This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

LEA^2: A Lightweight Ensemble Adversarial Attack via Non-overlapping Vulnerable Frequency Regions

Yaguan Qian *[1], Shuke He[1], Chenyu Zhao[1], Jiaqiang Sha[1], Wei Wang[2], and Bin Wang[3]

[1]School of Science, Zhejiang University of Science and Technology, Hangzhou, China
[2]Beijing Key Laboratory of Security and Privacy in Intelligent Transportation, Beijing Jiaotong University, China
[3]Zhejiang Key Laboratory of Multidimensional Perception Technology, Application and Cybersecurity, China

Abstract

Recent work shows that well-designed adversarial examples can fool deep neural networks (DNNs). Due to their transferability, adversarial examples can also attack target models without extra information, called black-box attacks. However, most existing ensemble attacks depend on numerous substitute models to cover the vulnerable subspace of a target model. In this work, we find three types of models with non-overlapping vulnerable frequency regions, which can cover a large enough vulnerable subspace. Based on this finding, we propose a lightweight ensemble adversarial attack named LEA^2, integrated by standard, weakly robust, and robust models. Moreover, we analyze Gaussian noise from the perspective of frequency and find that Gaussian noise is located in the vulnerable frequency regions of standard models. Therefore, we substitute standard models with Gaussian noise to ensure the use of high-frequency vulnerable regions while reducing attack time consumption. Experiments on several image datasets indicate that LEA^2 achieves better transferability under different defended models compared with extensive baselines and state-of-the-art attacks.

1. Introduction

Convolutional neural networks (CNNs) have been successfully employed in image classification, but recent works have shown that even the most advanced CNNs are vulnerable to adversarial examples [12, 29, 26, 1]. Adversarial attacks deliberately impose small perturbations to the benign input to mislead a model. In general, adversarial attacks can be divided into white-box attacks and black-box attacks. White-box attacks need to know all the information about the target model (*e.g.*, model structure, and model parameters), which is usually impracticable because an adversary cannot obtain all the information about the target model in reality. In contrast, black-box attacks do not require knowing the internal information of the target model, which can be further separated into query-based and transfer-based ones. Query-based attacks require extensive queries of the output of the target model [19, 21, 27, 32], which not only makes the target model suspect but also increases query costs. Transfer-based attack treats the target model as a pure black box, which crafts adversarial examples using a substi-

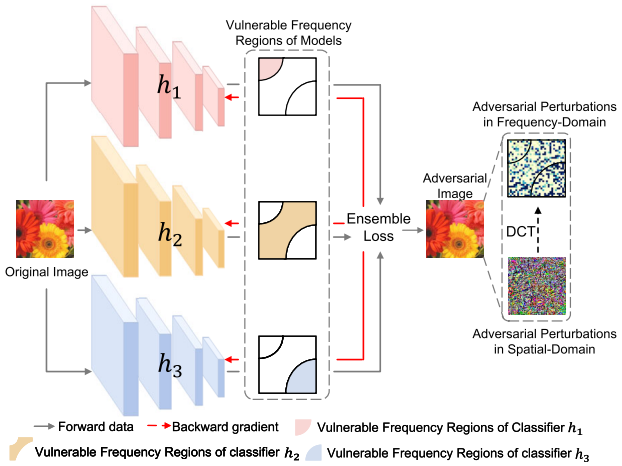


Figure 1. An example of our ensemble attacks, where three substitute models are used to craft adversarial examples. Classifier h_1 is an adversarial trained model, h_2 is a weakly robust model, and h_3 is a standard model.

*Corresponding author: qianyaguan@zust.edu.cn

tute model (white-box model) [10, 40, 7, 8, 23, 41].

In this paper, we focus on transfer-based black-box attacks. Depending on whether one or more substitute models are utilized, transfer-based attacks are further divided into single-model-based attacks and ensemble attacks. However, single-model-based attacks may lead adversarial examples to overfit the substitute model and decrease the attack success rate [16]. An efficient method to address this issue is to combine multiple substitute models. In order to approximate the target model, existing ensemble attacks have to use many models with different structures as substitute models [7, 8, 23, 41], which is generally intuitive and empirical. Moreover, training a large number of substitute models and crafting adversarial examples on them is seriously time-consuming [39, 24].

To facilitate the description of our ideas, we formalize the definition of vulnerable subspace, which is a set of adversarial examples. Each model has its special vulnerable subspace. The effectiveness of transfer-based attacks depends on how close the vulnerable subspace of the substitute model is to the target model. However, it is hard to characterize the vulnerable subspace due to its high dimensions. The transformation to the frequency domain makes it easier to study since the frequency domain of an image is two dimensions. Some previous work showed that adversarial perturbations added to the high-frequency regions of images are more effective than other frequency regions [35, 38], while other work showed that low-frequency regions are also vulnerable [13, 30].

In this work, we focus on the vulnerable frequency regions of different models. Our experiments showed that three types of models (standard, weakly robust, and robust) have distinct distributions of the vulnerable frequency regions. For standard models, they have vulnerable high-frequency regions; for weakly robust models, they have vulnerable mid-frequency regions; and for robust models, they have vulnerable low-frequency regions. Due to the reversibility of the Fourier Transform, we infer that the union of their corresponding vulnerable subspaces is large enough to cover that of the target model. Based on this assumption, we propose a lightweight ensemble adversarial attack, namely LEA^2, which only includes three types of substitute models regardless of the kind of target models. Figure 1 illustrates the process of our ensemble attack. Moreover, we find that Gaussian noise is located in the vulnerable frequency regions of standard models. Therefore, we use the Gaussian noise to replace standard models to reduce time consumption further. To sum up, our main contributions are summarized as follows:

- We investigate the weakly robust model in the frequency domain and find that the mid-frequency perturbations achieve the highest attack success rate.
- From the perspective of frequency, we construct an ensemble only including three types of substitute models, which significantly reduces the time cost and boosts the transferability of adversarial examples.
- Extensive experiments on three popular datasets demonstrate that LEA^2 significantly boosts the transferability of adversarial examples while reducing time consumption compared to state-of-the-art ensemble attacks.

2. Related Work

Transfer-based attacks are based on the fact that despite the substitute and target models adopting diverse architectures, they may share similar decision boundaries [8, 23]. According to the attack mechanism adopted by the adversary, there are two types of transfer-based attacks. One is the single-model-based attack [10, 8, 40], and the other one is the ensemble attack. Since the former tends to overfit the substitute models, we focus on ensemble attacks.

