Frequency Domain Model Augmentation for Adversarial Attack

Yuyang Long[1], Qilong Zhang[1], Boheng Zeng[1], Lianli Gao[1], Xianglong Liu[2], Jian Zhang[3], and Jingkuan Song[1]*

[1] Center for Future Media, University of Electronic Science and Technology of China
[2] Beihang University [3] Hunan University
yuyang.long@outlook.com,{qilong.zhang,boheng.zeng}@std.uestc.edu.cn,
lianli.gao@uestc.edu.cn,xlliu@buaa.edu.cn,jianzh@hnu.edu.cn,
jingkuan.song@gmail.com

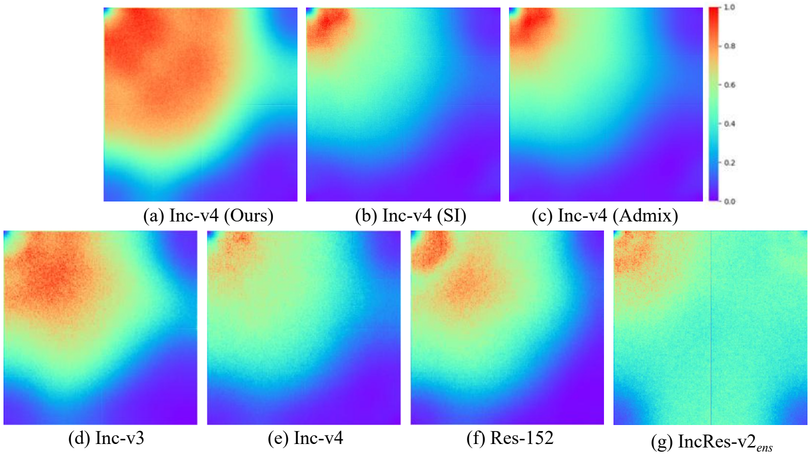
Abstract. For black-box attacks, the gap between the substitute model and the victim model is usually large, which manifests as a weak attack performance. Motivated by the observation that the transferability of adversarial examples can be improved by attacking diverse models simultaneously, model augmentation methods which simulate different models by using transformed images are proposed. However, existing transformations for spatial domain do not translate to significantly diverse augmented models. To tackle this issue, we propose a novel spectrum simulation attack to craft more transferable adversarial examples against both normally trained and defense models. Specifically, we apply a spectrum transformation to the input and thus perform the model augmentation in the frequency domain. We theoretically prove that the transformation derived from frequency domain leads to a diverse spectrum saliency map, an indicator we proposed to reflect the diversity of substitute models. Notably, our method can be generally combined with existing attacks. Extensive experiments on the ImageNet dataset demonstrate the effectiveness of our method, *e.g.*, attacking nine state-of-the-art defense models with an average success rate of **95.4%**. Our code is available in https://github.com/yuyang-long/SSA.

Keywords: Adversarial examples, Model augmentation, Transferability

1 Introduction

In recent years, deep neural networks (DNNs) have achieved a considerable success in the field of computer vision, *e.g.*, image classification [15,16,55], face recognition [40,43] and self-driving [2,34]. Nevertheless, there are still many concerns regarding the stability of neural networks. As demonstrated in prior works [39,12], adversarial examples which merely add human-imperceptible perturbations on clean images can easily fool state-of-the-art DNNs. Therefore, to help improve the robustness of DNNs, crafting adversarial examples to cover as many blind spots of DNNs as possible is necessary.

*Corresponding author



(a) Inc-v4 (Ours) (b) Inc-v4 (SI) (c) Inc-v4 (Admix)

(d) Inc-v3 (e) Inc-v4 (f) Res-152 (g) IncRes-v2$_{ens}$

Fig. 1: Visualization of the spectrum saliency maps (average of all images) for normally trained models Inc-v3 [38], Inc-v4 [37], Res-152[15] and defense model IncRes-v2$_{ens}$ [42]. **(a)**: the result for our transformation images ($N = 5$) conducted in frequency domain. **(b$\sim$c)**: the result for scale-invariant ($m_1 = 5$) [22] and Admix ($m_1 = 5$, $m_2 = 3$) [46] transformations conducted in spatial domain. **(d$\sim$g)**: the results for raw images on four different models.

In general, settings of adversarial attacks can be divided into white-box and black-box. For the former [3,28,59,26], an adversary has access to the model, *e.g.*, model architecture and parameters are known. Therefore, adversarial examples can be directly crafted by the gradient (w.r.t. the input) of the victim model, and thus achieving a high success rate. However, white-box attack is usually impracticable in real-world applications because an adversary is impossible to obtain all information about a victim model. To overcome this limitation, a common practice of black-box attacks [4,53,10] turns to investigate the inherent cross-model transferability of adversarial examples. Typically, an adversary crafts adversarial examples via a substitute model (*a.k.a.* white-box model), and then transfers them to a victim model (*a.k.a.* black-box model) for attacking.

However, the gap between the substitute model and the victim model is usually large, which manifests as the low transferability of adversarial examples. Although attacking diverse models simultaneously can boost the transferability, collecting a large number of diverse models is difficult and training a model from scratch is also time-consuming. To tackle this issue, *model augmentation* [22] is proposed. In particular, typical model augmentation approaches [53,5,22] aim to simulate diverse models by applying loss-preserving transformations to inputs. Yet, all of existing works investigate relationships of different models in spatial domain, which may overlook the essential differences between them.

To better uncover the differences among models, we introduce the spectrum saliency map (see Sec. 3.2) from a frequency domain perspective since representation of images in this domain have a fixed pattern [1,54], *e.g.*, low-frequency components of an image correspond to its contour. Specifically, the spectrum saliency map is defined as the gradient of model loss function w.r.t. the frequency spectrum of input image. As illustrated in Figure 1 (d∼g), spectrum saliency maps of different models significantly vary from each other, which clearly reveals that each model has different interests in the same frequency component.

Motivated by this, we consider tuning the spectrum saliency map to simulate more diverse substitute models, thus generating more transferable adversarial examples. To that end, we propose a spectrum transformation based on (discrete cosine transform) DCT and (inverse discrete cosine transform) IDCT techniques [1] to diversify input images. We theoretically prove that this spectrum transformation can generate diverse spectrum saliency maps and thus simulate diverse substitute models. As demonstrated in Figure 1 (a∼c), after averaging results of diverse augmented models, only our resulting spectrum saliency map can cover almost all those of other models. This demonstrates our proposed spectrum transformation can effectively narrow the gap between the substitute model and victim model. To sum up, our main contributions are as follows:

1) We discover that augmented models derived from the spatial domain transformations are not significantly diverse, which may limit the transferability of adversarial examples.

2) To overcome this limitation, we introduce the spectrum saliency map (based on a frequency domain perspective) to investigate the differences among models. Inspired by our finds, we propose a novel Spectrum Simulation Attack to effectively narrow the gap between the substitute model and victim model.

3) Extensive experiments on the ImageNet dataset highlight the effectiveness of our proposed method. Remarkably, compared to state-of-the-art transfer-based attacks, our method improves the attack success rate by 6.3%∼12.2% for normally trained models and 5.6%∼23.1% for defense models.

