Enhancing Transferability of Audio Adversarial Example for Both Frequency- and Time-domain

Zilin Tian[1] , **Yunfei Long**[1] , **Liguo Zhang**[1*] and **Jiahong Zhao**[2]
[1]Harbin Engineering University
[2]University of Southampton
{tzl, 2016064109, zhangliguo}@hrbeu.edu.cn, jz3e23@soton.ac.uk

Abstract

Audio adversarial examples impose acoustically imperceptible perturbations to clean audio examples, fooling classification models into producing incorrect results. Transferability is a critical property of audio adversarial examples, making black-box attacks applicable in practice and attracting increasing interest. Despite recent studies achieving transferability across models within the same domain, they consistently fail to achieve transferability across different domains. Given that time-domain and frequency-domain models are the two predominant approaches in audio classification, we observe that adversarial examples generated for one domain demonstrate significantly constrained transferability to the other. To address this limitation, we propose an Adaptive Inter-domain Ensemble (AIE) attack, which integrates transferable adversarial information from both domains and dynamically optimizes their contributions through adaptive weighting, improving the cross-domain transferability of audio adversarial examples. Extensive evaluations on diverse datasets consistently demonstrate that AIE outperforms existing methods, establishing its effectiveness in enhancing adversarial transferability across domains.

1 Introduction

Deep neural networks (DNNs) have achieved exceptional performance in audio classification tasks [Sadovsky *et al.*, 2023; Liu *et al.*, 2024; Choi *et al.*, 2025; Gazneli *et al.*, 2022; Berg *et al.*, 2021]. However, they have been demonstrated to be vulnerable to audio adversarial examples, which are maliciously crafted by imposing acoustically imperceptible perturbations on clean audio examples to result in erroneous predictions [Goodfellow *et al.*, 2015; Hai *et al.*, 2023; Huang *et al.*, 2023; Ma *et al.*, 2023a; Jin *et al.*, 2024]. Numerous works suggest adversarial examples exhibit cross-model transferability, i.e., those crafted against one surrogate model can also mislead other models. This property makes black-box attacks practically effective, posing significant threats to safety-critical applications. This paper focuses on investigating the transferability of audio adversarial examples because of their practicality.

There are two approaches, gradient-optimization [Fang *et al.*, 2024; Zhu *et al.*, 2023; Zhang *et al.*, 2023] and model ensemble [Tang *et al.*, 2024; Chen *et al.*, 2023a; Ma *et al.*, 2023b] to improve the transferability of adversarial examples in the image field. Transferable audio adversarial attacks are inspired by these similar methods [Tripathi and Mishra, 2022; Abdoli *et al.*, 2019; Koerich *et al.*, 2020]. However, we observe that the generated audio adversarial examples face a critical limitation, i.e., they exhibit low transferability across different domains, specifically the time and frequency domains [Chen *et al.*, 2024]. Typically, audio classification models are categorized into time-domain models [Wang *et al.*, 2023; Zeng *et al.*, 2021; Abdallah *et al.*, 2022] and frequency-domain models [Gong *et al.*, 2021; Liu *et al.*, 2024; He *et al.*, 2024] according to the input form. We find that audio adversarial examples generated by attacking a time domain surrogate model exhibit very limited transferability to frequency domain black-box models, as illustrated in Figure 1a. This limitation could be attributed to changes in data distribution introduced by the time-frequency transformation and the architectural differences between models. Moreover, the nonlinearity of the time-frequency transformation makes the exact inversion of perturbations from spectrograms back to audio intractable, consequently, adversarial examples crafted directly in the frequency domain are less likely to transfer to time-domain models, as shown in Figure 1b.

The above observations motivate us to dive into the question: *How can collective transferable adversarial information from both domains be effectively captured and directly used to generate perturbations on the audio waveform?* To cope with this problem, we investigate performing an inter-domain ensemble strategy, which fuses outputs from both domains to get an ensemble loss. We demonstrate that the time-frequency transformation, such as transforming a Mel-spectrogram back to a waveform, is differentiable, enabling the gradient of the ensemble loss to be propagated back to the audio waveform. Therefore, the generated adversarial examples would be optimized to converge toward a common adversarial space among both domains. However, we observe that simply averaging the outputs of both domains causes ad-

* Corresponding author.

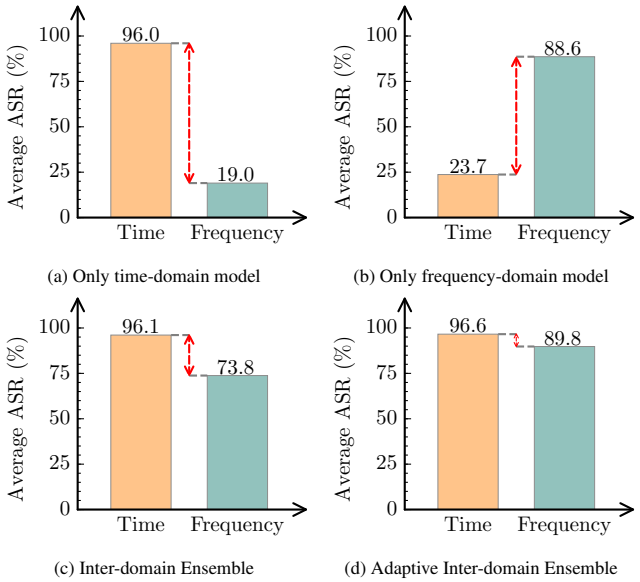


Figure 1: The attack success rates with different surrogate models. (a) and (b) are the MI-FGSM using a surrogate model from the time and frequency domains, respectively. (c) and (d) are the Inter-domain Ensemble (IE) and Adaptive Inter-domain Ensemble (AIE), which use the time and frequency surrogate models. The dataset is UrbanSound8k.

versarial examples to predominantly transfer to one domain, while their transferability to the other domain decreases compared to ones crafted against the uni-domain case, as depicted in Figure 1c. Such a phenomenon violates the intention of improving transferability by integrating information from both domains. Analyzing the evolution tendency of adversarial examples during the update process, we argue that this update progression is often dominated by one domain, suppressing the transferability to the other one. Hence, we further propose an adaptive domain weight adjustment method. Specifically, the weight assigned to each domain's output is dynamically adjusted by monitoring the ratio of the cosine similarity between the potential outputs of each domain and the ensemble potential output. We term the proposed inter-domain ensemble attack with adaptive domain weight adjustment as the **Adaptive Inter-domain Ensemble (AIE) attack**, as illustrated in Figure 2. It mitigates the above imbalance and enhances cross-domain transferability, as shown in Figure 1d.

Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first work aimed at enhancing the cross-domain transferability of audio adversarial examples to simultaneously deceive classifiers in both the time and frequency domains.
- We propose an Adaptive Inter-domain Ensemble attack that captures collective transferable adversarial information from both domains while dynamically assigning the weight to each domain, enhancing the cross-domain transferability of audio adversarial examples.
- Extensive experiments on diverse datasets demonstrate that AIE consistently outperforms existing methods, the generated audio adversarial examples exhibit higher cross-domain transferability. Notably, AIE can be integrated with various gradient optimization attacks.

2 Overview

2.1 Preliminaries

Given a clean audio x with its corresponding ground-truth label y, the audio classification model f is expected to predict label $\text{argmax} f(x) = y$ with high confidence, where $f(x)$ represents logit output. Let $\mathcal{B}_\epsilon(x) = \{\hat{x} : \|\hat{x} - x\|_p \leq \epsilon\}$ be an ϵ-ball around x, where $\epsilon > 0$ is a predefined perturbation magnitude, and $\|\cdot\|_p$ denotes the L_p-norm (e.g., L_1-norm). Transfer-based adversarial attacks aim to find an example $x^{adv} \in \mathcal{B}_\epsilon$ that misleads the model prediction $\text{argmax} f(x^{adv}) \neq y$, and then transfer x^{adv} to directly attack the black-box target model. The objective can be formulated as the following constrained optimization problem:

$$\max_{x^{adv} \in \mathcal{B}_\epsilon(x)} \mathcal{L}\left(f\left(x^{adv}\right), y\right), \tag{1}$$

where $\mathcal{L}(\cdot)$ is the cross-entropy function. The typical process of generating adversarial examples through gradient iterations can be described as:

$$x_{t+1}^{adv} = \Phi_\epsilon\left[x_t^{adv} + \alpha \cdot \text{sign}(\nabla_{x_t^{adv}} \mathcal{L}(f(x_t^{adv}, y))\right], \tag{2}$$

where $x_0^{adv} = x$, t is the current number of iteration, α is the step size, and Φ_ϵ projects x_t^{adv} into the ϵ-ball around x.

2.2 Motivation

We observe that existing methods for crafting transferable audio adversarial examples face a critical limitation, i.e., they overlook the transferability across different domains, specifically the time and frequency domains. Typically, audio classification models are categorized into the following two types based on input form: time-domain model, denoted as f_w, which uses raw **w**aveforms of audio x as input, and frequency-domain model, denoted f_s, which utilizes **s**pectrograms $\mathbf{F}(x)$ obtained through the time-frequency transformation $\mathbf{F}(\cdot)$. Normally, for a clean audio x, the predictions of the two models are expected to be consistent, which can be expressed as:

$$\text{argmax} f_w(x) = \text{argmax} f_s(\mathbf{F}(x)) = y. \tag{3}$$

Both time-domain and frequency-domain models are vulnerable to adversarial attacks. However, due to differences in data distribution and model structures, existing attacks demonstrate low cross-domain transferability. We notice that the adversarial example x^{adv} generated by attacking the time-domain model f_w exhibits very limited transferability to the frequency-domain model f_s after being converted into spectrograms $\mathbf{F}(x^{adv})$. As the exact inversion of perturbations on spectrograms is intractable due to the nonlinearity of the time-frequency transformation, crafting audio adversarial examples by attacking frequency-domain models remains a largely unsolved challenge. From a practical standpoint, the initial

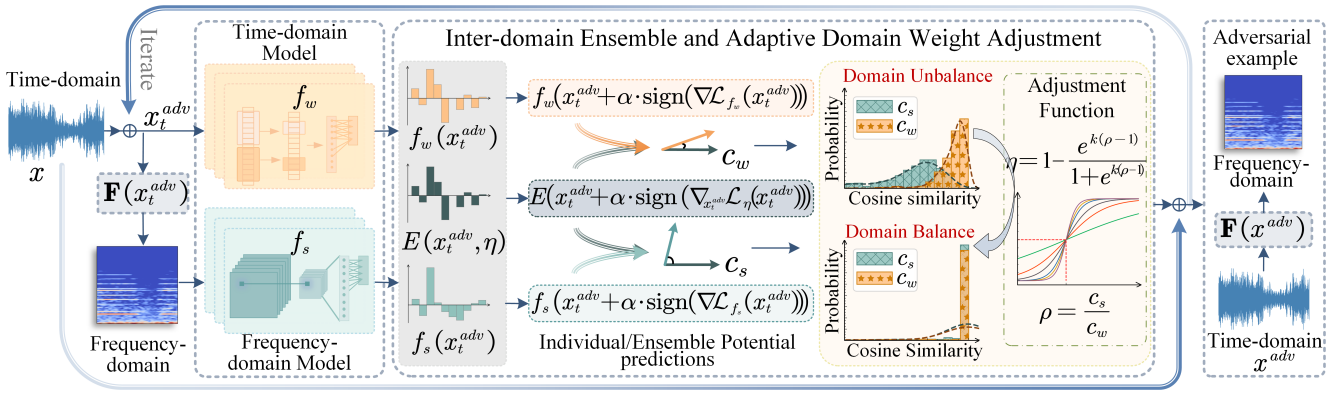


Figure 2: The pipeline of an Adaptive Inter-domain Ensemble attack. It captures collective transferable adversarial information from both time and frequency domains while dynamically assigning weights to each domain through ADWA (Section 3.3), addressing the optimization imbalance problem (Section 3.2), and enhancing the cross-domain transferability of audio adversarial examples. Additionally, we further propose augmenting AIE with an intra-domain ensemble (Section 3.4).

form of input to all audio classification models is raw waveforms. Therefore, a true black-box attack on the audio classification task must directly generate perturbations on waveforms and be able to fool target models across different domains. The final attack objective can be described as:

$$\text{argmax} f_w(x^{adv}), \text{argmax} f_s(\mathbf{F}(x^{adv})) \neq y. \quad (4)$$

