# Enhancing Transferability of Adversarial Examples Through Mixed-Frequency Inputs

Yaguan Qian, Kecheng Chen, Bin Wang, Zhaoquan Gu, *Member, IEEE*, Shouling Ji, *Member, IEEE*, Wei Wang, *Member, IEEE*, and Yanchun Zhang, *Member, IEEE*

***Abstract*— Recent studies have shown that Deep Neural Networks (DNNs) are easily deceived by adversarial examples, revealing their serious vulnerability. Due to the transferability, adversarial examples can attack across multiple models with different architectures, called transfer-based black-box attacks. Input transformation is one of the most effective methods to improve adversarial transferability. In particular, the attacks fusing other categories of image information reveal the potential direction of adversarial attacks. However, the current techniques rely on input transformations in the spatial domain, which ignore the frequency information of the image and limit its transferability. To tackle this issue, we propose Mixed-Frequency Inputs (MFI) based on a frequency domain perspective. MFI alleviates the overfitting of adversarial examples to the source model by considering high-frequency components from various kinds of images in the process of calculating the gradient. By accumulating these high-frequency components, MFI acquires a more steady gradient direction in each iteration, leading to the discovery of better local maxima and enhancing transferability. Extensive experimental results on the ImageNet-compatible datasets demonstrate that MFI outperforms existing transform-based attacks with a clear margin on both Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), which proves MFI is more suitable for realistic black-box scenarios.**

***Index Terms*— Security vulnerability, adversarial examples, transfer-based attack, Fourier transform.**

## I. Introduction

DEEP neural networks (DNNs) have shown outstanding performance in various fields of computer vision and permeate our daily lives [1]. However, it has been shown that DNNs are threatened by adversarial examples [2], [3], which deceive the model by adding well-designed noise, leading to the misclassification of DNNs. This is also called *adversarial attacks*, which poses a great threat to security-sensitive application scenarios, such as face recognition [4], speech recognition system [5], and autonomous driving systems [6]. Therefore, we need a deep understanding of DNNs [7], [8], [9] to better identify the vulnerability of models through more powerful attacks. It is the first step towards improving the robustness of deep learning models against adversarial attacks [10], [11], [12].

Adversarial attacks are generally divided into two categories: white-box attacks [3], [13], [14], [15] and black-box attacks [16], [17], [18], [19]. In white-box attacks, the adversary has access to the full information of the target model. In contrast, in the black-box attack scenario, the adversary can only access the output of the target model. Compared with white-box attacks, black-box attacks are more consistent with the real world and more applicable to real-world systems. Thus, we focus on black-box attacks to align realistic scenarios in daily life.

For black-box attacks, there are two basic approaches: query-based [20], [21] and transfer-based [22], [23], [24]. Query-based attacks interact with the target model to generate adversarial examples, which require a large number of queries to the target model and bring heavy time costs. In contrast, transfer-based attacks utilize an intriguing property of adversarial examples called *transferability* [25], [26], [27], [28], in which the crafted adversarial examples can attack across multiple models successfully. Specifically, it creates adversarial examples using a source model without requesting the target model. Therefore, compared with white-box attacks and query-based attacks, transfer-based attacks have attracted huge attention due to their high efficiency, and become a popular attack technique.

In the subsequent research, improving the transferability of adversarial examples becomes a critical issue, which mainly focuses on gradient optimization [13], [14], ensemble and model augmentation [23], [26], input transformation [17], [18], [22], [25], and advanced loss functions [29], [30]. One of the most effective methods is input transformation, including random resizing and padding [25], translation [17], scaling [18], etc. These methods transform the input image and then feed it into a classifier for diversity processing, similar to data augmentation techniques. However, they are applied to a single input image, which limits the transferability of adversarial examples [22]. In this paper, we consider multiple

Manuscript received 15 January 2024; revised 23 April 2024, 17 May 2024, and 11 July 2024; accepted 12 July 2024. Date of publication 18 July 2024; date of current version 21 August 2024. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3102100; in part by the National Natural Science Foundation of China under Grant 92167203, Grant 62372137, and Grant U21A20463; and in part by Zhejiang Provincial Natural Science Foundation of China under Grant LZ22F020007. The associate editor coordinating the review of this article and approving it for publication was Prof. Yanjiao Chen. *(Corresponding authors: Bin Wang; Zhaoquan Gu.)*

Yaguan Qian and Kecheng Chen are with the School of Big-Data Science, Zhejiang University of Science and Technology, Hangzhou 310023, China (e-mail: qianyaguan@zust.edu.cn; 222209252015@zust.edu.cn).

Bin Wang is with Zhejiang Key Laboratory of Artificial Intelligence of Things (AIoT) Network and Data Security, Hangzhou 310053, China (e-mail: wbin2006@gmail.com).

Zhaoquan Gu is with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), Shenzhen 518071, China (e-mail: guzhaoquan@hit.edu.cn).

Shouling Ji is with the College of Computer Science and Technology, Zhejiang University, Hangzhou 310058, China (e-mail: sji@zju.edu.cn).

Wei Wang is with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: wangwei1@bjtu.edu.cn).

Yanchun Zhang is with the School of Computer Science and Technology, Zhejiang Normal University, Jinhua 321004, China, and also with the School of Computer Science and Mathematics, Victoria University, Melbourne, VIC 8001, Australia (e-mail: yanchun.zhang@vu.edu.au).

Digital Object Identifier 10.1109/TIFS.2024.3430508

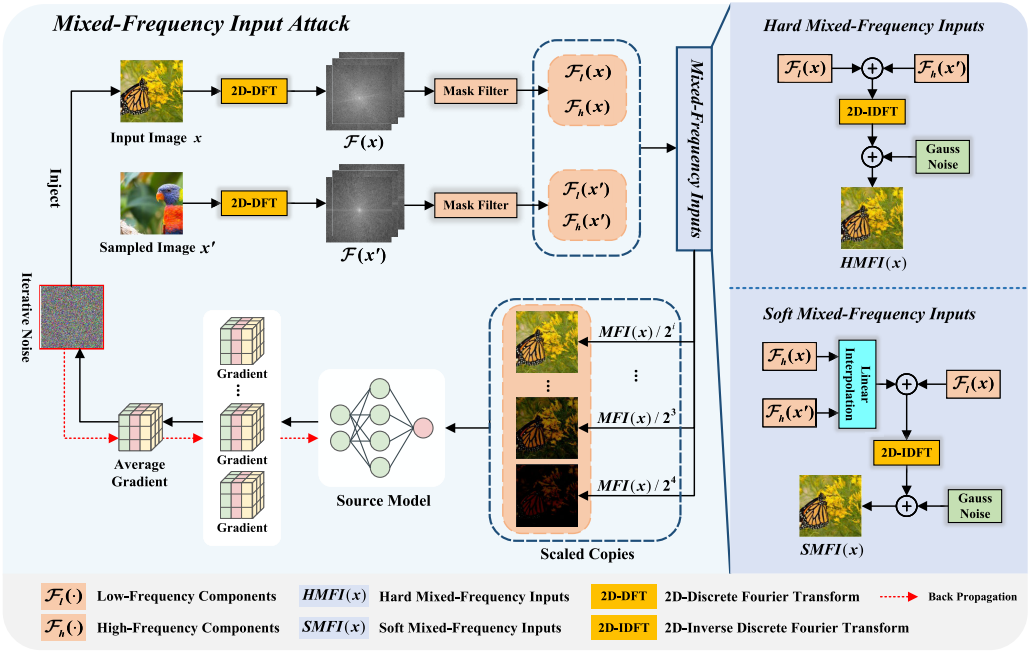



Fig. 1. Overview of the proposed Mixed-Frequency Input (MFI) attack. The overall pipeline of the MFI attack is presented on the left. The core of image processing for MFI is shown on the right, including hard mixed-frequency inputs (HMFI) and soft mixed-frequency inputs (SMFI). First, it exchanges the high-frequency components of each input image with randomly sampled images from a set. Then, the transformed image is scaled uniformly, generating multiple scale copies. Finally, the gradient is calculated by all scaled images.

inputs and strive for more diverse and efficient transformation methods.

Mixup [31] is a data augmentation technology to improve the generalization of standard training, which creates new synthetic examples while softening the labels in a probabilistic manner. Drawing inspiration from Mixup, Wang et al. [22] propose a new transfer-based attack, Admix, which calculates the gradient on the admixed image combined with the original input and images randomly picked from other categories in the spatial domain. However, from the view of frequency domains, they ignored the negative impact of high-frequency features in input images, resulting in a certain loss of transferability. Wang et al. [32] show that DNN is biased towards high-frequency patterns of images to some extent [32]. Furthermore, the learning priority of DNNs shifts quickly from low frequency to high frequency [33], and throughout the training phase, the model mostly learns high-frequency features. Since DNNs excessively focus on high-frequency features, it may cause the model to be overly sensitive to noise or slight variations in the data. Therefore, Admix still cannot completely avoid overfitting the model.

To address the above challenge, we propose a novel input transformation method based on the frequency domain perspective, called mixed-frequency inputs (MFI). MFI biases the input image towards other classes by adding high-frequency components of other images to the original. Besides, we embed MFI into the adversarial attack framework and present a black-box attack method. The whole procedure of the MFI attack is exhibited in Fig. 1.

To evaluate our approach, we conduct experiments on the ImageNet-compatible dataset, which contains 1000 images provided by the NIPS 2017 adversarial attack competition [34]. Extensive experiments confirm that the average attack success rate of our method is significantly improved in the target models, especially adversarial training models and advanced defense models, reaching 93.7% and 92.7% on average, respectively. Finally, we demonstrate our approach to the real Google Cloud Vision service successfully.

To sum up, our main contributions are summarized as follows:

- We devise a novel input transformation strategy based on the frequency domain perspective, *i.e.*, mixed-frequency inputs (MFI). MFI alleviates the high-frequency pitfall of the origin image and improves the diversity of the transformed image.
- We integrate MFI into the adversarial attack framework and propose a new black-box attack method. The adversarial examples generated by this innovative approach obtain enhanced transferability while exhibiting remarkable ability in the white-box setting.
- The extensive experiments on the ImageNet-compatible dataset show that MFI attack has significant advantages over the baseline, particularly against adversarial training models and advanced defense models in ensemble settings, achieving an average attack success rate of 93.7% and 92.7%, respectively. Additionally, MFI is compatible with other input transformation techniques.

## II. Related Work

To identify the vulnerability of DNNs, various attack methods have been proposed recently, such as optimization-based attacks [2], [13], query-based attacks [20], [21], and transfer-based attacks [22], [23], [24], etc. Among these, transfer-based attacks do not access anything from the target model, making them applicable to attack any model in the physical world.

In this work, we focus on generating more transferable adversarial examples and briefly introduce the mainstream methods of transfer-based attacks, as well as the research on adversarial examples in the frequency domain in recent years.

### A. Transfer-Based Attacks

Transfer-based attacks have received increasing attention due to their high efficiency. In order to improve the transferability of adversarial examples, there are two main ways described in previous works. One is the gradient optimization, and the other is the input transformation. Their commonality is to avoid the generated adversarial examples falling into the local optimum of the source model, which is crucial to transferability.

