Data-free Defense of Black Box Models Against Adversarial Attacks

Gaurav Kumar Nayak
University of Central Florida
Orlando, Florida, USA
gauravkumar.nayak@ucf.edu

Inder Khatri
New York University
New York, USA
ik2535@nyu.edu

Ruchit Rawal, Anirban Chakraborty
Indian Institute of Science
Bangalore, India
rawalruchit22@gmail.com,
anirban@iisc.ac.in

Abstract

*Several companies often safeguard their trained deep models (i.e. details of architecture, learnt weights, training details etc.) from third-party users by exposing them only as 'black boxes' through APIs. Moreover, they may not even provide access to the training data due to proprietary reasons or sensitivity concerns. In this work, we propose a novel defense mechanism for black box models against adversarial attacks in a data-free set up. We construct synthetic data via a generative model and train surrogate network using model stealing techniques. To minimize adversarial contamination on perturbed samples, we propose 'wavelet noise remover' (WNR) that performs discrete wavelet decomposition on input images and carefully select only a few important coefficients determined by our 'wavelet coefficient selection module' (WCSM). To recover the high-frequency content of the image after noise removal via WNR, we further train a 'regenerator' network with an objective to retrieve the coefficients such that the reconstructed image yields similar to original predictions on the surrogate model. At test time, WNR combined with trained regenerator network is prepended to the black box network, resulting in a high boost in adversarial accuracy. Our method improves the adversarial accuracy on CIFAR-*10 *by* 38.98% *and* 32.01% *against the state-of-the-art Auto Attack compared to baseline, even when the attacker uses surrogate architecture (Alexnet-half and Alexnet) similar to the black box architecture (Alexnet) with same model stealing strategy as defender.*

1. Introduction

Deep neural networks, applied in computer vision [40, 49], machine translation [3, 33], speech recognition [15, 29], have exhibited success but face unreliability due to adversarial attacks causing erroneous predictions [2, 22, 45, 48]. These attacks can be categorized into either black box [7, 18, 37, 51] and white box [9, 16, 26, 30] attacks based on access to model parameters. Black box attacks, more practical and realistic than white-box attacks, involve stealing the functionality of target models by training surrogate models using pairs of (image, predictions). Adversarial samples crafted using surrogate models can also exploit the property of transferability to attack the target model. Hence, immediate attention is required to protect against such attacks.

To make it harder for the adversary to craft black box attacks, companies prefer not to release the training dataset and keep them proprietary. However, recent works have shown that model stealing can compromise the confidentiality of black-box models even without the training data. Existing works perform generative modeling either with proxy data [4, 36, 43] or without proxy data [23, 47, 52] and train the surrogate model with the synthesized data for model stealing. However, their focus is more on obtaining highly accurate surrogate models. In contrast to existing works, we inquire about an important question regarding the safety of the black box models - "*how to defend against black box attacks in data-free (absence of training data) set up*?"

In order to tackle this problem, our proposed method '*DBMA*' (i.e., **d**efending **b**lack box **m**odels against **a**dversarial attacks in data-free setup) leverages the wavelet transforms [12]. We observe difference between wavelet transform on adversarial sample and original sample (shown in Fig. 1 (B)), and notice that detail coefficients in high-frequency regions (LH, HL and HH regions) are majorly corrupted by adversarial attacks and the approximate coefficients (LL region) is least affected for level 1 decomposition. Similar observation holds even for decomposition on other levels. To improve the adversarial accuracy, a naive way would be to completely remove the detail coefficients which can minimize the contamination in adversarial samples but it can lead to a huge drop in clean accuracy as the model predictions are highly correlated with high frequencies [50]. To avoid that, we assign importance to each of the detail coefficients based on magnitude. The least perturbed LL region usually contains higher magnitude coefficients than other regions (Fig. 1 (A)). So, we prefer high magnitude detail coefficients. However, taking a large number of such coefficients can lead to good clean accuracy but at the cost of low adversarial accuracy due to more contamination. On the other hand, taking very few such coefficients can allow lower contamination but results in low clean accuracy. Our method judiciously takes care of this trade-off, and carefully selects the required important detail coefficients (discussed in Sec. 4.2) using the proposed wavelet coefficient selection module (WCSM).

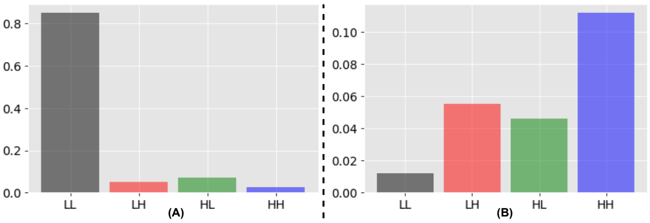

Figure 1. The average absolute magnitude of approximate (LL) and detail coefficients (LH, HL and HH) (via wavelet decomposition) across samples on a) clean data and b) normalized difference between wavelet decomposition of clean and corresponding adversarial image. In (a) the lesser contaminated LL coefficients have higher magnitude. In (b) LL are least affected.

The wavelet noise remover (WNR) removes noise coefficients by filtering out only the top-$k\%$ high magnitude coefficients where optimal k is selected using the WCSM module. As a side-effect, a lot of high-frequency content of the image gets lost which reduces the overall discriminability and ultimately results in suboptimal model's performance. To cope up with this reduced discriminability, we introduce a U-net-based regenerator network(Sec. 4.3), that takes the spatial samples corresponding to selected coefficients as input and outputs the reconstructed image. The regenerator network is trained by regularizing the feature and input space of the reconstructed image on the surrogate model. In the feature space, we apply cosine similarity and KL divergence losses to ensure proper reconstruction on clean and adversarial data respectively. Besides this, we also regularize the input space using our spatial consistency loss. Finally, the WNR module combined with the trained regenerator network is appended before the black-box model. The attacker has black-box access to the complete end to end model containing the defense components.

We summarize our contributions as follows:

- In this work, we investigate the largely underexplored yet critical challenge of defending against adversarial attacks on a *black box model without access to the network weights and in the absence of original training samples*.
- We propose a novel strategy to provide adversarial robustness against data-free black box attacks by introducing two key defense components:

  i.) We propose a wavelet-based noise remover (WNR) containing selective wavelet coefficient module (Sec. 4.2) that aims to remove coefficients corresponding to high frequency components, which are most likely to be corrupted by adversarial attack.

  ii.) We propose a U-Net-based regenerator network (Sec. 4.3) that retrieves the coefficients that are lost after the noise removal (via WNR) so that the clean high-frequency image content can be restored.
- We demonstrate the efficacy of our method via extensive experiments and ablations on both the components of proposed framework, viz., wavelet noise remover (Sec. 5.1, 5.2) and the regenerator network (Sec. 5.3), which are appended before the black box target model. The resulting combined model used as the new black box (as seen by attacker) yields high clean and adversarial accuracy on test data (Sec. 5.4).

