electronics

MDPI

Article

A High-Transferability Adversarial Sample Generation Method Incorporating Frequency Domain Transformations

Sijian Yan, Zhengjie Deng *, Jiale Dong and Xiyan Li

School of Information Science and Technology, Hainan Normal University, Haikou 571158, China; 202312083900001@hainu.edu.cn (S.Y.); 202212083900002@hainu.edu.cn (J.D.); 920220@hainnu.edu.cn (X.L.)
* Correspondence: zjdeng@hainnu.edu.cn

Abstract: Adversarial attack methods have achieved satisfactory results in white-box attack scenarios, but their performance declines when transferred to other deep neural network (DNN) models. Currently, there are many methods to improve the transferability of adversarial samples, and enhancing transferability through input transformations is an effective approach. However, most existing input transformations are performed in the spatial domain, neglecting transformations in the frequency domain. Therefore, this paper proposes a novel input transformation-based attack: the frequency domain enhancement (FDE) method, which performs input transformations in the frequency domain to increase input diversity. Specifically, this method processes input images in the frequency domain, suppresses high-frequency information in the input images, and then randomly amplifies certain frequency domain information, generating adversarial samples with stronger transferability. Experimental results show that adversarial samples generated through FDE demonstrate significant improvement in transferability on both undefended and defended models on the ImageNet dataset. Notably, this method can be combined with many existing techniques to further enhance the transferability of adversarial samples.

Keywords: adversarial examples; transferability; perturbation; frequency domain

check for updates

Citation: Yan, S.; Deng, Z.; Dong, J.; Li, X. A High-Transferability Adversarial Sample Generation Method Incorporating Frequency Domain Transformations. *Electronics* **2024**, *13*, 4480. https://doi.org/10.3390/electronics13224480

Academic Editor: Stefanos Kollias

Received: 14 October 2024
Revised: 10 November 2024
Accepted: 12 November 2024
Published: 15 November 2024

Copyright: © 2024 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) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In recent years, DNN [1–4] have achieved tremendous success in the field of computer vision, including applications such as autonomous driving, facial recognition, and object detection. However, the stability of DNN remains a concern for the general public. Researchers such as Szegedy and Goodfellow [5,6] have highlighted the existence of adversarial attacks, where carefully designed but imperceptible perturbations are added to natural images, enabling adversarial samples to successfully deceive DNN models. Thus, designing effective attack methods to verify the robustness of DNN models before deploying them in critical domains, such as autonomous driving, is of paramount importance.

Adversarial attacks can be categorized into white-box and black-box attacks. In white-box attacks, the victim model is fully transparent to the attacker, who has access to all information, including the victim model architecture and parameters, resulting in a high success rate for the adversarial samples. However, because it is challenging to gain access to all information about the victim model, white-box attacks are not feasible in real-world environments. Due to the inherent transferability of adversarial samples across various DNN models, many researchers have focused on black-box attack scenarios. In black-box attacks, adversarial samples are crafted using a reference white-box model and then transferred to the black-box model for attacking. When the differences between the white-box and black-box models are substantial, or the adversarial samples are overly fitted to the original reference white-box model, their transferability decreases.

To address this issue, researchers have proposed several techniques to enhance the transferability of adversarial samples, including gradient calculation-based methods [7,8],

input transformations [7,9–13], and feature-level attacks [14,15]. Among these, input transformation has emerged as one of the most effective methods, which is conceptually similar to data augmentation techniques used in model training. This is also the main focus of our research. The objective of applying transformations to input images is to reduce the overfitting of the generated adversarial samples to the original reference white-box model, thereby improving transferability.

Existing input transformations are mostly conducted in the spatial domain, but transformations in the frequency domain, as shown in traditional image processing research, can often achieve similar effects to those in the spatial domain. This paper proposes a frequency domain enhancement (FDE) method, which processes images in the frequency domain to generate adversarial samples with stronger transferability. Firstly, FDE performs image transformations in the frequency domain by suppressing high-frequency information while preserving low-frequency information, and then partially enhancing specific components in the frequency domain. Experimental results demonstrate that adversarial samples generated by FDE achieve higher transferability when attacking multiple recognition models. Overall, the main contributions of this paper are as follows:

- This study finds that the patterns in the frequency domain of images are relatively consistent, allowing for more convenient modifications to specific regions of an image by altering its frequency domain. Such modifications are difficult to achieve in the spatial domain, and they can help generate adversarial samples with higher transferability.
- This paper proposes a novel method of frequency domain transformation and finds that suppress high-frequency information in the input image, while enhancing the frequency domain information of specific regions, is beneficial for improving the transferability of generated adversarial samples.
- This paper conducts extensive experiments to demonstrate the superiority of the frequency domain enhancement (FDE) method, which exhibits excellent transferability across both standard models and defense models. Furthermore, combining FDE with existing methods can enhance the transferability of the generated adversarial samples.

2. Related Work

Szegedy et al. [5] highlighted the vulnerability of DNN to adversarial samples, which can cause the model to misclassify with a high probability. They employed the LBFGS method to generate adversarial samples. Since then, adversarial samples have been extensively studied.

