remote sensing

MDPI

Article

Learning Adversarially Robust Object Detector with Consistency Regularization in Remote Sensing Images

Yang Li [1,2,3], Yuqiang Fang [4], Wanyun Li [1,3], Bitao Jiang [1,2,*], Shengjin Wang [5] and Zhi Li [4]

[1] Department of Space Information, Space Engineering University, Beijing 101416, China; yangli.cs@outlook.com (Y.L.)
[2] Beijing Institute of Remote Sensing Information, Beijing 100192, China
[3] Graduate School, Space Engineering University, Beijing 101416, China
[4] Science and Technology on Complex Electronic System Simulation Laboratory, Space Engineering University, Beijing 101416, China; lizhizys@139.com (Z.L.)
[5] Department of Electronic Engineering, Tsinghua University, Beijing 100084, China; wgsgj@tsinghua.edu.cn
* Correspondence: bitao_jiang@163.com

Abstract: Object detection in remote sensing has developed rapidly and has been applied in many fields, but it is known to be vulnerable to adversarial attacks. Improving the robustness of models has become a key issue for reliable application deployment. This paper proposes a robust object detector for remote sensing images (RSIs) to mitigate the performance degradation caused by adversarial attacks. For remote sensing objects, multi-dimensional convolution is utilized to extract both specific features and consistency features from clean images and adversarial images dynamically and efficiently. This enhances the feature extraction ability and thus enriches the context information used for detection. Furthermore, regularization loss is proposed from the perspective of image distribution. This can separate consistent features from the mixed distributions for reconstruction to assure detection accuracy. Experimental results obtained using different datasets (HRSC, UCAS-AOD, and DIOR) demonstrate that the proposed method effectively improves the robustness of detectors against adversarial attacks.

Keywords: robust detector; adversarial example; object detection; adversarial training; remote sensing image

Citation: Li, Y.; Fang, Y.; Li, W.; Jiang, B.; Wang, S.; Li, Z. Learning Adversarially Robust Object Detector with Consistency Regularization in Remote Sensing Images. *Remote Sens.* **2023**, *15*, 3997. https://doi.org/10.3390/rs15163997

Academic Editor: Adrian Stern

Received: 30 May 2023
Revised: 4 August 2023
Accepted: 7 August 2023
Published: 11 August 2023

1. Introduction

Object detection in RSIs has broad applications in urban and agricultural fields. In recent years, there has been significant progress in convolutional-neural-network-based (CNN-based) detectors. Compared with traditional methods, CNN-based detectors have better feature extraction ability, faster speeds, and higher accuracy [1,2]. Adversarial attacks introduce small and imperceptible texture perturbations to clean images through gradient descending or adversarial generation and then generate adversarial examples. CNNs are vulnerable to attacks because they make final predictions by promoting semantic understanding according to the corresponding texture and color of the input images. However, adversarial attacks disrupt the texture information and destroy the semantic understanding of the network, eventually leading to incorrect predictions [3]. Therefore, it is important to improve the robustness of CNN-based detectors and mitigate performance degradation caused by adversarial attacks, especially to ensure the reliable deployment of security sensitive applications.

Early works on adversarial attacks [4,5] focus on the classifier and invalidate the classification model. These works mainly generate the most effective adversarial example while making minimal changes to the original image by adding disturbances to the input. Thus, these changes mislead the network and greatly reduce the accuracy of the classifier. Specifically, gradient-based methods compute the minimum necessary perturbation by

maximizing the prediction error of the network to produce adversarial examples. Transfer-based attacks exploit the use of training data to generate adversarial alternatives. As an extension of the classification task, object detection [6–8] outputs the classification label and the location at the same time. In this way, the detection process is more complex and difficult to attack. Some works [9,10] significantly reduce the accuracy of models by generating adversarial patches, exposing the weak robustness of object detection under adversarial attacks.

Compared with natural images, RSIs have a larger size, higher spatial resolution, and more detailed spectral information. A small noise perturbation can successfully attack the model and change the output [11]. Specifically, adversarial attacks in the digital world generate adversarial patches to cover important parts [12], which can also be applied in the physical world to evade or deceive object detectors. Defense methods under attack have often been ignored, especially in military scenarios. Attackers apply adversarial perturbation to the target (such as an aircraft or ship); then, the adversarial examples can directly deceive the CNN-based model, introducing high-level risks to the object detection system.

Current defense methods used in object detection to achieve robustness improvement can be generally classified into two classes. One type includes methods developed for model vulnerability [13], which find the potential adversarial patch area and recover it for detection. Unfortunately, these models pose considerable challenges to the authenticity and accuracy of data recovery, which are inapplicable to remote sensing scenarios. The other type of widely used methods are those that obtain robust models through training. Through adversarial training [14], the MTD model learns both clean and adversarial images, pays more attention to the robust features in adversarial samples, and ignores the non-robust features. However, adversarial training always leads to a robustness bottleneck [15]. Specifically, to resist adversarial attacks, the detector has to sacrifice the performance on clean images to detect adversarial images. This is mainly because of the introduction of adversarial features. The opposing training effects between the clean image and adversarial image make it harder to distinguish clean and adversarial features, resulting in a decrease in performance during detection. As shown in Figure 1, we use a classical single-shot multibox detector (SSD) [6], a robust detector using multi-task domains (MTD) [14], a robust detector with class-wised adversarial training (CWAT) [16], and our proposed detector to test the performance before and after the adversarial attack Projected Gradient Descent (PGD) [17]. The results show that the accuracy is improved after the attack to some extent, but the performance on clean images is degraded inevitably for current robust detectors.

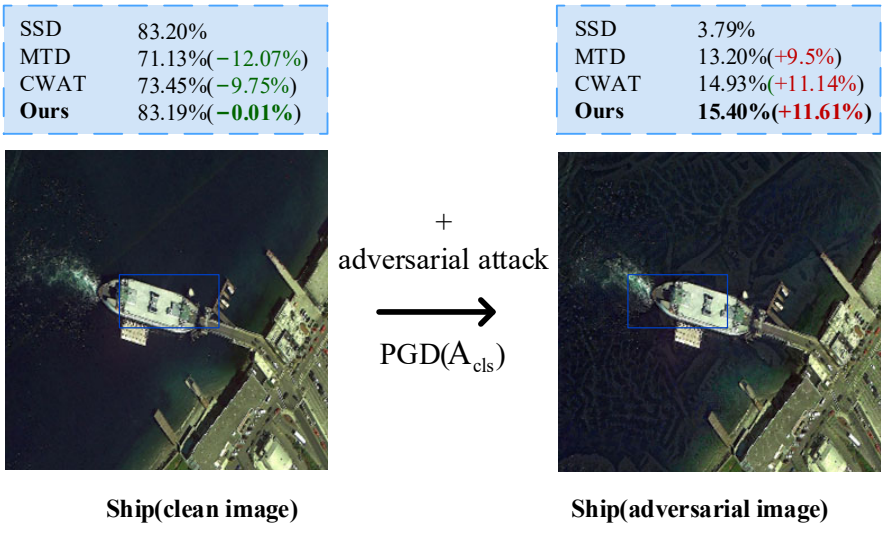

Figure 1. Performance comparison of the standard SSD [6], robust detector MTD [14], CWAT [16], and our model after a classification attack by PGD [17].

This paper provides effective solutions to the above issues. Firstly, a multi-dimensional convolution kernel is proposed to learn the robust features of the clean image and the corresponding adversarial image efficiently and dynamically. It can obtain rich context information from RSIs and significantly improve the feature extraction ability of the detector. Furthermore, a regularization loss was designed to constrain the consistent feature reconstruction process from the perspective of the internal distribution of the image. It reduces the interference of adversarial attacks and further increases the robustness of the object detector. The key contributions of this paper are summarized as follows:

1. A multi-dimensional adversarial convolution (MAConv) kernel is proposed to extract features adaptively and dynamically. It effectively extracts both shared features and specific features from clean images and adversarial images under attacks and thus enhances the ability of feature extraction.
2. From the perspective of image distribution, a consistency regularization loss is proposed to extract consistent features from the mixture distribution under attacks. By reconstructing clean features for detection, our method successfully improves the robustness of the object detector.
3. We performed extensive experiments under different adversarial attacks on three public datasets, HRSC, UCAS-AOD, and DIOR. The experimental results show superior performance in both single-class and multi-class object detection compared with current robust object detectors.

