Boundary Defense Against Black-box Adversarial Attacks

Manjushree B. Aithal and Xiaohua Li
Department of Electrical and Computer Engineering
Binghamton University
Binghamton, NY 13902, USA
Email: {maithal1, xli}@binghamton.edu

***Abstract*—Black-box adversarial attacks generate adversarial samples via iterative optimizations using repeated queries. Defending deep neural networks against such attacks has been challenging. In this paper, we propose an efficient Boundary Defense (BD) method which mitigates black-box attacks by exploiting the fact that the adversarial optimizations often need samples on the classification boundary. Our method detects the boundary samples as those with low classification confidence and adds white Gaussian noise to their logits. The method's impact on the deep network's classification accuracy is analyzed theoretically. Extensive experiments are conducted and the results show that the BD method can reliably defend against both soft and hard label black-box attacks. It outperforms a list of existing defense methods. For IMAGENET models, by adding zero-mean white Gaussian noise with standard deviation 0.1 to the logits of those images with classification confidence less than 0.3, the defense reduces the attack success rate to almost 0 while limiting the classification accuracy degradation to around 1 percent.**

I. Introduction

Deep neural networks (DNNs) have achieved increasing demand in many practical applications [1][2][3]. However, studies over the past few years have also shown the intriguing issue that DNN models are very sensitive and vulnerable to adversarial samples [4][5], implying potential security threats to their applications.

One of the widely studied adversarial attacks is the evasion attack, where the main aim of the attacker is to cause misclassification in the DNN model. Black-box evasion attacks have attracted increasing research interests recently, where black-box means that the attacker does not know the DNN model but can query the model to get the DNN inference outputs, either the detailed confidence score or just a classification label [6][7][8][9][10][11][12][13][14][15][16]. If the attacker has access to the full output logit values, they can apply soft-label attack algorithms such as [8][9][10][6][16]. On the other hand, if the attacker has access to only the classification label, they can apply hard-label attack algorithms such as [7][13][15].

Along with the surge of attack algorithm research, there has been an increase in the development of defense algorithms such as Adversarial Training (AT) [17], input transformation [18][19], gradient obfuscation [20], and stochastic defense via randomization [21][22][23][24][25][26][27]. However, limitations of existing defense techniques have also been observed [28][29][30]. It has been proven that stochastic defense suffers from large degradation of DNN performance or limited defense performance. Gradient obfuscation method has also been proven to be ineffective.

In this work, we develop an efficient and more effective method to defend the DNN against black-box attacks. During the adversarial attack's optimization process, there is a stage that the adversarial samples are on the DNN's classification boundary. Boundary Defense BD(θ, σ), our method, detects these boundary samples as those with the classification confidence score below the threshold θ and adds white Gaussian noise with standard deviation σ to their softmax logits. This will prevent the attackers from optimizing their adversarial samples and maintain low DNN performance degradation.

Major contributions of this work are:

- A new boundary defense algorithm BD(θ, σ) is developed, which can be implemented efficiently and mitigate reliably both soft and hard label black-box attacks.
- Theoretical analysis is conducted to study the impact of the parameters θ and σ on the classification accuracy.
- Extensive experiments are conducted, which demonstrate that BD(0.3, 0.1) (or BD(0.7, 0.1)) reduces attack success rate to almost 0 with around 1% (or negligible) classification accuracy degradation over the IMAGENET (or MNIST/CIFAR10) models. The defense performance is shown superior over a list of existing defense algorithms.

The organization of this paper is as follows. In Section II, related works are introduced. In Section III, the BD method is explained. In Section IV, experiment results are presented. Finally, conclusions are given in Section V.

II. Related Work

Black-box adversarial attacks can be classified into soft-label and hard-label attacks. In soft-label attacks like AutoZOOM [8][12] and NES-QL [10], the attacker generates adversarial samples using the gradients estimated from queried DNN outputs. In contrast, SimBA [16], GenAttack [6] and Square Attack [31] resort to direct random search to obtain the adversarial sample. Hard-label attacks like NES-HL [10], BA (Boundary Attack) [7], Sign-OPT [12][13], and HopSkipJump [15] start from an initial adversarial sample and iteratively reduce the distance between the adversarial sample and original sample based on the query results.

For the defense against black-box attacks, a lot of methods are derived directly from the defense methods against white-box attacks, such as input transformation [32], network randomization [33] and adversarial training [34]. Defenses designed specifically for black-box attacks include denoised smoothing [35], malicious query detection [36][37][38], and random-smoothing [39][40]. Nevertheless, their defense performance is not reliable and defense cost or complexity is too high. Random noise defense has been studied recently as a low-cost approach to defend against black-box attacks, where white noise is added to the DNN inputs [41][23][33], weights [42][43][21], or outputs [44]. Nevertheless, a challenge is that heavy noise is needed to defend against some attacks, especially hard-label attacks (in order to change hard labels), but heavy noise leads to severe degradation of DNN accuracy. Our proposed BD method exploits the efficient noise defense principle similarly. But we add noise to the outputs of boundary samples only instead of all samples, with which we can apply heavy noise to mitigate attacks effectively while avoiding significant degradation of DNN accuracy.

III. Boundary Defense

A. Black-box attack model

Consider a DNN that classifies an image $\mathbf{X}_0$ into class label c within N classes. The DNN output is softmax logit (or confidence score) $F(\mathbf{X}_0)$. The classification result is $c = \arg\max_i F_i(\mathbf{X}_0)$, where F_i is the prediction of the ith class, $i = 0, \cdots, N-1$. The attacker does not know the DNN model but can send samples $\mathbf{X}$ to query the DNN and get either $F(\mathbf{X})$ or just c. The objective of the attacker is to generate an adversarial sample $\mathbf{X} = \mathbf{X}_0 + \Delta\mathbf{x}$ such that the output of the classifier is $t = \arg\max_i F_i(\mathbf{X}) \neq c$, where the adversary $\Delta\mathbf{x}$ should be as small as possible.

Soft-Label Black-box Attack: The attacker queries the DNN to obtain the softmax logit output tensor $F(\mathbf{X})$. With this information, the attacker minimizes the loss function $f_{\mathrm{SL}}(\mathbf{X})$ to generate the adversarial sample $\mathbf{X}$ [9],

$$f_{\mathrm{SL}}(\mathbf{X}) = \mathcal{D}(\mathbf{X}, \mathbf{X}_0) + \lambda \mathcal{L}(F(\mathbf{X}), t), \tag{1}$$

where $\mathcal{D}(\cdot,\cdot)$ is a distance function, e.g., $\|\mathbf{X} - \mathbf{X}_0\|_p$, and $\mathcal{L}(\cdot, t)$ is the loss function, e.g., cross-entropy [10].

Hard-Label Black-box Attack: The attacker does not use $F(\mathbf{X})$ but instead uses the class label $\arg\max_i F_i(\mathbf{X})$ to optimize the adversarial sample $\mathbf{X}$. A common approach for the attacker is to first find an initial sample $\mathbf{X}_{t,0}$ in the class $t \neq c$, i.e., $\arg\max_i F_i(\mathbf{X}_{t,0}) = t$. Then, starting from $\mathbf{X}_{t,0}$, the attacker iteratively estimates new adversarial samples $\mathbf{X}$ in the class t so as to minimize the loss function $f_{\mathrm{HL}}(\mathbf{X}) = \mathcal{D}(\mathbf{X}, \mathbf{X}_0)$.

The above model is valid for both targeted and untargeted attacks. The attacker's objective is to increase attack success rate (ASR), reduce query counts (QC), and reduce sample distortion $\mathcal{D}(\mathbf{X}, \mathbf{X}_0)$. In this paper, we assume that the attacker has a large enough QC budget and can adopt either soft-label or hard-label black-box attack algorithms. Thus, our proposed defense's main objective is to reduce the ASR to 0.

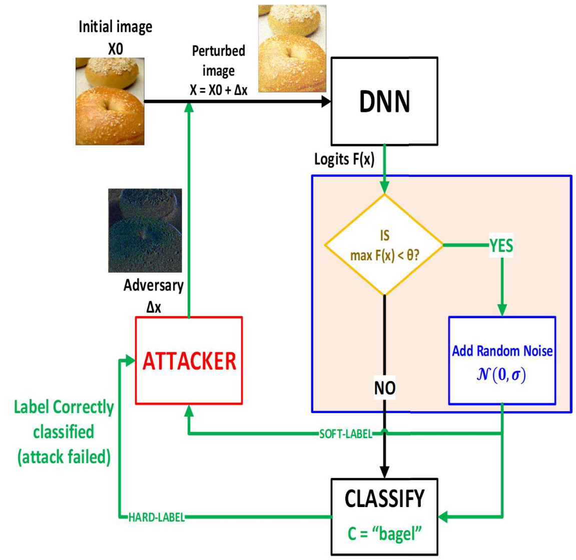


Fig. 1. Schematic representation of the proposed Boundary Defense algorithm $\mathrm{BD}(\theta, \sigma)$ (highlighted region) to prevent soft/hard-label black-box attacks from generating adversarial samples.

