remote sensing

MDPI

Article

Adversarial Robustness Enhancement of UAV-Oriented Automatic Image Recognition Based on Deep Ensemble Models

Zihao Lu, Hao Sun * and Yanjie Xu

College of Electronic Science, National University of Defense Technology, Changsha 410073, China; luzihao21@nudt.edu.cn (Z.L.); xuyanjie@nudt.edu.cn (Y.X.)
* Correspondence: sunhao@nudt.edu.cn

Abstract: Deep neural networks (DNNs) have been widely utilized in automatic visual navigation and recognition on modern unmanned aerial vehicles (UAVs), achieving state-of-the-art performances. However, DNN-based visual recognition systems on UAVs show serious vulnerability to adversarial camouflage patterns on targets and well-designed imperceptible perturbations in real-time images, which poses a threat to safety-related applications. Considering a scenario in which a UAV is suffering from adversarial attack, in this paper, we investigate and construct two ensemble approaches with CNN and transformer for both proactive (i.e., generate robust models) and reactive (i.e., adversarial detection) adversarial defense. They are expected to be secure under attack and adapt to the resource-limited environment on UAVs. Specifically, the probability distributions of output layers from base DNN models in the ensemble are combined in the proactive defense, which mainly exploits the weak adversarial transferability between the CNN and transformer. For the reactive defense, we integrate the scoring functions of several adversarial detectors with the hidden features and average the output confidence scores from ResNets and ViTs as a second integration. To verify their effectiveness in the recognition task of remote sensing images, we conduct experiments on both optical and synthetic aperture radar (SAR) datasets. We find that the ensemble model in proactive defense performs as well as three popular counterparts, and both of the ensemble approaches can achieve much more satisfactory results than a single base model/detector, which effectively alleviates adversarial vulnerability without extra re-training. In addition, we establish a one-stop platform for conveniently evaluating adversarial robustness and performing defense on recognition models called AREP-RSIs, which is beneficial for the future research of the remote sensing field.

Keywords: deep neural network; adversarial defense; deep ensemble model; unmanned aerial vehicle; remote sensing; image recognition

Citation: Lu, Z.; Sun, H.; Xu, Y. Adversarial Robustness Enhancement of UAV-Oriented Automatic Image Recognition Based on Deep Ensemble Models. *Remote Sens.* **2023**, *15*, 3007. https://doi.org/10.3390/rs15123007

Academic Editor: Gwanggil Jeon

Received: 27 April 2023
Revised: 31 May 2023
Accepted: 7 June 2023
Published: 8 June 2023

1. Introduction

Over the past several decades, an abundance of remote sensing images (RSIs) have been continuously collected from UAVs with massive and detailed information that allows researchers to observe the Earth more precisely. Nevertheless, the mode of image interpretation, which relies only on expert knowledge and handcrafted features, can no longer meet the requirements of higher accuracy and efficiency. Fortunately, the substantial progress of DNNs [1] in computer vision has achieved the state-of-the-art performances in the various tasks of remote sensing field and supported on-device inference for the real-time demands. Well-trained DNNs can be deployed on UAVs for the tasks including image recognition, object detection, image matching and so on, which enables quick feedback with useful analysis for both military (e.g., target acquisition [2–5], battlefield reconnaissance [6], communications [7–9]) and civilian (e.g., land surveys [10], delivery service [11,12], medical rescue [13,14]) use.

However, hidden dangers lurk in the working process of UAV, and a great diversity of counter-UAV attacks have been extensively developed that are targeted at its vulnerability,

which mainly exists in the cyber, sensing, and kinetic domains [15]. Distribution drifts [16–18] and common corruptions such as blur, weather and noise [19] also interfere with the automatic interpretations of RSIs in the image domain. Meanwhile, a new kind of threat has emerged due to the security and reliability issues with DNN models [20–22], which is known as *adversarial vulnerability* and potentially has devastating effects on the UAVs with autonomous visual navigation and recognition systems. For example, when such a UAV carries out a target recognition task, particularly for the non-cooperative vehicles on military missions, the suspicious vehicles with carefully designed camouflage patterns (i.e., physical adversarial attacks) or a leakage of real-time images with malicious perturbations (i.e., digital adversarial attacks) can mislead the DNNs on UAVs to make wrong predictions and violate the integrity of the outputs. In this way, the enemy's targets are likely to evade the automatic recognition, causing a severe disadvantage to the battlefield reconnaissance. Thus, the harmful effects of adversarial vulnerability in DNN models need to be taken more seriously for modern UAVs. Moreover, compared with the natural images such as ImageNet [23], not as many RSIs are labeled in a dataset. Therefore, the trained DNNs in the remote sensing field tend to be sensitive to adversarial attacks [24], which puts forward a higher requirement on the adversarial robustness.

Under threat from the adversarial attacks, researchers are motivated to propose effective defense methods mainly in the context of natural images. The defense strategies can be divided into two categories. The first is *proactive defense* to generate robust DNNs aimed at correctly classifying all the attacked images. Adversarial training (AT) [25] is a commonly used approach belonging to this category, which minimizes the training loss with online-generated adversarial examples. However, standard AT counts on prior knowledge with no awareness of new attacks and can decrease the accuracy of benign data. So, many improved versions such as TRADES [26], FAT [27], and LAS-AT [28] are developed. In addition, an attack designed for one DNN model may not confuse another DNN, which makes ensemble methods [29–32] an attractive defense strategy while bridging the gap between benign and adversarial accuracy. Ensemble methods against adversarial attacks often combine the output predictions or fuse the features extracted from the intermediate layers of several DNNs.

However, given the fact that obtaining a sufficiently robust DNN against any kind of attack is not realistic, some research efforts have been turned to *reactive defense*, namely detecting the input image whether it has been attacked or not. The detection strategy can be classified into three categories, including statistical [33–38], prediction inconsistency-based [39,40] and auxiliary model [41–44] strategies. In reactive defense, we do not modify the original victim models during the detection and train a detector with a certain strategy as a 3rd-party entity. Moreover, the reactive defense is valuable when the output of a baseline DNN does not agree with the one from a robust DNN strengthened by a proactive defense method [45].

In this article, we consider the case that the DNN-based visual navigation and recognition systems on UAVs are suffering from adversarial attacks when performing an important task after take-off. Aimed at this intractable scenario and several analyzed motives, we propose to investigate the ensemble strategy to address the problem for both proactive and reactive defense **only using base DNN models**:

- In proactive defense, standard AT and its variants need re-training and model updates if UAVs meet unknown attacks, which does not suit the environment of edge devices with limited resources (e.g., latency, memory, energy); thus, an ensemble of base DNN models can be an alternative strategy. Intuitively, an ensemble is expected to be more robust than an individual model, as the adversary needs to fool the majority of the sub-models. As the representative models of CNNs and transformers, ResNet [46] and Vision Transformer (ViT) [47] have different network architectures and mechanisms in extracting discriminative features. We also verify that the adversarial examples of RSIs show weak transferability between CNNs and transformers. Therefore, we combine the probability distributions of output layers from CNNs and transformers with standard

supervised training for a better performance under adversarial attacks in the recognition of RSIs.

- In terms of reactive defense, we consider a case study with the framework of ENsemble Adversarial Detector (ENAD) [48], which combines scoring functions computed by multiple adversarial detection algorithms with intermediate activation values in a well-trained ResNet. Based on the original framework, we further integrate the scoring functions from ViT with the ones from ResNet, forming a connection with the ensemble method in proactive defense. Therefore, the ensemble has two levels of meaning: one is combining layer-specific values from multiple adversarial detection algorithms, and the other is integrating the results from CNNs and transformers. Different detection algorithms with different network architectures can exploit distinct statistical features of the images, so this ensemble strategy is highly suitable for RSIs with rich information.

Both of the defenses in the form of an ensemble will be activated when the controller realizes that the outputs from the system on UAVs are obviously manipulated. The supposed scenarios and the role of ensemble defense are illustrated as Figure 1. To verify their effectiveness, we conduct a series of experiments with the datasets including optical and SAR RSIs. For proactive defense, we compare the performances regarding the *Attack Success Rate* of an ensemble of base ResNets and ViTs for different adversarial attack algorithms with three other proactive defense to improve the robustness of base DNN models. In terms of reactive defense, we compare the ensemble framework with three stand-alone adversarial detectors, which are also the components in the ensemble framework. The metrics of detection are the Area Under the Receiver Operating Curve (AUROC) and the Area Under Precision Recall (AUPR).

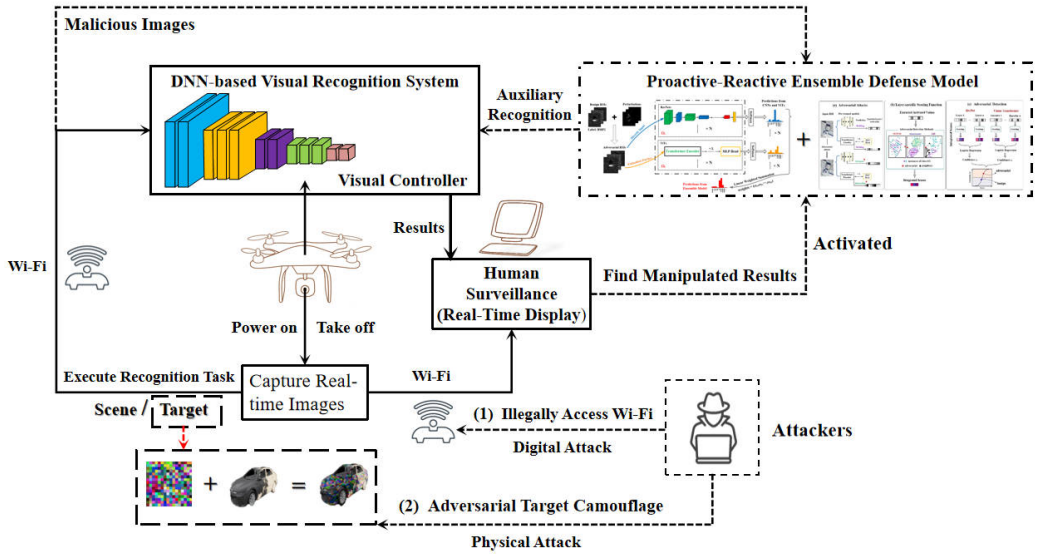

Figure 1. The threat scenarios caused by adversarial vulnerability in modern UAVs and the role of our adversarial ensemble defense (blue lines: general working mode of UAVs; red lines: confrontation with adversarial attacks).

From the experimental results, we find that an ensemble of base ResNets and ViTs demonstrates good defensive capability in most experimental configurations of proactive defense. It does not need a re-training but can be on a par with the methods based on AT. Moreover, an ensemble framework modified from ENAD can yield AUROC and AUPR of over 90 in gradient-based attacks of optical datasets. The performances of the ensemble method slightly decrease on Deepfool, C&W and adversarial examples of SAR RSIs, but it is still generally better than the stand-alone adversarial detectors.