2. Related Works

2.1. Attacks for Object Detection

Object detectors are vulnerable to adversarial attacks, which can lead to wrong predictions. The current adversarial attack methods mainly fall into two types, white-box and black-box attacks. Specifically, white-box attacks assume knowledge of the structure and parameters of the model. Typical attacks, including FGSM [4] and PGD [17], use one-step and multi-step gradient descent at each pixel to create adversarial examples. On the other hand, black-box methods only need to know the output of the attacked model [18–20]. With regard to object detection tasks, Lu et al. [21] attacked the object-level features by introducing perturbations to whole images. Li et al. [22] used adversarial perturbations to interfere with the region proposal network in the network. Zhang et al. [23] performed feature-level attacks on each neuron in the middle layer of the network to provide adversarial examples. Czaja et al. [24] simulated physical attacks in digital space and conducted experiments on remote sensing classification tasks for the first time. Adhikari et al. [25] apply patch-based adversarial attacks to unmanned aerial surveillance, camouflaging the aircraft with a small adversarial patch to prevent automatic detection. Li et al. [26] discuss adversarial examples and propose model vulnerability and attack selectivity in RSIs. Xu et al. [27] showed that there are also adversarial examples in the hyperspectral domain. Du et al. [28] applied adversarial patches on cars and suggest new experiments and metrics to prove the effectiveness of physical adversarial attacks on detectors in aerial images. Xu et al. [29] conducted black-box attacks by finding common vulnerabilities by attacking shallow features in the proxy model. Zhang et al. [30] propose a scale factor to provide a generic adversarial patch and act on different scales of targets in RSIs. Yolo series algorithms were used to verify this method in practical applications.

2.2. Adversarial Training and Robustness Detector

Adversarial training is a mainstream and effective strategy to make robust models. Hendrycks et al. [31] first propose ways to reinforce perturbation robustness for image classifiers. Zhang [14] generalize adversarial training from classification to detection, and analyze the relationship between different tasks' losses to improve robustness.

Tsipras et al. [32] indicated that there is a trade-off between the accuracy and robustness of adversarial models, due to the intrinsic differences between the feature representations learned by robust detectors and standard detectors in adversarial training. Maini [33]

developed a generalization of standard PGD-based methods to incorporate different perturbations without complex architectures. Rice et al. [34] found that early stopping benefits robust performance in adversarial training. Chen et al. [16] designed class-aware adversarial training leveraging a class weighted loss to strengthen the model's robustness. Pang et al. [35] investigated the impact of implementation details, which are usually ignored in adversarial training. They concluded that these settings should also be fine-tuned carefully during robustness research. Tack et al. [36] propose auxiliary consistency regularization to force the predictive distributions after attacking to increase the accuracy. Gowal et al. [37] artificially expanded the scale of the original training set and introduce generative models. Dong et al. [15] focused on extracting robust features from both clean and adversarial samples to enhance detector robustness.

3. Proposed Methods

3.1. Network Framework

The model robustness bottleneck mainly lies in the conflicts of learning clean and adversarial images, which makes it difficult to identify robust and non-robust features. For the purpose of further improving the robustness of object detectors, we propose a robust object detector for RSIs. Consistent with the current mainstream robust detectors, we also used the standard SSD as the base model. In the feature extraction stage, the model learns the objective difference between clean images and adversarial images through our proposed multi-dimensional dynamic convolution. It is specifically applied to the layers before the conv4_3 layer of the SSD backbone to effectively improve the feature extraction ability while controlling the network scale, as shown in Figure 2. To further ensure that robust features can be extracted effectively from the adversarial images and that they are consistent with the clean image, influenced by VAE [38] and RobustDet [15], a decoder structure is utilized to reconstruct clean images from the features extracted by the front-end network. Meanwhile, a consistency regularization loss is designed to constrain the reconstructed image to the similar predictions with the original clean image. Therefore, the model can learn robust features and improve the robustness performance effectively.

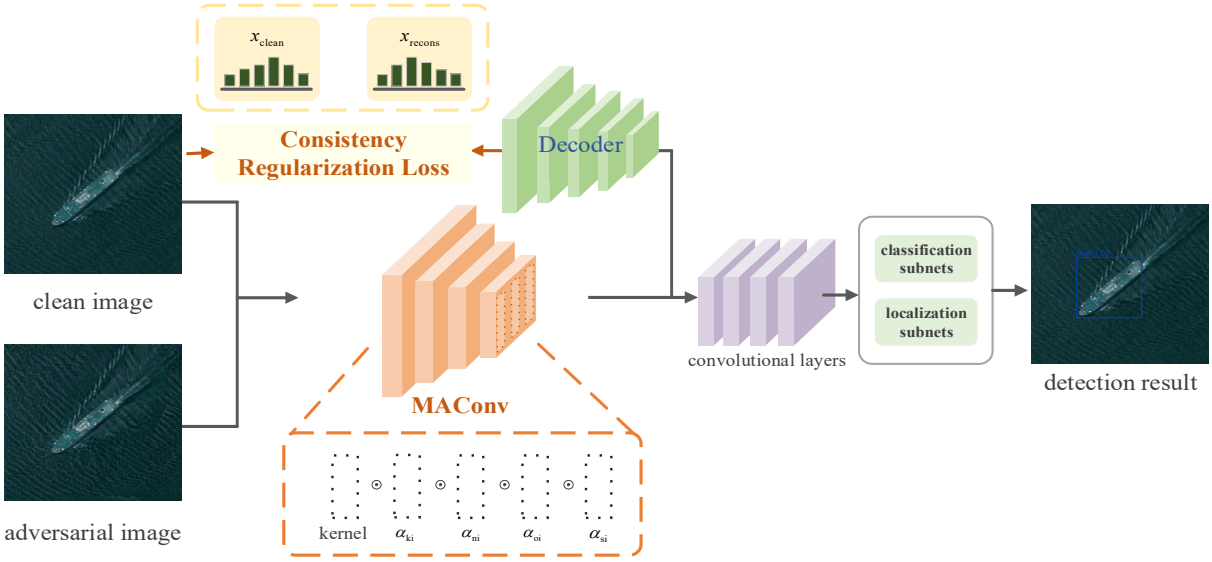

Figure 2. Network architecture of the proposed robust detector.

3.2. Multi-Dimensional Adversarial Convolution Feature Extraction

Compared with natural images, objects are captured from a top-down view in RSIs, which account for a smaller proportion of the whole image. The background is also more cluttered, which puts forward higher requirements for feature extraction. When the image is disturbed by adversarial attacks, features in the adversarial image change, thus

making it difficult to distinguish robust features using regular kernel learning. Unlike regular convolutional layers, which use the same convolution kernel for all inputs, dynamic convolution [39,40] can learn a linear combination of multiple kernels weighted with the attention conditioned on the input. Specifically, dynamic convolution operations can be described as $y = (\sum_{i=1}^{n} \pi_{ki} \cdot \mathrm{k_i}) * x$, where n is the number of dynamic convolutions; x, y denote the input and output features; k is the convolution kernel; and π_{ki} denotes the convolution weight calculated by the attention function, which is the dividual attention vector assigned to each kernel k_i. These dynamic convolutions only focus the weights on the convolution kernel dimension but make no difference to other dimensions.