*1) Gradient-Based Optimization Attacks:* The objective of gradient optimization is to escape from bad local optima by adjusting the iteration direction of the vanilla gradient when generating adversarial examples. A typical method is MI-FGSM, proposed by Dong et al. [26], which integrates the momentum term into I-FGSM [14] to improve its adversarial transferability. Based on this, Lin et al. [18] made use of Nesterov to accelerate the gradient to accumulate momentum to achieve better transferability. Different from the previous iterative attacks, Gao et al. [19] reconsidered the drawbacks of direct clipping noise and created adversarial examples by patch-wise noise. Besides, Wang and He [27] utilized the gradient variance along the momentum optimization path to avoid overfitting. Zhu et al. [35] explored the fluctuation phenomenon on the plus-minus sign of the adversarial perturbations and proposed gradient relevance frameworks aimed at uncovering potential neighbor information surrounding the input.

*2) Input-Transformation-Based Attacks:* Methods based on input transformation are considered to be an effective way to avoid overfitting source models during training. It aims to prevent adversarial examples from overfitting the white-box model and failing to transfer to the black-box model. Xie et al. [25] first proposed to apply a variety of input patterns and random scaling and padding to calculate gradients. Dong et al. [17] translated the input, generated a series of translation images, and approximately estimated the overall gradient to alleviate the problem of excessive dependence on source models. Lin et al. [18] discovered the scale-invariant property of DNNs and proposed a loss-preserving transformation. Inspired by this, Long et al. [23] simulated model augmentation by enhancing the spectrum saliency map in the frequency domain. Recently, Wang et al. [24] summarize transformations on singular input and hypothesized that the more diverse transformed images result in better transferability.

Furthermore, beyond the manipulation of individual images, various techniques have made strides in handling mixed images. Wang et al. [22] averaged gradients across a set of mixed images, which modify the input images by blending other categories of images while maintaining the input image labels. Zhao et al. [36] achieved multiple data augmentation by leveraging gradients from previous iterations and images from other categories in the same iteration. Besides, Xu and Ghamisi [37] first generated virtual examples that do not belong to any category by mixing image slices of different categories and then crafted adversarial examples by iteratively maximizing the loss function.

### B. Interpretability of DNN in Frequency Domain

In recent years, numerous investigations have looked into the frequency characteristics of models and the sensitivity of DNNs [38], [39], [40]. Tsuzuku and Sato [38] first proposed a frequency framework by studying the sensitivity of DNNs to different Fourier bases. Yin et al. [39] investigated the sensitivity of the model to high-frequency and low-frequency corruptions through perturbation analysis in the Fourier domain. Abello et al. [40] divided the spectrum into incoherent disks based on the energy distribution and found that medium and high frequency are particularly important for DNNs. Wang et al. [32] found that DNNs can capture the high-frequency components of images that are almost imperceptible to humans, indicating that high-frequency components play an important role in improving the accuracy of DNNs.

## III. Motivation

Adversarial examples generated by traditional white-box attacks often suffer from overfitting to the source model, resulting in lower transferability [26]. In the context of transfer-based black-box attacks, it is crucial to mitigate this overfitting on the source model (substitute model). In other words, our goal is to enhance the transferability between the source model and the target model. One effective approach to achieve this is through input transformation. Input transformation has been recognized as a viable method to alleviate overfitting on the source model. Guo et al. [41], [42] have demonstrated that when adversarial examples undergo a simple transformation, their attack potency diminishes. This finding suggests that input transformation can effectively reduce the reliance of adversarial examples on the specific characteristics of the source model.

However, the existing methods of input transformation primarily focus on spatial domain implementations [17], [18], [25]. These methods enhance the source model by applying loss-preserving transformations in the spatial domain [18], which may overlook the fundamental differences between models and limit the diversity of the source models. Moreover, models tend to exhibit distinct behaviors in the frequency domain compared to the spatial domain. Recent studies [32], [39] have revealed that different models often rely on different frequency components of input images while making decisions. This observation motivates us to explore input transformation techniques in the frequency domain as a potential solution.

Previous studies have demonstrated that the essential content-defining information in natural images primarily resides in the low-frequency components, while high-frequency signals often correspond to data distribution and natural noise [32], [39]. According to [32], DNNs have the ability to capture high-frequency components (HFCs) and largely relies on HFC to make predictions, which implies an

inseparable correlation between HFC and the labels. In addition, the learning priority of DNNs shifts from low frequency to high frequency within a short time and dedicates a significant portion of their learning capacity to fit high-frequency features during the natural training phase [33].

However, we traditionally craft adversarial examples on naturally trained models, which forces us to rethink the impact of high-frequency features. We hypothesize that breaking the alignment of high-frequency features and class labels can effectively alleviate the problem of overfitting the source model in the iterative process of adversarial examples. To this end, we proposed MFI, a methodology designed to mitigate the adverse effects caused by the high-frequency attributes present in the original image on DNNs.

## IV. Methodology

In this section, we describe our method in detail. First, we provide a formal description of transfer-based adversarial attacks in Section IV-A. Then we introduce the 2-dimensional discrete Fourier transform (2D-DFT) in Section IV-B, which is an important tool for processing the frequency domain of images. In this paper, we utilize 2D-DFT to realize mixing frequency. Next, we illustrate an input transformation strategy called mixed-frequency inputs (MFI) in Section IV-C. Finally, we integrate MFI into the attack framework to simulate model augmentation in the frequency domain and achieve better transferability.

### A. Transfer-Based Adversarial Attacks

Let $x$ be a benign image, $y$ be the corresponding ground-truth label, and $f(x;\theta)$ be a DNN model with parameters $\theta$. Let $\mathcal{J}$ denote the loss function (*e.g.*, cross-entropy loss) of $f$. The goal of adversarial attacks is to find a tiny perturbation $\delta$, fooling the classifier $f(x^{adv};\theta) \neq y$, where adversarial examples $x^{adv} = x+\delta$. Generally, to guarantee that adversarial examples resemble benign images as closely as feasible, the norm of $\delta$ is required to satisfy $\|\delta\|_p \leq \epsilon$, where $\epsilon$ is named perturbation budget, $\|\cdot\|_p$ is the $p$-norm, and $p$ could be 1, 2, $\infty$. In this paper, we adopt $p = \infty$ to align with previous works. Accordingly, a *white-box adversarial attack* can be modeled as a constrained optimization problem as follows:

$$\arg\max_{\delta} \mathcal{J}(f(x+\delta;\theta), y), \ \ s.t.\ \|\delta\|_\infty \leq \epsilon. \tag{1}$$

where $f$ is called the *source model* in the white-box attack scenario.

As mentioned in the Introduction section, white-box attacks are generally impractical in real-world scenarios. Therefore, in this paper, our focus is on transfer-based black-box attacks. In this scenario, we assume the presence of two models: the *substitute model* $f_s(\cdot)$ and the *target model* $f_t(\cdot)$. The substitute model can be fully controlled by the adversary locally, while the internal information (e.g., model parameters) of the remote target model remains inaccessible. To realize the cross-model attack, we leverage the inherent property of transferability in adversarial examples. This property describes the phenomenon where an adversarial example that fools the model $f_s(\cdot)$ can also fool the model $f_t(\cdot)$. Specifically, we consider the substitute model as the source model for conducting a white-box attack to generate an adversarial example. We then utilize this adversarial example to attack the remote target model $f_t(\cdot)$. Therefore, the essence of a transfer-based black-box attack can be viewed as a white-box attack (as shown in Eq. (1)) performed on the substitute model, simulating an attack on the remote target model:

$$\begin{aligned} &\arg\max_{\delta} \mathcal{J}(f_s(x+\delta;\theta), y), \\ &s.t.\ \|\delta\|_\infty \leq \epsilon \wedge f_t(x+\delta) \neq y. \end{aligned} \tag{2}$$

Unfortunately, if the generated adversarial examples overfit the substitute model, it will significantly reduce the success rate of attacks on the target model, since the substitute model is usually difficult to align with the target model. Therefore, the key point is to avoid this overfitting in transfer-based attacks.

### B. Discrete Fourier Transform

In this paper, we adopt 2D-DFT as a tool to extract deep features in the frequency domain. 2D-DFT is a transformation method that converts an image from the spatial domain to the frequency domain. Converting an image from the spatial domain to the frequency domain enables more intuitive observation and processing of the image, and it is also more conducive to the filtering operation of the frequency domain. Assuming that the input image $x \in \mathbb{R}^{C\times H\times W}$, for each channel $C$, its 2D-DFT $\mathcal{F}(.,.)$ can be expressed as follows:

$$\mathcal{F}(u,v) = \frac{1}{HW}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\mathcal{F}^{-1}(h,w)e^{-j2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)}, \tag{3}$$

where $u = 0, 1, \ldots, H-1$, $v = 0, 1, \ldots, W-1$, $j$ represents an imaginary unit. $H$ and $W$ denote the height and width of the input image, respectively. Remarkably, the spectrum after 2D-DFT is consistent with the dimension of the input image. Accordingly, the 2-dimensional inverse discrete Fourier transform (2D-IDFT) $\mathcal{F}^{-1}(.,.)$ transforms an image from the frequency domain back to the spatial domain:

$$\mathcal{F}^{-1}(h,w) = \sum_{u=0}^{H-1}\sum_{v=0}^{W-1}\mathcal{F}(u,v)e^{j2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)}. \tag{4}$$

By employing 2D-DFT and a mask filter, we can thus derive the low-frequency components (LFC) denoted by $\mathcal{F}_l(x)$ and high-frequency components (HFC) $\mathcal{F}_h(x)$ of the input image:

$$\mathcal{F}_l(x) = \mathcal{F}(u,v) \odot M(r), \tag{5}$$

$$\mathcal{F}_h(x) = \mathcal{F}(u,v) \odot (I - M(r)), \tag{6}$$

where $\odot$ is the Hadamard product, $M(r)$ represents the mask filter, and $I$ denotes the identity matrix. The mask filter $M(r)$ is defined as follows:

$$M(r) = \begin{cases} 1, & if\sqrt{(u-u_o)^2+(v-v_o)^2} \leq r \\ 0, & otherwise \end{cases}, \tag{7}$$

where $(u_o, v_o)$ is the centroid coordinates of the spectrogram and $r \in \left[0, \sqrt{\left(\frac{w}{2}\right)^2+\left(\frac{H}{2}\right)^2}\right]$ is the mask radius.

