Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.082629

ARTICLE

Enhancing the Transferability of Adversarial Samples through Frequency-Domain Attenuation

Li Peng[1,2], **Xiangbing Li**[1,2], **Kun Zou**[1], **Yong Liu**[1,2,*] **and Haibo Huang**[1]

[1]School of Artificial Intelligence, Hubei University of Automotive Technology, Shiyan, China
[2]Shiyan Key Laboratory of Electromagnetic Induction and Energy-Saving Technology, Hubei University of Automotive Technology, Shiyan, China

*Corresponding Author: Yong Liu. Email: liuyong@huat.edu.cn

Received: 19 March 2026; Accepted: 15 May 2026; Published: 23 July 2026

ABSTRACT: In recent years, the transferability of adversarial examples has attracted significant attention. To improve the effectiveness of black-box attacks, a frequency-domain decay constraint is introduced, inspired by weight decay and regularization techniques commonly employed during model training. By treating adversarial perturbations as inputs in an optimization process, this constraint aims to mitigate the excessive reliance on low-frequency components during adversarial example generation, thereby enhancing transferability. Fourier heatmaps are utilized to analyze the sensitivity of input samples, enabling a decomposition of the frequency spectrum into low-frequency and high-frequency components. Based on this analysis, low-frequency attenuation is applied in the Fourier domain to suppress dominant low-frequency information, followed by reconstruction of the perturbed inputs. The proposed frequency-domain attenuation strategy enjoys good compatibility with existing algorithms, and increases the attack success rate by approximately 1.53%–8.68% relative to the original method. Extensive experimental results show that the proposed method surpasses existing iterative attack methods and generates more transferable adversarial examples, demonstrating its effectiveness and superiority.

KEYWORDS: Adversarial transferability; black-box attack; adversarial examples; frequency domain attenuation

1 Introduction

In recent years, the emergence of image adversarial samples has created a serious threat to the security of deep learning, and the problem of adversarial samples has become a hot spot in current research [1–4]. Understanding how adversarial examples are created is essential for assessing the robustness of deep neural networks, identifying their vulnerabilities, and evaluating potential security risks. Methods for crafting these adversarial inputs typically fall into two categories: white-box and black-box strategies. In a white-box setting, adversaries have complete information about the model, including its architecture and learned parameters [1]. In contrast, black-box techniques operate under the assumption that the attacker has no insight into the model's internal workings. While white-box strategies achieve elevated efficacy by leveraging gradient-based optimization to craft perturbations, their real-world applicability is constrained by the rarity of comprehensive model transparency. To address this limitation, research efforts have shifted to analyzing the cross-model applicability of adversarial perturbations in black box environments, making it possible for the designed attack methods to have better transferability [3,5].

Due to the tendency of adversarial examples to migrate between models, a typical strategy involves applying white-box attack methods in a transfer-based black-box setting [6]. Nevertheless, this approach can cause the adversarial images to overfit to the particular weaknesses of the source model [7]. Consequently, while carefully crafted adversarial examples may attain nearly 100% success against white-box models [8], their performance often drops when attacking different models.

In this study, we enhance the transferability of iterative black-box attack strategies through the utilization of frequency attenuation on input images during the generation of adversarial examples. Drawing a parallel between model training and adversarial attacks from an optimization standpoint, iterative attacks can be considered as a training process for the input, whereas model training involves training the network's parameters. We are inspired by techniques such as regularization and weight decay, which effectively prevent overfitting in deep neural networks during model training by avoiding overfitting to the training set. Our focus is on the frequency-domain decay of the input image to minimize overfitting to the white-box model, thereby enhancing the mobility of the adversarial examples. Furthermore, our Fourier frequency-domain perturbation analysis reveals that low frequencies have a greater impact on the model's output. That is, the output of a CNN remains largely unchanged even with attenuated low frequencies. To improve mobility, we propose attenuating the low frequencies of the input, which can be integrated as a general regularization term into existing iterative attacks, such as MI-FGSM [9], DI-FGSM [10], TI-FGSM [7], and S^2I-FGSM [11], to further enhance the transferability and effectiveness of adversarial examples.

The main contributions of this work are summarized as follows:

- We propose to generate adversarial samples by attenuating the low-frequency components in the input image. This approach reduces the excessive reliance on the low-frequency components of the sample, thereby enhancing the transferability of the generated samples.
- We propose a classification method that utilizes Fourier heatmaps to partition the frequency domain. This objective strategy avoids the arbitrariness of empirical partitioning and enables more targeted frequency attenuation. The proposed frequency-domain attenuation strategy exhibits good compatibility with existing algorithms, and the integrated method achieves a higher adversarial attack success rate than the original method.
- Both visualization results and ablation studies demonstrate the feasibility and effectiveness of the proposed approach.

2 Related Work

This section first briefly introduces white-box adversarial attacks and black-box adversarial attacks, and then introduces common defense methods against adversarial attacks.

2.1 White-Box Adversarial Attack Methods

There exist two fundamental categories of adversarial attacks, namely white-box and black-box attacks, which are delineated by the extent of prior model information available to the adversary. Specifically, white-box attacks are executed under the condition that the attacker possesses complete, unrestricted knowledge of the target model's architectural design and parameter values, whereas black-box attacks are performed without any prior understanding of the model's internal characteristics. Within the white-box attack setting, adversarial sample generation can be achieved with minimal difficulty, as the attacker has full access to the detailed internal structure and complete parameter set of the targeted model. A suite of well-established algorithms has been developed for white-box adversarial attacks, with representative implementations including FGSM [1], BIM [12], PGD [13], DeepFool [14], JSM [15], and CW [16], among

others. As a pioneering work in this field, FGSM [1] represents the first formal introduction of gradient-based methodology for adversarial attack construction. Despite the high computational efficiency of FGSM in generating adversarial samples, its single-step attack framework results in low precision in perturbation calculation, which consequently leads to a relatively low attack success rate. The PGD attack [13], which has been widely adopted across adversarial machine learning research, operates by iteratively updating the gradient of each pixel along the gradient ascent direction, while simultaneously enforcing hard constraints to confine the generated perturbations within a predefined bounded range. This iterative optimization approach delivers a substantial enhancement in overall attack performance.

2.2 Black-Box Adversarial Attack Methods

Within the context of black-box adversarial attacks, adversaries possess access solely to the target model's outputs. Such scenarios involve two primary strategies: query-driven assaults and transfer-based methods. This investigation centers on transfer-based attacks under black-box conditions. Prominent among transferable adversarial techniques are Fast Gradient Sign Method (FGSM) inspired approaches [12,17,18], which exploit the transferability of adversarial instances. Iterative techniques target surrogate models, generating adversarial examples through stepwise updates of perturbations based on gradient data. Kurakin et al. [12] refined FGSM via the Basic Iterative Method (BIM), also referred to as Iterative Fast Gradient Sign Method (I-FGSM). By dividing perturbation computation into incremental stages and restricting pixel values within defined bounds via clipping, the I-FGSM approach amplifies FGSM's efficacy. This adjustment led to improved adversarial success rates. Nevertheless, I-FGSM tends to produce adversarial samples susceptible to local overfitting, thereby diminishing their transferability. To mitigate this limitation, Dong et al. [9] integrated momentum into the I-FGSM framework, formulating the Momentum Iterative Fast Gradient Sign Method (MI-FGSM).

Alongside the exploration of advanced optimization strategies to strengthen defenses against adversarial attacks, model augmentation has gained prominence as a widely used and effective method [17]. To address model overfitting associated with the I-FGSM framework, Xie et al. [10] introduced diverse image transformations, coining their method the Diverse Iterative Fast Gradient Sign Method (DI-FGSM). Dong et al. [7] proposed input data transformations to reduce dependence on surrogate models, synthesizing image sequences and estimating global gradients. Leveraging the scale invariance property of deep neural networks (DNNs), Lin et al. [17] computed averaged gradients across varying scales to refine adversarial example generation. Building on DI-FGSM [10], Zou et al. [19] adapted the algorithm to produce multiscale gradient information, thereby enhancing TI-FGSM [7]. At the same time, Wang and He [20] examined gradient variance along the momentum-based optimization path to mitigate overfitting. Although these approaches demonstrate strong results in untargeted adversarial settings, their effectiveness significantly declines in scenarios requiring precise target alignment.

Over the past few years, researchers have proposed various methods based on frequency-domain analysis to improve the transferability of adversarial attacks in black-box settings. Liu et al. [8] analyzes the model's sensitivity to different frequencies in the frequency domain and imposes perturbations only on key frequency bands, thereby improving the transferability and efficiency of adversarial attacks. Long et al. [11] adopts frequency-domain data augmentation to generate diverse inputs, and jointly optimizes adversarial examples over multiple spectral distributions to enhance the cross-model generalization capability of attacks. Li et al. [21] introduces frequency-domain regularization constraints into the iterative adversarial attack process to regulate the spectral structure of perturbations, rendering the attack more stable and natural.

In addition, Li et al. [22] abandons the traditional multi-architecture model ensemble, and implements ensemble attacks by leveraging checkpoints of a single model at different epochs. By fusing early and late

model snapshots, it generates adversarial examples with high black-box transferability at low computational cost. Ren et al. [23] proposes an attack strategy based on forward propagation refinement. By imposing AMD perturbations on attention maps and adopting momentum smoothing on token embeddings, it produces transferable adversarial examples with strong robustness across CNN and ViT architectures. For targeted black-box attacks, Liang et al. [24] introduces learnable subtle perturbations into the feature space and mixes them with randomly purified features, generating robust targeted adversarial examples that can cross the decision boundaries of different black-box models.

