



Article

Distillation-Based Cross-Model Transferable Adversarial Attack for Remote Sensing Image Classification

Xiyu Peng ^{1,2}, Jingyi Zhou ^{1,2} and Xiaofeng Wu ^{1,2,*}

¹ Key Laboratory for Information Science of Electromagnetic Waves (MoE), Fudan University, Shanghai 200433, China; 19307140045@fudan.edu.cn (X.P.); 19307140040@fudan.edu.cn (J.Z.)

² Image and Intelligence Laboratory, School of Information Science and Technology, Fudan University, Shanghai 200433, China

* Correspondence: xiaofengwu@fudan.edu.cn

Abstract: Deep neural networks have achieved remarkable performance in remote sensing image (RSI) classification tasks. However, they remain vulnerable to adversarial attack. In practical applications, classification models are typically unknown black-box models, requiring substitute models to generate adversarial examples. Feature-based attacks, which aim to exploit the importance of intermediate features in a substitute model, are a commonly used strategy to enhance adversarial transferability. Existing feature-based attacks typically rely on a single surrogate model, making the generated adversarial examples less transferable across different architectures, such as transformer, Mamba, and CNNs. In this paper, we propose a high-transferability feature-based attack method, DMFAA (Distillation-based Model with Feature-based Adversarial Attack), specifically designed for an RSI classification task. The DMFAA framework enables the surrogate model to learn knowledge from Mamba, achieving enhanced black-box attack performance across different model architectures. The method consists of two stages: the distillation-based surrogate model training stage and the feature-based adversarial attack stage. In the training stage, DMFAA distills features from Mamba to train surrogate model, ensuring that it retains its own structure while incorporating features from other models. In the attack stage, we calculate the aggregate gradient of shallow feature through frequency-domain transformation and white-box attack, while using input transformation tailored for RSI. Experiments on the AID, UC, and NWPU datasets demonstrate the effectiveness of DMFAA, which significantly outperforms existing adversarial attack methods. The average success rate of the DMFAA attack exceeds that of state-of-the-art black-box attacks by more than 3%, 7%, and 13% on the AID, UC, and NWPU datasets, respectively, while maintaining a high success rate in white-box attacks.

Keywords: adversarial examples; remote sensing image; deep neural network; scene classification



Academic Editor: Mohammad Awrangjeb

Received: 7 March 2025

Revised: 7 May 2025

Accepted: 9 May 2025

Published: 13 May 2025

Citation: Peng, X.; Zhou, J.; Wu, X. Distillation-Based Cross-Model Transferable Adversarial Attack for Remote Sensing Image Classification. *Remote Sens.* **2025**, *17*, 1700. <https://doi.org/10.3390/rs17101700>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Remote sensing images are widely used in mapping, environmental monitoring, battlefield detection, and other fields [1,2]. To prevent possible risks, remote sensing image recognition technology not only needs to be accurate, but also needs to be highly reliable and safe. In the field of remote sensing, scene classification is a fundamental research direction. Current studies focus not only on optical images but also on synthetic aperture radar (SAR) [3] and hyperspectral images [4]. Deep neural networks (DNNs) have achieved excellent performance in remote sensing image classification. In the early stages, convolution neural networks (CNNs) played a key role in image feature extraction [5]. Since

the emergence of vision transformer (ViT) in 2020, many variants of ViT have been proposed to enhance its classification performance on satellite and aerial images [6]. Unlike the convolution operation with a local receptive field, ViT typically transforms the input image into a series of patches and uses the self-attention mechanism to capture global relationships, which helps in capturing the entire contextual information between and within the object. The vision Mamba, proposed in 2024, is capable of processing high-resolution remote sensing images [7], extracting large-scale spatial features in multiple directions, and is more efficient in terms of GPU memory usage and inference time compared to transformer-based models [8].

Many studies have shown that DNNs face safety threats from adversarial attack. Adversarial attack alters the original image by introducing carefully designed perturbations that cause the deep neural network to produce false predictions [9,10]. In the application scenarios of remote sensing, adversarial attacks can bring serious consequences [11]. For example, adversarial attack could fabricate environmental disasters to mislead rescue decision-making. Additionally, attackers can obscure military forces and infrastructure, disrupting battlefield situational awareness. Therefore, researching and defending against adversarial attacks targeting remote sensing images is of significant practical importance and academic value [12–15].

The field of adversarial attacks in remote sensing has been widely studied. However, there are three limitations in the field of adversarial attacks in RSI. The first aspect is that most adversarial attacks are primarily conducted and validated on CNN-based models, with limited exploration of cross-architecture transferable black-box attacks [16–18]. Although some studies have investigated adversarial attacks on ViT-based models, these approaches often take advantage of architectural properties specific to ViT-based models, including class tokens or self-attention mechanisms [19,20]. Therefore, these ViT-specific attack methods are difficult to migrate to models with different architectures. Secondly, like CNNs and ViTs, Mamba also faces challenges from adversarial perturbations [21]. However, to the best of our knowledge, no existing research has explored the adversarial robustness of Mamba models in RSI classification. To improve the reliability of Mamba in remote sensing applications, it is necessary to find its weakness through adversarial attacks. Finally, transfer-based black-box attacks require a surrogate model to generate adversarial examples. However, existing studies on transfer-based attacks in remote sensing primarily focus on crafting perturbations, with little emphasis on the selection and enhancement of substitute models [22]. Our experiments provide additional insights into improving substitute models for better attack performance.

To address the aforementioned research gaps, we propose a distilled surrogate model with feature-based adversarial attack (DMFAA) for remote sensing domain. Our method enhances the adversarial transferability of RSI classification models, enabling them to achieve better performance across models with different architectures. As shown in Figure 1, models with different architectures exhibit differences in their attention regions for RSI classification. The adversarial examples generated by the DMFAA can alter the model's attention regions on the image, leading to misclassification. Firstly, to ensure the adversarial examples have good transferability on models with different architectures, we propose a model distillation approach to train a new surrogate model. This approach ensures that the shallow features from the surrogate model incorporate information from various teacher models. We use the trained student model as an surrogate model and generate adversarial examples by feature-based attack. Figure 1 shows that in the original image, the student model with the ResNet50 architecture focuses on features more similar to those of the Mamba model in the fourth column. This indicates that the distilled student model can learn feature representations from models with different architectures, thus

improving the ability to generate cross-architecture attack examples. Secondly, we improve the feature-based adversarial attack method by extracting feature importance from both the semantic features and non-semantic features. We obtain the aggregated semantic features by transforming the image in the low-frequency region through frequency domain transformation. At the same time, we use white-box attack to fabricate adversarial example sets and obtain an aggregated non-semantic feature gradient. Finally, we adopt a data enhancement strategy for remote sensing images, including rotation transformation and random occlusion, and optimize the perturbation with momentum iteration. To evaluate cross-architecture black-box attacks, we selected models from four different architectures, including CNN, ViT, CLIP, and Mamba as target models. We conducted our experiments on the AID [23], UC [24], and NWPU [2] datasets to verify the validity of the proposed approach, which achieves state-of-the-art performance across model architectures in a black-box setup.

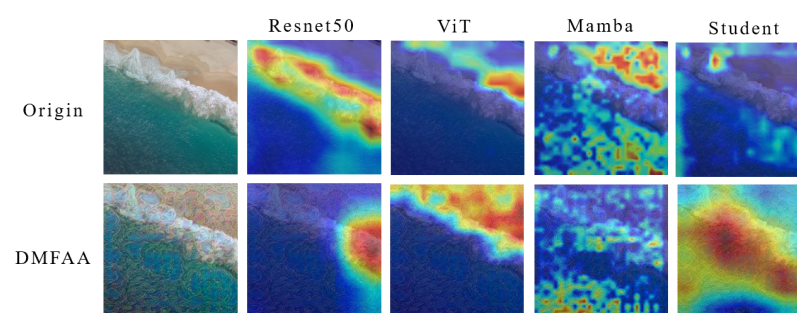


Figure 1. The attention of different models to remote sensing target recognition. We use Grad-CAM [25] to visualize the attention on image features, where red indicates regions with stronger attention. DMFAA shifts the model’s high-attention regions from the key feature locations in the original image to other regions, leading to misclassification.

The main contributions of this paper are summarized as follows:

1. A new framework, DMFAA, is proposed, in which Mamba is used as a teacher model to distill a surrogate model and improve the feature-based attack, thereby enhancing the transferability of adversarial examples in the RSI classification task.
2. The feature-based attack method is improved, differing from existing feature-based adversarial attacks by calculating both semantic and non-semantic aggregate features. This method focuses on the transferability of adversarial examples across models with different architectures.
3. An input transformation technique is developed specifically for remote sensing target detection. This method utilizes regional sequence-independent features of RSI to perform input variations, thereby effectively improving the transferability of adversarial examples across black-box model.
4. The performance of DMFAA is tested on 14 different models, including CNNs, ViTs, CLIP, and Mamba. To the best of our knowledge, this is also the first evaluation of adversarial attacks on Mamba in RSI.

The remainder of this paper is organized as follows. Section 2 reviews works relevant to our research. Section 3 provides an overview of the problem setup and offers a detailed explanation of our proposed method. Section 4 presents the experimental results, conducts ablation studies, and analyzes the findings of the proposed approach. Section 5 discusses the work and suggests directions for future research. Section 6 concludes the paper.

2. Related Work

In this section, we explore the concepts of adversarial attack and introduce several general adversarial attack methods, including white-box attack and black-box attack. Then, we introduce the adversarial attack in remote sensing.

2.1. White-Box Attack

Adversarial example refers to the image after adding slight disturbance to the original sample, which can lead to certain changes in the recognition results after they pass through the victim model [26]. Adversarial examples can lead to the increase in misjudgment rate of deep learning models in image, natural language, 3D point cloud, and other fields [27–29]. Given the input x , data label y , and target model $f(\cdot)$, adversarial example $\tilde{x}$ satisfies the condition $f(x) = y$, $f(\tilde{x}) \neq y$, and $\|x - \tilde{x}\| \leq \epsilon$, where $\|\cdot\|_p$ is the distance metric between the adversarial example and the original image and ϵ is a small constant that controls the magnitude of the perturbation.

Goodfellow et al. [30] proposed the Fast Gradient Sign Method (FGSM), which uses the gradient of the loss function to calculate the disturbance. FGSM attacks only need to make one forward and backward pass in the network to generate adversarial examples. Kurakin et al. [31] proposed the Iterative Gradient Sign Method (BIM). BIM changes the single-step attack to multiple iterations of smaller disturbance, further improving the attack effect of FGSM. Madry et al. [32] proposed Projected Gradient Descent (PGD) method. PGD updates the adversarial example by several iterations; each iteration carries out a small disturbance and clips the disturbed image to the normal range. The definition of PGD is as follows:

$$x_{t+1} = \Pi_{(x+s)}(x_t + \alpha \cdot \text{sign}(\nabla_x J(\theta, x, y))) \quad (1)$$

where α is the step size for each iteration, θ represents the parameters of the target model, and s denotes the number of iterations for the attack. $J(\theta, x, y)$ is the loss function of the model. $\text{sign}(\cdot)$ is the sign function, which outputs -1 , 0 , or 1 . ∇_x represents the gradient with respect to x . x_t denotes the image with perturbation added after the t -th iteration, and when $t = 0$, x_0 is the original image.