2 Related Works

2.1 Adversarial Attacks

Since Szegedy *et al.* [39] discover the existence of adversarial examples, various attack algorithms [12,18,3,28,32,27,59,30,6,20,58,50,56,57,24] have been proposed to investigate the vulnerability of DNNs. Among all attack branches, FGSM-based black-box attacks [12,18,4,53,10,11,49,9] which resort to the transferability of adversarial examples are one of the most efficient families. Therefore, in this paper, we mainly focus on this family to boost adversarial attacks.

To enhance the transferability of adversarial examples, it is crucial to avoid getting trapped in a poor local optimum of the substitute model. Towards this end, Dong *et al.* [4] adopt the momentum term at each iteration of I-FGSM [18] to stabilize update direction. Lin *et al.* [22] further adapt Nesterov accelerated gradient [31] into the iterative attacks with the aim of effectively looking ahead. Gao *et al.* [10] propose patch-wise perturbations to better cover the discriminate region of images. In addition to considering better optimization algorithms, *model augmentation* [22] is also an effective strategy. Xie *et al.* [53] introduce a random transformation to the input, thus improving the transferability. Dong *et al.* [5] 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. Lin *et al.* [22] leverage the scale-invariant property of DNNs and thus average the gradients with respect to different scaled images to update adversarial examples. Zou *et al.* [60] modify DI-FGSM [53] to promote TI-FGSM [5] by generating multi-scale gradients. Wang *et al.* [45] consider the gradient variance along momentum optimization path to avoid overfitting. Wang *et al.* [47] average the gradients with respect to feature maps to disrupt important object-aware features. Wang *et al.* [46] average the gradients of a set of admixed images, which are the input image admixed with a small portion of other images while maintaining the label of the input image. Wu *et al.* [50] utilizes an adversarial transformation network to find a better transformation for adversarial attacks in the spatial domain.

2.2 Frequency-based Analysis and Attacks

Several works [54,44,36,6,48] have analyzed DNNs from a frequency domain perspective. Wang *et al.* [44] notice DNNs' ability in capturing high-frequency components of an image which are almost imperceptible to humans. Dong *et al.* [54] find that naturally trained models are highly sensitive to additive perturbations in high frequencies, and both Gaussian data augmentation and adversarial training can significantly improve robustness against high-frequency noises.

In addition, there also exists several adversarial attacks based on frequency domain. Guo *et al.* [13] propose a LF attack that only leverages the low-frequency components of an image, which shows that low-frequency components also play a significant role in model prediction as high-frequency components. Sharma *et al.* [36] demonstrate that defense models based on adversarial training are less sensitive to high-frequency perturbations but still vulnerable to low-frequency perturbations. Duan *et al.* [6] propose the AdvDrop attack which generates adversarial examples by dropping existing details of clean images in frequency domain. Unlike these works that perturb a subset of frequency components, our method aims to narrow the gap between models by frequency-based analysis.

2.3 Adversarial Defenses

To mitigate the threat of adversarial examples, numerous adversarial defense techniques have been proposed in recent years. One popular and promising way is adversarial training [12,25] which leverages adversarial examples to augment the training data during the training phase. Tramèr *et al.* [42] introduce ensemble adversarial training, which decouples the generation of adversarial examples from the model being trained, to yield models with stronger robustness to black-box attacks. Xie *et al.* [52] inject blocks that can denoise the intermediate features into the network, and then end-to-end train it on adversarial examples to learn to reduce perturbations in feature maps.

Although adversarial training is the most effective strategy to improve the robustness of the model at present, it inevitably suffers from time-consuming training costs and is expensive to be applied to large-scale datasets and complex DNNs. To avoid this issue, many works try to cure the infection of adversarial perturbations before feeding to DNNs. Guo *et al.* [14] utilize multiple input transformations (*e.g.*, JPEG compression [7], total variance minimization [33] and image quilting [8]) to recover from the adversarial perturbations. Liao *et al.* [21] propose high-level representation guided denoiser (HGD) to suppress the influence of adversarial perturbation. Xie *et al.* [51] mitigate adversarial effects through random resizing and padding (R&P). Cohen *et al.* [17] leverage the classifier with Gaussian data augmentation to create a provably robust classifier.

In addition, researchers also try to combine the benefits of adversarial training and input pre-processing methods to further improve the robustness of DNNs. NeurIPS-r3 solution [41] propose a two-step procedure which first process images with a series of transformations (*e.g.*, rotation, zoom and sheer) and then pass the outputs through an ensemble of adversarially trained models to obtain the overall prediction. Naseer *et al.* [29] design a Neural Representation Purifier (NRP) model that learns to clean adversarial perturbed images based on the automatically derived supervision.

3 Methodology

In this section, we first give the basic definition of our task in Sec. 3.1, and then introduce our motivation in Sec. 3.2. Based on the motivation, we provide a detailed description of the proposed method - Spectrum Transformation (Sec. 3.3). Finally, we introduce our overall attack algorithm in Sec. 3.4.

3.1 Preliminaries

Formally, let $f_\theta : \boldsymbol{x} \rightarrow y$ denotes a classification model, where θ, $\boldsymbol{x}$ and y indicate the parameters of the model, input clean image and true label, respectively. Our goal is to craft an adversarial perturbation $\boldsymbol{\delta}$ so that the resulting adversarial example $\boldsymbol{x'} = \boldsymbol{x} + \boldsymbol{\delta}$ can successfully mislead the classifier, *i.e.*, $f_\theta(\boldsymbol{x'}) \neq y$ (*a.k.a.* non-targeted attack). To ensure an input is minimally changed, an adversarial example should be in the ℓ_p-norm ball centered at $\boldsymbol{x}$ with radius ϵ. Following previous works [4,53,5,10,46,47,9], we focus on the ℓ_∞-norm in this paper. Therefore, the generation of adversarial examples can be formulated as the following optimization problem:

$$\arg\max_{\boldsymbol{x'}} J(\boldsymbol{x'}, y; \theta), \quad \text{s.t. } \|\boldsymbol{\delta}\|_\infty \leq \epsilon, \tag{1}$$

where $J(\boldsymbol{x'}, y; \theta)$ is usually the cross-entropy loss. However, it is impractical to directly optimize Eq. 1 via the victim model f_θ under the black-box manner because its parameter θ is inaccessible. To overcome this limitation, a common practice is to craft adversarial examples via the accessible substitute model f_ϕ

and relying on the transferability to fool the victim model. Taking I-FGSM [18] as an example, adversarial examples at iteration $t+1$ can be expressed as:

$$\boldsymbol{x'}_{t+1} = \mathrm{clip}_{\boldsymbol{x},\epsilon}\{\boldsymbol{x'}_t + \alpha \cdot \mathrm{sign}\left(\nabla_{\boldsymbol{x'}_t} J\left(\boldsymbol{x'}_t, y; \phi\right)\right)\}, \tag{2}$$

where $\mathrm{clip}_{\boldsymbol{x},\epsilon}(\cdot)$ denotes an element-wise clipping operation to ensure $\boldsymbol{x'} \in [\boldsymbol{x} - \epsilon, \boldsymbol{x} + \epsilon]$, and α is the step size.