Inspired by [41], our model proposes multi-dimensional adversarial convolution (MAConv), which can distinguish more dimensions to refine clean and adversarial images, as illustrated in Figure 3. By adapting parameters and kernels to different images, our model learns specific features efficiently. In order to balance the computational burden and the effectiveness of feature extraction, we apply MAConv in layers before the conv4_3 layer of the VGG16 backbone. Mathematically, convolutional weights are computed in parallel for all four dimensions:

$$y = (\sum_{i=1}^{n} \pi_{ki} \odot \pi_{ni} \odot \pi_{oi} \odot \pi_{si} \odot k_i) * x \tag{1}$$

where $\pi_{ki}, \pi_{ni}, \pi_{oi}, \pi_{si}$ denote the weights of different dimensions, including convolution kernels, the input channel dimension, the output channel dimension, and the spatial dimension, respectively. Accordingly, π_{ki} denotes the weight for n dynamic convolutions; π_{ni} is applied to the input channel dimension of each convolutional filter; π_{oi} is applied to the output channel dimension of each filter; π_{si} is applied to the $k \times k$ spatial locations of each kernel; and $\odot$ denotes the multiplication operation along the corresponding dimensions of the filter. In this way, the attention weight acting on full dimensions can significantly enhance the feature extraction ability of convolution. Thus, the model can obtain rich context information in RSIs.

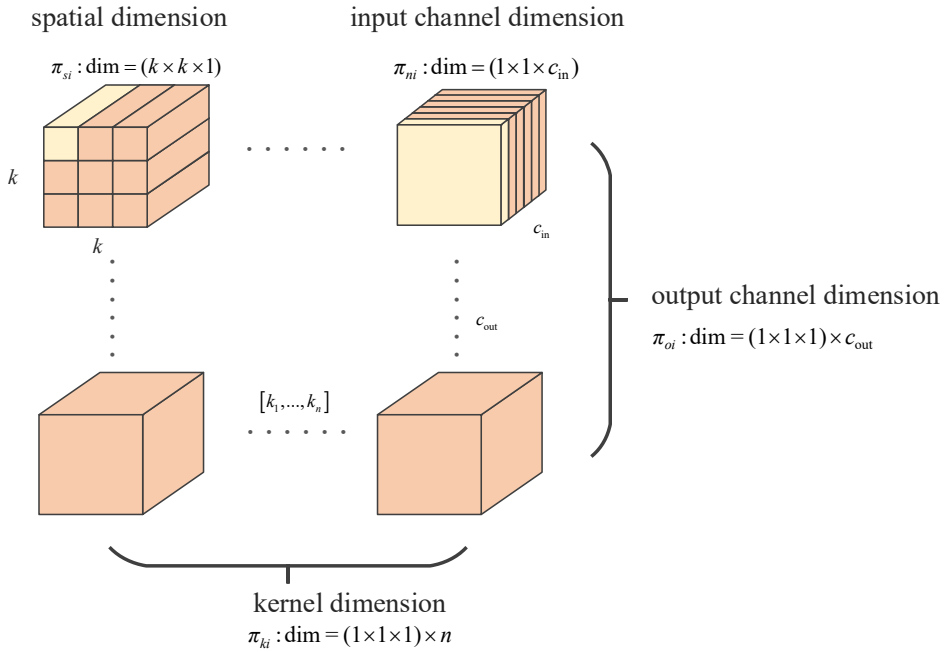

Figure 3. Illustration of weight calculations for different dimensions.

To calculate the weights, the input x is first passed through channel-wise average pooling (AP) and transformed into a feature vector of length c_{in}, followed by a fully

connected (FC) layer for further dimensionality reduction. Then, a rectified linear unit (ReLU) is connected, which can be denoted as follows:

$$B = ReLU(FC(AP(x))) \tag{2}$$

Later, four branches are added in parallel to calculate the weights in four different dimensions. In each branch, the FC and Sigmoid layers are utilized to perform dimension transformation and output the weights in different sizes, which are $1 \times 1 \times 1 \times n$, $1 \times 1 \times c_{in}$, $1 \times 1 \times 1 \times c_{out}$, and $k \times k \times 1$, corresponding to the convolution kernels, the input channel dimension, the output channel dimension, and the spatial dimension, respectively, which can be expressed as:

$$\Pi = Sigmoid(FC(B_i)) \tag{3}$$

where $\Pi = \{\pi_{ki}, \pi_{ni}, \pi_{oi}, \pi_{si}\}$. Finally, we can obtain the output of the MAConv:

$$y = (\sum k \odot \Pi) * x \tag{4}$$

MAConv adaptively extracts features from multiple dimensions, which is suitable for RSIs with complex feature distribution and can effectively learn robust features. Compared with the regular convolution layer, it dynamically combines full dimensions to perform convolution operations to the input and adapts different kernels and weights for each dimension. Through various weight computing, both shared features and the specific features in clean and adversarial images can be accurately extracted. Thus, the model has the ability to address the robustness bottleneck of the detection task under adversarial attacks.

3.3. Consistency Regularization Loss for Reconstruction

For the classical assumption of machine learning, models can correctly classify data that are independent and well-distributed within the training set. From the perspective of distribution, detection models often make identical predictions for images in consistent distribution. Xie et al. [42] and Zhu et al. [43] also showed that under adversarial attacks, the distribution of the generated adversarial images will change and mislead the target model.

To visualize how clean and adversarial images are distributed, we project their high-dimensional representations into a 2D space via t-SNE. In Figure 4, the orange represents clean images of the typical ship targets from the HRSC dataset, and the green describes the adversarial examples derived from the class attack. It can be seen that the attack results in misalignment in the domain compared with the original images. Such adversarial interference also causes inherent changes in the original data and misleads the detector into obtaining the wrong results. Clean and adversarial images have different inference distributions within different domains, which leads to the inconsistency of gradient updating direction in adversarial training. Such conflict results in a performance bottleneck in the robust model. Figure 5 depicts the estimated probability distribution of clean and adversarial images severally and in a mixture.

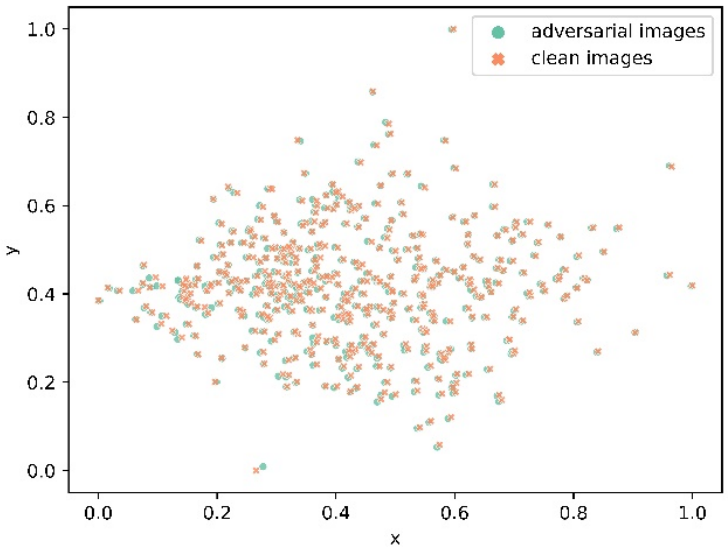

Figure 4. Distribution of clean and adversarial images from the HRSC dataset using t-SNE.

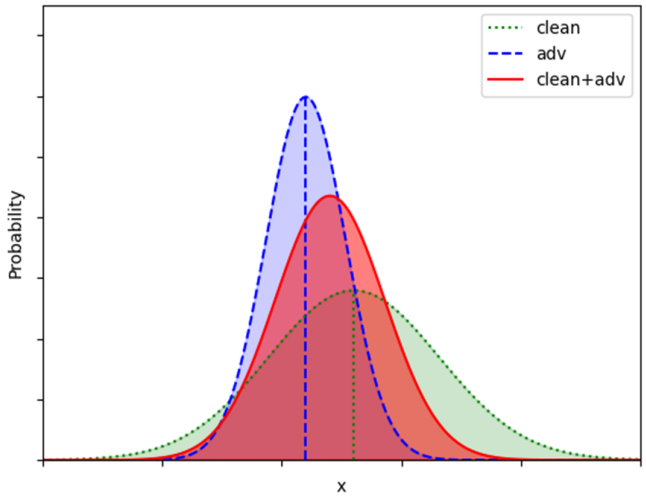

Figure 5. Estimated probability statistics of clean and adversarial images.