Dong et al. [33] proposed the Momentum Iterative Fast Gradient Sign Method (MIM), which is based on momentum iterative gradients. The formal definition of the MIM attack is as follows:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_x J(x_t, y)}{\|\nabla_x J(x_t, y)\|} \quad (2)$$

$$x_{t+1} = x_t + \alpha \cdot \text{sign}(g_{t+1}) \quad (3)$$

where g_t is the gradient of t -th iteration and $g_0 = 0$. μ is the momentum term that controls the accumulation of the gradients. α is the size of the perturbation added at each iteration, defined as $\alpha = \epsilon/T$, where ϵ is the maximum perturbation magnitude and T is the number of iterations for the adversarial attack.

In addition to gradient-based adversarial attacks, the researchers also found other ways to build adversarial examples [34–36]. The C&W attack proposed by Carlini N et al. [34] is also a white-box attack method, which minimizes the objective function by optimizing the algorithm to find the most suitable disturbance. Moosavi et al. [9] proposed Deepfool based on hyperplane classification to find the minimum perturbation that can cause the classifier to misjudge.

2.2. Black-Box Attack

White-box attack methods usually have a high attack success rate, but their practical utility is limited because it is difficult to obtain enough information from the target model in the real world. In response to this situation, researchers have proposed black-box attack methods [37], which assume that the information about the target model is only partially known. Compared to white-box attacks, black-box attacks are more difficult, but have more practical significance. There are three main types of black-box attacks: query-based attacks [38], transfer-based attacks [39], and decision boundary-based attacks [40]. In our study, we mainly focus on transfer-based attack.

Transfer-based attacks train a surrogate model using the same input data as the target model, and then use it to generate adversarial examples. Even if the attacker has no knowledge of the internal structure of the target model, these adversarial examples can still effectively attack the target model. The main directions of transfer-based black-box attacks include model ensemble, input transformation, and feature attack. Liu Y et al. [41] analyzed the transferability of adversarial attack methods and proposed an ensemble-based approach to generate adversarial examples, which enabled adversarial examples to be transferred between multiple models for the first time. Xiong et al. [42] proposed random variance reduce ensemble-based attack, which considers the gradient variance of different models and stabilizes the gradient updates of the ensemble-based attack by reducing the variance. H. Huang et al. [43] proposed a self-integrated black-box attack, which requires only one victim model to achieve a highly adversarial transferable attack against the target detector. For attacks based on input transformation, Xie et al. [44] introduced input diversity to enhance the transferability. K. Wang et al. [45] showed that breaking the internal relationships of images can destroy the attention heat map of depth models, and proposed block shuffling and rotation (BSR) attacks to destroy the attention heat map. For feature-based attacks, Z. Wang et al. [39] discovered that features were overfitting to specific noise in models and proposed the feature importance perception attack (FIA). By disrupting the key object perception features that dominate decision-making across different models, FIA enhances the transferability of adversarial examples. J. Zhang et al. [46] deployed neuron attribution methods to better measure the importance of neurons. Wang et al. [47] proposed black-box feature attack (BFA), where attacking the insensitive features of an image on the surrogate model can better improve transferability.

2.3. Adversarial Attack on Remote Sensing

Adversarial attacks have been widely studied in remote sensing, including semantic segmentation, target recognition, scene classification, and other tasks [17,48,49]. Czaja et al. [50] studied adversarial examples on RSI for the first time and focused on targeted adversarial attacks on deep learning models. Xu et al. [16] used feature similarity to generate remote sensing adversarial examples. Deng et al. [13] proposed that pasting rust-style adversarial patches on RSI can play a role in real scenes. Lin et al. [17] proposed a shallow feature attack method targeting SAR images. Hu et al. [22] conducted research on RSI and proposed an optimization-based attack method specifically designed for a multi-modal RSI. Fu et al. [18] proposed an input transformation method combining spatial domain and frequency domain (SFCoT). Zheng et al. [14] proposed a contrastive learning-based method to improve the transferability of adversarial examples, enabling black-box attacks on SAR. Zhang et al. [51] proposed a ViT-based Jacobian matrix regularization method to improve block-attack in physical scenarios.

At present, there are relatively few existing studies on transferable adversarial attack across models with different architectures in RSI classification. This paper intends to use the

above related research methods as the benchmark methods. Table 1 shows the publication dates and attributes of these benchmark methods.

Table 1. Mainstream adversarial attack algorithm.

Method	Transferability	Algorithm	Year
FGSM [30]	white-box	gradient-based	2014
PGD [32]	white-box	gradient-based	2017
MIM [33]	white-box	gradient-based	2017
DIM [44]	black-box	transfer-based	2019
FIA [39]	black-box	feature-based	2021
NAA [46]	black-box	feature-based	2022
BSR [45]	black-box	transfer-based	2024
BFA [47]	black-box	feature-based	2024
SFCoT [18]	black-box	transfer-based	2024

3. Method

In this section, we propose DMFAA for RSI classification with the aim of generating robust and transferable adversarial examples. The method mainly consists of two parts. First, we show how to train a surrogate model so that the student model and the target black-box model have a higher similarity in the hidden space. Then, we describe how to use the student model for feature-based adversarial attack. We introduce the detail in the following subsections. The overall framework is shown in Figure 2.

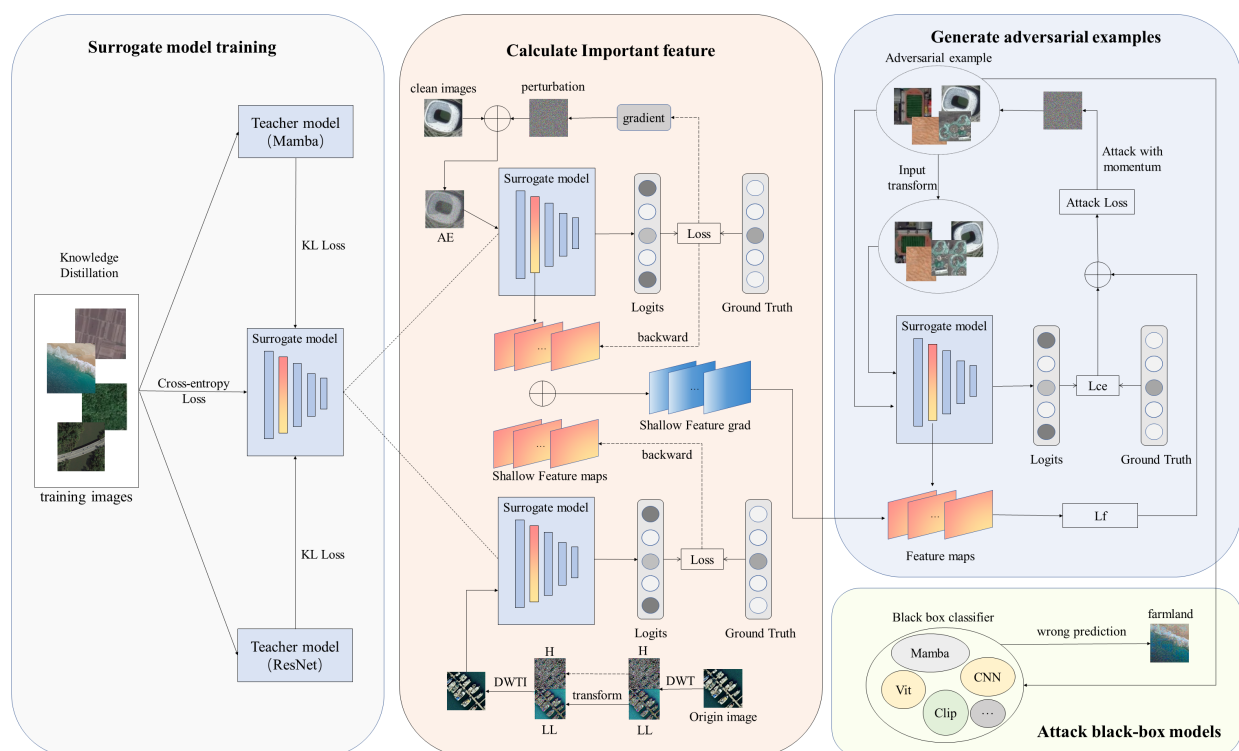


Figure 2. The flow of the proposed DMFAA method. During the surrogate model training stage, the student model distills knowledge from Mamba and ResNet. Then, we calculate the important feature map of the surrogate model through frequency domain transformation and white-box attack. H and LL represent the high-frequency and low-frequency components of the image after DWT transformation, respectively. During the attack stage, perturbations are added to the original image to produce adversarial examples. Lce and Lf denote the cross-entropy loss and feature loss used during the attack stage.

3.1. Improving Surrogate Model

DNNs tend to extract semantic features that align with human attention and are crucial for object recognition and classification, effectively improving classification accuracy. Apart from semantic features, due to the differences in model architectures, the model also focus on some non-semantic features. Existing transfer-based adversarial attacks on RSI typically modify the gradient at the model's decision layer to induce misclassification, without considering the selection and processing of the surrogate model. Adversarial examples generated under such attacks modify both shared semantic features and surrogate model-specific non-semantic features. The former effectively enhances attack success on models with similar architectures, while cross-architecture models are less vulnerable due to differences in semantic feature focus. Meanwhile, modifications to non-semantic features may lead to overfitting to the surrogate model, trapping adversarial examples in its local optimum and eventually reducing their transferability.

To address these limitations, we propose a novel distillation-based adversarial attack method. This is based on student model distillation for RSI classification with the goal of generating robust and transferable adversarial examples. We choose the Mamba as the teacher model to guide the training of the ResNet-based surrogate model. Our aim is to enable the student model to not only focus on the specific features emphasized by the CNN-based architecture but also capture a broader range of features learned from Mamba model architectures. Using the distilled student model for adversarial attack can prevent adversarial examples from falling into the optimal solution of the CNN-based model, thus improving the transferability of adversarial attack.

Mamba-Distilled Student Model

In recent years, SSM-based architectures like Mamba have been able to surpass CNNs and ViTs in image classification. However, the adversarial attack of Mamba has not been studied in the field of remote sensing. In Section 4, we compare the robustness of CNN, ViT, and Mamba models against adversarial attacks. The experimental results indicate that Mamba has superior robustness to multi-step attacks such as PGD, MIM, and DIM compared to CNN. Since most mainstream transfer-based attacks are multi-step attacks, many of the mainstream attack methods designed for CNNs do not perform well against Mamba.

