*big data and cognitive computing*

MDPI

*Article*

# Enhancing Adversarial Transferability via Fourier-Based Input Transformation

**Zilin Tian [1], Xin Wang [2], Yunfei Long [1,*] and Liguo Zhang [1]**

[1] College of Computer Science and Technology, Harbin Engineering University, Harbin 150001, China
[2] Hangzhou Institute of Technology, Xidian University, Hangzhou 311231, China
* Correspondence: 2016064109@hrbeu.edu.cn

**Abstract**

Adversarial transferability makes black-box attacks practical and exposes weaknesses of deep neural networks for computer vision, image recognition, and visual understanding. Among various transferability-enhancing methods, input transformation is one of the most effective strategies. However, existing methods often ignore the decoupling of style and semantics in the input image, as well as the need for customized transformation strategies, resulting in limited performance gains or suboptimal outcomes. In this paper, we propose a novel Fourier-based perspective for input transformation generalization in the context of vision adversarial attacks. The main observations are that the Fourier amplitude captures stylistic information and the phase encompasses richer semantics which are crucial for visual understanding. Motivated by this, we develop a Fourier-based strategy, which performs a stylistic transform and semantic mixup on the input examples to improve transferability. To avoid inconsistent semantics of augmented images for the surrogate model, we mix the original images with the augmentations to maintain semantic consistency and mitigate imprecise gradients. Extensive experiments on ImageNet-compatible datasets demonstrate that our method consistently outperforms existing input transformation attacks.

**Keywords:** adversarial examples; transferability; input transformation; Fourier-based transformation

## 1. Introduction

The vulnerability of deep neural networks (DNNs) to adversarial examples has attracted great attention [1,2]. The adversarial examples are maliciously crafted by adding imperceptible perturbations to natural examples, which can arbitrarily change the prediction of remarkable DNNs [3]. Transferability is key to the effectiveness of adversarial examples in the black-box attack settings, allowing adversarial examples generated by attacking white-box surrogate models to mislead other multiple black-box target models and helping to evaluate the robustness. Hence, studying transfer-based attacks contributes to offering valuable guidance for designing more effective defense strategies have garnered increasing attention.

To improve adversarial transferability, many existing methods utilize gradient optimization [4–6], model ensemble [7,8], or input transformation [9] and have shown promising results. Gradient-optimization attacks iteratively refine perturbations by stabilizing and accelerating convergence within the loss landscape [10,11]. Input-transformation methods enhance transferability by introducing spatial or stylistic variations that mitigate

Check for updates

Academic Editor: Fabrizio Messina

Received: 20 March 2026
Revised: 13 April 2026
Accepted: 24 April 2026
Published: 27 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) license.

overfitting to surrogate models [9,12,13]. Ensemble-based approaches [7,8] integrate multiple surrogates to capture a more comprehensive approximation of the target model's decision boundary, which enhances adversarial transferability. Among them, input transformation methods are widely adopted as an effective approach for mitigating the overfitting of adversarial examples to surrogate models. These methods can be broadly categorized into two groups. The first performs spatial-domain transformation on input, which changes their spatial structure through resizing and padding (DIM [14]), translation (TIM [15]), scaling (SIM [16]), and block-wise shuffle (BSR [13]). The second applies feature-domain transformation to increase input diversity, which modifies the feature distribution through style transfer (ATTA [17], STM [9], SSA [12]), semantic mixing (Admix [18], MFI [19]). In practice, these methods insufficiently decouple style from semantics and fail to design transformation strategies for high-level and low-level features, respectively.

In this paper, we introduce a novel **F**ourier-based **i**nput **t**ransformation (FIT) for improving adversarial transferability. Our motivation comes from a well-known property of the Fourier transformation, i.e., the phase component of the Fourier spectrum preserves high-level semantic content of the original image, and the amplitude component contains low-level stylistic features. This property has been extensively validated in classical signal processing [20–23] and later adopted in computer vision and deep learning contexts [24–26]. This decomposition provides a principled way to perform that: (1) stochastic perturbations on the amplitude spectrum generate diverse style variants without destroying semantic content; (2) mixing phase-only reconstructions from other classes accurately introduces their semantic content. Inspired by the above observations, we propose a Fourier-based framework for unified feature-domain transformation that achieves both stylistic transfer and semantic mixup. Figure 1 illustrates the flowchart of this framework. First, the input images are transformed into the frequency domain using, e.g., the discrete Fourier transform (DFT), and the stochastic masking operation is applied to their amplitude spectrum to generate stylistic variants. Then, we purify the semantic information by using the phase spectrum only and setting the amplitude component to a constant. Finally, combining stylistic variants with phase-only semantic reconstruction to produce augmented examples. In the attack process, we mix up the input with its transformed augmented examples to maintain semantic content and avoid imprecise gradients.

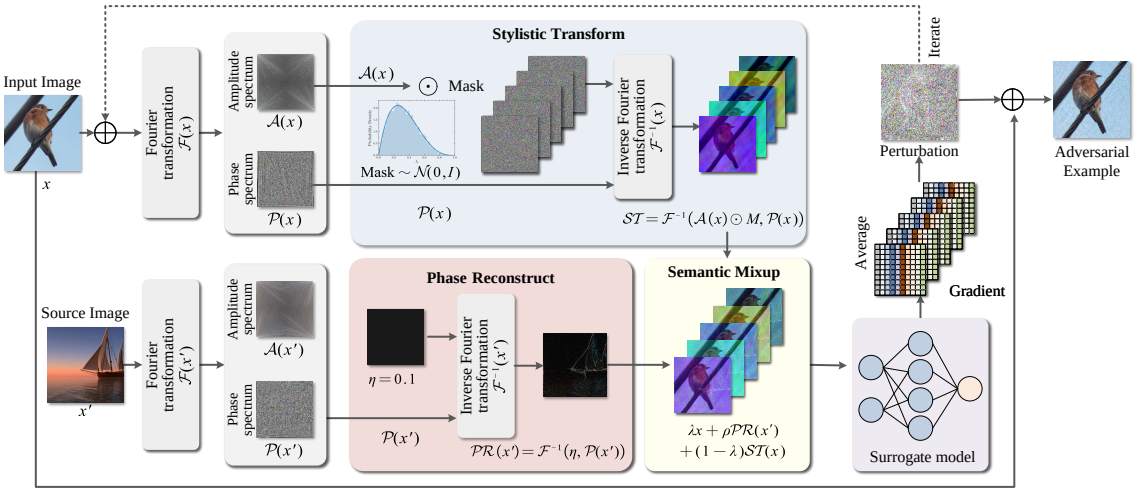



**Figure 1.** Flowchart of our Fourier-based input transformation. First, the input images are transformed into the frequency domain. Then, we generate stylistic variants and phase-only semantic reconstruction and combine them to produce augmented examples. Finally, we mix up the input with its transformed augmented examples in the attack process.

Our main contributions are summarized as follows:

- We propose a novel Fourier-based input transformation (FIT) strategy. FIT manipulates amplitude and phase components to achieve both stylistic transformation and semantic mixup.
- We integrate FIT into the adversarial attack framework and propose a new black-box attack method. The adversarial examples generated by this innovative approach obtain enhanced transferability.
- The extensive experiments conducted on the ImageNet-compatible dataset demonstrate that the FIT attack has significant advantages over the baseline.

## 2. Related Work

### 2.1. Transfer-Based Attack

In transfer-based attacks, perturbations are generated using surrogate models, and their transferability is exploited to target models with unknown internal details. This property makes them suitable for real-world black-box scenarios [27–29]. Typically, such attacks can generally be grouped into the following:

Gradient-optimization attacks refine gradient calculation to avoid adversarial examples getting trapped in suboptimal local maxima. Goodfellow et al. [10] propose the earliest gradient-based attack, the Fast Gradient Sign Method (FGSM), which employs a single-step update to maximize the loss function and generate adversarial examples. Kurakin et al. [11] subsequently introduced an iterative variant, named the Iterative Fast Gradient Sign Method (I-FGSM), as a refinement of the original approach. Dong et al. [4] introduced the Momentum Iterative Fast Gradient Sign Method (MI-FGSM), where a momentum factor is incorporated to preserve direction consistency over iterations, enhancing attack success rates and transferability. Building upon this, the Nesterov Iterative Fast Gradient Sign Method proposed by Lin et al. [16] applies Nesterov-accelerated gradient descent, which further improves the convergence speed and attack performance. Wang et al. [5] developed the Variance Tuning Momentum Iterative Fast Gradient Sign Method, wherein gradient modulation is performed using variance estimates derived from adjacent previous iterations, preventing unstable adjustments. In a similar vein, Wang et al. [6] devised the Enhanced Momentum Iterative Fast Gradient Sign Method, which accumulates gradients from multiple locations along earlier directions, leading to perturbations with greater transferability. Qin et al. [30] proposed the Reverse Adversarial Perturbation method, setting the adversarial attack target as a max–min problem. In each iteration, the loss is minimized within the local neighborhood of the current adversarial sample to find the reverse adversarial perturbation. A local worst-case scenario is determined along this perturbation direction, and the gradient of the local worst-case scenario is used for momentum iteration, guiding the adversarial sample into a flatter, high-loss region. Ge et al. [31] proposed the Penalized Gradient Norm method, explicitly adding a gradient norm penalty term to the adversarial attack target and using the joint gradient of the adversarial loss and gradient norm for momentum iteration. Gan et al. [32] proposed the Gradient Aggregation Attack, which further assumes that the flat maxima are subject to three well-defined constraints, decomposing the adversarial attack target into the adversarial loss, the neighborhood worst-case loss, and the difference between the two losses. ResPA [33] models from a global direction perspective, using residual gradients to construct perturbation directions, making the optimization process more focused on the overall loss trend. Lin et al. [29] proposed semantic-style joint perturbations (SSEPs), which establish a style loss based on the kernel function from the feature space of the surrogate model and inject it into gradient-based attacks to form a semantics-style joint Loss for generating perturbations.