Based on the above work, we establish a one-stop integrated platform for evaluating the adversarial robustness of DNNs trained with optical or SAR RSIs and conducting adversarial defenses on the models called *Adversarial Robustness Evaluation Platform for*

Remote Sensing Images (AREP-RSIs). Users can operate just on AREP-RSIs to perform a complete robustness evaluation with all necessary procedures, including training, adversarial attacks, tests for recognition accuracy, proactive defense and reactive defense. AREP-RSIs can be deployed on the edge devices such as UAVs and connected with cameras for real-time recognition as well. Equipped with various network architectures, several training paradigms, and classical defense methods, to the best of our knowledge, AREP-RSIs is the first platform for adversarial robustness improvements and evaluations in the remote sensing field. More importantly, the framework of AREP-RSIs is flexibly extendable. Users can add the model architecture files, load their own weight configurations, and register new attack and defense methods for a customized DNN, which greatly facilities designing robust DNN-based recognition models in the remote sensing field for the future research. The AREP-RSIs can be available at Github (https://github.com/ZeoLuuuuuu/AREP-RSIs, accessed on 26 April 2023).

In summary, the main contributions of this paper are as follows.

- We innovatively analyze the adversarial vulnerability from a scenario in which the edge-deployed DNN-based system for visual navigation and recognition on a modern UAV is suffering from adversarial attacks produced by the physical camouflage patterns or digital imperceptible perturbations.
- To cope with the intractable condition, we investigate the ensemble of ResNets and ViTs for both proactive and reactive defense for the first time in the remote sensing field. We conduct experiments with optical and SAR remote sensing datasets to verify that the ensemble strategies have good efficacy and show a favorable prospect against adversarial vulnerability in the DNN-based visual recognition task.
- We finally integrate all the procedures of performing adversarial defenses and evaluating adversarial robustness into a platform called AREP-RSIs. Equipped with various network architectures, several training paradigms, and defense methods, users can verify if a specific model has good adversarial robustness or not just through this one-stop platform AREP-RSIs.

The rest of this paper is organized as follows. Section 2 introduces the background knowledge, related works and threat model utilized in this article. Section 3 tells why we use the ensemble strategy, specific methods and our developed platform in detail. Section 4 reports on the experimental results and provides an analysis. Finally, the conclusions are given in Section 5.

2. Background and Related Works

This section briefly reviews the causes of adversarial vulnerability in image recognition tasks and existing research of the adversarial vulnerability in the remote senisng field and DNN-based UAVs. Finally, we provide a threat model including the potential approaches of attacking the automatic recognition systems of UAVs with adversarial examples, some possible goals and the access level of models for attackers.

2.1. Causes of Adversarial Vulnerability in Image Recognition

To better learn the adversarial vulnerability in an image recognition system, its possible causes are discussed theoretically. Sun et al. [49] give a comprehensive analysis, and based on their work, we briefly review the reasons why adversarial vulnerability is a common problem for image recognition.

- **Dependency on Training Data**: The accuracy and robustness of an image recognition model are highly dependent on the quantity and quality of training data. During the training process, DNN models only learn the correlations from data, which tend to vary with data distribution. In many security-sensitive fields, the severe scarcity of large-scale high-quality training data and the problem of category imbalance in the training datasets can exacerbate the risk of adversarial vulnerability of DNN models.
- **High-Dimensionality of Input Space**: The training dataset only covers a very small part of the input space portion, and a large amount of potential input data are not

utilized. Moreover, hundreds of parameters are optimized during the training process, and the space formed by parameters is also huge. Therefore, the generalized decision boundaries in the input space are just roughly approximated by DNNs, which cannot completely overlap with the ground-truth decision boundaries. The adversarial examples may exist in the gap between them.

- **Black-box property of DNNs**: Due to the complex network architectures and optimization process, it is hard to directly translate the internal representation of a DNN into a tool for understanding its wrong outputs under an adversarial attack. So, this black-box property of DNNs makes it more difficult to design a universal defense technique against adversarial perturbations from the perspective of the model itself.

2.2. Adversarial Vulnerability in DNN-based UAVs

In recent years, as DNNs are increasingly applied to the visual navigation and recognition systems on UAVs, the security threat produced by adversarial attacks has been a formidable problem, which can be utilized by the attackers with motives for maliciously permeating into the working process of these DNN-based UAVs.

Previous research has indicated that this security problem exists in DNN models for RSI recognition, which poses a threat to the modern UAVs. Most of them still focus on the digital attacks, which directly manipulate the pixel values in RSIs and suppose full access to the images for attackers. In terms of scene recognition, Li et al. [50] and Xu et al. [51] both used various adversarial attacks to fool multiple high-accuracy models trained on different scene datasets. In another article, Xu et al. also provided a black-box universal dataset with adversarial examples called UAE-RS [52], which serve as a benchmark to design DNNs with higher robustness. Even further, Li et al. [53] proposed a soft threshold defense for scene recognition to judge whether an input RSI is adversarial or not. Focused on SAR target recognition, Li et al. [54] mounted white-box attacks on SAR images and proposed a new metric to successfully explain the phenomenon of attack selectivity. Du et al. [55] proposed a fast C&W algorithm for DNN-based SAR classifiers, using a deep coded network to replace the search process in the original C&W algorithm. Zhou et al. [56] focused on the sparsity of SAR images and applied the sparse attack methods on the MSTAR dataset to verify their effectiveness in SAR target recognition.

In addition, there are also explorations into physical adversarial attacks applied to RSIs. Czaja et al. [57] conducted attacks through adversarial patches to confuse the victim DNN among four scene classes, and den Hollander et al. [58] generated the patches for the task of object detection. However, they only restricted their patches to the digital domain and did not print them. The most relevant to our assumed scenario is the work of Du et al. [59], in which they optimized, fabricated and installed their designed patches on or around a car to significantly reduce the efficacy of a DNN-based detector on a UAV. They also experimented under different atmospheric factors (lighting, weathers, seasons) and distance between the camera and target. Their results indicated the realistic threat of adversarial vulnerability on DNN-based intelligent systems on UAVs.

Moreover, some research has discussed the adversarial vulnerability in the context of UAVs. Doyle et al. [15] considered two common operations for a navigation system of UAVs: follow and waypoint missions to develop a threat model from the perspective of attackers. They sketched state diagrams and analyzed the potential attacks for each state transition. Torens et al. [60] give a comprehensive review for the verification and safety of machine learning in UAVs. Tian et al. [61] proposed two adversarial attacks for the regression problems of predicting steering angles and collision probabilities from real-time images in UAVs. They also investigated standard AT and defensive distillation against the two designed attacks.

2.3. Threat Model

We denote a real-time image captured and processed by the sensors as $x \in \mathbb{R}^{h \times w \times c}$ with h, w, c representing height, width and channel (c = 3 for optical images and c = 1

for SAR images), which is also the input of a DNN-based visual recognition system $\mathcal{M}(\cdot)$ deployed on UAVs. In addition, each image has a potential groundtruth label $y \in \mathcal{Y} = \{0, 1, \ldots, K-1\}$ where K is the number of recognizable categories for the system. A well-trained system $\mathcal{M}(\cdot)$ can correctly recognize the scene or targets for most of x, namely $\mathcal{M}(x) = y$.

We suppose two possible approaches that attackers can exploit to attack the DNN-based visual recognition system on UAVs.

(1) The first approach is to illegally access the Wi-Fi communication between the sensors (i.e., cameras) and the controller for UAVs. The attackers can spoof imperceptible perturbations ρ to the images provided by the sensors to craft adversarial examples $\hat{x} = x + \rho$ through the communication link. The wrong predictions $y' = \mathcal{M}(\hat{x}) \neq y$ for most of $\hat{x}$ can influence the next commands and actions for UAVs.

(2) The second approach is physically realizing the perturbations as "ground camouflage" based on adversarial patches [62], especially for the task of target recognition. An adversarial patch is generally optimized in the form of sub-images by modifying the pixel values within a confined area, and the attacker then prints the patch as a sticker or poster. Ref. [59] gives a real-world experiment for this approach by pasting designed patches on top of or around vehicles to highly reduce the probabilities of detection and recognition rates. Even if the patterns are noticeable to our human eyes, they can effectively confuse the recognition system.

There are several reasons why attackers hope to do harm to the visual navigation and recognition system on a UAV. For scene recognition, attackers can mislead UAVs to incorrect situational awareness for military use. In addition, the misclassification of the scene may make the navigation system confuse the current environment, become lost, and hover in the air. For target recognition, once non-cooperative targets of high military value are camouflaged, UAVs will not be able to accurately detect and recognize them, which aims at evading aerial reconnaissance or targeted strikes in the battlefield.

The access level of the victim DNN models for attackers is an important factor. White-box attackers are the strongest in all conditions. They can obtain the network structures, weights and even the training data. In contrast, black-box attackers only query the outputs at each attempt, craft adversarial examples against a substitute model or search randomly. Moreover, whether they mislead DNNs to a specified class distinguishes an attack as a targeted or untargeted one. In our threat model, we consider both white-box and black-box settings during our experiments with the more general untargeted condition.

3. Methodology

This section will briefly analyze the motives of exploiting the ensemble strategy in Section 3.1. Then, it will present the proactive–reactive defensive ensemble framework in detail in Section 3.2 and finally introduce our edge-deployed platform AREP-RSIs for adversarial robustness improvements and evaluations in Section 3.3.

3.1. Motives of Ensemble

As the most representative models of CNNs and transformers, ResNet and ViT are mainly discussed within the defensive ensemble framework. Before a detailed description of the defense method, we start with the reasons why the ensemble strategy should be selected and attempted in the supposed scenario of this article.

3.1.1. Different Mechanisms for Feature Representations

Recently, ViTs have drawn great attention as a fundamentally new model structure offering impressive performances in image recognition and robustness benefits as well [63]. Compared with CNNs, ViTs have striking differences in their feature representations [64].

Specifically, CNNs share kernels in each convolution layer (Conv) that locally perceive a small part of the input image (i.e., receptive field) to extract features. The powerful inductive bias of translation equivariance and locality correlation within the convolutional

layers make CNNs excellent in learning general-purpose visual representations. However, the receptive fields are limited with a fixed size, which is not conducive to obtaining global information. In contrast, ViT processes an image as a sequence of image patches, and each patch is linearly projected into a representation vector with a positional embedding. Moreover, a learnable class token is also attached for the image. As the main component in ViT, multi-head self-attention modules (MSAs) are then connected for an aggregation of the information from all patches to have an entire view of the image.