Depending on the level of access to model information, current attack methods can be classified into white-box and black-box attacks. White-box attacks, such as the fast gradient sign method (FGSM) [6], suggest that the existence of adversarial samples is due to the linear characteristics of DNN and use gradient ascent to maximize the loss. Iterative FGSM (I-FGSM) [16] extends FGSM by iteratively applying small perturbations in the direction of gradient increase. DeepFool [17] generates minimal norm adversarial perturbations through an iterative calculation method, pushing images beyond the classification boundary until misclassification occurs. In practice, attackers usually do not have access to the internal information of the model, making black-box attacks more relevant. Black-box attacks can be broadly divided into query-based [18,19] and transfer-based approaches. Query-based attacks focus on estimating the target model's gradient by interacting with it. Transfer-based attacks include gradient calculation-based methods, input transformation-based methods and feature-level attacks. Momentum Iterative FGSM (MI-FGSM) [8] proposes an iterative generation method based on momentum to avoid getting trapped in local optima. Diverse input FGSM (DI-FGSM) [9] randomly adjusts the size of the input image and applies padding. Translation-invariant FGSM (TI-FGSM) [20] smooths gradients using a Gaussian kernel. Feature importance-aware (FIA) [14] introduces aggregated gradients to capture feature importance, averaging gradients over the feature maps of the source model. Spectrum simulation attack (SSA) [21] converts images to the frequency domain and applies random masks, generating diverse spectral saliency maps that reflect the

diversity of substitute models. Other transfer-based methods include scale-invariant (SI) [7], which leverages the scale-invariance property of CNN by computing the average gradient over multiple scaled images, and the Nesterov iterative (NI) [7], which incorporates the Nesterov accelerated gradient method into iterative gradient-based attacks to mitigate local optima during the optimization process. Object diversity-based input (ODI) [10] renders images on different 3D objects, drawing adversarial images on these objects. Admix [11] randomly samples images from other classes and mixes them with the original image for gradient calculation. Linear backpropagation (LinBP) [22] removes certain ReLU layers from the source model to reduce their impact during forward and backward computations, making the model more linear. Backward propagation attack (BPA) [23] identifies that non-linear layers (ReLU, max-pooling) truncate gradients during backpropagation, reducing adversarial sample transferability. BPA uses a non-monotonic function as the derivative of ReLU and combines softmax with a temperature parameter to smooth the derivative of max-pooling. Affordable and generalizable substitute (AGS) [24] believes that current adversarial attacks are all based on pre-trained models used as source models; however, these models are not specifically developed for adversarial attacks. Therefore, they propose a training architecture tailored for adversarial attacks. By employing various techniques, these methods enhance the transferability of adversarial samples across different DNN models, providing insights into their robustness and adaptability in both standard and defense models.

3. Methodology

This paper focuses on the transferable attacks of adversarial samples. It begins with an introduction to basic attack methods [6,8,16,21], followed by a detailed description of the frequency domain enhancement (FDE) method and its underlying motivation.

3.1. Preliminary

First, we define $f(\cdot)$ as a neural network classifier, θ as the model parameters, and x as the input image with the true label y. Let J denote the loss function. The goal of an adversarial attack is to generate an adversarial perturbation δ such that the adversarial example $x^{\mathrm{adv}} = x + \delta$ leads to misclassification by the model, i.e., $f(x+\delta) \neq \mathrm{y}$. To ensure that the perturbation δ is imperceptible to the human eye, an $\ell_\infty-$norm constraint is applied, represented as $\| \delta \|_\infty \leq \epsilon$, where ϵ denotes the maximum perturbation magnitude. Therefore, the process of generating an adversarial perturbation can be viewed as an optimization problem, as shown in Equation (1):

$$x^{adv} = \underset{\|\delta\|_\infty \leq \epsilon}{argmax} \mathrm{J}\left(x^{adv}, \mathrm{y}; \theta\right) \tag{1}$$

where $J\left(x^{adv}, y; \theta\right)$ represents the cross-entropy loss function.

In black-box attacks, the model parameters θ are not accessible. Therefore, the common approach is to solve the optimization problem on a substitute model to generate adversarial samples. Here, we introduce some other attack methods. The adversarial samples generated by the fast gradient sign method (FGSM) [6] are formulated as follows:

$$x^{adv} = x + \epsilon \cdot sign(\nabla_x J(x,\ y; \theta)) \tag{2}$$

where $\nabla_x J(x, y; \theta)$ represents the gradient of the loss function with respect to x, sign$(\cdot)$ is the sign function.

The iterative fast gradient sign method (I-FGSM) [16] is an improved version of FGSM. Instead of adding a single-step perturbation in the direction of the gradient as in FGSM, I-FGSM introduces multiple small perturbations iteratively in the direction of the

gradient. The magnitude of each perturbation is controlled by α, and the gradient direction is recalculated at each step. The formula is as follows:

$$x_0^{adv} = x \tag{3}$$

$$x_{t+1}^{adv} = \mathrm{Clip}_x^{\varepsilon}\left\{x_t^{adv} + \alpha \cdot sign\left(\nabla_x J\left(x_t^{adv}, y; \theta\right)\right)\right\} \tag{4}$$

where x_t^{adv} represents the adversarial example generated at the t-th iteration, $\mathrm{Clip}_x^{\varepsilon}(\cdot)$ denotes an element-wise clipping operation to ensure $x_t^{adv} \in [x - \varepsilon, x + \varepsilon]$, and α is the step size.