Input-transformation methods aim to reduce overfitting of adversarial examples to the surrogate models by introducing a variety of transformations before perturbation generation. Xie et al. [14] employ random resizing and padding operations to enhance the transferability and propose the Diverse Inputs Method (DIM). Dong et al. [15] propose the Translation-Invariant Method (TIM), which optimizes perturbations over translated inputs and efficiently approximates this process via convolution of gradients with a predefined kernel. Lin et al. [16] introduce the Scale-Invariant Method (SIM), which improves adversarial transferability by optimizing perturbations over multiple scale-transformed copies of the input. Both TIM and SIM incorporate geometric invariances. Wu et al. [17] put forward an Adversarial Transformation-enhanced Transfer Attack (ATTA), in which an auxiliary adversarial transformation network is trained to perform input transformations. Adversarially Mixup (Admix) was proposed by Wang et al. [18] to integrate cross-class information into the input, achieved by mixing it with samples originating from different categories in the spatial domain. The Spectrum Simulation Attack (SSA) proposed by Long et al. [12] enhances adversarial transferability by performing model augmentation in the frequency domain through spectrum transformations. The Style Transfer Method (STM) [9] applies style transfer to generate cross-domain image variants while preserving semantic consistency, thereby enhancing input diversity. The Block Shuffle and Rotation (BSR) [13] randomly shuffles and rotates image blocks to disrupt attention maps, generating diverse transformed inputs for gradient averaging. Qian et al. [19] proposed Mixed-Frequency Inputs (MFI), which biases the semantics of the input towards other classes by adding high-frequency components of other images to the original. Transformed examples produced by existing input transformation adversarial attacks are shown in Figure 2.

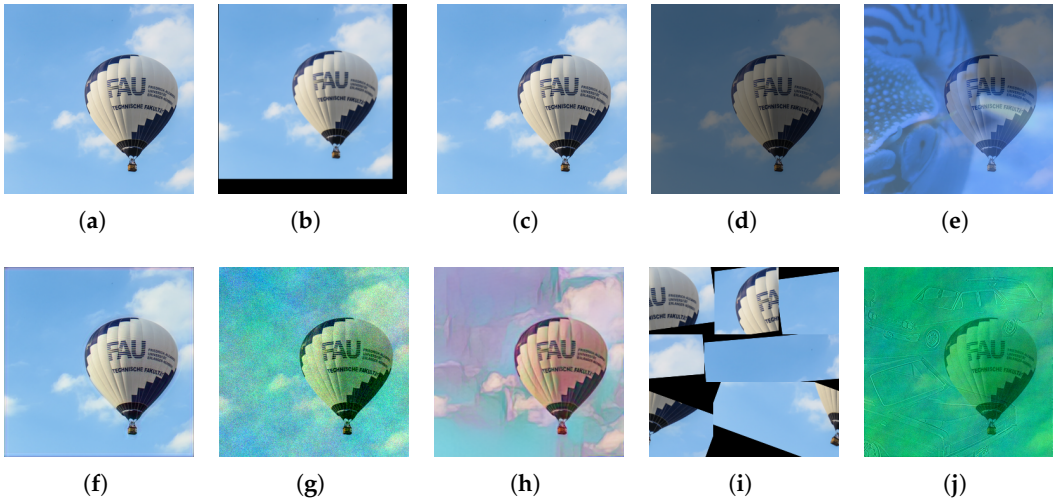



**Figure 2.** A clean example and its transformed images generated by existing input transformation methods and our proposed FIT method. (**a**) Clean; (**b**) DIM [14]; (**c**) TIM [15]; (**d**) SIM [16]; (**e**) Admix [18]; (**f**) ATTA [17]; (**g**) SSA [12]; (**h**) STM [9]; (**i**) BSR [13]; (**j**) Our FIT.

Ensemble-based Attacks enhance the transferability by combining information from multiple surrogate models during the attack process. Dong et al. [4] propose the MI-Ens-FGSM, which enhances the MI-FGSM by combining weighted losses or logits across multiple surrogate models to guide the optimization process. The Stochastic Variance Reduced Ensemble (SVRE) proposed by Xiong et al. [34] reduces variance across ensemble models during iterative optimization, leading to more stable updates and better exploitation of ensemble-based attacks. The adaptive ensemble attack (AdaEA) proposed by Chen et al. [35] adaptively weights surrogate model contributions based on their discrepancy ratios and applies a disparity-reduced filter to better align gradient update directions,

which enhances transferability. The Common Weakness Attack (CWA) [7] improves adversarial transferability by promoting flat loss landscapes and aligning perturbations with shared local optima across ensemble models to exploit their common weaknesses. Furthermore, the Stochastic Mini-batch black-box attack with Ensemble Reweighing (SMER) [8] exploits perturbation diversity across surrogate models and adaptively reweights their contributions via reinforcement learning to maximize attack effectiveness and boost adversarial transferability.

### 2.2. Frequency-Based Analysis and Attacks

In recent years, increasing attention has been devoted to understanding the sensitivity of deep neural networks to frequency-domain characteristics. Wang et al. [36] notice CNN's ability in capturing the high-frequency components of images, which are almost imperceptible to a human. Yin et al. [37] investigated the sensitivity of the model to high-frequency and low-frequency additive noise aligned on different Fourier basis vectors. Building on these observations, some scholars have proposed adversarial attack methods based on the frequency domain. Sharma et al. [38] show that both adversarially trained and undefended models remain vulnerable to low-frequency perturbations. Duan et al. [39] propose the AdvDrop attack, which generates adversarial examples by dropping imperceptible details of images in the frequency domain. More recently, Wang et al. [40] observe unequal sensitivity of DNNs to different frequency components and design attacks that selectively target specific frequency bands to enhance attack effectiveness and transferability. However, existing methods are mainly focused on manipulating specific frequency bands based on sensitivity observations, without leveraging the intrinsic decoupling between style and semantics in the Fourier domain. Consequently, frequency components are treated largely as low-level signals rather than as carriers of semantically meaningful factors, which limits the potential for further improving adversarial transferability.

## 3. Methodology

In this section, we present the proposed adversarial attack method, which incorporates the Fourier-based Input Transformation (FIT) strategy. We first review the preliminaries of adversarial attacks in Section 3.2. Next, we introduce the FIT strategy in Section 3.3, which consists of two steps: stylistic transformation and semantic mixup. Finally, we integrate FIT into the attack framework to generate adversarial examples with enhanced transferability in Section 3.4. For clarity, we summarize the main notations used throughout this paper in Table 1.

**Table 1.** The main notations.

| | |
|---|---|
| $x, y$ | A clean image and its ground truth label. |
| $x_t^{adv}$ | The generated adversarial example at the $t$-th iteration. |
| $f$ | The classify model mapping input variables to label variables. |
| $\mathcal{L}\left(f\left(x_t^{adv}\right), y\right)$ | The cross-entropy loss of the classify model w.r.t $x_t^{adv}$. |
| $\epsilon$ | The adversarial perturbation. |
| $\alpha$ | The step size. |
| $g_{t+1}$ | The calculated gradient at the $t$-th iteration |
| $\mathcal{M}_t$ | The accumulation of gradients at the $t$-th iteration. |
| $\mathcal{F}$ | The discrete Fourier transform (DFT) functions. |
| $\mathcal{F}^{-1}$ | The inverse discrete Fourier transform (IDFT) functions. |
| $\mathcal{A}(x)$ | The amplitude spectrum of $x$. |
| $\mathcal{P}(x)$ | The phase spectrum of $x$. |

**Table 1.** *Cont.*

| | |
|---|---|
| $M$ | A random mask sampled from Gaussian distribution. |
| $\eta$ | A small constant replaces the original amplitude spectrum. |
| $\lambda$ | The mixing coefficient in the Fourier-based input attack. |
| $\mu$ | The decay factor. |
| $\rho$ | The weight controlling the semantic mixup. |

### 3.1. Motivation

Transfer-based adversarial attacks aim to generate perturbations that transfer to unseen black-box models. Although existing input transformation methods improve input diversity to mitigate surrogate overfitting, they insufficiently decouple low-level stylistic information from high-level semantic structures. Without dedicated transformation strategies for each feature level, the resulting perturbations fail to simultaneously influence shallow and deep layers of DNNs, which exhibit different sensitivities to style and semantic information [41,42]. Hence, it is reasonable to argue that separately transforming different feature levels can greatly increase the input diversity and be deceptive to different layers of DNNs.

From a Fourier perspective, an image can be decomposed into an amplitude spectrum encoding stylistic information and a phase spectrum preserving semantic content. This property enables a principled decoupling strategy: stochastic perturbation on the amplitude introduces diverse style variations without damaging semantics, while phase-based reconstructions provide precise high-level semantics. Motivated by these observations, we design a Fourier-based input transformation that jointly performs stylistic diversification and semantic mixup during iterative optimization, enhancing adversarial transferability.