### C. Mixed-Frequency Inputs

In [22], they propose Admix, a variation of Mixup [31], which modifies the input image $x$ by adding other randomly sampled images $x'$ with a certain strength in the spatial domain but keeps the label unchanged. This increases input diversity by loss-preserving transformation [18], and thus improves transferability while maintaining white-box attack performance. An input transformation $T(x)$ in spatial domain is defined as follows [22]:

$$\begin{aligned} T(x) &= \gamma \cdot x + \tau \cdot x' \\ &= \gamma \cdot (x + \eta \cdot x'), \end{aligned} \tag{8}$$

where $\gamma$ and $\tau$ denote the mixed parameters, $\gamma \in (0, 1]$, $\tau < \gamma$, and $\eta = \frac{\tau}{\gamma}$ controls the strength of mixed other categories. Since $\gamma$ is randomly sampled to ensure loss-preserving transformation [18], we can abbreviate this input transformation $T(x)$ more essentially:

$$T(x) = x + \eta \cdot x'. \tag{9}$$

However, it augments images in the spatial domain, ignoring the frequency domain characteristics of the input image. A widely accepted fact is that naturally trained models would easily fall into the trap of overfitting the image's high-frequency features during the training phase. In [22], they alleviate the overfitting of adversarial examples to the source model by incorporating additional randomly sampled images $x'$ with a certain strength. However, they did not fully consider the side effects of high-frequency features of the origin $x$, which may lead to getting stuck in local maxima during the generation of adversarial examples, potentially resulting in the failure to cross the decision boundary.

Thus, we investigate this input transformation $T(x)$ from the perspective of the frequency domain. According to the linear property of DFT, Eq. (9) can be decomposed as follows:

$$\begin{aligned} T(x) &= \mathcal{F}^{-1}\big(\mathcal{F}(x + \eta \cdot x')\big) \\ &= \mathcal{F}^{-1}\big(\mathcal{F}(x) + \eta \cdot \mathcal{F}(x')\big) \\ &= \mathcal{F}^{-1}\big(\underbrace{\mathcal{F}_l(x) + \mathcal{F}_h(x)}_{\text{Item (i)}} + \underbrace{\eta \cdot \mathcal{F}_l(x') + \eta \cdot \mathcal{F}_h(x')}_{\text{Item (ii)}}\big). \end{aligned} \tag{10}$$

Based on our motivation proposed in Section III, we argue that Admix suffers a potential overfitting pitfall caused by the HFC of the origin $x$ in Item (i). Therefore, we hypothesize that breaking the alignment between high-frequency features and class labels can effectively alleviate the issue of adversarial examples overfitting the substitute model during the iterative process. To this end, we further analyze and optimize input transformation $T(x)$. Specifically, we discard a component in Item (i) which impairs the transferability, *i.e.*, $\mathcal{F}_h(x)$. Meanwhile, we remove the coefficient $\eta$ on $\mathcal{F}_h(x')$ in Item (ii) to amplify the benefits from the HFCs of different images. In addition, the $\eta \cdot \mathcal{F}_l(x')$ in Item (ii) indicates the downsizing of the LFCs by $\eta$, which can be simply achieved by the Gaussian noise since it concentrates most of the energy around the central frequency.

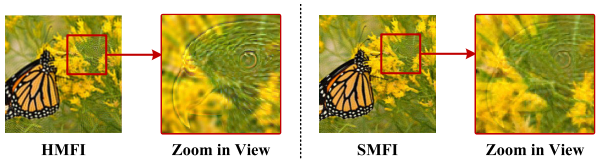



Fig. 2. Local view of MFI. SMFI perturbs each frequency component uniformly through linear interpolation, which is less radical than HMFI. **Left**: Zoom in view of HMFI. **Right**: Zoom in view of SMFI.

From the above analysis, we propose an input transformation method based on the frequency domain, called mixed-frequency inputs (MFI), denoted as $MFI(x)$. MFI can exploit the HFC of images across multiple classes to alleviate overfitting the substitute model further effectively. To convert the HFC, we employ two distinct techniques: direct exchange for HMFI and linear interpolation for SMFI.

*1) Hard Mixed-Frequency Inputs:* HMFI creates an entirely novel image by swiftly swapping the HFC of any two images and reintegrating them with the LFC of the original image. HMFI is defined as follows:

$$HMFI(x) = \mathcal{F}^{-1}\big(\mathcal{F}_l(x) + \mathcal{F}_h(x')\big) + \xi, \tag{11}$$

where $\xi \sim N(0, \sigma^2)$ and the strength of the Gaussian noise perturbation is controlled by $\sigma$.

*2) Soft Mixed-Frequency Inputs:* We suspect that directly replacing HFC may be too aggressive, which will confuse the model classification. Thus, we propose soft mixed-frequency inputs (SMFI), a conservative version of HMFI. As shown in Fig. 2, SMFI linearly interpolates the HFC of the two images and then recombines them with the LFC of the original image. SMFI is defined as follows:

$$\begin{aligned} SMFI(x) = \mathcal{F}^{-1}\Big(&\mathcal{F}_l(x) + (1 - \lambda) \cdot \mathcal{F}_h(x) \\ &+ \lambda \cdot \mathcal{F}_h(x')\Big) + \xi, \end{aligned} \tag{12}$$

where $\lambda \sim U(0, \rho)$ and the parameter $\rho$ controls the strength of linear interpolation of frequency components.

### D. Mixed-Frequency Input Attack

Through the above analysis, we propose the MFI attack method to improve the transferability of adversarial examples. MFI attack contains two parts, a mixup strategy for MFI, and scale transformation, the same as previous work [22]. By its nature, the MFI attack aggregates the gradients of HFC from other classes, alleviating the problem of a single transformation leading to limited diversity while still overfitting alternative models. The MFI attack utilizes the extra image $x'$ from different categories or samples the value of $\gamma$ to compute the average gradient on a set of mixed images of the input $x$ in Eqs. (13) and (14):

$$\overline{g}_{t+1} = \frac{1}{m \cdot n} \sum_{x' \in X'} \sum_{i=0}^{m} \sum_{i=0}^{n-1} \nabla_{x_t^{adv}} \mathcal{J}\Big(\gamma_i \cdot HMFI(x), y; \theta\Big), \tag{13}$$

$$\overline{g}_{t+1} = \frac{1}{m \cdot n} \sum_{x' \in X'} \sum_{i=0}^{m} \sum_{i=0}^{n-1} \nabla_{x_t^{adv}} \mathcal{J}\Big(\gamma_i \cdot SMFI(x), y; \theta\Big), \tag{14}$$

**Algorithm 1** Mixed-Frequency Input Attack

**Require:**
A clean example $x$ with ground-truth $y$, classifier model $f(x;\theta)$, loss function $\mathcal{J}(x, y;\theta)$, maximum perturbation budget $\epsilon$, iteration number $T$, decay factor $\mu$, sampled images $m$, scale copies $n$, hyperparameters $r$, $\sigma$, $\rho$;

**Ensure:**
An adversarial example $x^{adv}$.

$\alpha = \epsilon/T$; $g_0 = 0$; $\overline{g}_0 = 0$; $x_0^{adv} = x$;
**for** $t = 0$ to $T-1$ **do**
  Randomly sample a set $X'$ of $m$ images from another category;
  Apply 2D-DFT to the clean example and selected images by Eq. (3);
  Get LFC and HFC by Eq. (5) and Eq. (6);
  Apply HMFI by Eq. (11) or SMFI by Eq. (12);
  Calculate the average gradient $\overline{g}_{t+1}$ by Eq. (13) or Eq. (14);
  Update the enhanced momentum $g_t$:
    $g_{t+1} \leftarrow \mu \cdot g_t + \frac{\overline{g}_{t+1}}{\|\overline{g}_{t+1}\|_1}$
  Update $x_{t+1}^{adv}$ by applying the gradient sign:
    $x_{t+1}^{adv} \leftarrow Clip_x^{\epsilon}\{x_t^{adv} + \alpha \cdot sign(g_{t+1})\}$
**end for**
**Return** $x^{adv} \leftarrow x_T^{adv}$

where $\mathcal{J}(\cdot)$ denotes cross-entropy loss, $X'$ is the set of $m$ randomly selected images from different categories, and $n$ is the number of MFI mixed images of other categories $x'$. Alternatively, MFI could be combined with input transformation techniques other than SIM and any gradient-based attack. We enumerate the MFI embedded in the MI-FGSM algorithm in Algorithm 1 (referred to as MFI without ambiguity in the subsequent discussion). As shown in Fig. 3, MFI obtains a more reasonable gradient direction in each iteration, finding better local maxima and exhibiting higher transferability.

Here, we make a brief analysis of the time complexity for the input transformation strategy in our proposed MFI (*i.e.*, Steps 3 to 6 in Algorithm 1), since the computation regarding gradient updates for adversarial examples remains consistent with prior work [22]. Assuming that the input image $x \in \mathbb{R}^{C\times H\times W}$ where $C$ represents the number of channels. $H$ and $W$ respectively represent the image's height and width. Let $k = H \times W$. In general, $C$ is a constant, which can be neglected in algorithm analysis. Therefore, the time complexity of HMFI is $\mathcal{T}(k) = \mathcal{O}\big(1 + k\log k + k\big) \approx \mathcal{O}(k\log k)$, and the time complexity of SMFI is $\mathcal{T}(k) = \mathcal{O}\big(1 + k\log k + k + k^{\frac{\log 7}{2}}\big) \approx \mathcal{O}(k^{\frac{\log 7}{2}})$. Here, $\mathcal{O}(k\log k)$ represents the time complexity of the fast Fourier transform (FFT) algorithm used in Step 4 and Step 6, and $\mathcal{O}(k^{\frac{\log 7}{2}})$ represents the time complexity of Strassen matrix multiplication in Step 6 of SMFI.

### E. DFT Versus DCT

Discrete Fourier transform (DFT) and discrete cosine transform (DCT) are widely employed in image signal processing as transformation methods. These techniques enable the conversion of image signals from the spatial domain to the frequency domain. By decomposing the image into sub-bands in the frequency domain, these transforms facilitate the analysis of each sub-band to extract the desired image information. Notably, DCT is a specific form of DFT. In cases where the function being expanded is a real even function, the Fourier series representation exclusively comprises cosine terms. Consequently, discretizing this Fourier series yields DCT. Hence, DCT can be considered a subset of DFT, as both methods entail an analytical transformation of the entire signal.

Natural images exhibit a notable spatial correlation, meaning that adjacent pixels tend to show strong interdependencies. This characteristic implies the presence of redundancy within the input image. DCT effectively captures and encodes this redundancy, resulting in improved compression of image data. Specifically, DCT partitions the image into small patches containing various frequencies, which are subsequently subjected to quantization. During the quantization process, the HFC (imaginary part) is discarded, while the LFC (real part) is retained for image compression and subsequent reconstruction. On the other hand, DFT does not explicitly consider spatial correlation. Instead, it converts the entire image into a frequency domain representation, preserving the original frequency domain information of the image in its entirety.

While DCT offers advantages in terms of extraction efficiency and computational speed compared to DFT, our research focuses on the extraction of high-frequency information from images. Consequently, for our proposed method, we prioritize the attainment of comprehensive and precise frequency domain data, making DFT a more suitable choice.