3 Adaptive Inter-domain Ensemble Attack

3.1 Uniformly Inter-domain Ensemble

To make generated adversarial examples more likely to transfer to models from both time and frequency domains, we propose an Inter-domain Ensemble attack, which fuses outputs from the time and frequency domains to get an ensemble loss, finding a common adversarial space among them. We begin by simply ensembling two models f_w and f_s. For the x^{adv}, the fusion output can be represented as

$$E(x^{adv}, \eta) = \eta f_w\left(x^{adv}\right) + (1-\eta) f_s\left(\mathbf{F}(x^{adv})\right), \quad (5)$$

where $\eta \in [0,1]$ is the domain weight, initially set to 0.5. Intuitively, only the gradient of f_w can be backpropagated to x^{adv}. However, we introduce a key property of the time-frequency transformation: differentiability.

Proposition 1. ***Differentiability of time-frequency transformation***. *Given an audio x with its corresponding ground-truth label y, the frequency-domain model f_s takes the spectrogram $\mathbf{F}(x)$ as input. The gradient propagation of the loss $\mathcal{L}(f_s(\mathbf{F}(x)), y)$ with respect to x can be expressed as a chain as follows:*

$$\nabla_x \mathcal{L}\left(f_s\left(\mathbf{F}\left(x\right)\right), y\right) = \frac{\partial \mathcal{L}\left(f_s\left(\mathbf{F}\left(x\right)\right), y\right)}{\partial \mathbf{F}\left(x\right)} \frac{\partial \mathbf{F}\left(x\right)}{\partial x}. \quad (6)$$

The gradients of the time-domain and frequency-domain models in the ensemble loss can be reliably propagated backward. Consequently, the iterative process of inter-domain ensemble attack can be formulated as:

$$x_{t+1}^{adv} = \Phi_\epsilon\left[x_t^{adv} + \alpha \cdot \text{sign}(\nabla_{x_t^{adv}} \mathcal{L}(E(x_t^{adv}, \eta))\right]. \quad (7)$$

The proposed attack paradigm can leverage collective vulnerabilities of the time-domain model and frequency-domain model and generate perturbation on raw waveform, effectively boosting the cross-domain transferability of audio adversarial examples. However, we observe that simply averaging the outputs of the time domain and frequency domain, i.e., $\eta = 0.5$, causes the transferable information to be dominated by one domain, ultimately leading to suboptimal results. Let us take Figure 1 as an example again, the attack success rate decreases by 26.0% when uniformly attacking f_w and f_s, compared to attacking only f_s ($\eta = 0$). To mitigate this problem, we further propose an adaptive adjustment strategy.

3.2 Attack Imbalance Analysis

Audio data typically exhibits high dimensionality (e.g., 22,050 Hz). Due to the curse of dimensionality, gradient similarity is unsuitable as a metric to observe the discrepancy in adapting to different domains during the updating process of the inter-domain ensemble attack, e.g., their cosine similarity tends to approach zero. Therefore, we consider calculating the similarity between outputs for evaluating the discrepancy. Specifically, for models f_w and f_s, we define the individual potential outputs using the probability outputs on adversarial examples generated through one-step attacks, expressed as,

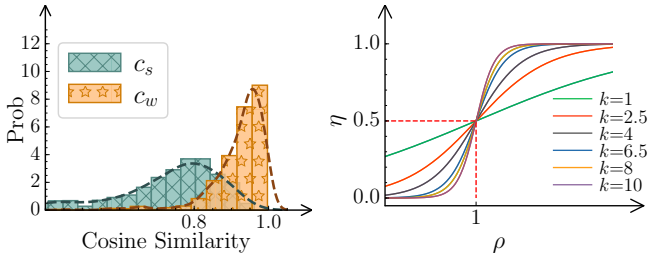
$$\begin{cases} \mathbf{p}_s = \text{softmax}\left(f_s(\mathbf{F}(x_t^{adv} + \alpha \cdot \text{sign}(\nabla_{x_t^{adv}} \mathcal{L}(f_s(\mathbf{F}(x_t^{adv})), y))))\right) \\ \mathbf{p}_w = \text{softmax}\left(f_w(x_t^{adv} + \alpha \cdot \text{sign}(\nabla_{x_t^{adv}} \mathcal{L}(f_w(x_t^{adv}), y)))\right). \end{cases} \quad (8)$$

We then define the ensemble potential output as the average probability output on adversarial examples generated using the gradient of the inter-domain ensemble loss with respect to the current adversarial examples x_t^{adv}, expressed as,

$$\mathbf{p} = \text{softmax}(E(x_t^{adv} + \alpha \cdot \text{sign}(\nabla_{x_t^{adv}} \mathcal{L}(E(x_t^{adv}, \eta), y)), \eta)), \quad (9)$$

where η is initially set to 0.5, indicating that the time domain and frequency domain are given equal weight.

The discrepancy between the individual potential outputs and the ensemble potential output reflects the extent of dom-



(a) Attack imbalance problem (b) Adaptive adjust function

Figure 3: (a) shows the probability densities of c_s and c_w, which is imbalance. (b) illustrates the influence of the parameter of k on the adjustment function.

inance of models in the attack process. We evaluate the discrepancy using the cosine similarity,

$$\begin{cases} c_w &= \cos\left(\mathbf{p}_w, \mathbf{p}\right) \\ c_s &= \cos\left(\mathbf{p}_s, \mathbf{p}\right). \end{cases} \tag{10}$$

We statistically compared the probability densities of c_w and c_s in the normal attack process, as shown in Figure 3a. Obviously, the updating progress of the adversarial example is often dominated by one domain (e.g., time domain).

3.3 Adaptive Domain Weight Adjustment

To address the above optimization imbalance problem, we propose to adaptively adjust the weights of each domain η via monitoring the ratio between c_w and c_s. Here, we define the ratio as $\rho = c_s/c_w$. With ρ to dynamically monitor the dominant discrepancy between time and frequency domains, we are able to adaptively modulate the domain weight η via:

$$\eta = e^{k(\rho-1)} \cdot \left(1 + e^{k(\rho-1)}\right)^{-1}, \tag{11}$$

where k is a hyperparameter that controls the slope of the function. Figure 3b intuitively illustrates how the value of η varies with the ρ under the different values of k. Constantly, η is constrained to the interval $[0, 1]$ following its intrinsic properties. In particular, when $\rho = 1$, indicating an equal contribution from both domains, η converges to 0.5.