In order to solve the problem of separating effective consistency features from mixed distributions and reduce the negative impact of adversarial perturbations on the detection results, the decoder structure is utilized to reconstruct the consistency features from both clean and adversarial images after the feature extraction stages. The reconstruction process is constrained in the perspective of image distribution, and the reconstructed features are used for subsequent detection.

Inspired by VAE [38] and RobustDet [15], we used a decoder structure to reconstruct the consistent features of clean and adversarial images after feature extraction. As the reconstruction process essentially involves two distributions, the important step is to disentangle consistent features from this mixture distribution. To further constrain the process, we controlled the model by the intrinsic image distribution and ensured the reconstruction distributed in the adjacent areas of the corresponding clean image. Thus, the reconstructions can obtain the correct results like the original image.

Specifically, our model utilizes temperature scaling [44], an extension of the parametric calibration method to measure the image distribution. Formally, the temperature-scaled distribution is described as:

$$\hat{f}(x;\tau) = Softmax(z(x)/\tau) \tag{5}$$

where $z(x)$ denotes the logit vector produced before the softmax layer for input x, and τ is the temperature hyperparameter and is set to 1.0. To regularize the distributions of clean images and reconstructed images to make them consistent, Jensen–Shannon (JS) divergence [45] was employed to measure the distance of different distributions and we used a generalized weighted form to keep it geometrically symmetric. The consistency regularization loss is:

$$\begin{array}{l} L_{cr} = JS(\hat{f}(x_{clean};\tau)\|\hat{f}(x_{recons};\tau)) = \frac{1}{2}KL(\hat{f}(x_{clean};\tau)\|\frac{1}{2}(\hat{f}(x_{clean};\tau)+\hat{f}(x_{recons};\tau))) \\ +\frac{1}{2}KL(\hat{f}(x_{recons};\tau)\|\frac{1}{2}(\hat{f}(x_{clean};\tau)+\hat{f}(x_{recons};\tau))) \end{array} \tag{6}$$

where x_{clean} denotes the clean image; x_{recons} denotes the reconstructed image; and $KL(\cdot)$ represents the basic distribution distance of the Kullback–Leibler (KL) divergence [46].

3.4. Overall Adversarial Loss Function

The loss of object detection task L_{det} includes classification loss L_{cls} and location loss L_{loc}, which is:

$$L_{\text{det}}(x,l,b;\theta) = L_{cls}(x,l;\theta) + L_{loc}(x,b;\theta) \tag{7}$$

where x represents the training sample, l is the label of object class, and b is the bounding box. θ denotes the model parameters. For object detection, the adversarial attack includes classification attack A_{cls} and localization attack A_{loc}, which can be described as:

$$A_{cls} = \underset{\overline{x}\in S}{\text{argmax}} L_{cls}(\overline{x},l;\theta),\ A_{loc} = \underset{\overline{x}\in S}{\text{argmax}} L_{loc}(\overline{x},b;\theta) \tag{8}$$

where $\overline{x}$ is the corresponding adversarial sample of x. Specifically, it satisfies the condition that $\|\overline{x}-x\|_\infty \leq \varepsilon$, where ε is an adversarial perturbation. S is the adversarial sample space. During the adversarial training process of the robust object detector, both clean and adversarial samples are mixed together. The overall aim of the adversarial training is:

$$\underset{\theta}{\text{argmin}} L_{\text{det}}(\overline{x},l,b;\theta) + L_{\text{det}}(x,l,b;\theta) \tag{9}$$

In this paper, for the reconstructed features from the decoder, a reconstruction loss is also utilized for constraint, which can be defined as follows:

$$L_{\text{recons}} = \|x_{recons} - x\|^2 \tag{10}$$

where $\|\cdot\|^2$ indicates the L2-norm. Finally, the total training objective loss L_{total} combines the classification loss, the localization loss, the reconstruction loss L_{recons}, and the reconstruction consistency loss L_{cr}, which is described as:

$$L_{\text{total}} = L_{\text{det}} + \lambda_1 L_{recons} + \lambda_2 L_{\text{cr}} \tag{11}$$

where λ_1 and λ_2 are the hyperparameters and are set to 0.16 and 5, respectively. Algorithm 1 illustrates the details of the overall training algorithm.

Algorithm 1 Overall Adversarial Training

Input: Dataset D, epochs P, batch size B, and adversarial perturbation ε
Output: Model parameters θ
for $epoch \in \{1, 2, \ldots, P\}$ **do**
  for $i \in \{1, 2, \ldots, B\}$ **do**
Sample clean image x_i with label l and bounding box b
    Generate adversarial sample $\overline{x}_i$: $\|\overline{x} - x\|_\infty \leq \varepsilon$
    Select $\overline{x}_i$, which leads to the max adversarial training loss:
    $L_{\text{recons}} \leftarrow \|x_{recons} - x_i\|^2$
    $L_{cr} \leftarrow JS(\hat{f}(x_i; \tau) \| \hat{f}(x_{recons}; \tau))$
    $\underset{\theta}{\text{argmin}} L_{\text{det}} + \lambda_1 L_{recons} + \lambda_2 L_{\text{cr}}$
  end for
end for
Return θ

4. Experiments and Analysis

4.1. Datasets

In this work, three RSI datasets with multi-resolution, including one-class, two-class, and multi-class datasets, were employed to validate the detection robustness and to evaluate the effectiveness of our proposed method. The details of the three datasets are introduced as follows:

The HRSC [47] dataset contains images of two scenarios, including offshore ships and nearshore ships. The data are all obtained from Google Earth and contain 2976 targets in all. The image resolution ranges from 0.4 to 2 m, and the size covers 300 × 300 to 1500 × 900. The dataset is split into training, validation, and test sets, which contain 436 images, 181 images, and 444 images, respectively.

UCAS-AOD [48] has 1510 aerial images with two categories of 14,596 instances. The size of the images is about 659 × 1280 pixels, and the object categories are vehicle and plane. We randomly selected 1110 images for training and 400 for testing.

DIOR [49] is a large-scale, publicly available benchmark for object detection in remote sensing scenarios. It contains 23,463 images and 192,472 instances in total. The image size is 800 × 800. The dataset covers 20 object classes: airplane, airport, baseball field, basketball court, bridge, chimney, dam, expressway service area, expressway toll station, golf course, ground track field, harbor, overpass, ship, stadium, storage tank, tennis court, train station, vehicle, and windmill. Our experiments selected 11,725 images as the training dataset and 11,738 as the test dataset.

4.2. Implementation Details

Our experiments were realized in Pytorch and run with NVIDIA V100 GPUs. Experiments on three typical datasets of RSIs were conducted to evaluate the universality of our method. In order to show the applicability of the model, we also conducted an experiment on the general dataset PASCAL VOC [50]. For robustness evaluation, all of the settings in this paper are consistent with current robust detectors such as MTD [14], CWAT [16], and RobustDet [15]. Specifically, our proposed method is embedded in the one-stage detector SSD [6] with VGG16 as the backbone. We used three common attack algorithms for comparison including PGD [17], CWA [16], and DAG [9]. The visualizations of the attacks on different datasets are shown in Figure 6. Due to the complex background of remote sensing images, the changes of the images cannot be easily found by naked eyes. Here, the texture changes can be seen obviously on the 'sea' part in the HRSC examples. The PGD-20 attacker is set with a budget $\varepsilon = 8$ to generate adversarial examples. For the DAG attack, we conducted 150 steps for an effective attack. We employed stochastic gradient descent (SGD) with a learning rate of 10^{-5}. The Pascal VOC mean average precision (mAP) with the IoU threshold 0.5 was used as the evaluation metric.