To achieve better cross-architecture transferability, we adopt a dual-teacher distillation approach to train a student model. We hope that the student model can enhance its adversarial robustness by aligning with Mamba during distillation since Mamba exhibits stronger robustness against adversarial attacks. When using a more robust model as the surrogate model, only adversarial examples with stronger attack capabilities can successfully induce misclassifications. Therefore, we expect that the adversarial examples generated by the distilled student model during the attack phase will have an improved effect. Figure 3 shows the distillation process of the surrogate model. Given that Mamba has the ability to capture global information, the student model can focus on a larger range of features during training. Our goal is to enable the student model to learn more comprehensive features, thereby enhancing its cross-architecture transferability between different architectures. We compared different teacher models in Section 4. The experimental results verified the effectiveness of Mamba.

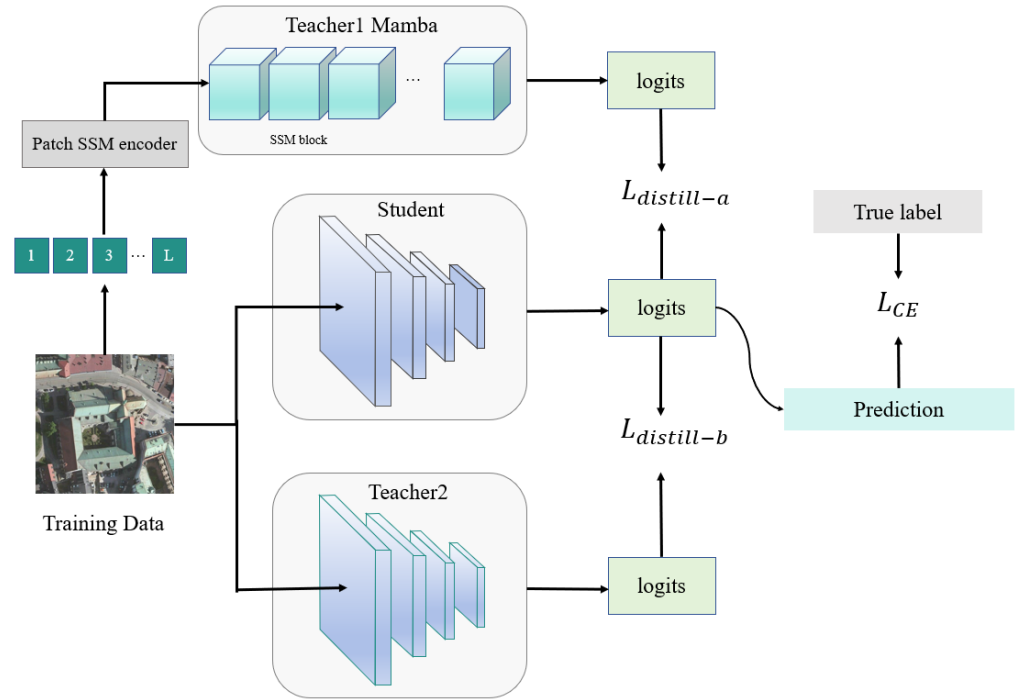


Figure 3. Structure of the distillation method used for surrogate model training. The student model aligns its logits with two teacher models and optimizes cross-entropy with ground-truth labels.

Model distillation is a technique used to transfer knowledge from the teacher model to the student model. It takes the teacher's output as a soft target for training the student, enabling the student model to better generalize to unseen data. We denote the Mamba-l and ResNet50 models as T_1 and T_2 , and the student model, which is also a ResNet50, as S . The Kullback–Leibler (KL) divergence measures the distance between the probability distributions generated by two models. Let p_s and p_t represent the predicted probability distributions of the student model and the teacher model for a given input. The distillation loss function $\mathcal{L}_s$ for the student model is defined as follows:

$$\mathcal{L}_s = \lambda_1 \mathcal{L}_{KL}(S, T_1) + \lambda_2 \mathcal{L}_{KL}(S, T_2) + \mathcal{L}_{CE}(S) \quad (4)$$

where λ_1 and λ_2 are hyperparameters that control the contributions of the different teacher models. The term $\mathcal{L}_{KL}(S, T) = \text{KL}(p_s(x) \parallel p_t(x))$ represents the KL divergence between the student model and the teacher model, while $\mathcal{L}_{CE}$ is the cross-entropy loss between the student model and the true labels. By simultaneously utilizing the information provided by two teacher models with different architectures, the student model can learn more effectively, thereby improving its performance. By balancing knowledge from different teacher models, the student model can acquire cross-architecture knowledge and shift its focus, thereby achieving better attack performance.

3.2. Feature-Based Attack

As mentioned before, neural network models typically extract two distinct types of features from data: semantic features and non-semantic features, both of which are difficult for humans to detect. The former represents common traits that different models will focus on, while the latter corresponds to model-specific feature representations resulting from architectural differences. As shown in Figure 4, different neural networks typically retain more detailed semantic information in shallow feature spaces, such as object contours and textures. And these shallow features often exhibit significant similarity across different

networks. Therefore, by changing the shallow feature space of the image, we can influence the key features that have a greater impact on the final decision of the model.

Feature-based attack methods typically enhance attack effectiveness by designing diverse example inputs to the model, aggregating shallow features to reduce noise interference. FIA uses simple mask transformations on the original image to extract aggregated gradients. However, the experiments in Section 4 have shown that such aggregated gradients may perform poorly when the model architecture spans a wide range from the surrogate model. To separately aggregate the semantic features extracted by the model and the non-semantic features that are imperceptible to humans, we designed two distinct approaches. We created transformed datasets using frequency-domain transformed images and images generated through white-box attacks, respectively, and used these two methods to extract two types of aggregated feature gradients within the model.

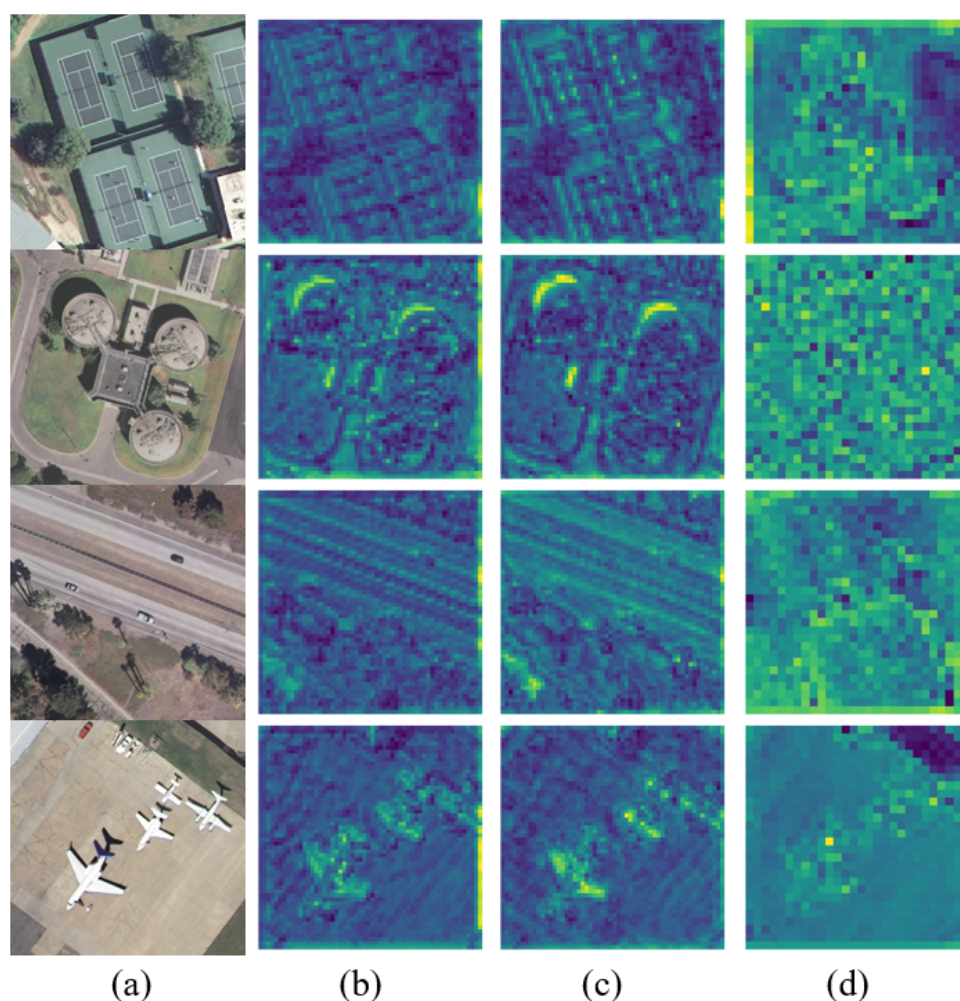


Figure 4. Original input image and average feature maps extracted by three surrogate models, respectively. (a) Origin. (b) Student. (c) ResNet. (d) Mamba.

3.2.1. Aggregate Semantic Feature with Frequency Domain Transformation

The traditional feature-based attack methods only perform random masking transformations on the original image to extract aggregated features. Inspired by SFCoT, we argue that important semantic information is concentrated in the low-frequency components of the image, and transforming this part can effectively alter the semantic content. Discrete wavelet transform (DWT) is a technique for image compression and feature extraction that decomposes an image into sub-bands of different frequencies and spatial resolutions. The

low-frequency components retain the main texture information, while the high-frequency components capture structural details and edges.

Through this approach, we aim to selectively apply image transformations in the low-frequency components during the feature aggregation stage, thereby reducing unnecessary perturbations in the spatial domain and enhancing the extraction of semantic features. We first apply a Bernoulli random mask to manipulate the image, and then use DWT to decompose the image, applying random scaling to the low-frequency components. Our method can be expressed as follows:

$$DWT(x \odot M_1) = \{x_L, x_H\} \quad (5)$$

$$x^{TE} = IDWT((x_L \odot M_2), x_H) \quad (6)$$

where the input image is denoted as x , with the low-frequency and high-frequency components represented as x_L and x_H , respectively. The mask is denoted as M , and the output image after frequency-domain transformation is represented as x^{TE} . The operator $\odot$ refers to the Hadamard product. $M_1 \sim \text{Bernoulli}(1 - \alpha)$ is a random mask applied to the image, setting a portion of the pixels to zero. $M_2 \sim U(1 - r, 1 + r)$ represents a random scaling factor applied to each pixel in the low-frequency component. By applying transformations to the low-frequency components of the image, we distort the details while preserving the spatial structure of the image.

