

Article

Improving Transferability of Adversarial Attacks via Frequency-Consistent Regularization

Tengfei Shi ^{1,2} , Shihai Wang ^{1,2,*}  and Bin Liu ^{1,2}

¹ School of Reliability and Systems Engineering, Beihang University, Beijing 100191, China; shitfing@buaa.edu.cn (T.S.); liubin@buaa.edu.cn (B.L.)

² Science and Technology on Reliability and Environmental Engineering Laboratory, Beijing 100191, China

* Correspondence: wangshihai@buaa.edu.cn

Abstract

Adversarial examples have revealed the vulnerability of deep neural networks, and their transferability makes black-box attacks particularly concerning. However, perturbations crafted on a surrogate model often do not remain sufficiently effective on unseen target models. In this paper, we revisit this issue from a frequency-domain perspective and observe that perturbation optimization can become overly dependent on specific spectral patterns, which weakens cross-model transfer. To address this problem, we propose frequency-consistent regularization (FCR), a simple plug-in strategy that can be combined with existing iterative attacks. FCR introduces multiple low-frequency preserving views with randomly sampled frequency ranges at each iteration and optimizes perturbations across these varied views. In this way, the generated perturbations are less tied to a specific frequency configuration and show improved transferability. Experimental results show that FCR consistently improves the transfer performance of various iterative attacks. The improvement is observed not only in standard target models but also in adversarially trained models, where the gain is often more pronounced.

Keywords: adversarial examples; deep neural networks; black-box adversarial attack; transferable attack; iterative attacks; transferability

1. Introduction

Deep neural networks (DNNs) have shown strong performance in a broad range of visual recognition tasks and have been widely adopted in real-world applications [1,2]. Despite their impressive performance, it has been well established that DNNs remain vulnerable to adversarial examples (AEs), namely, visually imperceptible perturbations that can induce incorrect predictions [3–6]. This vulnerability raises serious security concerns, especially in practical scenarios where image classifiers are used in safety-sensitive or security-critical systems [7–9]. As a result, adversarial attacks [10] and defenses [11–15] have become important topics in the study of trustworthy deep learning.

Among existing threat models, black-box attacks [16] are of particular practical interest because the attacker typically has no access to the target model's internal parameters. In this setting, a common strategy is to generate adversarial examples on one or several surrogate models and then transfer them to unseen target models. Such transfer-based attacks [17–19] do not rely on query access and therefore provide a realistic and efficient way to evaluate model vulnerability. Although adversarial examples often transfer across models to some extent, their effectiveness usually drops substantially when the target architecture differs



Academic Editor: Mohamed Benbouzid

Received: 16 March 2026

Revised: 6 April 2026

Accepted: 8 April 2026

Published: 11 April 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

from the surrogate or when the target model is equipped with adversarial training or defensive preprocessing [5,11]. Improving transferability, therefore, remains a central challenge in adversarial attack research.

A large body of prior work has attempted to address this problem from the perspective of optimization, stabilization, and input transformation [2,20]. Representative strategies include momentum-based updates [18], translation-invariant gradients [21], input diversity [19,22], scale transformations [23], and various structure-preserving perturbation schemes [24,25]. These methods improve transferability by smoothing gradients, enlarging the effective attack region, or reducing overfitting to a specific surrogate decision boundary. More recently, stronger transformation-based attacks, such as the spectrum simulation attack (SSA) [26] and the structure-invariant transformation (SIA) [24], have further shown that carefully designed perturbation views can substantially enhance black-box attack performance. Nevertheless, even with these improvements, the generated perturbations often remain fragile under frequency filtering, compression, or other preprocessing-based defenses [27,28], suggesting that existing methods may still rely excessively on unstable spectral components.

These findings suggest that frequency may play a more important role in adversarial transfer than is often assumed. Traditional adversarial attacks are primarily gradient-based, in which perturbations are directly optimized in pixel space using model gradients to maximize the attack objective. In contrast, frequency-based perspectives consider the spectral composition of perturbations by analyzing or manipulating their Fourier components, leveraging the fact that neural networks exhibit different sensitivities to different frequency bands. From our analysis, clean-image classification mainly relies on low-frequency components that retain the overall semantic structure. In contrast, adversarial perturbations generated by iterative attacks depend more on components that are easier to disrupt [29]. This difference may explain why many attacks perform well on the surrogate model but degrade noticeably after transfer, a phenomenon widely observed in prior studies [19,26,30]. Although the optimized perturbation is effective for the source network, it may still rely on spectral details that are not stable across model architectures or defensive preprocessing. As a result, the attack can be sensitive to operations such as filtering [29] or compression [28], even when the visual appearance of the adversarial example changes only slightly. This suggests that further improving transferability requires more than simply strengthening the attack itself. A more important step is to guide the optimization toward perturbations that remain effective under different frequency conditions, rather than overfitting to unstable spectral patterns.

In this paper, we propose a simple yet effective plug-and-play strategy, termed frequency-consistent regularization (FCR), to improve the transferability of iterative adversarial attacks. The key idea is to constrain the perturbation so that it remains effective across multiple low-frequency-preserving views of the input image during optimization. At each iteration, instead of computing the gradient from a single image representation, we generate multiple frequency-filtered views by applying DCT-based [26] low-frequency masking with randomly sampled ratios. The gradients from these views are individually normalized and then aggregated to produce a more stable update direction. In this way, the attack is encouraged to optimize perturbations that are effective under diverse frequency conditions, rather than overfitting to a narrow set of surrogate-sensitive spectral patterns.

The proposed FCR framework is general and can be seamlessly integrated with different transferable attacks. In particular, it does not modify the underlying attack objective or transformation mechanism but regularizes the gradient computation stage through multi-view frequency sampling. This makes FCR compatible with a variety of representative baselines, including TIM [21], DIM [19], SIM [18], SEM [23], SSA [26], and

SIA [24]. This makes the resulting perturbations less sensitive to spectral distortions, and thus more likely to remain effective when transferred to other architectures or evaluated under preprocessing-based defenses.

To evaluate the proposed method, we conduct experiments in single-model [10,23], ensemble-model [24], and defense settings [13] on an ImageNet-compatible benchmark [31]. The results show that FCR can be effectively integrated into different baseline attacks and consistently improves their transfer performance on both standard and adversarially trained target models [11]. In addition, experiments against several defense methods further demonstrate that the proposed regularization produces perturbations that are more robust to spectral distortions. We also provide frequency analysis and ablation results to examine how low-frequency view sampling contributes to the overall improvement.

The contributions of this paper can be summarized as follows:

- We analyze transferable adversarial perturbations in the frequency domain and observe that they do not align well with the frequency characteristics of natural image semantics. This mismatch provides one explanation for why their effectiveness often drops under model transfer and defensive preprocessing.
- We propose frequency-consistent regularization (FCR), a plug-in module that can be combined with existing iterative attacks by introducing multiple randomly sampled low-frequency views during optimization.
- We adopt a normalized multi-view gradient aggregation strategy to obtain a more stable update direction and to reduce the attack's dependence on narrow surrogate-specific spectral components.
- Extensive experiments verify the effectiveness of FCR under single-model transfer, ensemble-model transfer, and defense settings.

2. Related Work

2.1. Adversarial Attacks

Deep neural networks (DNNs) have been shown to be vulnerable to adversarial examples [3,10], which are intentionally perturbed inputs that induce incorrect predictions. Despite their small magnitude, these perturbations can significantly affect model outputs while remaining visually indistinguishable from the original inputs. This vulnerability has motivated extensive research on adversarial attacks. In particular, transfer-based attacks have received significant attention because they enable effective black-box attacks without requiring access to the target model's internal parameters or gradients.

Formally, given a classifier $f(\cdot)$ and an input image x with its ground-truth label y , the goal of an adversarial attack is to generate a perturbed sample $x^{adv} = x + \delta$ within a bounded perturbation budget such that the prediction of the model is altered. An adversarial attack can be formulated as:

$$\max_{\delta \in \mathbb{R}^d} J(f(x + \delta), y) \quad \text{s.t.} \quad \|\delta\|_p \leq \epsilon. \quad (1)$$

where $\delta \in \mathbb{R}^d$ denotes the perturbation, $\epsilon > 0$ controls its magnitude under the ℓ_p -norm.

Early adversarial attacks were formulated as constrained optimization problems. A representative example is the box-constrained L-BFGS [3] attack, which generates adversarial examples by solving an optimization problem under perturbation constraints. Subsequently, the Fast Gradient Sign Method (FGSM) [4] proposed an efficient gradient-based approach that produced adversarial perturbations through a single update along the gradient direction. To further improve attack effectiveness, the Basic Iterative Method (BIM) [5] extended the FGSM by replacing the single large update step with multiple small iterative updates, resulting in more stable and accurate adversarial perturbations.

Later studies revealed that adversarial examples generated on one model could also mislead other models, a phenomenon known as adversarial transferability. This property has motivated extensive research on transfer-based attacks, where adversarial examples crafted on a surrogate model are expected to remain effective against unseen target models. To enhance the transferability of iterative attacks, the Momentum Iterative Method (MIM) [30] introduces a momentum term into the gradient update process, stabilizing optimization trajectories and helping escape poor local optima. Building upon this idea, the Nesterov Iterative Fast Gradient Sign Method (NI-FGSM) [18] further improves optimization by incorporating Nesterov accelerated gradients to obtain more informative update directions.

Another line of work improves transferability through input transformation strategies. The Diverse Input Method (DIM) [19] applies stochastic transformations to the input image at each iteration, such as random resizing and padding, so that the perturbation is optimized over multiple transformed inputs rather than a single image. Similarly, the Translation-Invariant Method (TIM) [21] enhances transferability by optimizing perturbations over translated versions of the input image. This is implemented efficiently by convolving the gradients with a predefined kernel that approximates the gradients of multiple translated inputs.