2.3 Adversarial Defense

While adversarial training remains the predominant approach for enhancing model robustness, its practical application to large-scale datasets and intricate deep neural networks is hindered by substantial computational demands and prolonged training durations. To circumvent these challenges, recent research has shifted focus toward preprocessing techniques designed to neutralize adversarial perturbations prior to model input. For instance, Guo et al. [25] developed a method leveraging diverse input transformations including JPEG compression [26] and total variance minimization [27] to restore adversarial examples. Similarly, Liao et al. [28] introduced the high-level representation-guided denoiser (HGD), which attenuates adversarial interference by aligning perturbations with semantic features. Complementary strategies, such as randomized resizing and padding (R&P) proposed by Xie et al. [29], further diminish adversarial effects through stochastic input modifications. Additionally, Jia et al. [30] demonstrated that classifiers augmented with Gaussian noise during training can achieve verifiable robustness against adversarial perturbations. Yan et al. [31] utilizes Wavelet Transform in the frequency domain to regularize deep neural networks, so as to enhance the model robustness against adversarial attacks.

3 Proposed Method

This section begins by providing a detailed overview of the main notations used throughout our proposed method, as summarized in Table 1. Subsequently, we will elaborate on our proposed method. Fig. 1 illustrates the schematic framework of the proposed method.

Table 1: The main notations used in the proposed method and their corresponding explanations.

Notation	Brief Description of the Notations
x	The sample images input into the deep neural network model
x_{adv}	The adversarial sample images obtained by the algorithm
δ	The perturbation of the adversarial sample
$f_\theta(x, y)$	The substitution model
y_{ori}	The original label of the sample image
ϵ	The maximum perturbation
$F(x)$	The discrete fourier transform of the input image
$F^{-1}(x)$	The inverse discrete fourier transform of the input image
$T_F(x)$	The frequency domain attenuation of the input image
λ	The constraint factor
Z_L	The low frequency region
Z_H	The high frequency region
g	The gradient information during backpropagation
T	The number of iterations

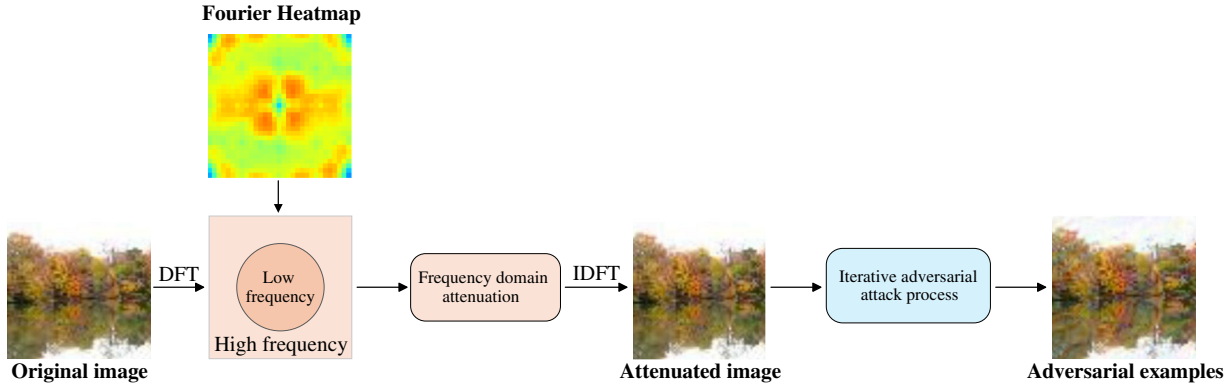


Figure 1: Schematic framework of the proposed method. DFT stands for discrete fourier transform, and IDFT denotes inverse discrete fourier transform.

3.1 Theoretical Analysis of Frequency Attenuation and Transferability

In this section, we provide theoretical insights into why the proposed frequency-domain low-frequency attenuation improves the transferability of adversarial examples.

3.2 Preliminary

We begin by establishing a general classification framework $f_\theta(x, y) : x \in X \to y \in Y$ with θ, where θ denotes the network's learnable parameters, x corresponds to clean image data, and y indicates the predicted class labels. The input images $x \in \mathbb{R}^{c\times h\times \mathrm{w}}$ are characterized by three dimensions: channel count c, vertical resolution h, and horizontal resolution w. The output labels $y \in \mathbb{R}^{\mathrm{C}}$ reside in a vector space with dimensionality C, equivalent to the total categories in the training corpus. Here, y_{ori} signifies the original ground truth annotation for a given image sample. The objective of supervised learning for classification can thus be mathematically defined as

$$\theta = \arg\min_{\theta} \mathcal{L}(f_\theta(x, y), y_{ori}), \tag{1}$$

where $\mathcal{L}(f_\theta(x, y), y_{ori})$ can be denote as the loss function of the neural network classification model.

The core of an adversarial attack lies in creating a perturbation, represented as δ, which successfully deceives the classifier to produce an adversarial example x_{adv}. This procedure can be formulated as an optimization task aimed at maximizing the loss function $J(x_{adv}, y; \theta)$, while ensuring that the ℓ_p-norm of the perturbation δ does not exceed a predefined threshold ε. The optimization task can thus be described as follows:

$$x_{adv} = x + \delta, \tag{2}$$

where x_{adv} represents the generated adversarial sample, and the constraints on the adversarial sample can be formally expressed as:

$$\arg\max_{x_{adv}} \mathcal{L}(x_{adv}, y; \theta), \quad s.t. \quad |\delta|_p \leq \epsilon. \tag{3}$$

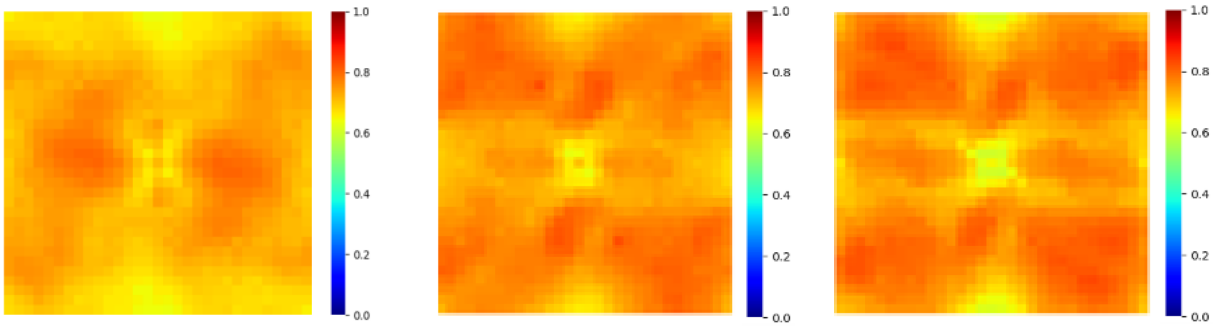
In essence, $\mathcal{L}(x_{adv}, y; \theta)$ seeks to amplify the cross-entropy loss. To guarantee that the crafted adversarial example stays closely aligned with the original input, exhibiting only minor perturbations, its deviation must be confined within an ℓ_p-norm ball of radius ϵ centered at x.

Generalization refers to the capacity of deep learning algorithms to adapt to unseen data samples. In Eq. (2), we optimize the model parameters θ through learning from input samples x. To mitigate overfitting, various regularization techniques are commonly employed, such as L2 regularization, dropout, and early stopping, aimed at preventing the parameters θ from fitting too closely to the training set and thereby enhancing the classifier's generalization performance. The classifier's generalization capability gauges its performance on unseen inputs, and likewise, we can also define the generalization capability of adversarial samples. Particularly, in transfer-based adversarial attack approaches, adversarial samples are crafted using a surrogate model(i.e., white-box model). In essence, adversarial samples x_{adv} are optimized based on specific model parameters θ and utilized to deceive unknown models. The generalization capacity of adversarial samples can serve as a metric for assessing their ability to attack other black-box models. If adversarial samples achieve a high attack success rate against unknown models, it suggests strong generalization ability.

3.3 Deep Neural Network Frequency Domain Analysis

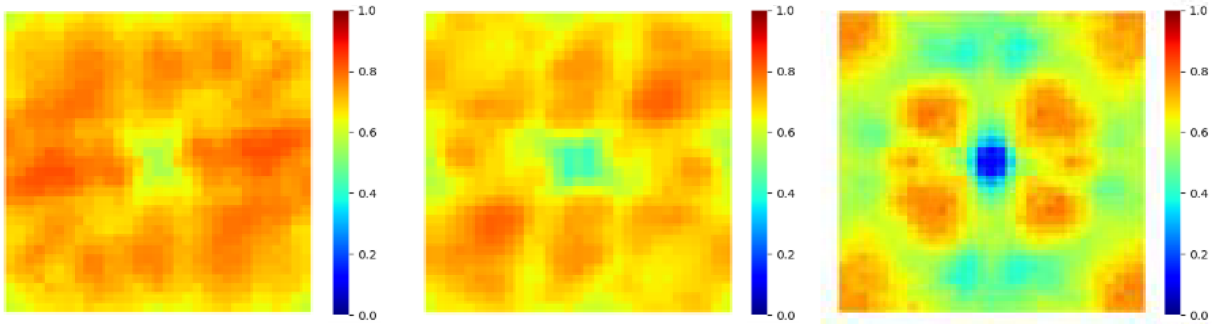
A growing body of recent research [32,33] has performed comprehensive analyses on the robustness and generalization properties of deep neural networks (DNNs) through the lens of frequency domain theory. Among these contributions, the work presented in [32] uncovered the frequency principle, a core rule describing that the generalization performance of neural networks is preferentially focused on low-frequency components during the model training procedure, and that DNNs generally fit the target function following a progression from low-frequency to high-frequency components. Relevant literature [32,33] has demonstrated both theoretically and experimentally the importance of low frequencies on model classification output results. The low-frequency information contributes more to the model fit than the high-frequency information, i.e., the low-frequency component has a greater impact on the model output.