3.2.2. Aggregate Non-Semantic Feature with White-Box Attack

The images based on frequency domain transform dissolve the influence of specific features of the model, and the aggregated gradients reflect the important features brought by semantic information. In addition to the shared semantic features, we also want the adversarial examples to have the ability to attack non-semantic features of the black-box model. Inspired by the BFA method, we generate adversarial examples to aggregate the non-semantic features of the black-box model. We use FGSM to generate adversarial examples:

$$x_{t+1}^{AE} = x + \frac{\nabla_x J_f(x_t^{AE}, y)}{\|\nabla_x J_f(x_t^{AE}, y)\|} \cdot r \quad (7)$$

where x^{AE} is the fitted image, t is the number of iterations, y is the label of the sample, and r is the perturbation strength parameter.

White-box adversarial examples introduce noise into the original image, causing the model's attention to shift further away from the original feature regions. Since the added perturbation is small, it primarily disrupts the semantic features of the original image while allowing non-semantic features to aggregate under different noise patterns. As a result, the aggregated features derived from the set of adversarial examples can capture non-semantic black-box features of the surrogate model.

3.2.3. Aggregate Gradient Calculation

As shown in Figure 5, to enable the model to successfully misclassify adversarial examples of the target label using our method, we transform the clean image into two types of images: one based on frequency domain transformation and the other based on adversarial attack. The important features of these images are then computed within the classification model. We obtain the aggregated features of a specified layer by calculating the dot product between the feature map extracted from that layer and the gradients backpropagated to the same layer. The features propagated through the model to the output at layer k are represented as f_k .

$$G_1 = G_1 + \nabla_{f_k(x)} J_f(x_t^{TE}, y) \quad (8)$$

$$G_2 = G_2 + \nabla_{f_k(x)} J_f(x_t^{AE}, y) \quad (9)$$

where G_1 and G_2 represent the aggregated gradients generated by two different methods. The final aggregated gradient G is given by the following:

$$G = \alpha \cdot \frac{G_1}{\|G_1\|} + \beta \cdot \frac{G_2}{\|G_2\|} \quad (10)$$

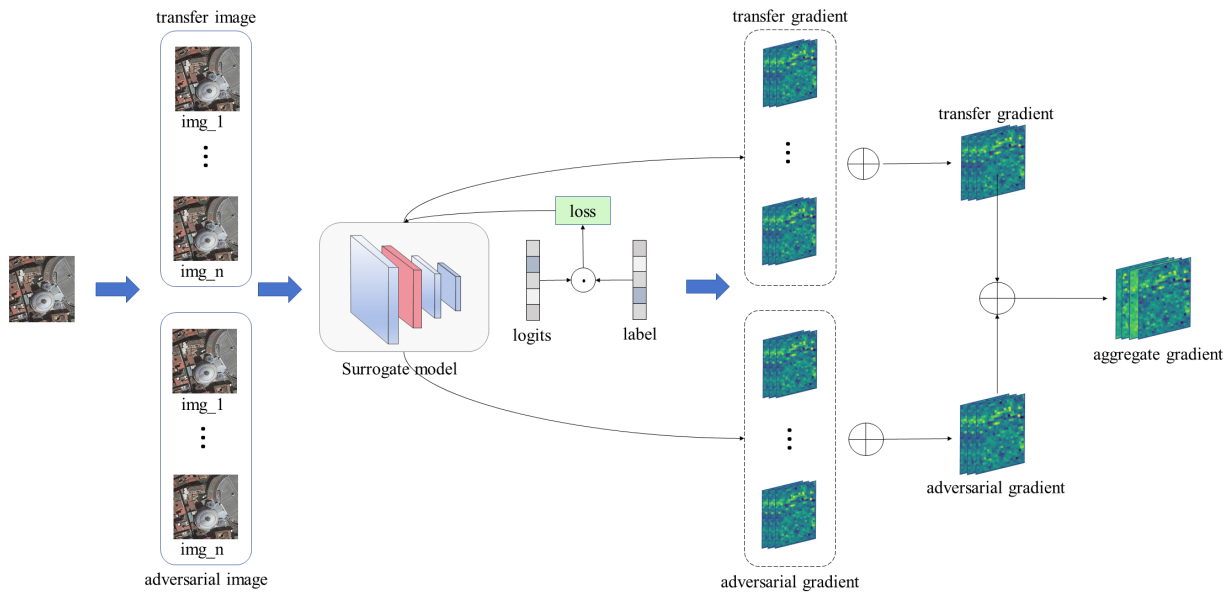


Figure 5. Aggregate feature extraction flow chart. Images undergo frequency domain transformation and white-box attack, and the resulting images are used to conduct feature-based attack on the surrogate model.

3.3. Total Loss Function

Our overall loss consists of two parts. The first part is the feature-based attack loss function L_F , which can be expressed as follows:

$$L_F(x^{adv}) = \sum (G \cdot f_k(x^{adv})) \quad (11)$$

In addition to attacks based on shallow features, we also want the logit output of the alternative model to be far away from the real label of the image. In the second part of the loss function, we perform random rotation, and flip and crop operations on the remote sensing image. These operations not only increase the robustness of the model to image transformations, but also help the model better learn the diversity of images. Different from images of objects like cats and birds, remote sensing images have geographical and spatial structural features, such as buildings, roads, and rivers. Even if the image structure is rotated or flipped, the classification accuracy of the model can still remain at a high level. Therefore, we add cross-entropy loss to the loss function, which measures the gap between the data-enhanced adversarial example and the true label. The loss function L_{CE} for adversarial examples is given by the following:

$$L_{CE}(x^{adv}, y) = \sum_{i=1}^C y_i \log(f(T(x^{adv}))) \quad (12)$$

where $T(\cdot)$ represents a random transformation applied to the image. Therefore, our total loss function is designed as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \eta \cdot \mathcal{L}_{\text{F}} \quad (13)$$

where η is a parameter. Similar to existing feature-based attack methods, DMFAA can update momentum in each iteration of the adversarial example generation process to achieve better results.

$$g_{t+1} = g_t + \frac{\nabla L_{\text{total}}}{\|\nabla L_{\text{total}}\|} \quad (14)$$

$$x_{t+1} = \text{Clip}(x_t + \alpha \text{sign}(g_{t+1})) \quad (15)$$

The overall procedure of the attack algorithm that employs the proposed method to iteratively generate adversarial examples is summarized in Algorithm 1.

Algorithm 1 DMFAA

Input: student model f , teacher model T_1, T_2 , Object function J , origin image x , label y , parameter $\lambda_1, \lambda_2, \alpha, \beta, \eta$

Output: The adversarial example x^{adv}

```

1: for  $i = 0$  to  $I-1$  do
2:   calculate the student model loss function by Equation (4)
3:   Update the weights of student model
4: end for
5: for  $n = 0$  to  $N-1$  do
6:   calculate aggregated gradient  $G$  by Equation (10)
7: end for
8: for  $t = 0$  to  $T-1$  do
9:   construct optimization objective by Equation (13)
10:  Update the momentum by Equation (14)
11:  Update the generated adversarial example by Equation (15)
12: end for
13: return  $x^{adv}$ 

```

4. Experiments

4.1. Experimental Setup

4.1.1. Datasets

We used three RSI datasets for scenario classification, AID [23], UC Merced [24], and NWPU [2]. In the experiment, all images were adjusted to a size of 224×224 .

(1) AID dataset: AID contains 30 categories, each with a different image count, ranging from a minimum of 220 to a maximum of 400. It consists of 10,000 images with a resolution of 600×600 pixels each and is saved in PNG format.

(2) UC dataset: The UC dataset consists of a high-resolution RSI with 21 different scenario categories. Each image in the UC dataset has a resolution of 256×256 pixels.

(3) NWPU dataset: The NWPU dataset consists of RSI with 45 different scenario categories. Each image in the NWPU dataset has a resolution of 256×256 pixels.

4.1.2. Model Architectures

In terms of model selection, we used 14 typical DNN classification models to evaluate attack performance, including ResNet50, Resnet18, DenseNet161, VGG19, MobileNet, efficient-b0, ViT, Swin Transformer, Maxvit, Clip-Resnet, Clip-Vit, Mamba-b, and Mamba-l. As shown in Table 2, all models have more than 70% accuracy with the AID, UC, and NWPU datasets excluding Clip-ResNet50.

Table 2. The accuracy of the classification model on datasets.

Model	AID	UC	NWPU
Resnet50 [52]	100.00	100.00	95.11
Resnet18 [52]	100.00	99.05	96.89
densnet161 [53]	100.00	98.10	97.11
VGG [54]	95.00	98.57	92.67
Mobilenet [55]	97.33	98.57	94.67
efficient-b0 [56]	98.33	97.62	94.22
ViT [57]	78.67	91.90	70.44
swin transformer [58]	97.00	100.00	76.89
maxvit [59]	80.33	92.86	84.67
clip-resnet50 [60]	75.67	64.29	50.22
clip-vit [60]	82.67	82.38	72.89
Mamba-b [7]	98.67	99.52	95.11
Mamba-l [7]	98.00	99.52	96.22
Average [7]	92.44	94.03	85.93

4.1.3. Comparison Methods

In the comparative experiment, we compare DMFAA with seven state-of-the-art attack methods, including PGD, MIM, FIA, NAA, BSR, BFA, and SFCoT. Table 1 shows the publication times and reference scenarios for these methods.

4.1.4. Evaluation Indicator

In the study of adversarial attacks, a commonly used metric for evaluating attack effectiveness is the attack success rate (ASR). ASR is calculated as follows:

$$\text{ASR} = \frac{n_{\text{wrong}}}{n_{\text{total}}} \quad (16)$$

where n_{total} denotes the total number of images in the test dataset and n_{wrong} represents the number of images misclassified by the target model. A higher ASR indicates a more effective attack method.

4.1.5. General Implementation

We randomly select images from the test set as clean original images to generate adversarial examples. During the training of the surrogate model, the number of iterations I and the loss function parameters λ_1 and λ_2 are set to 20, 0.5, and 0.1. In the feature extraction stage, the random mask α and random scaling r in the frequency domain transformation are set to 0.05 and 0.1, respectively, using the db3 wavelet transform. In the white-box adversarial example generation stage of feature extraction, the iteration strength r is set to 28, and the number of iterations t is set to 30. When generating adversarial examples, the perturbation constraint λ , step size α , number of iterations T , and loss function η are set to 16, 1, 30, and 3, respectively. The size of the transformed image set N is set to 5 for gradient computation. The experiment was conducted under the PyTorch framework that supports Nvidia RTX 3090 GPU.

4.2. Result

Quantitative results: We conducted a comprehensive evaluation of the attack performance on different models in a black-box environment, with the results shown in Tables 3–5. Tables 3–5 present the results on the AID dataset, UC dataset, and NWPU dataset, respectively. The target models and attack methods are listed in rows and columns, respectively. For other methods, ResNet50 trained on the original dataset is used as the surrogate model, while for DMFAA, the ResNet50 model distilled with Mamba is used as the surrogate

model. The average accuracies of the target models when trained on the AID, UC, and NWPU dataset were 92.44%, 94.03%, and 85.93%, respectively. After applying the DMFAA attack algorithm proposed in this paper, the average ASR of 14 target models decreased to 14%, 24.84%, and 9.85%.