3.2 Spectrum Saliency Map

In order to effectively narrow the gap between models, it is important to uncover the essential differences between them. Recently, various attack methods [12,18,4,53,5,10,60,22,45,46] have been proposed to boost the transferability of adversarial examples. Among these algorithms, *model augmentation* [22] is one of the most effective strategies. However, existing works (*e.g.*, [5,22]) usually augment substitute models by applying loss-preserving transformations in the spatial domain, which might ignore essential differences among models and reduce the diversity of substitute models. Intuitively, different models usually focus on similar *spatial regions* of each input image since location of key objects in images is fixed. By contrast, as demonstrated in previous work [48,54,44], different models usually rely on different *frequency components* of each input image when making decisions.

Motivated by this, we turn to explore correlations among models from a perspective of frequency domain. Specifically, we adopt DCT to transform input images $\boldsymbol{x}$ from the spatial domain to the frequency domain. The mathematical definition of the DCT (denoted as $\mathcal{D}(\cdot)$[4] in the following) can be simplified as:

$$\mathcal{D}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}\boldsymbol{A}^{\mathrm{T}}, \tag{3}$$

where $\boldsymbol{A}$ is an orthogonal matrix and thus $\boldsymbol{A}\boldsymbol{A}^{\mathrm{T}}$ is equal to the identity matrix $\boldsymbol{I}$. Formally, low-frequency components whose amplitudes are high tend to be concentrated in the upper left corner of a spectrum, and high-frequency components are located in the remaining area. Obviously, the pattern of frequency domain is more fixed compared with diverse representations of images in spatial domain (more visualizations can be found in supplementary Sec. D.1). Therefore, we propose a spectrum saliency map $\boldsymbol{S}_\phi$ to mine sensitive points of different models f_ϕ, which is defined as:

$$\boldsymbol{S}_\phi = \frac{\partial J(\mathcal{D}_{\mathcal{I}}(\mathcal{D}(\boldsymbol{x})), y; \phi)}{\partial \mathcal{D}(\boldsymbol{x})}, \tag{4}$$

where $\mathcal{D}_{\mathcal{I}}(\cdot)$ denotes the IDCT which can recover the input image from frequency domain back to spatial domain. Note that both the DCT and the IDCT are lossless, *i.e.*, $\mathcal{D}_{\mathcal{I}}(\mathcal{D}(\boldsymbol{x})) = A^{\mathrm{T}}\mathcal{D}(\boldsymbol{x})A = \boldsymbol{x}$.

From the visualization result of $\boldsymbol{S}_\phi$ shown in Figure 1, we observe that frequency components of interest usually varies from model to model. Hence, the spectrum saliency map can serve as an indicator to reflect a specific model.

[4] In the implementation, DCT is applied to each color channel independently.

3.3 Spectrum Transformation

The analysis above motivates us that if we can simulate augmented models with a similar spectrum saliency map to victim model, the gap between the substitute model and victim model can be significantly narrowed and adversarial examples can be more transferable.

Lemma 1. *Assume both $\boldsymbol{B}_1$ and $\boldsymbol{B}_2$ are n-by-n matrix and $\boldsymbol{B}_1$ is invertible, then there must exist an n-by-n matrix $\boldsymbol{C}$ that can make $\boldsymbol{B}_1 \times C = \boldsymbol{B}_2$.*

Lemma 1 shows that it is possible to make two matrices (note the essence of spectrum saliency map is also a matrix) equal in the form of a matrix transformation. However, the spectrum saliency map of vicitm model is usually not available under black-box setting. Moreover, spectrum saliency map of substitute model is high-dimensional and not guaranteed to be invertible. To tackle this problem, we propose a random spectrum transformation $\mathcal{T}(\cdot)$ which decomposes matrix multiplication into matrix addition and Hadamard product to get diverse spectrums. Specifically, in combination with the DCT/IDCT, our $\mathcal{T}(\cdot)$ can be expressed as:

$$\mathcal{T}(\boldsymbol{x}) = \mathcal{D}_\mathcal{I}((\mathcal{D}(\boldsymbol{x}) + \mathcal{D}(\boldsymbol{\xi})) \odot \boldsymbol{M}), \tag{5}$$

$$= \mathcal{D}_\mathcal{I}(\mathcal{D}(\boldsymbol{x} + \boldsymbol{\xi}) \odot \boldsymbol{M}) \tag{6}$$

where $\odot$ denotes Hadamard product, $\xi \sim \mathcal{N}(0, \sigma^2 I)$ and each element of $M \sim \mathcal{U}(1-\rho, 1+\rho)$ are random variants sampled from Gaussian distribution and Uniform distribution, respectively. In practice, common DCT/IDCT paradigm [6,19], *i.e.*, splitting the image into several blocks before applying DCT, not works well for boosting transferability (see the ablation study for experimental details). Therefore, we apply DCT on the whole image in our experiments and visualization of transformation outputs can be found in supplementary Sec. D.2.

Formally, $\mathcal{T}(\cdot)$ is capable of yielding diverse spectrum saliency maps (we also provide proof in supplementary Sec. A) which can reflect the diversity of substitute models, and meanwhile, narrowing the gap with the victim model. As illustrated in Figure 1, previously proposed transformations [22,46] in the spatial domain (*i.e.*, (b & c)) is less effective for generating diverse spectrum saliency maps, which may lead to a weaker model augmentation. In contrast, with our proposed spectrum transformation, resulting spectrum saliency map (*i.e.*, (a)) can cover almost all those of other models.

3.4 Attack Algorithm

In Sec. 3.3, we have introduced our proposed spectrum transformation. This method could be integrated with any gradient-based attacks. For instance, in combination with I-FGSM [18] (*i.e.*, Eq. 2), we propose the Spectrum Simulation Iterative Fast Gradient Sign Method (S^{2}I-FGSM). The algorithm is detailed in Algorithm 1. Technically, our attack can be mainly divided into three steps. First, in lines 3-6, we apply our spectrum transformation $\mathcal{T}(\cdot)$ to the input image $\boldsymbol{x}_t'$ so that the gradient $\boldsymbol{g}_i'$ obtained from the substitute model is approximately equal

```
Algorithm 1: S$^2$I-FGSM
   Input : A classifier $f$ with parameters $\phi$; loss function $J$; a clean image $\boldsymbol{x}$
           with ground-truth label $y$; iterations $T$; $L_\infty$ constraint $\epsilon$; spectrum
           transformation times $N$; tunning factor $\rho$; std $\sigma$ of noise $\boldsymbol{\xi}$.
   Output: The adversarial example $\boldsymbol{x'}$
 1 $\alpha = \epsilon/T$, $\boldsymbol{x'_0} = \boldsymbol{x}$
 2 for $t = 0 \rightarrow T-1$ do
 3     for $i = 1 \rightarrow N$ do
 4         Get transformation output $\mathcal{T}(\boldsymbol{x'_t})$ using Eq. 6
 5         Gradient calculate $\boldsymbol{g'_i} = \nabla_{\boldsymbol{x'_t}} J(\mathcal{T}(\boldsymbol{x'_t}), y; \phi)$
 6     end
 7     Average gradient: $\boldsymbol{g'} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{g'_i}$
 8     $\boldsymbol{x'_{t+1}} = \text{clip}_{\boldsymbol{x},\epsilon}\{\boldsymbol{x'_t} + \alpha \cdot sign(\boldsymbol{g'})\}$
 9     $\boldsymbol{x'_{t+1}} = \text{clip}(\boldsymbol{x'_{t+1}}, 0, 1)$
10 end
11 $x' = \boldsymbol{x'_T}$
12 return $x'$
```