Yin et al. [34] investigated the effect of different frequencies on model robustness by Fourier perturbation analysis and evaluated the test error of different frequencies on the model output results by visualization of Fourier heatmaps, these visualizations are called Fourier heatmaps of the test model. Fig. 2 presents the Fourier heat map of ResNet56 on CIFAR-10 during the idle training period. Colors ranging from red to blue correspond to test errors from 1 to 0, reflecting the model's sensitivity to different frequency components. With increasing iterations, the central low-frequency regions transition from red to blue first, followed by gradual changes in the surrounding high-frequency regions. This indicates that the DNN fits the objective function from low to high frequencies during training, and the low-frequency components contribute more significantly to the model's output.



(a) 5 epochs (b) 30 epochs (c) 60 epochs

Figure 2: (Continued)



(d) 120 epochs (e) 180 epochs (f) 360 epochs

Figure 2: Fourier heat maps of the classifier at different training epochs. Subfigures (**a**–**f**) present the Fourier heat maps of the neural network model obtained after 5, 30, 60, 120, 180 and 360 training epochs, respectively.

3.4 Frequency Domain Attenuation Analysis Based on Frequency Domain Sensitivity

In the context of image classification, the high dimensionality of input images often results in the creation of numerous intricate adversarial examples and the overfitting of white-box neural networks. To address this challenge, weight decay and regularization methods have been developed as efficient approaches to reduce overfitting. Similarly, we introduce frequency-domain attenuation constraints to generate more diverse adversarial examples. Our aim is to mitigate the complexity of adversarial instances and enhance the transferability of black-box models through this attenuating input feature constraint. As demonstrated by the Fourier sensitivity analysis in the preceding section, we observe that the CNN output predominantly relies on the low-frequency components of the input image, indicating a correlation between the CNN output and the low-frequency characteristics of the input. Building upon this observation, we propose a frequency-domain low-frequency attenuation constraint method based on sensitive frequency domain analysis to enhance the generalization ability of adversarial examples. Specifically, given an input x and its potential true label y, along with a classifier $f_\theta(x, y)$, where θ is trained and employed to learn the adversarial manipulation of x, we enforce the constraint that the input $T_F(\cdot)$ satisfies:

$$\hat{x} = T_F(x), \quad s.t. \quad x \in X, \tag{4}$$

where $\hat{x}$ represents the image post-attenuation in the frequency domain. This constraint regulates input features and reduces the complexity of adversarial samples, enabling the white-box network to generate more universal adversarial manipulations due to its robust learning capability. By restricting inputs, the neural network learns stronger perturbations to maximize the loss function iteratively throughout the adversarial attack. Our proposed approach exhibits behavior analogous to the regularization term utilized during network training.

During model training, it can be added to the loss function for adaptive regularisation due to the high number of iterations and epochs. When conducting iterative attacks, we typically employ a smaller number of iterations. Therefore, we propose directly incorporating regularization constraints onto the inputs rather than integrating them into the loss function.

Leveraging the significance and consistency of low frequencies in convolutional neural networks, we establish a low-frequency decay constraint denoted as $T_F(X)$ for iterative attacks conducted in the frequency domain. For the sake of simplicity, we term the method employing frequency-domain decay constraints as FDL. Specifically, we determine that the inverse discrete Fourier transform described in Eq. (6) constitutes a suitable form of constraint within the frequency domain. Consequently, our method can be formally delineated as follows.

$$F(x) = F(u, v), \tag{5}$$

where $F(x)$ represents the discrete fourier transform, and the attenuation constraint can be expressed as

$$T_F(x) = F^{-1}\left(\lambda \cdot Z_L + Z_H\right). \tag{6}$$

The crux of frequency domain attenuation constraints lies in the selection of the constraint factor λ. Intuitively, this factor must meet two criteria. Firstly, it should be generalizable across networks, as tailored values for specific networks lack broad applicability. Secondly, it must genuinely restrict the low frequency of the input; in most iterations, the value cannot reach extremes of 0 or 1. When the value is 1, low frequencies remain unattenuated, and when it is 0, it becomes meaningless. A straightforward approach is to use a constant constraint factor across all instances. However, for enhanced the performance, we opt to randomly select the constraint factor for each iteration from the interval $(0, 1)$ and uniformly sample across this range, centered at 0.5. This random selection aims to introduce diversity, akin to the dropout technique's random node deactivation. The formulation of the constraint factor is outlined below:

$$\lambda = \text{random}(0, 1), \tag{7}$$

where $\text{random}(0, 1)$ represents a uniformly random sampling function that yields a number within the interval $(0, 1)$.

Another crucial consideration is determining the boundary between low and high frequencies. Different from the global regularization scheme in [21], our method specifically targets low-frequency components, which is driven by the inherent frequency bias of DNNs. Furthermore, we introduce a data-driven frequency partitioning strategy based on Fourier sensitivity heatmaps, instead of relying on fixed or empirical frequency division rules. To achieve this, we utilize the Fourier heatmap [34] generated from an alternative model, enabling the differentiation of these frequency bands. Fig. 3 shows the Fourier heat map of the sensitivity frequencies of different classification models on different datasets. As shown in Fig. 3, the sensitive frequencies vary across different models operating on the same dataset. This variability implies differences in the frequencies deemed sensitive and robust, consequently affecting the delineation of the model's low-frequency region. Therefore, classifying the low-frequency and high-frequency regions of different agent models based on an analysis of sensitive frequencies proves to be reasonable and effective. This approach distinguishes itself from arbitrary empirical or random divisions of the frequency domain into low and high-frequency components.

Our specific methodology involves two primary steps. Firstly, we assess the sensitivity of each frequency point location by extracting values from the Fourier heat map, representing the degree of sensitivity of that frequency to the model. Subsequently, we establish a threshold value to identify the least sensitive frequency points. Next, we compute the mean distance from all identified least sensitive frequency points to the center of the spectrum, serving as the foundation for delineating low frequencies. Here the distance from the insensitive frequency points to the center of the spectrogram $(\frac{h}{2}, \frac{w}{2})$ is defined as d_L. Then $d_i(u, v)$ can be formalized as

$$d_i(u, v) = \sqrt{\left(u - \frac{h}{2}\right)^2 + \left(v - \frac{w}{2}\right)^2}, \quad s.t. \quad H(u, v) \leq \varepsilon_1 \tag{8}$$

where $H(u, v)$ is the heat value of each frequency point in the Fourier heat map, i.e., the value of model sensitivity at each frequency point. The parameter ϵ_1 is the threshold for judging insensitivity. Then, the formula for d_L can be expressed as

$$d_L = \frac{1}{N}\sum_{i=0}^{N} d_i(u,v), \tag{9}$$

where N is the total number of insensitive frequency points. According to the obtained low-frequency distance, we can divide the frequency domain into low-frequency region z_L and high-frequency region z_H, which can be formulated as

$$\begin{cases} Z_L = F(u,v), & if \quad d(u,v) \le d_L \\ Z_H = F(u,v), & if \quad d(u,v) > d_L \end{cases} \tag{10}$$

where z_L is the low frequency region and z_H represents the high frequency region.