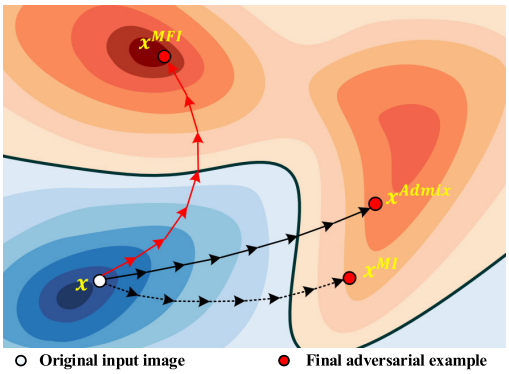



Fig. 3. Illustration of our proposed MFI attack (indicated by the solid red arrow), the Admix attack (indicated by the solid black arrow), and the traditional momentum-based iterative attack (indicated by the dashed black arrow). The MFI attack mainly modifies and stabilizes the gradient direction by aggregating the feature of the HFC from other images during each iteration.

## V. Experiments

### A. Experimental Setup

*1) Dataset:* Following previous works [17], [19], [26], we conduct our experiments on the ImageNet-compatible dataset, which contains 1000 images provided by the NIPS 2017 adversarial attack competition [34]. Each image in the dataset has an officially specified ground truth for a fair comparison.

*2) Models:* We treat both the normally trained models and defended models as the target models. For normally trained models, we select six CNN-based models with different architectures, containing Inception-v3 (Inc-v3) [43], Inception-v4 (Inc-v4) [44], Inception-Resnet-v2 (IncRes-v2) [44], and Resnet-v2-50 (Res-50), Resnet-v2-101 (Res-101) and Resnet-v2-152 (Res-152) [8], [45], as well as four transformer-based models, *i.e.*, ViT [46], PiT [47], Visformer [48], Swin [49]. For the defense models, we consider the adversarially trained models and the advanced defense models. We chose four adversarially trained models, containing individual adversarial trained model as well as ensemble adversarial training from multiple models [15], [50]. The selected defense models include Inc-v3$_{adv}$, Inc-v3$_{ens3}$, Inc-v3$_{ens4}$, and IncRes-v2$_{ens}$. In addition, we include five advanced defense models that are robust enough and resistant to black-box attacks on the ImageNet dataset. These defenses include R&P [42], Bit-Red [51], JPEG [41], FD [52], ComDefend [53], HGD [54], and NRP [55]. For R&P, we adopt IncRes-v2$_{ens}$ as the target model. For all other defense methods, we adopt Inc-v3$_{ens3}$ as the target model.

*3) Baselines:* To verify the effectiveness of our proposed MFI, we compare it with various attack methods. These include MI-FGSM [26], NI-FGSM [18], SI-FGSM [18], PI-FGSM [19], VMI-FGSM [27], Admix [22], S$^2$I-FGSM [23], and GRA [35].

*4) Parameter & Attack Settings:* In all experiments, the maximum perturbation $\epsilon = 16$, the iteration T $= 10$, and the step size $\alpha = \epsilon/T = 1.6$. For MI-FGSM [26] and NI-FGSM [18], we set the decay factor $\mu = 1.0$. For SI-FGSM [18], we set the number of scaled versions $n = 5$. For DI-FGSM [25], we set the transformation probability $p = 0.5$. For TI-FGSM [17], we deploy the $7 \times 7$ Gaussian kernels. For PI-FGSM [19], we set the amplification factor $\beta_1 = 10$, the project factor $\gamma = 16$, and the kernel length $k_w = 3$ for normally trained models, $k_w = 7$ for defense models. For VT-FGSM [27], we set the hyper-parameter $\beta_2 = 1.5$, number of sampling examples is 20. For Admix [22], we set the number of copies $m_1 = 5$, the sample number $m_2 = 3$, and the admix ratio $\eta = 0.2$. For S$^2$I-FGSM [23], we set the tuning factor $p = 0.5$, $\sigma$ is equal to the value of $\epsilon$, and the number of spectrum transformations $N = 20$. For GRA [35], we set the sample quantity $\mathcal{M} = 20$, the upper bound factor of sample range $\beta_3 = 3.5$. For our proposed MFI, we follow the settings provided by Admix [22] for a fair comparison and set the mask filter radius $r = 100$, the strength of Gaussian noise perturbation $\sigma = 32$, and the strength of linear interpolation $\rho = 1$. The parameter settings for the combined version are the same. All experiments in this paper were conducted with the NVIDIA® RTX 2080TI GPU. Fig. 4 illustrates the nine pairs of randomly selected original images and their corresponding adversarial examples crafted by HMFI. These perturbations of adversarial examples are virtually imperceptible to humans.

*5) Metrics:* We evaluate the performance of attack methods by the attack success rate (ASR) of the target model as follows:

$$ASR = \frac{n_{f_t(x+\delta)\neq y}}{n_{f_s(x+\delta)\neq y}} \times 100\%, \tag{15}$$


Fig. 4. Visualization of randomly selected original images and their corresponding adversarial examples. In each set, the original image is shown on the left and the adversarial example is shown on the right. All these adversarial examples are crafted by our proposed method HMFI on Inc-v3 with the maximum perturbation of each pixel $\epsilon = 16$.

where $n_{f_t(x+\delta)\neq y}$ represents the number of misclassified adversarial examples by the target model and $n_{f_s(x+\delta)\neq y}$ represents the total number of adversarial examples crafted by the substitute model.

### B. Attacks on Normally Trained Models

*1) Attacks on CNN-Based Models:* We initiate this by subjecting a single neural network to advanced adversarial attacks, thereby gauging the efficacy of these attacks across the six normally trained models under consideration. Table I illustrates the success rate of these attacks, indicating the misclassification rate observed on each model when exposed to adversarial examples. In the first column of Table I, we denote the source models utilized during the crafting of adversarial examples: specifically, Inc-v3, Inc-v4, IncRes-v2, and Res-101. Meanwhile, the first row showcases the six target models employed in emulating the black-box architecture. Additionally, the second column delineates the prevailing advanced baseline attacks alongside our proposed method.

A first glance shows that our proposed HMFI and SMFI outperform the well-known baseline attacks on all black-box models. To illustrate, when targeting Inc-v3 as the source model, MI-FGSM, SI-FGSM, and PI-FGSM achieve a mere 40.3%, 38.2%, and 44.9% success rate, respectively, in transferring adversarial examples to ResNet-50. In stark contrast, our HMFI and SMFI methods notably enhance ASRs to 80.7% and 82.4%, respectively. This convincingly validates the efficiency of our proposed methods relative to normally trained models and helps to gain more insight into the vulnerability of the models. Considering the practical advantages of the black-box attack task, we assert that the marginal decline in white-box performance remains well within an acceptable range. Even though our proposed method might lead to partial alterations to the original image and a slight reduction in white-box performance, the practical gains in black-box attack efficiency validate this compromise.

*2) Attacks on Transformer-Based Models:* The Transformer demonstrates impressive performance in the field of image classification. Due to its unique architecture, exhibiting less

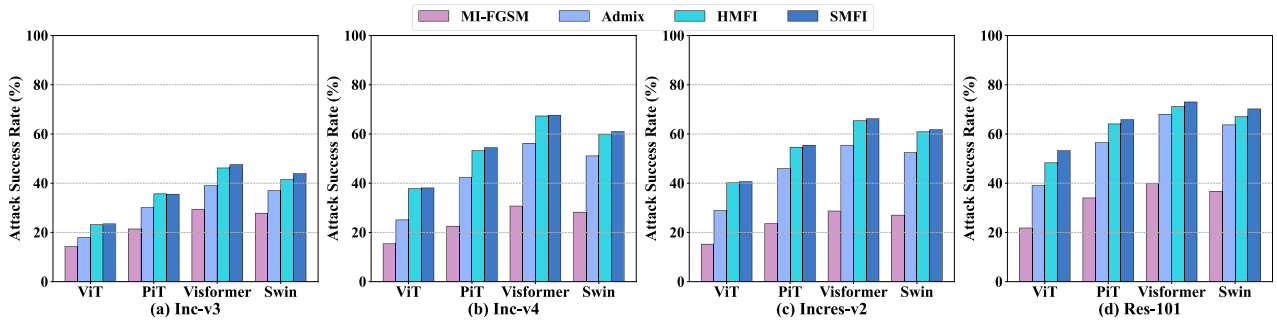



Fig. 5. ASRs (%) on four transformer-based models. The adversarial examples are crafted via Inc-v3, Inc-v4, IncRes-v2, and Res-101, respectively.

TABLE I

ASRs (%) ON SIX NORMALLY TRAINED MODELS. THE ADVERSARIAL EXAMPLES ARE CRAFTED VIA INC-V3, INC-V4, INCRES-V2, AND RES-101, RESPECTIVELY. "*" INDICATES WHITE-BOX ATTACKS. THE BEST RESULTS ARE SHOWN IN BOLD

