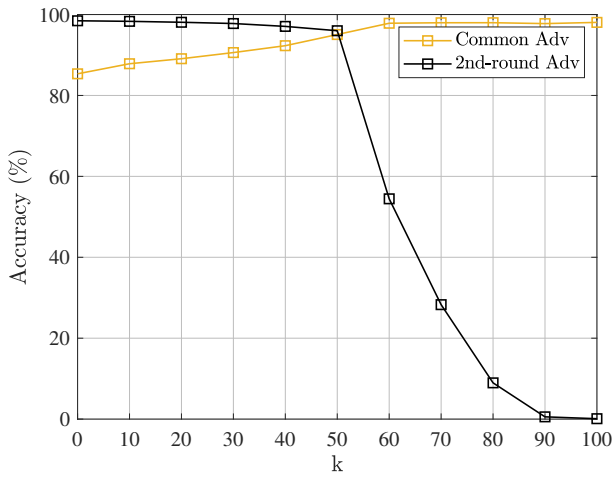


Fig. 6. The detection accuracies on common adversarial examples and second-round adversarial examples targeting the CNN detector change with the parameter k.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2014, pp. 580–587.
[5] R. Girshick, "Fast r-cnn," in Proceedings of the IEEE international conference on computer vision, 2015, pp. 1440–1448.
[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in Advances in neural information processing systems, 2015, pp. 91–99.
[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, 2013.
[8] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, 2014.
[9] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," arXiv preprint arXiv:1611.01236, 2016.
[10] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, 2017.
[11] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2574–2582.
[12] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in 2017 IEEE Symposium on Security and Privacy (SP). IEEE, 2017, pp. 39–57.
[13] J. Rony, L. G. Hafemann, L. S. Oliveira, I. B. Ayed, R. Sabourin, and E. Granger, "Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4322–4330.
[14] J. Chen, M. I. Jordan, and M. J. Wainwright, "Hopskipjumpattack: A query-efficient decision-based attack," in 2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020. IEEE, 2020, pp. 1277–1294. [Online]. Available: https://doi.org/10.1109/SP40000.2020.00045
[15] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in 2016 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2016, pp. 372–387.
[16] A. Modas, S.-M. Moosavi-Dezfooli, and P. Frossard, "Sparsefool: a few pixels make a big difference," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 9087–9096.
[17] F. Croce and M. Hein, "Sparse and imperceivable adversarial attacks," in Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 4724–4732.
[18] F. Tramèr, A. Kurakin, N. Papernot, I. Goodfellow, D. Boneh, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," arXiv preprint arXiv:1705.07204, 2017.

[19] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in Proceedings of the 2017 ACM on Asia conference on computer and communications security. ACM, 2017, pp. 506–519.

[20] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in 2016 IEEE Symposium on Security and Privacy (SP). IEEE, 2016, pp. 582–597.

[21] A. N. Bhagoji, D. Cullina, C. Sitawarin, and P. Mittal, "Enhancing robustness of machine learning systems via data transformations," in 2018 52nd Annual Conference on Information Sciences and Systems (CISS). IEEE, 2018, pp. 1–5.

[22] N. Das, M. Shanbhogue, S.-T. Chen, F. Hohman, L. Chen, M. E. Kounavis, and D. H. Chau, "Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression," arXiv preprint arXiv:1705.02900, 2017.

[23] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1778–1787.

[24] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, "Deflecting adversarial attacks with pixel deflection," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 8571–8580.

[25] D. Hendrycks and K. Gimpel, "Early methods for detecting adversarial images," arXiv preprint arXiv:1608.00530, 2016.

[26] A. N. Bhagoji, D. Cullina, C. Sitawarin, and P. Mittal, "Enhancing robustness of machine learning systems via data transformations," in 2018 52nd Annual Conference on Information Sciences and Systems (CISS). IEEE, 2018, pp. 1–5.

[27] X. Ma, B. Li, Y. Wang, S. M. Erfani, S. Wijewickrema, G. Schoenebeck, D. Song, M. E. Houle, and J. Bailey, "Characterizing adversarial subspaces using local intrinsic dimensionality," arXiv preprint arXiv:1801.02613, 2018.

[28] A. Bendale and T. E. Boult, "Towards open set deep networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 1563–1572.

[29] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," arXiv preprint arXiv:1610.02136, 2016.

[30] K. Lee, K. Lee, H. Lee, and J. Shin, "A simple unified framework for detecting out-of-distribution samples and adversarial attacks," in Advances in Neural Information Processing Systems, 2018, pp. 7167–7177.

[31] B. Liang, H. Li, M. Su, X. Li, W. Shi, and X. Wang, "Detecting adversarial image examples in deep neural networks with adaptive noise reduction," IEEE Transactions on Dependable and Secure Computing, 2018.

[32] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," arXiv preprint arXiv:1704.01155, 2017.

[33] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (statistical) detection of adversarial examples," arXiv preprint arXiv:1702.06280, 2017.