B. Boundary Defense Algorithm

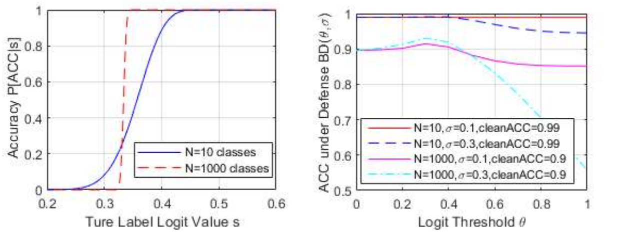
In this work, we propose a Boundary Defense method that defends the DNN against black-box (both soft and hard label, both targeted and untargeted) attacks by preventing the attacker's optimization of $f_{\mathrm{SL}}(\mathbf{X})$ or $f_{\mathrm{HL}}(\mathbf{X})$. As illustrated in Fig. 1, for each query of $\mathbf{X}$, once the defender finds that the classification confidence $\max F(\mathbf{X})$ is less than certain threshold θ, the defender adds zero-mean white Gaussian noise $\mathcal{N}(0, \sigma^2)$ with a certain standard deviation σ to all the elements of $F(\mathbf{X})$. The DNN softmax logits thus become

$$F_{\mathrm{BD}}(\mathbf{X}) = \begin{cases} F(\mathbf{X}), & \text{if } \max F(\mathbf{X}) > \theta \\ F(\mathbf{X}) + V, & \text{otherwise} \end{cases} \tag{2}$$

where V is white Gaussian noise tensor. The DNN output is either softmax logits $\mathrm{clip}\{F_{\mathrm{BD}}(\mathbf{X}),\ 0,\ 1\}$ (soft label) or classification label $\arg\max_i F_{\mathrm{BD},i}(\mathbf{X})$ (hard label).

We call it the $\mathrm{BD}(\theta, \sigma)$ algorithm because samples with low confidence scores are usually on the classification boundary. For a well-designed DNN, the clean (non-adversarial) samples can usually be classified accurately with high confidence scores. Those with low confidence scores happen rarely and have low classification accuracy. In contrast, when the attacker optimizes $f_{\mathrm{SL}}(\mathbf{X})$ or $f_{\mathrm{HL}}(\mathbf{X})$, there is always a stage that the adversarial samples $\mathbf{X}$ have small $\max F(\mathbf{X})$ values. Our BD method exploits this weakness of black-box attacks by detecting these samples and scrambling their query results to prevent the attacker's optimization.

For example, in the soft-label black-box targeted attacks, the attacker needs to maximize the tth logit value $F_t(\mathbf{X})$ by minimizing the cross-entropy loss $\mathcal{L}(F(\mathbf{X}), t) = -\log F_t(\mathbf{X})$. Initially $F_t(\mathbf{X}_0)$ is very small and $F_c(\mathbf{X}_0)$ is large. The optimization increases $F_t(\mathbf{X})$ while reducing $F_c(\mathbf{X})$. There



(a) (b)

Fig. 2. Impact of parameters θ and σ to classification accuracy (ACC). (a) ACC as function of true logit value $s = F_c(\mathbf{X})$. (b) ACC when boundary defense $\mathrm{BD}(\theta, \sigma)$ is applied. CleanACC is the DNN's ACC without attack/defense.

is a stage that all logit values are small, which means $\mathbf{X}$ is lying on the classification boundary.

As another example, a typical hard-label black-box targeted attack algorithm first finds an initial sample inside the target class t, which we denote as $\mathbf{X}_{t,0}$. The algorithm often uses line search to find a boundary sample $\mathbf{X} = \alpha\mathbf{X}_{t,0} + (1-\alpha)\mathbf{X}_0$ that maintains label t, where α is the optimization parameter. Then the algorithm randomly perturbs $\mathbf{X}$, queries the DNN, and uses the query results to find the direction to optimize $f_{\mathrm{HL}}(\mathbf{X})$. Obviously, $\mathbf{X}$ must be on the decision boundary so that the randomly perturbed $\mathbf{X}$ will lead to changing DNN hard-label outputs. Otherwise, all the query results will lead to a constant output t, which is useless to the attacker's optimization process.

One of the advantages of the $\mathrm{BD}(\theta, \sigma)$ algorithm is that it can be implemented efficiently and inserted into DNN models conveniently with minimal coding. Another advantage is that the two parameters (θ, σ) make it flexible to adjust the BD method to work reliably. Large σ leads to small ASR but severe DNN performance degradation. Some attacks are immune to small noise (small σ), such as the HopSkipJump hard-label attack [15]. Some other attacks such as SimBA [16] are surprisingly immune to large noise in boundary samples, so simply removing boundary samples or adding extra large noise to boundary samples (suggested in [15]) does not work. In contrast, the flexibility of (θ, σ) makes the BD method effective to resolve such complicated issues.

C. Properties of Boundary Samples

In this section, we study $\mathrm{BD}(\theta, \sigma)$'s impact on the DNN's classification accuracy (ACC) when there is no attack, which provides useful guidance to the selection of θ and σ.

Consider a clean sample $\mathbf{X}$ with true label c and confidence $s = F_c(\mathbf{X})$. Since the DNN is trained with the objective of maximizing $F_c(\mathbf{X})$, we can assume that all the other logit values $F_i(\mathbf{X})$, $i = 0, \cdots, N-1$, $i \neq c$, are independent and identically distributed uniform random variables with values within 0 to $a = (1-s)/(N-1)$, i.e., $F_i(\mathbf{X}) \sim U(0, a)$. Without loss of generality, let $F_0(\mathbf{X})$ be the maximum among these $N-1$ values. Then $Y = \sum_{i=1, i\neq c}^{N-1} F_i(\mathbf{X})$ follows Irwin-Hall distribution with cumulative distribution function (CDF)

$$P_Y[y < x] = \frac{1}{(N-2)!} \sum_{k=0}^{\lfloor x/a \rfloor} (-1)^k \begin{pmatrix} N-2 \\ k \end{pmatrix} \left(\frac{x}{a} - k\right)^{N-2}. \tag{3}$$

When N is large, the distribution of Y can be approximated as normal $\mathcal{N}\left(\frac{(1-s)(N-2)}{2(N-1)}, \left(\frac{1-s}{N-1}\right)^2 \frac{N-2}{12}\right)$. We denote its CDF as $\Phi(x)$. Since the sample $\mathbf{X}$ is classified accurately if and only if $s > F_0(\mathbf{X})$, the classification accuracy $P[\mathrm{ACC}|s]$ can be derived as

$$P[\mathrm{ACC}|s] = P_Y[y > 1-2s] = 1 - \Phi(1-2s). \tag{4}$$

Using (4), we can calculate $P[\mathrm{ACC}|s]$ for each s, as shown in Fig. 2(a) for $N = 10$ and 1000 classes. It can be seen that for $N = 1000$, if $s < 0.32$, then the sample's classification ACC is almost 0. This means that we can set $\theta \leq 0.32$ to safely scramble all those queries whose maximum logit value is less than 0.32 without noticeable ACC degradation.

Next, to model the ACC when $\mathrm{BD}(\theta, \sigma)$ is applied, we assume the true label c's logit value s follow approximately half-normal distribution, whose probability density function is

$$f_S(s) = \begin{cases} \frac{\sqrt{2}}{\nu\sqrt{\pi}} e^{-\frac{(1-s)^2}{2\nu^2}}, & s \leq 1 \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

with the parameter ν. The ACC of the DNN without attack or defense (which we call cleanACC) is then

$$\mathrm{cleanACC} = \int_0^1 P[\mathrm{ACC}|s] f_S(s) ds. \tag{6}$$

Using (5)-(6), we can find the parameter ν for each clean ACC. For example, for $N = 1000$ and a DNN with clean ACC 90%, the distribution of true logit s follows (5) with $\nu = 0.41$.