Table 3. Adversarial attacks on the AID dataset. The best result on each model is highlighted in bold.

Model	Attack	PGD	MIM	FIA	NAA	BSR	BFA	SFCoT	DMFAA
CNN	Resnet50	81.00	92.33	93.00	92.33	100.00	99.33	100.00	94.67
	Resnet18	30.67	61.00	81.33	79.33	90.33	93.33	95.33	94.00
	densnet161	16.33	48.67	76.33	72.00	88.67	90.33	94.67	93.33
	VGG	40.33	72.33	80.33	85.00	90.00	93.33	95.33	96.67
	Mobilenet	40.67	67.33	79.00	82.33	88.00	91.00	90.00	95.67
	Effi-b0	29.67	63.00	81.33	80.33	86.33	91.00	93.67	88.67
Vit	ViT	46.33	52.67	57.33	59.67	67.00	65.00	67.00	77.00
	swint	27.00	50.33	72.00	72.33	80.00	85.33	88.67	87.33
	maxvit	51.00	70.67	66.33	69.00	76.00	78.67	79.00	83.33
Clip	clip-res50	60.33	71.33	85.00	80.00	88.67	91.00	90.00	91.33
	clip-vit	25.67	44.33	59.67	58.33	62.67	69.33	70.00	78.67
Mit	Mamba-b	12.00	26.00	40.67	42.00	46.67	56.00	56.67	70.67
	Mamba-l	11.00	24.67	35.33	38.00	44.33	53.33	52.33	66.67
Average	Average	36.31	57.28	69.82	70.05	77.41	81.31	82.51	86.00

Table 4. Adversarial attacks on the UC dataset.

Model	Attack	PGD	MIM	FIA	NAA	BSR	BFA	SFCoT	DMFAA
CNN	Resnet50	82.86	84.76	83.81	80.48	96.67	96.67	95.24	85.71
	Resnet18	24.76	47.62	70.48	61.90	86.67	88.10	88.10	88.57
	Densnet161	37.62	50.48	74.76	61.43	85.24	85.24	86.19	80.48
	VGG	31.43	54.29	76.19	66.67	85.24	87.62	87.62	84.29
	Mobilenet	30.95	55.71	76.67	66.67	85.71	88.10	86.67	82.38
	Effi-b0	36.19	55.24	69.52	60.95	83.33	80.00	88.10	78.10
Vit	ViT	42.86	43.33	45.71	42.38	45.71	46.19	47.62	53.33
	Swint	6.67	13.33	31.90	31.43	41.90	47.62	49.52	67.14
	Maxvit	31.43	46.67	39.52	44.29	45.71	50.95	52.86	72.38
Clip	clip-res50	66.67	69.05	83.81	75.71	84.29	88.57	86.19	91.90
	clip-vit	31.43	47.62	50.95	46.67	62.86	61.90	60.95	80.48
Mit	Mamba-b	5.71	9.05	12.38	10.48	15.71	20.95	24.76	55.71
	Mamba-l	4.29	9.05	14.29	18.10	19.05	26.19	25.24	56.67
Average	Average	33.30	45.09	56.15	51.32	64.47	66.74	67.62	75.16

The experimental results demonstrate that DMFAA exhibits strong cross-model attack performance on average. For the AID and UC datasets, the results indicate that the distillation-based surrogate model does not perform well on CNN, but achieves higher ASR on ViTs, CLIP, and Mamba. In the NWPU dataset, ASR improved on almost all target models except the white-box surrogate model ResNet50. This may be because the NWPU dataset contains more image categories, allowing the distilled surrogate model to capture richer data information and extract more generalizable architectural features. Our average ASR on the AID, UC, and NWPU datasets exceeds that of state-of-the-art methods, with improvements of 3.49%, 7.54%, and 13.09%, respectively. This demonstrates

that our method significantly enhances the transferability of adversarial examples across different architectures.

Table 5. Adversarial attacks on the NWPU dataset.

Model	Attack	PGD	MIM	FIA	NAA	BSR	BFA	SFCoT	DMFAA
CNN	Resnet50	59.11	77.33	92.89	86.44	99.33	98.00	99.78	92.89
	Resnet18	22.22	54.00	74.00	70.00	81.78	88.67	86.44	98.22
	Densnet161	26.22	49.11	81.11	74.67	80.44	90.00	92.44	92.44
	VGG	31.11	52.22	77.33	76.22	77.56	85.56	87.78	94.89
	Mobilenet	34.44	57.56	79.11	78.22	80.44	91.11	90.00	94.00
	Effi-b0	25.78	47.33	75.11	72.22	77.56	84.22	86.22	87.78
Vit	ViT	55.33	60.44	63.33	65.78	64.44	68.00	65.56	87.11
	Swint	51.33	62.22	65.11	64.67	67.11	72.89	72.44	92.67
	Maxvit	47.33	74.44	72.67	70.22	78.44	83.33	78.22	93.56
Clip	clip-res50	68.00	79.11	89.56	87.11	86.67	92.00	88.44	95.33
	clip-vit	39.33	50.22	59.56	61.78	61.56	69.33	64.00	85.56
Mit	Mamba-b	9.11	20.00	32.89	40.00	28.22	41.11	36.89	76.89
	Mamba-l	11.11	23.33	30.00	39.11	32.89	37.56	38.67	80.67
Average	Average	36.96	54.41	68.67	68.19	70.50	77.06	75.91	90.15

Adversarial defense: To further assess the robustness of the proposed DMFAA method in adversarial defense scenarios, we selected several commonly used defense techniques for evaluation. First, we applied the adversarial attack strategies mentioned earlier to generate adversarial samples. Then, we used three typical adversarial defense methods—Total Variance Minimization (TVM), JPEG Compression (JC), and Quantization-based Defense (QT)—to process these adversarial examples. After applying these defense methods, we re-evaluated the success rate of the adversarial attacks.

The parameter settings for these defense methods are as follows: for TVM, the Bernoulli distribution probability is set to 0.3, and the norm and lambda parameters are set to 2 and 0.5, respectively; for QT, the depth parameter is set to 3. The experimental results are shown in Tables 6–9. After defense processing, the ASR of the DMFAA attack reached its highest value on both datasets. Furthermore, compared to the scenario without defense methods, the ASR of DMFAA decreased to less than that of some other attacks. This indicates that the proposed DMFAA attack method remains highly effective, even in adversarial defense scenarios.

Table 6. Defence of TVM on AID dataset.

Attack	R50	R18	dens	VGG	Mob	effi	ViT	swin	maxvit	clipr	clipv	mb	ml	Avg
PGD	54.00	19.33	14.67	27.33	25.33	24.33	45.67	19.67	49.00	57.33	26.33	22.00	21.33	31.26
MIM	84.00	43.00	37.33	60.67	54.00	52.00	52.33	42.67	66.00	75.00	42.33	31.33	26.33	51.31
FIA	80.33	70.67	68.67	68.33	65.33	72.67	55.00	59.00	63.00	79.67	53.33	46.67	42.33	63.46
NAA	87.00	67.67	64.00	69.33	67.33	70.67	59.67	62.33	65.33	79.33	52.67	50.67	47.00	64.85
BSR	99.00	83.67	77.00	83.00	77.33	75.00	68.00	69.67	71.33	87.67	61.33	47.00	44.00	72.62
BFA	98.67	87.67	82.67	88.67	83.67	86.33	64.00	78.33	76.33	86.00	66.33	57.67	55.00	77.79
SFCoT	99.00	91.33	89.67	90.33	86.00	87.00	66.33	84.00	76.33	90.33	68.00	53.67	53.00	79.62
DMFAA	92.67	89.00	89.00	92.00	90.33	84.33	74.33	81.00	81.67	90.67	75.67	69.67	64.33	82.67

Table 7. Defence of QT on AID dataset.

Attack	R50	R18	dens	VGG	Mob	effi	ViT	swin	maxvit	clipr	clipv	mb	ml	Avg
PGD	67.00	29.67	19.67	42.00	44.00	27.33	49.33	26.67	57.67	63.67	30.00	17.33	14.67	37.62
MIM	88.67	57.67	48.33	71.00	66.33	60.00	53.00	47.67	72.00	72.67	45.00	27.33	24.00	56.44
FIA	85.67	79.00	75.33	80.33	79.00	79.33	57.00	71.67	70.00	83.67	56.33	41.33	37.00	68.90
NAA	92.67	80.00	73.33	85.33	81.33	80.33	59.33	72.33	71.67	81.33	55.33	40.33	37.00	70.03
BSR	100.00	86.67	83.33	90.00	84.00	82.00	67.00	76.67	78.00	85.33	64.33	46.67	41.33	75.79
BFA	99.33	93.00	89.67	94.00	90.33	90.33	66.33	87.00	80.00	87.67	69.67	55.33	50.67	81.03
SFCoT	100.00	94.00	91.33	94.67	87.00	92.00	67.00	87.67	80.33	89.33	68.33	56.67	53.67	81.69
DMFAA	94.00	93.67	94.33	96.67	95.00	89.00	77.00	86.67	83.67	91.00	77.33	70.33	65.33	85.69

Table 8. Defence of TVM on UC dataset.

Attack	R50	R18	dens	VGG	Mob	effi	ViT	swin	maxvit	clipr	clipv	mb	ml	Avg
PGD	62.86	14.76	30.48	22.38	21.43	21.43	42.86	5.71	29.05	63.33	28.10	6.67	5.24	27.25
MIM	79.05	38.10	45.71	49.05	46.67	41.90	44.76	10.48	41.43	72.38	40.48	10.48	10.00	40.81
FIA	80.00	60.95	68.10	68.10	65.24	55.71	43.81	27.62	36.67	80.00	51.90	14.76	17.62	51.58
NAA	75.24	51.43	54.29	62.38	61.43	55.71	43.81	28.10	39.52	75.24	43.81	14.29	19.52	48.06
BSR	95.71	83.81	83.33	82.86	81.43	80.48	46.19	37.14	49.05	86.19	57.14	15.24	19.05	62.89
BFA	94.76	83.81	79.05	81.43	80.95	73.81	46.19	41.90	45.71	87.62	60.95	20.48	23.81	63.11
SFCoT	94.29	86.67	83.33	82.86	82.86	85.24	47.14	45.71	50.48	87.14	59.52	25.24	27.14	65.97
DMFAA	86.19	87.14	80.00	83.81	83.33	74.76	50.95	65.24	72.38	89.05	80.00	51.43	54.29	73.74

Table 9. Defence of JC on UC dataset.