| Model | Attack | Inc-v3 | Inc-v4 | IncRes-v2 | Res-50 | Res-101 | Res-152 | Average |
|---|---|---|---|---|---|---|---|---|
| Inc-v3 | MI-FGSM [26] | **100.0*** | 45.8 | 41.0 | 40.3 | 36.6 | 34.2 | 49.7 |
| | NI-FGSM [18] | **100.0*** | 51.5 | 50.7 | 46.9 | 41.4 | 39.4 | 55.0 |
| | SI-FGSM [18] | **100.0*** | 42.6 | 35.5 | 38.2 | 31.6 | 30.3 | 46.4 |
| | PI-FGSM [19] | **100.0*** | 52.0 | 43.3 | 44.9 | 41.5 | 39.9 | 53.6 |
| | VMI-FGSM [27] | 97.9* | 69.8 | 65.2 | 60.4 | 56.9 | 56.8 | 67.8 |
| | Admix [22] | 98.1* | 75.9 | 74.5 | 71.8 | 68.4 | 65.5 | 75.7 |
| | $S^2$I-FGSM [23] | 99.7* | 63.6 | 59.6 | 57.1 | 53.8 | 51.0 | 64.1 |
| | GRA [35] | 98.4* | 85.4 | 84.2 | 79.5 | 78.5 | 76.5 | 83.8 |
| | HMFI (Ours) | 98.6* | 84.8 | 84.6 | 80.7 | 79.1 | 80.1 | 84.7 |
| | SMFI (Ours) | 98.4* | **86.5** | **84.7** | **82.4** | **80.6** | **80.3** | **85.5** |
| Inc-v4 | MI-FGSM [26] | 61.7 | **99.9*** | 46.9 | 45.2 | 43.5 | 43.0 | 56.7 |
| | NI-FGSM [18] | 67.7 | **99.9*** | 55.2 | 52.1 | 47.4 | 45.0 | 61.2 |
| | SI-FGSM [18] | 59.2 | **99.9*** | 44.0 | 43.3 | 37.9 | 39.5 | 54.0 |
| | PI-FGSM [19] | 60.8 | **99.9*** | 45.8 | 48.6 | 44.2 | 43.5 | 57.1 |
| | VMI-FGSM [27] | 79.5 | 98.4* | 70.7 | 65.1 | 63.6 | 63.2 | 73.4 |
| | Admix [22] | 85.9 | 98.6* | 78.6 | 75.8 | 72.5 | 72.2 | 80.6 |
| | $S^2$I-FGSM [23] | 71.6 | 99.6* | 54.1 | 55.7 | 47.2 | 48.6 | 62.8 |
| | GRA [35] | **89.1** | 97.8* | 83.4 | 80.7 | **79.2** | 80.5 | 85.1 |
| | HMFI (Ours) | 87.4 | 97.5* | 83.2 | 79.8 | 78.3 | 79.1 | 84.2 |
| | SMFI (Ours) | 88.1 | 97.4* | **84.7** | **81.2** | 78.3 | **81.7** | **85.2** |
| IncRes-v2 | MI-FGSM [26] | 60.3 | 53.0 | 99.3* | 50.0 | 46.2 | 44.9 | 59.0 |
| | NI-FGSM [18] | 66.9 | 56.8 | 99.6* | 52.0 | 47.9 | 45.3 | 61.4 |
| | SI-FGSM [18] | 59.0 | 49.7 | **99.8*** | 45.9 | 41.3 | 40.7 | 56.1 |
| | PI-FGSM [19] | 64.8 | 57.6 | **99.8*** | 52.9 | 48.2 | 47.3 | 61.8 |
| | VMI-FGSM [27] | 80.7 | 77.2 | 99.3* | 69.7 | 69.7 | 65.6 | 77.0 |
| | Admix [22] | 89.3 | 86.5 | 99.7* | 84.7 | 82.5 | 80.7 | 87.2 |
| | $S^2$I-FGSM [23] | 76.4 | 69.1 | 98.3* | 60.9 | 58.7 | 56.6 | 70.0 |
| | GRA [35] | 87.8 | 86.5 | 98.0* | 82.9 | 81.5 | 82.6 | 86.6 |
| | HMFI (Ours) | 89.5 | 86.8 | 97.1* | 84.5 | 84.1 | 83.3 | 87.6 |
| | SMFI (Ours) | **89.7** | **87.3** | 97.4* | **85.3** | **84.6** | **84.3** | **88.1** |
| Res-101 | MI-FGSM [26] | 60.5 | 53.0 | 50.6 | 90.0 | 99.6* | 87.6 | 73.6 |
| | NI-FGSM [18] | 67.2 | 59.7 | 56.8 | 92.4 | 99.6* | 91.7 | 77.9 |
| | SI-FGSM [18] | 48.7 | 41.5 | 36.5 | 89.6 | **99.9*** | 86.2 | 67.1 |
| | PI-FGSM [19] | 65.3 | 56.0 | 51.5 | 86.7 | 99.6* | 84.6 | 74.0 |
| | VMI-FGSM [27] | 75.1 | 69.5 | 68.4 | 93.3 | 97.6* | 92.5 | 82.7 |
| | Admix [22] | 76.7 | 71.4 | 71.0 | **95.7** | 98.1* | 94.5 | 84.6 |
| | $S^2$I-FGSM [23] | 71.0 | 62.1 | 60.8 | 95.0 | 99.7* | 94.9 | 80.6 |
| | GRA [35] | **85.5** | 80.6 | **82.1** | 95.1 | 98.2* | 95.1 | **89.4** |
| | HMFI (Ours) | 85.2 | 79.5 | 79.8 | 95.2 | 98.1* | 95.0 | 88.8 |
| | SMFI (Ours) | 85.0 | **80.8** | 80.2 | 95.0 | 98.2* | **95.3** | 89.1 |

TABLE II

ASRs (%) ON FOUR ADVERSARIALLY TRAINED MODELS. THE ADVERSARIAL EXAMPLES ARE CRAFTED VIA INC-V3, INC-V4, INCRES-V2, AND RES-101, RESPECTIVELY. THE BEST RESULTS ARE SHOWN IN BOLD

| Model | Attack | Inc-v3$_{adv}$ | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes-v2$_{ens}$ | Average |
|---|---|---|---|---|---|---|
| Inc-v3 | MI-FGSM [26] | 24.4 | 19.3 | 15.9 | 8.2 | 17.0 |
| | NI-FGSM [18] | 24.6 | 18.6 | 16.3 | 8.9 | 17.1 |
| | SI-FGSM [18] | 18.3 | 17.3 | 17.8 | 8.4 | 15.5 |
| | PI-FGSM [19] | 36.1 | 35.6 | 37.0 | 27.3 | 34.0 |
| | VMI-FGSM [27] | 39.4 | 34.7 | 35.8 | 21.3 | 32.8 |
| | Admix [22] | 41.6 | 38.2 | 38.3 | 20.9 | 34.8 |
| | $S^2$I-MI-FGSM [23] | 62.1 | 58.2 | 58.0 | 36.6 | 53.7 |
| | GRA [35] | 65.5 | 58.5 | 57.6 | 38.4 | 55.0 |
| | HMFI (Ours) | **69.9** | **63.5** | 62.0 | 40.5 | **59.0** |
| | SMFI (Ours) | 69.2 | 62.2 | **62.6** | **41.7** | 58.9 |
| Inc-v4 | MI-FGSM [26] | 24.6 | 19.5 | 18.6 | 11.3 | 18.5 |
| | NI-FGSM [18] | 25.0 | 18.3 | 18.4 | 10.2 | 18.0 |
| | SI-FGSM [18] | 19.2 | 18.2 | 21.7 | 12.0 | 17.8 |
| | PI-FGSM [19] | 36.6 | 36.5 | 37.3 | 28.4 | 34.7 |
| | VMI-FGSM [27] | 39.7 | 40.8 | 40.3 | 25.3 | 36.5 |
| | Admix [22] | 48.9 | 50.3 | 46.8 | 31.1 | 44.3 |
| | $S^2$I-MI-FGSM [23] | 58.7 | 57.3 | 55.6 | 35.9 | 51.9 |
| | GRA [35] | 65.4 | 66.9 | 64.2 | 51.9 | 62.1 |
| | HMFI (Ours) | **68.1** | **70.5** | **68.8** | 52.6 | **65.0** |
| | SMFI (Ours) | 67.6 | 69.5 | 68.6 | **53.2** | 64.7 |
| IncRes-v2 | MI-FGSM [26] | 27.4 | 21.2 | 21.5 | 14.0 | 21.0 |
| | NI-FGSM [18] | 26.7 | 19.7 | 19.3 | 12.6 | 19.6 |
| | SI-FGSM [18] | 23.5 | 25.2 | 23.2 | 17.5 | 22.4 |
| | PI-FGSM [19] | 40.6 | 39.5 | 41.0 | 35.9 | 39.3 |
| | VMI-FGSM [27] | 45.9 | 48.7 | 44.6 | 38.2 | 44.4 |
| | Admix [22] | 62.1 | 64.6 | 55.3 | 49.2 | 57.8 |
| | $S^2$I-MI-FGSM [23] | 69.8 | 69.2 | 62.4 | 55.9 | 64.3 |
| | GRA [35] | 71.5 | 71.7 | 67.8 | 66.1 | 69.3 |
| | HMFI (Ours) | 77.1 | 77.0 | 73.1 | 66.7 | 73.5 |
| | SMFI (Ours) | **77.2** | **77.6** | **73.7** | **67.4** | **74.0** |
| Res-101 | MI-FGSM [26] | 30.6 | 30.4 | 25.9 | 16.6 | 25.9 |
| | NI-FGSM [18] | 32.3 | 29.3 | 27.2 | 16.5 | 26.3 |
| | SI-FGSM [18] | 20.4 | 24.6 | 23.5 | 14.6 | 20.8 |
| | PI-FGSM [19] | 43.5 | 43.0 | 44.1 | 36.6 | 41.8 |
| | VMI-FGSM [27] | 45.6 | 47.6 | 44.2 | 32.5 | 42.5 |
| | Admix [22] | 45.6 | 45.1 | 39.8 | 28.7 | 39.8 |
| | $S^2$I-MI-FGSM [23] | 69.6 | 67.8 | 63.2 | 48.3 | 62.2 |
| | GRA [35] | **70.8** | 70.5 | **67.8** | **57.4** | **66.6** |
| | HMFI (Ours) | 69.7 | 71.7 | 67.2 | 56.2 | 66.2 |
| | SMFI (Ours) | 70.1 | **72.2** | 67.4 | 56.0 | 66.4 |

bias towards local texture features. Hence, we conduct additional extensive experiments on transformer-based models. The results are summarized in Fig. 5. Both representations of MFI exhibit superior performance on the transformer architecture compared to Admix.

### C. Attacks on Adversarial Trained Models

Adversarial training [3] could alleviate the vulnerability of the model to a certain extent, and the model after adversarial training is more robust than the normally trained model. Thus, we evaluated the effectiveness of the attack on four adversarial training models. Through Table II, we can observe that HMFI and SMFI substantially lead the extensive baseline attack methods on all adversarial training models.

In addition, several input transformation techniques have been previously proposed in the literature [17], [18], [25], which are integrated into gradient-based attack methods, such as SI-FGSM [18], DI-FGSM [25], TI-FGSM [17]. These methods have demonstrated considerable enhancements in the transferability of adversarial examples. We consolidate these input transformation methodologies into our attack approach as displayed in Table III.

### D. Attacks on Ensemble of Models

According to prior research by Dong et al. [26], attacking several source models at once and calculating the average of the logit outputs is a workable strategy to increase the adversarial transferability. It integrates the vulnerabilities of different architectural models by combining the comprehensive outputs of all source models. To be more specific, we select source models that align with the criteria outlined in the previous experiments and average the logit outputs of these

TABLE III

ASRs (%) COMPATIBLE WITH VARIOUS INPUT TRANSFORMATION TECHNIQUES. THE ADVERSARIAL EXAMPLES ARE CRAFTED VIA INC-V3, INC-V4, INCRES-V2, AND RES-101, RESPECTIVELY. THE BEST RESULTS ARE SHOWN IN BOLD

| Model | Attack | Inc-v3$_{adv}$ | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes-v2$_{ens}$ | Average |
|---|---|---|---|---|---|---|
| Inc-v3 | MI-SI-DI-TI-FGSM [26] | 65.4 | 65.5 | 64.9 | 47.8 | 60.9 |
| | Admix-DI-TI [22] | 68.9 | 68.2 | 66.8 | 49.6 | 63.4 |
| | HMFI-DI-TI (Ours) | 80.3 | 78.1 | 77.8 | 65.6 | 75.5 |
| | SMFI-DI-TI (Ours) | **80.7** | **79.9** | **78.2** | **66.3** | **76.3** |
| Inc-v4 | MI-SI-DI-TI-FGSM [26] | 65.8 | 68.9 | 66.5 | 56.2 | 64.4 |
| | Admix-DI-TI [22] | 70.4 | 72.7 | 69.8 | 58.6 | 67.9 |
| | HMFI-DI-TI (Ours) | **77.0** | 78.1 | 76.9 | **69.6** | 75.4 |
| | SMFI-DI-TI (Ours) | 76.7 | **79.2** | **77.9** | 69.4 | **75.8** |
| IncRes-v2 | MI-SI-DI-TI-FGSM [26] | 77.1 | 79.7 | 75.4 | 73.2 | 76.4 |
| | Admix-DI-TI [22] | 78.4 | 81.2 | 77.7 | 73.0 | 77.6 |
| | HMFI-DI-TI (Ours) | **84.2** | **85.3** | 81.9 | 79.8 | 82.8 |
| | SMFI-DI-TI (Ours) | 83.9 | **85.3** | **82.2** | **80.9** | **83.1** |
| Res-101 | MI-SI-DI-TI-FGSM [26] | 71.4 | 72.8 | 68.9 | 61.1 | 68.6 |
| | Admix-DI-TI [22] | 73.6 | 74.8 | 72.1 | 62.1 | 70.7 |
| | HMFI-DI-TI (Ours) | 80.4 | 80.1 | 77.5 | 71.4 | 77.4 |
| | SMFI-DI-TI (Ours) | **81.2** | **81.5** | **79.1** | **72.7** | **78.6** |