When noise is added, each $F_i(X)$ becomes $F_i(X) + v_i$ with noise $v_i \sim \mathcal{N}(0, \sigma^2)$. Following similar derivation of (3)-(4), we obtain the ACC of the noise perturbed $F_c(X) + v_c$ as

$$P[\mathrm{ACC}|s, \sigma] = 1 - \tilde{\Phi}(1-2s), \tag{7}$$

where $\tilde{\Phi}$ is the CDF of the new normal distribution $\mathcal{N}\left(\frac{(1-s)(N-2)}{2(N-1)}, \left(\frac{1-s}{N-1}\right)^2 \frac{N-2}{12} + (N+2)\sigma^2\right)$. The ACC under the defense is then

$$\mathrm{ACC} = \int_0^\theta P[\mathrm{ACC}|s, \sigma] f_S(s) ds + \int_\theta^1 P[\mathrm{ACC}|s] f_S(s) ds. \tag{8}$$

Fig. 2(b) shows how the defense ACC (8) degrades with the increase of θ and σ. We can see that with $\sigma = 0.1$, there is almost no ACC degradation for $N = 10$. For $N = 1000$, ACC degradation is very small when $\theta < 0.4$ but grows to 5% when $\theta > 0.6$. Importantly, under $\theta < 0.4$ we can apply large noise $\sigma = 0.3$ safely without obvious ACC degradation. This shows the importance of adding noise to boundary samples only. Existing defenses [23][44] scramble all the samples, which corresponds to $\theta = 1$, and thus suffer from significant ACC degradation.

IV. Experiments

A. Experiment Setup

In the first experiment, we evaluated the degradation of classification accuracy of a list of popular DNN models when there is no attack but our proposed BD method is applied. We used the full validation datasets of MNIST (10,000 images), CIFAR10 (10,000 images), and IMAGENET (50,000 images).

In the second experiment, with 1000 validation images of MNIST/CIFAR10 and 100 validation images of IMAGENET, we evaluated the defense performance of our BD method against several state-of-the-art black-box attack methods, including soft-label attacks **AZ** (AutoZOOM) [9], **NES-QL** (query limited) [10], **SimBA** (SimBA-DCT) [16], and **GA** (GenAttack) [6], as well as hard-label attacks **NES-HL** (hard label) [10], **BA** (Boundary Attack) [7], **HSJA** (HopSkipJump Attack) [15], and **Sign-OPT** [13]. We adopted their original source codes with the default hyper-parameters and just inserted our BD(θ, σ) as a subroutine to process $F(\mathbf{X})$ after each model prediction call. These algorithms used the InceptionV3 or ResNet50 IMAGENET models. We considered the l_2 norm setting throughout the experiment.

Next, we compared our BD method with some representative black-box defense methods, including **JPEG** compression, **Bit-Dept**, and **TVM** (Total Variation Minimization), whose data were obtained from [45], for soft-label attacks, and **DD** (Defensive Distillation) [20], **Region-based** classification [46], and **AT** (Adversarial Training) [47] for hard-label attacks. Especially, we compared our method with two similar noise defense methods **RND** [23] and **NP** (Noise Perturbation) [44] which add noise to DNN inputs and outputs, respectively.

In order to have a more persuasive and comprehensive study of the robustness of the proposed BD method, we also performed experiments using Robust Benchmark models [48], such as **RMC** (Runtime Masking and Cleaning) [49], **RATIO** (Robustness via Adversarial Training on In- and Out-distribution) [50], **RO** (Robust Overfitting) [51], **MMA** (Max-Margin Adversarial) [52], **ER** (Engstrom Robustness) [53], **RD** (Rony Decoupling) [54], and **PD** (Proxy Distribution) [55] models, over the CIFAR10 dataset for various attack methods.

As the primary performance metrics, we considered **ACC** (DNN's classification accuracy) and **ASR** (attacker's attack success rate). The ASR is defined as the ratio of samples with $\arg\max_i F_i(\mathbf{X}) = t \neq c$, without sample distortion constraint. On the other hand, since most hard-label attack/defense papers use the ASR defined as the ratio of samples satisfying both $\arg\max_i F_i(\mathbf{X}) = t \neq c$ and median l_2 distortion ($\sqrt{\|\mathbf{X} - \mathbf{X}_0\|^2/M}$ when $\mathbf{X}_0$ has M elements) less than a certain threshold, we will also report our results over this ASR, which we called **ASR2**.

We show only the results of targeted attacks in this section. Supplementary material shows experiments of untargeted attacks (Section B), extra experiment data of targeted attacks (Section A) and result discussions (Section C).

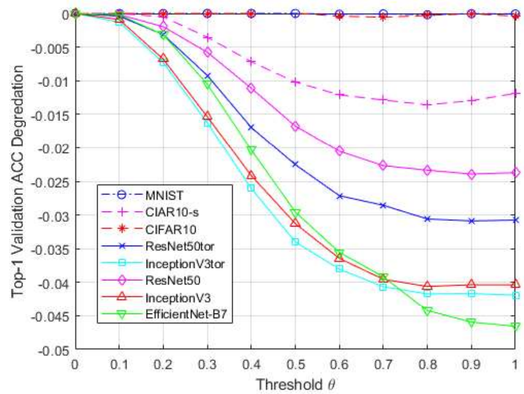


Fig. 3. Top-1 classification accuracy degradation (Defense ACC − Clean ACC) versus θ. $\sigma = 0.1$.

B. ACC Degradation Caused by Boundary Defense

We applied our BD algorithm to modify the DNN outputs and evaluated classification ACC degradation. For MNIST, we trained a 5-layer convolutional neural network (CNN) with clean ACC 99%. For CIFAR10, we trained a 6-layer CNN with clean ACC 83% and also applied the pre-trained model of [56] with ACC 97%, which are called **CIFAR10-s** and **CIFAR10**, respectively. For IMAGENET, we used standard pre-trained models from the official Tensoflow library (**ResNet50**, **InceptionV3**, **EfficientNet-B7**) and the official PyTorch library (**ResNet50tor**, **InceptionV3tor**), where "-tor" indicates their PyTorch source.

It can be observed from Fig. 3 that even adding heavy noise ($\sigma = 0.1$), by making $\theta \leq 0.3$ we can keep the loss of ACC around just 1% for IMAGENET models (from 0.5% of ResNet50 to 1.5% of InceptionV3). Existing noise defense methods such as [44] are equivalent to $\theta = 1$ and would result in up to 5% ACC degradation. For MNIST and CIFAR10 models, the ACC shows almost no degradation except that CIFAR10-s has limited 1.5% ACC degradation for large θ. This fits well with the analysis results shown in Fig. 2(b).

C. Performance of BD Defense against Attacks

1) ASR of soft-label black-box attacks: Table I shows the ASR of soft-label black-box attack algorithms under our proposed BD method. To save the space we have shown the data of $\sigma = 0.1$ only. Results regarding varying θ and σ are shown in Fig. 4.

From Table I we can see that with the increase in θ, the ASR of all the attack algorithms drastically reduced. Over the IMAGENET dataset, the BD method reduced the ASR of all the attack algorithms to almost 0 with $(\theta, \sigma) = (0.3, 0.1)$. For MNIST/CIFAR10 datasets, the BD(0.5, 0.1) was enough. Fig. 4 shows a consistent decline of ASR over the increase in noise level σ. This steady decline indicates robust defense performance of the BD method against the soft-label attacks.

2) ASR of hard-label black-box attacks: We have summarized the ASR and median l_2 distortion of hard-label attacks in

TABLE I

ASR (%) OF TARGETED SOFT-LABEL ATTACKS. $\sigma = 0.1$.

Dataset	Attacks	No defense	$\theta_1 = 0.5$	$\theta_2 = 0.7$
MNIST	AZ	100	8	8
	GA	100	0	0
	SimBA	97	3	0
CIFAR10	AZ	100	9	9
	GA	98.76	0	0
	SimBA	97.14	23	15
Dataset	Attacks	No Defense	$\theta_1 = 0.1$	$\theta_2 = 0.3$
IMAGENET	AZ	100	0	0
	NES-QL	100	69	8
	GA	100	0	0
	SimBA	96.5	6	2

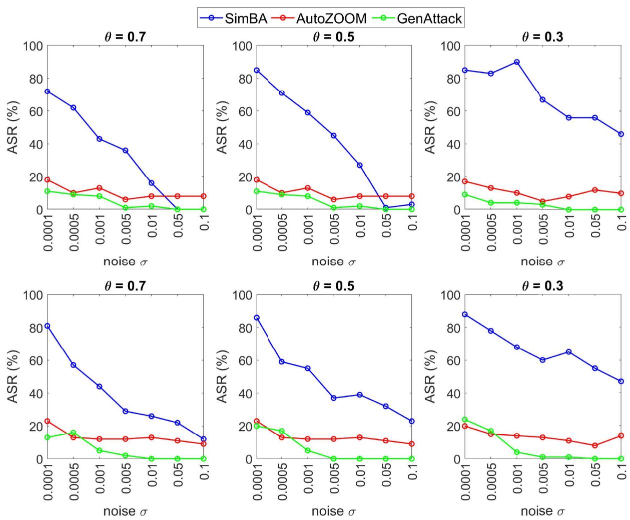


Fig. 4. ASR(%) vs noise level σ for various boundary threshold θ. The top row is for MNIST, and the bottom row is for CIFAR10.