2. Related Works

Our work is closely related to model stealing and wavelets, hence we briefly discuss their related works.

Data-efficient Model stealing: Based on the availability of training data, we categorize model stealing works below.
Training data - On full training data, knowledge distillation [21] is used to extract knowledge using soft labels obtained from the black box model. With few training samples, Papernot *et al.* [37] generates additional synthetic data in the directions (computed using jacobian) where model's output varies in the neighborhood of training samples.
Proxy data - In the absence of training data, either natural or synthetic images are used as proxy data. Orekondy *et al.* [36] query the black box model on the natural images using adaptive strategy via reinforcement learning to get output predictions and use them to replicate the functionality of the black box model. Barbalau *et al.* [4] use evolutionary framework to learn image generation on a proxy dataset where the generated images are enforced to exhibit high confidence on the black box model. Sanyal *et al.* [43] use the GAN framework with a proxy dataset composed of either related/unrelated data or synthetic data.
Without Proxy data - Kariyappa *et al.* [23] proposed an alternate training mechanism between generator and surrogate model, where generator is trained to produce synthetic samples to maximize the discrepancy between the predictions of the surrogate and the black box model. Truong *et al.* [47] also train generator and surrogate model alternatively but they replace discrepancy loss computed using KL divergence with L1 norm over logits approximated from softmax. Similarly, Zhou *et al.* [52] also formulate a min-max adversarial game but they additionally enforce the synthetic data to be equally distributed among the classes using a conditional generator.

Unlike these existing works, we use model stealing only as a means to obtain the surrogate model and synthetic data. Unlike them where they steal as an adversary, our goal is to provide robustness against black box attacks i.e. to reduce the effects of the adversarial samples on the black box model that are crafted using the surrogate model.

Wavelet in CNNs: Before the CNNs, the wavelets had been used for noise reduction and denoising [13, 14]. Prakash *et al.* [38] used pixel deflection technique followed by adaptive thresholding on the wavelets coefficients for denoising. Unlike these works, our setup is more challenging due to no access to both the training data and the model weights. Mustafa *et al.* [34] utilized wavelet denoising with image superresolution as a defense mechanism against adversarial attacks in grey-box settings. Different from this work, our approach utilizes the wavelet with a proposed regenerator network for defense against adversarial attacks in a black-box setting.

Data and Training Efficient Adversarial Defense: Adversarial Defense techniques can be broadly classified into two categories: Adversarial Training (AT) and Non-Adversarial Training (Non-AT) based methods. AT-based methods [8, 16, 28, 30] rely on adversarial samples during training to improve performance on perturbed samples. However, these methods are computationally expensive and often require high-capacity networks for achieving significant gains in robustness [53]. On the other hand, Non-AT-based methods like JARN[5], BPFC[1], and GCE[6] offer faster training but perform inadequately against a wide range of strong attacks [20].

In recent years, researchers have proposed adversarial defense methods that do not require re-training a model for providing robustness, thus making them train efficiently. This has immense practical benefits as unlike most state-of-the-art defenses that necessitate model re-training, such approaches can be seamlessly integrated with already deployed models. One such method, namely Magnet[32], first classifies inputs as clean or adversarial and then transforms the adversarial inputs closer to the clean image manifold. Another approach, Defense-GAN by Samangouei et al.[42], uses Generative Adversarial Networks to learn the distribution of clean images and generate samples similar to clean images from inputs corrupted with adversarial noise. Sun et al.[44] introduced the Sparse Transformation layer (STL) which maps input images to a low-dimensional quasi-natural image space, suppressing adversarial contamination and making adversarial and clean images indistinguishable. Theagarajan et al.[46] presented a defense protocol for black-box facial recognition classifiers consisting of a Bayesian CNN-based adversarial attack detector and image purifiers trained using the data from similar domain to the original training data. They used ensemble of image purifiers for removing the adversarial noise and attack detector for validating the purified image. However, a key limitation of these methods is their reliance on the original training-data for training the defense components, which hinders their use in scenarios where training-data/statistics are not freely available due to proprietary to privacy reasons, etc.

Consequently, recent advancements have redirected focus towards tackling the limitations of training data dependency in defending neural networks against adversarial attacks. For instance, Mustafa et al.[34] provide defense on pretrained network without retraining or accessing training data, but have additional dependency on pretrained image super-resolution networks. Moreover, their approach can only be integrated with the target networks that are capable of handling multi-scale inputs, thus making them infeasible on the pretrained networks incapable of handling multi-scale inputs. To avoid these problems, Qiu et al.[39] proposed the RDG (Random Distortion over Grids) preprocessing operation, randomly distorting input images by dropping and displacing pixels. Guesmi et al.[17] presented SIT (Stochastic Input Transformation), applying random transformations to eliminate adversarial perturbations while maintaining similarity to the original clean images. However, the stochastic nature of SIT without guidance negatively affects the clean accuracy of the target model. In contrast, our proposed Data-free Black-Box defense method, DBMA, better preserves clean-accuracy while also achieving higher robustness.

3. Preliminaries

Notations: The black box model is denoted by B_m which is trained on the proprietary training dataset O_d^{train}. We denote the surrogate model by S_m. The generator G produces synthetic data $S_d = \{x_s^i\}_{i=1}^N$ containing N samples. The logit obtained by the model S_m on input x is $S_m(x)$. The softmax and the label predictions on sample x by model S_m are represented by $soft(S_m(x))$ and $label(S_m(x))$.

The set $A_a = \{A_a^p\}_{p=1}^P$ contains P different adversarial attacks. The adversarial sample corresponding to the n^{th} sample of test dataset O_d^{test} (i.e. x_o^n) is denoted by x_{oa}^n which is crafted with a goal to fool the network B_m. Similarly, S_{da} is the set of crafted adversarial samples corresponding to the synthetic data S_d where the adversarial sample $x_{sa}^n \in S_{da}$ is obtained by perturbing the synthetic sample $x_s^n \in S_d$ using an adversarial attack $A_a^j \in A_a$.

We denote the discrete wavelet transform and its inverse operation by $DWT(.)$ and $IDWT(.)$ respectively. The wavelet coefficient selection module is denoted by WCSM. The regenerator network is represented by R_n.