The momentum iterative fast gradient sign method (MI-FGSM) [8] integrates the technique of I-FGSM by accumulating a velocity vector in the direction of the gradient of the loss function during the iteration process to accelerate the gradient descent algorithm. The formula is as follows:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_{x_t^{adv}} J\left(x_t^{adv}, y; \theta\right)}{\| \nabla_{x_t^{adv}} J\left(x_t^{adv}, y; \theta\right) \|_1} \tag{5}$$

$$x_{t+1}^{adv} = \mathrm{Clip}_x^{\varepsilon}\left\{x_t^{adv} + \alpha \cdot sign(g_{t+1})\right\} \tag{6}$$

where g_t is the accumulated gradient at the t-th iteration, and $g_t = 0$. μ represents the decay factor, which is typically set to 1.

3.2. Frequency Domain Enhancement

Previous work [1,2,25,26] has shown that different models often rely on different frequency components of each input image when making decisions. From a frequency domain perspective, Long at al. [21] explores the correlation between models by enhancing the image in the frequency domain after applying a discrete cosine transform (DCT). After a DCT transformation, the low-frequency components of an image are concentrated in the upper-left corner of the spectrum. Compared to the spatial domain, information in the frequency domain is more stable. The formula for transforming an image signal from the spatial domain to the frequency domain using DCT is as follows:

$$DCT(x)_{[u,v]} = \frac{2}{E} C(u) C(v) \sum_{i=0}^{E-1} \sum_{j=0}^{E-1} x[i,j] cos\left[\frac{(2i+1)u\pi}{2E}\right] cos\left[\frac{(2j+1)v\pi}{2E}\right] \tag{7}$$

$x[i, j]$ represents the value of the image at the coordinate $[i, j]$, and E denotes the size of the image, while $C(u)$ and $C(v)$ are compensation coefficients designed to ensure the orthogonality of the DCT matrix.

JPEG [27] compression technology achieves a high compression ratio with minimal degradation of image quality by retaining the low-frequency coefficients that are important to human vision while setting most of the high-frequency coefficients to zero. Inspired by this, we propose a new spectral transformation method, frequency domain enhancement (FDE), which improves upon the SSA [21] through a modified random spectral transformation. The formula for SSA is as follows:

$$T(x) = IDCT(DCT(x + \xi) \odot M) \tag{8}$$

where DCT stands for discrete cosine transform, and IDCT stands for inverse discrete cosine transform, and $\odot$ represents the Hadamard product. The values in $\xi \sim \mathcal{N}(0, \sigma^2 I)$ and each element of $M \sim \mathcal{U}(1 - \rho, 1 + \rho)$ represent random elements sampled from Gaussian and uniform distributions, respectively. In the paper, σ = 16 and ρ = 0.5. It is important to note that both DCT and IDCT are lossless transformations.

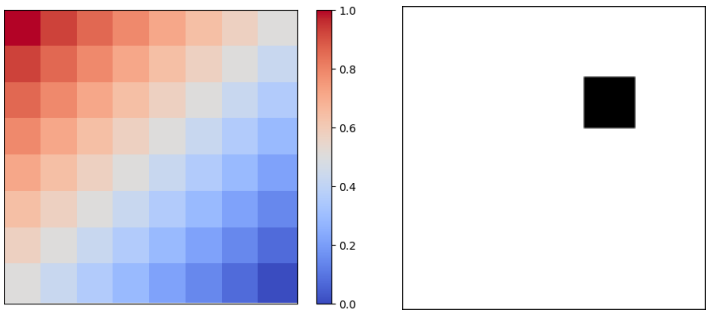
The proposed frequency domain enhancement (FDE) method filters certain frequency components of the image using a weight matrix. One of the designed weight matrices, M_1

has the same size as the image, with values linearly decreasing from 1 at the top-left corner to 0 at the bottom-right corner. Specifically, the top-left 1/64 region has a value of 1, and the bottom-right 1/64 region has a value of 0, as shown in Figure 1a, where different colors represent the distribution of M_1 values. By applying the Hadamard product between M_1 and the spectrum, the resulting spectrum attenuates certain high-frequency components. To increase the variety of processed images, the method also designs another weight matrix, M_2, composed solely of 0 and 5. This matrix is used for random frequency enhancement operations. Figure 1b shows a selected weight matrix, where the black areas have a value of 5 and the white areas have a value of 1. The enhancement operation is performed by applying the Hadamard product between M_2 and the spectrum elements. Finally, the inverse discrete cosine transform (IDCT) is applied to convert the image from the frequency domain back to the spatial domain, yielding the enhanced image. The formula is as follows:

$$P(x) = DCT(x + \xi) \odot M \tag{9}$$

$$F(x) = IDCT(P(x) \odot M_1 \odot M_2(S, K)) \tag{10}$$

S represents the area of frequency domain enhancement, with the enhanced region being randomly selected, and K denotes the enhancement coefficient in the frequency domain; $\odot$ represents the Hadamard product.



(**a**) (**b**)

Figure 1. (**a**) M_1 illustrates values that display with the variation in color. (**b**) M_2 shows values where the black areas have a value of 5 and the white areas have a value of 1.