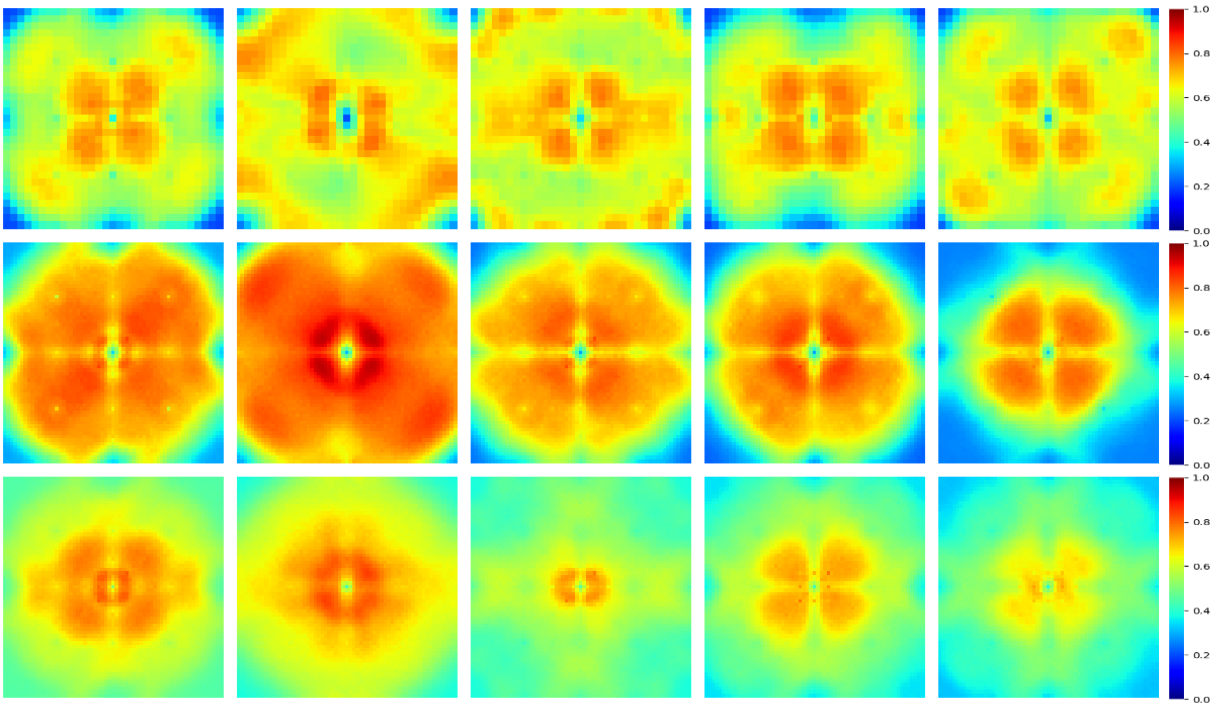


Figure 3: Visualization of sensitive frequencies via Fourier heat maps, generated for different models tested on three separate datasets. The first, second and third rows present the fourier heat map outcomes acquired on the Cifar10, Cifar100 and Tiny-ImageNet datasets, in corresponding order.

3.5 Theoretical Analysis of Frequency Attenuation and Transferability

In this section, we provide theoretical insights into why the proposed frequency-domain low-frequency attenuation improves the transferability of adversarial examples.

Given a surrogate model $f_s(\cdot)$ and loss function $\mathcal{L}(\cdot, y)$, the adversarial perturbation δ is optimized by:

$$\delta = \arg\max_{\|\delta\| \le \epsilon} \mathcal{L}(f_s(T_F(x+\delta)), y). \tag{11}$$

Compared to standard attacks, the optimization is performed over a modified input distribution induced by T_F, which alters the gradient structure.

From the perspective of generalization perspective, deep neural networks are known to exhibit a frequency bias toward low-frequency components. As a result, the gradient $\nabla_x \mathcal{L}(f_s(x), y)$ is typically dominated by low-frequency signals. Standard iterative attacks therefore produce perturbations that are highly aligned with these model-specific gradients, leading to overfitting to the surrogate model.

By attenuating low-frequency components during each iteration, the proposed method modifies the gradient as:

$$\nabla_x \mathcal{L}(f_s(T_F(x)), y) \approx \lambda \nabla_{x_L} \mathcal{L} + \nabla_{x_H} \mathcal{L}, \tag{12}$$

which reduces the dominance of low-frequency gradients. This encourages the optimization to explore perturbations that rely less on model-specific features, thereby improving cross-model transferability.

From an information-theoretic perspective, low-frequency components typically encode structured and model-dependent semantic information. Let $I(x_L; f_s(x))$ denote the mutual information between the low-frequency component and the model output. Since x_L contributes significantly to model predictions, we have:

$$I(x_L; f_s(x)) \gg I(x_H; f_s(x)). \tag{13}$$

The attenuation operation reduces the contribution of x_L, thereby lowering the mutual information:

$$I(\tilde{x}; f_s(x)) < I(x; f_s(x)). \tag{14}$$

This forces adversarial perturbations to depend on features that are less specific to the surrogate model and more universally shared across different models, which enhances transferability.

From the perspective of its association with regularization, the proposed method can be interpreted as an input-space regularization mechanism. While weight decay constrains the parameter space by penalizing large weights, our approach constrains the effective input distribution by suppressing dominant frequency components. Both mechanisms reduce overfitting, but they operate in complementary domains.

3.6 Frequency Domain Attenuation Algorithm

Our method imposes a direct constraint on the input sample, implying its compatibility for integration into existing iterative methods like FGSM,MI-FGSM, DI-FGSM, TI-FGSM, etc. We demonstrate the combination of FDL and DI-FGSM in Algorithm 1, called FDL-MI-DI-FGSM. In this research work, the method of MI-FGSM can be simply formalised as:

$$\begin{cases} g_{t+1} = \mu \cdot g_t + \dfrac{\nabla_x J\left(x_{adv}^t, y\right)}{\left\|\nabla_x J\left(x_{adv}^t, y\right)\right\|_1}, \\ x_{adv}^{t+1} = \mathrm{Clip}_x^{\epsilon}\left\{x_{adv}^t + \alpha \cdot \mathrm{sign}\left(g_{t+1}\right)\right\}. \end{cases} \tag{15}$$

Here, $\mathrm{Clip}_x^{\epsilon}$ indicates that the output image is constrained within the ε-sized ball around the original image x. The term g_t accumulates gradient information with a decay factor of μ, while α denotes the step size. Generally, the parameter μ is assigned a default value of 1.0.

The DI-FGSM method is to apply an image transformation $T(\cdot)$ to the input with probability p in each iteration of I-FGSM [35] to mitigate overfitting. The variation of the gradient information of the original

DI-FGSM can be represented as:

$$T_D\left(X_{adv}^n; p\right) = \begin{cases} T_D\left(X_{adv}^n\right), & if \quad p = 0.5 \\ X_{adv}^n, & if \quad 1 - p = 0.5 \end{cases} \tag{16}$$

where $T\left(X_{adv}^n; p\right)$ is the stochastic transformation function. Here the probability p is 0.5 by default.

When applying our proposed approach to MI-DI-FGSM, the gradient information updating process can be formalized as follows:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J\Big(T_D\big(T_F\big(x_{adv}^t\big), y\big)\Big)}{\Big\|\nabla_x J\Big(T_D\big(T_F\big(x_{adv}^t\big), y\big)\Big)\Big\|_1}. \tag{17}$$

Algorithm 1: The overall algorithm of FDL-MI-DI-FGSM

Input: Given a white box network $f_\theta(x, y)$ parameterized by θ, input image x, ground truth label y, loss function $J(\cdot)$, the frequency domain attenuation constraint factor λ, adversarial perturbation δ, the size of perturbation ε, step size α, number of iterations T.

Output: Adversarial sample x_{adv}

Initialize $\alpha = \varepsilon/T, g_0 = 0, x_{adv}^0 = x$

1: **for** $i = 0 \to T - 1$ **do**

2: Obtaining the frequency attenuation factor using $\lambda = \text{random}(0, 1)$;

3: Obtaining the constrained input image $T_F\left(x_{adv}^t\right)$ using Eq. (6);

4: Applying an image transformation $T_D(\cdot)$ to the input sample;

5: Obtaining the gradient $\nabla_x J\left(T_D\left(T_F(x_{adv}^t)\right), y\right)$;

6: Updating the gradient information g_{t+1} using Eq. (15);

7: Updating adversarial sample x_{adv}^{t+1} by applying the sign gradient as $x_{adv}^{t+1} = \text{Clip}_x^{\epsilon}\left\{x_{adv}^t + \alpha \cdot \text{sign}\left(g_{t+1}\right)\right\}$;

8: **end for**

9: **return** Adversarial example $x_{adv} = x_{adv}^T$.

4 Experimental Results

4.1 Experiment Settings

In this section, we conducted a large number of experimental analyses on commonly used standard benchmark datasets. Based on the experimental results and visualized analysis, we evaluated the performance of the proposed method. Additionally, we performed ablation experiments on our method to further verify the effectiveness of our proposed approach.

Dataset settings. This investigation employed three widely recognized image classification datasets, namely CIFAR-10 [36], CIFAR-100 [36], and Tiny-ImageNet [37]. For experiments involving non-targeted attacks, images were randomly selected from the test set, contingent on the prerequisite that all models had accurately classified them beforehand. A consistent set of pristine samples was then utilized to benchmark various adversarial techniques under equivalent conditions.

Comparison methods. The baseline classification performance of different models on the three datasets is shown in Table 2. The results indicate that all neural network models adopted in the experiment achieve

promising performance on the image classification task. For the purpose of validating the performance effectiveness of our proposed method, we carried out a comprehensive comparative analysis with multiple mainstream state-of-the-art attack algorithms, covering DI-FGSM [10], S^2I-FGSM [11], as well as hybrid variant designs including MI-DI-FGSM and TI-DI-FGSM. In detail, MI-DI-FGSM, first put forward by Xie et al. [10], delivers an enhancement to the momentum iterative FGSM (MI-FGSM) [9] through the introduction of diverse input strategies (DI). In comparison, TI-DI-FGSM, established by Dong et al. [7], integrates the translation-invariant attack mechanism (TI) into the foundational architecture of DI-FGSM.

Table 2: The baseline classification performance of different models on three datasets.

Dataset	VGG19	Res18	Res50	DN121	W-Res50
Cifar10	90.3%	81.6%	82.5%	83.0%	83.8%
Cifar100	72.5%	71.2%	77.3%	75.9%	73.0%
Tiny-ImageNet	61.6%	56.2%	65.8%	61.7%	68.1%

Parameter settings. We performed uniform hyperparameter configuration for all algorithms across experiments implemented on the Cifar100 [36] and Tiny-ImageNet [37] datasets, with detailed settings listed as below: we set the maximum allowable perturbation to $\varepsilon = 20/255$, fixed the total iteration count at $T = 10$, adopted a step size of $\alpha = \varepsilon/T = 2.0/255$, and assigned a value of 20 to the parameter K.