Scale-based transformation strategies have also been widely explored. The Scale-Invariant Method (SIM) [18] aggregates gradients computed from multiple scaled copies of the input image, encouraging the generated perturbations to remain effective across different image resolutions. The Scale Ensemble Method (SEM) [23] further improves this idea by integrating gradients from a set of scale transformations into the optimization process.

In addition, several recent methods introduce more complex transformations to improve cross-model generalization. For example, the Admix method [25] enhances transferability by mixing the current image with randomly selected images from other categories, thereby producing more diverse gradients. The Structure Invariant Attack (SIA) [24] generates multiple transformed samples that preserve the original image's structural information, thereby encouraging perturbations to remain effective across various structural variations.

Recent studies have explored adversarial perturbation generation in the frequency domain, where the spectral characteristics of perturbations are explicitly manipulated to improve transferability. For example, the Spectrum Simulation Attack (SSA) [26] introduces spectral-domain sampling to simulate perturbations with diverse frequency compositions. Specifically, the SSA transforms the input into the frequency domain and perturbs spectral coefficients via stochastic sampling, enabling the attack to explore a broader range of frequency patterns during optimization. By incorporating multiple spectral variants in each iteration, the SSA reduces the dependence on a specific perturbation pattern and improves the transferability of adversarial examples.

2.2. Adversarial Defenses

Adversarial defense methods have been widely studied to improve the robustness of deep neural networks against adversarial perturbations. One major direction is adversarial training (AT) [11,32], which enhances robustness by incorporating adversarial examples during training, enabling the model to learn more stable decision boundaries. Variants such as ensemble adversarial training [33] further improve robustness by introducing perturbations transferred from multiple surrogate models.

Another important class of defenses relies on preprocessing techniques that suppress adversarial perturbations in the input space. Compression-based methods, including JPEG compression [34], and Feature Distillation (FD) [27], reduce adversarial noise by removing high-frequency artifacts while preserving semantic content.

Randomized transformations have also been explored to mitigate adversarial effects. Random Resize and Padding (R&P) [35] applies stochastic geometric transformations to the input image, while bit-depth reduction (Bit-Red) [28] compresses the color representation of images to suppress small perturbations. Random Smoothing (RS) [29] further improves robustness by averaging model predictions over randomly perturbed inputs.

In addition, several approaches attempt to restore clean image representations through purification or reconstruction mechanisms. Pixel Deflection (PD) [36] redistributes local pixel values to disrupt adversarial patterns. High-Level Representation Guided Denoiser (HGD) [37] and Neural Representation Purifier (NRP) [14] learn purification networks guided by feature representations of pretrained models. Image Super-Resolution (SR) [38] has also been explored for recovering image structure and mitigating adversarial artifacts.

3. Approach

3.1. Notions and Definitions

Given a trained classification model $f(x) : x \in \mathcal{X} \rightarrow y \in \mathcal{Y}$, the model predicts a label y for an input sample x , where the ground-truth label of x is denoted by y^{true} . Adversarial attacks aim to construct a small perturbation δ added to the clean input such that the classifier produces an incorrect prediction while the perturbation remains imperceptible. Formally, the adversarial example can be written as $x^{\text{adv}} = x + \delta$.

To ensure that the perturbation remains visually imperceptible, it is typically constrained within a bounded region under an ℓ_p -norm, where $p \in \{1, 2, \infty\}$. The generation of adversarial examples can therefore be formulated as the following optimization problem:

$$\begin{aligned} \arg \max_{x^{\text{adv}}} J(f(x^{\text{adv}}), y^{\text{true}}) \\ \text{s.t. } \|x^{\text{adv}} - x\|_p \leq \epsilon, \end{aligned} \quad (2)$$

where $J(\cdot)$ denotes the loss function of the classifier, typically the cross-entropy (CE) loss, and ϵ controls the maximum perturbation magnitude.

In practice, adversarial examples are often generated through iterative optimization. Starting from the original input x , the adversarial example is progressively updated along the gradient direction of the loss function while enforcing the perturbation constraint at each step. The iterative update can be generally expressed as

$$x_{t+1}^{\text{adv}} = \Pi_{\mathcal{B}(x, \epsilon)} \left(x_t^{\text{adv}} + \alpha \cdot \text{sign} \left(\nabla_x J(f(x_t^{\text{adv}}), y^{\text{true}}) \right) \right), \quad (3)$$

where α denotes the step size, and $\Pi_{\mathcal{B}(x, \epsilon)}(\cdot)$ represents the projection operator that ensures the perturbation remains within the ℓ_p -norm constraint.

3.2. Frequency Characteristics of Adversarial Perturbations

To better characterize adversarial perturbations, we analyzed the frequency distributions of clean and adversarial images. In general, the main semantic content of natural images lies in low-frequency components, whereas high-frequency components mainly reflect finer details. This motivated us to investigate how model predictions changed when different frequency components were progressively removed.

We first analyzed the classification performance of several ImageNet models when only the low-frequency components of input images were preserved. To this end, we used the two-dimensional Discrete Cosine Transform (DCT) to represent the input image in the frequency domain, enabling analysis of information at different spatial frequencies.

Given an input image $x \in \mathbb{R}^{H \times W}$, the 2D DCT is defined as

$$X(u, v) = \alpha(u)\alpha(v) \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x(i, j) \cos\left(\frac{\pi(2i+1)u}{2H}\right) \cos\left(\frac{\pi(2j+1)v}{2W}\right), \quad (4)$$

where $X(u, v)$ denotes the frequency coefficient at location (u, v) , and $\alpha(\cdot)$ is a normalization factor. In the DCT spectrum, low-frequency components are concentrated near the top-left region, while higher-frequency components appear toward the bottom-right. For RGB images, the above transformation is applied independently to each color channel.

To control the amount of preserved frequency information, we introduced a low-pass frequency mask M_r parameterized by a ratio $r \in [0, 1]$. Specifically, we retained only the coefficients within a square region of size $rH \times rW$ and set the remaining coefficients to zero:

$$\tilde{X}(u, v) = M_r(u, v) \cdot X(u, v), \quad (5)$$

where the mask M_r is defined as

$$M_r(u, v) = \begin{cases} 1, & u < rH, v < rW, \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

The filtered spectrum $\tilde{X}$ was then transformed back to the spatial domain using the inverse DCT (IDCT):

$$\tilde{x} = \text{IDCT}(\tilde{X}), \quad (7)$$

which yielded a reconstructed image using only the frequency coefficients retained by the mask M_r .

By varying the ratio r , we progressively controlled the number of preserved frequency components and evaluated the classification accuracy of several ImageNet models on the reconstructed images. As shown in Figure 1, when the low-frequency ratio increases, the model accuracy rapidly recovers and stabilizes. Even when only a small portion of low-frequency components is retained, the models can already achieve relatively high classification accuracy. This observation suggests that the semantic information required for image recognition is largely preserved in the low-frequency spectrum.

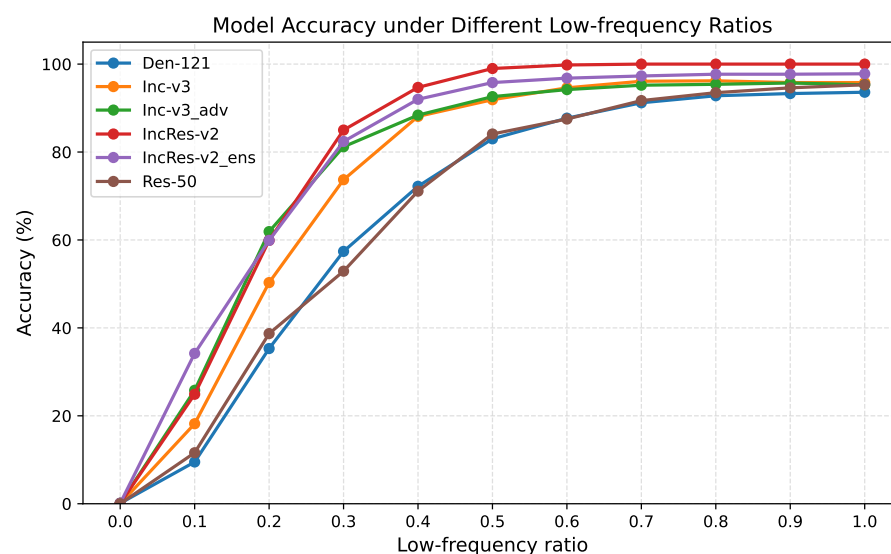


Figure 1. Classification accuracy of ImageNet models on clean images under different low-frequency preservation ratios in the DCT domain. Model performance rapidly recovers as more low-frequency components are preserved.

We further performed the same frequency analysis on adversarial examples generated by the Momentum Iterative Method (MIM). The resulting adversarial images were then transformed into the frequency domain using the same DCT transformation described above. A low-pass frequency mask with ratio r was applied to the DCT coefficients, and the filtered spectrum was reconstructed via inverse DCT to obtain frequency-controlled adversarial images.

Figure 2a shows the untargeted attack success rate under different low-frequency ratios. When the ratio r is very small, only a limited portion of low-frequency components is retained, and the semantic content of the image is severely degraded. Here, a high untargeted success rate does not necessarily mean a strong perturbation. It may simply be caused by the loss of too much image information, which makes the original class hard to recognize. As r increases, more low-frequency semantic structure is gradually restored, improving the image's recognizability and decreasing the untargeted attack success rate. As r increases, the recovered image content begins to offset the remaining perturbation, and the success rate rises again after a turning point. When r gets close to 1, more high-frequency details are brought back, so the success rate gradually moves back toward that of the original adversarial examples. This pattern suggests that low-frequency components mainly retain semantic content, while the perturbations responsible for misclassification are more concentrated in the high-frequency part.