Model Stealing: Model stealing involves extracting the black box model's (B_m) functionality by inputting images into its API to gather outputs, used subsequently to train a surrogate model (S_m). When no training data O_d^{train} is accessible, this scenario is termed data-free model stealing.

Adversarial Attacks: An adversarial attack $A_a^i \in A_a$ is a human-imperceptible noise ($\|\delta\|_\infty < \epsilon$) crafted to alter model's predictions on the perturbed sample (i.e. adversarial sample) x_{oa} from the original sample x_o. In black box adversarial attacks, the surrogate model S_m creates adversarial samples, transferable to the black box model B_m.

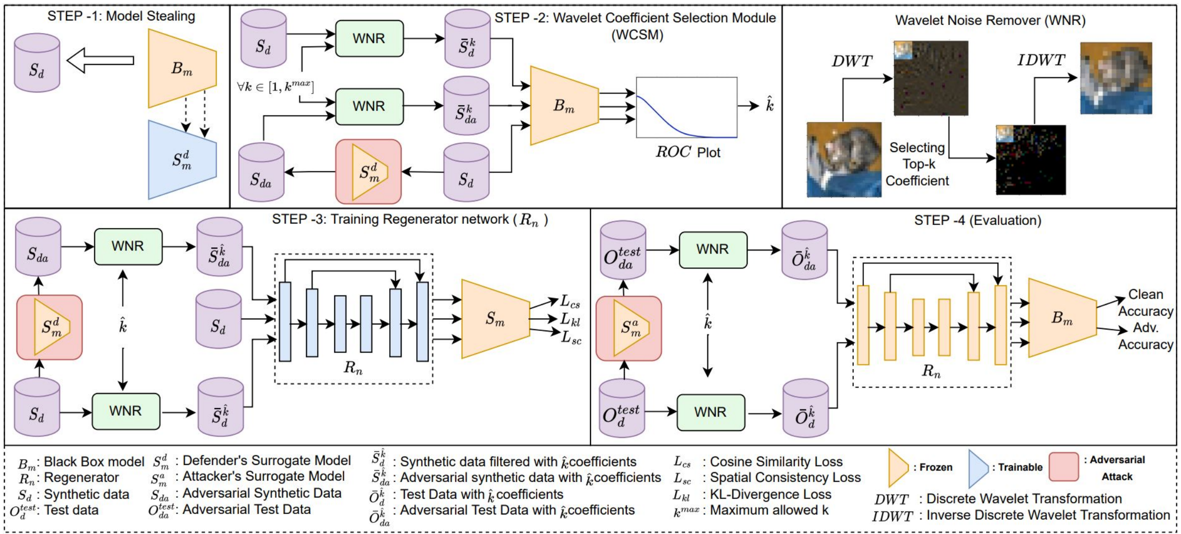
STEP -1: Model Stealing
S_d
B_m
S_m^d
STEP -2: Wavelet Coefficient Selection Module (WCSM)
S_d
WNR
$\bar{S}_d^k$
$\forall k \in [1, k^{max}]$
WNR
$\bar{S}_{da}^k$
B_m
$\hat{k}$
ROC Plot
S_{da}
S_m^d
S_d
Wavelet Noise Remover (WNR)
DWT
IDWT
Selecting Top-k Coefficient
STEP -3: Training Regenerator network (R_n)
S_{da}
WNR
$\bar{S}_{da}^k$
S_m^d
$\hat{k}$
S_d
S_d
WNR
$\bar{S}_d^k$
R_n
S_m
L_{cs}
L_{kl}
L_{sc}
STEP -4 (Evaluation)
O_{da}^{test}
WNR
$\bar{O}_{da}^k$
S_m^a
$\hat{k}$
O_d^{test}
WNR
$\bar{O}_d^k$
R_n
B_m
Clean Accuracy
Adv. Accuracy
B_m: Black Box model
R_n: Regenerator
S_d : Synthetic data
O_d^{test}: Test data
S_m^d : Defender's Surrogate Model
S_m^a : Attacker's Surrogate Model
S_{da} : Adversarial Synthetic Data
O_{da}^{test}: Adversarial Test Data
$\bar{S}_d^k$: Synthetic data filtered with $\hat{k}$coefficients
$\bar{S}_{da}^k$: Adversarial synthetic data with $\hat{k}$coefficients
$\bar{O}_d^k$: Test Data with $\hat{k}$ coefficients
$\bar{O}_{da}^k$: Adversarial Test Data with $\hat{k}$coefficients
L_{cs} : Cosine Similarity Loss
L_{sc} : Spatial Consistency Loss
L_{kl} : KL-Divergence Loss
k^{max} : Maximum allowed k
: Frozen
: Trainable
: Adversarial Attack
DWT : Discrete Wavelet Transformation
IDWT : Inverse Discrete Wavelet Transformation

Figure 2. An overview of our proposed approach DBMA. In step 1, we obtain the defender's surrogate model S_m^d and synthetic data S_d by model stealing from the victim model B_m. In step 2, we use the Wavelet Coefficient Selection Module (WCSM) that gives the optimal % of coefficients ($\hat{k}$) to be selected by the Wavelet Noise Remover (WNR) which are likely to be least corrupted by adversarial attacks. In step 3, we train a regenerator network R_n using different losses (L_{cs}, L_{kl}, L_{sc}) such that the model S_m^d yields features on the regenerated data (clean $R_n(\bar{S}_d^k)$ and adversarial $R_n(\bar{S}_{da}^k)$) similar to the features on clean data S_d. Finally in step 4, we evaluate our DBMA approach on test clean (O_d^{test}) and adversarial samples (O_{da}^{test}) where the WNR (with $k = \hat{k}$) and trained R_n are prepended to B_m.

Wavelet Transforms: Wavelets represent the time series signal using linear combinations of an orthogonal basis depending on which there are different types of wavelets such as Haar, Cohen and Daubechies [11]. The 2D DWT on an i^{th} image x^i for level 1 yields low pass subband (i.e approximation coefficients - denoting by LL_1^i) and high pass subband (i.e. detail-level coefficients - denoting by LH_1^i, HL_1^i, HH_1^i). In multi-level DWT, the approximation subband is further decomposed (for e.g. on a 2-level decomposition, LL_1^i is also decomposed to LL_2^i, LH_2^i, HL_2^i, HH_2^i).