### 3.2. Preliminaries

Given a clean image $x$ with its corresponding ground-truth label $y$, the deep model $f$ is designed to output the predicted probability with high confidence, where $f(x)$ represents the probability output. The highest probability predicted label is denoted as $argmax f(x) = y$. The adversarial attacks aim to find an adversarial example $x^{adv}$ that can mislead the classifier $f$ to make unexpected predictions. In order to make the adversarial sample as close as possible to the original sample, the added perturbation is usually limited to an $L_p$-norm ball (e.g., $L_\infty$-norm) around $x$ with radius $\epsilon$, which can be denoted as $\mathcal{B}_\epsilon(x) = \{\hat{x} : \|\hat{x} - x\|_p \leq \epsilon\}$. The typical MI-FGSM [4] generates adversarial examples via the following iterative optimization updates

$$\begin{aligned} &x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot \text{sign}\left(\nabla_{x_t^{adv}} \mathcal{L}\left(f(x_t^{adv}), y\right)\right) \\ &s.t.\ x_{t=0}^{adv} = x, \left\|x_t^{adv} - x\right\|_p \leq \epsilon \end{aligned} \tag{1}$$

where step size $\alpha = \epsilon/T$, $T$ is the total number of iterations, $p$ denotes the $L_p$-norm (e.g., $L_\infty$-norm), and $\epsilon$ represents the maximum allowable perturbation.

### 3.3. Fourier-Based Input Transformation

Let $\mathcal{F}(\cdot)$ and $\mathcal{F}^{-1}(\cdot)$ denote the discrete Fourier transform (DFT) and inverse DFT (IDFT) functions. Typically, for an image $x \in \mathbb{R}^{C\times H\times W}$, DFT is independently applied to each channel in the pixel space as follows:

$$\mathcal{F}(x)(u,v) = \sum_{h=1}^{H}\sum_{w=1}^{W} x(h,w)e^{-j2\pi\left(u\frac{h}{H}+v\frac{w}{W}\right)} \tag{2}$$

where $(h, w)$ denotes the pixel coordinates of $x$, and $(u, v) \in [H] \times [W]$ signifies the corresponding coordinates in the frequency domain. Then, the amplitude spectrum $\mathcal{A}(x)$ and phase spectrum $\mathcal{P}(x)$ are as follows:

$$\begin{aligned} \mathcal{A}(x) &= \sqrt{\mathrm{Re}^2(\mathcal{F}(x)) + \mathrm{Im}^2(\mathcal{F}(x))}, \\ \mathcal{P}(x) &= \arctan\Big(\mathrm{Im}(\mathcal{F}(x)) \cdot \mathrm{Re}^{-1}(\mathcal{F}(x))\Big), \end{aligned} \tag{3}$$

where $\mathrm{Re}(\mathcal{F}(x))$ and $\mathrm{Im}(\mathcal{F}(x))$ represent the real and imaginary parts of $\mathcal{F}(x)$, respectively.

The amplitude spectrum $\mathcal{A}(x)$ primarily encodes the distribution of energy across different frequency components, which is closely related to low-level visual attributes such as texture, color distribution, and contrast—these are typically associated with what is perceived as "style." In contrast, the phase spectrum $\mathcal{P}(x)$ determines how these frequency components are spatially arranged, which directly governs the structural layout of the image, including object shapes and contours. As a result, phase plays a dominant role in preserving high-level semantic content. With the stylistic and semantic information contained in the Fourier amplitude and phase components, respectively, we propose a Fourier-based framework with two transformation steps.

(1) Stylistic Transformation: To alter the distribution of low-level stylistic features in an image, a natural choice is perturbing its amplitude information via a multiplicative stochastic modulation. In contrast to purely additive noise, multiplicative modulation preserves the global image structure while substantially altering gradient directions, thereby enhancing attack transferability [13]. Moreover, random scaling enables effective exploration of different networks' sensitivities to frequency components. Accordingly, we apply a random mask to the amplitude spectrum

$$\mathcal{A}(x)' = \mathcal{A}(x) \odot M, M \sim \mathcal{N}(0, I), \tag{4}$$

where $\odot$ represents element-wise multiplication, $M$ is a random mask sampled from Gaussian distribution, and $I$ has the same dimensions as $x$. Then, to preserve semantic consistency between the stylized image and the original input, the masked amplitude spectrum is combined with the original phase spectrum to form a new Fourier representation, which is then fed into the inverse Fourier transform to generate the augmented image.

$$\mathcal{ST}(x) = \mathcal{F}^{-1}(\mathcal{A}(x)', \mathcal{P}(x)) \tag{5}$$

where $\mathcal{ST}(x)$ defines the stylized image of the original image via stylistic transformation. As shown in Figure 3, we present the original images and several results produced by the proposed stylistic transformation. Compared with existing style transfer methods that rely on pretrained networks, our method requires only manipulating the amplitude spectrum in the frequency domain.

(2) Semantic Mixup: Incorporating information from other categories into the input is a diverse and effective data augmentation technology [18,43]. Admix creates mixed-category examples via linear interpolation in the spatial domain. In addition to introducing semantic content, this strategy also introduces redundant information. We propose a precise mix-semantic transformation that considers the input image and the phase-only reconstructions of images randomly sampled from other categories. To effectively filter out the stylistic information while preserving the semantic content in the phase, we replace the original amplitude spectrum with a constant $\eta \in \mathbb{R}^{C \times H \times W}$, and then reconstruct a phase-only version of the input via the inverse Fourier transform:

$$\mathcal{PR}(x) = \mathcal{F}^{-1}(\eta, \mathcal{P}(x)). \tag{6}$$

where $\mathcal{PR}(x)$ defines the phase-only reconstruction of the original image. As $\eta$ is small, $\mathcal{PR}(x)$ carries the object contours encoded by phase while eliminating color variations, as presented in Figure 4. For the semantic mixup, FIT takes the stylized image $\mathcal{ST}(x)$ as the primary image and mixes it with a phase-only reconstruction $\mathcal{PR}(x)$, where $x'$ is randomly picked from other categories, as illustrated in Figure 5.




**Figure 3.** The stylistic transformation masks and results.

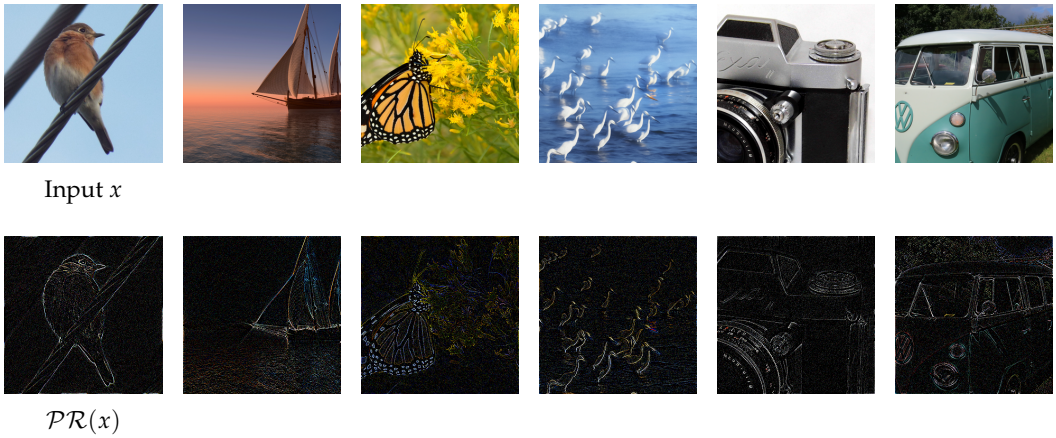



**Figure 4.** The phase-only reconstructions of the clean images.

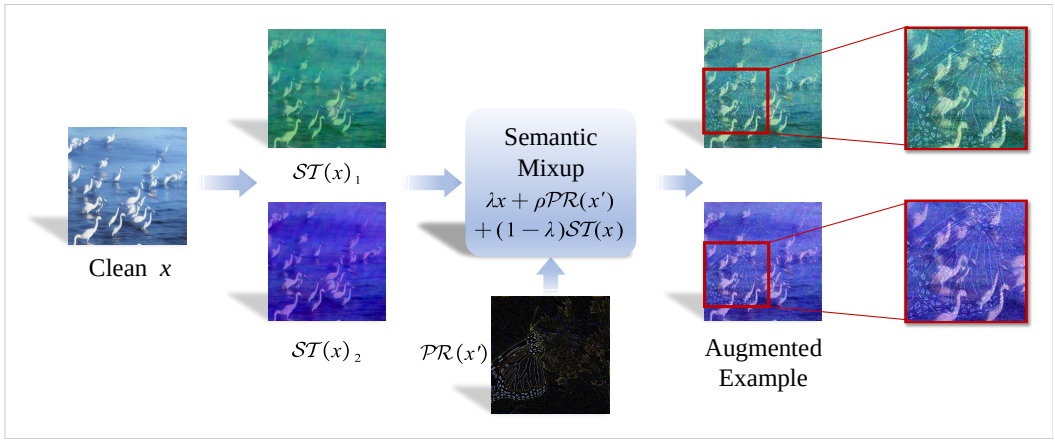



**Figure 5.** The semantic mixup process, which combines the input image with phase-only reconstructions of images randomly sampled from other categories.

The proposed FIT provides a unified framework for feature-domain augmentation. Unlike existing methods that operate either in the spatial domain or perform partial manipulations in the frequency domain, our approach explicitly leverages the amplitude–phase

decomposition to decouple and control stylistic and semantic factors. This enables us to perform both stylistic transformation and semantic mix-up within a single, coherent representation. Such a design not only avoids the entanglement of multiple factors in prior transformations but also offers a more principled way to enhance input diversity by targeting different levels of visual information. Consequently, FIT goes beyond treating style perturbation and semantic mixup as independent strategies, and instead integrates them through a unified Fourier-domain perspective.