4.2 The Evaluation of Adversarial Attack Transferability

For the purpose of examining the cross-architecture generalization ability of multiple adversarial attack frameworks, we selected four widely used deep learning models to serve as our surrogate networks, namely Wide ResNet50, ResNet50, DenseNet 121, and VGG19. Following the generation of adversarial perturbations on these surrogate models, we then applied these crafted perturbations to a range of distinct black-box target models. They include a series of Resnet models, namely ResNet18 (Res18), ResNet50 (Res50), Wide-ResNet50 (WRes50), DenseNet121 (DN121), as well as the commonly used VGG19 model. A white box scenario arises if the target network coincides with the one used for attack generation. To maintain experimental consistency, every method was evaluated using the same set of randomly sampled images drawn from the dataset. This analysis concentrates on untargeted adversarial attacks, with primary emphasis on black box configurations.

Tables 3–5 present the transferability assessment results of adversarial attacks in a non-target attack environment. The information in the table indicates that, compared with the existing iterative attack methods, the method we proposed achieves a higher attack success rate on and Tiny-ImageNet datasets, proving the effectiveness of our approach.

Table 3: Experimental results on transferability of CIFAR-10 dataset. $*$ denotes the result of the white-box attack. The best results are in bold.

Model	Attack Method	ResNet18	VGG19	DN121	Res50	WRes50
Res50	DI-FGSM [10]	65.1%	64.7%	69.6%	88.5% *	67.2%
	TI-DI-FGSM [7]	56.6%	43.3%	56.5%	86.3% *	58.5%
	MI-DI-FGSM [10]	69.6%	65.7%	73.0%	90.8% *	74.45%
	S^2I-FGSM [11]	77.2%	73.3%	78.2%	96.0% *	76.1%
	FDL-S^2I-FGSM	**78.5%**	**75.3%**	**81.4%**	**98.2%** *	**78.5%**

(Continued)

Table 3 (continued)

Model	Attack Method	ResNet18	VGG19	DN121	Res50	WRes50
VGG19	DI-FGSM [10]	42.6%	93.8% *	48.3%	42.8%	45.4%
	TI-DI-FGSM [7]	33.6%	90.7% *	43.5%	37.9%	34.5%
	MI-DI-FGSM [10]	46.2%	90.4% *	50.6%	45.1%	47.3%
	S^2I-FGSM [11]	33.7%	95.8% *	39.4%	41.5%	42.3%
	FDL-S^2I-FGSM	**58.9%**	**97.6%** *	**57.5%**	**59.7%**	**63.2%**
WRes50	DI-FGSM [10]	41.8%	32.3%	50.6%	35.4%	83.5% *
	TI-DI-FGSM [7]	35.1%	18.2%	37.7%	26.4%	71.6% *
	MI-DI-FGSM [10]	52.5%	40.8%	56.3%	42.7%	84.6% *
	S^2I-FGSM [11]	70.4%	60.5%	75.2%	60.9%	93.3% *
	FDL-S^2I-FGSM	**75.6%**	**60.5%**	**77.8%**	**64.4%**	**95.7%** *
DN121	DI-FGSM [10]	70.2%	65.6%	90.3% *	61.1%	76.4%
	TI-DI-FGSM [7]	59.7%	40.3%	87.6% *	47.5%	65.8%
	MI-DI-FGSM [10]	61.3%	58.7%	78.4% *	58.9%	67.2%
	S^2I-FGSM [11]	78.6%	77.4%	92.8% *	69.5%	83.2%
	FDL-S^2I-FGSM	**85.5%**	**77.6%**	**96.3%** *	**74.7%**	**92.4%**

Table 4: Experimental results on transferability of CIFAR-100 dataset. $*$ denotes the result of the white-box attack. The best results are in bold.

Model	Attack Method	ResNet18	VGG19	DN121	Res50	WRes50
Res50	DI-FGSM [10]	77.0%	79.4%	82.4%	99.5% *	45.3%
	TI-DI-FGSM [7]	75.6%	78.7%	81.8%	99.3% *	51.7%
	MI-DI-FGSM [10]	86.1%	85.6%	88.4%	99.2% *	64.2%
	S^2I-FGSM [11]	87.8%	88.3%	89.7%	99.6% *	65.9%
	FDL-S^2I-FGSM	**89.3%**	**88.7%**	**94.1%**	**100.0%** *	**66.6%**
VGG19	DI-FGSM [10]	69.9%	99.0% *	74.1%	64.3%	39.7%
	TI-DI-FGSM [7]	71.0%	99.2% *	75.9%	71.3%	47.2%
	MI-DI-FGSM [10]	72.4%	99.0% *	84.2%	72.5%	49.0%
	S^2I-FGSM [11]	84.3%	**100.0%** *	82.9%	79.2%	55.1%
	FDL-S^2I-FGSM	**85.0%**	**100.0%** *	**88.0%**	**84.0%**	**60.5%**
WRes50	DI-FGSM [10]	88.7%	87.9%	88.2%	87.6%	99.0% *
	TI-DI-FGSM [7]	87.1%	87.3%	88.6%	87.9%	**100.0%** *
	MI-DI-FGSM [10]	90.0%	85.2%	90.6%	90.3%	**100.0%** *
	S^2I-FGSM [11]	94.6%	91.6%	94.3%	92.0%	**100.0%** *
	FDL-S^2I-FGSM	**95.3%**	**92.0%**	**98.5%**	**94.6%**	**100.0%** *

(Continued)

Table 4 (continued)

Model	Attack Method	ResNet18	VGG19	DN121	Res50	WRes50
DN121	DI-FGSM [10]	74.2%	77.3%	99.3% *	72.9%	40.5%
	TI-DI-FGSM [7]	75.3%	76.4%	99.0% *	76.8%	47.6%
	MI-DI-FGSM [10]	80.9%	80.1%	99.5% *	78.2%	40.3%
	S^2I-FGSM [11]	84.1%	82.6%	**100.0%** *	80.0%	47.1%
	FDL-S^2I-FGSM	**85.0%**	**84.9%**	**100.0%** *	**81.2%**	**48.3%**

Table 5: Experimental outcomes of adversarial attack transferability evaluated on the Tiny-ImageNet dataset. $*$ denotes the result of the white-box attack. The best results are in bold.

Model	Attack Method	ResNet18	VGG19	DN121	Res50	WRes50
Res50	DI-FGSM [10]	78.0%	62.6%	37.3%	**100.0%** *	61.7%
	TI-DI-FGSM [7]	76.3%	58.7%	37.6%	**100.0%** *	56.9%
	MI-DI-FGSM [10]	84.8%	69.7%	49.5%	99.5% *	69.2%
	S^2I-FGSM [11]	83.5%	66.0%	**55.2%**	99.6% *	**78.5%**
	FDL-S^2I-FGSM	**85.0%**	**70.1%**	54.3%	99.8% *	73.9%
VGG19	DI-FGSM [10]	71.5%	99.6% *	34.4%	52.7%	46.6%
	TI-DI-FGSM [7]	72.2%	99.7% *	33.6%	48.9%	42.9%
	MI-DI-FGSM [10]	76.3%	**100.0%** *	46.5%	64.4%	57.2%
	S^2I-FGSM [11]	79.9%	**100.0%** *	54.8%	68.1%	61.0%
	FDL-S^2I-FGSM	**80.7%**	**100.0%** *	**56.9%**	**69.5%**	**62.8%**
WRes50	DI-FGSM [10]	75.8%	63.2%	41.1%	71.3%	**100.0%** *
	TI-DI-FGSM [7]	76.3%	60.1%	41.0%	65.7%	**100.0%** *
	MI-DI-FGSM [10]	78.7%	68.0%	49.6%	78.7%	99.3% *
	S^2I-FGSM [11]	84.0%	65.5%	56.5%	83.0%	**100.0%** *
	FDL-S^2I-FGSM	**85.5%**	**72.6%**	**58.3%**	**83.8%**	99.6% *
DN121	DI-FGSM [10]	94.0%	88.9%	99.0% *	91.8%	91.4%
	TI-DI-FGSM [7]	90.5%	83.6%	**100.0%** *	86.8%	86.3%
	MI-DI-FGSM [10]	91.6%	91.0%	**100.0%** *	91.3%	90.0%
	S^2I-FGSM [11]	94.2%	91.2%	99.00% *	92.0%	91.1%
	FDL-S^2I-FGSM	**94.9%**	**91.8%**	**100.0%** *	**94.6%**	**93.7%**

4.3 The Compatibility of FDL with Other Attack Methods

In this section, we combine low-frequency frequency-domain attenuation with existing methods to further analyze the effectiveness of our approach.

In particular, we utilize advanced variants of DI-FGSM and TI-FGSM. The attack strategies that integrate frequency-domain regularization techniques are labeled as FDL-DI-FGSM, FDL-TI-DI-FGSM, FDL-MI-DI-FGSM, and FDL-S^2I-FGSM, respectively. Table 6 demonstrates the performance results of our approach when combined with current techniques. An examination of Table 6 shows that our suggested method preserves a strong attack success rate in white-box situations. Furthermore, when paired with

existing approaches, it achieves significantly improved attack success rates in black-box settings. This suggests that the adversarial examples produced using our low-frequency domain reduction technique possess superior generalization abilities. From a quantitative perspective, when the alternative model is VGG19, compared with the existing methods, our approach achieves a higher average attack success rate in the black-box scenario, thereby demonstrating the effectiveness of our method. Specifically, When the surrogate model is VGG19, FDL-DI-FGSM, FDL-TI-DI-FGSM, FDL-MI-DI-FGSM, and FDL-S^{2}I-FGSM outperform DI-FGSM, TI-DI-FGSM, MI-DI-FGSM, and S^{2}I-FGSM by an average of approximately 7.23%, 8.68%, 4.58%, and 1.53%, respectively.