TABLE II

ASR (%) AND MEDIAN l_2 DISTORTION OF TARGETED HARD-LABEL ATTACKS. $\sigma = 0.1$. "-" MEANS NO l_2 DISTORTION DATA DUE TO ABSENCE OF ADVERSARIAL SAMPLES.

Dataset	Attacks	ASR/l_2 No defense	ASR/l_2 $\theta_1 = 0.5$	ASR/l_2 $\theta_2 = 0.7$
MNIST	Sign-OPT	100/0.059	4/0.12	0/-
	BA	100/0.16	17/0.55	9/0.56
	HSJA	100/0.15	38/0.14	7/0.15
CIFAR10	Sign-OPT	100/0.004	4/0.08	0/-
	HSJA	100/0.05	18/0.05	7/0.05
Dataset	Attacks	ASR/l_2 No Defense	ASR/l_2 $\theta_1 = 0.1$	ASR/l_2 $\theta_2 = 0.3$
IMAGE-NET	NES-HL	90/0.12	0/-	0/-
	Sign-OPT	100/0.05	14/0.4	0/-
	BA	100/0.08	0/-	0/-
	HSJA	100/0.03	34/0.11	0/-

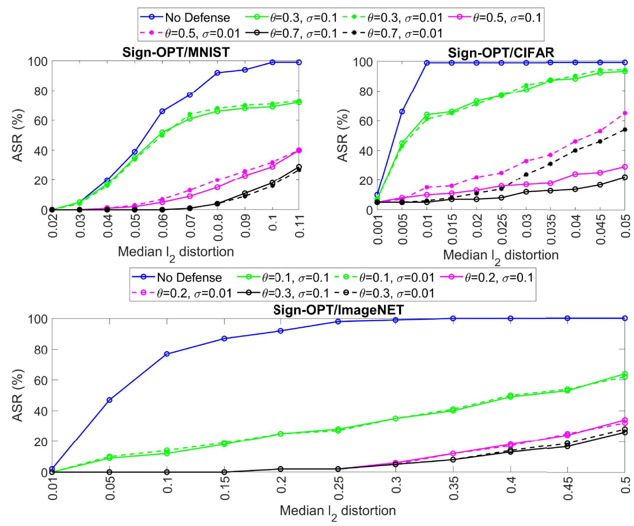


Fig. 5. ASR (%) versus median l_2 distortion of the Sign-OPT attack under the proposed BD method.

presence of our proposed BD method in Table II. Surprisingly, the BD method performed extremely well against the hard-label attacks that were usually challenging to conventional defense methods. In general, BD(0.3, 0.1) was able to reduce ASR to 0% over the IMAGENET dataset, and BD(0.7, 0.1) was enough to reduce ASR to near 0 over MNIST and CIFAR10.

Figure 5 shows how ASR2 varies with the pre-set l_2 distortion threshold when the BD method was used to defend against the Sign-OPT attack. We can see that ASR2 reduced with the increase of either θ or σ, or both. BD(0.7, 0.1) and BD(0.3, 0.1) successfully defended against the Sign-OPT attack over the MNIST/CIFAR10 and IMAGENET datasets, respectively.

D. Comparison with Other Defense Methods

To demonstrate the advantage of our proposed BD method, we first compare it with two similar and representative noise defense methods RND and NP. Results are shown in Table III, which clearly demonstrated our method was superior. It mitigated successfully both the soft-label attack GA and hard-label attacks BA, HSJA and Sign-OPT with zero ASR and minor ACC degradation. NP needed heavy noise which led to higher ACC degradation. Results of RND showed that adding noise to inputs was not as effective as adding noise to outputs.

Next, Table IV compares the BD method with the DD, AT, and Region-based defense methods against hard-label attacks. It can be seen that our BD method outperformed all these defense methods with lower ASR. For soft-label attacks, Table V shows that our BD method also outperformed a list of existing defense methods.

Finally, we also ran experiments using RobustBench models. The defense performance over the CIFAR10 dataset is reported in Table VI. ASR is used as our preliminary evaluation criteria because for an attacker the higher the ASR the more robust the attack method is against all the defenses. From Table VI, we can see that our method had the most superior

TABLE III
PERFORMANCE OF NOISE DEFENSE METHODS BD(θ, σ), RND [23] AND NP [44] AGAINST TARGETED BLACK-BOX ATTACKS. DNN MODEL WAS IMAGENET INCEPTIONV3.

Noise Defense Method	Degradation of ACC (%)	Attack Algorithm's ASR (%)			
		GA	BA	HSJA	Sign-OPT
RND($v = 0.02$)	-1.2	97	100	100	100
RND($v = 0.05$)	-4.5	98	100	100	100
RND($v = 0.1$)	-9.5	96	100	100	100
NP($\sigma = 10^{-6}$)	0.0	70	90	100	60
NP($\sigma = 0.1$)	-4.1	0	0	0	0
BD (0.3, 0.1)	**-1.5**	**0**	**0**	**0**	**0**

TABLE IV
COMPARISON OF BD METHOD WITH OTHER DEFENSE METHODS AGAINST TARGETED HARD-LABEL ATTACKS IN TERMS OF ASR (%).

Dataset	Defense	HSJA	BA	SimBA-DCT
MNIST	DD [20]	98	80	-
	AT [47]	100	50	4
	Region-based [46]	100	85	-
	BD ($\theta = 0.5, \sigma = 0.1$)	**38**	**17**	**3**
	BD ($\theta = 0.7, \sigma = 0.1$)	**7**	**9**	**0**

defense performance.

E. Robust defense performance against adaptive attacks

To evaluate the robustness of the defense, it is crucial to evaluate the defense performance against adaptive attacks [34]. For example, the attacker may change the query limit or optimization step size. In this subsection, we show the effectiveness of our BD defense against 2 major adaptive attack techniques: 1) adaptive query count (QC) budget; and 2) adaptive step size.

First, with increased attack QC budget, the results obtained are summarized in Table VII. We observe that when the

TABLE V
COMPARISON OF BD METHOD ($\theta = 0.5, \sigma = 0.1$) WITH OTHER DEFENSE METHODS AGAINST TARGETED SOFT-LABEL GENATTACK IN TERMS OF ASR (%).

Dataset	Attack	Bit-Depth	JPEG	TVM	BD
MNIST	GA	95	89	-	**0**
CIFAR10	GA	95	89	73	**0**

TABLE VI
COMPARE ASR (%) OF PROPOSED BD METHOD WITH THE ROBUST BENCH DEFENSE MODELS. CIFAR10 DATASET.

RobustBench Defense	Sign-OPT	SimBA	HSJA
RMC [49]	100	83	100
RATIO [50]	100	59	100
RO [51]	100	85	100
MMA [52]	100	83	100
ER [53]	100	92	100
RD [54]	-	80	-
PD [55]	100	71	100
BD ($\theta = 0.5, \sigma = 0.1$)	**4**	**23**	**18**
BD ($\theta = 0.7, \sigma = 0.1$)	**0**	**15**	**7**

TABLE VII
ASR(%) OF ADAPTIVE BLACK-BOX ATTACKS UNDER THE PROPOSED BD METHOD

Dataset	Attack	query 10^4 (preset)	budget 10^8	10^{10}
CIFAR10	GA	0	0	0
	HSJA	3	4	0
	Sign-OPT	5	8	8
ImageNET	NES-QL	2	12	8
	Boundary	0	0	0
	HSJA	0	0	0
	Sign-OPT	0	0	0

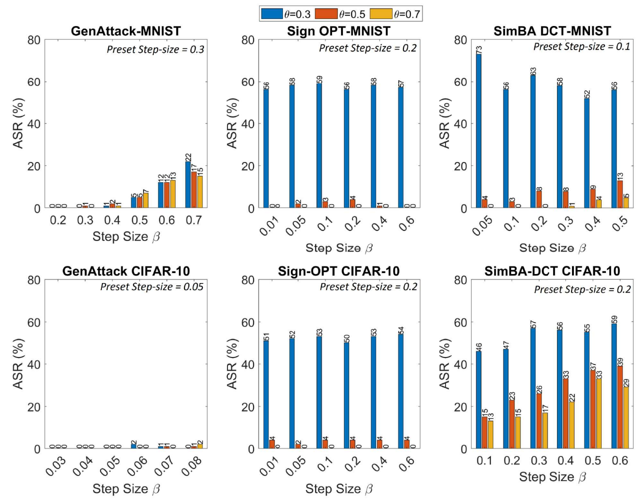


Fig. 6. ASR (%) versus step size of the adaptive attacks. Note that for GenAttack we considered $\sigma = 0.01$, since the ASR for $\sigma = 0.1$ was always 0 for all threshold values. For Sign-OPT & SimBA we considered $\sigma = 0.1$.

attacker increased QC from 10^4 to 10^{10}, there was no significant increase in ASR. Next, we adjusted the optimization (or gradient estimation) step size of the attack algorithms (such as β of the Sign-OPT algorithm), and evaluated the performance of BD. The ASR data are shown in Fig 6. We can see that there was no significant change of ASR when the attack algorithms adopted different optimization step sizes.