### *3.4. Fourier-Based Input Attack*

We integrate the proposed FIT into the iterative gradient-based attack, e.g., MI-FGSM [4], and propose a novel Fourier-based Input attack. Let $x_t^{adv}$ denote the current adversarial example at the $t$-th iteration, where $x_0^{adv} = x$ is the original image and $0 \leq t \leq T$ is the total number of iterations. We mix up $x_t^{adv}$ with its transformed images via FIT to maintain semantic content and avoid imprecise gradients. The augmented example is given by

$$\tilde{x} = \lambda x_t^{adv} + (1-\lambda)\mathcal{ST}(x_t^{adv}) + \rho\mathcal{PR}(x'), \tag{7}$$

where $\lambda \sim \text{Beta}(\tau, \tau)$, which is a mixing coefficient sampled from a Beta distribution for each sample. For instance, $\tau = 1$ is equivalent to sampling uniform distribution $U(0,1)$, and $\rho$ is a weight controlling the semantic mixup. Mixing up the original image with its FIT-transformed images allows the augmented image to introduce features from different domains while preserving the original semantic labels to avoid generating imprecise gradient information during the iterative optimization process.

For a more stable gradient update, we adopt the average gradient of multiple augmented images with random noise to update the perturbation

$$g_t = 1/K \sum_{k=1}^{K} \nabla_{\tilde{x}_k} \mathcal{L}\big(f(\tilde{x}_k + \xi), y\big), \tag{8}$$

where $\mathcal{L}$ represents the loss function of surrogate model $f$, $K$ represents the number of augmented images, $\xi \sim U(-\epsilon \cdot \beta, \epsilon \cdot \beta)$ is uniform additive noise, $\epsilon$ is a pre-defined attack radius, and $\beta$ controls the noise upper bound.

We integrate $g_t$ into MI-FGSM [4] and propose the Fourier-based Input attack (denoted as FIT without ambiguity in the following), which generates adversarial examples of iterative optimization updates as follows:

$$\begin{aligned} \mathcal{M}_{t+1} &= \mu\mathcal{M}_t + \frac{g_t}{\|g_t\|_1}, \\ x_{t+1}^{adv} &= \textstyle\prod_{\mathcal{B}_\epsilon(x)} \Big[ x_t^{adv} + \alpha \cdot \text{sign}(\mathcal{M}_{t+1}) \Big], \end{aligned} \tag{9}$$

where $\mu$ is the decay factor, $\mathcal{M}$ represents the accumulation of gradients, $\alpha$ is the step size and $\prod_{\mathcal{B}_\epsilon(x)}$ constraints $x^{adv}$ in the $\mathcal{B}_\epsilon(x) = \{x' : \|x' - x\|_\infty \leq \epsilon\}$ around $x$. More details of the FIT are summarized in Algorithm 1.

**Algorithm 1:** Fourier-based input transformation (FIT) attack method

**Input:** Surrogate model $f$; natural example $x$ with label $y$.

**Parameters:** The perturbation magnitude $\epsilon$; iteration number $T$; step size $\alpha$; the amplitude constant $\eta$; the number of randomly sampled examples $N$; the noise upper bound $\beta$; the decay factor $\mu$; the distribution parameter $\tau$; the mixing coefficient $\lambda$; the number of transformed images $K$; the semantic mixup weight $\rho$;

**Output:** An adversarial example $x^{adv}$;

**Initialize:** $\alpha = \epsilon/T$; $\mathcal{M}_0 = 0$; $x_0^{adv} = x$;

**for** $t = 1$ *to* $T$ **do**

  Set $g_t = 0$;

  Apply the DFT to the example and get $\mathcal{A}(x_t^{adv})$ and $\mathcal{P}(x_t^{adv})$ by Equation (3);

  **for** $k = 1$ *to* $K$ **do**

    Obtain a random stylistic transformed image $\mathcal{ST}(x_t^{adv})$ by Equation (5);

    Randomly sample an example $x'_k$, then apply semantic extraction and get $\mathcal{PR}(x'_k)$ by Equation (6);

    Sample $\lambda \sim \text{Beta}(\tau, \tau)$ and noise $\xi \sim U(-\epsilon\beta, \epsilon\beta)$,

    Obtain mixed input by Equation (7):
    $\tilde{x} = \lambda x_t^{adv} + (1-\lambda)\mathcal{ST}(x_t^{adv}) + \rho\mathcal{PR}(x')$;

    Calculate the gradient $g_k = \nabla_{\tilde{x}_k}\mathcal{L}\big(f(\tilde{x}_k + \xi), y\big)$;

    Accumulate: $g_t = g_t + 1/K g_k$;

  **end**

  Update momentum

  Get $\mathcal{M}_{t+1} = \mu\mathcal{M}_t + \frac{g_t}{\|g_t\|_1}$;

  Update perturbation

  $x_{t+1}^{adv} = \prod_{\mathcal{B}_\epsilon(x)}\Big[x_t^{adv} + \alpha \cdot \text{sign}(\mathcal{M}_{t+1})\Big]$;

**end**

**return:** $x^{adv} = x_T^{adv}$.

## 4. Experiments

### *4.1. Experimental Setup*

Datasets: We conduct experiments on the classical and extensively used transfer-based adversarial attack studies dataset [5,6,30], which is named the ImageNet-compatible benchmark dataset [44]. This dataset contains 1000 labeled images, each with dimensions of $299 \times 299 \times 3$.

Target models: Our evaluation is performed under black-box attack settings, where both standard and adversarially defended models are selected as targets, which are pre-trained from the Timm library [45]. The normally trained models are DenseNet-121 (Dense-121) [46], ResNet-152 (Res-152) [47], Swin-Base (Swin-B) [48], Inception-ResNet-v2 (IncRes-v2) [49], and DeiT-Base (DeiT-B) [50]. Furthermore, the adversarially defended models include Inc-Res-v2$*ens$, Inc-v3$*ens$4, Inc-v3$*ens$3, and Inc-v3$*adv$ [51]. All aforementioned networks are obtained from the Timm library [45]. In addition, we consider three defense methods: DiffPure [52], RDC [53], and NRP [54], which employ ResNet-101 as a preprocessing-based model [30,31].

Baselines: We compare our method with nine representative input transformation-based attacks: TIM [15], DIM [14], SIM [16], ATTA [17], Admix [18], SSA [12], BSR [13], STM [9], and SSEPs [29]. These attacks are all integrated into the MI-FGSM [4]. The experi-

ments are repeated five times using different random seeds, and the mean performance is reported.

Hyperparameters: For the proposed FIT method, the amplitude parameter is set to $\eta = 0.2$, and the mixing coefficient is $\lambda \sim \text{Beta}(1, 1)$. The perturbation budget is set to $\epsilon = 16/255$, the step size is $\alpha = \epsilon/T$, the iteration number is $T = 10$, and the decay factor for MI-FGSM is 1.0, which follows the common experimental settings in prior work [4,31]. For the baseline methods, we employ the best hyperparameter settings as provided in their papers. In addition, when compared with STM, we set the noise bound $\beta = 2$ and the number of transformed samples $K = 20$ for fairness. All experiments are implemented in PyTorch 2.1.2 and conducted on a system equipped with an Intel(R) Core(TM) i9-14900K CPU and an NVIDIA GeForce RTX 4090 GPU (24 GB memory).

Metrics: The attack performance is measured using the attack success rate (ASR), where "Avg." denotes the average success rate across all evaluated models.

### *4.2. Attacks on Normally Trained Models*

We evaluate the effectiveness of the proposed method on normally trained models. We first adopt four surrogate models, including Inception-v3 (Inc-v3), Inception-v4 (Inc-v4) [55], ResNet-101 (Res-101) [47], and ViT-Base (ViT-B) [56], respectively, to generate adversarial examples using various attack methods. The generated adversarial examples are then used to attack five normally trained models, with ASRs summarized in Table 2. The first column lists the baselines used for comparison. Overall, our method consistently achieves higher ASRs than the baselines across all target models. For instance, when using Inc-v3 as the surrogate model, the ASRs of our FIT on Res-152, Dense-121, IncRes-v2, Swin-B, and DeiT-B are 66.7%, 85.5%, 90.0%, 38.2%, and 45.7%, respectively. These results achieve improvements of 24.4%, 19.2%, 23.8%, 21.2%, and 19.8% over DIM. These findings verify the effectiveness of our FIT on normally trained models and provide further insights into their vulnerabilities to adversarial attack. We also provide visual comparisons of adversarial examples generated by different methods; as illustrated in Figure 6, the perturbations remain visually imperceptible and comparable to existing approaches.