[34] Z. Gong, W. Wang, and W.-S. Ku, "Adversarial and clean data are not twins," arXiv preprint arXiv:1704.04960, 2017.

[35] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," arXiv preprint arXiv:1702.04267, 2017.

[36] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 2017, pp. 3–14.

[37] X. Yu, K. Chen, Y. Wang, W. Li, W. Zhang, and N. Yu, "Robust adaptive steganography based on generalized dither modulation and expanded embedding domain," Signal Processing, vol. 168, p. 107343, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168419303962

[38] T. Qiao, S. Wang, X. Luo, and Z. Zhu, "Robust steganography resisting jpeg compression by improving selection of cover element," Signal Processing, vol. 183, p. 108048, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168421000876

[39] S. Li, D. Ye, S. Jiang, C. Liu, X. Niu, and X. Luo, "Attack on deep steganalysis neural networks," in International Conference on Cloud Computing and Security. Springer, 2018, pp. 265–276.

[40] W. Tang, B. Li, S. Tan, M. Barni, and J. Huang, "Cnn-based adversarial embedding for image steganography," IEEE Transactions on Information Forensics and Security, vol. 14, no. 8, pp. 2074–2087, 2019.

[41] J. Liu, W. Zhang, Y. Zhang, D. Hou, Y. Liu, H. Zha, and N. Yu, "Detection based defense against adversarial examples from the steganalysis point of view," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 4825–4834.

[42] P. Wang, F. Liu, and C. Yang, "Towards feature representation for steganalysis of spatial steganography," Signal Processing, vol. 169, p. 107422, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0165168419304748

[43] J. Fridrich and J. Kodovsky, "Rich models for steganalysis of digital images," IEEE Transactions on Information Forensics and Security, vol. 7, no. 3, pp. 868–882, 2012.

[44] S. Tan, W. Wu, Z. Shao, Q. Li, B. Li, and J. Huang, "Calpa-net: Channel-pruning-assisted deep residual network for steganalysis of digital images," IEEE Transactions on Information Forensics and Security, vol. 16, pp. 131–146, 2020.

[45] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in 2009 IEEE conference on computer vision and pattern recognition. Ieee, 2009, pp. 248–255.

[46] J. Ye, J. Ni, and Y. Yi, "Deep learning hierarchical representations for image steganalysis," IEEE Transactions on Information Forensics and Security, vol. 12, no. 11, pp. 2545–2557, 2017.

[47] M. Boroumand, M. Chen, and J. Fridrich, "Deep residual network for steganalysis of digital images," IEEE Transactions on Information Forensics and Security, vol. 14, no. 5, pp. 1181–1193, 2018.

[48] 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 Transactions on Information Forensics and Security, vol. 15, pp. 1138–1150, 2019.

[49] H. Duda, P. Hart, and G. David, "Stork, pattern classification," ed: John Wiley & Sons, vol. 25, pp. 1150–1157, 2001.

[50] J.-c. Lu, F.-l. Liu, and X.-y. Luo, "Selection of image features for steganalysis based on the fisher criterion," Digital Investigation, vol. 11, no. 1, pp. 57–66, 2014.

[51] A. Athalye, N. Carlini, and D. A. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, ser. Proceedings of Machine Learning Research, J. G. Dy and A. Krause, Eds., vol. 80. PMLR, 2018, pp. 274–283. [Online]. Available: http://proceedings.mlr.press/v80/athalye18a.html

[52] J. Kodovsky, J. Fridrich, and V. Holub, "Ensemble classifiers for steganalysis of digital media," IEEE Transactions on Information Forensics and Security, vol. 7, no. 2, pp. 432–444, 2011.

[53] M. Goljan, J. Fridrich, and R. Cogranne, "Rich model for steganalysis of color images," in 2014 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2014, pp. 185–190.

[54] T. Filler, J. Judas, and J. Fridrich, "Minimizing additive distortion in steganography using syndrome-trellis codes," IEEE Transactions on Information Forensics and Security, vol. 6, no. 3, pp. 920–935, 2011.

[55] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7132–7141.

[56] A. Krizhevsky, G. Hinton et al., "Learning multiple layers of features from tiny images," 2009. [Online]. Available: http://www.cs.toronto.edu/~kriz/cifar.html

[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[58] N. Papernot, P. D. McDaniel, I. J. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, AsiaCCS 2017, Abu Dhabi, United Arab Emirates, April 2-6, 2017, R. Karri, O. Sinanoglu, A. Sadeghi, and X. Yi, Eds. ACM, 2017, pp. 506–519. [Online]. Available: https://doi.org/10.1145/3052973.3053009

[59] F. Tramèr, A. Kurakin, N. Papernot, I. J. Goodfellow, D. Boneh, and P. D. McDaniel, "Ensemble adversarial training: Attacks and defenses," in 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018. [Online]. Available: https://openreview.net/forum?id=rkZvSe-RZ

[60] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12368. Springer, 2020, pp. 484–501. [Online]. Available: https://doi.org/10.1007/978-3-030-58592-1_29