Ensemble attacks. Liu et al. [23] first propose an ensemble attack that prevents the noise from overfitting a single model architecture and thus bolsters the transferability. Dong et al. [7] further investigate three manners of organizing the base models and demonstrate that the ensemble of averaging logits outperforms the others for boosting the attack effectiveness. Hang et al. [15] propose two types of ensemble-based black-box attack strategies to produce adversarial examples with more powerful transferability. Xie et al. [41] propose an ensemble attack DI-FGSM by employing random transformations to the input examples to enhance the transferability. Dong et al. [8] shift the input to create a series of translated images and approximately estimate the overall gradient to mitigate the problem of over-reliance on the substitute model. Li et al. [22] apply feature-level perturbations to an existing model to potentially create a huge set of diverse models and propose a longitudinal ensemble method specifically for their networks. Xiong et al. [42] reduce the gradient variance among various models to boost ensemble attacks. Che et al. [2] divide a large number of pre-trained source models into several batches and introduce long-term gradient memories in their new ensemble algorithm for specific networks or tasks(e.g., pix-to-pix image translation). Long et al. [25] proposed a frequency domain data augmentation for training, and this can significantly improve transferability. However, these ensemble attacks have to use a large number of substitute models to craft adversarial examples.

Adversarial Examples in Frequency Domain. It has been increasingly common in recent years to investigate the essential characteristics of adversarial examples from the frequency domain. Tsuzuku & Sato [34] first proposed a frequency framework by studying the sensitivity of CNN's

for different Fourier bases. Wang et al. [35] show that high-frequency components play significant roles in promoting CNN'S accuracy and conclude that smoothing the CNN kernels helps to enforce the model to use features of low frequencies. Guo et al. [13] propose a low-frequency attack (LA) that successfully fools defended models, which shows that low-frequency components also play a significant role in model prediction. Deng & Karam [5] proposed a method of generating adversarial attacks in the frequency domain itself. Chen et al. [3] revealed that CNN classifiers rely on the amplitude spectrum of images rather than the phase spectrum, whereas humans rely more on the phase spectrum. However, these studies did not analyze the differences between the different types of models from the frequency perspective.

3. Motivation

Szegedy et al. [33] first observed the adversarial examples in DNNs. Let $\mathcal{X}$ be an input space and $\mathcal{Y}$ be a label set. A classification model is a mapping function $g_\theta : \mathcal{X} \to \mathcal{Y}$. Given an original image $x \in \mathcal{X}$ with the ground truth label $y \in \mathcal{Y}$, $g_\theta(x) = y$, where θ is the parameter of the classification model. The purpose of adversarial attacks is to find a tiny perturbation δ, fooling the classifier $g_\theta(x') \neq y$, where an adversarial example $x' = x + \delta$. In general, to ensure that the adversarial example is as similar as possible to the original image, δ is required to be less than a specific value ϵ as $\|x' - x\|_p \leq \epsilon$, where $\|\cdot\|$ is a norm and p could be 1, 2, ∞. In this work, we formally define vulnerable subspace as follows:

Definition 1 *Vulnerable Subspace. Given a perturbation budget ϵ, there exist a set $\mathcal{A} \subset \mathcal{X}$ such that $\mathcal{A} = \{x' | x' = x + \delta \wedge \|\delta\|_p \leq \epsilon \wedge g(x') \neq y\}$. We say $\mathcal{A}$ is the vulnerable subspace of the classification model g. For convenience, we denote it by $\mathcal{A}_g$.*

As shown in [33], a non-target attack is modeled as a constraint optimization problem:

$$\arg\max_{\delta} \mathcal{L}\left(g\left(x+\delta\right), y; \theta\right), \ s.t. \left\|\delta\right\|_\infty \leq \epsilon, \tag{1}$$

where $\mathcal{L}$ is the loss function. However, it is impractical to directly optimize Eqn (1) via the target model g under black-box settings because its parameter θ is inaccessible. To address this problem, a common approach is to train a local substitute model h simulating the target model g. The effectiveness (*i.e.*, transferability) relies on the overlap between the vulnerable subspace $\mathcal{A}_h$ and $\mathcal{A}_g$ of the substitute model h and the target model g. Due to the black-box property, the vulnerable subspace of the target model is unknown. An intuitive way [23] is to collect many substitute models $h_i, i = 1, 2, ..., M$ with different architectures to cover the target model's all possible vulnerable subspace, *i.e.*, $\mathcal{A}_g \subseteq \mathcal{A}_{h_1} \cup \mathcal{A}_{h_2} \cdots \cup \mathcal{A}_{h_M}$. Thus, ensemble attacks were proposed [7, 41, 24], which is modeled as the following optimization:

$$\arg\max_{x'} \quad -\log\left(\left(\sum_{i=1}^{M} w_i S_i\left(x'\right)\right) \cdot \mathbf{1}_y\right), \tag{2}$$

where $S_i(x')$ is the softmax outputs of the i-th substitute model, w_i is the ensemble weight with $w_i \geq 0$ and $\sum_{i=1}^{M} w_i = 1$, and $\mathbf{1}_y$ is the one-hot encoding of y.

However, collecting a large number of models is generally inefficient and time-consuming to train so many substitute models [24]. In this paper, we hope to cover the vulnerable subspace of the target model with as few substitute models as possible.

Unfortunately, it is difficult to accurately characterize a vulnerable subspace due to its high dimensions. Meanwhile, the target model is unknowable. In the frequency domain, no matter what kind of target model, its vulnerable subspace is in certain 2-D frequency regions. From this perspective, we construct a frequency-based ensemble attack.

4. Methodology

4.1. Vulnerable Frequency Regions

Fourier analysis provides another view to investigate the properties of images. Some previous works showed that adversarial perturbations added to the high-frequency regions of images are more effective than other frequency regions [35, 38]. In contrast, other works showed that low-frequency regions are also vulnerable to adversarial examples [13, 30]. Similar to the vulnerable subspace defined in the spatial domain, we formally define vulnerable frequency regions as follows:

Definition 2 *Vulnerable Frequency Regions. Given the model g's vulnerable subspace $\mathcal{A}_g$, there exists a vulnerable frequency region of g correspondingly in the frequency domain: $\mathcal{B}_g = \{f | x + \delta_f \in \mathcal{A}_g\}$ where δ_f is the specific perturbation corresponding to the frequency f.*