By employing the domain weight adjustment strategy, the attack process against both domain models can be effectively regulated, thereby mitigating the imbalance issue and enhancing transferability to time and frequency domains. Our AIE method can integrate existing transfer-based gradient attacks to enhance cross-domain attack performance. For example, AIE incorporating MI-FGSM [Liu *et al.*, 2016] is summarized in Algorithm 1.

3.4 Intra-domain Ensemble Enhancement

Following Algorithm 1, AIE improves the cross-domain transferability of audio adversarial examples. While utilizing a single model from each domain demonstrates sufficient performance, we further propose augmenting AIE with an intra-domain ensemble. This approach averages the outputs of multiple models within the same domain, capturing additional intrinsic transferable adversarial information. It transforms Eq. 5 into the following representation:

$$H(x^{adv}, \eta) = \eta \mathbb{E}\left[f_w^i\left(x^{adv}\right)\right] + (1-\eta)\,\mathbb{E}\left[f_s^i\left(x^{adv}\right)\right], \tag{12}$$

Algorithm 1 AIE with MI-FGSM attack

Input: Surrogate models f_w, f_s; A natural audio example x with label y
Parameter: The perturbation magnitude ϵ; the number of iteration T; the decay factor μ; the hyper-parameter k
Output: An adversarial example x^{adv}

Initialize: $\alpha = \epsilon/T$; $\mathcal{M}_0 = 0$; $x_0^{adv} = x$;
for $t = 1$ to T **do**
  Initialize set $\eta = 0.5$
  # Calculate discrepancy ratio between the individual potential outputs and the ensemble potential output
  Calculate the individual potential outputs $\mathbf{p}_s$ and $\mathbf{p}_w$ using Eq. 8
  Calculate the ensemble potential output $\mathbf{p}$ using Eq. 9
  Calculate the discrepancy ratio $\rho = \frac{\cos(\mathbf{p}_s, \mathbf{p})}{\cos(\mathbf{p}_w, \mathbf{p})}$
  # Adaptively adjust the domain weights based on the discrepancy ratio
  Update the domain weight η using Eq. 10
  Calculate the inter-domain ensemble loss $\mathcal{L}\left(E(x^{adv}, \eta), y\right)$ with updated η
  # Update momentum using the gradient of inter-domain ensemble loss
  Get $\mathcal{M}_{t+1} = \mu\mathcal{M}_t + \frac{\nabla_{x_t^{adv}} \mathcal{L}\left(x_t^{adv}, \eta\right)}{\|\nabla_{x_t^{adv}} \mathcal{L}\left(x_t^{adv}, \eta\right)\|_1}$
  # Update adversarial example
  $x_{t+1}^{adv} = \prod_{\mathcal{B}_\epsilon(x)} \left[x_t^{adv} + \alpha \cdot \text{sign}\left(\mathcal{M}_{t+1}\right)\right]$
end for
return $x^{adv} = x_T^{adv}$.

where f_w^i and f_s^i represent i-th surrogate model from time domain and frequency domain, respectively, and $\mathbb{E}(\cdot)$ represents the average of gradients.

Moreover, such an ensemble can be seamlessly integrated with AIE without conflict, as the weight assigned to the inter-domain ensemble output can be adaptively optimized. While extending AIE to incorporate a multi-model ensemble within each domain is straightforward, this extension further enhances overall transferability compared to single-model AIE, particularly in scenarios involving cross-architectural transfer. Empirical results in Section 4.3 validate the effectiveness of this ensemble-based approach.

4 Experiments

4.1 Experimental Setup

Dataset. To comprehensively evaluate the effectiveness of the proposed method, we conduct extensive experiments on two widely recognized datasets for audio classification tasks: UrbanSound8k [Salamon *et al.*, 2014] for environmental sound classification and ShipsEar [Santos-Domínguez *et al.*, 2016] for underwater acoustic target identification.

Pre-processing. All audio examples are standardized to a length of 1 second. From each dataset, we randomly select 1000 clean audio examples, ensuring that each is correctly classified by all evaluated models and preventing data overlap. The results of frequency-domain attacks are reported us-

Base	Domain	Time-domain models							Frequency-domain models						
		Res18_T*	Res50_T	Effi_T	Dens_T	LSTM_T	Res_T^{adv}	Avg.	Res18_F*	Res50_F	Shuf_F	Dens_F	AST_F	Res_F^{adv}	Avg.
MI	IE	96.1	62.3	35.1	40.0	36.9	5.9	46.0	73.8	56.4	42.5	44.6	28.4	5.0	41.8
	AIE	**96.6**	**62.6**	**35.4**	**40.3**	**37.4**	**6.8**	**46.5**	**89.8**	**68.9**	**45.5**	**46.8**	**29.0**	**6.1**	**47.7**
NI	IE	96.1	62.0	34.3	40.9	39.6	5.8	46.4	77.8	59.9	42.0	45.7	27.5	4.9	43.0
	AIE	**96.4**	**62.1**	**38.5**	**42.2**	**41.5**	**7.0**	**48.0**	**96.6**	**69.0**	**47.2**	**50.9**	**32.7**	**6.5**	**50.5**
VMI	IE	96.1	63.2	32.1	40.4	40.7	6.3	46.5	73.7	57.1	42.1	44.6	32.5	5.4	42.6
	AIE	**97.1**	**69.2**	**38.0**	**42.0**	**41.4**	**8.2**	**49.3**	**91.5**	**69.3**	**48.6**	**53.1**	**37.0**	**7.9**	**49.7**
EMI	IE	100.0	67.5	38.7	40.2	37.5	6.9	48.5	81.0	64.8	46.7	48.8	32.2	5.9	46.6
	AIE	**100.0**	**68.3**	**40.8**	**41.8**	**40.6**	**8.8**	**50.5**	**93.9**	**69.8**	**48.9**	**53.8**	**37.5**	**8.1**	**52.0**

Table 1: The attack success rates (%) of adversarial examples generated under different domain settings (Inter-domain Ensemble (IE) and Adaptive Inter-domain Ensemble (AIE)), and various baseline methods, evaluated against time and frequency domain models. The bolded numbers indicate the best results. The dataset is UrbanSound8k.