More importantly, [65] revealed that the MSAs in ViT exhibit opposite behaviors with the Convs in ResNet by performing the Fourier analysis of feature maps from both models. The Convs act like a low-pass filter that tends to reduce low-frequency signals, while MSAs are high-pass filters that are robust against high-frequency noise in images. In addition, [64] found that ViT incorporates more global information and has more uniform representations with greater similarity throughout the layers. There have been many hybrid architectures that combine CNNs and transformers to inherit both of their advantages [66–69]. Therefore, to some extent, ViT can be complementary to ResNet, which intuitively enlightens us about the selection of network architectures in the ensemble.

3.1.2. Weak Adversarial Transferability

Reducing the adversarial transferability among base models in an ensemble can achieve good robustness without sacrificing benign accuracy [70–72]. To further verify the differences between ResNet and ViT in the context of remote sensing, we found empirical evidence that the adversarial examples of RSIs tend to have weak transferability between CNNs and ViTs, which facilities constructing ensemble classifiers to generate a more robust model. For the details of transferability experiments, we trained a set of various CNNs (including ResNet-18, ResNet-50, DenseNet-121, DenseNet-201, MobileNet-V2 and ShuffleNet-V2) and two ViT variants (ViT-Base/16 and ViT-Large/16) with the same training setting on the MSTAR dataset. A white-box attack, PGD-$ł_\infty$ [25], is applied on the test set of MSTAR with a different attack radius against the victim models of ResNet-18 and ViT-Base/16, respectively. Then, both sets of generated adversarial examples are recognized under each well-trained DNN model. Similarly, we conducted the experiments on the UC Merced LandUse, which is an optical scene RSI dataset again. The results are illustrated in Figures 2 and 3. From the results, for both datasets, we observe that the adversarial examples crafted against ResNet-18 generally have much better performance of recognition accuracy in ViTs and vice versa.

3.1.3. Defects in AT and Our solution for Edge Environment

One of the most commonly used methods for improving adversarial robustness is still AT, which trained DNNs with both natural data and its corresponding adversarial variants. Even though previous research indicated that AT can force DNNs to learn robust features and gained better performance on adversarial robustness, the absence of non-robust features can lead to a drop in generalization and the accuracy on the benign data [73]. This trade-off between adversarial robustness and natural accuracy still needs to be considered when using AT. Moreover, AT sometimes heavily counts on such prior knowledge and cannot achieve a sufficient robustness against an unknown attack.

Generally, modern UAVs are equipped with different base DNN models instead of the DNN models trained with AT for standard automatic visual recognition. When the UAVs suffer from adversarial attacks in performing a recognition task, it is time-consuming to make an extra re-training to obtain a new robust model and replace the base models on the ground. Training on edge devices is also impractical because of the resource-limited environment. Therefore, our proposed solution for this problem is attempting an ensemble of base DNN models, especially DNNs with different network architectures and feature extraction mechanisms. Based on the analysis of CNNs and transformers above, we decide to use ResNet and ViT, which are two standard popular DNN architectures in the ensemble.

They will be trained solely with benign data to improve adversarial robustness while guaranteeing natural accuracy in our supposed scenario.

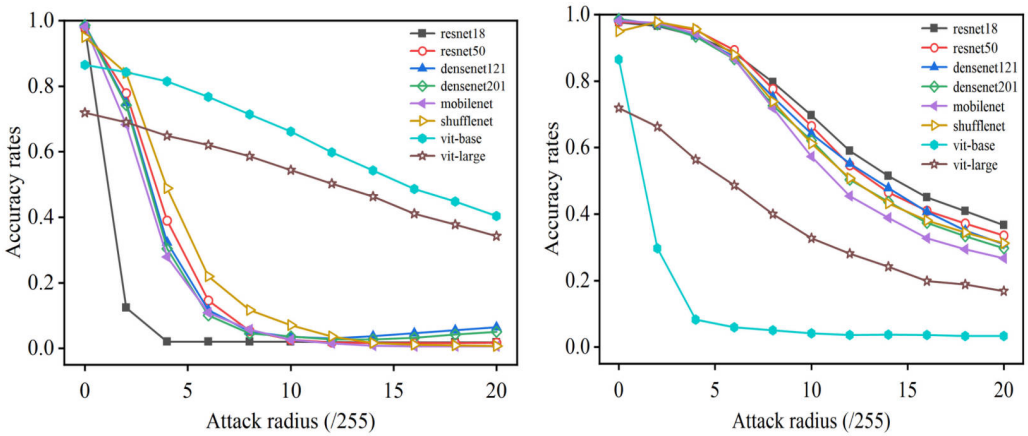

Figure 2. The transferability test of PGD on MSTAR against ResNet-18 (**left**) and ViT-Base/16 (**right**).

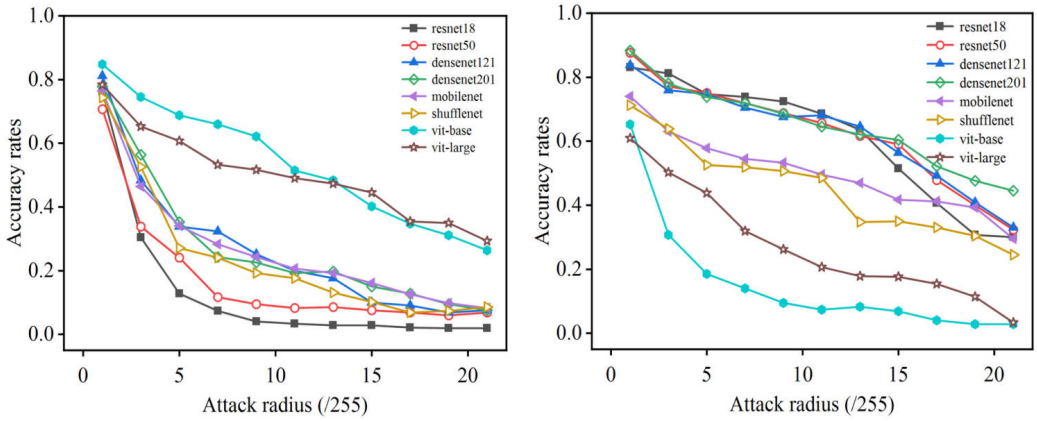

Figure 3. The transferability test of PGD on UC Merced LandUse against ResNet-18 (**left**) and ViT-Base/16 (**right**).

In addition, we report the computation and memory footprints of base DNN models used in our ensemble as shown in Table 1, including the number of parameters within network architecture (Params), floating point operations (FLOPs) and parameter memory footprint (Param. Mem). The specific network architectures consist of ResNet-18, ResNet-50, and ResNet-101 for CNN and ViT-Base/16, ViT-Large/16, and ViT-Base/32 for transformer.

Table 1. The computation and memory footprints of base DNN models used in our ensemble.

	Params (M)	FLOPs (GFLOPs)	Param. Mem (MB)
ResNet-18	11.69	2	45
ResNet-50	25.56	4	98
ResNet-101	44.55	8	170
ViT-Base/16	86.86	17.6	327
ViT-Base/32	88.30	8.56	336
ViT-Large/16	304.72	61.6	1053

As shown in Table 1, several of the network architectures we used such as ResNet-101 and ViT-Large/16 seem to be less suitable for the edge environment; however, our intention of this attempt is to first verify that the ensemble of CNNs and transformers can resist adversarial data of RSIs in both proactive and reactive defense. So, the two most commonly used DNNs are selected for the paradigm in our article. In practice, we can replace them

with more light-weight DNNs such as MobileNet, ShuffleNet, and Inception-V3 for CNN and ViT-Tiny/16, EfficientFormer for transformer.

3.2. Proactive–Reactive Defensive Ensemble Method

3.2.1. Proactive Defense

In the non-ensemble schemes, a single base model is provided to attackers, which can be attacked with the worst perturbations. However, based on the analyzed motives above, an ensemble of CNNs and transformers suits our supposed scenario better. Our defensive ensemble model includes both proactive and reactive defense. For proactive defense, an ensemble model is a weighted average of N random base ResNets with different depths denoted as $\Omega_1 = \{\mathcal{R}_1, \mathcal{R}_2, ..., \mathcal{R}_N\}$ and N base ViT variants denoted as $\Omega_2 = \{\mathcal{R}_{N+1}, \mathcal{R}_{N+2}, ..., \mathcal{R}_{2N}\}$. To confuse the whole ensemble model, an attacker has to design an attack against both types of DNN with more difficulty [74].

Specifically, we can train two sets of base DNNs including ResNets and ViTs with $N = 3$ (i.e., $\Omega_1 = \{$ResNet-18, ResNet-50, ResNet-101$\}$, $\Omega_2 = \{$ViT-Base/16, ViT-Large/16, ViT-Base/32$\}$). Ω_1 and Ω_2 form the overall set of base models Ω. We denote $\{\mathcal{D}_j\}_{j=1}^{2N}$ as a large set including the probability distributions $\{d_{jk}\}_{k=1}^{K}$ predicted by each base DNN model, where d_{jk} is the confidence score of category k predicted by the j^{th} base model and K denotes the number of recognizable categories. Therefore, the probability distribution for each base model can be expressed as (1).

$$\mathcal{D}_j = \{d_{j1}, d_{j2}, \ldots d_{jK}\},\ j = 1, 2, \ldots, 2N \tag{1}$$

Then, we can weight the $2N$ models with non-negative values $(\omega_1, \omega_2, ..., \omega_{2N})$ that add up to 1. Let a vector $\mathcal{W}$ denote these weights, and we can obtain a new probability distribution $\mathcal{D}'$ of the deep ensemble model with new confidence scores $\{d'_k\}_{k=1}^{K}$ by taking a linearly weighted summation as (2).

$$\mathcal{D}' = \mathcal{W} \cdot (\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_{2N})^T = (d'_1, d'_2, \ldots, d'_K) \tag{2}$$

In fact, the new probability distribution $\mathcal{D}'$ is a fusion on the decision level, integrating the opinions from CNNs and transformers. In addition, from the perspective of base models, we can also express the DNN-based ensemble model $\mathcal{M}(x, \mathcal{W})$ as (3).

$$\mathcal{M}(x, \mathcal{W}) = \mathcal{W} \cdot \Omega = \sum_{j=1}^{2N} \omega_j \cdot \mathcal{R}_j \tag{3}$$

The framework of this ensemble model for proactive defense is illustrated as Figure 4. As shown in Figure 4, if a modern UAV captures real-time RSIs with BMP2 vehicles but suffers from adversarial perturbations crafted for CNN architecture, these RSIs can be sent to the proposed deep ensemble model and inferred by all of the base models simultaneously. Even though the adversarial RSIs can mislead the predictions from CNNs, the outputs from transformers are still correct. The model will fuse the opinions of CNNs and transformers on the decision level, namely making a linear weighted summation as mentioned above, to obtain the final correct prediction.