Let $\mathcal{D} : \mathcal{X} \to \mathcal{F}$ be the 2-D Discrete Cosine Transform [28] (DCT, details of which are included in Appendix A.1) and $\mathcal{D}^{-1}$ is its corresponding inverse. Here we present the formula of the specific frequency perturbation δ_f as follows:

$$\delta_f = \alpha \cdot \mathrm{Sgn}\left(\mathcal{D}^{-1}\left(\mathcal{D}\left(\nabla_x \mathcal{L}\right) \odot \mathcal{M}\right)\right), \tag{3}$$

where $\mathrm{Sgn}(\cdot)$ is a sign function, $\mathcal{M} = \begin{cases} 1, & \mathcal{D}(x) \in S \\ 0, & \mathcal{D}(x) \notin S \end{cases}$ is a mask to select frequencies, $S = Span\left\{f_1, f_2, \ldots, f_N\right\}$, where $f_1, f_2, \ldots, f_N$ are orthogonal DCT modes and $\mathcal{L}$ is

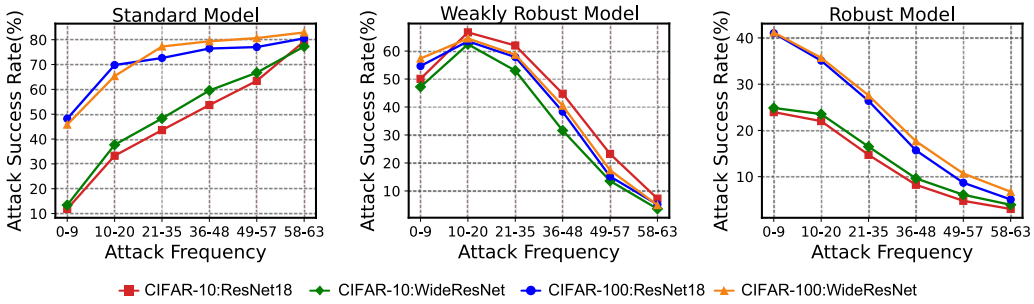


Figure 2. The attack success rate of δ_f on different models. Standard Model is standard trained ResNet18 (WideResNet), Weakly Robust Model uses PGD attack with $\epsilon = 4/255$, $\alpha = 2/255$ to train ResNet18 (WideResNet) for 20 epochs, and Robust Model uses PGD with $\epsilon = 8/255$, $\alpha = 2/255$ to train ResNet18 (WideResNet) for 50 epochs.

a loss function. When f selected by the mask $\mathcal{M}$ are frequency bands 36 to 63, δ_f are called the *high-frequency* perturbations, and when the selected f are frequency bands 10 to 35 and 0 to 9, then δ_f are called *mid-frequency* perturbations and *low-frequency* perturbations, respectively. The details of frequency bands are included in Appendix A.1.

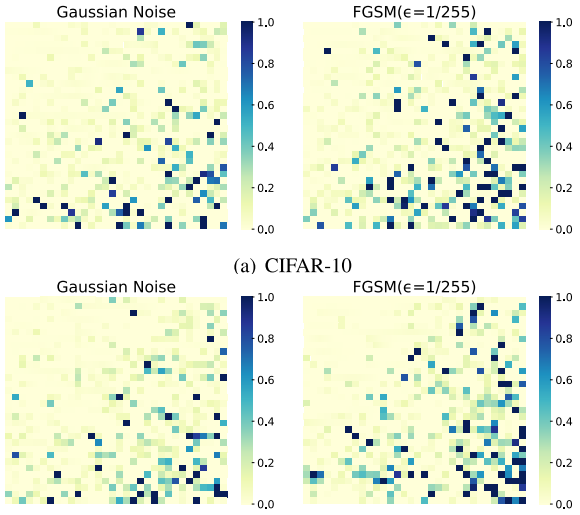
Our experiment shows that for *standard models*, high-frequency perturbations have a higher attack success rate; for *robust models*, the high-frequency perturbations can hardly fool it, while the low-frequency perturbations achieve a high attack success rate; for *weakly robust models*[1], the mid-frequency perturbations achieve the highest attack success rate (see Fig. 2). Though the standard model $h_{standard}$, weakly robust model h_{weak}, and robust model h_{robust} have the same structure, their vulnerable frequency regions are different but complementary. According to the vulnerable frequency regions, we further divide substitute models into standard, weakly robust, and robust for ensemble, equivalent to the ensemble of many randomly chosen models. Based on this idea, we construct a lightweight ensemble attack LEA[2] described in Section 4.3, and propose a remark that presumes:

Remark 1 *An ensemble of three types of models with non-overlapping vulnerable frequency regions,* i.e., $\mathcal{B}_{h_{standard}} \cap \mathcal{B}_{h_{weak}} \cap \mathcal{B}_{h_{robust}} = \phi$, *can achieve a large enough vulnerable subspace covered by an ensemble of many randomly chosen substitute models, i.e.,* $\mathcal{A}_{h_{standard}} \cup \mathcal{A}_{h_{weak}} \cup \mathcal{A}_{h_{robust}} \approx \mathcal{A}_{h_1} \cup \mathcal{A}_{h_2} \cdots \cup \mathcal{A}_{h_M}$, *where* $M \gg 3$.

4.2. Gaussian Noise Substitution

Previous research [6] indicated that Gaussian noise $r \sim N\left(0, \sigma^2\right)$ also has a significant impact on the classification performance of DNNs. Ford et al. [11] demonstrated that images with additive Gaussian noise and adversarial examples manifest the same underlying phenomenon. As

[1]The descriptions of the standard, robust, and weakly robust models are presented in Appendix A.2.



(a) CIFAR-10

(b) CIFAR-100

Figure 3. RCT maps for Gaussian noise and the FGSM adversarial examples [12] generated on the standard trained ResNet18. The upper left and lower right corners represent the lowest and highest frequency components in the DCT space, respectively. The deeper color indicates a greater change for a specific frequency component between the original and perturbed images, where $\sigma = 0.1$ for the Gaussian noise and the maximum perturbation $\epsilon = 1/255$ for FGSM.

we all know, generating adversarial examples is very time-consuming [39], which needs calculating the gradients by back-propagation, but generating the Gaussian noise only needs a random number generator. Based on this, we want to explore whether Gaussian noise can replace one of the three types of models above to reduce time consumption.

We identify the differences between the original image x and its perturbed x' in the frequency domain by calculating the average relative change of discrete cosine transform (RCT) [38]. It indicates which frequency regions the perturbations are mainly distributed in. RCT is defined as follows:

$$\mathrm{RCT} = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{\mathcal{D}\left(x_i'\right) - \mathcal{D}\left(x_i\right)}{\mathcal{D}\left(x_i\right)}\right|, \tag{4}$$

where $\mathcal{D}$ is the discrete cosine transform and N is the number of examples.