4. Proposed Approach

In this section, we first discuss the model stealing method (Sec. 4.1) that we use to train the surrogate model S_m (as proxy for black box model B_m) and generate synthetic data S_d (as proxy for original training data O_d^{train}). Next, we propose our method to remove the detail coefficients (Sec. 4.2) that can be most corrupted by an adversarial attack and select the important coefficients to preserve the signal strength in terms of retaining feature discriminability. We dub this approach as wavelet noise remover (WNR) and the coefficients are selected using the wavelet coefficient selection module (WCSM). To boost the performance, we propose a U-Net-based regenerator network (Sec. 4.3) that takes the output of WNR as input and is trained to output a regenerated image on which the surrogate model would yield similar features as the features of the clean sample. The different steps involved in our proposed method (*DBMA*) for providing data-free adversarial defense in the black box settings are shown in Fig. 2.

4.1. Obtain Proxy Model and Synthetic Data

Given a black box model B_m, our first step is to obtain a proxy model S_m which can allow gradient backpropagation. To steal the functionality of B_m, S_m can be trained using a model stealing technique. But we also do not have access to the original training samples O_d^{train}. Hence, we use a data-free model stealing technique [4] that trains a generator using proxy data to produce synthetic samples (S_d) on which the black box model B_m gives high-confident predictions. The surrogate model S_m is then trained on synthetic data S_d under the guidance of B_m, where the model S_m is enforced to mimic the predictions of model B_m. The trained S_m and the generated data S_d are used in next steps.

4.2. Noise Removal with Wavelet Coefficient Selection Module (WCSM)

For an i^{th} sample of the composed synthetic data S_d (i.e. x_s^i), its corresponding wavelet coefficients are obtained by DWT operation on it. The approximate coefficients are the low frequencies that are least affected by the adversarial attack (shown in Fig. 1 (B)). Thus, we retain these coefficients. For e.g. on level-2 discrete wavelet decomposition, LL_2^i (approximate coefficients for i^{th} sample) is kept. As the adversarial attack severely harms the detail coefficients, we determine the coefficients that can be most affected by it using WCSM for effective noise removal.

Based on our observation that the least affected approximate coefficients often have high magnitude coefficients, indicating that the high magnitude detail coefficients can be a good measure for choosing which coefficients to retain. Thus, we arrange the detail coefficients based on magnitude (from high to low order) and retain the top-k % coefficients. The efficacy of choosing top-k compared to different baselines (such as random-k and bottom-k) is empirically verified in Sec.1. However, determining the suitable value of k is a challenge and, if not properly chosen, can lead to a major bottleneck in clean/adversarial performance. To handle it, we propose a wavelet coefficient selection mechanism that carefully selects the value of k so that decent performance can be obtained on both clean and adversarial data.

We empirically estimate optimal k using all the crafted synthetic training samples S_d and their corresponding adversarial counterparts. We define a quantity label consistency rate (LCR^k) which is calculated for a particular value of k using the following steps:

1. Construct adversarial synthetic samples ($\{x_{sa}^i\}_{i=1}^N$) by using an adversarial attack $A_a^j \in A_a$ on the surrogate model S_m.
2. Obtain approximate and detail coefficients for each adversarial synthetic sample using $DWT(x_{sa}^i, l), \forall i \in (1, \ldots, N)$ where l is the decomposition level.
3. Craft spatial samples ($\bar{S}_{da}^k = \{\bar{x}_{sa}^i\}_{i=1}^N$) using $IDWT$ operation on complete approximate and selected top-k detail coefficients corresponding to each adversarial synthetic sample (i.e. $S_{da} = x_{sa}^i, \forall i \in (1, \ldots, N)$). For simplicity, we envelop the operations (2) and (3) and name them 'wavelet noise remover' (WNR). In general, for a given input image and k value, WNR applies DWT on input where the approximate coefficients and the chosen top-k % detail coefficients are retained, whereas the non-selected detail coefficients are made to zeros. These coefficients are then passed to $IDWT$ to obtain the filtered spatial image.
4. Perform WNR by repeating the steps 2 and 3 on clean samples x_s to obtain $\bar{S}_d^k$.
5. Compare the predictions of black box model B_m on samples of $\bar{S}_d^k$ and the corresponding samples in S_d. LCR_C^k denotes the fraction of clean samples whose predictions match when top-k % coefficients are selected.
6. Compare predictions of black box model B_m on samples of $\bar{S}_{da}^k$ and the corresponding samples in S_d. LCR_A^k denotes fraction of adversarial samples whose predictions match when top-k % coefficients are selected.
7. Compute $LCR^k = LCR_C^k + LCR_A^k$
8. Calculate the rate of change of LCR^k (i.e. ROC^k) as ROC^k = LCR^{k+1} - LCR^k

Using above steps, we calculate LCR for different values of $k \in [1, \cdots, k^{max}]$. As shown in Fig. 3, we observe that as we increase the value of k, initially LCR_A increases and reaches its maximum value, and then it starts decreasing. High LCR_A implies that the predictions of the B_m model on $\bar{S}_{da}^k$ and S_d have a low mismatch. Similarly, with the increase in value of k, LCR_C keeps increasing which implies as we add more coefficients, model discriminability increases. The value of LCR increases with the value of k, but the rate of increase of LCR keeps decreasing. Refer Fig. 4 where we plot the rate of change (ROC). We choose $\hat{k}$ at which ROC saturates. Here ROC is negligible at $k = 16$. This value of k gives the best trade-off between clean and adversarial accuracy (empirically validated in Sec. 5.1). Thus, wavelet noise remover (WNR) is applied on input at estimated optimal k ($\hat{k}$).

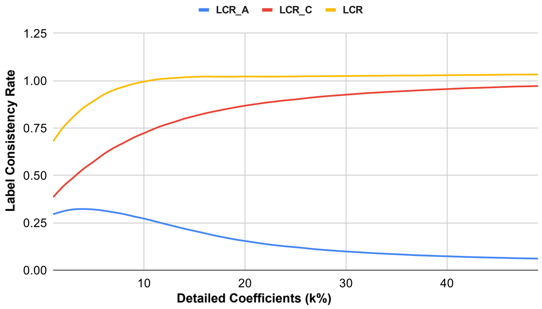

Figure 3. Label consistency rates (LCR_A , LCR_C and LCR) vs detail coefficients ($k\%$) plotted using prediction from black-box model B_m on CIFAR-10 dataset.