to the result obtained from a new model, *i.e.*, *model augmentation*. Second, in line 7, we average N augmented models' gradients to obtain a more stable update direction $\boldsymbol{g'}$. Finally, in line 8, we update adversarial examples $\boldsymbol{x'_{t+1}}$ of iteration $t+1$. In short, the above process can be summarised in the following formula:

$$\boldsymbol{x'_{t+1}} = \text{clip}_{\boldsymbol{x},\epsilon}\{\boldsymbol{x'_t} + \alpha \cdot sign(\frac{1}{N}\sum_{i=1}^{N} \nabla_{\boldsymbol{x'_t}} J(\mathcal{T}(\boldsymbol{x'_t}), y; \phi))\}. \tag{7}$$

The resulting adversarial examples are shown in Figure 2. Compared with I-FGSM [18] and SI-FGSM [22], our proposed S^{2}I-FGSM can craft more threatening adversarial examples for fooling black-box models.

4 Experiments

4.1 Experiment Setup

Dataset. Following previous works [4,5,9,10], we conduct our experiments on the ImageNet-compatible dataset[5], which contains 1000 images with resolution $299 \times 299 \times 3$.

Models. We choose six popular normally trained models, including Inception-v3 (Inc-v3) [38], Inception-v4 (Inc-v4) [37], Inception-Resnet-v2 (IncRes-v2) [37], Resnet-v2-50 (Res-50), Resnet-v2-101 (Res-101) and Resnet-v2-152 (Res-152) [15]. For defenses, we consider nine defense models (*i.e.*, Inc-v3$_{ens3}$, Inc-v3$_{ens4}$,

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

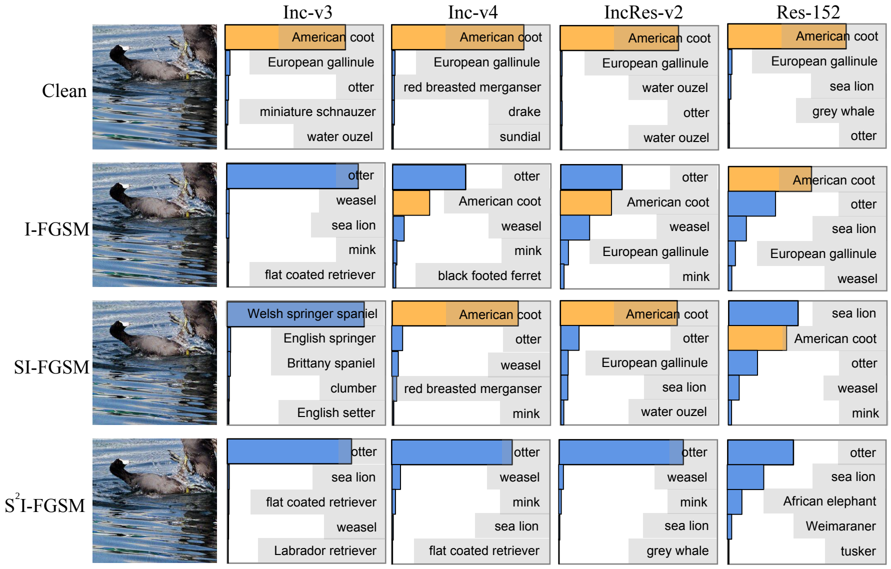


Fig. 2: The comparisons of attacks on Inc-v3 [38], Inc-v4 [37], IncRes-v2 [37] and Res152 [15]. The true label of clean image is *American coot* and marked as orange in the top-5 confidence distribution plots. The adversarial examples are crafted via Inc-v3 [38] by I-FGSM [18], SI-FGSM [22] and our proposed S^{2}I-FGSM, respectively. Remarkably, our method can attack the white-box model and all black-box models successfully.

IncRes-v2$_{ens}$ [42], HGD [21], R&P [51], NIPS-r3 [41], JPEG [14], RS [17] and NRP [29]) that are robust against black-box attacks.

Competitor. To show the effectiveness of our proposed spectrum simulation attack, we compare it with diverse state-of-the-art attack methods, including MI-FGSM [4], DI-FGSM [53], TI-FGSM [5], PI-FGSM [10], SI-NI-FGSM [22], VT-FGSM [45], FI-FGSM [47] and Admix [46]. Besides, we also compare the combined version of these methods, *e.g.*, TI-DIM (combined version of TI-FGSM, MI-FGSM and DI-FGSM) and SI-NI-TI-DIM.

Parameter Settings. In all experiments, the maximum perturbation $\epsilon = 16$, the iteration $T = 10$, and the step size $\alpha = \epsilon/T = 1.6$. For MI-FGSM, we set the decay factor $\mu = 1.0$. For DI-FGSM, we set the transformation probability $p = 0.5$. For TI-FGSM, we set the kernel length $k = 7$. For PI-FGSM, we set the amplification factor $\beta = 10$, project factor $\gamma = 16$ and the kernel length $k_w = 3$ for normally trained models, $k_w = 7$ for defense models. For SI-NI-FGSM, we set the number of copies $m_1 = 5$. For VT-FGSM, we set the hyper-parameter $\beta = 1.5$, number of sampling examples is 20. For FI-FGSM, the drop probability $p_d = 0.3$ for normally trained models and $p_d = 0.1$ for defense models, and the

Table 1: The attack success rates (%) on six normally trained models. The adversarial examples are crafted via Inc-v3, Inc-v4, IncRes-v2 and Res-152, respectively. "*" indicates white-box attacks.