As shown in Figure 3, the Gaussian noise is mainly distributed in the lower right corner of the RCT map, and the FGSM perturbation [12] generated on the standard model is also primarily concentrated in the same regions, which means that Gaussian noise $r \sim N\left(0, \sigma^2\right)$ is located in the vulnerable frequency regions of the standard model (*i.e.*, $\mathcal{D}(r) = \mathcal{B}_{h_{standard}}$). More experiments of Gaussian noise are shown in Appendix B. According to these finds, we can substitute standard models with Gaussian noise to reduce time consumption while maintaining the transferability of adversarial examples, as shown in the Remark 2:

Remark 2 *According to Remark 1, there exists a set* $\mathcal{A}_r = \{x' | x' = x + r, r \sim N\left(0, \sigma^2\right)\} = \mathcal{A}_{h_{standard}}$*, and replacing the standard model with Gaussian Noise can achieve an equally large vulnerable subspace, i.e.,* $\mathcal{A}_r \cup \mathcal{A}_{h_{weak}} \cup \mathcal{A}_{h_{robust}} \approx \mathcal{A}_{h_1} \cup \mathcal{A}_{h_2} \cdots \cup \mathcal{A}_{h_M}$*, where* $M \gg 3$*.*

4.3. Implementation

Algorithm 1 LEA^2

Input An original image x with ground-truth label y; robust models $h^1_{robust}, h^2_{robust}, \ldots, h^{M_1}_{robust}$, the ensemble weights w_i; weakly robust models $h^1_{weak}, h^2_{weak}, \ldots, h^{M_2}_{weak}$, the ensemble weights w_j; size of perturbation ϵ; iterations T; step size α

Output An adversarial example x'

$x'_{t=0} \leftarrow x + r$, where $r \sim N\left(0, \sigma^2\right)$
for $t = 0$ to $T - 1$ **do**
  Input $x_t{}'$ to h^i_{robust} and get $\mathcal{L}\left(h^i_{robust}\left(x_t{}'\right), y\right)$ for $i = 1, 2, \ldots, M_1$
  Input $x_t{}'$ to h^j_{weak} and get $\mathcal{L}\left(h^j_{weak}\left(x_t{}'\right), y\right)$ for $j = 1, 2, \ldots, M_2$
  Fuse the loss as $\mathcal{L}\left(x_t{}', y\right) \leftarrow \sum_{i=1}^{M_1} w_i \mathcal{L}\left(h^i_{robust}\left(x_t{}'\right), y\right) + \sum_{j=1}^{M_2} w_j \mathcal{L}\left(h^j_{weak}\left(x_t{}'\right), y\right)$
  Obtain the gradient $\nabla_x \mathcal{L}\left(x_t{}', y\right)$
  Update x'_{t+1} by applying the sign gradients as $x'_{t+1} \leftarrow \mathrm{clip}_{x,\epsilon}\left\{x_t{}' + \alpha \cdot \mathrm{Sgn}\left(\nabla_x \mathcal{L}\left(x_t{}', y\right)\right)\right\}$
end for
return x'_T

The analysis in Section 4.1 needs three types of substitute models for Eqn (2). Meanwhile, from the analysis in Section 4.2, the example subspace with Gaussian noise added (*i.e.*, $\mathcal{A}_r$) is similar to the vulnerable subspace of the standard model (*i.e.*, $\mathcal{A}_{h_{standard}}$). Therefore, we replace $\sum_{k=1}^{M_3} w_k S^k_{standard}\left(x + \delta\right)$ in Eqn (2) with Gaussian noise $r \sim N\left(0, \sigma^2\right)$ to further reduce the time consumption. In brief, our lightweight ensemble adversarial attack (LEA^2) is to solve the following optimization problems:

$$\begin{aligned}\arg\max_{\delta} \quad -\log\Bigg(\Bigg(&\sum_{i=1}^{M_1} w_i S^i_{robust}(x + r + \delta) + \\ &\sum_{j=1}^{M_2} w_j S^j_{weak}(x + r + \delta)\Bigg) \cdot \mathbf{1}_y\Bigg),\end{aligned} \tag{5}$$

where $r \sim N\left(0, \sigma^2\right)$ is the Gaussian noise, M_1 and M_2 are the number of robust models and weak robust models respectively (M_1 and M_2 are usually 1 or 2), S_{robust} and S_{weak} represent the softmax outputs of the robust model and weak robust model respectively, $\sum_{i=1}^{M_1} w_i + \sum_{j=1}^{M_2} w_j = 1$. The adversarial example x' generated by the above optimization is also limited by $\|x' - x\|_\infty \leq \epsilon$. The procedure of LEA^2 is presented in Algorithm 1.

We compared the differences of perturbations generated by the black-box MI-FGSM attack [7], white-box PGD attack [26], and our attack LEA^2 using Eqn (4) on CIFAR-10, as shown in Figure 4. The perturbations generated by LEA^2 are distributed throughout the entire frequency regions regardless of the maximum perturbation $\epsilon = 8/255$ or the larger perturbation $\epsilon = 16/255$. In contrast, the perturbations generated by MI-FGSM and PGD are more concentrated in the high-frequency regions, and almost no perturbation is generated in the low-frequency domains. Since the perturbations generated by LEA^2 can cover entire frequency regions, LEA^2 will attack successfully no matter the target model's vulnerable subspace. This can be explained from another view that all possible target models, regardless of their specific form, have their vulnerable frequency regions located in some specific frequency regions. So we gain a more deep insight than Remark 1 as follows:

Remark 3 *Three types of models with non-overlapping but complementary vulnerable frequency regions are chosen as substitute models for the ensemble, which can generate adversarial perturbations that cover almost all possible vulnerable subspaces of target models. i.e.,* $\mathcal{A}_r \cup \mathcal{A}_{h_{weak}} \cup \mathcal{A}_{h_{robust}} \approx \bigcup_i \mathcal{A}_{g_i}$ *where* g_i *represents target models.*

5. Experiments

5.1. Experiment Setup

Datasets. We conduct experiments on three general datasets, namely CIFAR-10 [20], CIFAR-100 [20], and ImageNet-30 [4]. In particular, CIFAR-10 contains 50K training examples and 10K testing examples with the size

Table 1. The attack success rate of various attacks on standard models with JPEG compression [14] on CIFAR-10. The best results are indicated in bold. Other results on CIFAR-100 are included in Appendix C.4.