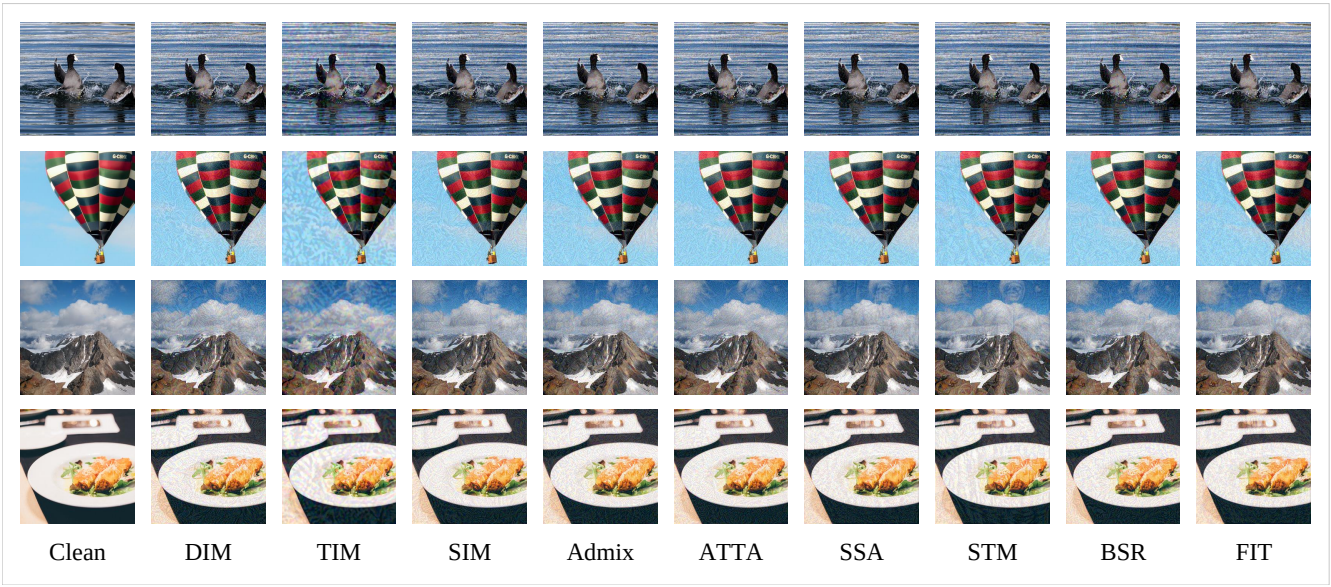



**Figure 6.** The visualization of adversarial examples generated by different methods. The surrogate model is Inc-v3.

**Table 2.** The ASRs (%) of various input transformation attacks on normally trained models. The bolded numbers indicate the best results.

| Surrogate Model | Target Models | Attack Success Rates (%) | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | DIM | TIM | SIM | Admix | ATTA | SSA | STM | BSR | SSEPs | FIT |
| Inc-v3 | Res-152 | 42.3 | 18.9 | 55.0 | 52.1 | 32.6 | 56.8 | 63.0 | 65.6 | 56.8 | **66.7** |
| | Dense-121 | 66.3 | 39.9 | 67.1 | 76.7 | 55.6 | 76.9 | 85.0 | **89.6** | 79.7 | 85.5 |
| | IncRes-v2 | 66.2 | 33.0 | 69.0 | 81.9 | 52.2 | 84.3 | 90.4 | 87.2 | 84.5 | **90.0** |
| | Swin-B | 17.0 | 9.0 | 18.3 | 22.1 | 14.4 | 29.2 | 36.1 | 28.9 | 25.4 | **38.2** |
| | DeiT-B | 25.9 | 17.6 | 26.9 | 30.6 | 18.3 | 35.0 | 46.5 | 36.1 | 33.3 | **45.7** |
| | Avg. | 43.5 | 23.7 | 44.6 | 52.7 | 34.6 | 56.4 | 64.2 | 61.5 | 55.9 | **65.2** |
| Inc-v4 | Res-152 | 46.8 | 22.7 | 54.3 | 64.7 | 38.7 | 65.4 | 70.0 | 62.7 | 65.8 | **70.1** |
| | Dense-121 | 67.5 | 41.5 | 75.0 | 80.7 | 57.0 | 81.2 | 85.5 | 86.6 | 84.5 | **86.9** |
| | IncRes-v2 | 68.8 | 32.2 | 73.6 | 85.9 | 52.4 | 83.8 | **89.3** | 78.5 | 86.4 | 88.4 |
| | Swin-B | 21.0 | 10.7 | 27.8 | 34.6 | 17.6 | 34.9 | 44.3 | 23.9 | 37.9 | **45.4** |
| | DeiT-B | 26.5 | 17.4 | 34.2 | 38.1 | 21.4 | 37.2 | 50.0 | 29.2 | 40.3 | **50.4** |
| | Avg. | 46.1 | 53.5 | 53.0 | 60.8 | 37.4 | 60.5 | 67.8 | 56.2 | 63.0 | **68.2** |
| Res-101 | Res-152 | 66.4 | 34.6 | 72.7 | 84.5 | 61.5 | 80.7 | 85.7 | **90.2** | 86.0 | 87.7 |
| | Dense-121 | 60.9 | 37.9 | 66.5 | 78.3 | 55.1 | 78.4 | 83.6 | **90.6** | 81.7 | 86.2 |
| | IncRes-v2 | 40.5 | 17.8 | 32.8 | 48.0 | 23.1 | 62.5 | 72.4 | 72.1 | 53.9 | **76.9** |
| | Swin-B | 21.3 | 10.0 | 19.9 | 24.6 | 13.8 | 38.4 | 32.2 | 31.7 | 29.8 | **49.2** |
| | DeiT-B | 22.1 | 17.3 | 18.8 | 23.9 | 14.4 | 36.7 | 33.5 | 36.8 | 27.1 | **50.5** |
| | Avg. | 42.2 | 62.3 | 42.1 | 51.9 | 33.6 | 59.3 | 61.5 | 64.3 | 55.7 | **70.1** |
| ViT-B | Res-152 | 40.3 | 22.3 | 38.3 | 41.1 | 31.1 | 45.6 | **49.7** | 49.3 | 43.3 | 49.6 |
| | Dense-121 | 55.1 | 40.2 | 55.2 | 60.2 | 46.9 | 63.5 | 64.6 | 68.5 | 62.6 | **68.9** |
| | IncRes-v2 | 41.8 | 19.4 | 37.4 | 44.3 | 29.8 | 53.2 | 57.5 | 64.3 | 48.9 | **64.6** |
| | Swin-B | 44.8 | 14.1 | 39.1 | 48.3 | 37.7 | 54.8 | 60.9 | 52.8 | 51.2 | **61.2** |
| | DeiT-B | 72.8 | 35.6 | 76.8 | 83.9 | 67.9 | 89.1 | 91.4 | 76.6 | 85.8 | **92.7** |
| | Avg. | 51.0 | 26.3 | 49.4 | 55.6 | 42.7 | 61.2 | 64.8 | 62.3 | 58.4 | **67.4** |

### *4.3. Attacks on Advanced Defense Methods*

We validate the performance of our FIT against various defenses in this section, including four adversarially trained models and three advanced defense methods. We choose Inc-v3 as the surrogate model. For a fair comparison, the maximum perturbation magnitude is set to 4/255 for DiffPure and 8/255 for RDC, consistent with their default parameter configurations. Table 3 and Figure 7 show that, despite the challenging nature of attacking adversarially trained black-box models and defense methods, our method achieves the strongest attack performance among the compared methods. For instance, ATTA, SSA, and BSR achieve ASRs of 29.8%, 59.2%, and 48.6% on Inc-v3$_{adv}$, while our method achieves an ASR of 73.5%. These results demonstrate that our FIT can perform well against defenses.

**Table 3.** The ASRs (%) of various input transformation attacks against adversarially trained models. The bolded numbers indicate the best results.

| Model | DEM | SIM | Admix | ATTA | SSA | STM | BSR | SSEPs | FIT |
|---|---|---|---|---|---|---|---|---|---|
| Inc-v3$_{adv}$ | 47.1 | 43.8 | 53.2 | 29.8 | 59.2 | 69.4 | 48.6 | 55.6 | **73.5** |
| Inc-v3$_{ens3}$ | 35.3 | 37.3 | 45.4 | 23.8 | 50.5 | 56.1 | 48.5 | 48.3 | **66.5** |
| Inc-v3$_{ens4}$ | 34.6 | 38.2 | 45.6 | 23.3 | 51.3 | 55.5 | 46.9 | 48.7 | **65.0** |
| Inc-v2$_{ens}$ | 19.3 | 21.7 | 26.4 | 11.8 | 29.8 | 33.1 | 28.1 | 30.2 | **42.7** |

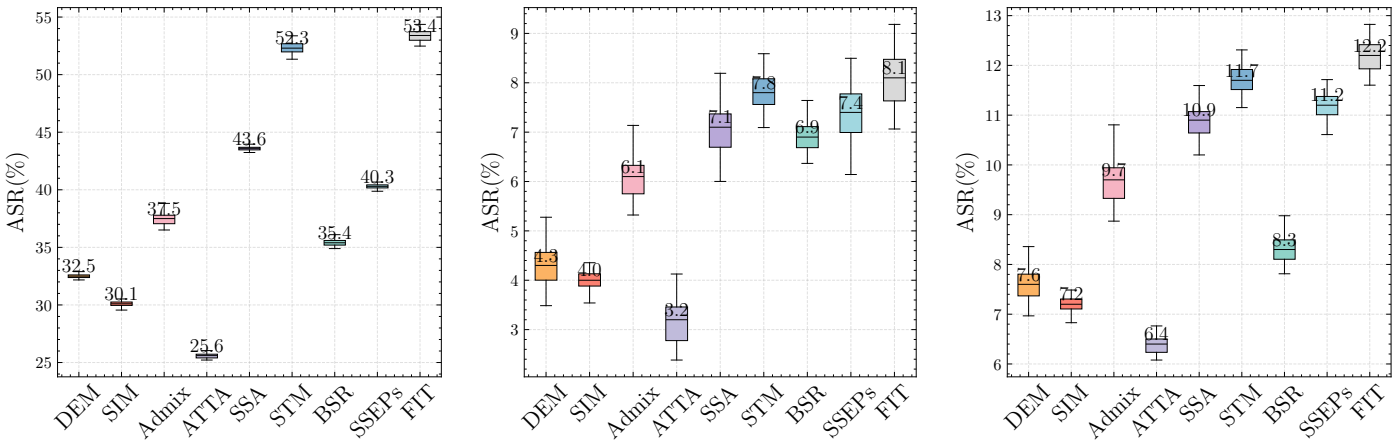



**Figure 7.** The ASRs (%) of various input transformation attacks against robust defense methods.

### *4.4. Attacks on Ensemble of Models*

We further validate the effectiveness of the proposed FIT in a model ensemble setting. The used surrogate ensemble includes Inc-v3, Inc-v4, and ResNet-101. Table 4 provides the ASRs on all normally and adversarially trained models. It can be observed that our method achieves an average ASR of 88.3% against the normally trained models, which is 18.5%, 51.1%, and 8.5% higher than DIM, TIM, and BSR, respectively. Meanwhile, its average ASR on the adversarially trained models is 94.9%, which is higher than the ASRs of SIM, Admix, and SSA (15.0%, 6.9%, and 5.4%, respectively). These results demonstrate the superior transferability of the proposed FIT. This further confirms the generalization capability and robustness of the proposed method.