Table 6: Experimental results on the compatibility of low-frequency frequency domain attenuation methods with other methods. Adversarial samples were generated by surrogate models Densenet121 and VGG19 on the Tiny-ImageNet dataset, respectively. $*$ denotes the result of the white-box attack. The best results are in bold.

Model	Attack Method	ResNet18	VGG19	DN121	Res50	WRes50
VGG19	DI-FGSM [10]	71.5%	99.6% *	34.4%	52.7%	46.6%
	TI-DI-FGSM [7]	72.2%	99.7% *	33.6%	48.9%	42.9%
	MI-DI-FGSM [10]	76.3%	**100.0%** *	46.5%	64.4%	57.2%
	S^{2}I-FGSM [11]	79.9%	**100.0%** *	54.8%	68.1%	61.0%
	FDL-DI-FGSM	77.6%	99.3% *	45.3%	58.5%	52.7%
	FDL-TI-DI-FGSM	78.3%	99.7% *	47.2%	56.3%	50.5%
	FDL-MI-DI-FGSM	80.8%	**100.0%** *	51.9%	68.1%	61.9%
	FDL-S^{2}I-FGSM	**80.7%**	**100.0%** *	**56.9%**	**69.5%**	**62.8%**
DN121	DI-FGSM [10]	94.0%	88.9%	99.0% *	91.8%	91.4%
	TI-DI-FGSM [7]	90.5%	83.6%	**100.0%** *	86.8%	86.3%
	MI-DI-FGSM [10]	91.6%	91.0%	**100.0%** *	91.3%	90.0%
	S^{2}I-FGSM [11]	94.2%	91.2%	99.00% *	92.0%	91.1%
	FDL-DI-FGSM	94.3%	89.5%	99.2% *	93.6%	92.9%
	FDL-TI-DI-FGSM	92.6%	86.7%	**100.0%** *	90.2%	90.4%
	FDL-MI-DI-FGSM	94.0%	91.1%	**100.0%** *	92.9%	92.7%
	FDL-S^{2}I-FGSM	**94.9%**	**91.8%**	**100.0%** *	**94.6%**	**93.7%**

4.4 Robustness Against Generic Defense Methods

In this section, we perform a systematic evaluation of how our proposed method performs when tested against mainstream adversarial defense strategies. For the purpose of assessing the effectiveness of adversarial attacks in bypassing defensive frameworks, we carry out a set of demonstrative experiments on the Tiny-ImageNet dataset, using three distinct defense methods built on input preprocessing [25,30]. The VGG19 network is chosen as our surrogate model for these experiments, where the preprocessing techniques are embedded into the first layer of the defense architecture and operated in coordination with the black-box target model. Among these tested defense methods, the R&C approach [25] implements random image resizing and padding operations, with the scale parameter set within the range of 0.5 and 0.8. This workflow involves randomly reducing the input image to 50%–80% of its original size, before rescaling the processed image back to its original dimensions. Additionally, the JPEG defense method [25] employs JPEG compression with a quality parameter fixed at 50%. In contrast, the R&S method [30] incorporates Gaussian smoothing via a Gaussian kernel with a size of 3 and a standard deviation set to 2.

The findings displayed in Table 7 highlight the effectiveness of adversarial attacks when confronted with different defense mechanisms. From the tabulated data, it is clear that, under non-targeted attack conditions, the success rate of transfer attacks for all methods is considerably reduced when applied to the defense model compared to a standard training model. This suggests that the defense strategies successfully reduce the influence of adversarial disturbances. Furthermore, among the three defense approaches, the introduced method demonstrates superior performance, surpassing the traditional iterative attack method in terms of success rate.

Table 7: Attack success rates on three defence methods. Adversarial examples were generated by the surrogate model VGG19 on Tiny-ImageNet. The best results are in bold.

Defence	Attack Method	ResNet18	VGG19	DN121	Res50	WRes50
R&C	DI-FGSM [10]	45.1%	86.6%	37.5%	45.9%	43.8%
	TI-DI-FGSM [7]	42.8%	85.4%	35.2%	41.1%	39.1%
	MI-DI-FGSM [10]	47.5%	87.6%	42.8%	48.5%	45.9%
	S^2I-FGSM [11]	49.3%	88.7%	36.9%	44.6%	44.3%
	FDL-S^2I-FGSM	**55.2%**	**91.0%**	**46.5%**	**53.0%**	**48.1%**
JPEG	DI-FGSM [10]	37.6%	98.3%	24.7%	31.5%	27.6%
	TI-DI-FGSM [7]	37.0%	97.2%	22.1%	29.7%	26.8%
	MI-DI-FGSM [10]	40.7%	96.9%	29.4%	40.7%	34.0%
	S^2I-FGSM [11]	48.9%	97.3%	33.7%	43.2%	43.1%
	FDL-S^2I-FGSM	**57.6%**	**98.0%**	**35.0%**	**45.1%**	**45.8%**
R&S	DI-FGSM [10]	25.6%	80.3%	21.7%	26.1%	25.6%
	TI-DI-FGSM [7]	32.2%	83.9%	**27.8%**	34.4%	28.3%
	MI-DI-FGSM [10]	27.3%	76.8%	21.6%	26.3%	28.0%
	S^2I-FGSM [11]	25.2%	74.1%	15.80%	23.5%	21.9%
	FDL-S^2I-FGSM	**37.5%**	**85.3%**	22.5%	**33.6%**	**28.9%**

4.5 CAM Attention Visualization

In this section, we utilize Gradient-weighted Class Activation Mapping [38] as a visualization tool to evaluate the effectiveness of our method in non-targeted attack situations.This tool can be abbreviated or abbreviated as Grad-CAM. Specifically, we utilized ResNet50 as a substitute for the initial black-box model and examined the model's behavior. In this evaluation, ResNet18 was designated as the initial black-box model.

Fig. 4 illustrates Grad-CAM heatmaps generated from adversarial examples under non-targeted attack conditions. Rows one and two display the metrics obtained on Cifar100, while rows three and four summarize those from Tinyâ ImageNet. In this visualization, the first and third rows contain images of the untouched original inputs, as well as the adversarial examples crafted by multiple distinct attack algorithms. On the contrary, the second and fourth rows present the attention heatmaps of the aforementioned adversarial samples, which are acquired by implementing Grad-CAM on the ResNet18 network. Through qualitative visual inspection, we observe that the target black-box model initially concentrates its focus on the most prominent regions of the unmodified clean images. Nevertheless, after we apply a variety of adversarial attack methods, the corresponding heatmaps show dramatic changes in the model's attention distribution. It is worth highlighting that when we compare the heatmaps generated by our proposed method with those

of the original clean inputs, we observe significant offsets, most notably in the areas highlighted in red. This phenomenon suggests that our attack strategy successfully diverts the model's attention from the key discriminative regions to irrelevant background features, which ultimately results in misclassification. These visual observations further validate the effectiveness of our proposed attack methodology.

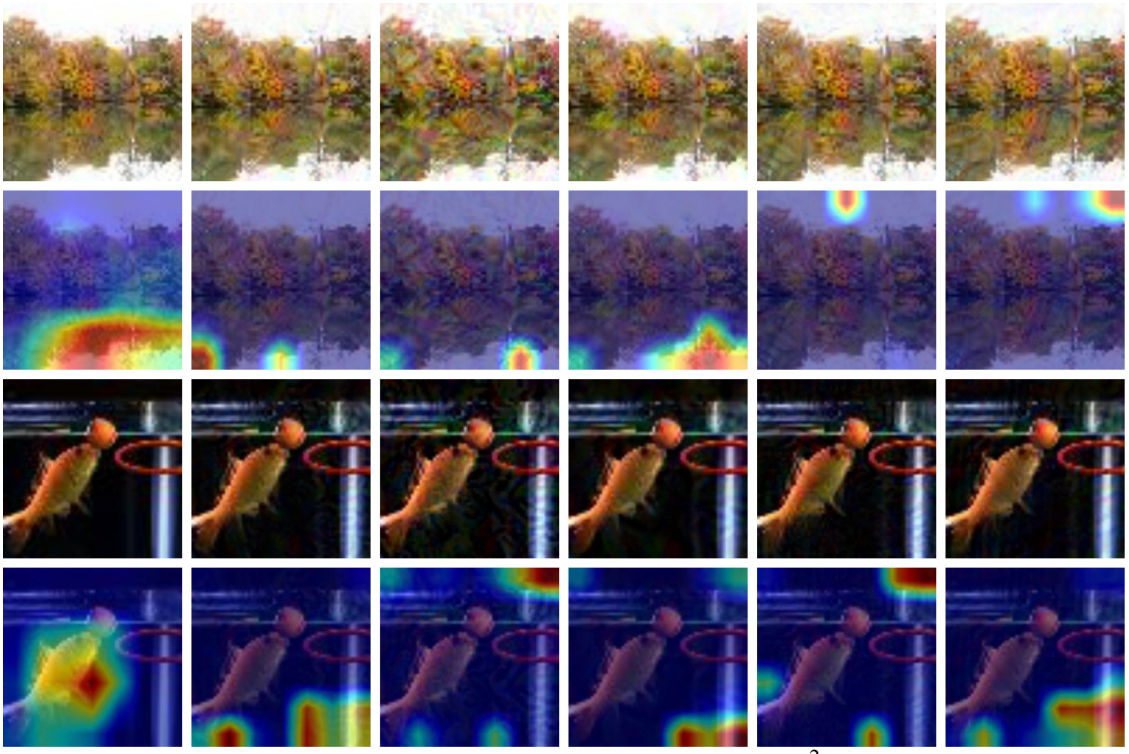


Figure 4: Visualization outcomes of Grad-CAM attention heat maps generated across the datasets of Cifar100 and Tiny-ImageNet.