Dataset	Attack	ResNet20			VGG16		
		Clean	JPEG-75	JPEG-50	Clean	JPEG-75	JPEG-50
CIFAR-10	TI-FGSM [8]	58.41%	42.86%	37.40%	54.25%	36.93%	33.30%
	MI-FGSM [7]	94.83%	64.07%	32.07%	88.91%	66.04%	26.72%
	DI-FGSM [41]	97.54%	75.95%	53.42%	96.30%	70.72%	47.00%
	MI-FGSMens [7]	**99.52%**	83.21%	58.32%	97.85%	77.53%	53.89%
	DI-FGSMens [41]	99.43%	**90.36%**	75.59%	**98.81%**	87.31%	72.37%
	LEA^2(ours)	94.93%	88.79%	**83.89%**	91.46%	**87.66%**	**84.73%**

Table 2. The attack success rate of various attacks on standard models with JPEG compression [14] on ImageNet-30. The best results are indicated in bold.

Dataset	Attack	WideResNet101			DenseNet121		
		Clean	JPEG-75	JPEG-50	Clean	JPEG-75	JPEG-50
ImageNet-30	TI-FGSM [8]	83.41%	82.06%	81.70%	78.45%	78.62%	78.27%
	MI-FGSM [7]	86.53%	81.83%	78.16%	69.84%	66.56%	61.94%
	DI-FGSM [41]	90.40%	88.03%	87.49%	82.68%	78.47%	77.19%
	MI-FGSMens [7]	97.24%	95.59%	93.67%	78.15%	76.41%	73.56%
	DI-FGSMens [41]	**98.66%**	96.33%	95.32%	92.14%	91.67%	89.74%
	LEA^2(ours)	96.95%	**96.87%**	**96.53%**	**95.16%**	**95.11%**	**94.92%**

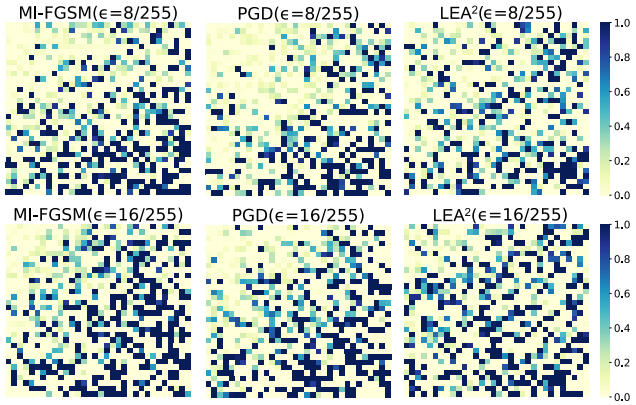


Figure 4. RCT map of various attack method on CIFAR-10. The upper left and lower right corners represent low and high frequency, respectively. More results on CIFAR-100 are included in Appendix C.3.

of 32×32 from 10 classes; CIFAR-100 has 100 classes, containing the same number of training (testing) examples as CIFAR-10; ImageNet-30 is a subset with 30 classes extracted randomly from the ImageNet dataset [4]. In experiments, we selected correctly classified images to evaluate various attacks, ensuring that the effectiveness of attacks is caused by the attacks themselves and not the model performance.

Models. For the standard models, we standardly train ResNet18, ResNet20, ResNet34 [17], WideResNet [44], and VGG16 [31] on CIFAR-10 and CIFAR-100 respectively. On ImageNet-30, we standardly train ResNet18, ResNet50, WideResNet101, Densenet121 [18], and VGG16. For the robust models, PGD [26] with $\epsilon = 8/255$ and $\alpha = 2/255$ is used for adversarial training for 50 epochs to obtain the robust models with two different structures, PGD-ResNet18 and PGD-WideResNet on CIFAR-10 and CIFAR-100. Trades [45] and Mart [37] are used to train for 50 epochs to obtain robust Trades-ResNet18 and Mart-ResNet18 on CIFAR-10 and CIFAR-100. On ImageNet-30, we train PGD-ResNet18, PGD-ResNet50, Trades-ResNet18, and Mart-VGG16 in the same way. For the weak robust models, we use PGD with $\epsilon = 4/255$ to train for 20 epochs to obtain Weak-ResNet18 on CIFAR-10, CIFAR-100, and ImageNet-30. The details of these models are provided in Appendix C.1.

Competitor. In order to show the effectiveness of our proposed LEA^2, we compare it with diverse state-of-the-art attack methods, including first-order attack FGSM [12], PGD [26], LA [13], MI-FGSM [7], DI-FGSM [41], TI-FGSM [8]. MI-FGSMens and DI-FGSMens represent the ensemble attacks mentioned in [7] and [41], respectively.

Implementation details. Among all attack methods, maximum perturbation $\epsilon = 16/255$, the iteration $T = 20$, and the step size $\alpha = 2/255$. For the convenience of comparison, white-box FGSM and PGD attacks are viewed as black-box attacks by using the substitute model (PGD-WideResNet) to craft adversarial examples when attacking various defended models on CIFAR-10 and CIFAR-100. Similarly, on ImageNet-30, Mart-VGG16 is used as the substitute model for FGSM and PGD. For the ensemble attacks MI-FGSMens and DI-FGSMens, we use two

standard models and two robust models as their substitute models according to the experiments in [7] and [41]. For CIFAR-10 and CIFAR-100, these models are PGD-WideResNet, Mart-ResNet18, ResNet18, and ResNet34, and the ensemble weights are set as 0.25 equally. These models are PGD-ResNet50, Mart-VGG16, ResNet18, and ResNet50 for ImageNet-30. For LEA^2, the Gaussian noise $r \sim N\left(0, \sigma^2\right)$ set $\sigma = 0.1$; the robust models on CIFAR-10/100 are PGD-WideResNet and Mart-ResNet18, while on ImageNet-30 are PGD-ResNet50 and Mart-VGG16; the weak robust model is Weak-ResNet18. The ensemble weights are set as 1/3 equally.

5.2. Attack Standard Trained Models

In this section, we present the performance of various black-box attacks on the standard models with JPEG defense [9] on CIFAR-10 and ImageNet-30 in Tables 1 and 2. JPEG is a defense method that removes high-frequency components to weaken adversarial examples and is often used in conjunction with models for defense [14]. TI-FGSM [8], MI-FGSM [7], and DI-FGSM [41] are first-order black-box attacks based on a single substitute model. MI-FGSMens [7] and DI-FGSMens [41] are ensemble attacks based on multiple substitute models. We use ResNet18 achieving 95.57% and 89.67% accuracy on CIFAR-10 and ImageNet-30, respectively, as their substitute model for the single-model-based black-box attacks. For ensemble attacks, the substitute models used by them are described in Section 5.1.