As shown in Figure 2, the transformation in the frequency domain differs from input transformations in the spatial domain. In the frequency domain, the transformation not only retains the essential semantic information but also involves color changes and effectively distinguishes important features from less significant ones. For example, in Figure 2, the main features—such as the panda, beetle, frog, and bird—are clearly separated from the background after the FDE transformation. This is because there is significant spatial correlation between pixels in the image, and DCT significantly reduces these correlations, concentrating the image's energy in the upper-left region, representing low-frequency information, while high-frequency information is concentrated in the lower-right region. High-frequency information corresponds to the edges of the image, which often overlap with key regions of the image. Our transformation suppresses high-frequency information, guiding the adversarial samples toward improved transferability.

Figure 2. *Cont.*

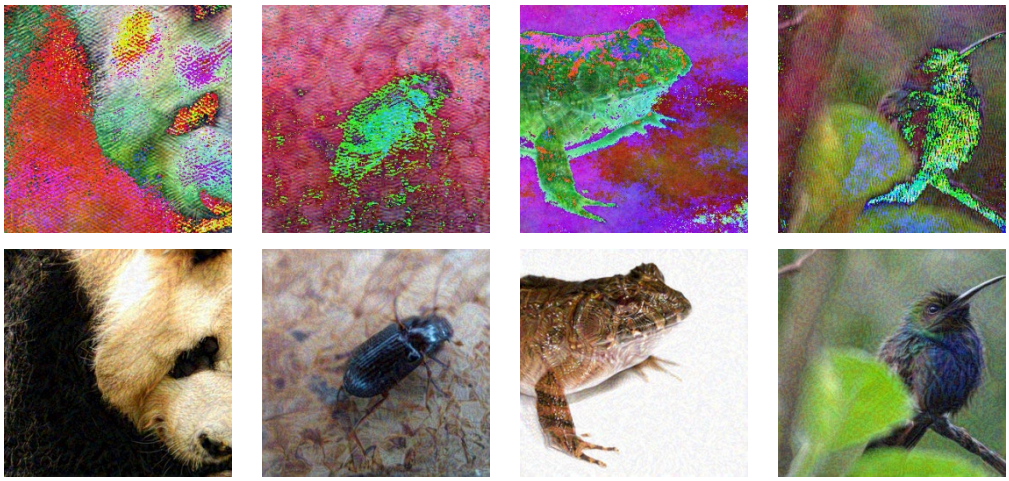


Figure 2. Through the visualization of the spectral transformation, the first row displays the clean images, the second row shows the visualized images after the frequency domain transformation, and the third row presents the generated adversarial samples.

3.3. Attack Algorithms

In the previous section, we introduced the spectral transformation method proposed in this paper. This transformation can be combined with any gradient-based attack. The attack algorithm designed in conjunction with FGSM in this paper is as follows (Algorithm 1):

Algorithm 1. FDE-FGSM

Input: A classifier $f(\cdot)$ with parameters θ, loss function J, clean image x with true label y, the maximum perturbation magnitude ϵ, Clip$(x, 0, 1)$ ensures that the generated pixel values remain within the range [0, 1] during the adversarial sample generation process, number of spectral transformations N, number of random enhancements Q, std σ of noise ξ, and number of iterations T.

Output: The adversarial example x^{adv}

$\alpha = \epsilon / T$, $x_0^{adv} = x$, $g_0 = 0$
for $t = 0 \rightarrow T - 1$ do
  for $i = 1 \rightarrow N$ do
    Get transformation output $x_t^{adv} = P\left(x_t^{adv}\right)$ using Equation (9)
    for $k = 1 \rightarrow Q$ do
      Get transformation output $F\left(x_t^{adv}\right)$ using Equation (10)
      Gradient calculate $g_{i,k} = \nabla_{x_t^{adv}} J\left(F\left(x_t^{adv}\right), y; \theta\right)$
    end for
  end for
  Average gradient: $\overline{g} = \frac{1}{N*Q} \sum_{i=1}^{N} \sum_{k=1}^{Q} g_{i,k}$
  $x_{t+1}^{adv} = \text{Clip}_x^{\varepsilon}\left\{x_t^{adv} + \alpha \cdot sign(\overline{g})\right\}$
  $x_{t+1}^{adv} = \text{Clip}\left(x_{t+1}^{adv}, 0, 1\right)$
end for
$x^{adv} = x_T^{adv}$
return x^{adv}

4. Experiments

4.1. Experiment Setup

The experiments in this paper are based on a dataset compatible with ImageNet, which contains 1000 images with a resolution of 299 × 299 × 3. This dataset was previously used in the NIPS 2017 adversarial competition. For model selection, six commonly used normally trained models were chosen, including Inception-v3 (Inc-v3) [1], Inception-v4 (Inc-v4) [2], Inception-Resnet-v2 (IncRes-v2) [2], Resnet-v2-50 (Res-50), Resnet-v2-101 (Res-101), and

Resnet-v2-152 (Res-152) [3]. Additionally, three defense models were employed: Inc_v3_{ens3}, Inc_v3_{ens4}, $IncRes\_v2_{ens}$ [28].

Comparison: To demonstrate the effectiveness of the spectral transformation attack proposed in this paper, we compared it with various state-of-the-art attack methods, including MI-FGSM [8], DI-FGSM [9], TI-FGSM [8], and S^2I-FGSM [21]. Additionally, we also compared the combined versions of these methods, such as SIM [7], DI-MI-FGSM, VMI-FGSM [29], and S^2I-MI-FGSM.