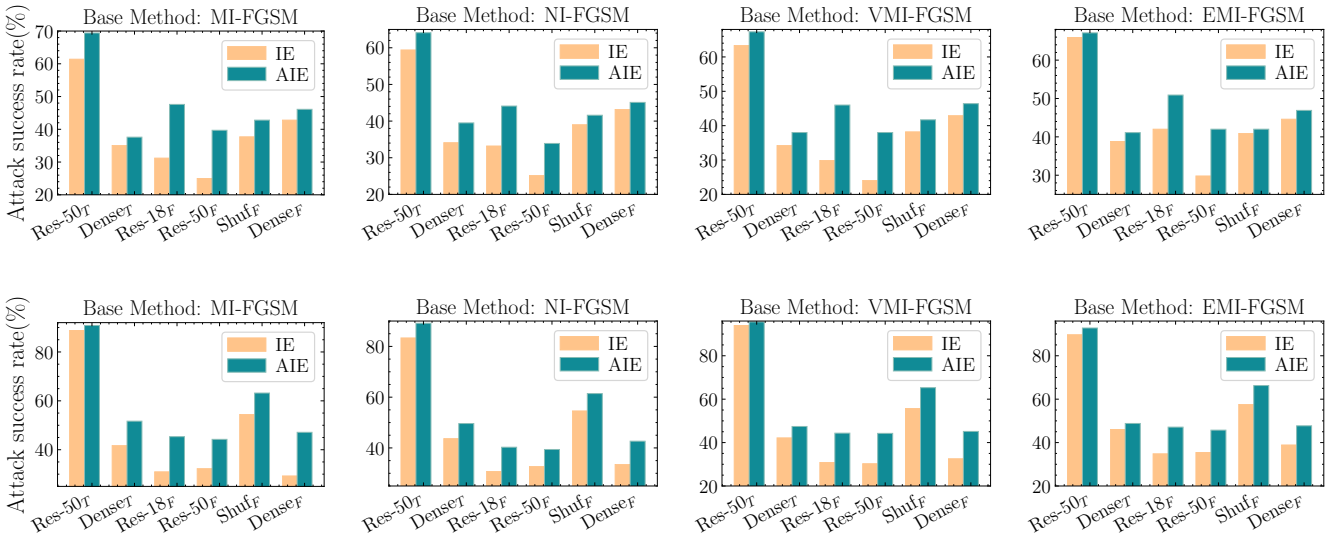


Figure 4: The attack success rates (%) against time and frequency domain models, exploring the transferability of adversarial examples generated by AIE with various baseline methods. The first row presents UrbanSound8k results, while the second row is on the ShipsEar.

ing the **Mel-spectrogram**, which is generated by applying the short-time Fourier transform to the audio and then mapping the output using the Mel scale.

Baselines. Our approach is integrated and evaluated against four advanced gradient-based adversarial attacks: MI [Liu *et al.*, 2016], NI [Lin *et al.*,], VMI [Wang and He, 2021], and EMI [Wang *et al.*, 2021]. Furthermore, we compare our method with state-of-the-art ensemble attacks, namely SVRE [Xiong *et al.*, 2022], AdaEA [Chen *et al.*, 2023a], and CWA [Chen *et al.*, 2023b], which have demonstrated strong transferability in the image field. These baseline methods are meticulously fine-tuned to ensure optimal performance in generating audio adversarial examples.

Models. We evaluate our approach on a diverse range of models in the time domain and frequency domain. These include Convolutional Neural Network (CNN)-based models such as Res18_T and Res18_F (ResNet18) [He *et al.*, 2016], Res50_T and Res50_F (ResNet50) [He *et al.*, 2016], Dens_T and Dens_F (DenseNet121) [Huang *et al.*, 2017], Effi_T (EfficientNet) [Tan and Le, 2019], and Shuf_F (ShuffleNetV2) [Ma *et al.*, 2018]. Additionally, we consider a Transformer-based model AST_F (Audio Spectrogram Transformer) [Gong *et al.*, 2021], and a Recurrent Neural Network-based model, LSTM_T [Sang *et al.*, 2018]. The subscripts $_T$ and $_F$ denote the models from the time domain and frequency domain, respectively. Furthermore, to assess robustness under adversarial conditions, we include adversarially trained models, Res_T^{adv} and Res_F^{adv} [Goodfellow *et al.*, 2015], which are based on the ResNet18 architecture.

Hyper-parameters. We empirically set the maximum perturbation to 0.01 ($l_\infty = 0.01$), the number of iterations $T = 10$, the step size $\alpha = 0.002$. For MI and NI, we set the decay factor $\mu = 1.0$. For VMI, we set the number of sampled examples $N = 20$ and the upper bound of neighborhood size