As shown in Tables 1 and 2, although the ensemble attacks MI-FGSMens and DI-FGSMens perform better against the completely defenseless standard models, the attack success rate drops by 9.07% to 41.20% in the presence of JPEG defense, whereas our attack method consistently maintains a high attack success rate (also maintains a high success rate against undefended standard models) that it only produces 0.05%$\sim$11.04% fluctuation with JPEG defense. These results further suggest that the perturbations generated by LEA^2 can cover the entire frequency region. Even if the high-frequency perturbations are removed by JPEG defense, the perturbations in the remaining frequency bands still play an important role. In contrast, the perturbations of other adversarial attacks are mainly concentrated in the high-frequency regions; therefore, the performance of adversarial examples is deeply weakened after JPEG compression.

5.3. Attack Defended Models

Although most of the attacks can easily fool standard models, they have a poor success rate when attacking the defensive models, especially in black-box settings. To further confirm the superiority of our attack, we first conduct a series of experiments on the advanced defensive models on CIFAR-10 and CIFAR-100 (see Table 3), including AT [26], Trades [45], JPEG [14], TVM [14], FS [43], and Spatial Smoothing [43]. FGSM, PGD, TI-FGSM, MI-FGSM, DI-FGSM, and LA are advanced first-order attacks based on a single model. In order to verify the transferability of these attacks, PGD-WideResNet is used as their substitute model. As described in Section 5.1, ensemble attacks MI-FGSMens and DI-FGSMens use two standard models and two robust models as their substitute models.

Table 3 shows the transferability of the above adversarial attacks on the advanced defended models. Compared with attacking standard trained models (see Tables 1 and 2), although the ensemble attacks MI-FGSMens and DI-FGSMens have high transferability when attacking the standard model, their performance on the defended models is deeply weakened. In contrast to them, our attack method LEA^2 has the highest attack success rate which is 3.8%$\sim$32.89% higher than other ensemble attacks. This is because LEA^2 produces perturbations covering entire vulnerable frequency regions, which can fool more target models regardless of where the target model's vulnerable subspace is located. More experiments on ImageNet-30 are provided in Appendix C.4.

For the comprehensive evaluation, more experiments are conducted on ImageNet-compatible dataset[2]. We compared our LEA^2 with three recent ensemble attacks: SVRE [42], VMI [36], and Ghost [22], where SVRE and VMI are generated on the ensemble of Res-101, IncRes-v2, Inc-v3, and Inc-v4, and Ghost is generated on the ensemble of Inc-v3, Inc-v4, IncRes-v2, IncRes-v2ens, and Res-v2-50. We also compared LEA^2 with a recent advanced frequency-based attack S^{2}I [25] generated on Adv-Inc-v3. As shown in Table 4, LEA^2 consistently outperforms the advanced ensemble attacks. The second column of Table 4 also shows that LEA^2 generates adversarial examples more efficiently than other advanced ensemble attacks [22, 42, 36]. Meanwhile, LEA^2 takes less time to generate adversarial examples than the frequency-based attack S^{2}I [25] with one model.

5.4. Ablation Study

In this section, we study the effect of Gaussian noise and mid-frequency perturbations on the transferability of our attack LEA^2. For the convenience of analysis, we only explored undefended ResNet20 and defended robust PGD-ResNet18 as our target models on CIFAR-10 and CIFAR-100.

Influence of Gaussian noise. As shown in the first two rows of each dataset in Table 5, we analyze the performance of LEA^2 when Gaussian noise $r \sim N\left(0, \sigma^2\right)$ is removed (*i.e.*, LEA^2 $- r$) or Gaussian noise is replaced by high-

[2] https://github.com/cleverhans-lab/cleverhans/tree/master/cleverhans_v3.1.0/examples/nips17_adversarial_competition/dataset

Table 3. The attack success rate of various attacks on advanced defenses models. The best results are indicated in bold.

Dataset	Attack	AT	Trades	JPEG-75	JPEG-50	TVM	FS	Spatial Smoothing
CIFAR-10	FGSM [12]	34.59%	33.44%	34.11%	34.00%	33.43%	34.56%	32.71%
	PGD [26]	47.66%	44.18%	45.02%	44.84%	39.79%	47.44%	41.48%
	TI-FGSM [8]	34.74%	35.10%	35.17%	36.65%	44.39%	34.47%	43.95%
	MI-FGSM [7]	46.13%	43.58%	44.76%	44.28%	40.14%	46.09%	41.62%
	DI-FGSM [41]	48.85%	47.00%	48.93%	50.26%	48.81%	48.68%	51.75%
	MI-FGSMens [7]	51.65%	45.96%	52.57%	50.59%	38.35%	41.15%	36.71%
	DI-FGSMens [41]	49.24%	52.65%	51.50%	49.76%	45.34%	43.00%	44.99%
	LA [13]	38.01%	36.43%	35.64%	35.97%	40.49%	38.13%	40.71%
	LEA^2(ours)	**59.40%**	**54.61%**	**57.31%**	**55.82%**	**50.21%**	**59.24%**	**54.13%**
CIFAR-100	FGSM [12]	57.42%	52.98%	56.27%	56.87%	53.57%	56.76%	55.20%
	PGD [26]	65.60%	59.18%	63.25%	61.99%	57.78%	63.83%	61.49%
	TI-FGSM [8]	50.17%	48.70%	51.91%	52.52%	55.97%	50.21%	55.64%
	MI-FGSM [7]	65.95%	60.54%	65.32%	64.24%	59.94%	64.71%	61.99%
	DI-FGSM [41]	62.92%	57.05%	63.89%	63.57%	64.13%	62.28%	64.57%
	MI-FGSMens [7]	62.14%	56.63%	39.83%	50.66%	48.86%	46.66%	43.15%
	DI-FGSMens [41]	59.73%	52.45%	32.75%	30.83%	42.62%	37.78%	38.84%
	LA [13]	57.57%	54.00%	56.63%	56.12%	60.36%	57.53%	60.86%
	LEA^2(ours)	**71.74%**	**65.92%**	**71.07%**	**69.55%**	**65.74%**	**71.64%**	**68.29%**

Table 4. The time of generating adversarial examples and black-box attack success rates (%) on ImageNet-compatible dataset. Among all attacks, maximum perturbation $\epsilon = 16/255$.