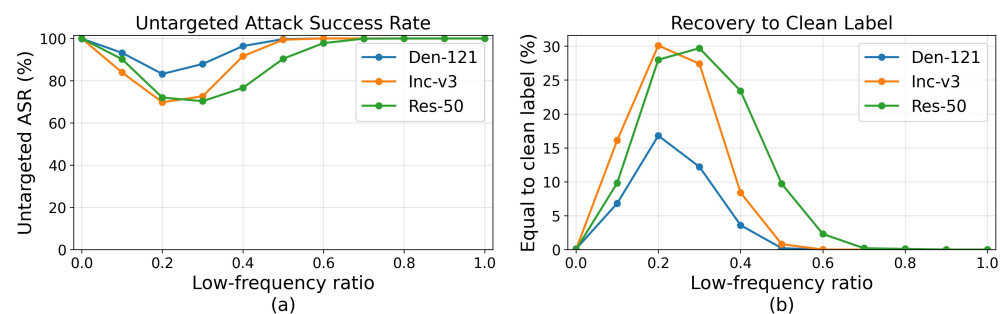


Figure 2. Effect of low-frequency preservation ratio on ImageNet model predictions in the DCT domain. (a) Untargeted attack success rate on low-frequency-preserved inputs. (b) Percentage of samples restored to the clean label. As the preserved low-frequency ratio increases, model predictions gradually recover.

Figure 2b further reports the proportion of adversarial samples whose predictions revert to the original clean labels after frequency filtering. When the low-frequency ratio r is very small, only a small portion of the image structure is preserved, and the semantic information is heavily degraded. In this case, the classifier cannot reliably recognize the original category, and therefore, the recovery rate remains close to zero. As r increases, more low-frequency semantic information is gradually restored, allowing the classifier to recover the correct predictions for a portion of the adversarial samples. Consequently, the recovery rate increases and reaches a peak when the restored semantic content becomes sufficiently recognizable, while the adversarial perturbations remain partially suppressed by the frequency filtering.

However, as r increases, more high-frequency components are reintroduced into the reconstructed image, thereby restoring the adversarial perturbations. As a result, the recovery effect gradually diminishes, and the recovery rate decreases toward zero when r approaches 1. This observation further supports the conclusion that the semantic structure of images is mainly preserved in low-frequency components, whereas the adversarial perturbations responsible for misclassification are largely concentrated in the high-frequency regions of the spectrum.

Taken together, the results show a clear mismatch in frequency between image semantics and adversarial perturbations. Classification-relevant semantic content is mainly found in low-frequency components, but gradient-based perturbations are more concentrated in high-frequency regions. This suggests that the attack optimization process tends to favor high-frequency signals that are weakly related to semantics but can still affect the model's output. These observations suggest that transferability is closely related to the stability of frequency components across models. This motivates the design of a mechanism that can reduce reliance on unstable spectral patterns, which we address in the following section.

3.3. Frequency-Consistent Regularization

The analysis in the previous subsection reveals a clear discrepancy in frequency between natural image semantics and adversarial perturbations. Specifically, model recognition on clean images can be largely preserved by low-frequency components. In contrast, the effectiveness of adversarial perturbations is much more sensitive to the removal of high-frequency signals. This suggests that conventional gradient-based attacks tend to over-rely on frequency components that are effective on the surrogate model but fragile under model transfer or defensive preprocessing. Motivated by this observation, we propose frequency-consistent regularization (FCR), which encourages adversarial perturbations to remain effective across multiple low-frequency-preserving views of the input. Rather than optimizing the perturbation on a single image representation, FCR regularizes the optimization process by introducing frequency diversity during each iteration. The proposed strategy is simple, model-agnostic, and can be seamlessly integrated into different iterative transfer attacks.

3.3.1. Interpretation from Gradient Alignment Perspective

While the proposed method is motivated by frequency-domain observations, its effectiveness can be further understood from the perspective of gradient alignment. In standard iterative attacks, perturbations are optimized using gradients from a surrogate model, which inherently reflect the model's specific spectral response. This often leads to surrogate-specific overfitting, in which the perturbation relies on frequency components that are not stable across architectures.

In particular, high-frequency components tend to exhibit significant variability across models, whereas low-frequency structures are more consistently aligned with semantic information. Such spectral variability leads to poor alignment between gradients across different models, limiting transferability.

The proposed frequency-consistent regularization addresses this issue by aggregating gradients over multiple low-frequency-preserving transformations:

$$\bar{g} = \mathbb{E}_T[\nabla_x J(f(T(x)), y)], \quad (8)$$

which can be interpreted as optimizing the perturbation with respect to an expectation over frequency-preserving transformations, rather than with respect to a single input instance.

Since these transformations preserve low-frequency semantics while perturbing high-frequency details, the aggregated gradient suppresses unstable high-frequency components and emphasizes frequency patterns that are consistent across views. This process reduces surrogate-specific spectral bias and improves cross-model gradient alignment, thereby enhancing transferability. From this perspective, the proposed method can be viewed as approximating a transformation-robust gradient that is less sensitive to spectral variations.

3.3.2. Random Low-Frequency Sampling

Given an input image $x \in \mathbb{R}^{C \times H \times W}$, we first transform it into the frequency domain using the two-dimensional Discrete Cosine Transform (DCT). As analyzed in the previous section, low-frequency components are concentrated in the top-left region of the DCT spectrum and account for most of the semantic structure required for recognition. To construct views preserving low frequencies, we apply a binary mask M_r Equation (6) parameterized by a ratio $r \in (0, 1]$. The corresponding low-frequency view is reconstructed by inverse DCT:

$$T_r(x) = \text{IDCT}(M_r \odot \text{DCT}(x)), \quad (9)$$

where $\odot$ denotes multiplication by an element.

Instead of fixing the same ratio throughout optimization, we randomly sample multiple low-frequency views at each iteration. Specifically, for the t th iteration, we first sample a minimum ratio $r_{\min}^{(t)}$ from

$$r_{\min}^{(t)} \sim \mathcal{U}(0.05, 0.15), \quad (10)$$

and then generate N random ratios

$$r_i^{(t)} \sim \mathcal{U}(r_{\min}^{(t)}, 1), \quad i = 1, \dots, N. \quad (11)$$

In our implementation, we set $N = 10$ for each iteration. This strategy introduces diverse frequency-preserving views ranging from strongly filtered inputs to nearly complete images, thereby reducing the risk that optimization becomes overly dependent on a narrow frequency band. We further examine the sensitivity of the proposed method to the choice of the $r_{\min}$ sampling range and the number of sampled views N in the ablation study.

3.3.3. Multi-View Gradient Aggregation

Based on the sampled ratios, we obtain a set of transformed inputs $\{T_{r_i^{(t)}}(x_t)\}_{i=1}^N$. Instead of computing the update direction from a single image view, FCR aggregates gradients over multiple sampled low-frequency views. Specifically, for each transformed input, we first compute the gradient of the attack objective and normalize it by its ℓ_1 norm:

$$\hat{g}_{t,i} = \frac{\nabla_x J(f(T_{r_i^{(t)}}(x_t)), y)}{\left\| \nabla_x J(f(T_{r_i^{(t)}}(x_t)), y) \right\|_1}, \quad i = 1, \dots, N, \quad (12)$$

where $f(\cdot)$ denotes the surrogate model, $J(\cdot)$ is the attack objective, and y is the ground-truth label. The normalized gradients from different views are then averaged to obtain the aggregated gradient:

$$g_t = \frac{1}{N} \sum_{i=1}^N \hat{g}_{t,i}. \quad (13)$$

This multi-view gradient aggregation encourages the perturbation to remain effective across varying degrees of frequency preservation, rather than relying excessively on a narrow set of frequency-sensitive components in the surrogate model. By normalizing each gradient with its ℓ_1 norm before averaging, we reduce magnitude bias across sampled views and obtain a more stable aggregated update direction. As a result, the optimized perturbation becomes less sensitive to frequency distortions and generalizes better across unseen architectures and preprocessing-based defenses. In the ablation experiments, we compare different gradient processing strategies, including direct averaging, ℓ_1 normalization, and ℓ_2 normalization.

3.3.4. Integration with Transferable Attacks

FCR is incorporated into a standard momentum-based iterative attack framework [30] and can be combined with different transfer-based baselines [18,23,24]. At iteration t , the momentum term is updated as

$$m_{t+1} = \mu m_t + \frac{g_t}{\|g_t\|_1}, \quad (14)$$

where μ is the decay factor and g_t is the aggregated gradient over the sampled low-frequency views. The adversarial example is then updated by

$$x_{t+1} = \Pi_\epsilon(x_t + \alpha \cdot \text{sign}(m_{t+1})), \quad (15)$$

where α is the step size, and $\Pi_\epsilon(\cdot)$ projects the result onto the ℓ_∞ ball of radius ϵ centered at the original image.

Importantly, FCR does not alter the core transformation mechanism of the underlying attack but only replaces the original single-view gradient computation with the proposed frequency-consistent multi-view aggregation. Therefore, it can be naturally integrated with representative transferable attacks such as the TIM [21], DIM [19], SIM [18], SEM [23], SSA [26], and SIA [24]. This plug-and-play property is also consistent with our experimental design, in which FCR is combined with various baseline attacks and evaluated under single-model transfer, ensemble-model transfer, and defense settings. Overall, by enforcing consistency across multiple randomly sampled low-frequency views, FCR guides adversarial optimization toward perturbations that are less sensitive to frequency distortions and more robust across different transfer scenarios.