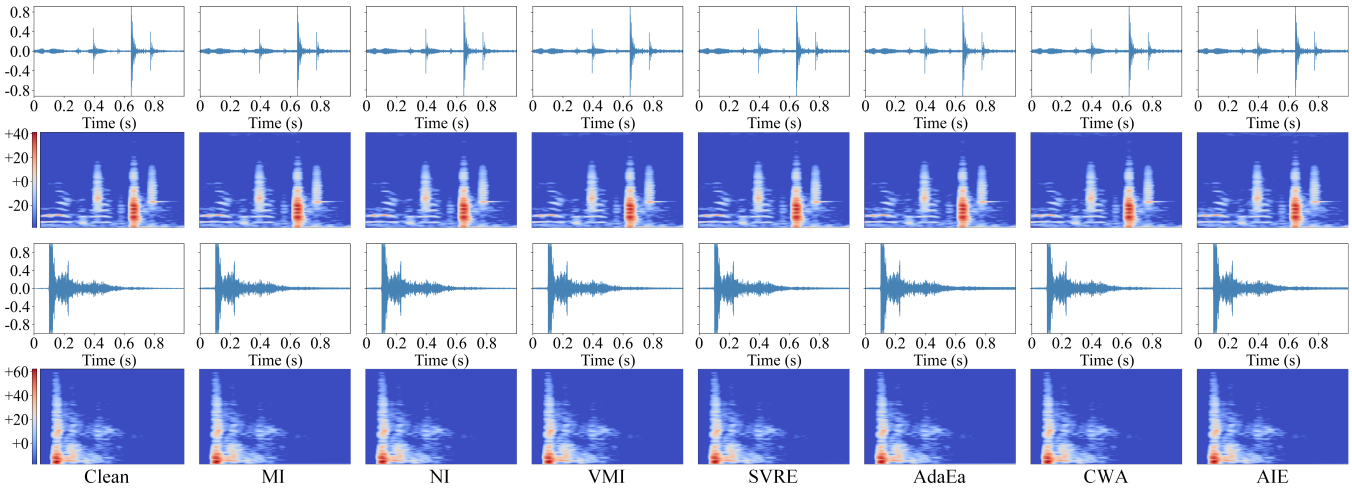


Figure 5: The waveforms and Mel-spectrograms of the audio adversarial examples generated by the gradient-optimization methods, model ensemble methods, and our AIE. The dataset is UrbanSound8k.

$\beta = 1.5 \times \epsilon$. For EMI, we set the number of sampled examples to 11, the sampling interval bound to 7, and adopt the linear sampling. The inner update time in SVRE is set to be four times the number of models. The tolerance threshold and temperature coefficient in AdaEa are set to be -0.3 and 10.

Metrics for evaluation. We adopt the attack success rate as the unified metric to evaluate the performance of various adversarial attack methods, where 'Avg.' denotes the average attack success rate across all target models.

4.2 Attack Results

We evaluate the attack performance of audio adversarial examples generated by integrating our proposed method with the MI, NI, VMI, and EMI methods under surrogate models from both time and frequency domains. Table 1 provides a comprehensive comparison of results across multiple victim models, using Res18_T and Res18_F—both CNN-based architectures—as surrogate models. We observe that attacks utilizing uniformly Inter-domain Ensemble (IE) achieve significantly lower average success rates in the frequency domain compared to the time domain, highlighting the inherent limitations of achieving unbalanced transferability. In contrast, our Adaptive Inter-domain Ensemble (AIE) method consistently achieves the highest attack success rates across both domains. Specifically, when VMI is employed as the baseline attack method, AIE enhances the average attack success rate by 7.1% in the frequency domain, with an even more substantial improvement of 17.8% when attacking Res18_F. These findings highlight the exceptional inter-domain transferability of adversarial examples generated by AIE and demonstrate its effectiveness in reducing the gap between transferability to time and frequency domains.

Furthermore, we evaluate the attack performance of adversarial examples generated using surrogate models with dissimilar architectures—Res18_T (CNN-based) for the time domain and AST_F (Transformer-based) for the frequency domain, as depicted in Figure 4. The first row of the figure corresponds to results on the UrbanSound8k dataset, while the second row represents the ShipsEar dataset. Figure 4 consistently demonstrates that our AIE method significantly achieves higher attack success rates in both the time and frequency domains. These results highlight the ability of AIE to effectively balance and enhance transferability to both domains, regardless of differences in the model architectures.

Additionally, Figure 5 visualizes the waveforms and Mel-spectrograms of audio adversarial examples generated by our proposed method and MI, NI, VMI, and EMI methods under the IE strategy. The first and third rows depict the waveforms, while the second and fourth rows present the corresponding Mel-spectrograms. Although adversarial examples generated by baseline methods are similar to the clean input in both time and frequency domains, they exhibit limited effectiveness in deceiving the black-box models in the frequency domain. In contrast, our method produces adversarial examples that are perceptually similar to those of baseline attacks in both domains but achieve higher cross-domain transferability.

4.3 Evaluate Intra-domain Ensemble

AIE effectively enhances cross-domain transferability by leveraging a single surrogate model for each domain. Previous studies [Xiong *et al.*, 2022; Chen *et al.*, 2023a; Chen *et al.*, 2023b] have demonstrated that incorporating more surrogate models is a powerful strategy for further improving attack transferability. To illustrate the scalability of our approach, this section explores the impact of varying the number of surrogate models on transferability. Table 2 presents the attack results using more surrogate models, including Res18_T, Dens_T, Res18_F, and AST_F, and compares these results with those of SVRE, AdaEA, and CWA—existing inter-domain ensemble methods. While these existing methods improve the transferability in the time domain compared to MI, they show lower success rates in the frequency domain. In contrast, our AIE method achieves substantial transferability improvements in both the time and frequency domains. Additionally, our method has a lower complexity of $O(6n)$ (n is audio size), compared to SVRE's $O(Mn)$ (M=16 is inner it-

Surrogate Models	Attack	Time-domain models						Frequency-domain models					
		Res18_T	Res50_T	Effi_T	Dens_T	LSTM_T	Avg.	Res18_F	Res50_F	Shuf_F	Dens_F	AST_F	Avg.
Res18_T Dens_T Res18_F AST_F	MI	97.3	62.1	36.7	91.7	36.7	64.9	73.4	56.9	42.9	46.6	62.6	56.5
	SVRE	97.2	62.8	38.1	91.8	35.4	65.1	84.1	65.7	46.1	51.5	75.6	64.6
	AdaEA	99.7	61.5	37.1	86.5	34.2	63.8	74.7	61.5	46.6	49.1	64.9	59.4
	CWA	87.5	54.5	35.4	90.1	35.4	60.6	87.1	68.8	47.3	53.5	81.4	67.6
	AIE	**99.8**	**63.3**	**38.6**	**92.2**	**38.9**	**66.6**	**90.3**	**69.9**	**48.1**	**55.2**	**82.7**	**69.2**
Res18_T Dens_T AST_F	MI	98.7	61.3	35.0	93.2	37.6	65.2	30.7	23.2	37.2	44.4	64.4	40.0
	SVRE	97.7	61.2	38.0	94.1	38.4	65.9	42.5	31.8	40.6	46.4	74.2	47.1
	AdaEA	95.7	56.3	35.0	87.8	35.8	62.1	43.5	36.8	40.4	42.5	79.9	48.6
	CWA	75.1	49.9	33.0	87.6	38.0	56.7	49.7	38.7	45.2	45.7	87.2	53.3
	AIE	**100.0**	**58.3**	**38.8**	**95.1**	**40.6**	**66.5**	**50.7**	**39.2**	**48.0**	**46.9**	**96.7**	**56.3**

Table 2: The attack success rates (%) of various model ensemble methods. The dataset is UrbanSound8k.

erations) and Adaea's $O(k(k+1)n)$ (k is number of models).