Attak	Time (min)	Adv-Inc-v3ens	Adv-Inc-v4ens	JPEG	TVM	FS
SVRE-MI-FGSM [42]	28.7	56.4	49.6	84.5	59.8	57.1
S^{2}I-FGSM [25]	24.3	31.6	30.4	83.1	40.4	35.1
VMI-FGSM [36]	27.7	36.6	32.9	79.5	50.4	52.5
Ghost-MI-FGSM [22]	118.4	42.9	41.2	71.7	48.5	44.7
MI-FGSMens [7]	30.2	48.8	40.6	70.8	46.3	51.8
DI-FGSMens [41]	34.8	52.4	45.2	74.2	57.7	55.3
LEA^2(ours)	**11.3**	**59.1**	**50.4**	**87.2**	**68.6**	**62.7**

frequency perturbations generated by the standard trained ResNet18 (*i.e.*, LEA^2 $-$ r $+$ ResNet18). (1) When Gaussian noise is removed, its success rate on the standard model decrease compared with LEA^2. (2) In the same way as state-of-the-art ensemble attacks select substitute models, the standard model is used as a substitute model rather than Gaussian noise (LEA^2 $-$ r $+$ ResNet18). When attacking the defended models, the attack success rate on CIFAR-10 and CIFAR-100 decreased by 17.93% and 16.12% respectively, and the attack time cost was dramatically raised since adding a substitute model. Therefore, it is effective to use Gaussian noise to replace the high-frequency perturbations generated based on the standard model.

Influence of mid-frequency vulnerable regions. According to the analysis in Section 4.1, Weak-ResNet18's vulnerable frequency regions are the mid-frequency regions. In order to examine whether mid-frequency vulnerable regions are useful for improving the transferability of adversarial examples, we conduct experiments on CIFAR-10 and CIFAR-100 that evaluate the attack's performance when LEA^2 removes Weak-ResNet18 (*i.e.*, LEA^2 $-$ Weak-ResNet18) replaced by standard trained ResNet18 (*i.e.*, LEA^2 $-$ Weak-ResNet18 $+$ ResNet18). As shown in the third and fourth rows of each dataset in Table 5, (1) when the mid-frequency vulnerable regions are removed, the attack success rate of attacking the standard model on CIFAR-10 and CIFAR-100 reduces by 24.99% and 13.09%, respectively. This is because the standard model not only relies on the high-frequency component for prediction but also the mid-frequency component plays an important role. (2) The transferability of adversarial examples on the adversarially trained model drastically reduces when the Weak-ResNet18 is replaced by the standardly trained ResNet18. It can be seen that mid-frequency vulnerable regions play a

Table 5. The time of generating adversarial examples and attack success rate (ASR) of LEA^2 using different substitute models on standard models and adversarially trained models.

Dataset	Attack	ResNet20		AT	
		Time (s)	ASR	Time (s)	ASR
CIFAR-10	LEA^2 $-$ r	418	86.86%	421	58.54%
	LEA^2 $-$ r + ResNet18	513	96.73%	512	41.47%
	LEA^2 $-$ Weak-ResNet18	325	66.94%	329	56.49%
	LEA^2 $-$ Weak-ResNet18 + ResNet18	418	94.57%	427	32.57%
	LEA^2	418	91.93%	427	59.40%
CIFAR-100	LEA^2 $-$ r	271	83.04%	271	71.68%
	LEA^2 $-$ r + ResNet18	360	93.75%	362	55.62%
	LEA^2 $-$ Weak-ResNet18	182	74.39%	184	70.72%
	LEA^2 $-$ Weak-ResNet18 + ResNet18	271	92.84%	273	46.12%
	LEA^2	271	90.48%	273	71.74%

Table 6. The attack success rate of existing ensemble attacks and ensemble attacks applying our ensemble strategy on the adversarially trained model. The best results are highlighted in bold.

Dataset	Attack	AT	Trades
CIFAR-10	MI-FGSMens [7]	41.19%	39.16%
	DI-FGSMens [41]	43.16%	41.11%
	LEA^2-MI-FGSMens(ours)	57.22%	52.18%
	LEA^2-DI-FGSMens(ours)	**58.25%**	**53.29%**
CIFAR-100	MI-FGSMens [7]	47.10%	44.87%
	DI-FGSMens [41]	38.13%	35.88%
	LEA^2-MI-FGSMens(ours)	**70.67%**	**64.52%**
	LEA^2-DI-FGSMens(ours)	67.12%	61.79%

significant role in boosting the transferability of adversarial examples.

5.5. Further Analysis

In this section, we analyze whether our ensemble strategy can be combined with existing ensemble attacks to significantly improve the transferability of adversarial examples. For convenience, we verify on advanced adversarial training defense, MI-FGSMens and DI-FGsmens are the same as the configuration in Section 5.1, LEA^2-MI-FGSMens and LEA^2-DI-FGSMens indicate that the Gaussian noise $r \sim N\left(0, \sigma^2\right)$ is added, PGD-WideResNet, Mart-ResNet18, and Weak-ResNet18 are used as substitute models. Applying our ensemble strategy to the existing ensemble attacks can significantly improve the transferability of adversarial examples on the robust model. As shown in Table 6, the attack success rate of LEA^2-MI-FGSMens increases by 13.02%$\sim$23.57% compared with MI-FGSMens, and LEA^2-DI-FGSMens increased by 12.18%$\sim$28.99% compared with DI-FGSMens.

6. Conclusion & Outlook

We find three types of models with non-overlapping vulnerable frequency regions, which can cover a large enough vulnerable subspace. Based on this finding, we propose a lightweight ensemble adversarial attack, LEA^2, integrated by standard, weakly robust, and robust models. In order to further reduce time consumption, we analyze Gaussian noise from the perspective of frequency and find that Gaussian noise is located in the vulnerable frequency regions of standard models. Therefore, we substitute standard models with Gaussian noise to ensure the use of high-frequency vulnerable regions while reducing attack time consumption. Compared with the black-box attacks and the ensemble attacks, extensive experiments demonstrate the significant effect of our method, which outperforms state-of-the-art transfer-based attacks by a large margin. Looking forward, more research directions about mid-frequency vulnerable regions could be exploited in future computer vision research. Meanwhile, we will explore how to apply vulnerable frequency regions on other tasks, such as the interpretability of DNNs. Moreover, effective ensemble defense strategies against LEA^2 will be another crucial and promising direction.