TABLE IV

ASRs (%) OF SEVEN MODELS UNDER ENSEMBLE ATTACK. THE ADVERSARIAL EXAMPLES ARE CRAFTED VIA AN ENSEMBLE OF INC-V3, INC-V4, INCRES-V2, AND RES-101. THE WEIGHT FOR EACH MODEL IS 1/4. "*" INDICATES WHITE-BOX ATTACKS. THE BEST RESULTS ARE SHOWN IN BOLD

| Attack | Inc-v3 | Inc-v4 | IncRes-v2 | Res-101 | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes-v2$_{ens}$ | Average |
|---|---|---|---|---|---|---|---|---|
| MI-FGSM [26] | 99.8* | 99.2* | 97.5* | **100.0*** | 43.6 | 43.2 | 27.4 | 73.0 |
| Admix [22] | **99.9*** | **99.4*** | **99.1*** | **100.0*** | 84.7 | 80.8 | 65.1 | 89.9 |
| HMFI (Ours) | 98.7* | 98.1* | 97.5* | 99.6* | 91.4 | 90.5 | 86.2 | 94.6 |
| SMFI (Ours) | 98.8* | 98.0* | 97.5* | 99.4* | **92.0** | **90.9** | **86.9** | **94.8** |
| MI-SI-DI-TI-FGSM [26] | **99.5*** | **99.3*** | **98.8*** | **99.8*** | 92.8 | 92.3 | 88.7 | 95.9 |
| Admix-DI-TI [22] | **99.5*** | 99.2* | **98.8*** | 99.6* | 93.7 | 93.2 | 90.5 | 96.4 |
| HMFI-DI-TI (Ours) | 99.1* | 98.9* | 98.7* | 99.1* | 94.3 | **94.4** | **92.3** | **96.7** |
| SMFI-DI-TI (Ours) | 99.1* | 98.7* | 98.3* | 99.2* | **94.8** | 94.2 | 91.7 | 96.6 |

models in each iteration. In this context, we assume that all attacks directed at the source models within the ensemble are considered white-box attacks. All the results are presented in Table IV.

### E. Attacks on Advanced Defense Models

While current attack methods could easily fool normally trained models, they may be helpless against models with defense mechanisms. Therefore, We consider models in the more challenging black-box scenario, which often improve their robustness with some kind of defense. To further verify the superiority of our method, we evaluate our method on models with advanced defenses, including R&P [42], Bit-Red [51], JPEG [41], FD [52], ComDefend [53], HGD [54], and NRP [55]. We choose MI-FGSM [26] and Admix [22] as the baseline attack methods, both of which could be compatible with a variety of input transformation techniques. Table V presents the results of our comprehensive experiments with single model attacks and ensemble attacks, respectively.

*1) Single-Model Attacks:* We examine the performance of adversarial examples generated by a single source model when transferred to the defense models. In Table V, the results demonstrate the substantial improvements achieved by our MFI, over existing attack techniques. For instance, the attack approach MI-SI-DI-TI-FGSM achieves an average ASR of 59.7% against the defense model. In contrast, our proposed HMFI achieves an average ASR of 73.5%, while SMFI achieves an average ASR of 74.5%. These results indicate a significant enhancement in transferability, with improvements of 13.8% and 14.8% achieved by HMFI and SMFI, respectively, compared to the baseline attack MI-SI-DI-TI-FGSM.

TABLE V

ASRs (%) ON ADVANCED DEFENSE MODELS. THE ADVERSARIAL EXAMPLES ARE CRAFTED VIA AN ENSEMBLE OF INC-V3, INC-V4, INCRES-V2, AND RES-101. THE WEIGHT FOR EACH MODEL IS 1/4. THE BEST RESULTS ARE SHOWN IN BOLD

| Model | Attack | JPEG-50 | JPEG-75 | R&P | Bit-Red | FD | ComDefend | HGD | NRP | Average |
|---|---|---|---|---|---|---|---|---|---|---|
| Inc-v3 | MI-SI-DI-TI-FGSM [26] | 74.7 | 74.0 | 49.4 | 48.7 | 71.5 | 69.0 | 55.1 | 35.4 | 59.7 |
| | Admix-DI-TI [22] | 79.5 | 78.9 | 52.3 | 52.3 | 73.4 | 73.7 | 62.6 | 41.7 | 64.3 |
| | HMFI-DI-TI (Ours) | 84.1 | 82.3 | 65.4 | 66.9 | 81.3 | 79.5 | 65.4 | **62.8** | 73.5 |
| | SMFI-DI-TI (Ours) | **84.9** | **83.5** | **67.2** | **68.6** | **82.1** | **80.5** | **67.0** | 62.5 | **74.5** |
| Ens | MI-SI-DI-TI-FGSM [26] | 94.2 | 94.8 | 90.0 | 78.8 | 90.0 | 93.0 | 92.8 | 71.7 | 88.2 |
| | Admix-DI-TI [22] | 94.9 | 96.0 | 91.5 | 80.1 | 91.6 | 94.4 | 94.1 | 75.4 | 89.8 |
| | HMFI-DI-TI (Ours) | 95.6 | 95.8 | 92.4 | 84.8 | 93.4 | 94.6 | 95.4 | 85.5 | 92.2 |
| | SMFI-DI-TI (Ours) | **96.2** | **96.3** | **92.5** | **85.6** | **93.6** | **94.9** | **95.6** | **86.7** | **92.7** |

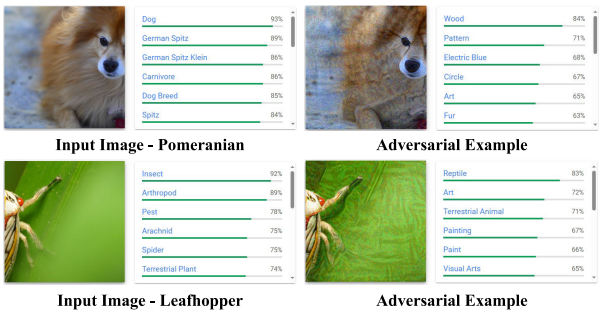



Fig. 6. The feedback of attacking the Google Cloud visual interface. All these adversarial examples are crafted by our proposed method HMFI on Inc-v3 with the maximum perturbation of each pixel $\epsilon = 16$.

These notable advancements validate the effectiveness of our proposed method when applied to defense models.

*2) Ensemble-Based Attacks:* We extend our experiments to evaluate the effectiveness of our method in attacking ensemble-based models. Specifically, we generate adversarial examples by leveraging an ensemble of Inc-v3, Inc-v4, IncRes-v2, and Res-101 models. Similar to the single-model attack, our SMFI-DI-TI, as shown in Table V, consistently demonstrates superior performance on the ensemble model. Both HMFI-DI-TI and SMFI-DI-TI consistently outperform the baseline attacks, achieving an average ASR of 92.2% and 92.7%, respectively. These results highlight the shortcomings of current defense models in terms of adversarial robustness, indicating that they do not meet the requirements for genuine security in real-world black-box scenarios.

### F. Attacks on Real-World Image Recognition System

Furthermore, we extended the application of MFI to real-world scenarios to substantiate its practical significance. In practical settings, we executed attacks on the Google Cloud Visual Interface and assessed our adversarial examples against it. The Google Cloud Vision system predicts a series of labels alongside corresponding confidence scores, exclusively presenting label annotations with confidence values exceeding 50%. This scenario operates in a fully black-box manner, where access to gradients and underlying system parameters is unavailable. Specifically, following the settings from prior experiments, we generated adversarial examples and inputted them into the interface to retrieve labels along with their associated confidence values. As depicted in Fig. 6, our proposed MFI demonstrated its effectiveness by adeptly altering the Top-K predicted labels within the Google Cloud Visual Interface, thereby achieving a successful attack.

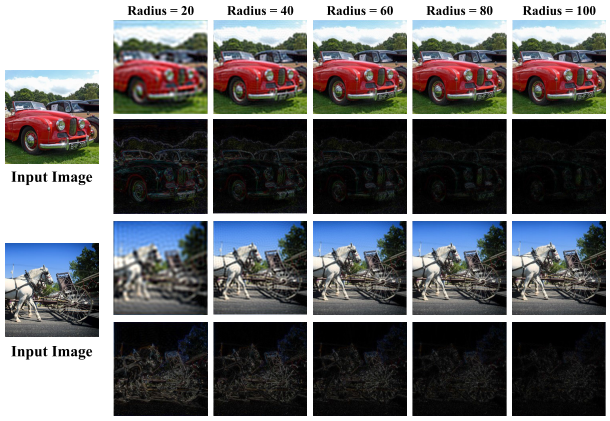



Fig. 7. Visualization of high-frequency and low-frequency reconstructed images under different mask radius $r$. Input images (left). LFC reconstructed images (1st and 3rd rows). HFC reconstructed images (2nd and 4th rows).

TABLE VI

ASRs (%) ON SELECTED SEVEN MODELS UNDER DIFFERENT MASK RADIUS $r$. THE ADVERSARIAL EXAMPLES ARE CRAFTED BY HMFI VIA INC-V3. "*" INDICATES WHITE-BOX ATTACKS. THE BEST RESULTS ARE SHOWN IN BOLD

| Mask Radius | Inc-v3 | Inc-v4 | IncRes-v2 | Res-101 | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes-v2$_{ens}$ |
|---|---|---|---|---|---|---|---|
| $r = 0$ | 100.0* | 42.6 | 35.5 | 31.6 | 17.3 | 17.8 | 8.4 |
| $r = 20$ | 78.7* | 41.3 | 35.2 | 34.6 | 19.8 | 18.9 | 10.3 |
| $r = 40$ | 90.1* | 49.3 | 43.5 | 41.1 | 25.7 | 23.9 | 13.1 |
| $r = 60$ | 99.1* | 66.9 | 63.4 | 58.0 | 35.1 | 35.5 | 19.4 |
| $r = 80$ | **99.3*** | 75.2 | 72.9 | 66.3 | **42.0** | 39.4 | 21.9 |
| $r = 100$ | 98.6* | **76.8** | **73.7** | **67.9** | 41.4 | **40.5** | **23.6** |
| $r = 120$ | 98.4* | 75.0 | 69.9 | 65.4 | 39.8 | 38.5 | 22.3 |
| $r = 140$ | 97.9* | 69.4 | 67.5 | 61.2 | 37.4 | 36.4 | 20.8 |

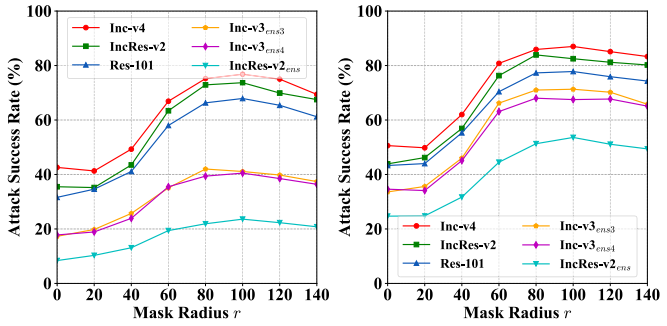



Fig. 8. ASRs (%) on the selected six models with adversarial examples generated by HMFI and HMFI-DI-TI on Inc-v3 under different mask radius $r$. **Left**: HMFI. **Right**: HMFI-DI-TI.