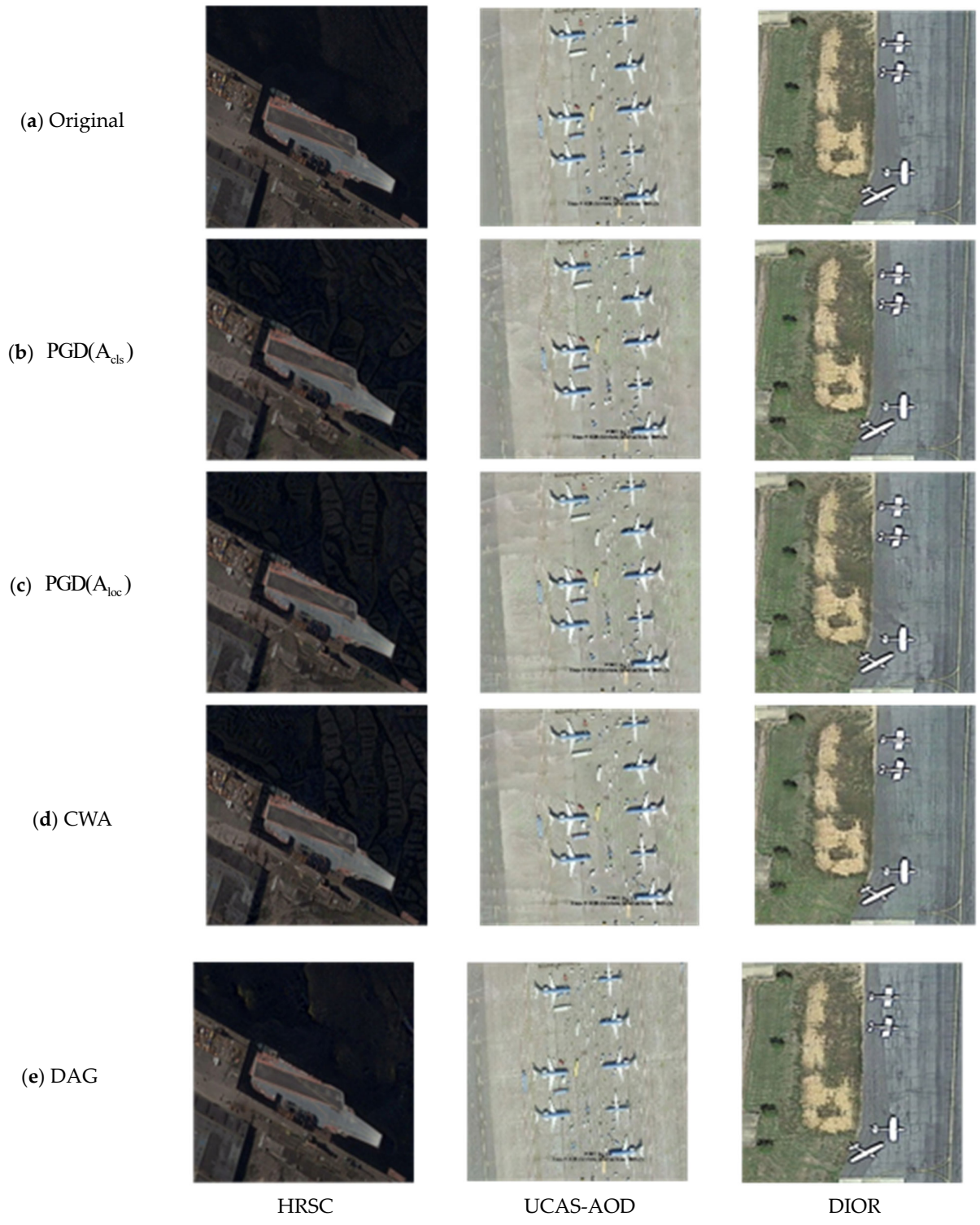

Figure 6. Visual comparison between the clean image and adversarial examples under several attacks on different datasets.

4.3. Ablation Study

We performed a series of ablation experiments to evaluate the performance of our method independently. 'Clean' represents the detector trained with normal training using clean images. Since object detection consists of two tasks: classification and location regression, we divided the PGD attack into classification attack A_{cls} and localization attack A_{reg}.

4.3.1. Ablation Test of MAConv

We assessed the effectiveness of MAConv on HRSC and employed the common DynamicConv [40] used in the current RobustDet for comparison. As shown in Table 1, the performance under different attacks was increased by 1.21%, 3.55%, 4.07%, 4.03%, and 3.90% compared with the original DynamicConv. In addition, MAConv effectively controls the parameter scale and maintains a proper balance between model accuracy and size. Compared with the DynamicConv, the number of convolution layers increased from 155 to 187 using MAConv, and the trainable parameters decreased from 114.73 M to 41.63 M. As MAConv is utilized in the early stage of the neural network to optimize the convolution, it can extract both shared and specific features from clean and adversarial images. Thus, MAConv reduces the training difficulty and improves the feature extraction ability. By increasing the attention dimension instead of simply stacking the network parameters and layers, the accuracy was greatly improved while ensuring the model magnitude. The results demonstrate that MAConv is highly effective at extracting feature information in RSIs.

Table 1. Ablation study of MAConv on HRSC.

Attack / Method	mAP (%)					Params
	Clean	A_{cls}	A_{reg}	CWA	DAG	
DynamicConv	80.98	10.85	11.57	11.12	13.54	114.73 M
MAConv	82.19 (+1.21)	14.40 (+3.55)	15.64 (+4.07)	15.15 (+4.03)	17.44 (+3.90)	41.63 M

4.3.2. Ablation Test of Consistency Regularization Loss

We analyzed the ability of consistency regularization loss under different attacks on the HRSC dataset. Table 2 shows the experimental results. Without regularization, the model exhibits poor robustness under attacks with only 10% to 14% mAP. In contrast, although slight degradation occurs on clean images, the detector achieves substantial performance improvement under all kinds of attacks using regularization loss. In particular, the highest improvement reached 35.87% under DAG attack with only 2.95% performance sacrifice on clean images. This demonstrates that the model predicts the correct distribution under the constraint of regularization loss and thus effectively improves the robustness of the detector. Compared with MAConv, which requires multiple layers to extract features, consistency regularization loss imposes more direct and stronger constraints on reconstructed features for subsequent detection, leading to a higher precision increase.

Table 2. Ablation study of consistency regularization loss on HRSC.

Attack / L_{cr}	mAP (%)				
	Clean	A_{ds}	A_{reg}	CWA	DAG
×	80.98	10.85	11.57	11.12	13.54
√	78.03 (−2.95)	23.53 (+12.68)	39.07 (+27.50)	31.26 (+20.14)	49.41 (+35.87)

4.3.3. Evaluation Using Different Network Architecture

Based on our proposed methods, we evaluated the model using different backbone networks. Besides the original VGG16 backbone in the SSD, we also utilized the ResNet50 backbone for our experiments. The results shown in Table 3 illustrate that our model can increase the accuracy of the robust detector by 3.55 to 4.07% under the VGG16 backbone. Meanwhile, the mAP on the clean image is also slightly increased on the HRSC dataset, about 1.21% compared with the original DynamicConv. For the ResNet50 backbone, MAConv was applied in the base layers of the SSD. The results show that the adversarial attacks still cause performance degradation after changing the backbone. Compared with the original 84.21%, 12.16%, 12.93%, 12.79%, and 14.77% mAP values, our model

improves the performance by 1.36%, 4.38%, 4.16%, 4.22%, and 3.78% under different attacks, respectively. The results illustrate that our model can consistently enhance the adversarial robustness of object detectors in different backbone networks.

Table 3. Ablation study of different network architecture on HRSC.

Backbone \ Attack		mAP (%)				
		Clean	**A_{cls}**	**A_{reg}**	**CWA**	**DAG**
VGG16	DynamicConv	80.98	10.85	11.57	11.12	13.54
	ours	82.19 (+1.21)	14.40 (+3.55)	15.64 (+4.07)	15.15 (+4.03)	17.44 (+3.90)
RESNET50	DynamicConv	84.21	12.16	12.93	12.79	14.77
	ours	85.57 (+1.36)	16.54 (+4.38)	17.09 (+4.16)	17.01 (+4.22)	18.55 (+3.78)

4.4. Overall Comparison

To evaluate the comprehensive performance, our detector was tested on the natural image set VOC07 and three remote sensing datasets: HRSC, UCAS-AOD, and DIOR. For fairness, all experiments were performed under the same conditions. Lastly, the evaluation results were comprehensively analyzed.