Acknowledgement

This work was supported by the National Natural Science Foundation of China under Grants 92167203 and U21A20463, Zhejiang Provincial Natural Science Foundation of China under Grant LZ22F020007, Foundation of Zhejiang Key Laboratory of Multi-dimensional Perception Technology Application and Cybersecurity under Grant HIK2022008, and the Science and Technology Innovation Foundation for Graduate Students of Zhejiang University of Science and Technology under Grant F464108M05.

References

[1] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57. IEEE Computer Society, 2017.

[2] Zhaohui Che, Ali Borji, Guangtao Zhai, Suiyi Ling, Jing Li, Xiongkuo Min, Guodong Guo, and Patrick Le Callet. SMGEA: A new ensemble adversarial attack powered by long-term gradient memories. *IEEE Trans. Neural Networks Learn. Syst.*, 33(3):1051–1065, 2022.

[3] Guangyao Chen, Peixi Peng, Li Ma, Jia Li, Lin Du, and Yonghong Tian. Amplitude-phase recombination: Rethinking robustness of convolutional neural networks in frequency domain. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, pages 448–457. IEEE, 2021.

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*, pages 248–255. IEEE Computer Society, 2009.

[5] Yingpeng Deng and Lina J. Karam. Frequency-tuned universal adversarial attacks on texture recognition. *IEEE Trans. Image Process.*, 31:5856–5868, 2022.

[6] Samuel F. Dodge and Lina J. Karam. A study and comparison of human and deep learning recognition performance under visual distortions. In *26th International Conference on Computer Communication and Networks, ICCCN 2017, Vancouver, BC, Canada, July 31 - Aug. 3, 2017*, pages 1–7. IEEE, 2017.

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 9185–9193. Computer Vision Foundation / IEEE Computer Society, 2018.

[8] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4312–4321. Computer Vision Foundation / IEEE, 2019.

[9] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M. Roy. A study of the effect of JPG compression on adversarial images. *CoRR*, abs/1608.00853, 2016.

[10] Lianli Gao, Qilong Zhang, Jingkuan Song, Xianglong Liu, and Heng Tao Shen. Patch-wise attack for fooling deep neural network. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVIII*, volume 12373 of *Lecture Notes in Computer Science*, pages 307–322. Springer, 2020.

[11] Justin Gilmer, Nicolas Ford, Nicholas Carlini, and Ekin D. Cubuk. Adversarial examples are a natural consequence of test error in noise. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 2280–2289. PMLR, 2019.

[12] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[13] Chuan Guo, Jared S. Frank, and Kilian Q. Weinberger. Low frequency adversarial perturbation. In Amir Globerson and Ricardo Silva, editors, *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pages 1127–1137. AUAI Press, 2019.

[14] Chuan Guo, Mayank Rana, Moustapha Cissé, and Laurens van der Maaten. Countering adversarial images using input transformations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[15] Jie Hang, Keji Han, Hui Chen, and Yun Li. Ensemble adversarial black-box attacks against deep learning systems. *Pattern Recognit.*, 101:107184, 2020.

[16] 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.

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016.

[18] Gao Huang, Zhuang Liu, Laurens van der Maaten, and Kilian Q. Weinberger. Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2261–2269. IEEE Computer Society, 2017.

[19] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Query-efficient black-box adversarial examples. *CoRR*, abs/1712.07113, 2017.

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[21] Huichen Li, Xiaojun Xu, Xiaolu Zhang, Shuang Yang, and Bo Li. QEBA: query-efficient boundary-based blackbox attack. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1218–1227. Computer Vision Foundation / IEEE, 2020.

[22] Yingwei Li, Song Bai, Yuyin Zhou, Cihang Xie, Zhishuai Zhang, and Alan L. Yuille. Learning transferable adversarial examples via ghost networks. In *The Thirty-Fourth*

AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020, pages 11458–11465. AAAI Press, 2020.

[23] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

[24] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, volume 13664 of *Lecture Notes in Computer Science*, pages 549–566. Springer, 2022.

[25] Yuyang Long, Qilong Zhang, Boheng Zeng, Lianli Gao, Xianglong Liu, Jian Zhang, and Jingkuan Song. Frequency domain model augmentation for adversarial attack. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part IV*, volume 13664 of *Lecture Notes in Computer Science*, pages 549–566. Springer, 2022.

[26] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[27] Ali Rahmati, Seyed-Mohsen Moosavi-Dezfooli, Pascal Frossard, and Huaiyu Dai. Geoda: A geometric framework for black-box adversarial attacks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8443–8452. Computer Vision Foundation / IEEE, 2020.

[28] Kamisetty Ramamohan Rao and Patrick C. Yip. *Discrete Cosine Transform - Algorithms, Advantages, Applications*. 1990.

[29] Huali Ren, Teng Huang, and Hongyang Yan. Adversarial examples: attacks and defenses in the physical world. *Int. J. Mach. Learn. Cybern.*, 12(11):3325–3336, 2021.

[30] Yash Sharma, Gavin Weiguang Ding, and Marcus A. Brubaker. On the effectiveness of low frequency perturbations. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 3389–3396. ijcai.org, 2019.

[31] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[32] Jiawei Su, Danilo Vasconcellos Vargas, and Kouichi Sakurai. One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput.*, 23(5):828–841, 2019.

[33] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun, editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[34] Yusuke Tsuzuku and Issei Sato. On the structural sensitivity of deep convolutional networks to the directions of fourier basis functions. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 51–60. Computer Vision Foundation / IEEE, 2019.

[35] Haohan Wang, Xindi Wu, Zeyi Huang, and Eric P. Xing. High-frequency component helps explain the generalization of convolutional neural networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 8681–8691. Computer Vision Foundation / IEEE, 2020.

[36] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 1924–1933. Computer Vision Foundation / IEEE, 2021.

[37] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[38] Zifan Wang, Yilin Yang, Ankit Shrivastava, Varun Rawal, and Zihao Ding. Towards frequency-based explanation for robust CNN. *CoRR*, abs/2005.03141, 2020.

[39] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[40] Weibin Wu, Yuxin Su, Xixian Chen, Shenglin Zhao, Irwin King, Michael R. Lyu, and Yu-Wing Tai. Boosting the transferability of adversarial samples via attention. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 1158–1167. Computer Vision Foundation / IEEE, 2020.

[41] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2730–2739. Computer Vision Foundation / IEEE, 2019.

[42] Yifeng Xiong, Jiadong Lin, Min Zhang, John E. Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 14963–14972. IEEE, 2022.

[43] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*. The Internet Society, 2018.

[44] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *CoRR*, abs/1605.07146, 2016.

[45] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 2019.