Attack	R50	R18	dens	VGG	Mob	effi	ViT	swin	maxvit	clipr	clipv	mb	ml	Avg
PGD	55.71	17.14	24.76	20.95	19.52	24.76	43.33	7.62	29.52	59.05	28.57	4.29	5.24	26.19
MIM	71.43	40.48	38.10	43.33	41.90	45.24	46.67	15.71	42.86	69.52	43.81	10.48	8.57	39.85
FIA	79.05	62.86	63.81	62.38	62.38	64.29	46.19	32.38	37.62	75.24	51.43	12.86	17.14	51.36
NAA	70.48	52.86	54.76	55.24	55.71	57.62	42.38	31.43	41.43	69.05	48.10	15.24	18.10	47.11
BSR	95.24	82.38	76.19	78.10	78.57	79.05	44.76	38.10	47.62	80.00	55.24	13.81	18.57	60.59
BFA	92.86	84.76	76.67	76.19	75.24	76.19	45.71	47.14	48.57	85.71	59.52	22.38	25.71	62.82
SFCoT	93.33	83.33	79.52	83.33	82.86	84.29	46.19	49.05	49.52	80.95	60.95	24.76	26.67	64.98
DMFAA	85.24	88.10	78.10	82.38	81.90	76.19	52.86	67.62	70.95	88.57	76.67	53.81	59.52	73.99

Qualitative results: In order to better illustrate the feasibility of our proposed method, we perform a visual analysis. Figure 6 shows the application of different attack methods on the two datasets and the effects of adversarial defense processing. It can be seen that the disturbance added to the graph by our proposed method interferes with human eye recognition, but it attacks the DNN models.

Imperceptibility of adversarial examples: Structural similarity (SSIM) is a commonly used metric to measure the imperceptibility of adversarial examples. When the SSIM is 1, it means that the two images are exactly the same. To ensure that the adversarial examples generated by our method maintain imperceptible, we compared the SSIM between adversarial examples and clean samples across three datasets for different attack methods. A higher SSIM means that the adversarial example is more structurally similar to the clean image. The results in Table 10 show that the adversarial examples generated by DMFAA have a similar SSIM than those produced by the state-of-the-art method SFCoT. Although the SSIM of PGD, FIA, and NAA is higher than that of DMFAA, their ASR are much lower. These results demonstrate that our proposed method improves ASR while maintaining imperceptibility.

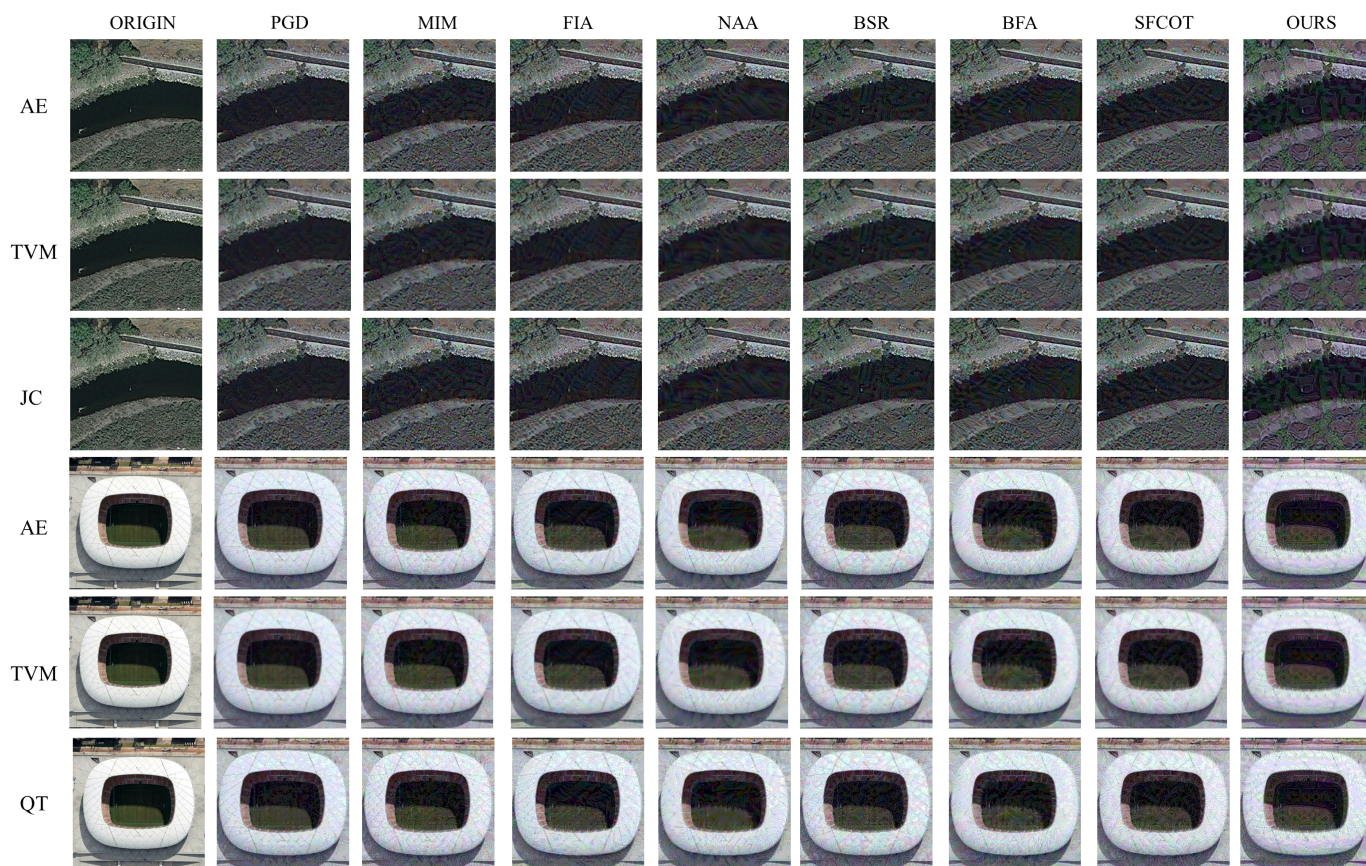


Figure 6. AEs are generated using different adversarial attack methods. Lines 1 and 4 are original adversarial examples from UC and AID datasets, respectively. Lines 2, 3, 5, and 6 are adversarial examples after adversarial defense.

Table 10. SSIM of three datasets.

	PGD	MIM	FIA	NAA	BSR	BFA	SFCoT	DMFAA
UC	0.192	0.165	0.165	0.174	0.138	0.138	0.139	0.147
AID	0.202	0.138	0.175	0.178	0.139	0.151	0.138	0.143
NWPU	0.231	0.155	0.207	0.219	0.156	0.177	0.177	0.162
Avg	0.208	0.152	0.182	0.190	0.144	0.155	0.151	0.151

4.3. Selection of Teacher Model

To emphasize the key role of Mamba model distillation, we conducted an ablation study. We first compare the ability of the CNN, ViT, and Mamba models to resist adversarial attacks. Figure 7 shows models that attack different architectures using four classical white-box attack algorithms. We used FGSM, PGD, MIM, and DIM to conduct white-box attacks against VGG, Resnet, SwinT, and Mamba, respectively. All four models that were used had a classification accuracy greater than 98% on clean images. The results show that the performance of all models decreases with the increase in disturbance intensity. Further analysis shows that the Mamba model is more robust than CNN against multi-step attacks, including PGD, MIM, and DIM. For single-step attacks such as FGSM, especially under large disturbance, Mamba is slightly less robust than CNN. We believe that this difference may be related to the correlation between image regions introduced in the Mamba model. In the Mamba architecture, the correlation between image regions allows FGSM to capture patterns across larger spatial regions better with fewer iterative gradients, which affects its response to perturbations.

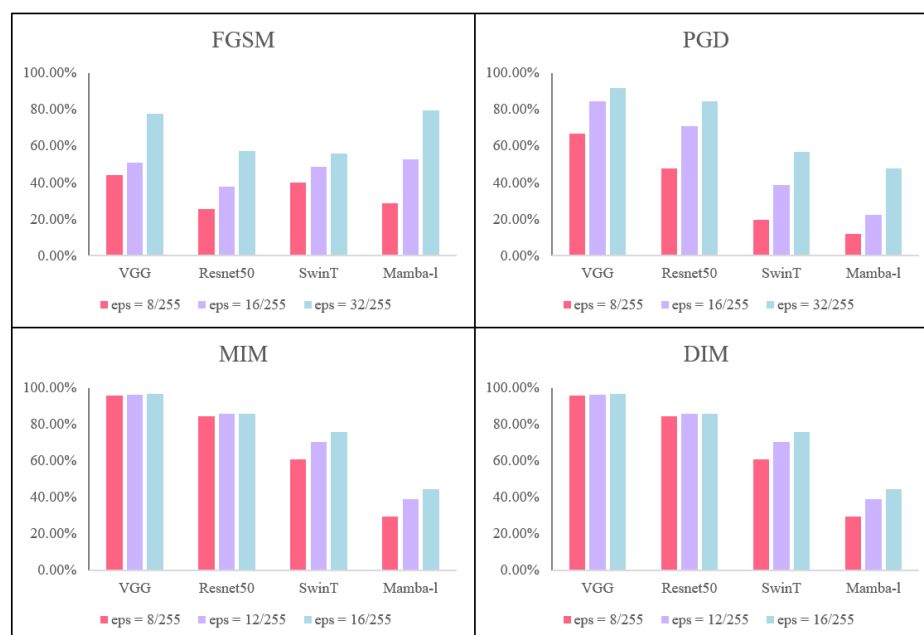


Figure 7. ASR of four white-box attacks on VGG, Resnet, Swint, and Mamba models. The number of iterations is 10.

Next, we attempt to distill the substitute model using teacher models with different architectures. We used an untrained ResNet50 as the student model, with ResNet18, ViT, and Mamba-l as teacher models T1, and the ResNet50 model as teacher model T2. Knowledge distillation was performed on the student model under the same parameter settings. The values of λ_1 and λ_2 were set to 0.5 and 0.1, respectively. Table 11 presents the classification accuracy of adversarial examples generated using student models trained with four different teacher models when evaluated on various target models. From the data in the table, it can be seen that when using only the MIM method, the average ASR is 53.82%. Adding Mamba as the teacher model increases the ASR by approximately 8%. Although the models distilled using ViT and ResNet improve the quality of adversarial example generation compared to the non-distilled MIM, their attack performance on most black-box models remains inferior to that of the Mamba model. This ablation study further indicates that Mamba model distillation, as a black-box attack method, can significantly improve the transferability of attacks.

Table 11. Method performance comparison for different teacher model.