3.4. Overall Framework and Implementation

As illustrated in Figure 3, the proposed FCR method generates transferable adversarial examples by iteratively optimizing a clipped perturbation in conjunction with a frequency-consistent regularization process. Starting from a clean image x , an additive perturbation δ is introduced and constrained within a prescribed budget through a clipping operation, producing an intermediate adversarial input $x_t^{adv} = x + \delta_t$. Instead of directly feeding this perturbed image into the surrogate model, FCR first passes x_t^{adv} through a frequency-consistent sampling module that preserves the low-frequency semantic structure while suppressing redundant high-frequency components. The reconstructed samples are then used as the actual inputs for surrogate-model optimization.

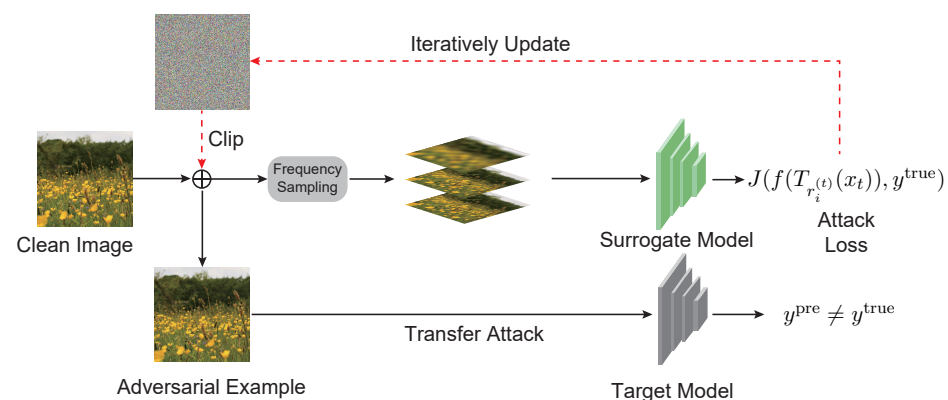


Figure 3. Framework of the Proposed FCR Method.

More specifically, at each iteration, the variants reconstructed by FCR are passed to the surrogate model, and the attack loss is computed with respect to the ground-truth

label in the untargeted setting. The resulting gradient is back-propagated to update the perturbation, and the updated perturbation is clipped to satisfy the constraint ℓ_∞ . This procedure is repeated iteratively until the perturbation converges or the maximum number of iterations is reached. Through this design, FCR does not simply optimize pixel-space perturbations; instead, it guides the attack toward frequency components that are more relevant to semantic content and more likely to generalize across model architectures. The flowchart of the proposed FCR method is shown in Figure 4, while the detailed algorithm is provided in Appendix A Algorithm A1.

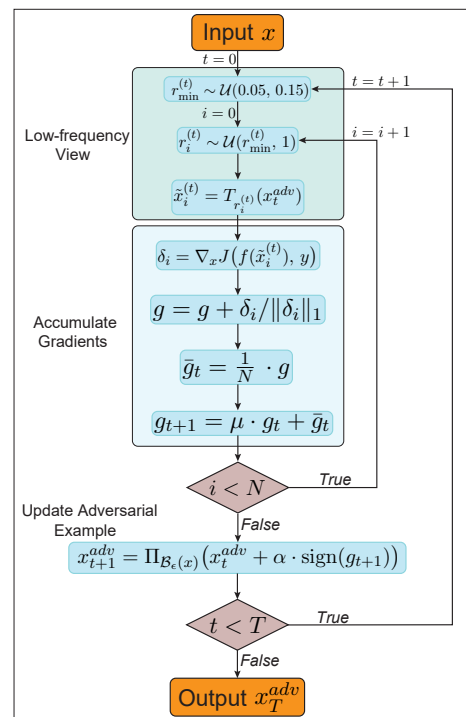


Figure 4. Flowchart of the proposed frequency-consistent regularization (FCR) method.

4. Experiments

This section first presents the experimental setup. We then report the results of the single-surrogate setting to evaluate the effectiveness of the proposed method. Next, we extend the evaluation to the ensemble-surrogate setting to examine its performance with multiple surrogate models. We further test the generated adversarial examples against several defense models to assess their robustness under defensive preprocessing. Finally, ablation studies are conducted to analyze the contribution of the main design choices.

4.1. Experiment Setup

Datasets. We conducted experiments on an ImageNet-compatible benchmark that contained 1000 natural images from the NIPS 2017 Adversarial Competition [31]. The images were sampled from the ImageNet validation set [39] and covered diverse object categories and visual scenes. Each image was paired with its ground-truth label, and all inputs were preprocessed to match the requirements of ImageNet-pretrained models [40]. This dataset provided a standard benchmark for evaluating the transferability of adversarial attacks across different model architectures.

Models. We evaluated the proposed method on six widely used ImageNet classification models, including Inc-v3 [41], Res-50 [42], Den-121 [43], IncRes-v2 [44], Inc-v3_{adv} [11], and IncRes-v2_{ens} [11]. The first four are standard classification models with different architectures, covering inception-based, residual, and densely connected networks. The last two

are adversarially trained models, which are more challenging targets for transferable attacks. All models were pretrained on ImageNet and are commonly used as benchmark models in studies of adversarial transferability. The source and target models were implemented using the PyTorch Image Models (timm) library (version 1.0.25) [40].

Baselines. We compared the proposed method with several representative transferable attacks, including the TIM [21], DIM [19], SIM [18], SEM [23], SSA [26], and SIA [24]. All these methods were implemented within the momentum-based iterative framework [30] and improve transferability through strategies such as translation-invariant gradients, input diversity, scale transformations, and structural or spectral perturbation mechanisms. These attacks provide strong baselines for evaluating the effectiveness of the proposed method.

Implementation Details. Unless otherwise specified, all attacks were conducted under the ℓ_∞ perturbation constraint with a maximum budget of $\epsilon = 16/255$. The step size was set to $\alpha = 1.6/255$, and each attack ran for 10 iterations. For all momentum-based methods, the momentum decay factor was fixed at 1.0.

For the baselines being compared, the TIM used Gaussian-smoothed gradients with a kernel size of seven and a standard deviation of 3.0. The DIM adopted the input diversity strategy with a transformation probability of 0.5. The SIM employed multi-size inputs with scale factors $\{0, 1, 2, 3, 4\}$, while SEM used the scale set $\{0.8, 0.9, 1.0, 1.1, 1.2\}$. For the SIA, the number of scales and shuffled blocks were set to five and three, respectively. For the SSA, we set $N = 10$, $\sigma = 16.0/255$, and $\rho = 0.5$.

For the proposed FCR, the regularization was integrated into each baseline attack during iterative optimization. At each iteration, we randomly sampled 10 frequency views for each input. Specifically, the minimum low-frequency ratio $r_{\min}$ was randomly drawn from the interval $[0.05, 0.15]$, and the corresponding frequency masks were applied in the DCT domain to generate multiple frequency-filtered samples. These samples were then transformed back to the spatial domain and jointly used to estimate the gradient update direction.

Evaluation Metrics. We used the attack success rate (ASR) as the evaluation metric. For a dataset of N images $\{x_i, y_i\}_{i=1}^N$, where x_i denotes the input image and y_i is its ground-truth label, let $f(\cdot)$ be the classifier and x_i^{adv} the generated adversarial example. In the untargeted setting, an attack is regarded as successful if the prediction of x_i^{adv} differs from the true label y_i . The ASR is defined as

$$ASR = \frac{1}{N} \sum_{i=1}^N \mathbf{1}(f(x_i^{adv}) \neq y_i), \quad (16)$$

where $\mathbf{1}(\cdot)$ is the indicator function. A higher ASR indicates stronger attack effectiveness and better transferability.

4.2. Single-Model Attacks

Table 1 reports the attack success rates (ASRs) of different methods under the single-surrogate setting. For each surrogate model, adversarial examples were generated on one surrogate network and then evaluated on multiple target models with different architectures. All statistics were computed over 1000 images. To ensure a fair comparison, all methods adopted momentum-based gradient updates and followed the same attack configuration. We compared several representative gradient-based attacks with their corresponding FCR-enhanced variants to examine the effectiveness of the proposed frequency-consistent regularization.

In general, incorporating FCR consistently improves the transferability of all baseline attacks. Across almost all settings, the FCR-enhanced variants achieve a higher ASR than their original counterparts, and the gains are particularly evident in many cross-model transfer scenarios. For relatively weaker baselines such as the TIM, the average ASR

increases from about 40–50% to 52–59% after introducing FCR. The SEM and DIM also show clear and stable improvements, typically on the order of 10–15 percentage points. Stronger attacks, including the SIM, SSA, and SIA, benefit even more from FCR, indicating that the proposed regularization remains effective even when combined with already competitive transfer-based attacks.

Table 1. Attack success rates (ASRs, %) of different attacks under various surrogate models. For each surrogate model, adversarial examples were generated using a single surrogate and evaluated on multiple target models. The symbol * indicates the white-box setting. Boldface highlights the better result within each attack pair.