In terms of the weights of base models $(\omega_1, \omega_2, \ldots, \omega_{2N})$ in the ensemble, one solution is to weight them with fixed values, and we can search for the better set of values manually. The other solution of deciding the models' weights is to make them learnable, so the weights can be adjusted automatically during the training time. In our following experiments, we choose the former for simplicity.

This deep ensemble model for proactive defense only exploits standard base models and does not need to require extra re-training such as AT, which constitutes a practical attempt for improving the adversarial robustness of automatic recognition systems on edge devices such as UAVs. It is also the first DNN-based ensemble model against adversarial

attacks in the remote sensing field. In this way, more adversarial examples are expected to be correctly recognized when confronting adversarial attacks. We will compare the ensemble of ResNets and ViTs with a victim model without any defense and trained with standard AT [25], Trades [26] and GAL [75] against malicious RSIs crafted by different adversarial attack methods. The experimental results will be collected in the next section. The Attack Success Rate (ASR) (i.e., the number of wrongly recognized RSIs divided by the number of RSIs in the whole test set) will be the metric for the proactive defense.

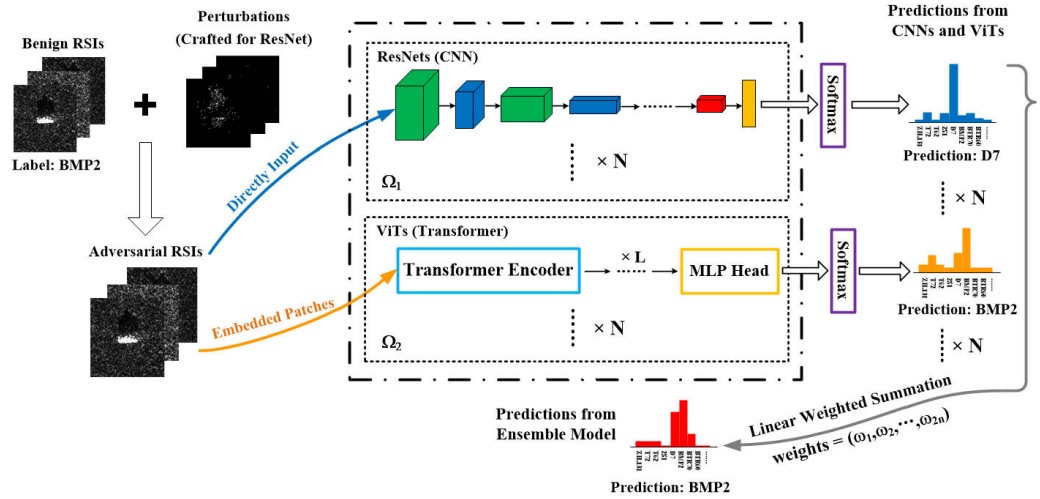

Figure 4. The process of proactive ensemble method with MSTAR dataset and victim model ResNet-18.

3.2.2. Reactive Defense

Considering the fact that it is not possible for proactive defense to classify all types of adversarial examples, detection-based methods (i.e., reactive defense) also deserve an exploration to indirectly enhance the adversarial robustness. To pursue better performance than individual adversarial detectors, we selected and modified an excellent deep ensemble framework called ENAD [48], which integrates the scoring functions from different adversarial detection algorithms on the hidden features of intermediate layers of CNNs. Our modified version repeats these procedures on ViTs with the features extracted from a transformer encoder and averages the detection outputs from both types of DNN at the end of the framework as the second integration. The structure of an ensemble including CNNs and transformers also matches our proposed model in the proactive defense.

The specific procedures of the ensemble model in reactive defense are illustrated in detail in Figure 5. A real-time RSI captured by UAVs, which can be either benign or maliciously attacked, is input to a well-trained ResNet and ViT. For ResNet, the activation values from several selected hidden layers are then extracted. Next, the model will compute layer-specific scores through three commonly used adversarial detection algorithms: Local Intrinsic Dimensionality (LID) [34], Kernel Density Estimation (KDE) [35] and Mahalanobis Distance (MD) [36]. Each detection algorithm measures the "distance" as the score based on each activation value of the real-time RSI with respect to training examples and the paradigm learned during the training time. The layer-specific scores for each detection algorithm are fused to obtain the detector-specific scores, namely three final scores, which are input to a logistic regression to compute a probability c_1 of classifying the test RSI as benign or adversarial. In the meantime, the above procedures are also performed in parallel on ViT with the activation values extracted from multi-head self-attention in several transformer encoders. The predicted probability from ViT is denoted as c_2. c_1 and c_2 from ResNet and ViT are averaged to obtain a final result p. The ensemble model will decide an RSI image as the adversarial one if c is greater than 0.5, and it is benign otherwise.

In terms of the individual detectors (i.e., LID, KDE and MD) in the ensemble model for reactive defense, there is one trick that needs to be considered. Apart from benign and

adversarial examples, we also craft noisy examples with Gaussian noise that are treated as benign examples during the training time for better generalization.

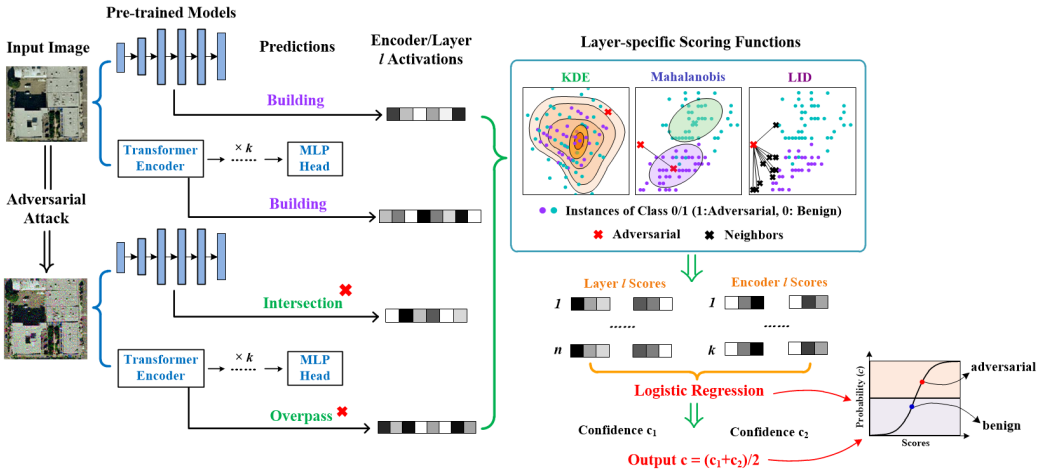

Figure 5. The procedures of reactive defense ensemble model modified from ENAD [48] (assuming the number of layers in ResNet is n and the number of transformer encoders in ViT is k).

The extracted activation values are high-dimensional in both types of DNN model, so different detection algorithms can use distinct statistical features of input images. The ensemble idea integrates the features and is expected to be perfectly suited for the adversarial detection of RSIs, because RSIs have rich information such as color, texture, spatial and spectral features. Moreover, the first integration of multiple adversarial detectors can better alleviate the problems of overfitting and generalization than just using one detector. The second integration benefits from different feature representations in ResNet and ViT. To evaluate the performances of detection, we take two standard metrics, Area Under the Receiver Operating Characteristic (AUROC) and Area Under Precision Recall (AUPR). The correctly detected adversarial and benign RSIs are true positives (TP) and true negatives (TN), respectively. On the contrary, the wrongly detected adversarial and benign RSIs are false negatives (FN) and false positives (FP), respectively.

3.3. Adversarial Robustness Evaluation Platform for Remote Sensing Images (AREP-RSIs)

Based on the above work, we further developed a one-stop platform for conducting adversarial defense and conveniently evaluating the adversarial robustness of a DNN-based visual recognition system on UAVs called *Adversarial Robustness Evaluation Platform for Remote Sensing Images* (AREP-RSIs). AREP-RSIs are multi-functional, and users can readily operate on this platform to evaluate the defensive performance for a DNN model trained with RSIs. In addition, if we load a well-trained DNN model, AREP-RSIs connected with cameras can predict the category of a scene or target for a real-time image and output the confidence scores in the main interface.

As shown in Figure 6, AREP-RSIs is built as a modular framework with 6 sub-modules including datasets, models, training, adversarial attack, test for recognition accuracy and adversarial defense. For example, the module of a test for recognition accuracy has two sub-models: single image test and batch images test. In the single image test, users can load an RSI and an arbitrary DNN model file to obtain the predicted category and the maximum confidence scores. If the selected RSI is detected as an adversarial example, the activated feature maps of this adversarial image and its corresponding original image are displayed. In the batch images test, a batch of RSIs is input to the selected DNN model, and the interface will show a confusion matrix to visualize the recognition performance. We can also know the recognition accuracy of this batch of RSIs.

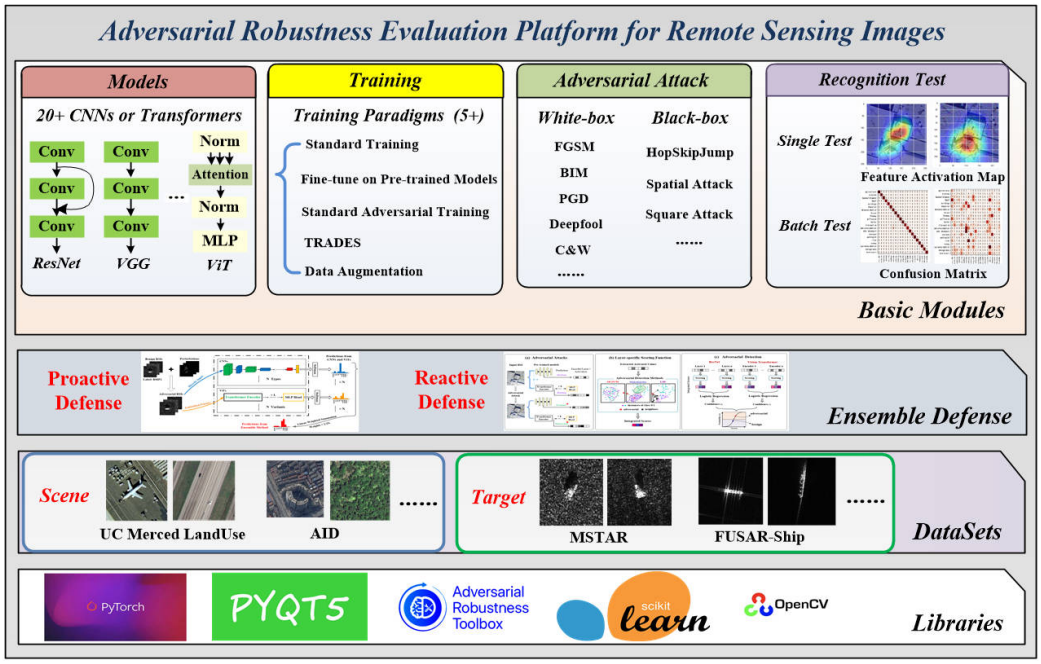

Figure 6. The overall framework of AREP-RSIs with 6 modules.