As a result, we can assert the robustness of the BD method against the black-box adversarial attacks.

V. CONCLUSIONS

In this paper, we propose an efficient and effective boundary defense method BD(θ, σ) to defend against black-box attacks. This method detects boundary samples by examining classification confidence scores and adds random noise to the query results of these boundary samples. BD$(0.3, 0.1)$ is shown to reduce the attack success rate to almost 0 with only about 1% classification accuracy degradation for IMAGENET models. Analysis and experiments were conducted to demonstrate that this simple and practical defense method could effectively defend the DNN models against state-of-the-art black-box attacks.

REFERENCES

[1] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2722–2730.

[2] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[4] 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.

[5] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[6] M. Alzantot, Y. Sharma, S. Chakraborty, H. Zhang, C.-J. Hsieh, and M. B. Srivastava, "Genattack: Practical black-box attacks with gradient-free optimization," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2019, pp. 1111–1119.

[7] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *International Conference on Learning Representations*, 2018.

[8] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 15–26.

[9] C.-C. Tu, P. Ting, P.-Y. Chen, S. Liu, H. Zhang, J. Yi, C.-J. Hsieh, and S.-M. Cheng, "Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 742–749.

[10] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2137–2146.

[11] S. Cheng, Y. Dong, T. Pang, H. Su, and J. Zhu, "Improving black-box adversarial attacks with a transfer-based prior," in *Advances in Neural Information Processing Systems*, 2019, pp. 10 934–10 944.

[12] M. Cheng, T. Le, P.-Y. Chen, H. Zhang, J. Yi, and C.-J. Hsieh, "Query-efficient hard-label black-box attack: An optimization-based approach," in *International Conference on Learning Representation (ICLR)*, 2019.

[13] M. Cheng, S. Singh, P. H. Chen, P.-Y. Chen, S. Liu, and C.-J. Hsieh, "Sign-opt: A query-efficient hard-label adversarial attack," in *International Conference on Learning Representations*, 2019.

[14] Y. Li, L. Li, L. Wang, T. Zhang, and B. Gong, "Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 3866–3876.

[15] 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)*. IEEE, 2020, pp. 1277–1294.

[16] C. Guo, J. Gardner, Y. You, A. G. Wilson, and K. Weinberger, "Simple black-box adversarial attacks," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2484–2493.

[17] F. Tramèr, D. Boneh, A. Kurakin, I. Goodfellow, N. Papernot, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings*, 2018.

[18] J. Buckman, A. Roy, C. Raffel, and I. Goodfellow, "Thermometer encoding: One hot way to resist adversarial examples," in *International Conference on Learning Representations*, 2018.

[19] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," in *International Conference on Learning Representations*, 2018.

[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] Z. He, A. S. Rakin, and D. Fan, "Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 588–597.

[22] X. Wang, S. Wang, P.-Y. Chen, Y. Wang, B. Kulis, X. Lin, and S. P. Chin, "Protecting neural networks with hierarchical random switching: Towards better robustness-accuracy trade-off for stochastic defenses," in *IJCAI*, 2019.

[23] Z. Qin, Y. Fan, H. Zha, and B. Wu, "Random noise defense against query-based black-box attacks," *Advances in Neural Information Processing Systems*, vol. 34, 2021.

[24] F. Nesti, A. Biondi, and G. Buttazzo, "Detecting adversarial examples by input transformations, defense perturbations, and voting," *arXiv preprint arXiv:2101.11466*, 2021.

[25] 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.

[26] W. Fan, G. Sun, Y. Su, Z. Liu, and X. Lu, "Integration of statistical detector and gaussian noise injection detector for adversarial example detection in deep neural networks," *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 20 409–20 429, 2019.

[27] B. Li, C. Chen, W. Wang, and L. Carin, "Certified adversarial robustness with additive noise," in *Advances in Neural Information Processing Systems*. Neural information processing systems foundation, 2019.

[28] A. Athalye, N. Carlini, and D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples," in *International conference on machine learning*. PMLR, 2018, pp. 274–283.

[29] 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.

[30] ——, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[31] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: a query-efficient black-box adversarial attack via random search," in *European Conference on Computer Vision*. Springer, 2020, pp. 484–501.

[32] G. K. Dziugaite, Z. Ghahramani, and D. M. Roy, "A study of the effect of jpg compression on adversarial images," *arXiv preprint arXiv:1608.00853*, 2016.

[33] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, "Mitigating adversarial effects through randomization," in *International Conference on Learning Representations*, 2018.

[34] F. Tramer, N. Carlini, W. Brendel, and A. Madry, "On adaptive attacks to adversarial example defenses," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[35] H. Salman, M. Sun, G. Yang, A. Kapoor, and J. Z. Kolter, "Denoised smoothing: A provable defense for pretrained classifiers," *arXiv preprint arXiv:2003.01908*, 2020.

[36] S. Chen, N. Carlini, and D. Wagner, "Stateful detection of black-box adversarial attacks," in *Proceedings of the 1st ACM Workshop on Security and Privacy on Artificial Intelligence*, 2020, pp. 30–39.

[37] H. Li, S. Shan, E. Wenger, J. Zhang, H. Zheng, and B. Y. Zhao, "Blacklight: Defending black-box adversarial attacks on deep neural networks," *arXiv preprint arXiv:2006.14042*, 2020.

[38] R. Pang, X. Zhang, S. Ji, X. Luo, and T. Wang, "Advmind: Inferring adversary intent of black-box attacks," in *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020, pp. 1899–1907.

[39] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *International Conference on Machine Learning*. PMLR, 2019, pp. 1310–1320.

[40] H. Salman, G. Yang, J. Li, P. Zhang, H. Zhang, I. Razenshteyn, and S. Bubeck, "Provably robust deep learning via adversarially trained smoothed classifiers," *arXiv preprint arXiv:1906.04584*, 2019.

[41] J. Byun, H. Go, and C. Kim, "Small input noise is enough to defend against query-based black-box attacks," *arXiv preprint arXiv:2101.04829*, 2021.

[42] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 2019, pp. 656–672.

[43] X. Liu, M. Cheng, H. Zhang, and C.-J. Hsieh, "Towards robust neural networks via random self-ensemble," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 369–385.

[44] M. B. Aithal and X. Li, "Mitigating black-box adversarial attacks via output noise perturbation," *IEEE Access*, vol. 10, pp. 12 395–12 411, 2022.

[45] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, "Countering adversarial images using input transformations," in *International Conference on Learning Representations*, 2018.

[46] X. Cao and N. Z. Gong, "Mitigating evasion attacks to deep neural networks via region-based classification," in *Proceedings of the 33rd Annual Computer Security Applications Conference*, 2017, pp. 278–287.

[47] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *stat*, vol. 1050, p. 20, 2015.

[48] F. Croce, M. Andriushchenko, V. Sehwag, E. Debenedetti, N. Flammarion, M. Chiang, P. Mittal, and M. Hein, "Robustbench: a standardized adversarial robustness benchmark," in *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2021. [Online]. Available: https://openreview.net/forum?id=SSKZPJCt7B

[49] Y.-H. Wu, C.-H. Yuan, and S.-H. Wu, "Adversarial robustness via runtime masking and cleansing," in *International Conference on Machine Learning*. PMLR, 2020, pp. 10 399–10 409.

[50] M. Augustin, A. Meinke, and M. Hein, "Adversarial robustness on in- and out-distribution improves explainability," in *European Conference on Computer Vision*. Springer, 2020, pp. 228–245.