Model	Attack	Inc-v3	Inc-v4	IncRes-v2	Res-152	Res-50	Res-101	AVG.
Inc-v3	MI-FGSM	**100.0***	50.6	47.2	40.6	46.9	41.7	54.5
	DI-FGSM	99.7*	48.3	38.2	31.8	39.0	33.8	48.5
	PI-FGSM	**100.0***	56.5	49.6	45.0	50.1	44.7	57.7
	S^{2}I-FGSM(ours)	99.7*	**65.0**	**58.9**	**50.3**	**56.2**	**53.3**	**63.9**
	SI-NI-FGSM	**100.0***	76.0	75.8	67.7	73.0	69.4	77.0
	VT-MI-FGSM	**100.0***	75.0	69.6	62.7	67.1	63.1	72.9
	FI-MI-FGSM	98.8*	83.6	80.0	72.7	80.2	74.9	81.7
	S^{2}I-MI-FGSM(ours)	99.6*	**88.2**	**85.8**	**81.0**	**83.4**	**81.3**	**86.6**
Inc-v4	MI-FGSM	62.0	**100.0***	46.2	41.4	47.7	42.8	56.7
	DI-FGSM	54.1	99.1*	36.3	31.4	33.7	30.4	47.5
	PI-FGSM	60.3	**100.0***	45.9	44.1	50.3	42.7	57.2
	S^{2}I-FGSM(ours)	**70.2**	99.6*	**57.1**	**48.1**	**56.5**	**47.7**	**63.2**
	SI-NI-FGSM	83.8	**99.9***	78.2	73.3	77.0	73.9	81.0
	VT-MI-FGSM	77.8	99.8*	71.5	64.1	65.7	64.4	73.9
	FI-MI-FGSM	84.9	94.7*	78.0	75.4	78.0	75.7	81.1
	S^{2}I-MI-FGSM(ours)	**90.3**	99.6*	**86.5**	**83.1**	**83.3**	**81.0**	**87.3**
IncRes-v2	MI-FGSM	60.4	52.8	99.4*	45.9	49.1	46.3	59.0
	DI-FGSM	56.5	49.1	97.8*	35.6	38.3	37.1	52.4
	PI-FGSM	62.6	57.9	**99.5***	47.0	51.4	47.9	61.1
	S^{2}I-FGSM(ours)	**76.0**	**67.7**	98.3*	**56.2**	**59.8**	**58.4**	**69.4**
	SI-NI-FGSM	86.4	82.3	**99.8***	76.8	79.6	76.4	83.4
	VT-MI-FGSM	79.3	75.6	99.5*	66.8	69.5	69.5	76.7
	FI-MI-FGSM	81.9	77.9	89.2*	72.3	75.2	75.0	78.6
	S^{2}I-MI-FGSM(ours)	**89.8**	**89.0**	98.4*	**84.9**	**86.0**	**84.3**	**88.7**
Res-152	MI-FGSM	54.7	50.1	45.5	99.4*	84.0	86.5	70.0
	DI-FGSM	57.3	51.5	47.2	99.3*	83.1	85.1	70.6
	PI-FGSM	63.2	55.1	47.8	**99.7***	82.8	84.8	72.2
	S^{2}I-FGSM(ours)	**66.8**	**62.8**	**57.4**	**99.7***	**92.8**	**94.4**	**79.0**
	SI-NI-FGSM	75.3	72.9	70.2	**99.7***	94.5	94.9	84.6
	VT-MI-FGSM	73.7	69.4	66.4	99.5*	93.1	93.8	82.7
	FI-MI-FGSM	83.7	82.1	78.6	99.4*	93.6	94.2	88.6
	S^{2}I-MI-FGSM(ours)	**88.1**	**86.9**	**86.3**	**99.7***	**97.5**	**97.6**	**92.7**

ensemble number is 30. For Admix, we set number of copies $m_1 = 5$[6] , sample number $m_2 = 3$ and the admix ratio $\eta = 0.2$. For our proposed S^{2}I-FGSM, we set the tuning factor $\rho = 0.5$ for $\boldsymbol{M}$, the standard deviation σ of $\boldsymbol{\xi}$ is simply set to the value of ϵ, and the number of spectrum transformations $N = 20$ (discussions about ρ, σ and N can be found in supplementary Sec. B). The parameter settings for the combined version are the same.

4.2 Attack Normally Trained Models

In this section, we investigate the vulnerability of normally trained models. We first compare S^{2}I-FGSM with MI-FGSM [4], DI-FGSM [53], PI-FGSM [10] to verify the effectiveness of our method in Table 1. A first glance shows that S^{2}I-FGSM consistently surpasses well-known baseline attacks on all black-box models. For example, when attacking against Inc-v3, MI-FGSM, DI-FGSM and PI-FGSM only successfully transfer 47.2%, 38.2% and 49.6% adversarial examples to IncRes-v2, while our S^{2}I-FGSM can achieve a much higher success rate

[6] Note that Admix is equipped with SI-FGSM by default.

Table 2: The attack success rates (%) on nine defenses. The adversarial examples are crafted via Inc-v3, Inc-v4, IncRes-v2 and Res-152, respectively.

Model	Attack	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	NIPS-r3	JPEG	RS	NRP	AVG.
Inc-v3	TI-DIM	43.2	42.1	27.9	36.0	30.2	37.4	56.7	55.8	22.0	39.0
	PI-TI-DI-FGSM	43.5	46.3	35.3	33.9	35.2	39.9	47.6	**74.9**	37.0	43.7
	SI-NI-TI-DIM	55.0	53.0	36.5	37.0	37.9	48.5	72.3	55.2	32.7	47.6
	VT-TI-DIM	61.3	60.4	46.6	53.9	47.8	53.3	68.3	62.4	36.1	54.5
	FI-TI-DIM	61.8	59.6	49.2	51.7	48.3	55.0	71.3	64.5	38.0	55.5
	Admix-TI-DIM	75.3	72.1	56.7	65.8	59.8	66.0	83.7	70.5	45.3	66.1
	S^{2}I-TI-DIM (ours)	81.5	81.2	69.8	77.8	70.1	77.2	86.7	71.8	**56.0**	74.7
	S^{2}I-SI-DIM (ours)	83.8	81.8	64.8	71.1	68.9	77.4	**91.8**	72.6	52.3	73.8
	S^{2}I-SI-TI-DIM (ours)	**88.6**	**87.8**	**77.9**	**81.1**	**77.6**	**83.3**	91.3	71.0	55.1	**79.3**
Inc-v4	TI-DIM	38.4	38.1	27.7	33.7	29.5	33.0	51.2	55.0	19.0	36.2
	PI-TI-DI-FGSM	42.3	43.8	32.5	33.0	33.9	36.7	46.0	74.8	32.3	41.7
	SI-NI-TI-DIM	60.2	56.9	43.8	46.0	46.5	52.7	73.7	56.3	32.5	52.1
	VT-TI-DIM	57.7	57.2	46.9	55.1	48.9	50.4	63.3	59.1	34.9	52.6
	FI-TI-DIM	61.0	58.4	50.6	53.6	51.7	55.1	67.7	62.6	38.6	55.5
	Admix-TI-DIM	77.3	74.1	63.8	73.4	67.1	71.4	82.6	67.2	48.0	69.4
	S^{2}I-TI-DIM (ours)	78.7	78.0	69.9	76.6	71.9	77.1	83.5	73.4	55.0	73.8
	S^{2}I-SI-DIM (ours)	86.0	83.7	72.4	78.4	76.8	81.7	**91.2**	73.9	**60.9**	78.3
	S^{2}I-SI-TI-DIM (ours)	**88.7**	**87.7**	**81.7**	**86.1**	**83.5**	**86.3**	90.8	**75.0**	59.6	**82.2**
IncRes-v2	TI-DIM	48.0	43.6	38.9	43.9	40.5	43.2	57.3	57.3	24.7	44.2
	PI-TI-DI-FGSM	49.7	51.1	46.0	40.1	45.9	47.8	50.6	78.0	41.0	50.0
	SI-NI-TI-DIM	71.8	62.8	55.6	53.2	59.6	64.7	82.0	60.6	41.0	61.3
	VT-TI-DIM	65.9	60.1	58.2	60.3	57.6	60.1	70.1	61.2	36.9	58.9
	FI-TI-DIM	58.1	54.4	53.5	52.6	52.2	56.8	64.2	64.4	39.8	55.1
	Admix-TI-DIM	85.3	82.0	79.5	82.4	79.6	82.4	85.9	74.2	59.7	79.0
	S^{2}I-TI-DIM (ours)	82.6	79.9	79.2	79.5	79.3	81.2	86.1	74.2	61.6	78.2
	S^{2}I-SI-DIM (ours)	90.3	88.6	83.7	86.6	84.1	86.9	92.0	75.5	69.0	84.1
	S^{2}I-SI-TI-DIM (ours)	**92.1**	**91.0**	**90.6**	**90.8**	**89.2**	**90.9**	**93.3**	**79.2**	**73.4**	**87.8**
Res-152	TI-DIM	55.1	52.3	42.5	55.6	46.5	52.3	64.9	61.2	32.2	51.4
	PI-TI-DI-FGSM	54.3	56.2	45.3	43.7	46.2	48.9	55.2	78.1	47.7	52.8
	SI-NI-TI-DIM	68.6	64.0	52.4	58.9	56.8	64.2	80.1	67.5	42.3	61.6
	VT-TI-DIM	64.3	61.4	54.9	60.7	54.8	59.4	69.3	67.9	41.2	59.3
	FI-TI-DIM	70.1	66.0	59.5	63.9	60.8	66.0	77.5	71.0	47.2	64.7
	Admix-TI-DIM	83.7	81.4	73.7	81.2	77.0	80.1	87.8	75.0	59.5	77.7
	S^{2}I-TI-DIM (ours)	86.6	83.9	79.0	85.3	81.8	85.5	90.6	80.9	66.1	82.2
	S^{2}I-SI-DIM (ours)	89.3	84.4	77.9	86.6	82.7	86.3	**92.8**	76.4	65.9	82.5
	S^{2}I-SI-TI-DIM (ours)	**92.5**	**88.6**	**85.3**	**88.6**	**87.8**	**89.8**	92.4	**83.6**	**72.0**	**86.7**