4.4.1. Results for the VOC07 Dataset

To verify the effectiveness and generality of the proposed model, we conducted experiments on the VOC07 dataset of natural images. As demonstrated in Table 4, for the clean SSD detector, the adversarial attacks lower the accuracies to less than 5%. The current robust detectors improve the performance under attacks but also degrade a lot on clean images. Our proposed model is theoretically effective at enhancing the detection robustness for natural images. The experimental results also show improvements of 2.8% and 2.7% under PGD attacks compared to clean the SSD. For CWA and DAG attacks, the performance improved by 2.4% and 2.6%, respectively. At the same time, it achieves an increase of 0.8% on clean images.

Table 4. Experimental results on the VOC07 dataset.

Detectors \ Attack	mAP (%)				
	Clean	**A_{cls}**	**A_{reg}**	**CWA**	**DAG**
Clean SSD	77.5	1.8	4.5	1.2	4.9
MTD	48.0	29.1	31.9	18.2	28.5
CWAT	51.3	22.4	36.7	19.9	50.3
RobustDet [15]	74.8	45.9	49.1	48.0	56.6
Ours	75.6	48.7	51.8	50.4	59.2

4.4.2. Results for the HRSC Dataset

Rectangular ships are the main targets in HRSC. We can summarize from the results in Table 5 that there was stable performance improvement in our model under different types of attack. The state-of-the-art robust detector RobustDet shows mAPs of 80.98%, 10.85%, 11.57%, 11.12%, and 13.54% in different cases. In comparison, our model shows enhancements of 2.29%, 4.57%, 4.99%, 4.53%, and 5.93%, respectively. At the same time, it also improves the performance by 2.21% on clean images. With comparable performance to the SSD on clean images, this proves the robustness of the proposed model in dealing with different attacks while surpassing the performance of the current robust detectors. Figure 7 visualizes an example of the detection results of the model on HRSC. In Figure 7, we can see the attack may lead to inaccurate localization of the densely arranged targets, resulting in overlapping bounding boxes.

Table 5. Experimental results on the HRSC dataset.

Attack / Detectors	mAP (%)				
	Clean	**A_{cls}**	**A_{reg}**	**CWA**	**DAG**
Clean SSD	83.20	3.79	6.50	10.07	29.74
MTD	71.13	13.20	14.55	13.77	16.83
CWAT	73.45	14.93	16.08	17.58	15.54
RobustDet	80.98	10.85	11.57	11.12	13.54
Ours	83.27	15.42	16.56	15.65	19.47

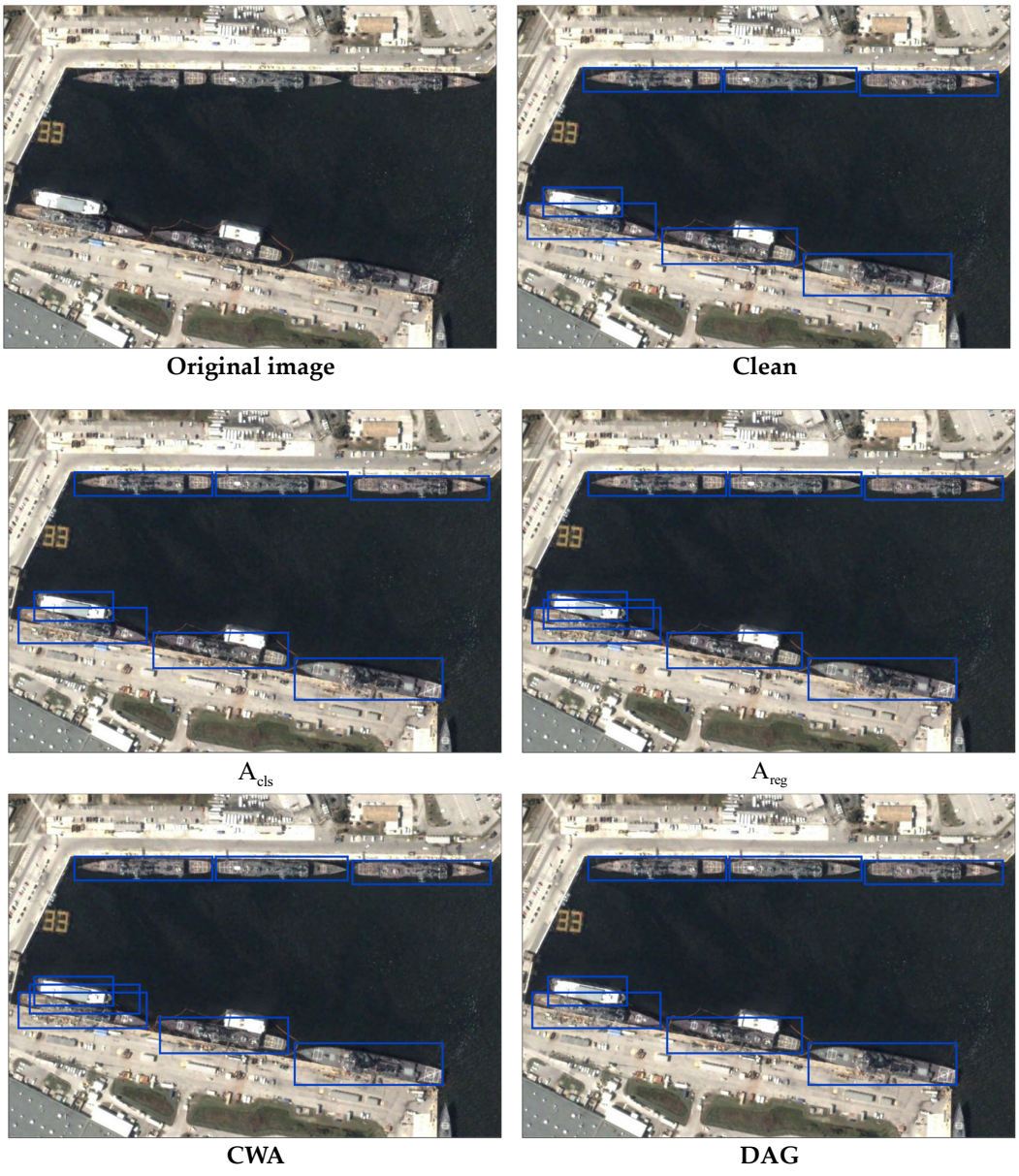

Figure 7. Visualization of the detection results under different attacks on a HRSC example.

4.4.3. Results for the UCAS-AOD Dataset

Table 6 depicts the results of current robust detectors for the UCAS-AOD dataset under mainstream adversarial attacks. The attacks significantly degrade the detector's performance. In particular, for small objects such as vehicles, the accuracy is worse than that

of aircraft, which reduces the overall detection accuracy. The results in Figure 8 demonstrate that adversarial attacks also make the localization of objects more difficult, especially for densely distributed small targets such as cars. Several bounding boxes overlap near the same target. For the clean SSD model, adversarial attacks decrease the mAP to 1.36%, making the detector almost ineffective. This requires robust detectors to improve the performance under attacks while maintaining robustness on clean images. Current robust detectors alleviate this problem to some extent. The results indicate that the proposed model increases the performance by up to 4.34% under different attacks compared with RobustDet, which shows excellent performance on RSIs.

Table 6. Experimental results on the UCAS-AOD dataset.