4.6 Ablation Studies

We systematically evaluated the contribution of fixed constraint parameters and random constraint parameters λ to the model performance through ablation experiments. We know that the adversarial samples suffer from the overfitting problem. In order to investigate the effect of the proposed frequency domain decay constraint strategy, we use our method to generate adversarial samples with a constant constraint factor $\lambda = 0.45$. In other words, rather than employing a random selection strategy, we opt for a constant constraint value in the proposed method. In contrast to the previous low-frequency fading method, this method adds fixed constraints to the input image, using fixed constraints instead of random constraint values. Fig. 5 shows the comparison results using the fixed frequency decay constraint and the random decay constraint.

As illustrated in Fig. 5, the fixed constraint factor approach achieves a markedly higher success rate in generating adversarial samples compared to traditional iterative techniques. This finding underscores the effectiveness of the frequency domain decay approach, utilizing a fixed constraint factor to constrain frequency domain inputs, in mitigating overfitting of adversarial samples to the white-box network. Furthermore, the results in Fig. 5 demonstrate that the generalization ability of adversarial samples can be further enhanced by replacing fixed constraints with our proposed randomized frequency domain decay

constraint strategy. This observation validates the superiority of employing a random frequency domain decay constraint strategy, which can further mitigate the overfitting of adversarial samples.

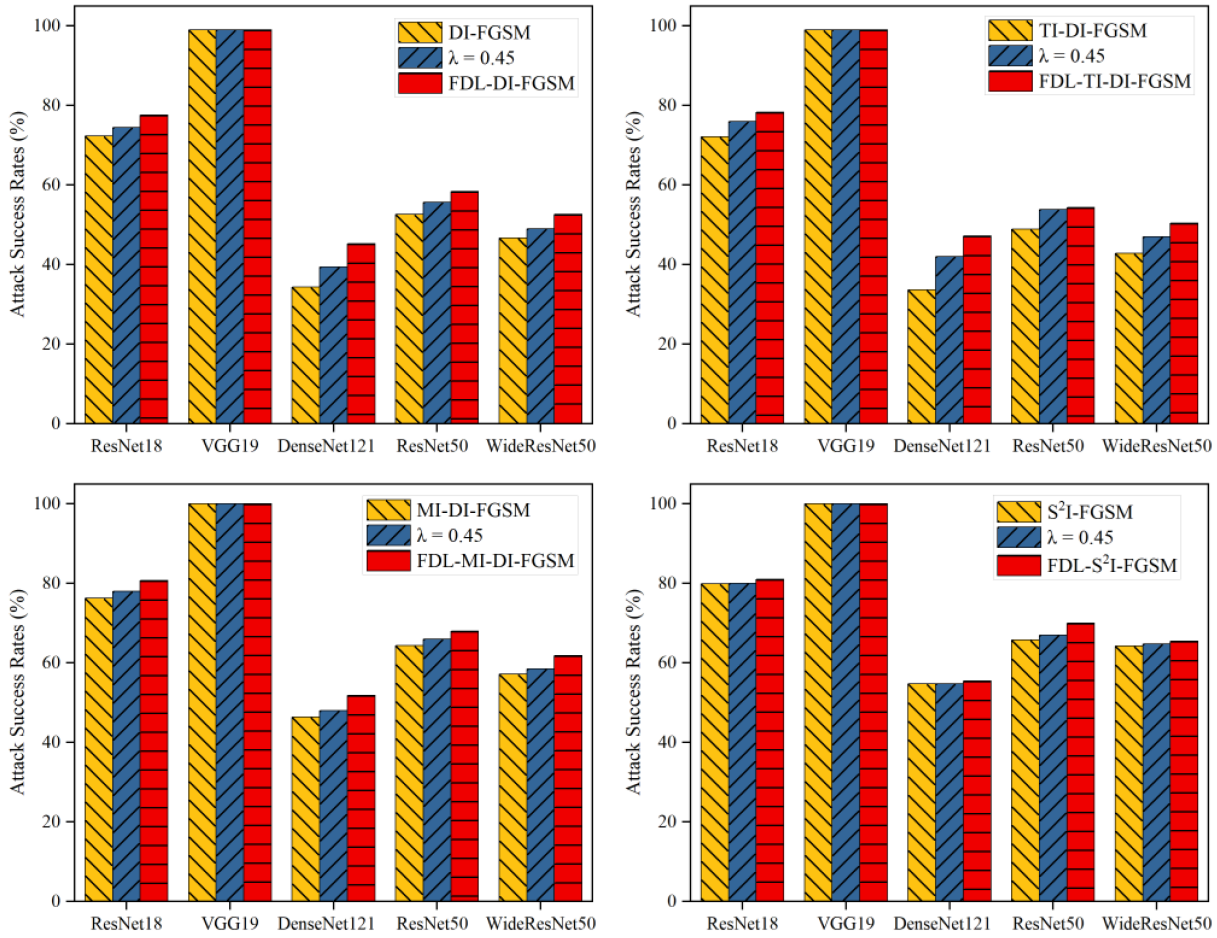


Figure 5: Comparative results using fixed frequency decay constraints and random decay constraint strategies on the Tiny-ImageNet dataset. The alternative model is VGG19.

Different thresholds have an impact on the performance of the proposed method. To determine the optimal threshold ϵ_1, we conduct an ablation analysis on the threshold parameter ϵ_1. Fig. 6 presents the attack success rates of the proposed method on the Tiny-ImageNet dataset under different threshold settings. As can be observed from Fig. 6, the proposed method achieves the best attack success rate when the threshold ϵ_1 is set to 0.7. Namely, under this threshold, our method is capable of generating adversarial examples with stronger transferability.

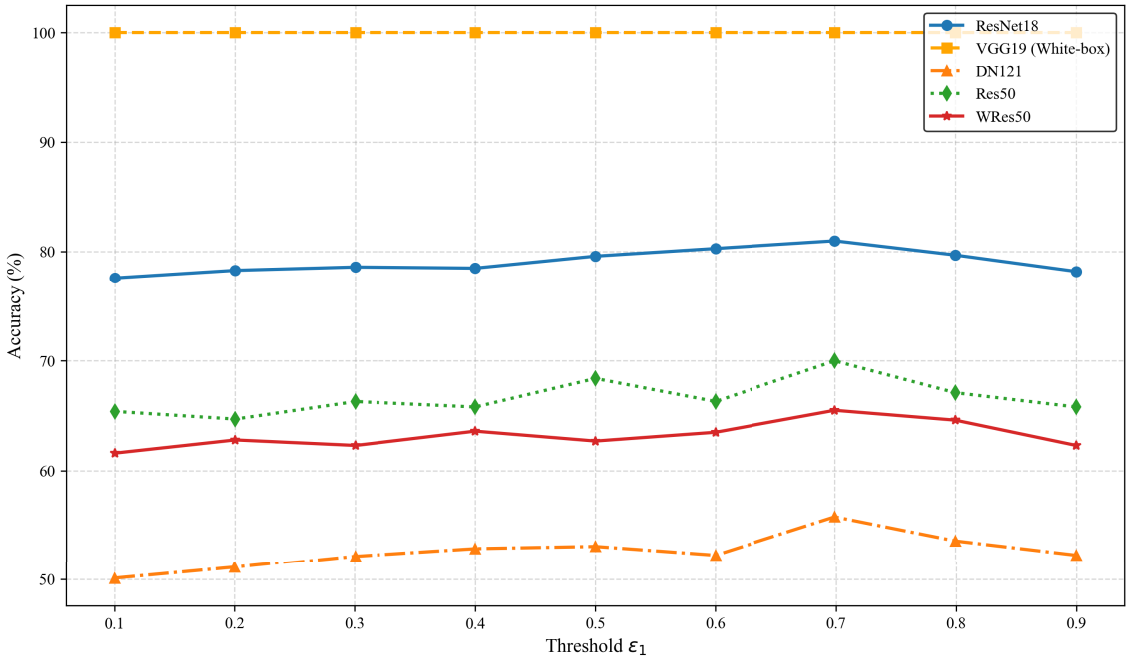


Figure 6: Comparative results of method FDL-S^{2}I-FGSM on the dataset Tiny-ImageNet under the different threshold ϵ_1. The surrogate model is VGG19. $*$ denotes the result of the white-box attack.