Parameter settings: In all experiments, the maximum perturbation is set to $\epsilon = 16$, with the number of iterations $T = 10$ and step size $\alpha = \epsilon/T = 1.6$. For MI-FGSM, we set the decay factor $\mu = 1.0$. For DI-FGSM, the transformation probability $p = 0.5$. For TI-FGSM, the kernel size k = 7. For SIM, the number of copies $m = 5$, and for VMI-FGSM, the neighborhood upper limit $\beta = 1.5 * \epsilon$. For S^2-FGSM, we set the number of spectral transformations $N = 20$. For FDE, we set the number of enhancements $Q = 2$, enhancement area $S = 70 \times 70$, and enhancement coefficient $K = 5$. For discarding high-frequency information, the generated weight matrix has the top-left 1/64 area set to 1, the bottom-right 1/64 area set to 0, and the values linearly decrease from 1 to 0 from the top-left to the bottom-right.

5. Attack Models

The effectiveness of the proposed method was evaluated by comparing it with benchmark methods (MI-FGSM, DI-FGSM, S^2I-FGSM). Adversarial examples generated by the proposed method were first created on four standard models and then evaluated for transferability on five standard models and three defense models. Table 1 shows the transferability of the proposed method compared to other methods. The proposed method outperforms other benchmarks in terms of transferability. For example, when Inc_v3 is used as the white-box model, the transfer success rates of MI-FGSM, DI-FGSM, and S^2I-FGSM on Inc_res_v2 are 46.50%, 47.60%, and 58.80%, respectively, while the success rate of the proposed method is 73.00%. This significant improvement demonstrates the effectiveness of the proposed method in enhancing transferability. Additionally, when combined with other attack methods, as shown in Table 2, the proposed method also exhibits superior performance. For instance, when combined with the momentum-based methods, the transfer success rates on Res_152 when using Inc_v3 as the white-box model are 69.00% for SIM, 68.30% for DI-MI, 61.00% for VMI, and 80.90% for S^2I-MI, whereas the proposed method achieves a success rate of 89.30%. This further highlights the effectiveness of our method in improving transferability. When combined with TIM and DIM, our method demonstrated even stronger transferability. When using Inc_v3 as the white-box model, the transferability on non-defense models improved by 2.7–5.1% compared to S^2I-DI-TI-MI, and on defense models, the transferability increased by 6.5–9.2%.

Table 1. The table below shows the success rates of ensemble attacks based on IFGSM, with adversarial examples generated from Inc-v3, Inc-v4, IncRes-v2, and Res_152. The best results are highlighted in bold.

Model	Attack	Inc_v3	Inc_v4	Inc_res_v2	Res_50	Res_101	Res_152	Inc_v3_{ens3}	Inc_v3_{ens4}	$IncRes\_v2_{ens}$
Inc_v3	MI-FGSM	**100.00**	51.50	46.50	48.30	42.90	41.40	22.80	21.30	11.00
	DI-FGSM	**100.00**	56.70	47.60	46.70	42.30	40.90	18.50	19.80	9.00
	S^2I-FGSM	99.70	63.70	58.80	57.50	52.60	48.60	31.20	33.00	17.10
	FDE-FGSM(our)	99.70	**75.60**	**73.10**	**69.70**	**62.80**	**61.50**	**42.60**	**43.30**	**24.00**
Inc_v4	MI-FGSM	60.90	**99.90**	45.50	46.30	42.70	43.10	19.80	18.40	10.70
	DI-FGSM	63.20	99.80	46.20	41.90	38.40	38.30	15.50	16.50	8.70
	S^2I-FGSM	70.70	99.70	55.50	55.40	49.90	48.60	30.90	31.80	17.60
	FDE-FGSM(our)	**78.40**	99.30	**64.50**	**63.70**	**57.70**	**57.60**	**38.90**	**38.50**	**24.90**
Inc_res_v2	MI-FGS	61.00	52.70	99.20	50.90	44.60	44.30	22.00	22.10	13.10
	MDI-FGSM	64.40	60.60	**99.60**	48.10	46.30	45.00	17.80	18.10	11.80
	S^2I-FGSM	76.40	68.00	98.30	60.50	58.30	56.20	37.60	33.90	28.40
	FDE-FGSM(our)	**85.60**	**78.10**	98.20	**73.70**	**69.50**	**66.80**	**51.50**	**46.90**	**39.90**

Table 1. *Cont.*

Model	Attack	Inc_v3	Inc_v4	Inc_res_v2	Res_50	Res_101	Res_152	Inc_v3_{ens3}	Inc_v3_{ens4}	$IncRes\_v2_{ens}$
Res_152	MI-FGSM	55.80	51.20	46.50	84.70	85.50	99.40	26.60	26.10	15.20
	DI-FGSM	63.10	60.30	57.70	89.80	91.90	**99.80**	26.30	24.10	15.20
	S^{2}I-FGSM	66.50	62.30	56.80	92.80	93.10	**99.80**	37.90	35.10	25.30
	FDE-FGSM(our)	**73.20**	**67.40**	**65.80**	**95.00**	**95.40**	99.10	**43.50**	**41.40**	**29.30**

Table 2. The table below presents the success rates of ensemble attacks based on MI-FGSM, with adversarial examples generated from Inc-v3, Inc-v4, IncRes-v2, and Res_152. The best results are highlighted in bold.