of **58.9%**. This convincingly validates the high effectiveness of our proposed method against normally trained models.

Besides, we also report the results for methods with the momentum term [4]. As displayed in Table 1, the performance gap between our proposed method and state-of-the-art approaches is still large. Notably, adversarial examples crafted by our proposed S^{2}I-MI-FGSM are capable of getting **88.8%** success rate on average, which outperforms SI-NI-FGSM, VT-MI-FGSM and FI-MI-FGSM by 7.3%, 12.2% and 6.3%, respectively. This also demonstrates that the combination of our method and existing attacks can significantly enhance the transferability of adversarial examples.

4.3 Attack Defense Models

Although many attack methods can easily fool normally trained models, they may fail in attacking models with the defense mechanism. To further verify the superiority of our method, we conduct a series of experiments against defense models. Given that the vanilla versions of attacks are less effective for defense

Table 3: The attack success rates (%) on nine defenses. The adversarial examples are crafted via an ensemble of Inc-v3, Inc-v4, IncRes-v2 and Res-152 and the weight for each model is 1/4.

Attack	Inc-v3$_{ens3}$	Inc-v3$_{ens4}$	IncRes-v2$_{ens}$	HGD	R&P	NIPS-r3	JPEG	RS	NRP	AVG.
TI-DIM	79.2	75.3	69.3	80.4	73.9	76.7	87.5	68.3	43.1	72.6
PI-TI-DI-FGSM	75.0	76.0	67.7	69.5	68.0	72.6	77.8	83.4	60.8	72.3
SI-NI-TI-DIM	90.2	87.9	80.0	83.2	83.5	87.8	94.3	81.4	59.2	83.1
VT-TI-DIM	85.0	82.3	78.3	83.9	79.4	81.9	88.5	74.5	59.7	79.3
FI-TI-DIM	83.1	83.6	74.6	84.9	76.5	78.6	90.2	72.2	61.2	78.3
Admix-TI-DIM	93.9	92.9	90.3	94.0	91.3	92.0	95.6	82.4	76.0	89.8
S^{2}I-TI-DIM (ours)	94.6	94.3	92.5	94.3	93.1	94.3	95.8	87.4	83.5	92.2
S^{2}I-SI-DIM (ours)	96.5	96.3	94.2	95.8	94.9	96.0	**97.4**	88.2	87.3	94.1
S^{2}I-SI-TI-DIM (ours)	**96.7**	**96.7**	**95.2**	**96.3**	**95.7**	**96.5**	96.9	**92.2**	**92.2**	**95.4**

models, we consider the stronger DIM, TI-DIM, PI-TI-DI-FGSM, SI-NI-TI-DIM, VT-TI-DIM, FI-TI-DIM and Admix-TI-DIM as competitors to our proposed S^{2}I-TI-DIM, S^{2}I-SI-DIM and S^{2}I-SI-TI-DIM.

Single-Model Attacks. We first investigate the transferability of adversarial examples crafted via a single substitute model. From the results of Table 2, we can observe that our algorithm can significantly boost existing attacks. For example, suppose we generate adversarial examples via Inc-v3, TI-DIM only achieves an average success rate of 39.0% on the nine defense models, while our proposed S^{2}I-TI-DIM can yield about 2× transferability, *i.e.*, outperforms TI-DIM by **35.7%**. This demonstrates the remarkable effectiveness of our proposed method against defense models.

Ensemble-based Attacks. We also report the results for attacking an ensemble of models simultaneously [23] to demonstrate the effectiveness of our proposed method. In particular, the adversarial examples are crafted via an ensemble of Inc-v3, Inc-v4, IncRes-v2 and Res-152. Similar to the results of Table 2, our S^{2}I-SI-TI-DIM displayed in Table 3 still consistently surpass state-of-the-art approaches. Remarkably, S^{2}I-SI-TI-DIM is capable of obtaining **95.4%** success rate on average, which outperforms SI-NI-TI-DIM, VT-TI-DIM, FI-TI-DIM and Admix-TI-DIM by 23.1%, 12.4%, 16.1%, 17.1% and 5.6%, respectively. This also reveals that current defense mechanisms are still vulnerable to well-design adversarial examples and far from the need of real security.

4.4 Ablation Study

In this section, we analyze the impact of different aspects of our method:

Frequency domain vs. Spatial domain. For our proposed S^{2}I-FGSM, transformation is applied in the frequency domain. To verify that frequency domain transformation (*i.e.*, our spectrum transformation) is more potent in narrowing the gap between models than spatial domain transformation (*i.e.*, remove the DCT/IDCT in spectrum transformation), we conduction an ablation study. As depicted in Figure 3 (left), regardless of what substitute models are attacked, the transferability of adversarial examples crafted based on frequency domain transformation is consistently higher than that of spatial domain transformation.

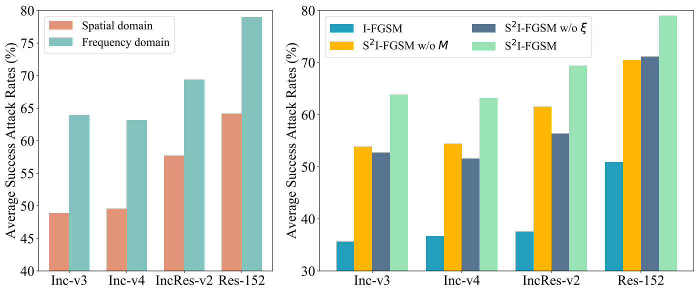


Fig. 3: The average attack success rates (%) on six normally trained models (introduced in Sec. 4.1). The adversarial examples are crafted via Inc-v3, Inc-v4, IncRes-v2 and Res-152, respectively. **Left**: Effect of frequency domain transformation. **Right**: Effect of $\boldsymbol{\xi}$ and $\boldsymbol{M}$.