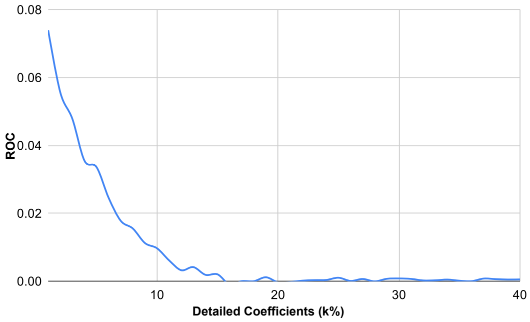

Figure 4. Rate of change (ROC) of LCR vs detail coefficients ($k\%$) plotted using prediction from black-box model B_m on CIFAR-10 data. As we increase value of k, ROC becomes negligible. At k = 16 it is close to zero.

4.3. Training of Regenerator Network

After the noise removal using the WNR at optimal k (i.e. $\hat{k}$ obtained by WCSM), there is also loss in the image signal information as a side effect. Thus, we further regenerate the coefficients using a regenerator network (R_n) that takes input as the output of WNR at $\hat{k}$ and yields reconstructed image which is finally passed to the black box model B_m for test predictions. The architecture of our R_n network is a U-net based architecture with skip connections which is inspired from [41]. For more details on the architecture of regenerator network, refer Supplementary (Sec.5).

$$\hat{k} = WCSM(S_d, S_m)$$
$$\bar{x}_s^i = WNR(x_s^i, \hat{k}); \;\; \bar{x}_{sa}^i = WNR(x_{sa}^i, \hat{k}) \tag{1}$$

For training, we feed the output obtained from the WNR at $\hat{k}$ as input to the network R_n and obtain a reconstructed image which is passed to the frozen surrogate model S_m for loss calculations. The losses used to train R_n are as follows:
Cosine similarity loss (L_{cs}): To enforce similar predictions from S_m on the regenerated synthetic sample and the corresponding original synthetic sample. $L_{cs} = CS(S_m(R_n(\bar{x}_s^i)), S_m(x_s^i))$.
KL divergence loss (L_{kl}): To align the predictions of S_m on the regenerated sample and its adversarial counterpart $L_{kl} = KL(soft(S_m(R_n(\bar{x}_{sa}^i))), soft(S_m(R_n(\bar{x}_s^i))))$.
Spatial consistency loss (L_{sc}): To make sure that the spatial reconstructed image (clean and adversarial) and the corresponding original synthetic image are similar in the image manifold $L_{sc} = \left\|R_n(\bar{x}_s^i) - x_s^i\right\|_1 + \left\|R_n(\bar{x}_{sa}^i) - x_s^i\right\|_1$.

Here, CS and KL denotes cosine similarity and KL divergence respectively. Overall loss used in training R_n:

$$L(R_n^\theta) = -\lambda_1 L_{cs} + \lambda_2 L_{kl} + \lambda_3 L_{sc} \qquad (2)$$

Finally, the black box model B_m is modified by prepending the WNR (with $k = \hat{k}$) and the trained R_n network to it. The resulting black box model defends the adversarial attacks which we discuss in detail in next section.

5. Experiments

In this section, we validate the effectiveness of our proposed method (DBMA) and perform ablations to show the importance of individual components. We use the benchmark classification datasets i.e. CIFAR-10 [24] and SVHN [35], on which we evaluate the clean and the adversarial accuracy against three different adversarial attacks (i.e., BIM [27], PGD[31] and Auto Attack [10]). Unless it is mentioned, we use Alexnet [25] as black box B_m (results on a larger black-box model are in Sec. 8 in supplementary) and Resnet-18 [19] as the defender's surrogate model S_m^d, which the defender uses to train the regenerator network R_n as explained in Sec. 4. In the black-box setting, attackers also do not have access to the black-box model's weights, thus restricting the generation of adversarial samples. So similar to the defender, we leverage the model stealing techniques to get a new surrogate model S_m^a, which the attacker uses for generating the adversarial samples. While evaluating against different attacks, we assume the attacker has access to defense components, i.e., the attacker uses model stealing methods to steal the functionality of the defense components along with the victim model.

We perform experiments with two different architectures for S_m^a: Alexnet-half and Alexnet, which are similar to the black-box model (Alexnet), making it tough for the defender. Ablation for different combination of S_m^d and S_m^a are in Sec .7 in supplementary. The attacker uses the same model stealing technique [4] as used by defender. It is important to note that our rigorous approach; we grant the attacker access to our exact model-stealing technique, architecture, and related details to ensure that any performance improvement is not attributed to differences in techniques or architecture between the attacker and defender. We use the Daubechies wavelet for both DWT and $IDWT$ operations. Refer to supplementary (Sec. 2) for experimental results on other wavelets. The decomposition level is fixed at 2 for all the experiments and ablations. The value of k^{max} is taken as 50. We assign equal weights to all the losses with weight as 1 (i.e. $\lambda_1 = \lambda_2 = \lambda_3 = 1$) in eq. 2.

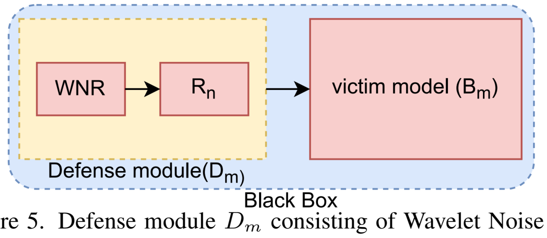

Figure 5. Defense module D_m consisting of Wavelet Noise Remover (WNR) and Regenerator R_n is prepended before the victim model B_m in our approach (DBMA). The D_m and B_m are combinedly considered as the black-box model by the attacker.

The defender constructs a defense module (D_m) using S_m^d. In subsections 5.1 and 5.2, the defense module only consists of the WNR , whereas subsections 5.3 onwards, both R_n and WNR are part of the defense module as shown in Fig. 5. We prepend the defense module before the B_m to create a new black box model that is used to defend against the adversarial attacks. To show the efficacy of defense components used in our method (DBMA), we consider the most challenging scenario, where the attacker uses the same model stealing technique as defender, and considers the defense module also a part of the black-box model while generating adversarial samples.

5.1. Ablation on quantity of coefficients

Table 1. Investigating the efficacy of our proposed WCSM in determining the quantity of detail coefficients to retain. The case (No-k) yields poor results justifying the need to preserve detail coefficients. Unlike, low and high values of k, we obtain better trade-off between clean and adversarial performance on our-k.