5 Conclusion

In this study, the analysis we conducted using convolutional neural networks for frequency-domain analysis and Fourier heat maps indicates that low frequencies contribute more to the model output during model training and fitting. To enhance the transferability of the black-box adversarial attack, we regard the iterative attack as a training process for the input data and introduce a new low-frequency attenuation method in the frequency domain. This method reduces the excessive reliance of adversarial sample generation on low-frequency information, thereby improving the transferability of adversarial samples. Our experiments on three datasets show that this technique achieves an excellent success rate of migration attacks, indicating that it can enhance the transferability of confrontation. The proposed frequency-domain attenuation method is highly dependent on Fourier domain interpretable analysis results, and variations in interpretable approaches can influence the enhancement of transferability. In future research, we will develop an adaptive frequency-domain attenuation method according to the results of Fourier heatmap analysis for each model, aiming to enhance the transferability of adversarial examples against diverse model architectures. By generating higher-quality opposing samples, we aim to evaluate the model robustness in actual scenarios more accurately and develop more robust neural network models.

Acknowledgement: Not applicable.

Funding Statement: This research was funded by the Hubei Provincial Natural Science Foundation under Grant 2025AFB416, the Open Fund Project of Shiyan Key Laboratory of Electromagnetic Induction and Energy Saving Technology SYZDK22024A02, the Doctoral Research Startup Project BK202513, and the Hubei Provincial Natural Science Foundation under Grant 2024AFB511.

Author Contributions: Conceptualization, Li Peng; methodology, Li Peng; software, Li Peng; validation, Xiangbing Li and Kun Zou; formal analysis, Kun Zou; investigation, Kun Zou; resources, Yong Liu; data curation, Xiangbing Li;

writing—original draft preparation, Li Peng and Xiangbing Li; writing—review and editing, Li Peng and Xiangbing Li; visualization, Yong Liu; supervision, Yong Liu, Haibo Huang and Kun Zou; project administration, Yong Liu; funding acquisition, Yong Liu. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: This study conducts empirical analyses on publicly accessible image classification benchmark datasets, the official sources of which are referenced and available as https://www.cs.toronto.edu/~kriz/cifar.html (accessed 10 February 2026) and http://cs231n.stanford.edu/tiny-imagenet-200.zip (accessed 10 February 2026). The source code can be obtained by contacting the author at the email: liuyong@huat.edu.cn.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Goodfellow IJ, Shlens J, Szegedy C. Explaining and harnessing adversarial examples. arXiv: 1412.6572. 2014.
2. Jin X, Liu Y, Sun G, Chen Y, Dong Z, Wu H. KRT-FUAP: key regions tuned via flow field for facial universal adversarial perturbation. Appl Sci. 2024;14(12):4973. doi:10.3390/app14124973.
3. Papernot N, McDaniel P, Goodfellow I, Jha S, Celik ZB, Swami A. Practical black-box attacks against machine learning. In: Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security; 2017 Apr 2–6; Abu Dhabi, United Arab Emirates. p. 506–19.
4. Xu L, Yao Z, Li H, Diao C, Zhou G, Zhao D, et al. CBAT-ASG: adversarial sample generation method based on CBA-transformer. In: Proceedings of the 2025 28th International Conference on Computer Supported Cooperative Work in Design (CSCWD); 2025 May 5–7; Compiegne, France. New York, NY, USA: IEEE; 2025. p. 1646–51.
5. Li C, Liu Y, Zhang X, Wu H. Exploiting frequency characteristics for boosting the invisibility of adversarial attacks. Appl Sci. 2024;14(8):3315. doi:10.3390/app14083315.
6. Liu Y, Chen X, Liu C, Song D. Delving into transferable adversarial examples and black-box attacks. arXiv:1611.02770. 2016.
7. Dong Y, Pang T, Su H, Zhu J. Evading defenses to transferable adversarial examples by translation-invariant attacks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA, USA. p. 4312–21.
8. Liu Y, Li C, Wang Z, Wu H, Zhang X. Transferable adversarial attack based on sensitive perturbation analysis in frequency domain. Inf Sci. 2024;678(1):120971. doi:10.1016/j.ins.2024.120971.
9. Dong Y, Liao F, Pang T, Su H, Zhu J, Hu X, et al. Boosting adversarial attacks with momentum. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 9185–93.
10. Xie C, Zhang Z, Zhou Y, Bai S, Wang J, Ren Z, et al. Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA, USA. p. 2730–9.
11. Long Y, Zhang Q, Zeng B, Gao L, Liu X, Zhang J, et al. Frequency domain model augmentation for adversarial attack. In: Proceedings of the Computer Vision-ECCV 2022: 17th European Conference; 2022 Oct 23–27; Tel Aviv, Israel. Berlin/Heidelberg, Germany: Springer; 2022. p. 549–66.
12. Kurakin A, Goodfellow IJ, Bengio S. Adversarial examples in the physical world. In: Artificial intelligence safety and security. Boca Raton, FL, USA: Chapman and Hall/CRC; 2018. p. 99–112.
13. Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. arXiv:1706.06083. 2017.
14. Moosavi-Dezfooli SM, Fawzi A, Frossard P. Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2016 Jun 27–30; Las Vegas, NV, USA. p. 2574–82.
15. Wiyatno R, Xu A. Maximal jacobian-based saliency map attack. arXiv:1808.07945. 2018.

16. Carlini N, Wagner D. Towards evaluating the robustness of neural networks. In: Proceedings of the 2017 IEEE Symposium on Security and Privacy (SP); 2017 May 22–26; San Jose, CA, USA. New York, NY, USA: IEEE; 2017. p. 39–57.
17. Lin J, Song C, He K, Wang L, Hopcroft JE. Nesterov accelerated gradient and scale invariance for adversarial attacks. arXiv:1908.06281. 2019.
18. Gao L, Zhang Q, Song J, Liu X, Shen HT. Patch-wise attack for fooling deep neural network. In: Proceedings of the Computer Vision–ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK. Berlin/Heidelberg, Germany: Springer; 2020. p. 307–22.
19. 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: Proceedings of the Computer Vision-ECCV 2020: 16th European Conference; 2020 Aug 23–28; Glasgow, UK. Berlin/Heidelberg, Germany: Springer; 2020. p. 563–79.
20. Wang X, He K. Enhancing the transferability of adversarial attacks through variance tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2021 Jun 20–25; Nashville, TN, USA. p. 1924–33.
21. Li T, Li M, Yang Y, Deng C. Frequency domain regularization for iterative adversarial attacks. Pattern Recognit. 2023;134(3):109075. doi:10.1016/j.patcog.2022.109075.
22. Li S, He C, Ma X, Zhu BB, Wang S, Hu H, et al. Enhancing adversarial transferability with checkpoints of a single model's training. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025 Jun 17–20; Nashville, TN, USA. p. 20685–94.
23. Ren Y, Zhao Z, Lin C, Yang B, Zhou L, Liu Z, et al. Improving adversarial transferability on vision transformers via forward propagation refinement. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025 Jun 17–20; Nashville, TN, USA. p. 25071–80.
24. Liang K, Dai X, Li Y, Wang D, Xiao B. Improving transferable targeted attacks with feature tuning mixup. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025 Jun 17–20; Nashville, TN, USA. p. 25802–11.
25. Guo C, Rana M, Cisse M, Van Der Maaten L. Countering adversarial images using input transformations. arXiv:1711.00117. 2017.
26. Dziugaite GK, Ghahramani Z, Roy DM. A study of the effect of jpg compression on adversarial images. arXiv:1608.00853. 2016.
27. Rudin LI, Osher S, Fatemi E. Nonlinear total variation based noise removal algorithms. Phys D Nonlinear Phenom. 1992;60(1–4):259–68. doi:10.1016/0167-2789(92)90242-f.
28. Liao F, Liang M, Dong Y, Pang T, Hu X, Zhu J. Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; 2018 Jun 18–23; Salt Lake City, UT, USA. p. 1778–87.
29. Xie C, Wu Y, Maaten LVD, Yuille AL, He K. Feature denoising for improving adversarial robustness. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019 Jun 15–20; Long Beach, CA, USA. p. 501–9.
30. Jia J, Cao X, Wang B, Gong NZ. Certified robustness for top-k predictions against adversarial perturbations via randomized smoothing. arXiv:1912.09899. 2019.
31. Yan J, Yin H, Zhao Z, Ge W, Zhang H, Rigoll G. Wavelet regularization benefits adversarial training. Inf Sci. 2023;649(8):119650. doi:10.1016/j.ins.2023.119650.
32. Xu ZQJ, Zhang Y, Xiao Y. Training behavior of deep neural network in frequency domain. In: Proceedings of the Neural Information Processing: 26th International Conference, ICONIP 2019; 2019 Dec 12–15; Sydney, NSW, Australia. Berlin/Heidelberg, Germany: Springer; 2019. p. 264–74.
33. Luo T, Ma Z, Xu ZQJ, Zhang Y. Theory of the frequency principle for general deep neural networks. arXiv:1906.09235. 2019.
34. Yin D, Gontijo Lopes R, Shlens J, Cubuk ED, Gilmer J. A fourier perspective on model robustness in computer vision. In: Advances in neural information processing systems. Vol. 32. San Diego, CA, USA: NeurIPS; 2019.
35. Kurakin A, Goodfellow I, Bengio S. Adversarial machine learning at scale. arXiv:1611.01236. 2016.

36. Krizhevsky A, Hinton G. Learning multiple layers of features from tiny images. Toronto, ON, Canada: Department of Computer Science, University of Toronto; 2009.
37. Brendel W, Rauber J, Kurakin A, Papernot N, Veliqi B, Mohanty SP, et al. Adversarial vision challenge. In: The NeurIPS'18 competition: from machine learning to intelligent conversations. Berlin/Heidelberg, Germany: Springer; 2020. p. 129–53.
38. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision; 2017 Oct 22–29; Venice, Italy. p. 618–26.