The graphic interface of this platform is designed with PyQt [76] and built upon necessary libraries such as Pytorch [77], Adversarial Robustness Toolbox (ART) [78], OpenCV [79] and Scikit-learn [80].AREP-RSIs includes several popular optical and SAR RSI datasets for scene/target recognition. We show the screenshots of a graphic interface of two modules in use, recognition test (single image test) and performing the proactive defense as shown in Figures 7 and 8.

Figure 7. The interface of a recognition test of a single RSI in AREP-RSIs.

Moreover, all of these modules are highly extendable, which greatly facilities designing robust DNN-based recognition models in the remote sensing field for future research. For instance, we can include other adversarial defense methods for RSI recognition, such as TRADES, GAL [75], and DVERGE [81] in proactive defense and FS [40] and DNR [42] in reactive defense into AREP-RSIs. New DNN model architecture, training paradigms and adversarial attacks can also be flexibly registered in AREP-RSIs for users to compare the adversarial robustness before and after performing a specific adversarial defense scheme to a base DNN model. In the current AREP-RSIs, we have embedded more than 20 types of DNNs with different training schemes and various mainstream adversarial attacks. We will

make the AREP-RSIs open source at Github (https://github.com/ZeoLuuuuuu/AREP-RSIs, accessed on 26 April 2023).

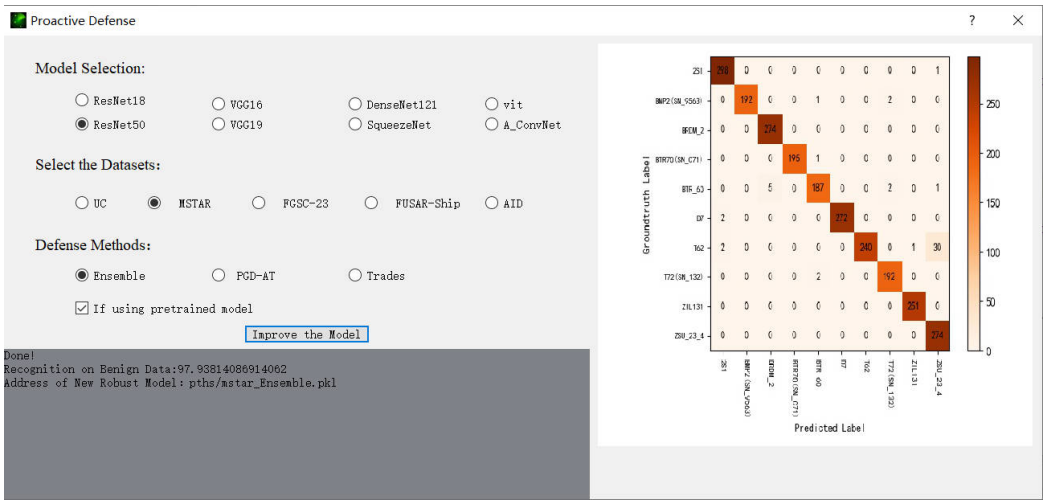

Figure 8. The interface of performing proactive defense to generate robust models in AREP-RSIs.

4. Experiments

4.1. Datasets

(1) Scene Recognition: Two high-quality datasets for scene classification, UCM [82] and AID [83], are selected for our experiments. Both of them include optical RSIs with scene only. The RSI examples for each dataset are illustrated in Figures 9 and 10.

Figure 9. RSI examples randomly selected for each class from UCM.

UCM: The UC Merced LandUse Dataset contains 2100 RSIs from 21 different land-use classes, each of which contains 100 256 × 256 images with a spatial resolution of 0.3 m per pixel in the RGB color space. The dataset is derived from the National Map Urban Area Imagery collection, which captures the scenes of nationwide towns across the United States.

AID: AID is a large RSI dataset that collects scene images from Google Earth. The dataset comprises 10,000 labeled RSIs containing 30 categories of scenes, approximately 200–420 images per category with an image size of 600 × 600 pixels. Even if the Google Earth images are post-processed using RGB renderings of the original aerial images, this does not affect its use in evaluating scene classification algorithms.

(2) Target Recognition: Two benchmark datasets for target recognition, MSTAR [84] and FUSAR-Ship [85], are also utilized in the experiments. The RSI examples for each dataset are illustrated in Figures 11 and 12.

Figure 10. RSI examples randomly selected for each class from AID.

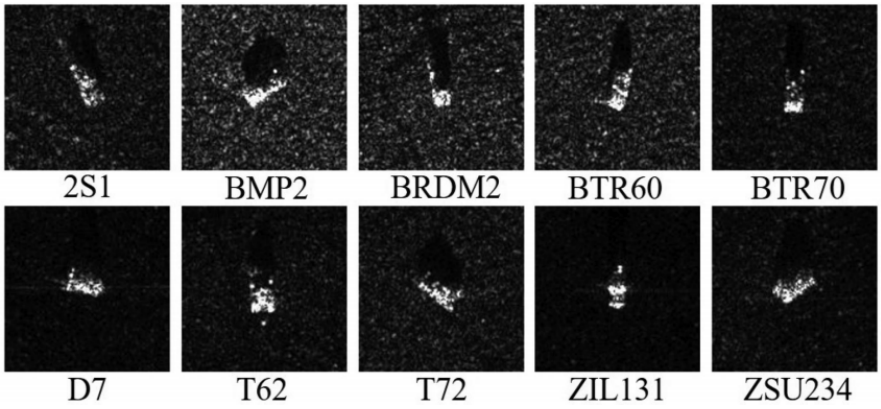

Figure 11. RSI examples randomly selected for each class from MSTAR.

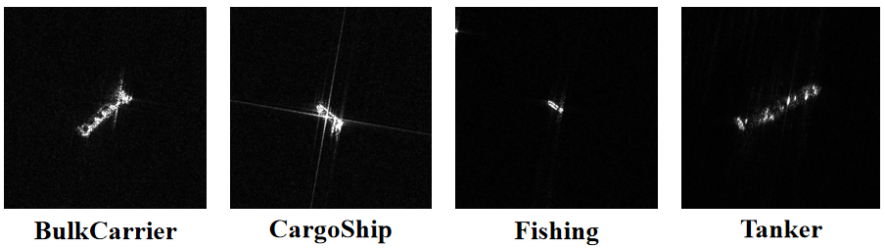

Figure 12. RSI examples randomly selected for each class from FUSAR-Ship.

MSTAR: MSTAR is from the publicly available Moving and Stationary Target Acquisition and Recognition (MSTAR) dataset produced by the US Defense Advanced Research Projects Agency. This dataset contains 5172 SAR sliced images of stationary vehicles with 10 categories acquired at various azimuths. The sensor is a high-resolution cluster SAR with a resolution of 0.3 m $\times$ 0.3 m, operating in the X-band.

FUSAR-Ship: FUSAR-Ship is a high-resolution SAR dataset obtained from GF-3 for ship detection and recognition. The maritime targets are divided into two branches, ship and non-ship. Here, we selected four sub-classes, bulk carrier, cargo ship, fishing and tanker from ship targets, collecting 420 images in total.

4.2. Experimental Setup and Results

We designed our experiment in a systematic manner to verify the adversarial robustness improvement of DNNs for RSI recognition after performing an ensemble strategy. In fact, our experiments include four procedures, which are training and testing base DNNs for recognition in RSIs, performing adversarial attacks with RSIs against the base models, improving adversarial robustness with the proactive ensemble model and detecting adversarial examples with the reactive ensemble model. All the experiments are implemented on a server equipped with an Intel Core i9-12900KF 3.19 GHz CPU, 32 GB of RAM and one NVIDIA GeForce RTX 3090 Ti GPU (24 GB Video RAM). The deep learning framework is Pytorch 1.8. All of the above experiments can be performed on the one-stop integrated platform AREP-RSIs, which makes it greatly convenient for users to evaluate the defensive effectiveness and adversarial robustness.

In this part, we collected all the quantitative results presented in the form of a graph or table, and in the following part, we analyzed the results adequately to verify if the ensemble models for both proactive and reactive defense are effective for the RSI recognition task.

In the first part, the training sets are randomly selected with 80% labeled images in each dataset, and the remaining images make up the test set. The trained base models are also the components in the following proactive ensemble model including ResNet-18, ResNet-50, ResNet-101, ViT-Base/16, ViT-Base/32 and ViT-Large/16. We train all models for 100 epochs with batch size = 32, and the optimizer as Adam [86]. We collected the recognition accuracy of the test set for these base models, as shown in Table 2.

Table 2. Recognition accuracy of base DNN models for test set of RSI datasets (the values below are averaged from 10 repeated experiments).

	UCM	AID	MSTAR	FUSAR-Ship
ResNet-18	96.19%	95.65%	94.80%	**81.40**%
ResNet-50	**96.67**%	96.05%	**97.73**%	80.95%
ResNet-101	92.38%	**97.90**%	93.32%	77.91%
ViT-Base/16	94.80%	92.70%	88.21%	79.76%
ViT-Base/32	91.80%	91.20%	82.64%	76.19%
ViT-Large/16	87.38%	93.34%	88.08%	78.57%

In terms of adversarial attacks, both white-box and black-box conditions are considered. Specifically, we choose 4 white-box and 2 black-box attack algorithms including the Fast Gradient Sign Method (FGSM) [25], Basic Iterative Method (BIM) [87], Carlini and Warger Attack (C&W Attack) [88], Deepfool [89], Square Attack (SA) [90] and HopSkipJump Attack (HSJA) [91]. The settings for attacks in our experiment are shown in Table 3. The victim model is ResNet-18 and ViT-Base/16.

Table 3. Important parameters of attack algorithms utilized in the experiments.

	Batch Size	Norm of Perturbation	Maximum Perturbation	Number of Iterations
FGSM	32	L_2	0.25	–
BIM	32	L_2	0.125	25
C&W	32	L_∞	–	20
Deepfool	8	–	–	50
SA	16	L_2	0.3	50
HSJA	16	L_2	–	50

In the part of proactive defense, we will recognize the generated adversarial data with the victim base models (i.e., ResNet-18 and ViT-Base/16). We set the weight of each base model in the ensemble as the same value, namely $1/2N$. The results of ASR from the victim model will be viewed as the performances before the defense. To evaluate the effectiveness of the ensemble model, we also conduct three counterparts in proactive defense of PGD-AT (adversarial training with PGD-perturbed RSIs), TRADES and GAL on the victim base models. The results for proactive defense are graphed as shown in Figures 13 and 14, and the victim model is labeled as *Without Defense* in both graphs.