Model	Attack	Inc_v3	Inc_v4	Inc_res_v2	Res_50	Res_101	Res_152	Inc_v3_{ens3}	Inc_v3_{ens4}	$IncRes\_v2_{ens}$
Inc_v3	SIM	**100.00**	76.30	74.90	72.60	68.60	69.00	39.90	38.00	23.80
	DI-MI	**100.00**	78.80	73.50	71.20	67.60	68.30	40.80	38.40	21.70
	VMI	**100.00**	73.90	68.50	64.90	59.90	61.00	38.80	38.70	23.30
	S^{2}I-MI	99.60	87.90	86.10	83.70	81.40	80.90	55.10	56.50	35.20
	S^{2}I-DI-TI-MI	99.10	92.00	91.20	87.80	86.80	87.40	81.70	80.40	69.60
	FDE-MI(our)	99.90	93.60	93.30	90.50	89.30	89.30	69.20	70.20	45.00
	FDE-DI-TI-MI(our)	99.70	**95.20**	**93.80**	**91.50**	**90.70**	**90.50**	**89.20**	**87.90**	**78.80**
Inc_v4	SIM	87.60	**100.00**	77.60	76.40	73.70	73.30	47.60	42.60	28.80
	DI-MI	83.10	99.90	75.30	68.90	64.80	65.50	35.90	33.70	19.70
	VMI	77.40	99.90	69.00	63.90	61.70	62.20	39.00	38.60	24.20
	S^{2}I-MI	91.20	99.40	86.30	83.50	82.60	81.80	57.40	56.40	36.50
	S^{2}I-DI-TI-MI	92.70	98.10	89.20	85.80	85.20	86.40	79.20	78.30	69.80
	FDE-MI(our)	94.20	99.30	90.20	88.10	86.60	85.80	68.20	65.20	46.00
	FDE-DI-TI-MI(our)	**95.60**	99.10	**92.60**	**91.30**	**88.60**	**88.50**	**85.90**	**84.40**	**77.10**
Inc_res_v2	SIM	86.20	83.70	**99.90**	79.30	77.60	76.30	55.70	48.70	39.70
	DI-MI	81.90	79.70	99.50	73.20	72.10	69.80	43.10	39.30	30.60
	VMI	78.70	74.50	98.80	66.80	65.60	63.00	45.80	41.70	34.30
	S^{2}I-MI	90.40	88.90	98.00	86.30	84.30	84.10	68.90	63.40	55.70
	S^{2}I-DI-TI-MI	90.40	89.10	97.30	85.70	84.50	84.40	80.00	76.50	76.30
	FDE-MI(our)	94.10	92.20	98.80	89.80	89.80	88.80	78.10	73.80	66.40
	FDE-DI-TI-MI(our)	**94.40**	**93.40**	97.90	**92.00**	**91.40**	**91.20**	**91.10**	**88.70**	**86.40**
Res_152	SIM	76.40	73.30	71.70	95.10	95.50	99.80	47.00	43.50	29.90
	DI-MI	85.10	83.90	80.10	95.30	96.00	**99.90**	51.70	48.20	34.60
	VMI	72.90	67.10	65.80	92.30	92.60	99.50	46.10	41.90	30.60
	S^{2}I-MI	87.80	86.90	85.50	97.50	97.40	99.70	62.90	59.70	46.40
	S^{2}I-DI-TI-MI	93.60	93.20	92.20	98.10	97.90	99.80	85.70	84.30	79.70
	FDE-MI(our)	91.40	89.30	90.40	97.70	97.90	99.50	71.30	68.80	55.00
	FDE-DI-TI-MI(our)	**94.70**	**93.30**	**93.80**	**98.20**	**98.20**	99.30	**89.80**	**87.40**	**83.80**

5.1. Ablation Study

In this section, we examine the impact of different parameters (such as the enhancement coefficient, area, and frequency) on the transferability of adversarial samples generated using the proposed method.

First, we analyze the effect of different frequency domain enhancement areas S on the performance of FDE-MI, as shown in Figure 3a. The adversarial samples were generated on Inc_v3 with an enhancement coefficient $K = 5$ and frequency domain enhancement iterations $Q = 2$. It can be observed that the highest transferability is achieved when S is set to 70×70. However, when S exceeds 70×70, the transferability begins to decline. This decline may be due to the larger enhancement area causing significant alterations to the image, potentially changing its semantic content.

In Figure 3b, we investigate the effect of different enhancement coefficients K on transferability. For this experiment, we fixed the enhancement area $S = 70 \times 70$ and the frequency domain enhancement iterations $Q = 2$. The results show that the best transferability is achieved when K is between 5 and 7. However, when K exceeds 7, the success rate begins to decrease, likely because an excessively large enhancement coefficient causes the image to lose its essential semantic information.