[51] L. Rice, E. Wong, and Z. Kolter, "Overfitting in adversarially robust deep learning," in *International Conference on Machine Learning*. PMLR, 2020, pp. 8093–8104.

[52] G. W. Ding, Y. Sharma, K. Y. C. Lui, and R. Huang, "Mma training: Direct input space margin maximization through adversarial training," *arXiv preprint arXiv:1812.02637*, 2018.

[53] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, B. Tran, and A. Madry, "Adversarial robustness as a prior for learned representations," *arXiv preprint arXiv:1906.00945*, 2019.

[54] 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/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4322–4330.

[55] V. Sehwag, S. Mahloujifar, T. Handina, S. Dai, C. Xiang, M. Chiang, and P. Mittal, "Improving adversarial robustness using proxy distributions," *arXiv preprint arXiv:2104.09425*, 2021.

[56] Y. Xu, L. Xie, X. Zhang, X. Chen, G.-J. Qi, Q. Tian, and H. Xiong, "Pc-darts: Partial channel connections for memory-efficient architecture search," *arXiv preprint arXiv:1907.05737*, 2019.

Supplementary Material

Appendix

A. Supplementary Experiment Results for Defense Against Targeted Attacks

1) Median l_2 Distortion of Soft-Label Targeted Attacks: In Table VIII, we have listed the median l_2 distortion of targeted soft-label attacks when our proposed BD(θ, σ) defense was applied. Note that when the ASR was 0%, then there were no successful adversarial samples generated, and thus the l_2 distortion was absent. From Table VIII, we can see that there is no significant change in the median l_2 distortion when comparing the cases with defense and without defense. This is the same as the result of hard-label attacks shown in Table II.

TABLE VIII
MEDIAN l_2 DISTORTION OF TARGETED SOFT-LABEL ATTACKS UNDER THE BD(θ, σ) DEFENSE. $\sigma = 0.1$

Dataset	Attacks	No defense	θ_1 = 0.5	θ_2 = 0.7
MNIST	AZ	0.0473	0.0478	0.0478
	GenAttack	4.8086	-	-
	SimBA	0.0957	0.0694	0.0823
CIFAR10	AZ	0.0672	0.0672	0.0672
	GenAttack	1.3874	-	-
	SimBA	0.0323	0.0236	0.0218
Dataset	Attacks	No Defense	θ_1 = 0.1	θ_2 = 0.3
IMAGENET	AZ	0.1566	-	-
	GenAttack	0.0377	-	-
	SimBA	0.0163	0.0092	0.0075

2) Visual Representation of Targeted Attack Samples: For the adversarial samples generated by the targeted attack algorithms under our proposed defense, we examine whether the distortion, i.e., the difference between the original sample and the adversarial sample, is imperceptible to the human eye. Some of the adversarial samples are shown in Fig. 7 and 8. Fig. 7 shows the MNIST adversarial samples generated by the AZ, GenAttack, and SimBA attack methods with our BD(θ, σ) defense. We can observe that when the proposed BD method was used, the adversarial images generated were extremely distorted.

In Section IV-E, we showed the robust performance of the proposed BD method against adaptive attacks, where we found that ASR stayed the same or just increased slightly even under the adaptive attacks. For adaptive attacks that changes the optimization step-size to increase ASR, some of the adversarial samples in the presence of the BD method are illustrated in Fig 8. We can see that the cost of slight increase of ASR was the heavy image distortion that can be detected by human perception easily.

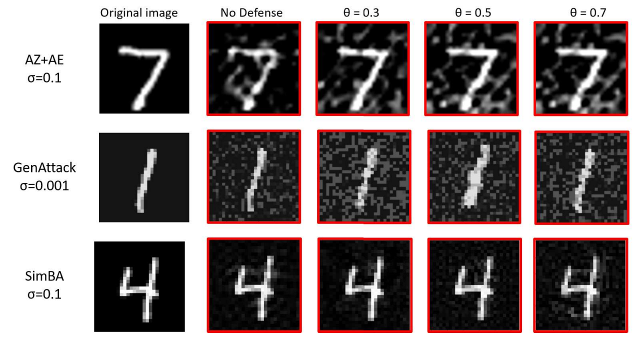


Fig. 7. Adversarial samples generated by the targeted soft-label attacks under the BD(θ, σ) defense over the MNIST dataset.

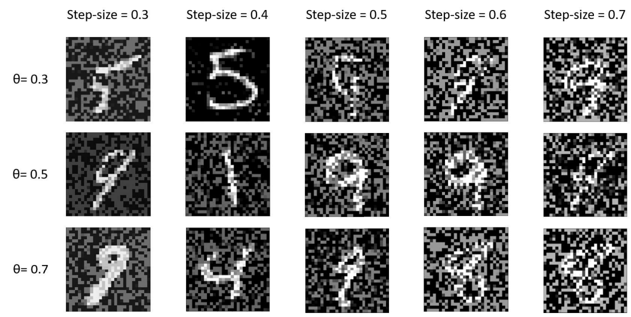


Fig. 8. Adversarial samples generated by the GenAttack with adaptive attack techniques (various optimization step-sizes) under the BD$(\theta, 0.1)$ defense. Note that the default step-size used in the GenAttack was 0.3 for the MNIST dataset.

B. Defense Performance against Untargeted Attacks

We evaluated the performance of the proposed BD method against untargeted attacks with 1000 validation images of MNIST and 100 validation images of IMAGENET. We considered the following untargeted black-box attacks: soft-label attacks - **AZ** [9], **SA**(Square Attack) [31], and **SimBA** [16]; hard-label attacks - **BA**(Boundary Attack) [7], **Sign-OPT** [13], and **HSJA**(HopSkipJump) [15]. Similar to the targeted case, we adopted their original source code with the default hyper-parameters and considered the l_2 norm setting throughout the experiment.

Besides evaluating the defense performance of our BD method against the above mentioned untargeted attacks, we have also compared the BD method's performance with other existing defense methods, including **NP** (Noise Perturbation) [44], **RND** (Random Noise Defense) [23], **GF** (Gaussian Augmentation Fine-Tuning) [23], **FD** (Feature Denoising) [23], and **AT** (Adversarial Training) [47].

An important aspect to consider for the untargeted attack case is the performance metric ASR. For the untargeted attacks, the attacker can simply add enough noise to the original sample to make the adversarial sample satisfy $\arg\max_i F_i(\mathbf{X}) \neq c$. Since all the defense methods in general make the attack algorithms to generate samples with heavy distortion or noise, the measure $\arg\max_i F_i(\mathbf{X}) \neq c$ tends to be satisfied by all the adversarial samples trivially. As a result, the first ASR definitions we used in the targeted attack case earlier in Section IV is not very meaningful. In this subsection,

TABLE IX
ASR (%) OF UNTARGETED SOFT-LABEL ATTACKS. σ = 0.1.

Dataset	Attacks	No defense	θ_1 = 0.5	θ_2 = 0.7
MNIST	AZ	97	0	0
	SA	92	89	10
	SimBA	82	31	0
Dataset	Attacks	No Defense	θ_1 = 0.3	θ_2 = 0.4
IMAGENET	AZ	100	0	0
	SA	100	38	30
	SimBA	100	96	0

TABLE X
COMPARISON OF THE BD METHOD WITH OTHER DEFENSE METHODS AGAINST UNTARGETED BLACK-BOX ATTACKS IN TERMS OF ACC DEGRADATION AND ASR

Dataset	Defense Method	Degradation of ACC (%)	SA ASR (%)	SimBA ASR (%)
IMAGE-NET	RND [23]	-1.9	51.9	38.8
	GF [23]	-0.2	95.8	88.8
	FD [23]	-20.8	51.8	38.7
	AT [47]	-13.4	47.2	35
	RND+AT [23]	-16.8	20	4.7
	NP [44]	-4.1	15	0
	BD(0.4, 0.1)	-2.0	30	0
	BD(0.6, 0.1)	-3.5	**15**	**0**

we use the second ASR definition, i.e., ASR2, which was defined in Section IV as the ratio of samples satisfying both $\arg\max_i F_i(\mathbf{X}) \neq c$ and median l_2 distortion less than a certain distortion threshold L.

To maintain a fair calculation of ASR with respect to the l_2 distortion threshold L for all the attack algorithms, we determined the threshold value L for all the attack methods from their no-defense results. Specifically, with their original source code and default parameters, we determined a minimum threshold that makes ASR to be near 100%. Then, we selected the L value to be multiple folds of this minimum threshold value. For example, for the AZ method we used distortion threshold $L = 0.5$ to calculate ASR for MNIST and $L = 1.5$ for IMAGENET. These L values were $10\times$ larger than the minimum threshold. The main objective was to guarantee almost 100% ASR for all the attack methods without defense. We used the same L value to calculate the ASR when the BD method was applied.