Amount of detail coefficients (k)	Black Box Model : Alexnet Surrogate Model (defense): Resnet-18			
	clean	BIM	PGD	Auto Attack
No k	31.19	5.88	4.68	9.35
low-k (k=1)	42.75	10.2	8.72	15.8
low-k (k=2)	50.17	15.37	14.14	21.92
low-k (k=4)	59.14	**17.54**	**16.03**	**25.08**
high-k (k=50)	82.58	5.58	3.33	10.44
our-k (k=16)	77.92	15.98	14.04	21.34

In Sec. 1 and 4, we discussed the importance of selecting the optimal number of detail coefficients ($\hat{k}$) and proposed the steps to find the value of $\hat{k}$ using the WCSM module. In this subsection, we do an ablation over the different choices for values of k (i.e., the number of detail coefficients to select) and analyze its impact on clean and adversarial accuracy. Specifically, we consider six distinct values of k across a wide range i.e. 0 (no detail coefficients, only approximate coefficients), $1, 2$ and 4(small k), 50(large k), $\hat{k}$ (optimal k given by WCSM). Fig. 4 shows the graph of the rate of change of LCR for different values of k. We select the value of k at which the ROC starts saturating, i.e. k with value 16 as $\hat{k}$. The results corresponding to the different values of k are shown in Table 1. We observe poor performance for both adversarial and clean samples when no detail coefficients are taken. On increasing the fraction of detail coefficients ($k = 1, 2, 4$), an increasing trend for both clean and adversarial performance is observed. Further, for a high value of k clean accuracy improves, but with a significant drop in the adversarial performance. Our choice of k (i.e., $\hat{k}$) indeed leads to better clean accuracy with decent adversarial accuracy, hence justifying the importance of the proposed noise removal using WCSM component.

5.2. Effect of wavelet noise remover with WCSM

In this section, we study the effect of prepending the WNR module to the black-box model B_m. WNR selects the approximate coefficients and optimal $\hat{k}\%$ detail coefficients (obtained by WCSM). It filters out the remaining detail coefficients, which helps in reducing the adversarial noise from the samples. We evaluate the performance of the B_m with and without the WNR Module and present the results in Table 2. When the attacker's surrogate model is Alexnet-half, adversarial accuracy improves by $\approx 19 - 22\%$ across attacks using the WNR module. Similarly, when the attacker's surrogate model is Alexnet, the adversarial accuracy improves by $\approx 11 - 12\%$.

Table 2. Our wavelet noise remover using WCSM yields improvement in adversarial accuracy with small drop in clean accuracy.

Surrogate model (attacker)	Noise Removal using WCSM	Black Box Model : Alexnet Surrogate Model (defense): Resnet-18			
		clean	BIM	PGD	Auto Attack
Alexnet-half	No	82.58	7.02	4.53	11.65
	(**Ours**) Yes	77.92	**26.66**	**24.55**	**34.02**
Alexnet	No	82.58	4.17	2.19	8.55
	(**Ours**) Yes	77.92	**15.98**	**14.04**	**21.34**

5.3. Ablation on losses

Until now, we performed experiments by using only the WNR defense module. Now, we additionally attach another defense module (R_n) to WNR (refer Fig 5). In this subsection, we perform ablation to demonstrate the importance of different losses used for training the Regenerator network R_n. As shown in eq. 2, the total loss L is the weighted sum of three different losses (i.e., L_{cs}, L_{kl}, and L_{sc}). To determine the effect of each of the individual losses, we train R_n using only the L_{cs} loss, L_{kl} loss and L_{sc} loss. Further to analyse the cumulative effect, we train R_n with different possible pairs of loss i.e., L_{cs} + L_{sc} loss, L_{cs} + L_{kl} loss, L_{kl} + L_{sc} loss, and finally with the total loss (L_{cs} + L_{kl} + L_{sc}) respectively. The results are displayed in Table 3.

Table 3. Contribution of different losses used for training R_n.The loss (L_{cs} + L_{kl} + L_{sc}) gives the best improvement in adversarial accuracy with decent clean accuracy.

Surrogate model (attacker)	Losses to train Regenerator network (R_n)	Black Box Model : Alexnet Surrogate Model (defense): Resnet-18			
		clean	BIM	PGD	Auto Attack
Alexnet-half	L_{cs}	78.96	26.33	24.75	33.81
	L_{sc}	78.85	27.38	25.75	35.51
	L_{kl}	9.82	6.56	6.54	8.98
	L_{cs} + L_{sc}	79.75	27.70	25.34	34.64
	L_{cs} + L_{kl}	62.06	36.03	35.93	43.05
	L_{kl} + L_{sc}	65.94	37.72	37.62	46.14
	L_{cs}+ L_{kl} + L_{sc}	73.77	**42.71**	**42.71**	**50.63**
Alexnet	L_{cs}	78.96	16.34	14.57	21.81
	L_{sc}	78.85	17.68	15.97	23.77
	L_{kl}	9.82	6.61	6.38	8.98
	L_{cs}+ L_{sc}	79.75	17.28	15.6	23.59
	L_{cs}+ L_{kl}	62.06	24.86	25.91	32.26
	L_{kl}+ L_{sc}	65.94	31.04	31.05	38.26
	L_{cs}+ L_{kl} + L_{sc}	73.77	**33.31**	**31.72**	**40.56**

Compared to the earlier best performance with WNR defense module (Table 2), we observe R_n trained with only L_{cs} loss gives no significant improvement in both the adversarial and clean accuracy. Similar trend is observed for R_n trained with L_{sc} loss. However, using only the L_{kl} loss shows a deteriorated clean and adversarial performance of R_n. Further, using the combination of both L_{cs} and L_{sc} loss also does not show much improvement. Combining the L_{kl} with L_{cs} and L_{sc} loss improves the adversarial performance of R_n appreciably, but with a drop in the clean accuracy. L_{kl} with L_{cs} loss shows a consistent improvement of $\approx 9 - 11\%$ in adversarial accuracy across attacks using both the Alexnet and Alexnet-half. Similarly, the combination of L_{kl} with L_{sc} loss improves the adversarial accuracy of R_n by $\approx 11 - 13\%$ and $\approx 15 - 17\%$ against the attacks using Alexnet-half and Alexnet respectively. When R_n is trained with all three losses gives the best adversarial accuracy across all the possible combinations. We observe an overall improvement of $\approx 16 - 19\%$ across different attacks using both Alexnet and Alexnet-half with a slight drop in clean accuracy ($\approx 4\%$). Regenerator network regenerates the lost coefficients, but as explained in section 4.2, detail coefficients also cause a decrease in adversarial accuracy. When we train a regenerator network using the combination of L_{cs}, L_{sc}, and L_{kl} loss, the L_{cs} and L_{sc} loss help to increase clean accuracy, but at the same time L_{kl} loss ensures regenerated coefficients do not decrease the adversarial accuracy. To achieve best tradeoff between clean and adversarial accuracy, R_n gets trained to increase adversarial accuracy at the cost of decreased clean accuracy compared to R_n network trained with only L_{cs} loss.