**Table 4.** The ASRs (%) of various input transformation attacks in a model ensemble setting. The numbers in bold indicate the best results.

| **Attack** | **Surrogate Models: Inc-v3, Inc-v4, Res-101** | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | **DIM** | **TIM** | **DEM** | **SIM** | **Admix** | **ATTA** | **SSA** | **STM** | **BSR** | **SSEPs** | **FIT** |
| Res-152 | 80.8 | 39.7 | 93.4 | 79.6 | 88.2 | 67.3 | 89.1 | 91.7 | 93.7 | 90.7 | **94.0** |
| Dens-121 | 86.2 | 58.4 | 95.9 | 88.5 | 93.8 | 76.4 | 93.7 | 95.5 | 97.4 | 94.9 | **97.7** |
| IncR-v2 | 87.4 | 48.5 | 98.1 | 86.9 | 93.6 | 72.7 | 95.2 | 96.7 | 93.9 | 96.8 | **98.2** |
| Swin-B | 40.6 | 13.1 | 45.1 | 42.7 | 53.4 | 30.1 | 61.0 | 65.8 | 51.3 | 57.2 | **71.9** |
| DeiT-B | 54.0 | 26.3 | 64.5 | 52.7 | 63.0 | 40.4 | 70.2 | 78.9 | 62.6 | 65.4 | **79.6** |
| Avg. | 69.8 | 37.2 | 79.4 | 70.1 | 78.4 | 57.4 | 81.8 | 85.7 | 79.8 | 81.0 | **88.3** |
| Inc-v3$_{adv}$ | 85.1 | 58.8 | 96.4 | 85.1 | 92.2 | 67.7 | 92.7 | 96.1 | 91.0 | 94.5 | **96.9** |
| Inc-v3$_{ens3}$ | 80.8 | 55.7 | 94.2 | 81.5 | 89.4 | 61.1 | 90.5 | 95.2 | 87.6 | 90.8 | **95.9** |
| Inc-v3$_{ens4}$ | 80.1 | 56.9 | 93.6 | 80.8 | 88.7 | 60.7 | 90.0 | 94.7 | 84.3 | 91.1 | **95.0** |
| Inc-v2$_{ens}$ | 73.4 | 47.4 | 88.3 | 72.1 | 81.7 | 50.2 | 84.9 | 91.0 | 77.1 | 85.5 | **91.6** |
| Avg. | 79.8 | 54.7 | 93.1 | 79.9 | 88.0 | 59.9 | 89.5 | 94.3 | 85.0 | 90.5 | **94.9** |

### *4.5. Efficiency Analysis*

The overall computational complexity of our FIT is aligned with existing input-transformation-based attacks. Specifically, the dominant cost in transfer-based attacks comes from gradient computation. In our framework, the number of gradient evaluations per iteration is determined by the number of augmented samples $K$, which is consistent with most input transformation attacks (e.g., STM, BSR). Therefore, from the perspective of gradient computation, our method does not introduce additional overhead beyond what is already required by comparable input transformation techniques. Furthermore, we analyze the additional cost introduced by the Fourier-based transformations. In each

iteration, our method applies the DFT and inverse DFT operations to generate stylized and phase-based variants. These operations are highly optimized on modern GPUs. Importantly, these transformations are applied at the input level and do not involve forward or backward passes through deep networks, making them relatively lightweight compared to gradient computation. To provide a more concrete comparison, we measure the average runtime per iteration (on a single batch with 10 $229 \times 229 \times 3$ images) across different input transformation attacks under the same hardware setting. The comparison includes representative methods such as ATTA, Admix, STM, BSR, and our proposed FIT. The results in Table 5 show that although our method introduces a modest increase in per-iteration time due to Fourier transforms, the overall runtime remains shorter than other advanced transformation-based attacks, especially those involving complex augmentations (e.g., style transfer network, adversarial transfer network). This demonstrates that the proposed method achieves improved transferability without incurring prohibitive computational costs.

**Table 5.** The comparison between the runtime of various methods.

| Attacks | Admix | ATTA | SSA | STM | BSR | FIT |
|---|---|---|---|---|---|---|
| Runtime (s) | 0.99 | 1.66 | 1.18 | 1.49 | 0.73 | 1.35 |

### *4.6. Visualization of Attack Influence*

To further evaluate the effectiveness of the adversarial examples generated by our proposed FIT method, we employ the Gradient-weighted Class Activation Mapping (Grad-CAM) visualization technique [57], as illustrated in Figure 8. Grad-CAM generates heatmaps that highlight the regions of an input image most influential to the model's decision-making process. In the figure, the first row displays the original clean images, where the heatmaps predominantly concentrate on the salient objects consistent with the model's accurate classification. In contrast, the second row shows the corresponding adversarial examples on Resnet101, where the heatmaps now emphasize extraneous features (e.g., background textures or peripheral artifacts) while disregarding the target object. This shift in attention distribution demonstrates that adversarial examples generated by FIT effectively mislead the model into focusing on unrelated regions, thereby resulting in incorrect predictions.

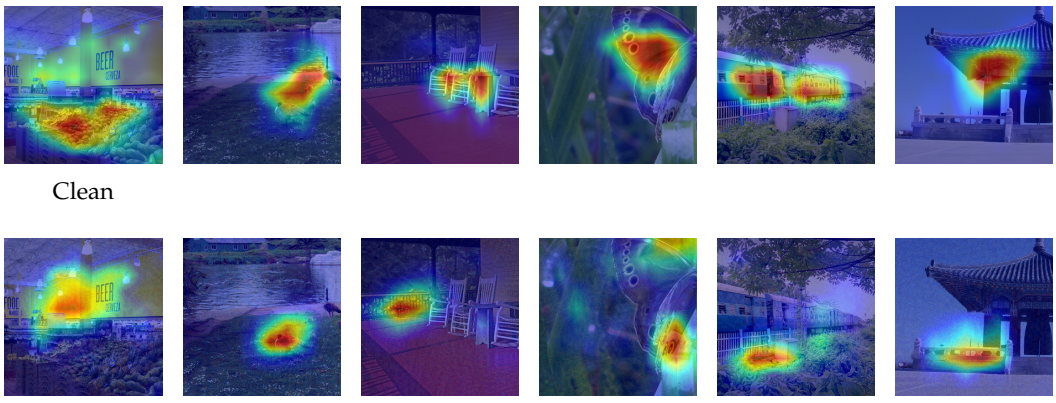



**Figure 8.** Visualization of heatmaps for clean images (first row) and their corresponding adversarial examples generated by FIT (second row).

### *4.7. Evaluation in Real World Applications*

In this section, we evaluate the performance of FIT in more challenging real-world application scenarios. Following our experimental setup, we adopt Inc-v3 as the surrogate model and set the maximum perturbation to 16/255. The generated adversarial examples are then used to attack the Google Cloud Vision API, which outputs predicted labels along with confidence scores. This setting is particularly challenging, as real-world vision systems do not provide access to gradients or internal model parameters, thereby constituting a fully black-box attack scenario. The attack results show a transfer success rate of 19.8%, demonstrating that our method remains effective even in a fully black-box, real-world setting. Figure 9 shows two examples of successful attacks. In the first row, the clean image is classified with the highest prediction probability as bird, whereas its adversarial example is misclassified as fish. Similarly, in the second row, the highest predicted label for the clean image is deer, which shifts to bat following the adversarial perturbation. These outcomes show that the adversarial examples generated by FIT successfully alter the predicted labels returned by Google Cloud Vision, demonstrating superior transferability.