Surrogate	Attack	Target Model						Avg
		Inc-v3	Res-50	Den-121	IncRes-v2	Inc-v3 _{adv}	IncRes-v2 _{ens}	
Inc-v3	TIM	99.7 *	28.6	37.8	26.0	31.0	21.0	40.7
	FCR-TIM	99.7 *	41.2	53.7	36.8	48.1	36.0	52.6
	SEM	100.0 *	40.8	48.8	51.5	29.1	12.3	47.1
	FCR-SEM	100.0 *	55.1	64.1	62.1	46.1	30.3	59.6
	DIM	99.9 *	54.0	59.8	64.1	35.2	14.4	54.6
	FCR-DIM	100.0 *	67.0	76.9	76.6	56.2	39.0	69.3
	SIM	100.0 *	56.9	64.1	68.1	41.0	20.2	58.4
	FCR-SIM	100.0 *	71.1	80.5	77.9	64.9	47.7	73.7
	SSA	99.1 *	66.6	72.3	83.0	55.2	23.6	66.6
	FCR-SSA	99.4 *	84.9	92.1	92.2	86.9	72.5	88.0
Res-50	SIA	99.8 *	73.9	78.8	86.2	43.6	19.1	66.9
	FCR-SIA	100.0 *	94.6	96.4	96.1	83.1	66.0	89.4
	TIM	37.5	99.2 *	42.1	22.9	31.7	24.3	43.0
	FCR-TIM	49.2	99.3 *	59.9	33.5	43.3	40.6	54.3
	SEM	52.3	100.0 *	71.0	44.9	39.0	25.6	55.5
	FCR-SEM	65.1	100.0 *	82.0	58.2	56.0	47.8	68.2
	DIM	64.7	100.0 *	77.4	57.9	46.4	32.7	63.2
	FCR-DIM	81.5	100.0 *	90.4	79.7	72.5	65.5	81.6
	SIM	63.1	100.0 *	79.3	57.0	46.5	33.0	63.2
	FCR-SIM	82.4	100.0 *	90.7	78.4	74.5	68.1	82.4
Den-121	SSA	82.1	99.8 *	89.8	78.3	67.4	52.3	78.3
	FCR-SSA	94.1	99.9 *	97.4	93.7	90.4	85.1	93.4
	SIA	79.8	99.5 *	91.4	72.6	50.5	30.5	70.7
	FCR-SIA	95.7	100.0 *	98.5	95.2	91.3	85.8	94.4
	TIM	47.9	45.0	99.9 *	30.8	43.1	36.6	50.6
	FCR-TIM	56.8	51.1	99.6 *	38.1	54.9	51.0	58.6
	SEM	70.5	81.8	100.0 *	58.5	58.1	44.1	68.8
	FCR-SEM	80.5	86.0	100.0 *	74.8	72.9	66.0	80.0
	DIM	80.5	85.9	100.0 *	72.6	68.3	51.7	76.5
	FCR-DIM	89.4	92.5	100.0 *	87.7	85.0	79.3	89.0
	SIM	78.1	87.6	100.0 *	68.6	66.0	50.8	75.2
	FCR-SIM	87.7	91.8	100.0 *	84.7	86.2	78.2	88.1
	SSA	89.2	93.8	99.8 *	86.1	81.8	65.5	86.0
	FCR-SSA	95.0	94.8	100.0 *	93.1	92.8	88.5	94.0
	SIA	91.0	96.2	100.0 *	86.7	79.1	59.0	85.3
	FCR-SIA	97.2	98.7	100.0 *	96.4	96.2	90.4	96.5

Another notable observation is that the gains from FCR are more pronounced for the two adversarially trained target models, Inc-v3_{adv} and IncRes-v2_{ens}, which are substantially harder to attack than standard models. On these robust targets, FCR often yields greater improvements, suggesting that frequency-consistent optimization helps produce perturbations less tied to fragile high-frequency artifacts and therefore transfer better under defensive training. Meanwhile, the relative performance of different attack backbones also reflects their design differences. The SSA generally performs better than the SIA on adversarially trained models, likely because the SSA introduces frequency-domain transformations that enrich spectral diversity during optimization, whereas the SIA mainly relies on structure-oriented spatial transformations. Since robust models tend to suppress unstable, high-frequency noise and place greater emphasis on semantically stable patterns, spectral-level perturbation design appears more effective in this setting. When combined with FCR, this advantage becomes even clearer, as the proposed regularization further reduces over-reliance on specific frequency components and improves transferability across both standard and robust target models.

4.3. Ensemble Model Attacks

To further evaluate the transferability of the proposed method, we conducted additional experiments in the ensemble-surrogate setting, where multiple surrogate models are jointly used to generate adversarial examples. Compared with single-surrogate attacks, ensemble attacks typically offer greater transferability because the optimization process can exploit gradients shared across multiple model architectures.

Table 2 reports the attack success rates (ASRs) of different methods under the ensemble-surrogate setting. Overall, the use of ensemble surrogates already yields a clear increase in attack performance, but FCR still provides consistent further improvements across almost all attack methods and surrogate combinations. For relatively weaker baselines such as the SEM and DIM, the improvement remains substantial, especially when the number of surrogate models increases. For example, under the four-model ensemble setting, the average ASR of the SEM rises from 80.4% to 92.0%, while that of the DIM improves from 83.8% to 95.5%. For stronger baselines such as the SIM, SSA, and SIA, the original attacks already achieve high transferability on standard models, yet FCR still provides non-negligible gains and further pushes the average ASR to above 96% in several cases.

Another clear pattern is that the benefit of FCR is particularly evident on the adversarially trained target models. Although the ASR on standard target models is already close to saturation in some ensemble settings, there is still considerable room for improvement on Inc-v3_{adv} and IncRes-v2_{ens}. In these cases, FCR consistently yields much larger gains than on naturally trained models. This again suggests that frequency-consistent regularization helps suppress attack patterns overly dependent on specific spectral artifacts and instead encourages perturbations that remain effective against stronger defenses.

A further observation is that the SSA and SIA remain the strongest attack backbones in the ensemble setting, and their FCR-enhanced versions achieve the best overall performance. Between them, FCR-SSA and FCR-SIA both show very strong transferability, while FCR-SIA is slightly better in several ensemble settings. Taken together, these results indicate that FCR is not limited to single-surrogate optimization but can be naturally combined with ensemble-based attacks to further improve the transferability of adversarial examples, especially against robust target models.

4.4. Performance Against Defense Models

We also evaluated the robustness of the generated adversarial examples against several input-preprocessing defenses. In this setting, three surrogate models, Den-121, Inc-v3, and

IncRes-v2, we used to generate adversarial examples, covering different model architectures. We focused on two strong transferable attacks, the SSA and SIA, which showed competitive performance in the previous experiments. The generated adversarial examples were then evaluated against the adversarially trained target model, IncRes-v2_{ens}, using six representative defense models.

Table 2. Attack success rates (%) under different ensemble-surrogate models. Boldface indicates the better result within each attack pair, and * denotes the white-box setting.

Surrogate	Attack	Target Model						Avg
		Inc-v3	Res-50	Den-121	IncRes-v2	Inc-v3 _{adv}	IncRes-v2 _{ens}	
IncRes-v2	SEM	69.5	54.6	58.1	99.0 *	33.6	23.0	56.3
	FCR-SEM	78.5	69.4	77.7	99.3 *	55.9	53.6	72.4
	DIM	78.0	63.5	68.1	97.3 *	42.0	29.3	63.0
	FCR-DIM	88.2	80.7	86.6	99.1 *	67.1	67.0	81.5
	SIM	85.4	72.2	78.2	99.6 *	58.8	47.1	73.6
	FCR-SIM	88.3	79.9	90.4	99.4 *	78.9	73.1	85.0
	SSA	88.0	74.0	79.6	97.5 *	62.3	48.9	75.1
	FCR-SSA	92.8	88.5	92.7	98.7 *	86.4	84.4	90.6
	SIA	92.9	81.4	84.8	98.8 *	52.9	36.8	74.6
	FCR-SIA	97.9	95.5	97.0	99.9 *	87.8	85.6	93.9
IncRes-v2 Den-121	SEM	81.4	84.9	100.0	98.3	45.7	36.5	74.5
	FCR-SEM	89.4	91.8	100.0	99.4	71.6	69.8	87.0
	DIM	86.8	89.0	100.0	97.6	55.6	46.9	79.3
	FCR-DIM	94.8	95.9	100.0	99.4	84.4	81.2	92.6
	SIM	92.1	92.4	100.0	98.7	69.7	60.2	85.5
	FCR-SIM	93.8	95.2	100.0	99.5	86.7	84.2	93.2
	SSA	92.6	93.4	99.7	96.3	78.5	66.0	87.8
	FCR-SSA	95.5	95.9	99.9	98.5	92.7	90.3	95.5
	SIA	96.7	96.2	99.8	99.5	70.6	55.2	86.3
	FCR-SIA	98.8	99.0	100.0	99.9	94.1	91.7	97.3
IncRes-v2 Inc-v3 Res-50 Den-121	SEM	99.7	100.0	100.0	98.3	48.2	36.2	80.4
	FCR-SEM	100.0	100.0	100.0	99.5	78.2	74.3	92.0
	DIM	99.4	100.0	99.4	97.8	59.3	47.0	83.8
	FCR-DIM	99.9	100.0	99.8	99.7	89.0	84.8	95.5
	SIM	99.9	100.0	100.0	98.7	71.6	60.5	88.5
	FCR-SIM	99.8	100.0	99.9	99.2	91.1	88.0	96.3
	SSA	98.6	99.8	99.0	97.6	83.9	70.9	91.6
	FCR-SSA	99.2	99.9	99.7	98.9	96.1	92.4	97.7
	SIA	99.3	99.8	99.3	98.8	71.1	52.1	86.7
	FCR-SIA	99.8	100.0	99.9	99.9	96.2	93.8	98.3

As shown in Table 3, incorporating FCR consistently improves the effectiveness of adversarial examples against all considered defense models. Across different surrogate models, the FCR-enhanced variants achieve substantially higher attack success rates than their original counterparts. For example, under the Den-121 surrogate, the average success rate of the SSA increases from 64.8% to 85.4%, while that of the SIA improves from 58.5% to 84.8%. Similar gains are also observed under the Inc-v3 and IncRes-v2 surrogates, indicating that the proposed regularization generalizes well across different attack backbones and surrogate architectures.