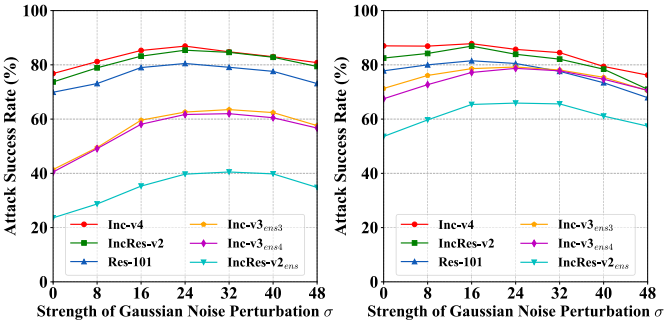



Fig. 9. ASRs (%) on the selected six models with adversarial examples generated by HMFI and HMFI-DI-TI on Inc-v3 under different strengths of Gaussian noise perturbation $\sigma$. **Left**: HMFI. **Right**: HMFI-DI-TI.

### G. Ablation Study

To investigate the effect of hyperparameters on our experiments, we first conduct a fairness setting as detailed below. For hyperparameters $m$ and $n$, we follow the setting of previous work [22] and choose $m = 3$ and $n = 5$. Here we conduct a series of ablation experiments to investigate the effect of the other two hyperparameters $r$ and $\xi$ used by HMFI and SMFI in the experiments. In addition, a separate experimental analysis is performed for the hyperparameter $\lambda$ in SMFI.

*1) Effect of Mask Radius:* The hyperparameter $r$ represents the size of the binary mask in Eq. (7), which controls the scale at which the LFC of the image is retained. We present the corresponding demonstration in Fig. 7. The range of the image to retain the LFC increases with increasing mask radius $r$, while the HFC of the other mixture categories decreases. In Table VI we report the attack success rate of HMFI for the selected seven models under different mask radii, where the strength of Gaussian noise perturbation $\sigma$ is set to 0. After reaching a peak at $r = 80$, the white-box ASR progressively drops. In contrast, practically all black-box models reach their peak performance when $r = 100$. We can see the shifting trend in Fig. 8. While there is a minor decrease in the white-box scenario, it is quite acceptable considering our objective is to increase black-box transferability. Thus, we choose $r = 100$ in our experiments.

*2) Effect of Gaussian Noise Perturbation:* We then study the impact of the strength of Gaussian noise perturbation $\sigma$ on the ASR in the black-box setting. By observing Fig. 9, we find that the strength of Gaussian noise perturbation does not perform the same on the normally trained models as on the robust models for either HMFI or HMFI-DI-TI. For HMFI, the transferability of adversarial examples on the normally trained model achieves the peak at $\sigma = 24$, and then gradually decreases. Conversely, $\sigma = 32$ yields the best results with the robust model. For HMFI-DI-TI, the transferability of adversarial examples on all models peaks at $\sigma = 16$ or $\sigma = 24$. Obviously, the HMFI compatible with DIM [25] and TIM [17] depends less on the strength of Gaussian noise. Since our ultimate goal is to improve the transferability of adversarial examples in black-box scenarios, the size of $\sigma$ does not increase the computational cost. Consequently, to better align with the realistic black-box robust scenario, we set $\sigma = 32$ in our experiments.

*3) Effect of Various Components:* As shown in Tab. VII, various components within MFI contribute to enhancing the transferability of adversarial examples. We observed that $\mathcal{F}_h(x')$ significantly improves the transferability to naturally trained models, while Gaussian noise $\xi$ exhibits notable enhancement for robust models. We suspect that such a significant difference is caused by the target model's frequency bias during the training phase. Naturally trained models tend to favor local features, whereas robust models exhibit the opposite preference, leaning towards global features. When all these components are combined, MFI achieves optimal performance, validating our rationale design.

*4) Effect of Linear Interpolation:* We conduct independent experimental analysis for the hyperparameter $\rho$ in SMFI. As depicted in Fig. 10, we discover that varying the strength of linear interpolation does not result in noticeable modifications. Across various models, their performance varies only slightly.

TABLE VII

EFFECT OF VARIOUS COMPONENTS IN MFI. ASRS (%) ON THE SELECTED SIX MODELS WITH ADVERSARIAL EXAMPLES GENERATED BY HMFI ON INC-V3. THE BEST RESULTS ARE SHOWN IN BOLD

| $\mathcal{F}_l(x)$ | $\mathcal{F}_h(x')$ | $\xi$ | Inc-v4 | IncRes-v2 | Res-101 | Inc-v3$_{ens3}$ | Inc-v3$_{ens4}$ | IncRes-v2$_{ens}$ | Average |
|---|---|---|---|---|---|---|---|---|---|
| ✔ | ✘ | ✘ | 64.0 | 61.4 | 56.1 | 34.3 | 35.3 | 18.9 | 45.0 |
| ✔ | ✔ | ✘ | 76.8 | 73.7 | 69.9 | 41.4 | 40.5 | 23.6 | 54.3 |
| ✔ | ✔ | ✔ | **84.8** | **84.6** | **79.1** | **63.5** | **62.0** | **40.5** | **69.0** |

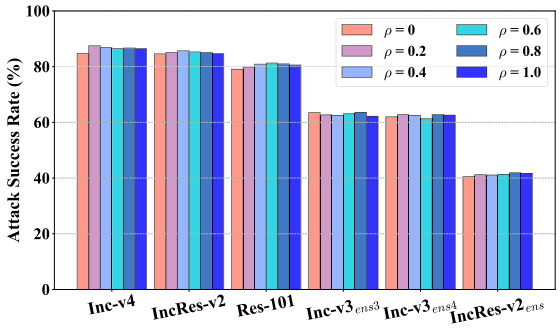



Fig. 10. Effect of the strength of linear interpolation $\rho$. ASRs (%) on the selected six models with adversarial examples generated by SMFI on Inc-v3 under different strengths of linear interpolation $\rho$.

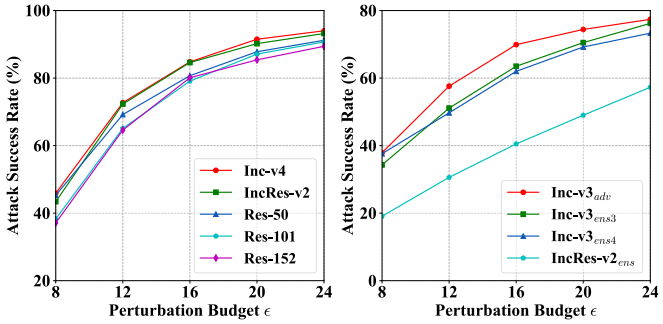



Fig. 11. ASRs (%) on the different models with adversarial examples generated by HMFI on Inc-v3 under different perturbation budgets ($8 \leq \epsilon \leq 24$). **Left**: Normally trained models. **Right**: Adversarial trained models.

This may be related to the frequency bias of the model during the training phase [33], [39], [40], which may explain the fluctuation of the ASRs of HMFI and SMFI in our experiment. In addition, the latest literature [56] shows that each frequency component of the image has a promoting or inhibiting effect on the confidence output of the model, which may be one of the explanations for this phenomenon. We set $\rho = 1$ in our experiments to seek a more HFC exchange. In summary, $r$ and $\sigma$ play a key role in improving the transferability of adversarial examples, while $\rho$ has a slight effect.

## VI. FURTHER STUDY

### A. Perturbation Budgets of MFI

Regarding step size and perturbation budget, we align with recent studies [23], [26], [35] to ensure a fair comparison ($\epsilon = 16$). Nevertheless, considering changes in the perturbation budget may offer adversarial examples with better generalization capabilities in black-box scenarios. We provided additional results in Fig. 11, which shows a larger perturbation budget leads to a higher attack success rate. However, a larger perturbation budget may lead to potential issues, *e.g.*, insufficient stealthiness. In real-world black-box scenarios, we need to hold a reasonable trade-off between the stealthiness and the deceptive effectiveness of adversarial examples.

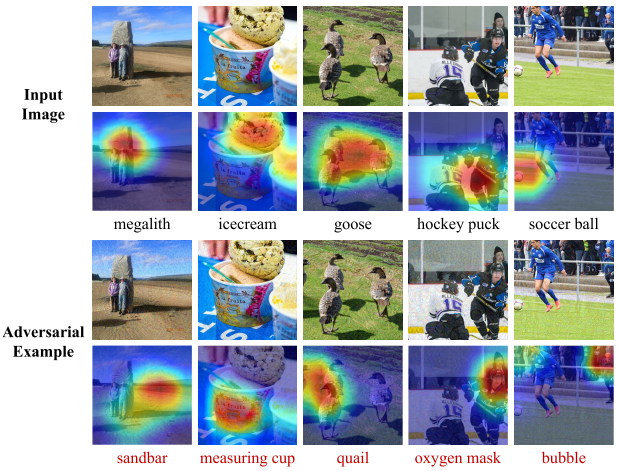



Fig. 12. Visualization for attention shift. We apply Grad-CAM [57] on Inc-v3 to visualize the attention maps of benign images (2nd row) and adversarial images (4th row). Adversarial examples are crafted by our HMFI through Inc-v3. The results show that our adversarial examples can sway the model's attention.

### B. Attention Shifting

We employ the visualization technique, *i.e.*, Grad-CAM [57], to compare the attention maps of clean images and their corresponding adversarial examples, as illustrated in Fig. 12. This visual comparison effectively showcases how our proposed method successfully redirects the model's attention from the main object to other unrelated regions. The contrasting attention maps highlight that our generated adversarial examples consistently induce the model to focus on erroneous features that are unrelated to the actual content of the image, leading to misclassification.