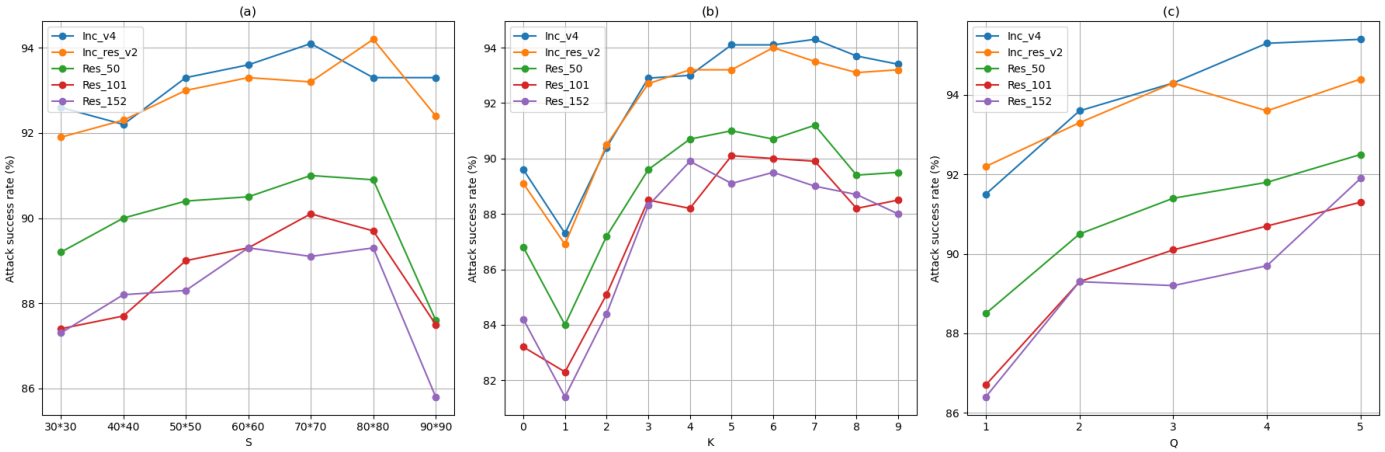


Figure 3. The attack success rates for adversarial samples generated using FDE with different parameters are presented. Figure (**a**) shows the effect of the enhancement area S on transferability, Figure (**b**) shows the effect of different enhancement coefficients K on transferability, and Figure (**c**) shows the effect of the number of enhancements Q on transferability. The adversarial samples were created on Inc_v3 and evaluated for their transferability across five models.

In Figure 3c, we study the impact of the number of enhancements Q on transferability, with the enhancement area S set to 70×70 and the enhancement coefficient K set to 5. The results indicate that performance improves as the number of enhancements Q increases. Notably, even after a single enhancement ($Q = 1$), as shown in Figure 4, FDE achieves success rates of 91.50%, 92.20%, 88.50%, 86.70%, and 86.40% on Inc_v4, Inc_res_v2, Res_50, Res_101, and Res_152, respectively, which are significantly higher than the corresponding success rates of 87.90%, 86.10%, 83.70%, 81.40%, and 80.90% achieved by S^2I-MI. This demonstrates the effectiveness of the proposed method.

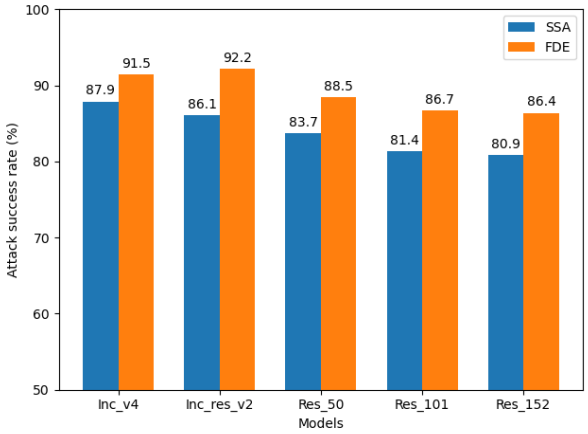


Figure 4. The comparison of attack success rates between FDE and SSA is shown when the enhancement frequency area $S = 70 \times 70$, enhancement coefficient $c = 5$, and number of enhancements $Q = 1$. The adversarial samples were generated on Inc_v3, with blue representing the SSA attack success rates and orange representing the FDE attack success rates.

In Figure 5, we study the impact of two types of frequency domain modifications in FDE on the transferability of generated adversarial samples. The FDE-FGSM method was applied on Inc_v3 with an enhancement coefficient $K = 5$, frequency domain enhancement iterations $Q = 2$, and enhancement area $S = 70 \times 70$. The performance was evaluated across five models. The results indicate that both types of frequency domain operations play a crucial role in enhancing the transferability of the generated adversarial samples using the proposed method.

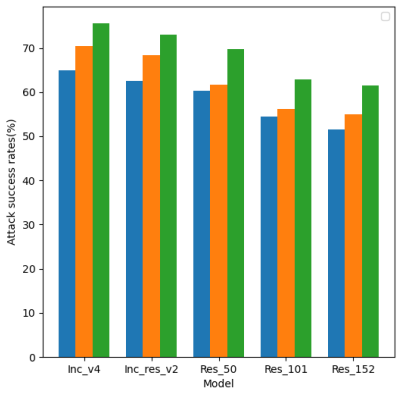


Figure 5. The adversarial samples were generated on Inc_v3 using FDE-FGSM. In the comparison, blue represents the success rates when applying the Hadamard product between the image's frequency domain and weight matrix M_1. Orange represents the success rates when applying the Hadamard product between the image's frequency domain and weight matrix M_2. Green represents the success rates when both frequency domain operations were applied simultaneously.

6. Conclusions

In this paper, we propose a novel frequency domain enhancement method, termed FDE, which enhances images from a frequency domain perspective to improve the transferability of generated adversarial samples. Specifically, for an input image, we suppress the high-frequency information while enhancing the frequency domain information in certain regions. Extensive experiments demonstrate that FDE outperforms baseline methods in terms of transferability and can be further combined with existing approaches to enhance transferability. In future work, we plan to continue exploring the frequency domain of images to develop more robust adversarial attack methods.