The improvements are especially clear under relatively stronger defenses such as RS and NRP, where the original attacks usually suffer noticeable degradation. For instance, when IncRes-v2 is used as the surrogate, the success rate of the SSA against NRP rises from 36.0% to 81.6% after incorporating FCR. In addition, the SSA generally performs slightly better than the SIA under defenses related to compression and frequency filtering, such as JPEG and FD, while both attacks benefit substantially from FCR. These results suggest that frequency-consistent regularization improves not only transferability across models but also the robustness of adversarial perturbations under diverse defensive preprocessing operations.

We further evaluated the robustness of FCR under varying strengths of preprocessing defenses, including different JPEG qualities and bit-depth reduction levels, as shown in Table 4. While stronger compression generally reduces attack success rates, FCR-enhanced variants consistently outperform their baselines across all settings. For example, FCR-SIA maintains an over 82% success rate under all JPEG qualities, whereas SIA remains below 50%. Similar improvements are observed under bit-depth reduction. These results indicate that FCR enhances the spectral robustness of adversarial perturbations, enabling them to remain effective under different levels of compression and quantization.

Table 3. Attack success rates (%) of adversarial examples against different defense models. Boldface indicates the better result within each attack pair.

Surrogate	Attack	Bit-Red	JPEG	FD	R&P	RS	NRP	Avg
Den-121	SSA	64.2	74.1	68.3	66.4	61.1	54.7	64.8
	FCR-SSA	88.4	90.0	89.5	88.6	76.5	79.7	85.4
	SIA	57.9	67.4	60.4	62.0	57.5	46.1	58.5
	FCR-SIA	90.1	91.1	90.9	91.3	69.4	75.8	84.8
Inc-v3	SSA	25.7	53.8	46.9	42.5	62.1	25.0	42.7
	FCR-SSA	71.2	83.4	85.2	79.6	74.0	56.2	74.9
	SIA	19.2	48.3	40.0	38.0	61.1	25.0	38.6
	FCR-SIA	63.8	85.0	84.8	83.0	70.5	52.8	73.3
IncRes-v2	SSA	48.9	72.0	67.6	62.8	62.9	36.0	58.4
	FCR-SSA	84.0	89.2	91.2	88.8	77.5	81.6	85.4
	SIA	34.0	64.7	58.8	57.8	60.3	37.4	52.2
	FCR-SIA	84.7	93.0	93.8	92.4	73.5	68.2	84.3

Table 4. Attack success rates (%) under different defense parameter settings. Boldface indicates the better result within each attack pair.

Attack	JPEG (65)	JPEG (75)	JPEG (95)	Bit-Red (3)	Bit-Red (5)	Bit-Red (7)
SSA	0.553	0.538	0.494	0.254	0.257	0.247
FCR-SSA	0.829	0.834	0.815	0.668	0.712	0.714
SIA	0.471	0.483	0.465	0.207	0.192	0.206
FCR-SIA	0.835	0.850	0.828	0.524	0.638	0.669

4.5. Ablation Study

4.5.1. Effect of the Minimum Low-Frequency Ratio

Figure 5 presents the effect of the minimum sampling ratio $r_{\min}$ on attack performance. For most FCR-based methods, better results are obtained when $r_{\min}$ is kept in a relatively low range, whereas the average ASR drops gradually as $r_{\min}$ increases. This indicates that relatively strong low-frequency filtering is useful for improving transferability during optimization. A small $r_{\min}$ allows the sampled views to cover a broader range of frequency

conditions, from heavily filtered inputs to nearly intact images. This makes the perturbation less tied to a specific spectral range and helps it remain effective as the degree of frequency preservation varies. In turn, the perturbation shows better transfer across target models. In contrast, as $r_{\min}$ increases, the sampled views become less diverse and more concentrated around weakly filtered inputs, thereby weakening the regularization effect of FCR.

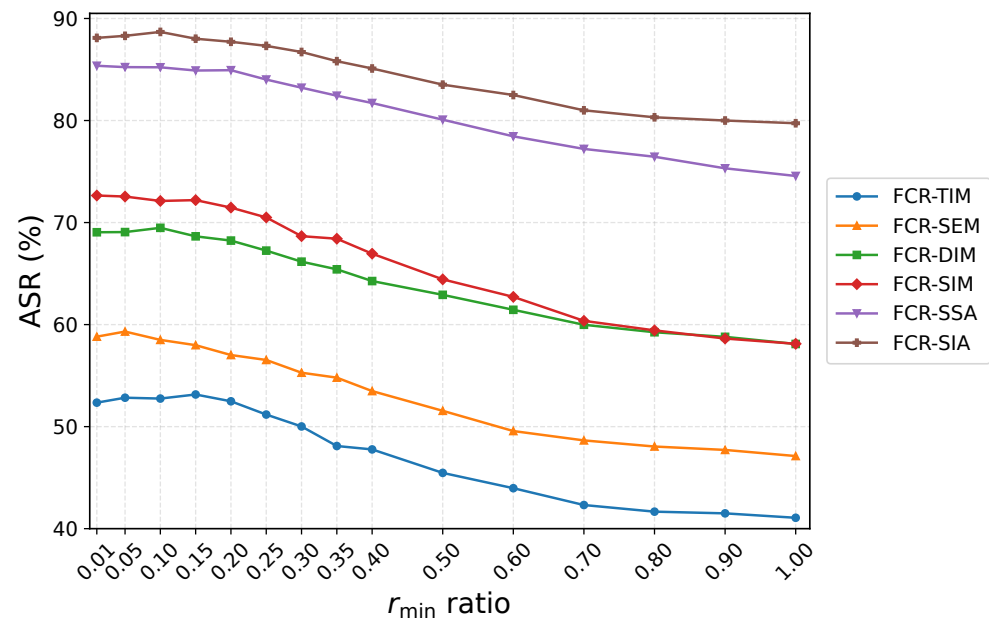


Figure 5. Effect of the minimum low-frequency ratio $r_{\min}$ in FCR under the single-model transfer setting, using Inc-v3 as the surrogate model. The reported values are average attack success rates (ASRs, %).

It is also worth noting that when $r_{\min}$ is extremely small, the ASR of some methods exhibits a slight increase before entering a clear downward trend. Including a small number of more aggressively filtered views appears to improve frequency-level regularization. Once the lower bound is pushed too low, however, some sampled views lose too much semantic information, which reduces the benefit of the added frequency diversity. Therefore, the best performance is generally achieved in a low-to-moderate range, rather than at the extreme minimum. This observation is consistent with the trade-off between semantic consistency and perturbation diversity in the frequency domain. Lower values of $r_{\min}$ enforce stronger low-frequency constraints but may limit optimization flexibility, while larger values introduce more high-frequency components that are less robust across models. Notably, the optimal range of $r_{\min}$ remains stable across different surrogate architectures, suggesting that the proposed frequency constraint aligns with general spectral properties of natural images rather than model-specific characteristics. In our method, instead of fixing a single value, we randomly sample the lower bound from $[0.05, 0.15]$ at each iteration.

4.5.2. Effect of the Number of Sampled Views

Figure 6a shows the effect of the number of sampled views N on attack performance. Overall, the average ASR increases consistently with more frequent views, indicating that multi-view optimization is important for improving transferability. When N is very small, the optimization sees only limited frequency perturbations, and the resulting perturbations are less stable across models. As N increases, the sampled views include a wider variety of frequency-preserving variations, thereby improving the perturbation's robustness to different spectral distortions. The gain is especially clear when increasing N from one to five or 10, at which point all methods show noticeable improvement. After that, the

increase in performance becomes smaller, suggesting that the benefit of adding more views gradually saturates. This means that a moderate number of sampled views is already sufficient to provide effective frequency regularization.

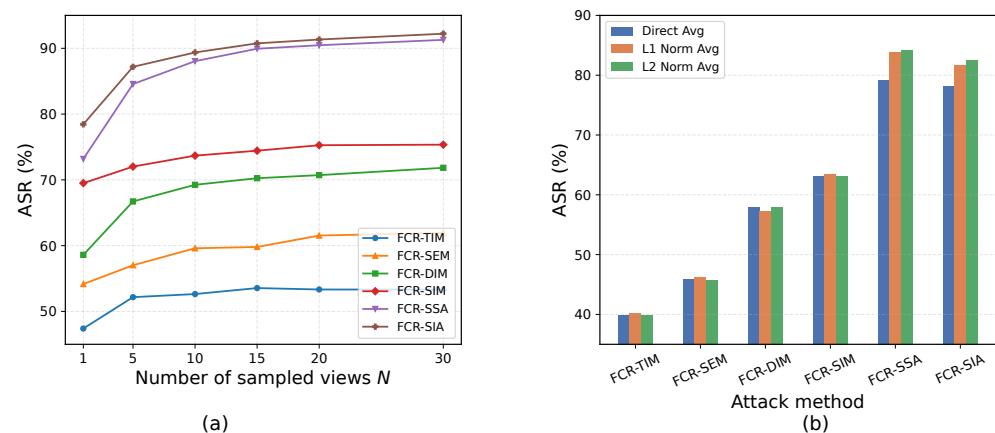


Figure 6. Ablation results of FCR under the single-model transfer setting using Inc-v3 as the surrogate model. (a) Effect of the number of sampled low-frequency views N . (b) Effect of different gradient aggregation schemes, including direct averaging, ℓ_1 -normalized averaging, and ℓ_2 -normalized averaging. The reported values are average attack success rates (ASRs, %).