Attack / Detectors	mAP (%)				
	Clean	**A_{cls}**	**A_{reg}**	**CWA**	**DAG**
Clean SSD	81.15	1.36	3.24	2.63	9.57
MTD	80.84	8.70	10.40	9.67	10.65
CWAT	83.32	17.69	19.55	18.56	20.42
RobustDet	85.49	13.51	16.39	15.99	17.13
Ours	85.34	17.85	19.87	18.45	20.53

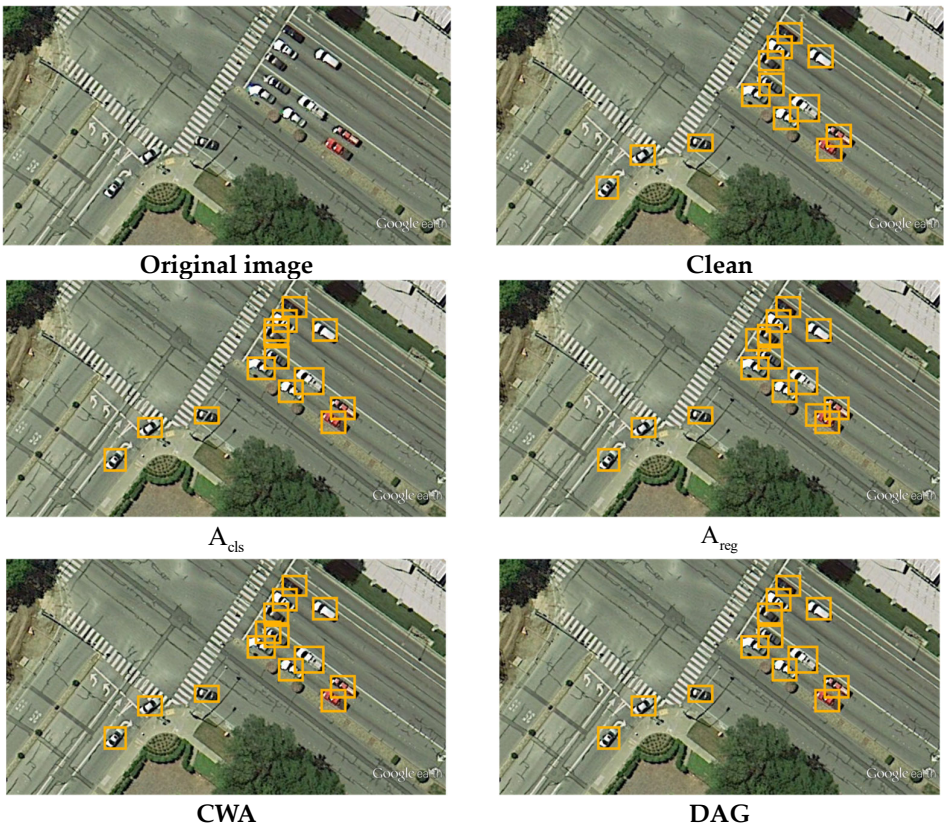

Figure 8. Visualization of the detection results under different attacks on a UCAS-AOD example.

4.4.4. Results for the DIOR Dataset

The DIOR dataset has more complex scenes and multiple target categories, containing dense and sparse objects of different scales. This not only increases the difficulty of detection, but also turns adversarial training into a challenge. The performance results are shown in Table 7. Despite the poor accuracy for some objects of small size, our proposed method still maintains performance and robustness advantages compared to other detectors.

Compared with RobustDet, our model improves mAP by 2.28% on clean images. At the same time, it also has advantages in robustness under different attacks, and the mAP increased by 1.47%, 1.32%, 1.58%, and 1.61%, respectively. The visualization results are presented in Figure 9. The results show that for small targets in normal arrangement, the attack may also make the bounding boxes overlap and lead to inaccurate localization. In general, our model achieves higher overall performance across datasets of different sizes and various types of objects.

Table 7. Experimental results on the DIOR dataset.

Attack / Detectors	mAP (%)				
	Clean	**A_{cls}**	**A_{reg}**	**CWA**	**DAG**
Clean SSD	53.92	0.16	0.32	1.47	2.81
MTD	33.84	1.21	2.07	1.68	2.23
CWAT	38.19	4.13	4.65	4.29	3.92
RobustDet	49.28	5.74	6.02	5.45	6.34
Ours	51.56	7.21	7.34	7.03	7.95

Figure 9. Visualization of the detection results under different attacks on the DIOR example.

5. Discussion

The results of this study show that our methods are effective at enhancing detection robustness. Compared with the latest robust detector RobustDet, the proposed method has a maximum improvement of 5.93%, 4.34%, and 1.61% on the remote sensing datasets HRSC2016, UCAS-AOD, and DIOR under different attacks. These results show that our proposed MAConv and consistency regularization loss play a key role in the robust detector. However, there are still many issues to be discussed for real-world applications. Firstly, in remote sensing scenarios, the coverage of attack is limited, and adversarial attacks on remote sensing targets often appear in the form of patches, which are inconspicuous and difficult to defend against. The specific attack implementation methods are still under development, and defense and robustness studies for such attacks still need to be further carried out in the future. In addition, for large datasets with many categories, especially for small targets with only a few samples, it is still hard to detect issues accurately. This brings difficulty to the adversarial training process of robust detectors. Optimizing the adversarial training process effectively and improving the feature learning strategy are also possible research directions for future works.

6. Conclusions

In this work, we perform a robust object detector for muti-scale objects in remote sensing data. With the aim of alleviating the robustness bottleneck under adversarial attacks, we proposed a multi-dimensional convolution to dynamically learn the specific and consistent features in clean and adversarial images. This method also enriches the feature extraction process for subsequent detection while controlling the network scale. Furthermore, to eliminate the interference of adversarial attacks on accuracy, the reconstruction of the clean image was carried out with a regularization constraint from the perspective of image distribution. The regularization loss contributes to extracting consistent features from the mixture distribution and eliminating the interference brought by adversarial attacks to conduct accurate object detection in complex remote sensing scenarios. The experimental results illustrate that our proposed method successfully increases the accuracy under adversarial attacks and maintains the performance on clean images at the same time. Thus, our method effectively enhances the robustness of the object detector and transcends current robust detectors in terms of performance on RSIs. However, for remote sensing applications, although current robust detectors can resist attacks to some extent, the accuracy under interference is still difficult to reach an application level. Tradeoff between accuracy and robustness of remote sensing object detectors for actual deployment is an important research direction in the future. Additionally, we will conduct the research on robust out-of-distribution detection to further improve the reliability of object detectors for RSIs.

Author Contributions: Conceptualization, Y.L. and Y.F.; methodology, Y.L. and W.L.; writing—original draft preparation: Y.L.; writing—review and editing, Y.F. and W.L.; resources, B.J. and S.W.; and project administration, B.J. and Z.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China under the Grant 61906213.

Data Availability Statement: The source code of the paper is available at https://github.com/yanglcs/AROD-RS (accessed on 10 August 2023).

Acknowledgments: The authors are grateful to the anonymous reviewers and the academic editors for providing valuable comments, which were very beneficial to improving the manuscript.