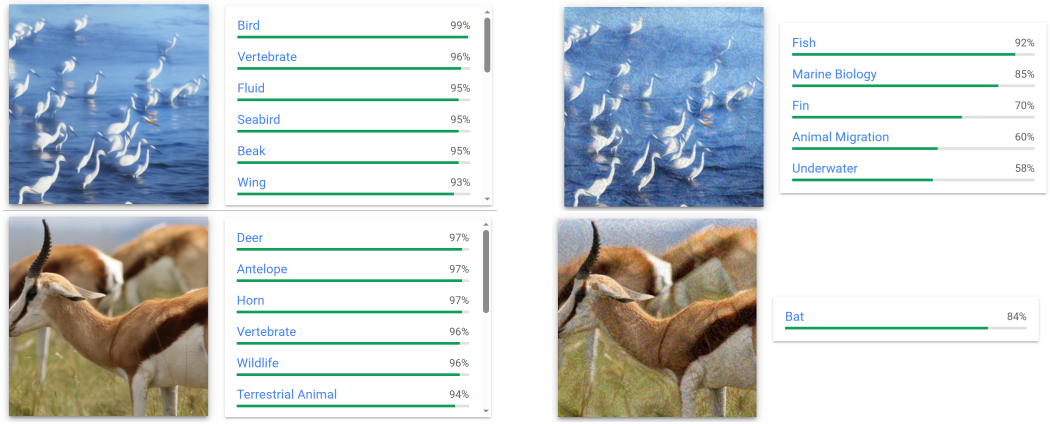


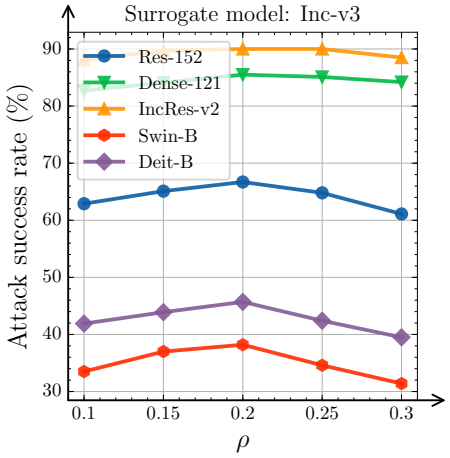

**Figure 9.** Feedback from attacks on the Google Cloud Vision API. The left column shows clean images, while the right column presents adversarial examples generated by FIT using the surrogate model Inc-v3.

### *4.8. Ablation Study*

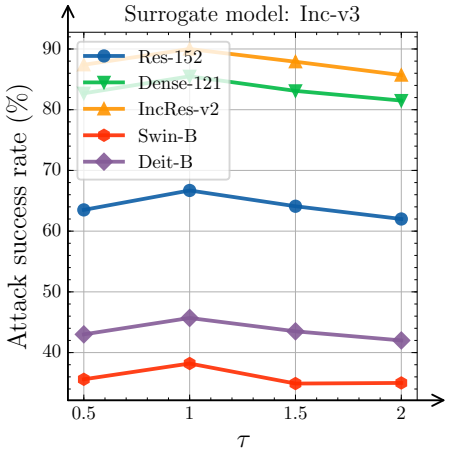
We analyze the impact of the hyperparameters: the semantic mixup weight $\rho$ and the mixing coefficient $\lambda \sim Beta(\tau, \tau)$. The surrogate models are Inc-v3 and ViT-B, respectively.

The semantic mixup weight $\rho$: Figure 10a presents the ASRs as a function of $\rho$. Performance increases when $\rho$ is raised from 0.1 to 0.2 but declines for values $\rho > 0.2$. As incorporating information from other categories into the input can effectively augment data, but higher values (e.g., $\rho = 0.3$) lead to over-perturbation and a drop in performance. Additionally, as shown in Figure 10c, we observe that the trend when using the ViT-B is highly consistent with that using the Inc-v3. Consequently, we select $\rho = 0.2$ as the default.

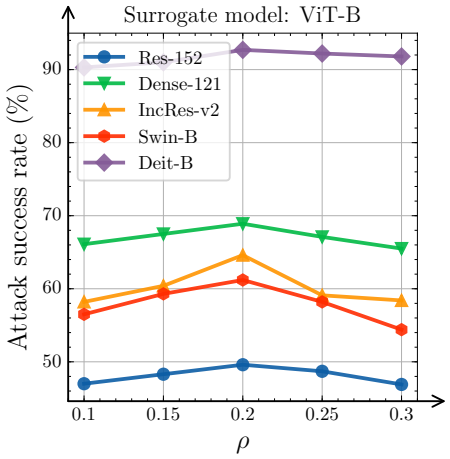
The mixing coefficient $\lambda \sim Beta(\tau, \tau)$: Compared to using fixed values, sampling $\lambda$ from a Beta distribution provides greater diversity and balance. We analyze the impact of the distribution parameter $\tau$ on the ASR. As shown in Figure 10b, for normally trained models, the ASRs reach their peak when $\tau = 1$, before declining for higher $\tau$ values. On the other hand, Figure 10d shows that the ASRs also reach their peak at $\tau = 1$ when the surrogate model is the ViT-B. Therefore, $\lambda$ is sampled from the Beta$(1, 1)$ distribution.



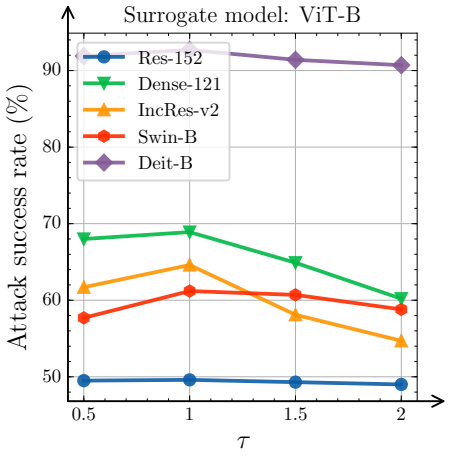

(**a**) Impact of $\rho$ on ASRs with Inc-v3 as the surrogate model.




(**b**) Impact of $\tau$ on ASRs with Inc-v3 as the surrogate model.




(**c**) Impact of $\rho$ on ASRs with ViT-B as the surrogate model.




(**d**) Impact of $\tau$ on ASRs with ViT-B as the surrogate model.

**Figure 10.** Impact of hyperparameters $\rho$ and $\tau$ on the transferability of adversarial examples.

## 5. Conclusions

In this paper, we propose the FIT method, a novel input transformation that integrates stylistic transform and semantic mixup in the frequency domain to increase input diversity. The stylistic transform component changes the style of the original images by modulating the amplitude spectrum, while the semantic mixup introduces high-level semantics by incorporating phase-only representations from other images. We integrate FIT in the iterative gradient-based attack for generating adversarial examples with high transferability. Experimental results on diverse models and defense mechanisms verify that our method achieves superior transfer performance. However, the current study focuses on image-domain adversarial attacks, and the proposed framework is designed and evaluated under visual recognition settings. Extending the framework to other modalities remains an open problem. Future work could extend FIT to the audio settings. For instance, FIT can be adapted to operate on time–frequency representations (e.g., spectrograms), where amplitude and phase correspond to timbral and structural information, respectively. This may enable analogous stylistic transformation and semantic-preserving mixup in the audio domain to improve the transferability of audio adversarial examples. Furthermore, since FIT explicitly exploits the decoupling of stylistic and semantic components, it suggests

that models may be vulnerable to perturbations that independently manipulate different feature levels. This observation indicates that potential defenses could benefit from jointly modeling low-level statistics and high-level structure to reduce such vulnerabilities.

**Author Contributions:** Conceptualization, Y.L.; methodology, Z.T.; data curation, L.Z.; writing—original draft preparation, Z.T.; writing—review and editing, Y.L.; visualization, X.W. 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:** ImageNet-compatible dataset is available at https://github.com/cleverhans-lab/cleverhans/tree/master/cleverhans_v3.1.0/examples/nips17_adversarial_competition/dataset (accessed on 2 March 2021).

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

## References

1. Zhan, Y.; Zheng, B.; Liu, D.; Deng, B.; Yang, X. Exploring black-box adversarial attacks on Interpretable Deep Learning Systems. *Comput. Vis. Image Underst.* **2025**, *259*, 104423. [CrossRef]
2. Riaz, K.; Anjum, M.L.; Hussain, W.; Manzoor, R. Targeted adversarial attack on classic vision pipelines. *Comput. Vis. Image Underst.* **2024**, *249*, 104140. [CrossRef]
3. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.J.; Fergus, R. Intriguing properties of neural networks. In Proceedings of the ICLR, Banff, AB, Canada, 14–16 April 2014 .
4. Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; Li, J. Boosting Adversarial Attacks With Momentum. In Proceedings of the CVPR, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9185–9193.
5. Wang, X.; He, K. Enhancing the Transferability of Adversarial Attacks Through Variance Tuning. In Proceedings of the CVPR, Virtual, 19–25 June 2021; pp. 1924–1933.
6. Wang, X.; Lin, J.; Hu, H.; Wang, J.; He, K. Boosting Adversarial Transferability through Enhanced Momentum. In Proceedings of the BMVC, Online, 22–25 November 2021; p. 272.
7. Chen, H.; Zhang, Y.; Dong, Y.; Yang, X.; Su, H.; Zhu, J. Rethinking Model Ensemble in Transfer-based Adversarial Attacks. In Proceedings of the ICLR, Vienna, Austria, 7–11 May 2024.
8. Tang, B.; Wang, Z.; Bin, Y.; Dou, Q.; Yang, Y.; Shen, H.T. Ensemble diversity facilitates adversarial transferability. In Proceedings of the CVPR, Seattle, WA, USA, 16–22 June 2024; pp. 24377–24386.
9. Ge, Z.; Shang, F.; Liu, H.; Liu, Y.; Wan, L.; Feng, W.; Wang, X. Improving the Transferability of Adversarial Examples with Arbitrary Style Transfer. In Proceedings of the ACM MM, Ottawa, ON, Canada, 29 October–3 November 2023; pp. 4440–4449.
10. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the ICLR, San Diego, CA, USA, 7–9 May 2015.
11. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In Proceedings of the ICLR 2017 Workshop Track, Toulon, France, 24–26 April 2017.
12. 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 ECCV, Tel Aviv, Israel, 23–27 October 2022; pp. 549–566.
13. Wang, K.; He, X.; Wang, W.; Wang, X. Boosting Adversarial Transferability by Block Shuffle and Rotation. In Proceedings of the CVPR, Seattle, WA, USA, 16–22 June 2024.
14. 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 CVPR, Long Beach, CA, USA, 16–20 June 2019; pp. 2730–2739.
15. Dong, Y.; Pang, T.; Su, H.; Zhu, J. Evading Defenses to Transferable Adversarial Examples by Translation-Invariant Attacks. In Proceedings of the CVPR, Long Beach, CA, USA, 16–20 June 2019; pp. 4312–4321.
16. Lin, J.; Song, C.; He, K.; Wang, L.; Hopcroft, J.E. Nesterov Accelerated Gradient and Scale Invariance for Adversarial Attacks. In Proceedings of the ICLR, Addis Ababa, Ethiopia, 26–30 April 2020.
17. Wu, W.; Su, Y.; Lyu, M.R.; King, I. Improving the Transferability of Adversarial Samples With Adversarial Transformations. In Proceedings of the CVPR, Virtual, 19–25 June 2021; pp. 9024–9033.
18. Wang, X.; He, X.; Wang, J.; He, K. Admix: Enhancing the Transferability of Adversarial Attacks. In Proceedings of the ICCV, Montreal, QC, Canada, 10–17 October 2021; pp. 16138–16147.