4.5.3. Effect of Gradient Aggregation Scheme

Figure 6b compares three ways of combining gradients from different frequency views: direct averaging, ℓ_1 -normalized averaging, and ℓ_2 -normalized averaging. Overall, normalized gradient aggregation outperforms direct averaging in most cases, especially for stronger attacks such as FCR-SSA and FCR-SIA. Averaging raw gradients directly can give too much weight to several views with large magnitudes. Normalization mitigates this issue and brings the contributions from different views closer together. Compared with direct averaging, both ℓ_1 and ℓ_2 normalization usually improve the final ASR, but ℓ_1 normalization gives more stable results across different baselines and is slightly better in most cases.

4.5.4. Effect of Perturbation Budget and Iteration Steps

To evaluate whether the effectiveness of FCR depended on specific attack settings, we further conducted experiments with different perturbation budgets and iteration counts under the ℓ_∞ constraint. As shown in Figure 7a, increasing the perturbation budget consistently improves the attack success rate for all methods. More importantly, the FCR-enhanced variants consistently outperform their corresponding baselines across all tested ϵ values. This indicates that the benefit of frequency-consistent regularization is not limited to a particular perturbation magnitude. Figure 7b presents the results under different iteration steps. While all methods benefit from more optimization steps, the proposed FCR method maintains a clear advantage over baseline attacks across all settings. Notably, the performance gain remains stable even as the number of iterations increases, suggesting that FCR does not rely on a specific optimization length. These results demonstrate that FCR consistently improves transferability across attack configurations, suggesting that its effectiveness is robust to varying perturbation budgets and optimization settings.

4.5.5. Efficiency–Performance Trade-Off

We analyzed the efficiency–performance trade-off of FCR by examining its computational cost and corresponding gains in attack success rate (ASR). As shown in Figure 8, FCR consistently improves the ASR across all baseline methods, at the expense of increased

computation due to multi-view sampling. In particular, the main overhead comes from repeated forward and backward passes across multiple sampled views, leading to an approximately linear increase in runtime with respect to the number of views. In practice, the computational cost of FCR is roughly proportional to the number of sampled views N , i.e., about N times that of the baseline per iteration. Meanwhile, additional memory consumption arises from maintaining gradients of multiple views during aggregation, but this overhead remains moderate and manageable in practice. Increasing the number of sampled views from $N = 5$ to $N = 10$ yields further ASR improvements but incurs noticeably higher runtime, with diminishing returns compared to the gain from the baseline to $N = 5$. The improvements are consistent across different attacks, indicating that FCR effectively complements existing transfer-enhancing strategies. Overall, FCR achieves a favorable efficiency–performance trade-off, with $N = 5$ providing a practical balance between effectiveness and computational overhead.

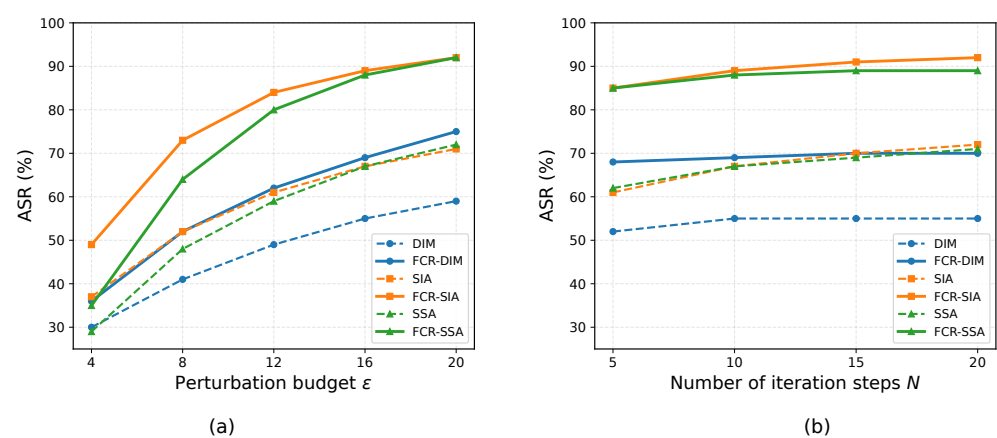


Figure 7. Ablation results of FCR under the single-model transfer setting using Inc-v3 as the surrogate model. (a) Effect of the perturbation budget ϵ . (b) Effect of the number of iteration steps N . The reported values are average attack success rates (ASRs, %).

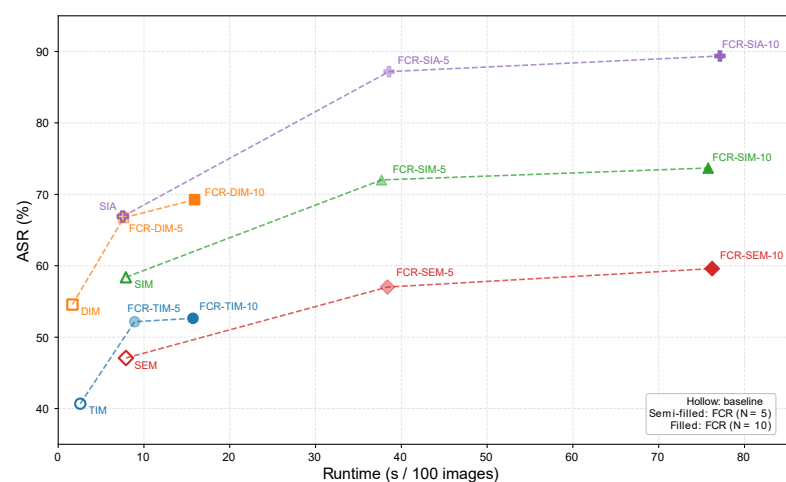


Figure 8. Efficiency-performance trade-off of baseline attacks and their FCR-enhanced variants under the single-surrogate setting (Inc-v3). FCR improves ASR consistently at the cost of increased runtime, with $N = 5$ providing a favorable balance between effectiveness and efficiency.

4.5.6. Perceptual Similarity Analysis

To evaluate whether FCR affected the visual quality of adversarial examples, we further measured perceptual similarity using SSIM [45] and LPIPS [46], where higher SSIM scores and lower LPIPS scores indicate a smaller perceptual distance between images.

The results are summarized in Table 5. As shown, FCR consistently improves perceptual quality across all attacks. Specifically, SSIM increases by 0.016–0.063, indicating better structural consistency with the original images. Meanwhile, LPIPS decreases by 0.006–0.026, suggesting that the adversarial perturbations become less perceptible to human observers. These results demonstrate that FCR does not introduce noticeable visual artifacts. Instead, enforcing frequency consistency across multiple views suppresses high-frequency noise and produces more natural-looking adversarial examples. Notably, this improvement is achieved consistently across all baseline methods, highlighting FCR’s robustness in maintaining visual fidelity.

Table 5. Perceptual similarity comparison between baseline attacks and FCR. Boldface indicates the better result between the baseline and FCR. $\uparrow$ and $\downarrow$ denote that higher and lower values are better, respectively.

Attack	SSIM (Base)	SSIM (FCR)	Δ SSIM $\uparrow$	LPIPS (Base)	LPIPS (FCR)	Δ LPIPS $\downarrow$
TIM	0.7647	0.7807	+0.0160	0.2931	0.2874	−0.0058
DIM	0.6652	0.7084	+0.0432	0.3258	0.3012	−0.0247
SIM	0.6678	0.7045	+0.0367	0.3069	0.2851	−0.0218
SEM	0.6716	0.7075	+0.0359	0.3150	0.2989	−0.0161
SIA	0.6595	0.7068	+0.0473	0.3330	0.3087	−0.0243
SSA	0.6541	0.7175	+0.0634	0.3058	0.2798	−0.0260

5. Conclusions and Discussion

In this paper, we revisited transferable adversarial attacks from a frequency perspective and showed that adversarial perturbations exhibited spectral characteristics different from those of natural image semantics. Our frequency analysis indicated that image semantics were mainly preserved in the low-frequency range, while adversarial behavior was more strongly affected by higher-frequency components. As a result, gradient-based optimization may favor surrogate-dependent spectral cues that do not transfer reliably.

Motivated by this observation, we proposed frequency-consistent regularization (FCR). FCR is a simple plug-in strategy that introduces multiple low-frequency-preserving views during each optimization step. Instead of updating the perturbation based on a single input view, it uses gradients computed from several randomly sampled, frequency-filtered samples. In this way, the perturbation is encouraged to remain effective across different frequencies, rather than relying too heavily on a narrow set of spectral cues.

Extensive experiments showed that FCR consistently improved the transfer performance of several representative iterative attacks under single-model, ensemble-model, and defense settings. From the ablation study, we found that the method worked better when the lower bound was set to a relatively small value.

Overall, this work highlights the role of frequency-consistent regularization in improving adversarial transferability and provides a spectral perspective on perturbation behavior. Nevertheless, several limitations of the proposed method should be acknowledged. FCR incurs additional computational overhead from multi-view sampling, increasing runtime and memory consumption. Although this overhead remains moderate and can be controlled by adjusting the number of sampled views, it may still become non-negligible in large-scale or resource-constrained scenarios. In addition, the effectiveness of FCR relies on the assumption that low-frequency components capture more transferable semantic information, while high-frequency components are more model-specific. While this assumption generally holds for natural images, it may not fully hold across all data distributions or model architectures. In particular, when high-frequency details contribute more significantly to model decisions, enforcing frequency consistency may reduce the flexibility of perturbation optimization. Furthermore, this work focuses on image classification, and the applicability of frequency-consistent regularization to other tasks, such as object detection

or semantic segmentation, remains to be further explored. Future work will further explore more adaptive frequency modeling strategies and extend the proposed framework to a broader range of tasks and datasets.