4.4 Ablation Studies

We further conduct ablation experiments to evaluate the effectiveness of the adaptive domain weight adjustment (ADWA). Additionally, we also analyze the impact of the hyperparameter k and perturbation ϵ. The experiments are conducted on the UrbanSound8k dataset.

Adaptively Domain Weight Adjustment (ADWA)

ADWA is a crucial strategy for leveraging collective vulnerabilities to enhance and balance cross-domain transferability. As illustrated in Figure 6, after applying ADWA, the cosine similarity of c_w and c_s both tend to 1, and their probability densities almost coincide. This demonstrates that the dominance levels of time and frequency domain models in the attack process tend to balance.

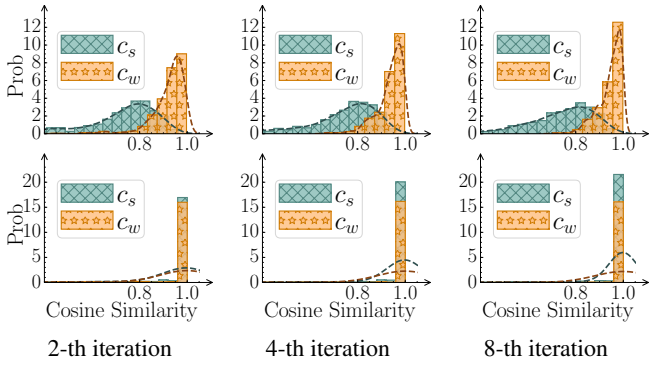


Figure 6: The probability densities of c_w and c_s across various iterations, without ADWA (first row) and with ADWA (second row), demonstrate that c_w and c_s tend to balance.

Parameters k on Adjustment Function

We introduce the hyperparameter k that controls the slope of the function of η with respect to ρ in this paper. Figure 7a presents the effect of different k on the average attack success rate of the adversarial examples. When k = 3, we obtain the highest average success rate in both the time and frequency domains. Therefore, k = 3 is set in our experiments.

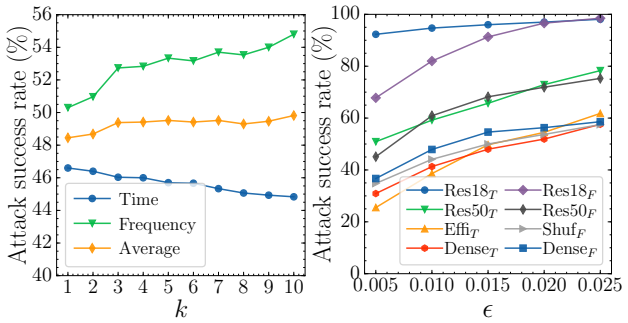


Figure 7: The success rates varies with different values of k and ϵ.

Perturbation Magnitude ϵ

The impact of perturbation magnitude ϵ on the attack success rates is illustrated in Figure 7b. We observe that a larger perturbation results in a higher attack success rate in both the time and frequency domains. To balance the performance and the imperceptibility [Chen *et al.*, 2023a; Chen *et al.*, 2023b], we set the ϵ to 0.01 in our experiments.

5 Conclusion

This work proposes the Adaptive Inter-domain Ensemble attack for generating adversarial audio examples with enhanced cross-domain transferability. AIE introduces an adaptive domain weight adjustment method that dynamically assigns weights to each domain, effectively addressing domain attack imbalance and significantly improving cross-domain transferability. Additionally, extending AIE to include model ensembles within each domain further enhances overall transferability. Extensive experiments on various datasets consistently demonstrate that our approach achieves higher attack success rates on time and frequency domain models than existing attack methods. Our work could shed light on the great potential of boosting cross-domain transferability through a better design of the adversarial attack methods and provide a reference for attacks on other signals (e.g., electromagnetic signals). We hope our work will inspire further in-depth investigations into cross-domain attacks.

Acknowledgments

This work was supported by National Natural Science Foundation of China under Grants (52071111) and Hainan Provincial Natural Science Foundation of China (623CXTD394).

References

[Abdallah *et al.*, 2022] Habib Ben Abdallah, Christopher J Henry, and Sheela Ramanna. 1-dimensional polynomial neural networks for audio signal related problems. *Knowledge-Based Systems*, 240:108174, 2022.

[Abdoli *et al.*, 2019] Sajjad Abdoli, Luiz G Hafemann, Jerome Rony, Ismail Ben Ayed, Patrick Cardinal, and Alessandro L Koerich. Universal adversarial audio perturbations. *arXiv preprint arXiv:1908.03173*, 2019.

[Berg *et al.*, 2021] Axel Berg, Mark O'Connor, and Miguel Tairum Cruz. Keyword transformer: A self-attention model for keyword spotting. In *Interspeech 2021*, pages 4249–4253. ISCA, 2021.

[Chen *et al.*, 2023a] Bin Chen, Jiali Yin, Shukai Chen, Bohao Chen, and Ximeng Liu. An adaptive model ensemble adversarial attack for boosting adversarial transferability. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4489–4498, 2023.

[Chen *et al.*, 2023b] Huanran Chen, Yichi Zhang, Yinpeng Dong, Xiao Yang, Hang Su, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. *arXiv preprint arXiv:2303.09105*, 2023.

[Chen *et al.*, 2024] Liangwei Chen, Xiren Zhou, and Huanhuan Chen. Audio scanning network: Bridging time and frequency domains for audio classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11355–11363, 2024.

[Choi *et al.*, 2025] Hyosun Choi, Li Zhang, and Chris Watkins. Dual representations: A novel variant of self-supervised audio spectrogram transformer with multi-layer feature fusion and pooling combinations for sound classification. *Neurocomputing*, page 129415, 2025.