1) ASR of Soft-Label Untargeted Attacks: The ASR of the soft-label untargeted black-box attacks under our proposed BD method is summarized in Table IX. We have shown the experiment data for $\sigma = 0.1$ and varying θ values. We can see that when defending against untargeted soft-label attacks, the performance of the BD method indicated the same trend as what we observed for targeted attacks. The ASR degraded drastically with the increase of θ. BD$(0.7, 0.1)$ defended the MNIST DNN models against untargeted attacks as effectively as against targeted attacks. However, for the IMAGENET DNN models, as compared to the targeted attack case, we had to increase the threshold θ. This is not surprising considering the fact that, compared with the targeted attack, the untargeted attack is usually easier for the attacker to be successful and is thus harder for the defender to mitigate. BD$(0.4, 0.1)$ reduced the ASR of AZ and SimBA to 0, and reduced the ASR of SA to 30%. Increasing θ further to 0.6 would reduce the ASR of SA to 15% as shown in Table X. Note that such a moderate increase of θ would not lead to too large ACC degradation, as we have evaluated in Section IV-B.

2) ASR of Hard-Label Untargeted Attacks: We have summarized the ASR of hard-label untargeted attacks in the presence of the proposed BD method in Fig. 9. As observed in the targeted attack case, the ASR of the untargeted attack methods display a steady decline with the increase of θ.

3) Comparison with Other Defense Methods against Untargeted Attacks: We have compared the performance of the BD method with some existing defense methods for defending against untargeted black-box attacks. The performance comparison data are illustrated in Table X, where most of the data of the existing defense methods were obtained from [23]. We can see that with 2 to 3.5% ACC degradation, the BD method could reduce ASR to near 0. The ASR of the Square Attack was comparatively higher and all other defence methods were not effective except RND+AT, NP and our BD method. However, RND+AT suffered from a heavy ACC degradation of -16.8%, which means this defense method is not useful in practice. Our BD method could suppress the ASR to 15% using $\theta = 0.6$ and $\sigma = 0.1$ with only a small -3.5% ACC degradation. Thus, we can claim that, in comparison with other defense methods, our BD method demonstrated effective and superior performance to defend against untargeted attacks.

4) Visual Representation of Untargeted Attack Samples: The visual representation of the adversarial samples generated by the untargeted attack methods in the presence of the BD method is shown in Fig 10-12. Fig. 10 shows the results of the soft-label attack SA, where the left column is the original image $\mathbf{X}_0$, the right column is the adversarial image $\mathbf{X}$, and the middle column shows the distortion $\Delta\mathbf{x} = \mathbf{X} - \mathbf{X}_0$. Note that $\theta = 0$ means no defense is applied. In contrast, Fig. 11 and 12 are for the hard-label attacks, where the first column shows the original images $\mathbf{X}_0$, and the rest columns show the distortion $\Delta\mathbf{x}$.

From these figures, we can see that for untargeted attacks, the attacker needed higher distortion in order to generate a successful adversarial image under our BD method. With the increase of θ, the distortion level added by the attacker increased rapidly, especially for hard-label attacks, which resulted in highly degraded adversarial images.

C. Discussions & Limitations

While the experiment results we obtained strongly corroborate our main hypothesis that **addition of a small random noise to the logits when the query sample reaches close to the decision boundary will successfully defend black-box attacks**, we have also found some interesting observations which will be helpful for the future study of the

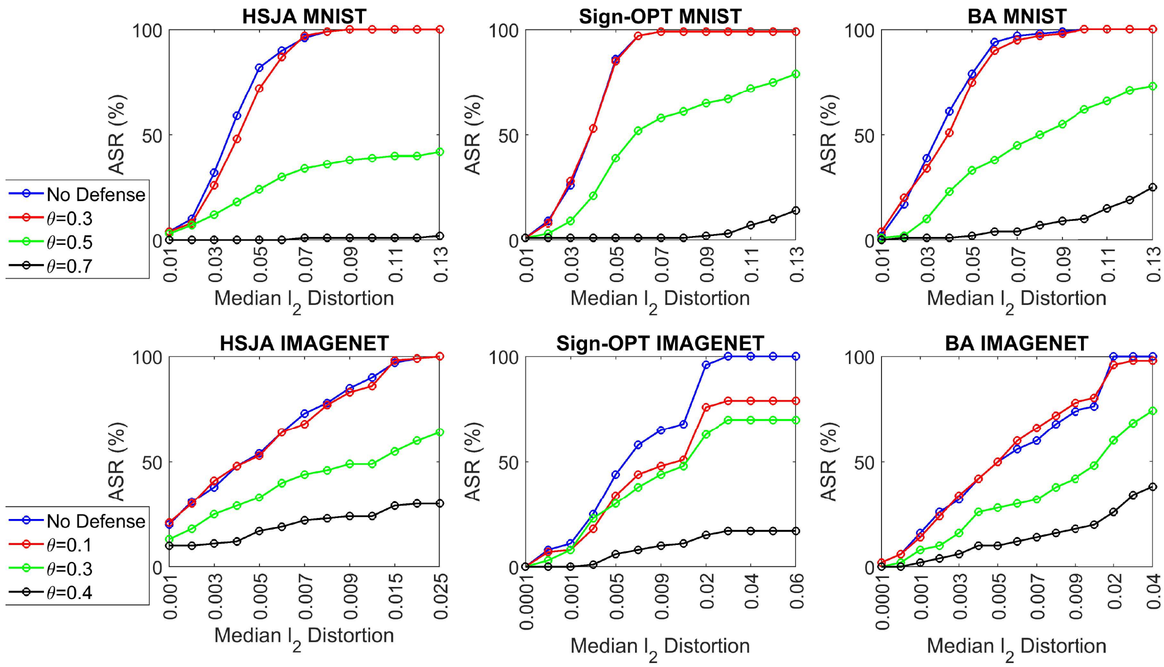


Fig. 9. ASR (%) vs. median l_2 distortion threshold L of untargeted hard-label attacks under the proposed BD method BD$(\theta, 0.1)$.

defense methods. Amongst various soft-label attacks that we experimented, we found that SimBA-DCT and NES-query-limited for CIFAR-10, and IMAGENET data, respectively, demonstrated the strongest performance against most of existing defense methods including our proposed BD method. Similarly, when evaluating the defense performance against hard-label attacks, we observed that the Boundary Attack and the HopSkipJump Attack performed comparatively stronger for datasets with small image sizes like MNIST & CIFAR-10 in the presence of the BD defense. However, from Table II, we can see that the BD method demonstrated superior performance against the HopSkipJump targeted attack over the IMAGENET dataset. We observed similar trend in the performance of the BD method against untargeted black-box attacks. While the untargeted attack setting is considered to be in the favor of the attacker, defending against untargeted attacks is considered to be tricky. The ASR of the untargeted attacks was comparatively higher than that of the targeted attacks for the same defense parameter setting. But when θ was increased, without too much ACC degradation, the ASR could still drastically reduce to almost 0%. Our proposed BD method has demonstrated robust performance against all the listed black-box attacks and we believe that the performance will be consistent for any other or new black-box attack methods. When compared with the existing defense methods, we observed that the BD method presented better and more reliable performance.

It will be very interesting to evaluate the augmented performance of the BD method in combination with some of the existing robust adversarial defenses. We leave this part of the study for future work. Note that our main focus of this paper is the query and search-based black-box attack method where both the model and dataset are unknown to the attacker. We do not cover the transfer-based attack method in this study (due to the inaccessibility of the training set used for training of the surrogate model in this scenario). We leave the study of the defense against transfer-based attack methods for future work.