Notably, when attacking against Inc-v3, the attack based on frequency domain transformation (*i.e.*, S^2I-FGSM) outperforms the attack based on spatial domain transformation by a large margin of **15.0%**. This convincingly validates that frequency domain can capture more essential differences among models, thus yielding more diverse substitute models than spatial domain.

Effect of $\boldsymbol{\xi}$ and $\boldsymbol{M}$. To analyze the effect of each random variant (*i.e.*, $\boldsymbol{\xi}$ and $\boldsymbol{M}$) in our spectrum transformation, we conduct the experiment in Figure 3 (right). From this result, we observe that both $\boldsymbol{\xi}$ and $\boldsymbol{M}$ are useful for enhancing the transferability of adversarial examples. It is because both of them can manipulate the spectrum saliency map to a certain extent, albeit from different aspects of implementation. Therefore, by leveraging them simultaneously, our proposed spectrum transformation can simulate a more diverse substitute model, thus significantly boosting attacks.

On the block size of DCT/IDCT. Previous works [6,19] usually started by splitting images into small blocks with size $n \times n$ and then apply DCT/IDCT. However, it is not clear that this paradigm is appropriate for our approach. Therefore, in this part, we investigate the impact of block size on the transferability. Specifically, we tune the block size from 8×8 to 299×299 (full image size) and report the attack success rates of S^2I-FGSM in Figure 4. From this result, we observe that larger blocks are more suited to our approach. Particularly, the attack success rates reach peak when the size of the block is the same as the full image size. Therefore, in our experiment, we do not split the image beforehand and directly apply DCT/IDCT on the full image to get its spectrum (we also provide time analysis of DCT/IDCT in supplementary Sec. C).

Attention shift. To better understand the effectiveness of our attack, we apply Grad-CAM [35] to compare attention maps of clean images with those of adversarial examples. As illustrated in Figure 5, our proposed method can effectively shift the model's attention from the key object to other mismatched regions. Consequently, the victim model inevitably captures other irrelevant features, thus leading to misclassification.

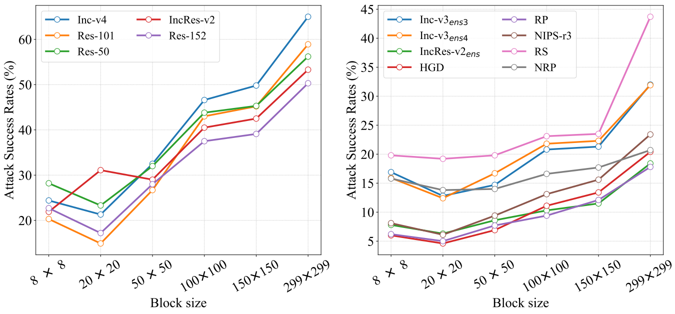


Fig. 4: The attack success rates (%) of S^{2}I-FGSM on normally trained models (**Left**) and defense models (**Right**) w.r.t. the block size of DCT/IDCT. Adversarial examples are generated via Inc-v3.

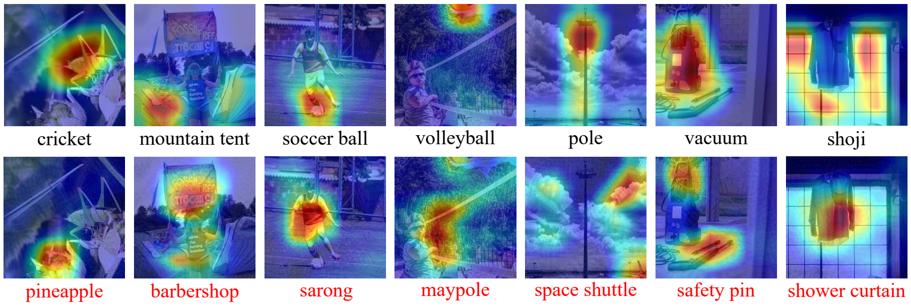


Fig. 5: Visualization for attention shift. We apply Grad-CAM [35] for Res-152 [15] to visualize attention maps of clean (1st row) and adversarial images (2nd row). Adversarial examples are crafted via Inc-v3 by our S^{2}I-FGSM. The result demonstrates that our adversarial examples are capable of shifting model's attention.

5 Conclusion

In this paper, we propose a Spectrum Simulation Attack to boost adversarial attacks from a frequency domain perspective. Our work gives a novel insight into model augmentation, which narrows the gap between the substitute model and victim model by a set of spectrum transformation images. We also conduct a detailed ablation study to clearly illustrate the effect of each component. Compared with traditional model augmentation attacks in spatial domain, extensive experiments demonstrate the significant effectiveness of our method, which outperforms state-of-the-art transfer-based attacks by a large margin.

Acknowledgments. This work is supported by the National Natural Science Foundation of China (Grant No. 62122018, No. 61772116, No. 61872064, No. U20B2063).

References

1. Ahmed, N., Natarajan, T.R., Rao, K.R.: Discrete cosine transform. IEEE Trans. Computers **23**, 90–93 (1974) 3
2. Bojarski, M., Testa, D.D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L.D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., Zieba, K.: End to end learning for self-driving cars. CoRR **abs/1604.07316** (2016) 1
3. Carlini, N., Wagner, D.A.: Towards evaluating the robustness of neural networks. In: Symposium on Security and Privacy (2017) 2, 3
4. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: CVPR (2018) 2, 3, 5, 6, 8, 9, 10, 11
5. Dong, Y., Pang, T., Su, H., Zhu, J.: Evading defenses to transferable adversarial examples by translation-invariant attacks. In: CVPR (2019) 2, 4, 5, 6, 8, 9
6. Duan, R., Chen, Y., Niu, D., Yang, Y., Qin, A.K., He, Y.: Advdrop: Adversarial attack to dnns by dropping information. In: ICCV (2021) 3, 4, 7, 13
7. Dziugaite, Karolina, G., Ghahramani, Z., Roy., D.M.: A study of the effect of jpg compression on adversarial images. CoRR **abs/1608.00853** (2016) 5
8. Efros, A.A., Freeman, W.T.: Image quilting for texture synthesis and transfer. In: SIGGRAPH (2001) 5
9. Gao, L., Cheng, Y., Zhang, Q., Xu, X., Song, J.: Feature space targeted attacks by statistic alignment. In: IJCAI (2021) 3, 5, 8
10. Gao, L., Zhang, Q., Song, J., Liu, X., Shen, H.T.: Patch-wise attack for fooling deep neural network. In: ECCV (2020) 2, 3, 5, 6, 8, 9, 10
11. Gao, L., Zhang, Q., Song, J., Shen, H.T.: Patch-wise++ perturbation for adversarial targeted attacks. CoRR **abs/2012.15503** (2020) 3
12. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. In: ICLR (2015) 1, 3, 4, 6
13. Guo, C., Frank, J.S., Weinberger, K.Q.: Low frequency adversarial perturbation. In: Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019. vol. 115, pp. 1127–1137 (2019) 4
14. Guo, C., Rana, M., Cissé, M., van der Maaten, L.: Countering adversarial images using input transformations. In: ICLR (2018) 5, 9
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR. pp. 770–778 (2016) 1, 2, 8, 9, 14
16. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: CVPR (2017) 1
17. Jia, J., Cao, X., Wang, B., Gong, N.Z.: Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. In: ICLR (2020) 5, 9
18. Kurakin, A., Goodfellow, I.J., Bengio, S.: Adversarial machine learning at scale. In: ICLR (2017) 3, 6, 7, 8, 9
19. Li, J., Ji, R., Liu, H., Liu, J., Zhong, B., Deng, C., Tian, Q.: Projection & probability-driven black-box attack. In: CVPR (2020) 7, 13
20. Li, X., Li, J., Chen, Y., Ye, S., He, Y., Wang, S., Su, H., Xue, H.: QAIR: practical query-efficient black-box attacks for image retrieval. In: CVPR (2021) 3
21. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: CVPR (2018) 5, 9
22. Lin, J., Song, C., He, K., Wang, L., Hopcroft, J.E.: Nesterov accelerated gradient and scale invariance for adversarial attacks. In: ICLR (2020) 2, 3, 4, 6, 7, 8, 9