[Fang *et al.*, 2024] Zhengwei Fang, Rui Wang, Tao Huang, and Liping Jing. Strong transferable adversarial attacks via ensembled asymptotically normal distribution learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24841–24850, 2024.

[Gazneli *et al.*, 2022] Avi Gazneli, Gadi Zimerman, Tal Ridnik, Gilad Sharir, and Asaf Noy. End-to-end audio strikes back: Boosting augmentations towards an efficient audio classification network. *arXiv preprint arXiv:2204.11479*, 2022.

[Gong *et al.*, 2021] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Interspeech 2021*, pages 571–575, 2021.

[Goodfellow *et al.*, 2015] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the 3rd International Conference on Learning Representations*, San Diego, CA, USA, May 2015.

[Hai *et al.*, 2023] Xuan Hai, Xin Liu, Yuan Tan, and Qingguo Zhou. Sifdetectcracker: An adversarial attack against fake voice detection based on speaker-irrelative features. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8552–8560, 2023.

[He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[He *et al.*, 2024] Wentao He, Yuchen Yan, Jianfeng Ren, Ruibin Bai, and Xudong Jiang. Multi-view spectrogram transformer for respiratory sound classification. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8626–8630. IEEE, 2024.

[Huang *et al.*, 2017] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[Huang *et al.*, 2023] Yihao Huang, Liangru Sun, Qing Guo, Felix Juefei-Xu, Jiayi Zhu, Jincao Feng, Yang Liu, and Geguang Pu. Ala: Naturalness-aware adversarial lightness attack. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2418–2426, 2023.

[Jin *et al.*, 2024] Zhibo Jin, Jiayu Zhang, Zhiyu Zhu, and Huaming Chen. Benchmarking transferable adversarial attacks. *CoRR*, 2024.

[Koerich *et al.*, 2020] Karl Michel Koerich, Mohammad Esmailpour, Sajjad Abdoli, Alceu de S Britto, and Alessandro L Koerich. Cross-representation transferability of adversarial attacks: From spectrograms to audio waveforms. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE, 2020.

[Lin *et al.*,] Jiadong Lin, Chuanbiao Song, Kun He, Liwei Wang, and John E Hopcroft. Nesterov accelerated gradient and scale invariance for adversarial attacks. In *International Conference on Learning Representations*.

[Liu *et al.*, 2016] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *arXiv preprint arXiv:1611.02770*, 2016.

[Liu *et al.*, 2024] Haohe Liu, Xubo Liu, Qiuqiang Kong, Wenwu Wang, and Mark D Plumbley. Learning temporal resolution in spectrogram for audio classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 13873–13881, 2024.

[Ma *et al.*, 2018] Ningning Ma, Xiangyu Zhang, Hai-Tao Zheng, and Jian Sun. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In *Proceedings of the European conference on computer vision (ECCV)*, pages 116–131, 2018.

[Ma *et al.*, 2023a] Qiaowei Ma, Jinghui Zhong, Yitao Yang, Weiheng Liu, Ying Gao, and Wing Ng. A lightweight and efficient model for audio anti-spoofing. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*, pages 1–7, 2023.

[Ma *et al.*, 2023b] Wenshuo Ma, Yidong Li, Xiaofeng Jia, and Wei Xu. Transferable adversarial attack for both vision transformers and convolutional networks via momentum integrated gradients. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4630–4639, 2023.

[Sadovsky *et al.*, 2023] Erik Sadovsky, Maros Jakubec, and Roman Jarina. Speech command recognition based on convolutional spiking neural networks. In *2023 33rd International Conference Radioelektronika (RADIOELEKTRONIKA)*, pages 1–5. IEEE, 2023.

[Salamon *et al.*, 2014] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014.

[Sang *et al.*, 2018] Jonghee Sang, Soomyung Park, and Junwoo Lee. Convolutional recurrent neural networks for urban sound classification using raw waveforms. In *2018 26th European Signal Processing Conference (EUSIPCO)*, pages 2444–2448. IEEE, 2018.

[Santos-Domínguez *et al.*, 2016] David Santos-Domínguez, Soledad Torres-Guijarro, Antonio Cardenal-López, and Antonio Pena-Gimenez. Shipsear: An underwater vessel noise database. *Applied Acoustics*, 113:64–69, 2016.

[Tan and Le, 2019] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019.

[Tang *et al.*, 2024] Bowen Tang, Zheng Wang, Yi Bin, Qi Dou, Yang Yang, and Heng Tao Shen. Ensemble diversity facilitates adversarial transferability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24377–24386, 2024.

[Tripathi and Mishra, 2022] Achyut Mani Tripathi and Aakansha Mishra. Adv-esc: Adversarial attack datasets for an environmental sound classification. *Applied Acoustics*, 185:108437, 2022.

[Wang and He, 2021] Xiaosen Wang and Kun He. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1924–1933, 2021.

[Wang *et al.*, 2021] Xiaosen Wang, Jiadong Lin, Han Hu, Jingdong Wang, and Kun He. Boosting adversarial transferability through enhanced momentum. In *Proceedings of the British Machine Vision Conference*, 2021.

[Wang *et al.*, 2023] Biao Wang, Wei Zhang, Yunan Zhu, Chengxi Wu, and Shizhen Zhang. An underwater acoustic target recognition method based on amnet. *IEEE Geoscience and Remote Sensing Letters*, 2023.

[Xiong *et al.*, 2022] Yifeng Xiong, Jiadong Lin, Min Zhang, John E Hopcroft, and Kun He. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14983–14992, 2022.

[Zeng *et al.*, 2021] Lingcai Zeng, Bing Sun, and Daqi Zhu. Underwater target detection based on faster r-cnn and adversarial occlusion network. *Engineering Applications of Artificial Intelligence*, 100:104190, 2021.

[Zhang *et al.*, 2023] Jianping Zhang, Jen-tse Huang, Wenxuan Wang, Yichen Li, Weibin Wu, Xiaosen Wang, Yuxin Su, and Michael R Lyu. Improving the transferability of adversarial samples by path-augmented method. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8173–8182, 2023.

[Zhu *et al.*, 2023] Hegui Zhu, Yuchen Ren, Xiaoyan Sui, Lianping Yang, and Wuming Jiang. Boosting adversarial transferability via gradient relevance attack. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4741–4750, 2023.