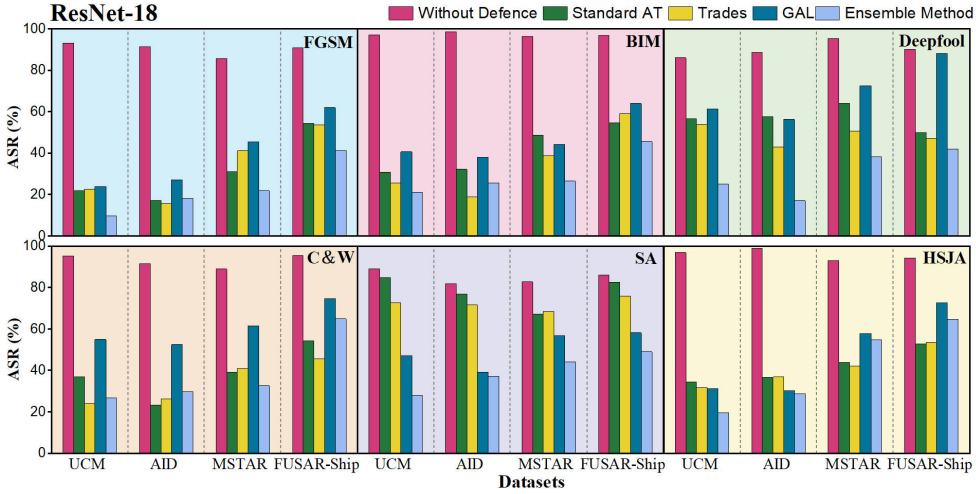

Figure 13. Comparisons in ASR of ensemble model in proactive defense with that of base model ResNet-18 and its three counterparts (the results are averaged from 10 repeated experiments).

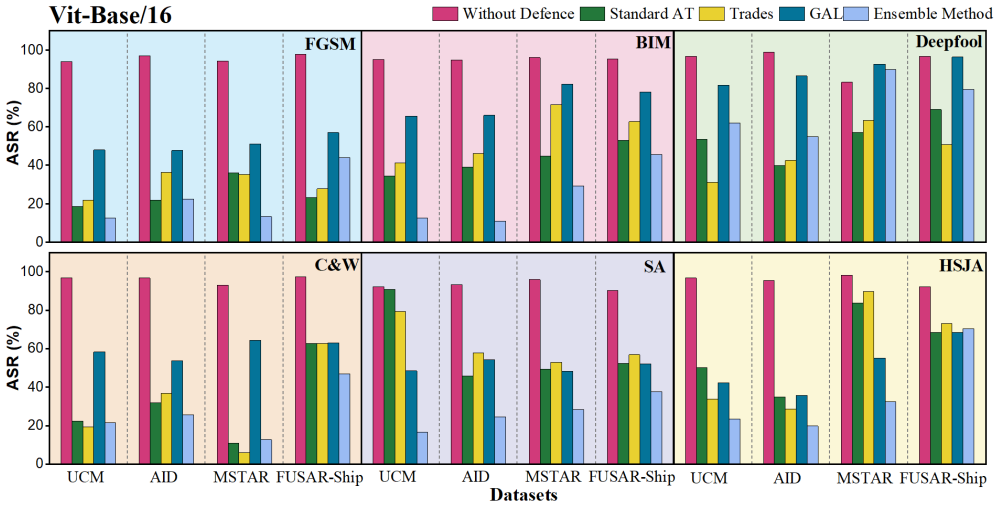

Figure 14. Comparisons in ASR of ensemble model in proactive defense with that of base model ViT-Base/16 and its three counterparts (the results are averaged from 10 repeated experiments).

In the last part of reactive defense, we compare the performances of the ensemble model with stand-alone detectors (i.e., KDE, LID and MD) in the ensemble framework. All four detectors exploit layer-specific scores from several intermediate layers of ResNet-18 and transformer encoders of ViT-Base/16 through logistic regression, and they detect if the input RSI is adversarial or benign. We only selected white-box attacks on UCM, AID and MSTAR for the experiments of this part because the RSIs in the test set of FUSAR-Ship are too inadequate to obtain stable data and analyze meaningful conclusions. The results of reactive defense are shown in Tables 4–6.

Table 4. Performances of ensemble method on UCM with three individual detection algorithms (the results below are averaged from 10 repeated experiments).

Dataset		FGSM		BIM		DeepFool		C & W	
		AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC
UCM	LID	88.89	90.30	88.52	89.53	57.91	65.22	64.27	72.35
	MD	92.95	88.51	90.24	84.83	67.45	74.28	76.44	83.42
	KDE	88.67	89.56	83.13	84.63	66.26	75.21	61.75	75.27
	Ensemble	**93.26**	**94.15**	**91.35**	**94.10**	**75.73**	**82.29**	**80.26**	**85.18**

Table 5. Performances of ensemble method on AID with three individual detection algorithms (the results below are averaged from 10 repeated experiments).

Dataset		FGSM		BIM		DeepFool		C & W	
		AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC
AID	LID	89.38	90.25	89.47	84.35	60.12	74.39	73.72	71.43
	MD	92.67	93.33	89.46	92.23	71.15	77.23	77.41	**85.63**
	KDE	87.59	83.84	80.93	83.30	68.18	78.32	61.51	73.77
	Ensemble	**95.73**	**95.93**	**93.37**	**95.10**	**74.05**	**81.08**	**80.40**	84.15

Table 6. Performances of ensemble method on MSTAR with three individual detection algorithms (the results below are averaged from 10 repeated experiments).

Dataset		FGSM		BIM		DeepFool		C & W	
		AUPR	AUROC	AUPR	AUROC	AUPR	AUROC	AUPR	AUROC
MSTAR	LID	81.58	83.17	81.13	82.79	71.80	74.83	68.47	71.61
	MD	**83.45**	84.85	85.73	77.67	67.25	72.23	75.41	78.24
	KDE	74.51	73.84	77.93	82.17	**73.60**	**78.74**	69.54	68.50
	Ensemble	82.43	**86.91**	**86.57**	**87.04**	72.13	77.37	**75.67**	**78.59**

4.3. Discussion

4.3.1. Recognition Performance of Base Models

First, for the base models in an ensemble of proactive defense, we trained them with the same setting and reported the recognition accuracy on the test sets. It can be observed that most of the 24 models yield very good performances with an accuracy of more than 85% except for the models on FUSAR-Ship. The reason for a drop in FUSAR-Ship is probably that the number of RSIs in FUSAR-Ship is scarce (only 420 RSIs in total) and the appearances of targets in four categories are similar, which makes it hard for the DNN model to learn the discriminative features to correctly distinguish them. The highest accuracy comes from ResNet-101 on AID, which can reach 97.86%. Models with deeper layers and more complex architectures perform a little bit worse such as ResNet-101 and ViT-Large/16 on UCM, which may be caused by a slight overfitting problem as the train data are not that sufficient. Nevertheless, all of these base models are well-trained and will be utilized in the later experiments of ensemble strategy for adversarial defense.

4.3.2. Analysis on Proactive Defense

We crafted adversarial examples against the ResNet-18 and ViT-Base/16, respectively, for each dataset with adversarial attack methods. The adversarial data are then recognized by the corresponding victim base model, our proposed ensemble model, and the victim base model is strengthened by three popular proactive defense methods. It can be noticed that in Figures 11 and 12, the height of all pink columns indicates that the ASR of these attacks reaches a very high level for the victim base model, which exhibits serious adversarial vulnerability and needs to be reduced urgently.

For adversarial examples generated against ResNet-18, we find that the ensemble of ResNets and ViTs performs well in optical datasets, especially with FGSM, BIM, Deepfool and HSJA attacks. In an optical setting, the ensemble can perform more consistently than other proactive defense methods. For example, ResNet-18 with Trades can correctly recognize more adversarial examples in BIM, but it has unsatisfactory performance in Deepfool. For the ensemble model, the best result is from the FGSM of UCM, with only 9.52% ASR. For SAR configurations, the ensemble of base models obtains better results in MSTAR than FUSAR-Ship, while it is worse than those from UCM and AID. In general, if we say an ASR below 30% is qualified, the ensemble has a good result in 15 out of 24 scenarios.

For adversarial examples generated against ViT-Base/16, the ensemble of ResNets and ViTs also maintains relatively low ASRs for most adversarial attack methods in optical RSI datasets. It is interesting to find that the ensemble model performs even worse than the base model without defense in Deepfool of MSTAR, but in C&W, another attack with very imperceptible noise, it yields decent values for MSTAR. Still, if we say an ASR below 30% is qualified, the ensemble has an acceptable result in 14 out of 24 scenarios.

Overall, compared with the models without defense under an adversarial attack, the ensemble strategy effectively improves the adversarial robustness and can rival or even perform better than the three other popular adversarial proactive defense methods.

4.3.3. Analysis on Reactive Defense

Last but not least, for reactive defense, we first discuss the results in optical RSI datasets. It can be observed that the ensemble method obtains the best AUPR or AUROC in 15 out of 16 scenarios. For gradient-based attacks of FGSM and BIM, the ensemble model can yield AUPR and AUROC values of more than 90%, which are obviously better than those from Deepfool and C&W. That is because Deepfool finds the shortest path to guide original RSIs across a decision boundary to generate adversarial examples, and C&W is an optimized-based attack with very small perturbations added to the original RSIs. The best result comes from the ensemble model in detecting FGSM on AID, with AUPR and AUROC values of 95.73 and 95.93, respectively. In addition, the results of FGSM are slightly better than those of the BIM attack, which is probably because the maximum perturbation in FGSM-perturbed RSIs is a little larger; thus, it leads to more obvious changes in feature representation. With respect to two harder situations, Deepfool and C&W, the ensemble model still shows better ability than stand-alone adversarial detection algorithms, especially with obvious improvements in Deepfool and C&W on UCM. MD only yields AUPR and AUROC values of Deepfool on UCM as 67.25 and 74.28, while our modified ENAD framework improve the metrics to 75.73 and 82.29. The results are not as good as those in gradient-based attacks, but compared with stand-alone detectors, these improvements show that the ensemble of detection algorithms and base DNN models has brought substantial benefits. In general, the ensemble framework has the potential to perform very well in RSI recognition for optical configuration.

In terms of results in MSTAR, the SAR dataset of target recognition, the values of output are generally lower than those of UCM and AID. The performances of the ensemble model are decreasing with the five best out of eight results. One possible reason for this phenomenon may lie in that the channel of SAR RSIs is 1 and most of an RSI in MSTAR is background without useful information, which inhibits the detector from extracting representative features except the target itself. Nevertheless, the detection of gradient-based attacks remains at a high level, with the AUPR and AUROC at around 85. The highest value comes from the BIM attack with 87.04 and the lowest is from the Deepfool attack with 73.60. The Deepfool and C&W attacks are still challenging situations with more imperceptible perturbations. In Deepfool, the results from the ensemble model are even lower than the stand-alone detector KDE, and in C&W, it performs at almost the same level as MD. Therefore, in such a case, an ensemble framework is not recommended, and it is worthwhile to further modify the ensemble model for a better detection in the SAR recognition dataset, especially for very imperceptible noise in the digital domain.