Teacher	R50	R18	dens	vgg	mob	effi	vit	swin	maxvit	clipr	clipv	mb	ml	avg
MIM	72.67	59.67	47.67	68.67	65.00	57.67	55.33	48.00	70.67	72.67	41.00	22.67	18.00	53.82
Vit	56.33	54.33	45.67	73.67	68.33	60.00	58.00	51.67	75.33	69.33	48.33	25.00	22.00	54.46
Resnet18	64.33	63.33	53.67	83.67	73.67	64.67	56.00	61.67	75.00	73.00	44.00	24.67	23.67	58.56
Mamba	67.33	67.67	59.33	79.67	77.00	71.67	58.67	61.00	73.00	77.67	46.00	33.67	30.00	61.74

4.4. Ablation Study

In this section, we focus on some of the hyperparameters that affect the performance of our methods.

Effectiveness of feature extraction. In the feature-based adversarial attack stage, the aggregate feature extraction process in our proposed method consists of two components: frequency domain transformation and white-box attack. The former aims to extract semantic information where the energy is concentrated in the image, while the latter seeks

to extract non-semantic features that are more vulnerable to attack. We conduct ablation experiments on the AID and UC datasets to demonstrate the ability of these two methods in enhancing adversarial transferability. Adversarial examples are generated on ResNet152 and tested on other models. The attack success rates for each dataset are shown in Figure 8. First, as shown in the table, compared to MIM, adversarial examples generated using frequency-domain transformations exhibit lower ASR on white-box models but achieve better performance on black-box models. This suggests that extracting low-frequency information from images helps identify regions with highly concentrated energy in semantic features, thereby enhancing semantic feature integration. Furthermore, adversarial examples generated under white-box attack can capture non-semantic features of black-box models, improving attack effectiveness. Finally, by integrating frequency-domain transformations, white-box attacks, and image transformations within the DMFAA framework, the best attack performance is achieved.

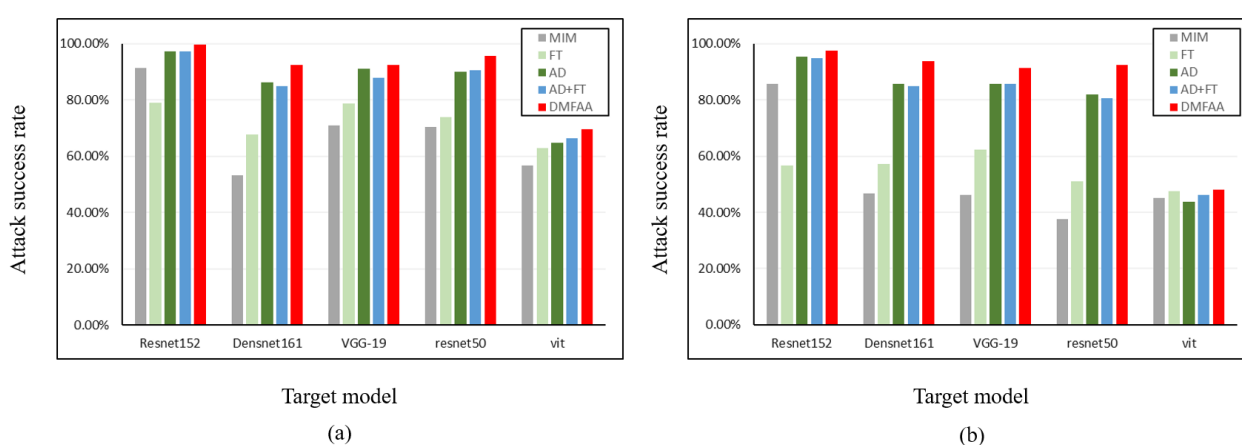


Figure 8. Ablation experiments for feature extraction in each method. (a) AID. (b) UC.

Selection of frequency domain. In DMFAA, we apply transformations in the low-frequency domain of the image to enhance the extraction of semantic information. We used ResNet152 as a surrogate model and performed experiments on nine target models using the UC dataset. The results in Table 12 show that applying transformations to different high-frequency directions performs worse than the origin image. Meanwhile, transforming in the low-frequency domain improves the average ASR by around 2% compared to origin image and 5% compared to the high frequency. These results demonstrate that transforming in the low-frequency space can effectively capture semantic information and enhance the transferability of adversarial examples.

Table 12. Ablation experiment with frequency domain selection. ALL refers to applying transformations on the original image. HH, HL, and LH represent the high-frequency diagonal, horizontal, and vertical direction, respectively. YL refers to the low-frequency domain.

	R152	R50	R18	dens	clipv	mob	vgg	swin	effi	avg
ALL	56.67	48.10	41.90	51.90	53.81	61.43	59.52	26.19	47.62	49.68
HH	51.43	42.86	39.05	49.05	49.05	60.48	60.00	23.33	47.62	46.98
HL	50.95	42.86	40.95	50.95	48.10	59.52	60.00	22.86	47.62	47.09
LH	49.05	43.81	40.95	50.00	48.10	60.00	60.95	21.90	48.10	46.98
YL	56.67	50.95	45.71	57.14	52.38	61.90	62.38	30.00	50.95	52.01

Parameters of the distillation stage. We conducted ablation studies on the teacher model parameters during the distillation phase. As shown in Figure 9, the ablation study

was conducted on the AID dataset. We fixed $\lambda_1 = 0.5$ and varied λ_2 among 0.9, 0.5, and 0.1. The results show that the average ASR gradually increases as λ_2 decreases. Furthermore, when $\lambda_2 = 0.1$, the success rate remains comparable across different values of λ_1 , including 0.3, 0.5, and 5. Based on the experiment, we chose $\lambda_1 = 0.5$ and $\lambda_2 = 0.1$ as the default parameters in our experiments.

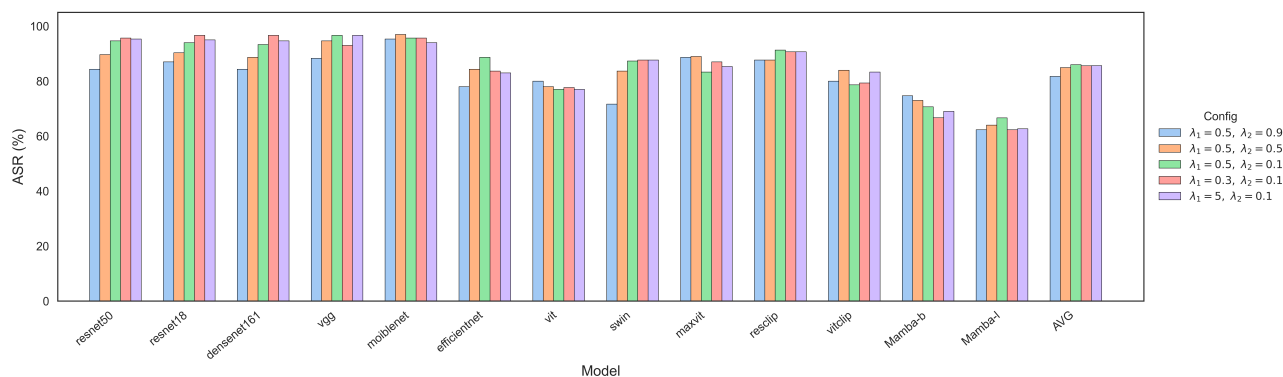


Figure 9. Attack success rate under different λ settings.

Number of iterations. Finally, we investigate the effect of the number of training iterations on the performance of adversarial examples. We conducted experiments with different iterations, and gave the accuracy rates for ResNet50, VGG16, MobileNet, Clip-ResNet50, Vit, and MaxVit, as shown in Figure 10. This helps us to understand how the number of training iterations affects the effectiveness of adversarial attacks. We can observe that from 10 iterations to 30 iterations, the accuracy of model recognition decreases, and the effect of transfer is gradually improved. After reaching 30 iterations, the mobility of the adversarial example reaches saturation or slightly decreases with additional iterations. In order to align previous experiments, we took 30 as the number of iterations.

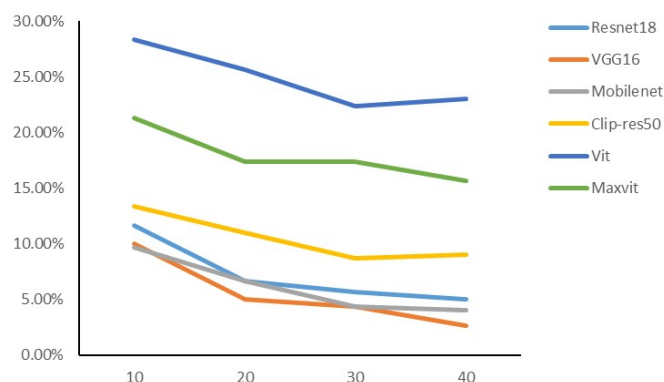


Figure 10. Integrated feature extraction flow chart.

5. Discussion

In summary, our extensive experimental results demonstrate that the proposed DMFAA exhibits outstanding performance across multiple datasets. Its performance not only surpasses classic adversarial attack methods such as PGD and FIA, but also outperforms attack methods like SFCoT that focus on remote sensing images. Additionally, to the best of our knowledge, we are the first to explore adversarial attacks and defense methods for Mamba in remote sensing images, highlighting its practical significance and potential for future development.

Although the effectiveness of the DMFAA in black-box attacks has been validated in our experiments, there remain many challenges and limitations in real-world applications.

On one hand, DMFAA requires model distillation for the surrogate model, which demands more time and computational resources compared to other attacks. On the other hand, the perturbations generated by DMFAA are not completely imperceptible to the human eye. Therefore, future algorithm improvements will focus on enhancing the imperceptibility of attack perturbations and reducing the resource consumption of the attack. In addition, our method has only been verified on optical RSI and has not been verified on other datasets such as SAR and hyperspectral. We will attempt to conduct further exploration in the future.

6. Conclusions

In this paper, we combine model distillation with feature-based attacks and propose DMFAA for remote sensing target recognition, which effectively enhances transferability between models. During the model distillation phase, we select Mamba as the teacher model, enabling the student model to learn cross-architecture features. In the feature attack phase, we selectively transform low-frequency components to reduce the gap between models. Simultaneously, we employ white-box attacks to extract the aggregated non-semantic features from the model. Finally, aggregate gradients are computed on the transformed image set, enabling the stable generation of adversarial examples through iterative refinement. We evaluate the proposed method on two remote sensing datasets to demonstrate the effectiveness, which outperforms state-of-the-art methods in cross-architecture black-box attacks. Beyond the performance advantages, the established benchmark for black-box attacks on remote sensing target recognition also provides insights into model vulnerabilities. In future work, our aim is to further investigate adversarial learning in remote sensing, including adversarial attacks and defenses.

Author Contributions: Conceptualization, X.P. and X.W.; methodology, X.P.; software, X.P.; validation, X.P., J.Z., and X.W.; formal analysis, X.P.; investigation, X.P.; writing—original draft preparation, X.P.; writing—review and editing, X.P., J.Z., and X.W.; visualization, X.P. and J.Z.; supervision, J.Z. and X.W.; project administration, X.P.; funding acquisition, X.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data associated with this research are available online. The AID dataset is available at <https://opendatalab.org.cn/OpenDataLab/AID>, accessed on 10 November 2024. The UC dataset is available at <https://aistudio.baidu.com/datasetdetail/51628>, accessed on 10 November 2024. NWPU dataset is available at <https://aistudio.baidu.com/datasetdetail/220767>, accessed on 1 December 2024. Citations are also provided in the article.