19. Qian, Y.; Chen, K.; Wang, B.; Gu, Z.; Ji, S.; Wang, W.; Zhang, Y. Enhancing transferability of adversarial examples through mixed-frequency inputs. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 7633–7645. [CrossRef]
20. Oppenheim, A.; Lim, J.; Kopec, G.; Pohlig, S. Phase in speech and pictures. In *Proceedings of the ICASSP'79. IEEE International Conference on Acoustics, Speech, and Signal Processing*; IEEE: Washington, DC, USA, 2–4 April 1979; Volume 4; pp. 632–637.
21. Oppenheim, A.V.; Lim, J.S. The importance of phase in signals. *Proc. IEEE* **2005**, *69*, 529–541. [CrossRef]
22. Piotrowski, L.N.; Campbell, F.W. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception* **1982**, *11*, 337–346. [CrossRef] [PubMed]
23. Hansen, B.C.; Hess, R.F. Structural sparseness and spatial phase alignment in natural scenes. *J. Opt. Soc. Am. A* **2007**, *24*, 1873–1885. [CrossRef] [PubMed]
24. Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; Tian, Q. A fourier-based framework for domain generalization. In Proceedings of the CVPR, Virtual, 19–25 June 2021; pp. 14383–14392.
25. Zhou, D.; Wang, N.; Yang, H.; Gao, X.; Liu, T. Phase-aware adversarial defense for improving adversarial robustness. In Proceedings of the ICML, Honolulu, HI, USA, 23–29 July 2023; pp. 42724–42741.
26. Li, F.; Li, K.; Wu, H.; Tian, J.; Zhou, J. DAT: Improving Adversarial Robustness via Generative Amplitude Mix-up in Frequency Domain. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 127099–127128.
27. Xiang, W.; Su, H.; Liu, C.; Guo, Y.; Zheng, S. Improving the robustness of adversarial attacks using an affine-invariant gradient estimator. *Comput. Vis. Image Underst.* **2023**, *229*, 103647. [CrossRef]
28. Guo, S.; Wang, B.; Li, X.; Zhu, P.; Wu, L.; Wu, W.; Jia, L. ALDA: Enhancing the transferability of adversarial attacks with attention-guided look-ahead and data augmentation. *Pattern Recognit.* **2025**, *172*, 112686. [CrossRef]
29. Lin, Z.; Wang, B.; Wang, X.; Zhang, Y.; Wang, X.; Deng, K.; Peng, A.; Tang, J.; Yang, X. Improving adversarial transferability via semantic-style joint expectation perturbations. *Pattern Recognit.* **2026**, *172*, 112474. [CrossRef]
30. Qin, Z.; Fan, Y.; Liu, Y.; Shen, L.; Zhang, Y.; Wang, J.; Wu, B. Boosting the Transferability of Adversarial Attacks with Reverse Adversarial Perturbation. In Proceedings of the NeurIPS, New Orleans, LA, USA, 28 November–9 December 2022.
31. Ge, Z.; Liu, H.; Xiaosen, W.; Shang, F.; Liu, Y. Boosting adversarial transferability by achieving flat local maxima. In Proceedings of the NeurIPS, New Orleans, LA, USA, 10–16 December 2023; pp. 70141–70161.
32. Gan, F.; Wo, Y. Boosting the Transferability of Adversarial Examples Through Gradient Aggregation. *IEEE Trans. Inf. Forensics Secur.* **2025**, *20*, 5563–5576. [CrossRef]
33. Peng, J.; Tao, Z.; Wang, H.; Wang, M.; Wang, Y. Boosting Adversarial Transferability via Residual Perturbation Attack. In Proceedings of the ICCV, Honolulu, HI, USA, 19–23 October 2025; pp. 1261–1270.
34. Xiong, Y.; Lin, J.; Zhang, M.; Hopcroft, J.E.; He, K. Stochastic Variance Reduced Ensemble Adversarial Attack for Boosting the Adversarial Transferability. In Proceedings of the CVPR, New Orleans, LA, USA, 18–24 June 2022; pp. 14963–14972.
35. Chen, B.; Yin, J.; Chen, S.; Chen, B.; Liu, X. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In Proceedings of the ICCV, Paris, France, 1–6 October 2023; pp. 4489–4498.
36. Wang, H.; Wu, X.; Huang, Z.; Xing, E.P. High-Frequency Component Helps Explain the Generalization of Convolutional Neural Networks. In Proceedings of the CVPR, Seattle, WA, USA, 13–19 June 2020; pp. 8681–8691.
37. Yin, D.; Lopes, R.G.; Shlens, J.; Cubuk, E.D.; Gilmer, J. A Fourier Perspective on Model Robustness in Computer Vision. In Proceedings of the NeurIPS, Vancouver, BC, Canada, 8–14 December 2019 ; pp. 13255–13265.
38. Sharma, Y.; Ding, G.W.; Brubaker, M.A. On the Effectiveness of Low Frequency Perturbations. In Proceedings of the IJCAI, 2019, Macao, China, 10–16 August 2019; pp. 3389–3396.
39. Duan, R.; Chen, Y.; Niu, D.; Yang, Y.; Qin, A.K.; He, Y. AdvDrop: Adversarial Attack to DNNs by Dropping Information. In Proceedings of the ICCV, Montreal, QC, Canada, 10–17 October 2021; pp. 7486–7495.
40. Wang, Y.; Wu, Y.; Wu, S.; Liu, X.; Zhou, W.; Zhu, L.; Zhang, C. Boosting the Transferability of Adversarial Attacks with Frequency-Aware Perturbation. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 6293–6304. [CrossRef]
41. Zeiler, M.D.; Fergus, R. Visualizing and understanding convolutional networks. In Proceedings of the ECCV, Zurich, Switzerland, 6–12 September 2014; pp. 818–833.
42. Baker, N.; Lu, H.; Erlikhman, G.; Kellman, P.J. Deep convolutional networks do not classify based on global object shape. *PLoS Comput. Biol.* **2018**, *14*, e1006613. [CrossRef] [PubMed]
43. Zhang, H.; Cissé, M.; Dauphin, Y.N.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. In Proceedings of the ICLR, Vancouver, BC, Canada, 30 April–3 May 2018.
44. K, A.; Hamner, B.; Goodfellow, I. NIPS 2017: Non-Targeted Adversarial Attack. Kaggle. 2017. Available online: https://kaggle.com/competitions/nips-2017-non-targeted-adversarial-attack (accessed on 2 October 2017 ).
45. Wightman, R. PyTorch Image Models. 2019. Available online: https://github.com/rwightman/pytorch-image-models (accessed on 30 December 2025).
46. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the CVPR, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.

47. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
48. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the ICCV, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
49. 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, San Francisco, CA, USA, 4–9 February 2017; pp. 2818–2826.
50. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the ICML, Virtual Event, 18–24 July 2021; pp. 10347–10357.
51. Tramèr, F.; Kurakin, A.; Papernot, N.; Goodfellow, I.J.; Boneh, D.; McDaniel, P.D. Ensemble Adversarial Training: Attacks and Defenses. In Proceedings of the ICLR, Vancouver, BC, Canada, 30 April–3 May 2018.
52. Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; Anandkumar, A. Diffusion Models for Adversarial Purification. In Proceedings of the ICML, Baltimore, MD, USA, 17–23 July 2022; pp. 16805–16827.
53. Chen, H.; Dong, Y.; Wang, Z.; Yang, X.; Duan, C.; Su, H.; Zhu, J. Robust Classification via a Single Diffusion Model. In Proceedings of the ICML, Vienna, Austria, 21–27 July 2024.
54. Naseer, M.; Khan, S.; Hayat, M.; Khan, F.S.; Porikli, F. A self-supervised approach for adversarial robustness. In Proceedings of the CVPR, Seattle, WA, USA, 13–19 June 2020; pp. 262–271.
55. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In Proceedings of the CVPR, Las Vegas, NV, USA, 27–30 June 2016; pp. 2818–2826.
56. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale. In Proceedings of the ICLR, Virtual Event, Austria, 3–7 May 2021.
57. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In Proceedings of the ICCV, Venice, Italy, 22–29 October 2017; pp. 618–626.


**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.