5. Conclusions

Stability and reliability are significant factors in the working process of modern UAVs with DNN-based visual navigation and recognition systems. However, there exists severe adversarial vulnerability when performing scene and target recognition tasks. We build a threat model when attackers maliciously access the communication link or place physical adversarial camouflage on targets. In the scenario, considering that AT is not adaptive for the resource-limited edge environment like UAVs and single adversarial detectors not performing well in reactive defense, we exploit the different mechanisms of feature extractions and weak adversarial transferability between the two mainstrean DNN models, CNN and transformer, to build deep ensemble models for both proactive and reactive adversarial defense only with base DNN models for the RSI recognition task. In addition, a one-stop platform for conducting adversarial defenses and evaluating adversarial robustness for DNN-based RSI recognition models called AREP-RSIs is developed, which can be edge-deployed to achieve real-time recognition and greatly facilitate designing more robust defense strategies in the remote sensing field for future research.

To evaluate the effectiveness of the two ensemble strategies, a series of experiments are conducted with both optical and SAR RSI datasets. We find that an ensemble of ResNets and ViTs can yield very satisfactory results in recognizing and detecting adversarial examples generated by gradient-based attacks such as FGSM and BIM. In proactive defense, compared with the three other popular defense methods, the ensemble can be more stable in different configurations. In reactive defense, our ensemble model integrates the scoring values from multiple detection algorithms and confidence scores from different base models, performing much better than stand-alone detectors in most experimental settings. Even though the proposed model does not perform as well on some attacks of SAR datasets, this ensemble strategy has shown the favorable potential to improve detection rates with the DNN models trained for RSI recognition.

In our future work, we will further optimize both of the deep ensemble frameworks, including exploring the defensive effectiveness against other types of adversarial attack in the RSI recognition task, replacing the current DNNs in the ensemble with more lightweight network architectures to suit the edge environment better and making the models' weights learnable during the training time to find the best combination. Therefore, as the first exploration of a deep ensemble method against adversarial RSIs in resource-limited environments, we need to conduct more experiments and report them in our next article. Finally, we will deploy the two deep ensemble models and AREP-RSIs on the edge devices to truly achieve a practical application.

Author Contributions: Methodology, Z.L.; software, Z.L. and Y.X.; validation, Z.L., H.S. and Y.X.; original draft preparation, Y.X.; writing—review and editing, Z.L.; supervision, H.S. and Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant 61971426.

Data Availability Statement: The UCM, AID, MSTAR and FUSAR-Ship dataset are available in the references of this paper.

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

References

1. LeCun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
2. He, H.; Wang, S.; Yang, D.; Wang, S. SAR target recognition and unsupervised detection based on convolutional neural network. In Proceedings of the 2017 Chinese Automation Congress (CAC), Jinan, China, 20–22 October 2017; pp. 435–438. [CrossRef]
3. Cho, J.H.; Park, C.G. Multiple Feature Aggregation Using Convolutional Neural Networks for SAR Image-Based Automatic Target Recognition. *IEEE Geosci. Remote Sens. Lett.* **2018**, *15*, 1882–1886. [CrossRef]
4. Wang, L.; Yang, X.; Tan, H.; Bai, X.; Zhou, F. Few-Shot Class-Incremental SAR Target Recognition Based on Hierarchical Embedding and Incremental Evolutionary Network. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5204111. [CrossRef]

5. Ding, B.; Wen, G.; Ma, C.; Yang, X. An Efficient and Robust Framework for SAR Target Recognition by Hierarchically Fusing Global and Local Features. *IEEE Trans. Image Process.* **2018**, *27*, 5983–5995. [CrossRef] [PubMed]
6. Deng, H.; Huang, J.; Liu, Q.; Zhao, T.; Zhou, C.; Gao, J. A Distributed Collaborative Allocation Method of Reconnaissance and Strike Tasks for Heterogeneous UAVs. *Drones* **2023**, *7*, 138. [CrossRef]
7. Li; Bin; Fei, Z.; Zhang, Y. UAV communications for 5G and beyond: Recent advances and future trends. *IEEE Internet Things J.* **2018**, *6*, 2241–2263. [CrossRef]
8. Khuwaja, A.A.; Chen, Y.; Zhao, N.; Alouini, M.S.; Dobbins, P. A survey of channel modeling for UAV communications. *IEEE Commun. Surv. Tutorials* **2018**, *20*, 2804–2821. [CrossRef]
9. Azari, M.M.; Geraci, G.; Garcia-Rodriguez, A.; Pollin, S. UAV-to-UAV communications in cellular networks. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 6130–6144. [CrossRef]
10. El Meouche, R.; Hijazi, I.; Poncet, P.A.; Abunemeh, M.; Rezoug, M. Uav Photogrammetry Implementation to Enhance Land Surveying, Comparisons and Possibilities. *Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.* **2016**, *42*, 107–114. [CrossRef]
11. Jung, S.; Kim, H. Analysis of amazon prime air uav delivery service. *J. Knowl. Inf. Technol. Syst.* **2017**, *12*, 253–266.
12. She, R.; Ouyang, Y. Efficiency of UAV-based last-mile delivery under congestion in low-altitude air. *Transp. Res. Part C Emerg. Technol.* **2021**, *122*, 102878. [CrossRef]
13. Thiels, C.A.; Aho, J.M.; Zietlow, S.P.; Jenkins, D.H. Use of unmanned aerial vehicles for medical product transport. *Air Med. J.* **2015**, *34*, 104–108. [CrossRef] [PubMed]
14. Konert, A.; Smereka, J.; Szarpak, L. The use of drones in emergency medicine: Practical and legal aspects. *Emerg. Med. Int.* **2019**, *2019*, 3589792. [CrossRef]
15. Michael, D.; Josh, H.; Keith, M.; Mikel, R. The vulnerability of UAVs: An adversarial machine learning perspective. In Proceedings of the Geospatial Informatics XI, SPIE, Online, 22 April 2021; Volume 11733, pp. 81–92. [CrossRef]
16. Barbu, A.; Mayo, D.; Alverio, J.; Luo, W.; Wang, C.; Gutfreund, D.; Tenenbaum, J.; Katz, B. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; pp. 9453–9463.
17. Hendrycks, D.; Dietterich, T. Benchmarking neural network robustness to common corruptions and perturbations. International Conference on Learning Representation. *arXiv* **2019**, arXiv:1903.12261.
18. Dong, Y.; Ruan, S.; Su, H.; Kang, C.; Wei, X.; Zhu, J. Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. Advances in Neural Information Processing Systems. *arXiv* **2022**, arXiv:2210.03895v1.
19. Hendrycks, D.; Lee, K.; Mazeika, M. Using pretraining can improve model robustness and uncertainty. *Int. Conf. Mach. Learn.* **2019**, *97*, 2712–2721.
20. Akhtar, N.; Mian, A. Threat of Adversarial Attacks on Deep Learning in Computer Vision: A Survey. *IEEE Access* **2018**, *6*, 14410–14430. [CrossRef]
21. Akhtar, N.; Mian, A.; Kardan, N.; Shah, M. Advances in Adversarial Attacks and Defenses in Computer Vision: A Survey. *IEEE Access* **2021**, *9*, 155161–155196. [CrossRef]
22. Khamaiseh, S.Y.; Bagagem, D.; Al-Alaj, A.; Mancino, M.; Alomari, H.W. Adversarial Deep Learning: A Survey on Adversarial Attacks and Defense Mechanisms on Image Classification. *IEEE Access* **2022**, *10*, 102266–102291. 2022.3208131. [CrossRef]
23. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Li, F.-F. ImageNet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255. [CrossRef]
24. Zhang, L.; Zhang, L. Artificial Intelligence for Remote Sensing Data Analysis: A review of challenges and opportunities. *IEEE Geosci. Remote Sens. Mag.* **2022**, *10*, 270–294. [CrossRef]
25. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* **2014**, arXiv:1412.6572.
26. Zhang, H.Y.; Yu, Y.D.; Jiao, J.T.; Xing, E.P.; Ghaoui, L.E.; Jordan, M. Theoretically principled trade-off between robustness and accuracy. *arXiv* **2019**, arXiv:1901.08573.
27. Zhang, J.F.; Xu, X.L.; Han, B.; Niu, G.; Cui, L.Z.; Sugiyama, M.; Kankanhalli, M. Attacks which do not kill training make adversarial learning stronger. *Int. Conf. Mach. Learn.* **2020**, *119*, 11278–11287.
28. Jia, X.J.; Zhang, Y.; Wu, B.Y.; Ma, K.; Wang, J.; Cao, X.C. LAS-AT: Adversarial Training with Learnable Attack Strategy. *arXiv* **2022**, arXiv:2203.06616.
29. Saligrama, A.; Leclerc, G. Revisiting Ensembles in an Adversarial Context: Improving Natural Accuracy. *arXiv* **2020**, arXiv:2002.11572.
30. Li, N.; Yu, Y.; Zhou, Z.H. Diversity regularized ensemble pruning. In Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Bilbao, Spain, 13–17 September 2012; pp. 330–345.
31. Wang, X.; Xing, H.; Hua, Q.; Dong, C.R.; Pedrycz, W. A study on relationship between generalization abilities and fuzziness of base classifiers in ensemble learning. *IEEE Trans. Fuzzy Syst.* **2015**, *23*, 1638–1654. [CrossRef]
32. Sun, T.; Zhou, Z.H. Structural diversity for decision tree ensemble learning. *Front. Comput. Sci.* **2018**, *12*, 560–570. [CrossRef]
33. Cohen, G.; Sapiro, G.; Giryes, R. Detecting adversarial samples using influence functions and nearest neighbors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 14453–14462.
34. Ma, X.J.; Li, B.; Wang, Y.S.; Erfani, S.M.; Wijewickrema, S.N.R.; Schoenebeck, G.; Song, D.; Houle, M.E.; Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. In Proceedings of the 6th International Conference on Learning Representations, ICLR, Vancouver, BC, Canada, 30 April–3 May 2018.