Author Contributions: Conceptualization, T.S. and S.W.; methodology, T.S.; software, T.S.; validation, T.S.; investigation, T.S., S.W. and B.L.; data curation, T.S. and S.W.; writing—original draft preparation, T.S.; writing—review and editing, T.S., S.W. and B.L.; visualization, T.S.; supervision, B.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original data presented in the study are openly available in the NIPS 2017 adversarial competition image dataset at <https://github.com/anlthms/nips-2017/tree/master/mmnta> (accessed on 10 March 2026).

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

Abbreviations

The following abbreviations are used in this manuscript:

FCR	Frequency-Consistent Regularization
DNNs	Deep Neural Networks
AEs	Adversarial Examples
AT	Adversarial Training
DCT	Discrete Cosine Transform
IDCT	Inverse Discrete Cosine Transform
CE	Cross-Entropy
ASR	Attack Success Rate

Appendix A

Algorithm A1 Frequency-consistent regularization algorithm.

Require: Clean image x , surrogate model f , loss function J , ground-truth label y , perturbation budget ϵ , number of iterations T , decay factor μ , number of sampled views N .

Ensure: Adversarial example x_T^{adv} .

```

1:  $\alpha = \epsilon / T$ 
2:  $g_0 = 0, x_0^{adv} = x$ 
3: for  $t = 0$  to  $T - 1$  do
4:   Sample  $r_{\min}^{(t)} \sim \mathcal{U}(0.05, 0.15)$ 
5:   Initialize aggregated gradient  $g = 0$ 
6:   for  $i = 1$  to  $N$  do
7:     Sample  $r_i^{(t)} \sim \mathcal{U}(r_{\min}^{(t)}, 1)$ 
8:     Obtain low-frequency view  $\tilde{x}_i^{(t)} = T_{r_i^{(t)}}(x_t^{adv})$ 
9:     Compute gradient  $\delta_i = \nabla_x J(f(\tilde{x}_i^{(t)}), y)$ 
10:    Accumulate gradients as  $g = g + \delta_i / \|\delta_i\|_1$ 
11:  end for
12:  Get average gradient as  $\bar{g}_t = \frac{1}{N} g$ 
13:  Update momentum term as  $g_{t+1} = \mu \cdot g_t + \bar{g}_t$ 
14:  Update adversarial example as
     $x_{t+1}^{adv} = \Pi_{B_\epsilon(x)}(x_t^{adv} + \alpha \cdot \text{sign}(g_{t+1}))$ 
15: end for
16: return  $x_T^{adv}$ 

```

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [\[CrossRef\]](#)
2. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaria, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [\[CrossRef\]](#)
3. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
4. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
5. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018; pp. 99–112.
6. Yuan, X.; He, P.; Zhu, Q.; Li, X. Adversarial examples: Attacks and defenses for deep learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2019**, *30*, 2805–2824. [\[CrossRef\]](#) [\[PubMed\]](#)
7. Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Xiao, C.; Prakash, A.; Kohno, T.; Song, D. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 1625–1634.
8. Singh, I.; Araki, T.; Kakizaki, K. Powerful physical adversarial examples against practical face recognition systems. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*; IEEE: Piscataway, NJ, USA, 2022; pp. 301–310.
9. Kilany, S.; Mahfouz, A. A comprehensive survey of deep face verification systems adversarial attacks and defense strategies. *Sci. Rep.* **2025**, *15*, 30861. [\[CrossRef\]](#)
10. Wang, X.; Ge, Z.; Liu, B.; Fang, Z.; Zhou, F.; Zhang, R.; Wang, S.; Luo, Y. Devling into Adversarial Transferability on Image Classification: Review, Benchmark, and Evaluation. *arXiv* **2026**, arXiv:2602.23117. [\[CrossRef\]](#)
11. Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. Ensemble adversarial training: Attacks and defenses. *arXiv* **2017**, arXiv:1705.07204.
12. Liu, X.; Yang, Y.; He, K.; Hopcroft, J.E. Parameter interpolation adversarial training for robust image classification. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 1613–1623. [\[CrossRef\]](#)
13. Mangussi, A.D.; Pereira, R.C.; Lorena, A.C.; Santos, M.S.; Abreu, P.H. Studying the robustness of data imputation methodologies against adversarial attacks. *Comput. Secur.* **2025**, *157*, 104574. [\[CrossRef\]](#)
14. Naseer, M.; Khan, S.; Hayat, M.; Khan, F.S.; Porikli, F. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2020; pp. 262–271.
15. Lu, Y.; Love, P.E.; Luo, H.; Fang, W. Mitigating adversarial attacks and building robust deep learning models for assessing risks in tunnel construction. *Reliab. Eng. Syst. Saf.* **2025**, *255*, 111491. [\[CrossRef\]](#)
16. Chen, P.Y.; Zhang, H.; Sharma, Y.; Yi, J.; Hsieh, C.J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*; ACM: Dallas, TX, USA, 2017; pp. 15–26.
17. Chen, J.; Feng, Z.; Zeng, R.; Pu, Y.; Zhou, C.; Jiang, Y.; Gan, Y.; Li, J.; Ji, S. Enhancing adversarial transferability with adversarial weight tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Philadelphia, PA, USA, 2025; Volume 39, pp. 2061–2069.
18. Lin, J.; Song, C.; He, K.; Wang, L.; Hopcroft, J.E. Nesterov accelerated gradient and scale invariance for adversarial attacks. *arXiv* **2019**, arXiv:1908.06281.
19. Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; Yuille, A.L. Improving transferability of adversarial examples with input diversity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2019; pp. 2730–2739.
20. Zhang, C.; Zhou, L.; Xu, X.; Wu, J.; Liu, Z. Adversarial attacks of vision tasks in the past 10 years: A survey. *ACM Comput. Surv.* **2025**, *58*, 1–42. [\[CrossRef\]](#)
21. 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*; IEEE: Piscataway, NJ, USA, 2019; pp. 4312–4321.
22. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2017**, arXiv:1706.06083.
23. Zhao, Y.; Zheng, J.; Gao, X.; Liu, L.; Zhang, Y.; Zhang, Q. Enhancing the transferability of adversarial attacks via Scale Enriching. *Neural Netw.* **2025**, *189*, 107549. [\[CrossRef\]](#)
24. Wang, X.; Zhang, Z.; Zhang, J. Structure invariant transformation for better adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2023; pp. 4607–4619.
25. Wang, X.; He, X.; Wang, J.; He, K. Admix: Enhancing the transferability of adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2021; pp. 16158–16167.

26. Long, Y.; Zhang, Q.; Zeng, B.; Gao, L.; Liu, X.; Zhang, J.; Song, J. Frequency domain model augmentation for adversarial attack. In *Proceedings of the European Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2022; pp. 549–566.
27. Liu, Z.; Liu, Q.; Liu, T.; Xu, N.; Lin, X.; Wang, Y.; Wen, W. Feature distillation: Dnn-oriented jpeg compression against adversarial examples. In *Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: Piscataway, NJ, USA, 2019; pp. 860–868.
28. Xu, W.; Evans, D.; Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv* **2017**, arXiv:1704.01155. [[CrossRef](#)]
29. Cohen, J.; Rosenfeld, E.; Kolter, Z. Certified adversarial robustness via randomized smoothing. In *Proceedings of the International Conference on Machine Learning*; PMLR: Long Beach, CA, USA, 2019; pp. 1310–1320.
30. Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; Li, J. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 9185–9193.
31. Kurakin, A.; Goodfellow, I.; Bengio, S.; Dong, Y.; Liao, F.; Liang, M.; Pang, T.; Zhu, J.; Hu, X.; Xie, C.; et al. Adversarial attacks and defences competition. In *The NIPS'17 Competition: Building Intelligent Systems*; Springer: Berlin/Heidelberg, Germany, 2018; pp. 195–231.
32. Jia, X.; Zhang, Y.; Wei, X.; Wu, B.; Ma, K.; Wang, J.; Cao, X. Improving fast adversarial training with prior-guided knowledge. *IEEE Trans. Pattern Anal. Mach. Intell.* **2024**, *46*, 6367–6383. [[CrossRef](#)]
33. Bin Kamilin, M.H.; Yamaguchi, S. Stackade Ensemble Learning for Resilient Forecasting Against Missing Values, Adversarial Attacks, and Concept Drift. *Appl. Sci.* **2025**, *15*, 8859. [[CrossRef](#)]
34. Guo, C.; Rana, M.; Cisse, M.; Van Der Maaten, L. Countering adversarial images using input transformations. *arXiv* **2017**, arXiv:1711.00117.
35. Xie, C.; Wang, J.; Zhang, Z.; Ren, Z.; Yuille, A. Mitigating adversarial effects through randomization. *arXiv* **2017**, arXiv:1711.01991.
36. Prakash, A.; Moran, N.; Garber, S.; DiLillo, A.; Storer, J. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 8571–8580.
37. 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*; IEEE: Piscataway, NJ, USA, 2018; pp. 1778–1787.
38. Mustafa, A.; Khan, S.H.; Hayat, M.; Shen, J.; Shao, L. Image super-resolution as a defense against adversarial attacks. *IEEE Trans. Image Process.* **2019**, *29*, 1711–1724. [[CrossRef](#)] [[PubMed](#)]
39. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [[CrossRef](#)]
40. Wightman, R. PyTorch Image Models. 2019. Available online: <https://github.com/rwightman/pytorch-image-models> (accessed on 10 March 2026).
41. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2016; pp. 2818–2826.
42. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2016; pp. 770–778.
43. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2017; pp. 4700–4708.
44. Szegedy, C.; Ioffe, S.; Vanhoucke, V.; Alemi, A. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*; AAAI Press: Philadelphia, PA, USA, 2017; Volume 31.
45. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)] [[PubMed](#)]
46. Zhang, R.; Isola, P.; Efros, A.A.; Shechtman, E.; Wang, O. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018; pp. 586–595.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.