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

References

1. Zhu, X.X.; Tuia, D.; Mou, L.; Xia, G.S.; Zhang, L.; Xu, F.; Fraundorfer, F. Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geosci. Remote Sens. Mag.* **2017**, *5*, 8–36. [\[CrossRef\]](#)
2. Cheng, G.; Han, J.; Lu, X. Remote sensing image scene classification: Benchmark and state of the art. *Proc. IEEE* **2017**, *105*, 1865–1883. [\[CrossRef\]](#)
3. Shi, J.; Wang, W.; Jin, H.; Nie, M.; Ji, S. A Lightweight Riemannian Covariance Matrix Convolutional Network for PolSAR Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–17. [\[CrossRef\]](#)
4. Zhou, J.; Sheng, J.; Ye, P.; Fan, J.; He, T.; Wang, B.; Chen, T. Exploring Multi-Timestep Multi-Stage Diffusion Features for Hyperspectral Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 1–16. [\[CrossRef\]](#)
5. Qin, P.; Cai, Y.; Liu, J.; Fan, P.; Sun, M. Multilayer feature extraction network for military ship detection from high-resolution optical remote sensing images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 11058–11069. [\[CrossRef\]](#)
6. Wang, D.; Zhang, Q.; Xu, Y.; Zhang, J.; Du, B.; Tao, D.; Zhang, L. Advancing plain vision transformer toward remote sensing foundation model. *IEEE Trans. Geosci. Remote Sens.* **2022**, *61*, 1–15. [\[CrossRef\]](#)

7. Chen, K.; Chen, B.; Liu, C.; Li, W.; Zou, Z.; Shi, Z. Rsmamba: Remote sensing image classification with state space model. *IEEE Geosci. Remote Sens. Lett.* **2024**, *21*, 8002605. [\[CrossRef\]](#)
8. Gu, A.; Dao, T. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv* **2023**, arXiv:2312.00752.
9. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
10. Shi, Y.; Wang, S.; Han, Y. Curls & whey: Boosting black-box adversarial attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 6519–6527.
11. Chen, L.; Xu, Z.; Li, Q.; Peng, J.; Wang, S.; Li, H. An empirical study of adversarial examples on remote sensing image scene classification. *IEEE Trans. Geosci. Remote Sens.* **2021**, *59*, 7419–7433. [\[CrossRef\]](#)
12. Bai, T.; Wang, H.; Wen, B. Targeted universal adversarial examples for remote sensing. *Remote Sens.* **2022**, *14*, 5833. [\[CrossRef\]](#)
13. Deng, B.; Zhang, D.; Dong, F.; Zhang, J.; Shafiq, M.; Gu, Z. Rust-style patch: A physical and naturalistic camouflage attacks on object detector for remote sensing images. *Remote Sens.* **2023**, *15*, 885. [\[CrossRef\]](#)
14. Zheng, S.; Han, D.; Lu, C.; Hou, C.; Han, Y.; Hao, X.; Zhang, C. Transferable Targeted Adversarial Attack on Synthetic Aperture Radar (SAR) Image Recognition. *Remote Sens.* **2025**, *17*, 146. [\[CrossRef\]](#)
15. Dang, Q.; Zhan, T.; Gong, M.; He, X. Boosting Adversarial Transferability by Batch-wise Amplitude Spectrum Normalization. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5608114. [\[CrossRef\]](#)
16. Xu, Y.; Ghamisi, P. Universal adversarial examples in remote sensing: Methodology and benchmark. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [\[CrossRef\]](#)
17. Lin, G.; Pan, Z.; Zhou, X.; Duan, Y.; Bai, W.; Zhan, D.; Zhu, L.; Zhao, G.; Li, T. Boosting adversarial transferability with shallow-feature attack on SAR images. *Remote Sens.* **2023**, *15*, 2699. [\[CrossRef\]](#)
18. Fu, Y.; Liu, Z.; Lyu, J. Transferable adversarial attacks for remote sensing object recognition via spatial-frequency co-transformation. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5636812. [\[CrossRef\]](#)
19. Zhang, J.; Huang, Y.; Wu, W.; Lyu, M.R. Transferable adversarial attacks on vision transformers with token gradient regularization. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 16415–16424.
20. Deng, H.; Fang, Y.; Huang, F. Enhancing Adversarial Transferability on Vision Transformer by Permutation-Invariant Attacks. In Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME), Niagara Falls, ON, Canada, 15–19 July 2024; pp. 1–6.
21. Du, C.; Li, Y.; Xu, C. Understanding robustness of visual state space models for image classification. *arXiv* **2024**, arXiv:2403.10935.
22. Hu, Q.; Shen, Z.; Sha, Z.; Tan, W. Multiloss Adversarial Attacks for Multimodal Remote Sensing Image Classification. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 4600813. [\[CrossRef\]](#)
23. Xia, G.S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L.; Lu, X. AID: A benchmark data set for performance evaluation of aerial scene classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [\[CrossRef\]](#)
24. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the 2016 International Conference on Computer, Information and Telecommunication Systems (Cits), Kunming, China, 6–8 July 2016; pp. 1–5.
25. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
26. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* **2013**, arXiv:1312.6199.
27. Brendel, W.; Rauber, J.; Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv* **2017**, arXiv:1712.04248.
28. Biggio, B.; Fumera, G.; Roli, F. Security evaluation of pattern classifiers under attack. *IEEE Trans. Knowl. Data Eng.* **2013**, *26*, 984–996. [\[CrossRef\]](#)
29. Liu, D.; Yu, R.; Su, H. Adversarial shape perturbations on 3d point clouds. In Proceedings of the Computer Vision–ECCV 2020 Workshops, Glasgow, UK, 23–28 August 2020; Proceedings, Part I 16; Springer: Berlin/Heidelberg, Germany, 2020; pp. 88–104.
30. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
31. Kurakin, A.; Goodfellow, I.; Bengio, S. Adversarial examples in the physical world. *arXiv* **2016**, arXiv:1607.02533.
32. Madry, A. Towards deep learning models resistant to adversarial attacks. *arXiv* **2017**, arXiv:1706.06083.
33. Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; Li, J. Boosting adversarial attacks with momentum. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 9185–9193.
34. Chen, J.; Wu, X.; Guo, Y.; Liang, Y.; Jha, S. Towards evaluating the robustness of neural networks learned by transduction. *arXiv* **2021**, arXiv:2110.14735.

35. Su, J.; Vargas, D.V.; Sakurai, K. One pixel attack for fooling deep neural networks. *IEEE Trans. Evol. Comput.* **2019**, *23*, 828–841. [\[CrossRef\]](#)
36. Wang, Y.; Liu, J.; Chang, X.; Rodríguez, R.J.; Wang, J. Di-aa: An interpretable white-box attack for fooling deep neural networks. *Inf. Sci.* **2022**, *610*, 14–32. [\[CrossRef\]](#)
37. Ilyas, A.; Engstrom, L.; Athalye, A.; Lin, J. Black-box adversarial attacks with limited queries and information. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 2137–2146.
38. Wang, Z.; Hao, Z.; Wang, Z.; Su, H.; Zhu, J. Cluster attack: Query-based adversarial attacks on graphs with graph-dependent priors. *arXiv* **2021**, arXiv:2109.13069.
39. Wang, Z.; Guo, H.; Zhang, Z.; Liu, W.; Qin, Z.; Ren, K. Feature importance-aware transferable adversarial attacks. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 7639–7648.
40. Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; Swami, A. Practical black-box attacks against machine learning. In Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security, Abu Dhabi, United Arab Emirates, 2–6 April 2017; pp. 506–519.
41. Liu, Y.; Chen, X.; Liu, C.; Song, D. Delving into transferable adversarial examples and black-box attacks. *arXiv* **2016**, arXiv:1611.02770.
42. Xiong, Y.; Lin, J.; Zhang, M.; Hopcroft, J.E.; He, K. Stochastic variance reduced ensemble adversarial attack for boosting the adversarial transferability. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14983–14992.
43. Huang, H.; Chen, Z.; Chen, H.; Wang, Y.; Zhang, K. T-sea: Transfer-based self-ensemble attack on object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 20514–20523.
44. Xie, C.; Zhang, Z.; Zhou, Y.; Bai, S.; Wang, J.; Ren, Z.; Yuille, A.L. Improving transferability of adversarial examples with input diversity. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 2730–2739.
45. Wang, K.; He, X.; Wang, W.; Wang, X. Boosting adversarial transferability by block shuffle and rotation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 16–22 June 2024; pp. 24336–24346.
46. Zhang, J.; Wu, W.; Huang, J.T.; Huang, Y.; Wang, W.; Su, Y.; Lyu, M.R. Improving adversarial transferability via neuron attribution-based attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14993–15002.
47. Wang, M.; Wang, J.; Ma, B.; Luo, X. Improving the transferability of adversarial examples through black-box feature attacks. *Neurocomputing* **2024**, *595*, 127863. [\[CrossRef\]](#)
48. Xu, Y.; Du, B.; Zhang, L. Assessing the threat of adversarial examples on deep neural networks for remote sensing scene classification: Attacks and defenses. *IEEE Trans. Geosci. Remote Sens.* **2020**, *59*, 1604–1617. [\[CrossRef\]](#)
49. Zhang, Y.; Zhang, Y.; Qi, J.; Bin, K.; Wen, H.; Tong, X.; Zhong, P. Adversarial patch attack on multi-scale object detection for UAV remote sensing images. *Remote Sens.* **2022**, *14*, 5298. [\[CrossRef\]](#)
50. Czaja, W.; Fendley, N.; Pekala, M.; Ratto, C.; Wang, I.J. Adversarial examples in remote sensing. In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 6–9 November 2018; pp. 408–411.
51. Zhang, Y.; Gong, Z.; Liu, W.; Wen, H.; Wan, P.; Qi, J.; Hu, X.; Zhong, P. Empowering Physical Attacks with Jacobian Matrix Regularization against ViT-based Detectors in UAV Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5628814. [\[CrossRef\]](#)
52. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
53. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
54. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
55. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
56. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 10–15 June 2019; pp. 6105–6114.
57. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929.
58. Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; Guo, B. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 10012–10022.

59. Tu, Z.; Talebi, H.; Zhang, H.; Yang, F.; Milanfar, P.; Bovik, A.; Li, Y. Maxvit: Multi-axis vision transformer. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; Springer: Berlin/Heidelberg, Germany, 2022; pp. 459–479.
60. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International Conference on Machine Learning, Online, 18–24 July 2021; pp. 8748–8763.

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