Table 4. Utility of each component used in our method (DBMA)- Wavelet Noise Remover (WNR) and Regenerator Network R_n on SVHN and CIFAR dataset. WNR with R_n yields huge gains in adversarial performance compared to baseline and WNR alone.

Surrogate Model (attacker)	Dataset	Method	Black Box Model : Alexnet Surrogate Model (defender) : Resnet-18			
			clean	BIM	PGD	Auto Attack
Alexnet-half	SVHN	Baseline	94.49	44.26	44.21	46.79
		SIT[17]	68.72	46.76 (↑ 2.5)	46.21 (↑ 2)	50.57 (↑ 3.78)
		RDG[39]	92.68	55.01 (↑ 10.75)	54.62 (↑ 10.41)	58.35 (↑ 11.56)
		WNR (**Ours**)	94.21	55.42 (↑ 11.16)	55.70 (↑ 11.49)	58.47(↑ 11.68)
		WNR+ R_n (**Ours**)	90.91	**68.63** (↑ 24.37)	**68.60** (↑ 24.39)	**71.71** (↑ 24.92)
	CIFAR-10	Baseline	82.58	7.02	4.53	11.65
		SIT[17]	51.16	22.05 (↑ 15.03)	21.68 (↑ 17.15)	29.08 (↑ 17.43)
		RDG[39]	67.58	19.99 (↑ 12.97)	18.97 (↑ 14.44)	29.64 (↑ 17.99)
		WNR (**Ours**)	77.92	26.66 (↑ 19.64)	24.55 (↑ 20.02)	34.02 (↑ 22.37)
		WNR+ R_n (**Ours**)	73.77	**42.71** (↑ 35.69)	**42.71** (↑ 38.18)	**50.63** (↑ 38.98)
Alexnet	SVHN	Baseline	94.49	38.14	38.19	40.16
		SIT[17]	68.72	43.48 (↑ 5.34)	43.77 (↑ 5.58)	47.63 (↑ 7.47)
		RDG[39]	92.68	50.28 (↑ 12.14)	50.29 (↑ 12.1)	53.79 (↑ 13.63)
		WNR (**Ours**)	94.21	48.98 (↑ 10.84)	49.02 (↑ 10.83)	51.49 (↑ 11.33)
		WNR+ R_n (**Ours**)	90.91	**63.13** (↑ 24.99)	**63.12** (↑ 24.93)	**66.18** (↑ 26.02)
	CIFAR-10	Baseline	82.58	4.17	2.19	8.55
		SIT[17]	51.16	18.64 (↑ 14.47)	18.43 (↑ 16.24)	26.23 (↑ 17.68)
		RDG[39]	67.58	15.54 (↑ 11.37)	14.50 (↑ 12.31)	24.01 (↑ 15.46)
		WNR (**Ours**)	77.92	15.98 (↑ 11.81)	14.04 (↑ 11.85)	21.34 (↑ 12.79)
		WNR+ R_n (**Ours**)	73.77	**33.31** (↑ 29.14)	**31.72** (↑ 29.53)	**40.56** (↑ 32.01)

5.4. Comparison with existing Data and Training efficient defense methods

In this subsection, we validate the efficacy of our proposed method DBMA by comparing it with two other state-of-the-art defense methods: SIT[17] and GD[39]. For comparison, we do experiments on two benchmark datasets, i.e., SVHN and CIFAR-10. We obtain the optimal $\hat{k}$ as 20 using the WCSM module for the black box model trained on the SVHN dataset. In Table 4, while defending with SIT, the adversarial accuracy improves by $\approx 2-3\%$ and $\approx 5-7\%$ across attacks crafted using different S_m^a (Alexnet and Alexnet-half) with a corresponding drop of $\approx 2-3\%$ in clean accuracy. On the other hand, defending with GD, improves adversarial accuracy by $\approx 2-3\%$ across attacks with a marginal drop of 1.66% in clean accuracy. While defending with only WNR in the defender module, the adversarial accuracy improves by $\approx 10-11\%$ across attacks crafted using different surrogate architectures S_m^a (Alexnet and Alexnet-half). The clean accuracy, however, experiences a minor drop of less than 1%. By utilizing both the WNR and R_n in the defender module, the adversarial performance further improves by $\approx 13-14\%$ across the attacks with a drop of $\approx 4\%$ in clean accuracy. Overall, we observe a gain of $\approx 24-26\%$ in adversarial accuracy compared to the baseline model, at the cost $\approx 5\%$ drop in clean accuracy. Similarly, for the CIFAR-10 dataset, we observe an overall improvement of $\approx 35-38\%$ and $\approx 29-32\%$ against attacks crafted using Alexnet-half and Alexnet, respectively. However, defending with SIT and RDG only improves the adversarial accuracy by $\approx 12-17\%$. Additionally, the clean accuracy drops by $\approx 31\%$ and $\approx 15\%$ when using SIT and RDG, respectively. In comparison, when using DBMA, we observed a relatively small drop of $\approx 8\%$ in clean accuracy, which is reasonable considering the challenging nature of our problem setup. Even in traditional adversarial training with access to full data, clean performance often drops at the cost of improving adversarial accuracy [31]. In our case, neither the training data nor the model weights are provided. Moreover, the black-box model is often obtained as APIs, and re-training the model from scratch becomes unfeasible. Considering these difficulties, the drop we observe on clean data is small with respectable overall performance.

6. Conclusion

We introduced DBMA, a novel defense strategy to defend black-box models from adversarial attacks without relying on training data. DBMA incorporates two defense components: a) Wavelet Noise Remover (WNR) that removes the most contaminated areas by adversarial attacks while preserving less affected regions b) a Regenerator network to restore lost information post WNR noise removal. Through various ablations and experiments, we demonstrated efficacy of each defense component. Our method DBMA significantly enhances robustness against data-free black box attacks across datasets and against diverse model-stealing methods. Similar to adversarial defenses in white-box setups, we observe that significant gains in robust accuracy come at the cost of a slight drop in clean accuracy. In future work, we plan to work on further mitigating this trade-off.

Acknowledgements: This work is partially supported by a Young Scientist Research Award (Sanction no. 59/20/11/2020-BRNS) to Anirban Chakraborty from DAE-BRNS, India.