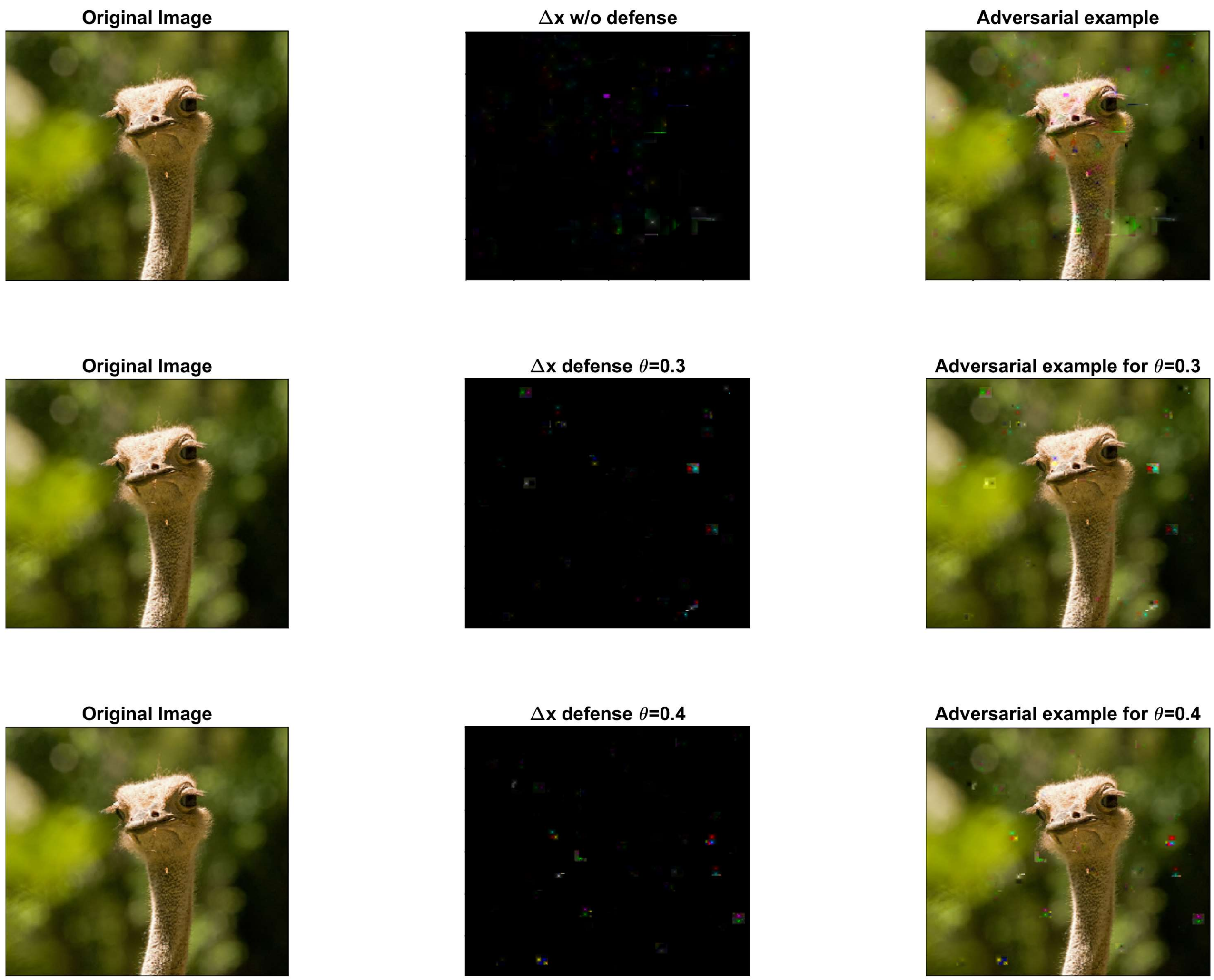


Fig. 10. Original images $\mathbf{X}_0$, distortions $\Delta\mathbf{x}$ and adversarial images $\mathbf{X}$ generated by the Square Attack (untargeted soft-label attack) under the proposed defense BD(θ, σ) for $\theta = \{0, 0.3, 0.4\}$, and $\sigma = 0.1$.

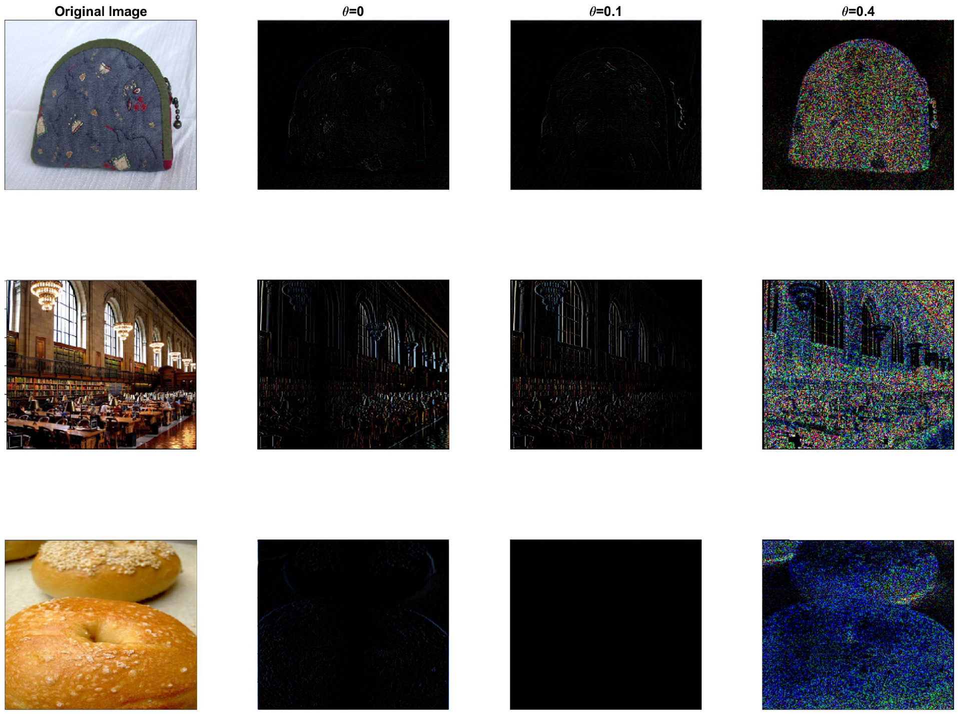


Fig. 11. Original images $\mathbf{X}_0$ and distortions $\Delta\mathbf{x}$ generated by the Boundary Attack (untargeted hard-label attack) under the proposed defense BD(θ, σ) for $\theta = \{0, 0.1, 0.4\}$, and $\sigma = 0.1$.

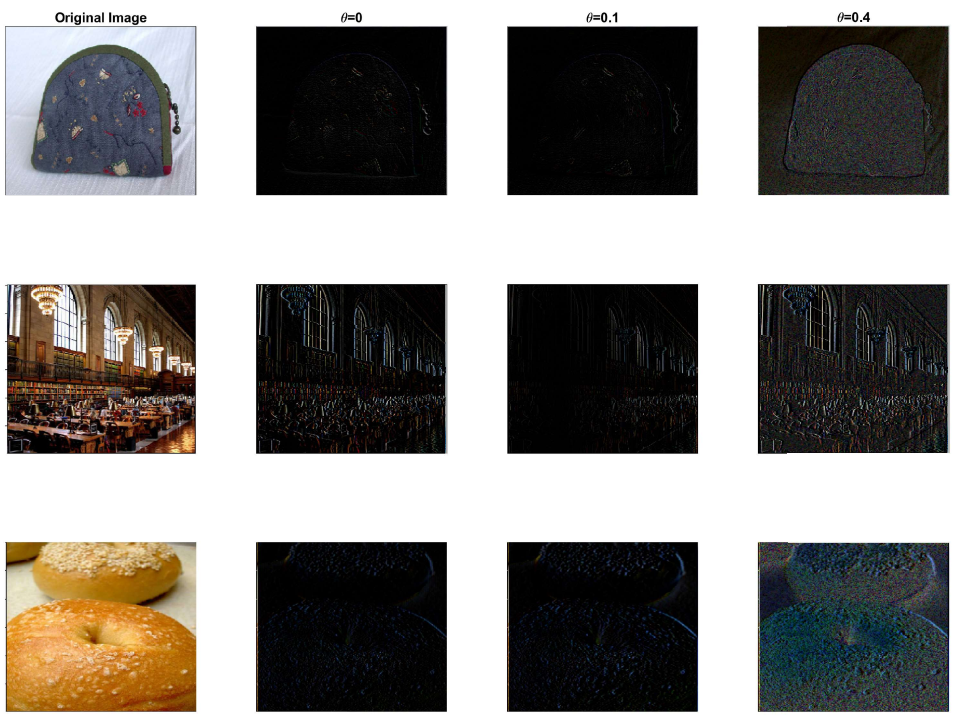


Fig. 12. Original images $\mathbf{X}_0$ and distortions $\Delta\mathbf{x}$ generated by the HopSkipJump Attack (untargeted hard-label attack) under the proposed defense BD(θ, σ) for $\theta = \{0, 0.1, 0.4\}$, and $\sigma = 0.1$.