## VII. CONCLUSION & OUTLOOK

In this study, we introduce a novel transform-based attack method, MFI, to enhance the transferability of adversarial examples in black-box scenarios. Unlike previous approaches, we leverage the components within the image to augment the data in the frequency domain. Through extensive experiments, we demonstrate that our proposed HMFI and SMFI methods significantly enhance the success rate of black-box attacks, while maintaining a high success rate in the white-box setting. In future research, we aim to explore additional innovative adversarial attacks in the frequency domain. Simultaneously,

we will develop corresponding countermeasures to ensure the secure deployment of DNNs in the real world.

## REFERENCES


[1] S. Pouyanfar et al., "A survey on deep learning: Algorithms, techniques, and applications," *ACM Comput. Surv.*, vol. 51, no. 5, pp. 92:1–92:36, 2019.

[2] C. Szegedy et al., "Intriguing properties of neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2014, pp. 1–10.

[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2015, pp. 1–11.

[4] X. Tu, Z. Ma, J. Zhao, G. Du, M. Xie, and J. Feng, "Learning generalizable and identity-discriminative representations for face anti-spoofing," *ACM Trans. Intell. Syst. Technol.*, vol. 11, no. 5, pp. 1–19, Oct. 2020.

[5] X. Zhang, X. Zhang, W. Liu, X. Zou, M. Sun, and J. Zhao, "Waveform level adversarial example generation for joint attacks against both automatic speaker verification and spoofing countermeasures," *Eng. Appl. Artif. Intell.*, vol. 116, Nov. 2022, Art. no. 105469.

[6] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[7] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2012, pp. 1106–1114.

[8] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, vol. 16, 2016, pp. 770–778.

[9] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst.*, 2017, pp. 5998–6008.

[10] Z. Liu, Y. Luo, L. Wu, S. Li, Z. Liu, and S. Z. Li, "Are gradients on graph structure reliable in gray-box attacks?" in *Proc. CIKM*, 2022, pp. 1360–1368.

[11] Z. Liu, Y. Luo, L. Wu, Z. Liu, and S. Z. Li, "Towards reasonable budget allocation in untargeted graph structure attacks via gradient debias," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 1–14.

[12] W. Wu, H. Xu, S. Zhong, M. R. Lyu, and I. King, "Deep validation: Toward detecting real-world corner cases for deep neural networks," in *Proc. 49th Annu. IEEE/IFIP Int. Conf. Dependable Syst. Netw. (DSN)*, Jun. 2019, pp. 125–137.

[13] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[14] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–14.

[15] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–23.

[16] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2017, pp. 1–14.

[17] Y. Dong, T. Pang, H. Su, and J. Zhu, "Evading defenses to transferable adversarial examples by translation-invariant attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4312–4321.

[18] J. Lin, C. Song, K. He, L. Wang, and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–12.

[19] L. Gao, Q. Zhang, J. Song, X. Liu, and H. Shen, "Patch-wise attack for fooling deep neural network," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 307–322.

[20] J. Du, H. Zhang, J. T. Zhou, Y. Yang, and J. Feng, "Query-efficient meta attack to deep neural networks," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2020, pp. 1–15.

[21] C. Ma, L. Chen, and J.-H. Yong, "Simulating unknown target models for query-efficient black-box attacks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11835–11844.

[22] X. Wang, X. He, J. Wang, and K. He, "Admix: Enhancing the transferability of adversarial attacks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 16138–16147.

[23] Y. Long et al., "Frequency domain model augmentation for adversarial attack," in *Proc. Eur. Conf. Comput. Vis.*, vol. 13664, 2022, pp. 549–566.

[24] X. Wang, Z. Zhang, and J. Zhang, "Structure invariant transformation for better adversarial transferability," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4607–4619.

[25] C. Xie et al., "Improving transferability of adversarial examples with input diversity," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2725–2734.

[26] Y. Dong et al., "Boosting adversarial attacks with momentum," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 9185–9193.

[27] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 1–10.

[28] J. Zhang et al., "Improving the transferability of adversarial samples by path-augmented method," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 8173–8182.

[29] M. Li, C. Deng, T. Li, J. Yan, X. Gao, and H. Huang, "Towards transferable targeted attack," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 638–646.

[30] Z. Zhao, Z. Liu, and M. A. Larson, "On success and simplicity: A second look at transferable targeted attacks," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 6115–6128.

[31] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–13.

[32] H. Wang, X. Wu, Z. Huang, and E. P. Xing, "High-frequency component helps explain the generalization of convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8681–8691.

[33] Z. Lin, Y. Gao, and J. Sang, "Investigating and explaining the frequency bias in image classification," in *Proc. Int. Joint Conf. Artif. Intell. (IJCAI)*, 2022, pp. 717–723.

[34] A. Kurakin et al., "Adversarial attacks and defences competition," 2018, *arXiv:1804.00097*.

[35] H. Zhu, Y. Ren, X. Sui, L. Yang, and W. Jiang, "Boosting adversarial transferability via gradient relevance attack," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 4718–4727.

[36] H. Zhao, L. Hao, K. Hao, B. Wei, and X. Cai, "Remix: Towards the transferability of adversarial examples," *Neural Netw.*, vol. 163, pp. 367–378, Jun. 2023.

[37] Y. Xu and P. Ghamisi, "Universal adversarial examples in remote sensing: Methodology and benchmark," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5619815.

[38] Y. Tsuzuku and I. Sato, "On the structural sensitivity of deep convolutional networks to the directions of Fourier basis functions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 51–60.

[39] D. Yin, R. G. Lopes, J. Shlens, E. D. Cubuk, and J. Gilmer, "A Fourier perspective on model robustness in computer vision," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 13255–13265.

[40] A. A. Abello, R. Hirata, and Z. Wang, "Dissecting the high-frequency bias in convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, 2021, pp. 863–871.

[41] C. Guo, M. Rana, M. Cissé, and L. van der Maaten, "Countering adversarial images using input transformations," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–12.

[42] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. L. Yuille, "Mitigating adversarial effects through randomization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–16.

[43] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[44] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2017, pp. 4278–4284.

[45] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2016, pp. 630–645.

[46] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–12.

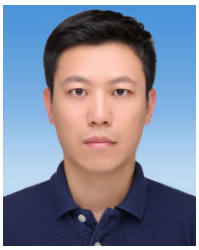
[47] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 11916–11925.

[48] Z. Chen, L. Xie, J. Niu, X. Liu, L. Wei, and Q. Tian, "Visformer: The vision-friendly transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 569–578.

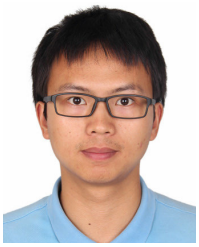
[49] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, 2021, pp. 9992–10002.
[50] F. Tramèr et al., "Ensemble adversarial training: Attacks and defenses," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2018, pp. 1–20.
[51] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proc. 25th Annu. Netw. Distrib. Syst. Secur. Symp.*, 2018, pp. 1–15.
[52] Z. Liu et al., "Feature distillation: DNN-oriented JPEG compression against adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 860–868.
[53] X. Jia, X. Wei, X. Cao, and H. Foroosh, "ComDefend: An efficient image compression model to defend adversarial examples," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6084–6092.
[54] F. Liao, M. Liang, Y. Dong, T. Pang, X. Hu, and J. Zhu, "Defense against adversarial attacks using high-level representation guided denoiser," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1778–1787.
[55] M. Naseer, S. Khan, M. Hayat, F. S. Khan, and F. Porikli, "A self-supervised approach for adversarial robustness," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 259–268.
[56] Y. Chen, Q. Ren, and J. Yan, "Rethinking and improving robustness of convolutional neural networks: A Shapley value-based approach in frequency domain," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2022, pp. 324–337.
[57] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.

**Yaguan Qian** received the Ph.D. degree in computer science from Zhejiang University in 2014. He is currently a Full Professor with the School of Big-Data Science, Zhejiang University of Science and Technology. He has authored or co-authored more than 40 papers in peer-reviewed international conferences and journals, including ICCV, ECCV, AAAI, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and TOPS. His main research interests include AI security, machine learning, pattern recognition, and machine vision.

**Kecheng Chen** received the B.S. degree in mathematics from Zhijiang College, Zhejiang University of Technology, in 2021. He is currently pursuing the M.S. degree with the School of Big-Data Science, Zhejiang University of Science and Technology. His research interests include deep learning, machine learning security, and computer vision.



**Bin Wang** received the Ph.D. degree from China National Digital Switching System Engineering and Technological Research and Development Center in 2010. He is currently a Professor with Zhejiang Key Laboratory of Artificial Intelligence of Things (AIoT) Network and Data Security and Zhejiang University. His main research interests include the Internet of Things security, cryptography, artificial intelligence security, and new network security architecture.



**Zhaoquan Gu** (Member, IEEE) received the bachelor's and Ph.D. degrees in computer science from Tsinghua University in 2011 and 2015, respectively. He is currently a Professor with the School of Computer Science and Technology, Harbin Institute of Technology (Shenzhen), China. He is also a Professor with the Department of New Networks, Peng Cheng Laboratory, Shenzhen, China. His research interests include cyberspace security, cyber range, big data analysis, and artificial intelligence security.

**Shouling Ji** (Member, IEEE) received the Ph.D. degree in electrical and computer engineering from Georgia Institute of Technology and the Ph.D. degree in computer science from Georgia State University. He is a Qiushi Distinguished Professor with the College of Computer Science and Technology, Zhejiang University. His current research interests include data-driven security and privacy, Al security, and software and system security. He is a member of ACM and a senior member of CCF. He was a Research Intern at the IBM T. J. Watson Research Center. He is the recipient of the 2012 Chinese Government Award for Outstanding Self-Financed Students Abroad and ten Best/Outstanding Paper Awards, including ACM CCS 2021.

**Wei Wang** (Member, IEEE) received the Ph.D. degree from Xi'an Jiaotong University, China, in 2006. He is currently a Full Professor with the Ministry of Education Key Laboratory for Intelligent Networks and Network Security, aka MOE KLINNS Lab, Xi'an Jiaotong University. He has authored or co-authored over 100 peer-reviewed papers in various journals and international conferences. His recent research interests include social networks and data security. He is an Editorial Board Member of *Computers & Security* and a Young AE of *Frontiers of Computer Science*. He is a Highly Cited Chinese Researcher in Elsevier.

**Yanchun Zhang** (Member, IEEE) received the Ph.D. degree in computer science from The University of Queensland in 1991. He is currently a distinguished Professor at Zhejiang Normal University. He is also a Professor with the Department of New Networks, Peng Cheng Laboratory, Shenzhen, China, and Emeritus Professor at Victoria University, Australia. He has published over 400 research papers in international journals and conference proceedings, authored/co-authored five monographs and edited a dozen of books in the related areas. His research interests include databases, data mining, social networking, web services and e-health / digital health. His research work has impacted health informatics and information security. He is a founding editor and editor-in-chief of *World Wide Web Journal* (Springer) and *Health Information Science and Systems Journal* (Springer). He has served as an expert panel member at various international research funding agencies such as Australia Research Council (ARC), UK's Medical Research Council (MRC) and Australia's National Health and Medical Research Council (NHMRC). He is a Fellow of The Royal Society of Medicine of UK (FRSM).