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

References

1. Li, Z.; Wang, Y.; Zhang, N.; Zhang, Y.; Zhao, Z.; Xu, D.; Ben, G.; Gao, Y. Deep Learning-Based Object Detection Techniques for Remote Sensing Images: A Survey. *Remote Sens.* **2022**, *14*, 2385. [CrossRef]
2. Liu, Y.; Li, Q.; Yuan, Y.; Du, Q.; Wang, Q. ABNet: Adaptive Balanced Network for Multiscale Object Detection in Remote Sensing Imagery. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5614914. [CrossRef]
3. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing Properties of Neural Networks. In Proceedings of the International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
4. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and Harnessing Adversarial Examples. In Proceedings of the International Conference on Learning Representations (ICLR), San Diego, CA, USA, 7–9 May 2015.
5. Song, Y.; Kushman, N.; Shu, R.; Ermon, S. Constructing Unrestricted Adversarial Examples with Generative Models. In Proceedings of the 32nd Conference on Neural Information Processing Systems (NeurIPS), Montreal, QC, Canada, 2–8 December 2018; pp. 8312–8323.
6. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A.C. SSD: Single Shot MultiBox Detector; Lecture Notes in Computer Science. In Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Volume 9905.
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137. [CrossRef] [PubMed]
8. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You Only Look One-Level Feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 13034–13043.
9. Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; Yuille, A. Adversarial Examples for Semantic Segmentation and Object Detection. In Proceedings of the 16th IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1378–1387.
10. Chen, S.; Cornelius, C.; Martin, J.; Chau, D.H.P. ShapeShifter: Robust Physical Adversarial Attack on Faster R-CNN Object Detector. In *Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2019; Volume 11051, p. 52.
11. 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]
12. Lu, M.; Li, Q.; Chen, L.; Li, H. Scale-Adaptive Adversarial Patch Attack for Remote Sensing Image Aircraft Detection. *Remote Sens.* **2021**, *13*, 4078. [CrossRef]
13. Xu, Z.; Yu, F.; Chen, X. LanCe: A Comprehensive and Lightweight CNN Defense Methodology against Physical Adversarial Attacks on Embedded Multimedia Applications. In Proceedings of the Asia and South Pacific Design Automation Conference, Beijing, China, 13–16 January 2020; pp. 470–475.
14. Zhang, H.; Wang, J. Towards Adversarially Robust Object Detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 421–430.
15. Dong, Z.; Wei, P.; Lin, L. Adversarially-Aware Robust Object Detector. In Proceedings of the 17th European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2022; pp. 297–313.
16. Chen, P.-C.; Kung, B.-H.; Chen, J.-C. Class-Aware Robust Adversarial Training for Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 10415–10424.
17. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards Deep Learning Models Resistant to Adversarial Attacks. In Proceedings of the 6th International Conference on Learning Representations (ICLR), Vancouver, BC, Canada, 30 April–3 May 2018.
18. Papernot, N.; McDaniel, P.; Goodfellow, I.; Jha, S.; Celik, Z.B.; Swami, A. Practical Black-Box Attacks against Machine Learning. In Proceedings of the ACM Asia Conference on Computer and Communications Security, Abu Dhabi, United Arab Emirates, 2–6 April 2017; pp. 506–519.
19. Dong, Y.; Su, H.; Wu, B.; Li, Z.; Liu, W.; Zhang, T.; Zhu, J. Efficient Decision-Based Black-Box Adversarial Attacks on Face Recognition. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 16–20 June 2019; pp. 7706–7714.
20. Yang, J.; Jiang, Y.; Huang, X.; Ni, B.; Zhao, C. Learning Black-Box Attackers with Transferable Priors and Query Feedback. In Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS), Virtual Online, 6–12 December 2020.
21. Lu, J.; Sibai, H.; Fabry, E. Adversarial Examples That Fool Detectors. *arXiv* **2017**, arXiv:1712.02494.
22. Li, Y.; Tian, D.; Chang, M.; Bian, X.; Lyu, S. Robust Adversarial Perturbation on Deep Proposal-Based Models. In Proceedings of the 29th British Machine Vision Conference (BMVC), Newcastle, UK, 3–6 September 2018.
23. 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 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 14973–14982.
24. 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.

25. Adhikari, A.; den Hollander, R.; 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 Conference on Artificial Intelligence and Machine Learning in Defense Applications II, Virtual, 21–25 September 2020.
26. Chen, L.; Zhu, G.; Li, Q.; Li, H. Adversarial Example in Remote Sensing Image Recognition. *arXiv* **2019**, arXiv:1910.13222.
27. Xu, Y.; Du, B.; Zhang, L. Self-Attention Context Network: Addressing the Threat of Adversarial Attacks for Hyperspectral Image Classification. *IEEE Trans. Image Process.* **2021**, *30*, 8671. [CrossRef] [PubMed]
28. 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 22nd IEEE/CVF Winter Conference on Applications of Computer Vision January, Waikoloa, HI, USA, 4–8 January 2022; pp. 3798–3808.
29. Xu, Y.; Ghamisi, P. Universal Adversarial Examples in Remote Sensing: Methodology and Benchmark. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5619815. [CrossRef]
30. 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]
31. Hendrycks, D.; Dietterich, T. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
32. Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; Madry, A. Robustness May Be at Odds with Accuracy. In Proceedings of the 7th International Conference on Learning Representations (ICLR), New Orleans, LA, USA, 6–9 May 2019.
33. Maini, P.; Wong, E.; Zico Kolter, J. Adversarial Robustness against the Union of Multiple Perturbation Models. In Proceedings of the 37th International Conference on Machine Learning (ICML), Virtual Online, 13–18 July 2020; pp. 6596–6606.
34. Rice, L.; Wong, E.; Kolter, J.Z. Overfitting in Adversarially Robust Deep Learning. In Proceedings of the 37th International Conference on Machine Learning (ICML), Virtual Online, 13–18 July 2020; pp. 8049–8074.
35. Pang, T.; Yang, X.; Dong, Y.; Su, H.; Zhu, J. Bag of Tricks for Adversarial Training. In Proceedings of the 9th International Conference on Learning Representations (ICLR), Virtual Online, 3–7 May 2021.
36. Tack, J.; Yu, S.; Jeong, J.; Kim, M.; Hwang, S.J.; Shin, J. Consistency Regularization for Adversarial Robustness. In Proceedings of the 36th AAAI Conference on Artificial Intelligence (AAAI), Virtual Online, 22 February–1 March 2022; pp. 8414–8422.
37. Gowal, S.; Rebuffi, S.-A.; Wiles, O.; Stimberg, F.; Calian, D.; Mann, T. Improving Robustness Using Generated Data. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS), Virtual Online, 6–14 December 2021; pp. 4218–4233.
38. Kingma, D.P.; Welling, M. Auto-Encoding Variational Bayes. In Proceedings of the 2nd International Conference on Learning Representations (ICLR), Banff, AB, Canada, 14–16 April 2014.
39. Yang, B.; Bender, G.; Le, Q.V.; Ngiam, J. CondConv: Conditionally Parameterized Convolutions for Efficient Inference. In Proceedings of the Advances in Neural Information Processing Systems 32 (Nips 2019), Vancouver, BC, Canada, 8–14 December 2019; p. 32.
40. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic Convolution: Attention over Convolution Kernels. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Online, 14–19 June 2020; pp. 11027–11036.
41. Li, C.; Zhou, A.; Yao, A. Omni-Dimensional Dynamic Convolution. In Proceedings of the 10th International Conference on Learning Representations (ICLR), Virtual Online, 25–29 April 2022.
42. Xie, C.; Tan, M.; Gong, B.; Wang, J.; Yuille, A.L.; Le, Q.V. Adversarial Examples Improve Image Recognition. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual Online, 14–19 April 2022; pp. 816–825.
43. Zhu, Y.; Chen, Y.; Li, X.; Chen, K.; He, Y.; Tian, X.; Zheng, B.; Chen, Y.; Huang, Q. Towards Understanding and Boosting Adversarial Transferability from a Distribution Perspective. *IEEE Trans. Image Process.* **2022**, *31*, 6487. [CrossRef] [PubMed]
44. Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K.Q. On Calibration of Modern Neural Networks. In Proceedings of the 34th International Conference on Machine Learning (ICML), Sydney, Australia, 6–11 August 2017; pp. 2130–2143.
45. Nielsen, F. On the Jensen—Shannon Symmetrization of Distances Relying on Abstract Means. *Entropy* **2019**, *21*, 485. [CrossRef] [PubMed]
46. Kullback, S.; Leibler, R. On Information and Sufficiency. *Ann. Math. Stat.* **1951**, *22*, 142–143. [CrossRef]
47. Liu, Z.; Yuan, L.; Weng, L.; Yang, Y. A High Resolution Optical Satellite Image Dataset for Ship Recognition and Some New Baselines. In Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM), Porto, Portugal, 24–26 February 2017; pp. 324–331.
48. Zhu, H.; Chen, X.; Dai, W.; Fu, K.; Ye, Q.; Jiao, J. Orientation Robust Object Detection in Aerial Images Using Deep Convolutional Neural Network. In Proceedings of the IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 27–30 September 2015; pp. 3735–3739.

49. Li, K.; Wan, G.; Cheng, G.; Meng, L.; Han, J. Object Detection in Optical Remote Sensing Images: A Survey and a New Benchmark. *ISPRS J. Photogramm. Remote Sens.* **2020**, *159*, 296. [CrossRef]
50. Everingham, M.; Van Gool, L.; Williams, C.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]