35. Feinman, R.; Curtin, R.R.; Shintre, S.; Gardner, A.B. Detecting Adversarial Samples from Artifacts. *arXiv* **2017**, arXiv:1703.00410.
36. Lee, K.; Lee, K.; Lee, H.; Shin, J. A simple unified framework for detecting out-of distribution samples and adversarial attacks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7167–7177.
37. Hendrycks, D.; Gimpel, K. Early methods for detecting adversarial images. In Proceedings of the 5th International Conference on Learning Representations, ICLR, Toulon, France, 24–26 April 2017.
38. Zheng, Z.H.; Hong, P.Y. Robust detection of adversarial attacks by modeling the intrinsic properties of deep neural networks. *Adv. Neural Inf. Process. Syst.* **2018**, *31*, 7913–7922.
39. Liang, B.; Li, H.C.; Su, M.Q.; Li, X.R.; Shi, W.C.; Wang, X.F. Detecting adversarial image examples in deep neural networks with adaptive noise reduction. *IEEE Trans. Dependable Secur. Comput.* **2021**, *18*, 72–85. [CrossRef]
40. Xu, W.L.; Evans, D.; Qi, Y.J. Feature squeezing: Detecting adversarial examples in deep neural networks. In Proceedings of the 25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, CA, USA, 18–21 February 2018.
41. Kherchouche, A.; Fezza, S.A.; Hamidouche, W.; Deforges, O. Detection of adversarial examples in deep neural networks with natural scene statistics. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, Glasgow, UK, 19–24 July 2020; pp. 1–7.
42. Sotgiu, A.; Demontis, A.; Melis, M.; Biggio, B.; Fumera, G.; Feng, X.Y.; Roli, F. Deep neural rejection against adversarial examples. *Eurasip J. Inf. Secur.* **2020**, *2020*, 5. [CrossRef]
43. Aldahdooh, A.; Hamidouche, W.; Deforges, O. Revisiting model's uncertainty and confidences for adversarial example detection. *arXiv* **2021**, arXiv:2103.05354.
44. Carrara, F.; Falchi, F.; Caldelli, R.; Amato, G.; Fumarola, R.; Becarelli, R. Detecting adversarial example attacks to deep neural networks. In Proceedings of the 15th International Workshop on Content-Based Multimedia, Florence, Italy, 19–21 June 2017.
45. Aldahdooh, A.; Hamidouch, W.; Fezza, S.A.; Deforges, O. Adversarial Example Detection for DNN Models: A Review and Experimental Comparison. *arXiv* **2021**, arXiv:2105.00203.
46. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]
47. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
48. Craighero, F.; Angaroni, F.; Stella, F.; Damiani, C.; Antoniotti, M.; Graudenzi, A. Unity is strength: Improving the detection of adversarial examples with ensemble approaches. *arXiv* **2021**, arXiv:2111.12631.
49. Sun, H.; Chen, J.; Lei, L.; Ji, K.; Kuang, G. Adversarial robustness of deep convolutional neural network-based image recognition models: A review. *J. Radars* **2021**, *10*, 571–594. [CrossRef]
50. Chen, L.; Zhu, G.; Li, Q.; Li, H. Adversarial example in remote sensing image recognition. *arXiv* **2019**, arXiv:1910.13222.
51. 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.* **2021**, *59*, 1604–1617. [CrossRef]
52. Xu, Y.; Ghamisi, P. Universal Adversarial Examples in Remote Sensing: Methodology and Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 1–15. [CrossRef]
53. Chen, L.; Xiao, J.; Zou, P.; Li, H. Lie to Me: A Soft Threshold Defense Method for Adversarial Examples of Remote Sensing Images. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 8016905. [CrossRef]
54. Li, H.; Huang, H.; Chen, L.; Peng, J.; Huang, H.; Cui, Z.; Mei, X.; Wu, G. Adversarial Examples for CNN-Based SAR Image Classification: An Experience Study. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 1333–1347. [CrossRef]
55. Du, C.; Huo, C.; Zhang, L.; Chen, B.; Yuan, Y. Fast C&W: A Fast Adversarial Attack Algorithm to Fool SAR Target Recognition with Deep Convolutional Neural Networks. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 4010005. [CrossRef]
56. Zhou, J.; Sun, H.; Lei, L.; Ke, J.; Kuang, G. Sparse Adversarial Attack of SAR Image. *J. Signal Process.* **2021**, *37*, 11.
57. 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, Washington, DC, USA, 6–9 November 2018; pp. 408–411.
58. Hollander, R.; Adhikari, A.; Tolios, I.; van Bekkum, M.; Bal, A.; Hendriks, S.; Kruithof, M.; Gross, D.; Jansen, N.; Perez, G.; et al. Adversarial patch camouflage against aerial detection. In Proceedings of the Artificial Intelligence and Machine Learning in Defense Applications II, International Society for Optics and Photonics, SPIE, Online, 20 September 2020; Volume 11543, pp. 77–86.
59. Du, A.; Chen, B.; Chin, T.-J.; Law, Y.W.; Sasdelli, M.; Rajasegaran, R.; Campbell, D. Physical adversarial attacks on an aerial imagery object detector. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022.
60. Torens, C.; Juenger, F.; Schirmer, S.; Schopferer, S.; Maienschein, T.D.; Dauer, J.C. Machine Learning Verification and Safety for Unmanned Aircraft-A Literature Study. In Proceedings of the AIAA Scitech 2022 Forum, San Diego, CA, USA, 3–7 January 2022.
61. Tian, J.; Wang, B.; Guo, R.; Wang, Z.; Cao, K.; Wang, X. Adversarial Attacks and Defenses for Deep-Learning-Based Unmanned Aerial Vehicles. *IEEE Internet Things J.* **2021**, *9*, 22399–22409. [CrossRef]
62. Brown, T.B.; Mané, D.; Roy, A.; Abadi, M.; Gilmer, J. Adversarial patch. *arXiv* **2017**, arXiv:1712.09665.
63. Gu, J.; Tresp, V.; Qin, Y. Are vision transformers robust to patch perturbations? In Proceedings of the Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, 23–27 October 2022; Proceedings, Part XII; Springer Nature: Cham, Switzerland, 2022.

64. Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; Dosovitskiy, A. Do vision transformers see like convolutional neural networks? *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 12116–12128.
65. Namuk, P.; Kim, S. How do vision transformers work? *arXiv* **2022**, arXiv:2202.06709.
66. Rao, Y.; Zhao, W.; Tang, Y.; Zhou, J.; Lim, S.N.; Lu, J. Hornet: Efficient high-order spatial interactions with recursive gated convolutions. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 10353–10366.
67. Si, C.; Yu, W.; Zhou, P.; Zhou, Y.; Wang, X.; Yan, S. Inception transformer. *arXiv* **2022**, arXiv:2205.12956.
68. Li, J.; Xia, X.; Li, W.; Li, H.; Wang, X.; Xiao, X.; Wang, R.; Zheng, M.; Pan, X. Next-vit: Next generation vision transformer for efficient deployment in realistic industrial scenarios. *arXiv* **2022**, arXiv:2207.05501.
69. Yang, T.; Zhang, H.; Hu, W.; Chen, C.; Wang, X. Fast-ParC: Position Aware Global Kernel for ConvNets and ViTs. *arXiv* **2022**, arXiv:2210.04020.
70. Cai, Y.; Ning, X.; Yang, H.; Wang, Y. Ensemble-in-One: Learning Ensemble within Random Gated Networks for Enhanced Adversarial Robustness. *arXiv* **2021**, arXiv:2103.14795.
71. Pang, T.; Xu, K.; Du, C.; Chen, N.; Zhu, J. Improving adversarial robustness via promoting ensemble diversity. In Proceedings of the 36th International Conference on Machine Learning, PMLR, Long Beach, CA, USA, 9–15 June 2019.
72. Teresa, Y.; Kar, O.F.; Zamir, A. Robustness via cross-domain ensembles. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021.
73. Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; Madry, A. Robustness may be at odds with accuracy. *arXiv* **2018**, arXiv:1805.12152.
74. Mahmood, K.; Mahmood, R.; van Dijk, M. On the Robustness of Vision Transformers to Adversarial Examples. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021; pp. 7818–7827. [CrossRef]
75. Sanjay, K.; Qureshi, M.K. Improving adversarial robustness of ensembles with diversity training. *arXiv* **2019**, arXiv:1901.09981.
76. Summerfield, M. *Rapid GUI Programming with Python and Qt: The Definitive Guide to PyQt Programming (Paperback)*; Pearson Education: London, UK, 2007.
77. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **2019**, *32*, 4970–4979.
78. Nicolae, M.-M.; Sinn, M.; Tran, M.N.; Buesser, B.; Rawat, A.; Wistuba, M.; Zantedeschi, V.; Baracaldo, N.; Chen, B.; Ludwig, H.; et al. Adversarial Robustness Toolbox v1.0.0. *arXiv* **2018**, arXiv:1807.01069.
79. Bradski, G. The openCV library. *Dr. Dobb's J. Softw. Tools Prof. Program.* **2000**, *25*, 120–123.
80. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.;Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
81. Yang, H.; Zhang, J.; Dong, H.; Inkawhich, N.; Gardner, A.; Touchet, A.; Wilkes, W.; Berry, H.; Li, H. DVERGE: Diversifying vulnerabilities for enhanced robust generation of ensembles. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 5505–5515.
82. Yang, Y.; Newsam, S. Bag-of-visual-words and spatial extensions for land-use classification. In Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems, San Jose, CA, USA, 2–5 November 2010; pp. 270–279.
83. Xia, G.-S.; Hu, J.; Hu, F.; Shi, B.; Bai, X.; Zhong, Y.; Zhang, L. AID: A Benchmark Data Set for Performance Evaluation of Aerial Scene Classification. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 3965–3981. [CrossRef]
84. Ross, T.D.; Worrell, S.W.; Velten, V.J.; Mossing, J.C.; Bryant, M.L. Standard SAR ATR evaluation experiments using the MSTAR public release data set. *Proc. SPIE* **1998**, *3370*, 566–573.
85. Hou, X.; Ao, W.; Song, Q.; Lai, J.; Wang, H.; Xu, F. FUSAR-Ship: Building a high-resolution SAR-AIS matchup dataset of Gaofen-3 for ship detection and recognition. *Sci. China Inf. Sci.* **2020**, *63*, 140303. [CrossRef]
86. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
87. Kurakin, A.; Goodfellow, I.J.; Bengio, S. Adversarial examples in the physical world. In *Artificial Intelligence Safety and Security*; Chapman and Hall/CRC: Boca Raton, FL, USA, 2018; pp. 99–112.
88. Carlini, N.; Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, New York, NY, USA, 3 November 2017.
89. Dezfooli, M.; Mohsen, S.; 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, New Orleans, LA, USA, 18–20 June 2016.
90. Andriushchenko, M.; Croce, F.; Flammarion, N.; Hein, M. Square attack: A query-efficient black-box adversarial attack via random search. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part XXIII; Springer International Publishing: Cham, Switzerland, 2020.
91. Chen, J.; Jordan, M.I.; Wainwright, M.J. Hopskipjumpattack: A query-efficient decision-based attack. In Proceedings of the 2020 IEEE Symposium on Security and Privacy (sp), IEEE, San Francisco, CA, USA, 18–21 May 2020.