References

[1] Sravanti Addepalli, Vivek B.S., Arya Baburaj, Gaurang Sriramanan, and R. Venkatesh Babu. Towards achieving adversarial robustness by enforcing feature consistency across bit planes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[2] Naveed Akhtar and Ajmal Mian. Threat of adversarial attacks on deep learning in computer vision: A survey. *Ieee Access*, 6:14410–14430, 2018. 1

[3] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, 2015. 1

[4] Antonio Barbalau, Adrian Cosma, Radu Tudor Ionescu, and Marius Popescu. Black-box ripper: Copying black-box models using generative evolutionary algorithms. *Advances in Neural Information Processing Systems*, 33:20120–20129, 2020. 1, 2, 4, 6

[5] Alvin Chan, Yi Tay, Yew Soon Ong, and Jie Fu. Jacobian adversarially regularized networks for robustness. In *International Conference on Learning Representations*, 2020. 3

[6] Hao-Yun Chen, Jhao-Hong Liang, Shih-Chieh Chang, Jia-Yu Pan, Yu-Ting Chen, Wei Wei, and Da-Cheng Juan. Improving adversarial robustness via guided complement entropy. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4881–4889, 2019. 3

[7] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017. 1

[8] Sihong Chen, Haojing Shen, Ran Wang, and Xizhao Wang. Towards improving fast adversarial training in multi-exit network. *Neural Networks*, 150:1–11, 2022. 3

[9] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 1

[10] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 6

[11] Ingrid Daubechies. Orthonormal bases of compactly supported wavelets. *Communications on Pure and Applied Mathematics*, 41(7):909–996, 1988. 4

[12] Ingrid Daubechies. *Ten Lectures on Wavelets*. Society for Industrial and Applied Mathematics, 1992. 1

[13] David L Donoho. De-noising by soft-thresholding. *IEEE transactions on information theory*, 41(3):613–627, 1995. 3

[14] David L Donoho and Jain M Johnstone. Ideal spatial adaptation by wavelet shrinkage. *biometrika*, 81(3):425–455, 1994. 3

[15] Haytham M Fayek, Margaret Lech, and Lawrence Cavedon. Evaluating deep learning architectures for speech emotion recognition. *Neural Networks*, 92:60–68, 2017. 1

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, 2015. 1, 3

[17] Amira Guesmi, Ihsen Alouani, Mouna Baklouti, Tarek Frikha, and Mohamed Abid. Sit: Stochastic input transformation to defend against adversarial attacks on deep neural networks. *IEEE Design & Test*, 39(3):63–72, 2021. 3, 8

[18] Lingguang Hao, Kuangrong Hao, Bing Wei, and Xue-song Tang. Boosting the transferability of adversarial examples via stochastic serial attack. *Neural Networks*, 150:58–67, 2022. 1

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 6

[20] Warren He, James Wei, Xinyun Chen, Nicholas Carlini, and Dawn Song. Adversarial example defenses: ensembles of weak defenses are not strong. In *Proceedings of the 11th USENIX Conference on Offensive Technologies*, pages 15–15, 2017. 3

[21] Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. 2

[22] Lifeng Huang, Chengying Gao, and Ning Liu. Defeat: Decoupled feature attack across deep neural networks. *Neural Networks*, 156:13–28, 2022. 1

[23] Sanjay Kariyappa, Atul Prakash, and Moinuddin K Qureshi. Maze: Data-free model stealing attack using zeroth-order gradient estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13814–13823, 2021. 1, 2

[24] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Canadian Institute for Advanced Research, 2009. 6

[25] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2012. 6

[26] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 1

[27] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018. 6

[28] Alex Lamb, Vikas Verma, Kenji Kawaguchi, Alexander Matyasko, Savya Khosla, Juho Kannala, and Yoshua Bengio. Interpolated adversarial training: Achieving robust neural networks without sacrificing too much accuracy. *Neural Networks*, 154:218–233, 2022. 3

[29] Jianjun Lei, Xiangwei Zhu, and Ying Wang. Bat: Block and token self-attention for speech emotion recognition. *Neural Networks*, 156:67–80, 2022. 1

[30] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learn-

ing models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 3

[31] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 6, 8

[32] Dongyu Meng and Hao Chen. Magnet: a two-pronged defense against adversarial examples. In *Proceedings of the 2017 ACM SIGSAC conference on computer and communications security*, pages 135–147, 2017. 3

[33] Chenggang Mi, Lei Xie, and Yanning Zhang. Improving data augmentation for low resource speech-to-text translation with diverse paraphrasing. *Neural Networks*, 148:194–205, 2022. 1

[34] Aamir Mustafa, Salman H Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29:1711–1724, 2019. 3

[35] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011. 6

[36] Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4954–4963, 2019. 1, 2

[37] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pages 506–519, 2017. 1, 2

[38] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8571–8580, 2018. 3

[39] Han Qiu, Yi Zeng, Qinkai Zheng, Shangwei Guo, Tianwei Zhang, and Hewu Li. An efficient preprocessing-based approach to mitigate advanced adversarial attacks. *IEEE Transactions on Computers*, 2021. 3, 8

[40] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural Computation*, 29:2352–2449, 2017. 1

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 5

[42] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018. 3

[43] Sunandini Sanyal, Sravanti Addepalli, and R Venkatesh Babu. Towards data-free model stealing in a hard label setting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15284–15293, 2022. 1, 2

[44] Bo Sun, Nian-hsuan Tsai, Fangchen Liu, Ronald Yu, and Hao Su. Adversarial defense by stratified convolutional sparse coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11447–11456, 2019. 3

[45] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations, ICLR 2014, Conference Track Proceedings*, 2014. 1

[46] Rajkumar Theagarajan and Bir Bhanu. Defending black box facial recognition classifiers against adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2020. 3

[47] Jean-Baptiste Truong, Pratyush Maini, Robert J Walls, and Nicolas Papernot. Data-free model extraction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4771–4780, 2021. 1, 2

[48] Petra Vidnerová and Roman Neruda. Vulnerability of classifiers to evolutionary generated adversarial examples. *Neural Networks*, 127:168–181, 2020. 1

[49] Athanasios Voulodimos, Nikolaos D. Doulamis, Anastasios D. Doulamis, and Eftychios E. Protopapadakis. Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*, 2018, 2018. 1

[50] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8684–8694, 2020. 1

[51] Hongzhi Zhao, Lingguang Hao, Kuangrong Hao, Bing Wei, and Xin Cai. Remix: Towards the transferability of adversarial examples. *Neural Networks*, 163:367–378, 2023. 1

[52] Mingyi Zhou, Jing Wu, Yipeng Liu, Shuaicheng Liu, and Ce Zhu. Dast: Data-free substitute training for adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 234–243, 2020. 1, 2

[53] Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 16443–16452, 2021. 3