23. Liu, Y., Chen, X., Liu, C., Song, D.: Delving into transferable adversarial examples and black-box attacks. In: ICLR (2017) 12
24. Liu, Y., Cheng, Y., Gao, L., Liu, X., Zhang, Q., Song, J.: Practical evaluation of adversarial robustness via adaptive auto attack. CoRR **abs/2203.05154** (2022) 3
25. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018) 4
26. Mao, X., Chen, Y., Wang, S., Su, H., He, Y., Xue, H.: Composite adversarial attacks. In: AAAI (2021) 2
27. Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., Frossard, P.: Universal adversarial perturbations. In: CVPR (2017) 3
28. Moosavi-Dezfooli, S., Fawzi, A., Frossard, P.: Deepfool: A simple and accurate method to fool deep neural networks. In: CVPR (2016) 2, 3
29. Naseer, M., Khan, S.H., Hayat, M., Khan, F.S., Porikli, F.: A self-supervised approach for adversarial robustness. In: CVPR (2020) 5, 9
30. Naseer, M., Khan, S.H., Khan, M.H., Khan, F.S., Porikli, F.: Cross-domain transferability of adversarial perturbations. In: NeurPIS (2019) 3
31. Nesterov, Y.: A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. In: Doklady AN USSR (1983) 3
32. Papernot, N., McDaniel, P.D., Goodfellow, I.J., Jha, S., Celik, Z.B., Swami, A.: Practical black-box attacks against machine learning. In: Karri, R., Sinanoglu, O., Sadeghi, A., Yi, X. (eds.) ACM (2017) 3
33. Rudin, L.I., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. Physica D: nonlinear phenomena (1992) 5
34. Sallab, A.E., Abdou, M., Perot, E., Yogamani, S.K.: Deep reinforcement learning framework for autonomous driving. CoRR **abs/1704.02532** (2017) 1
35. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017) 13, 14
36. Sharma, Y., Ding, G.W., Brubaker, M.A.: On the effectiveness of low frequency perturbations. In: IJCAI. pp. 3389–3396 (2019) 4
37. Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: AAAI (2017) 2, 8, 9
38. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: CVPR (2016) 2, 8, 9
39. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: ICLR (2014) 1, 3
40. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: CVPR (2014) 1
41. Thomas, A., Elibol, O.: Defense against adversarial attacks-3rd place. https://github.com/anlthms/nips-2017/blob/master/poster/defense.pdf (2017) 5, 9
42. Tramèr, F., Kurakin, A., Goodfellow, I.J., Boneh, D., McDaniel, P.D.: Ensemble adversarial training: Attacks and defenses. In: ICLR (2018) 2, 4, 9
43. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W.: Cosface: Large margin cosine loss for deep face recognition. In: CVPR (2018) 1
44. Wang, H., Wu, X., Huang, Z., Xing, E.P.: High-frequency component helps explain the generalization of convolutional neural networks. In: CVPR. pp. 8681–8691 (2020) 4, 6
45. Wang, X., He, K.: Enhancing the transferability of adversarial attacks through variance tuning. In: CVPR. pp. 1924–1933 (2021) 4, 6, 9
46. Wang, X., He, X., Wang, J., He, K.: Admix: Enhancing the transferability of adversarial attacks. In: ICCV (2021) 2, 4, 5, 6, 7, 9

47. Wang, Z., Guo, H., Zhang, Z., Liu, W., Qin, Z., Ren, K.: Feature importance-aware transferable adversarial attacks. In: ICCV (2021) 4, 5, 9
48. Wang, Z., Yang, Y., Shrivastava, A., Rawal, V., Ding, Z.: Towards frequency-based explanation for robust CNN. CoRR **abs/2005.03141** (2020) 4, 6
49. Wu, D., Wang, Y., Xia, S., Bailey, J., Ma, X.: Skip connections matter: On the transferability of adversarial examples generated with resnets. In: ICLR (2020) 3
50. Wu, W., Su, Y., Lyu, M.R., King, I.: Improving the transferability of adversarial samples with adversarial transformations. In: CVPR (2021) 3, 4
51. Xie, C., Wang, J., Zhang, Z., Ren, Z., Yuille, A.L.: Mitigating adversarial effects through randomization. In: ICLR (2018) 5, 9
52. Xie, C., Wu, Y., van der Maaten, L., Yuille, A.L., He, K.: Feature denoising for improving adversarial robustness. In: CVPR (2019) 4
53. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: CVPR (2019) 2, 3, 4, 5, 6, 9, 10
54. Yin, D., Lopes, R.G., Shlens, J., Cubuk, E.D., Gilmer, J.: A fourier perspective on model robustness in computer vision. In: NeurIPS (2019) 3, 4, 6
55. Zhang, J., Song, J., Gao, L., Liu, Y., Shen, H.T.: Progressive meta-learning with curriculum. IEEE Transactions on Circuits and Systems for Video Technology (2022) 1
56. Zhang, Q., Li, X., Chen, Y., Song, J., Gao, L., He, Y., Xue, H.: Beyond imagenet attack: Towards crafting adversarial examples for black-box domains. In: ICLR (2022) 3
57. Zhang, Q., Zhang, C., Li, C., Song, J., Gao, L., Shen, H.T.: Practical no-box adversarial attacks with training-free hybrid image transformation. CoRR **abs/2203.04607** (2022) 3
58. Zhang, Q., Zhu, X., Song, J., Gao, L., Shen, H.T.: Staircase sign method for boosting adversarial attacks. CoRR **abs/2104.09722** (2021) 3
59. Zhao, Z., Liu, Z., Larson, M.A.: Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In: CVPR (2020) 2, 3
60. Zou, J., Pan, Z., Qiu, J., Liu, X., Rui, T., Li, W.: Improving the transferability of adversarial examples with resized-diverse-inputs, diversity-ensemble and region fitting. In: ECCV (2020) 4, 6