Author Contributions: Conceptualization, methodology, software, validation, formal analysis, investigation, and writing—original draft preparation, S.Y.; Formal analysis, project administration, supervision, and writing—review and editing, Z.D.; Validation and resources, J.D.; Validation and writing—review and editing, X.L. All authors have read and agreed to the published version of the manuscript.

Funding: This work was financially supported by Natural Science Foundation of Hainan Province (No. 623RC480, 623QN236), Hainan Province Key R&D Program (No. WSJK2024MS234), the Major Science and Technology Project of Haikou City (No. 2020006), Hainan Province Higher Education Teaching Reform Research Funding Project (No. Hnjg2023-49, Hnjg2021-37, Hnjg2019-50), the Open Funds from Guilin University of Electronic Technology, Guangxi Key Laboratory of Image and Graphic Intelligent Processing (No. GIIP2012), and Haikou Science and Technology Plan Project (No. 2022-007).

Data Availability Statement: Image data can be obtained at https://github.com/cleverhans-lab/cleverhans/tree/master/cleverhans_v3.1.0/examples/nips17_adversarial_competition/dataset (accessed on 10 June 2024).

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

References

1. 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 2016, Las Vegas, NV, USA, 27–30 June 2016.
2. 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 2017, San Francisco, CA, USA, 4–9 February 2017.
3. 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 2016, Las Vegas, NV, USA, 27–30 June 2016.
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet Classification with Deep Convolutional Neural Networks. In Proceedings of the Advances in Neural Information Processing System, Lake Tahoe, NV, USA, 3–6 December 2012; Volume 25.

5. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.J.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, 14–16 April 2014.
6. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the 3rd International Conference on Learning Representations, San Diego, CA, USA, 7–9 May 2015.
7. 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.
8. 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 2018, Salt Lake City, UT, USA, 18–23 June 2018.
9. 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 2019, Long Beach, CA, USA, 15–20 June 2019.
10. Byun, J.; Cho, S.; Kwon, M.J.; Kim, H.S.; Kim, C. Improving the Transferability of Targeted Adversarial Examples through Object-Based Diverse Input. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2022, New Orleans, LA, USA, 18–24 June 2022.
11. 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 2021, Montreal, BC, Canada, 11–17 October 2021.
12. Deng, Z.; Xiao, W.; Li, X.; He, S.; Wang, Y. Enhancing the Transferability of Targeted Attacks with Adversarial Perturbation Transform. *Electronics* **2023**, *12*, 3895. [CrossRef]
13. Wang, X.; Zhang, Z.; Zhang, J. Structure Invariant Transformation for Better Adversarial Transferability. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2023, Paris, France, 2–3 October 2023.
14. Wang, Z.; Guo, H.; Zhang, Z.; Liu, W.; Qin, Z.; Ren, K. Feature Importance-Aware Transferable Adversarial Attacks. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2021, Montreal, BC, Canada, 11–17 October 2021.
15. Ganeshan, A.; BS, V.; Babu, R.V. Fda: Feature Disruptive Attack. In Proceedings of the IEEE/CVF International Conference on Computer Vision 2019, Seoul, Republic of Korea, 27 October–2 November 2019.
16. 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.
17. Moosavi-Dezfooli, S.M.; 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, Las Vegas, NV, USA, 27–30 June 2016.
18. 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 2017, Dallas, TX, USA, 3 November 2017.
19. Andriushchenko, M.; Croce, F.; Flammarion, N.; Hein, M. Square Attack: A Query-Efficient Black-Box Adversarial Attack Via Random Search. In Proceedings of the European Conference on Computer Vision 2020, Glasgow, UK, 23–28 August 2020.
20. 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, Long Beach, CA, USA, 15–20 June 2019.
21. 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 2022, Tel Aviv, Israel, 23–27 October 2022.
22. Guo, Y.; Li, Q.; Chen, H. Backpropagating Linearly Improves Transferability of Adversarial Examples. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 85–95.
23. Wang, X.; Tong, K.; He, K. Rethinking the Backward Propagation for Adversarial Transferability. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 1905–1922.
24. Wang, R.; Guo, Y.; Wang, Y. Ags: Affordable and Generalizable Substitute Training for Transferable Adversarial Attack. In Proceedings of the AAAI Conference on Artificial Intelligence 2024, Vancouver, BC, Canada, 20–27 February 2024.
25. Yin, D.; Gontijo Lopes, R.; Shlens, J.; Cubuk, E.D.; Gilmer, J. A Fourier Perspective on Model Robustness in Computer Vision. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32.
26. Wang, Z.; Yang, Y.; Shrivastava, A.; Rawal, V.; Ding, Z. Towards Frequency-Based Explanation for Robust Cnn. *arXiv* **2020**, arXiv:2005.03141.
27. Wallace, G.K. The Jpeg Still Picture Compression Standard. *Commun. ACM* **1991**, *34*, 30–44. [CrossRef]
28. Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.; Boneh, D.; McDaniel, P. Ensemble Adversarial Training: Attacks and Defenses. *arXiv* **2017**, arXiv:1705.07204.
29. 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, Nashville, TN, USA, 19–25 June 2021.

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.