remote sensing

MDPI

Article

A Region-Adaptive Local Perturbation-Based Method for Generating Adversarial Examples in Synthetic Aperture Radar Object Detection

Jiale Duan [1,2], Linyao Qiu [3,*], Guangjun He [4], Ling Zhao [1], Zhenshi Zhang [5] and Haifeng Li [1,2]

[1] School of Geosciences and Info-Physics, Central South University, South Lushan Road, Changsha 410083, China
[2] Xiangjiang Laboratory, Changsha 410205, China
[3] China Academy of Electronics and Information Technology, Shuangyuan Road, Beijing 100041, China
[4] State Key Laboratory of Space-Ground Integrated Information Technology, Beijing Institute of Satellite Information Engineering, Beijing 100086, China
[5] College of Basic Education, National University of Defense Technology, Changsha 410073, China
* Correspondence: choulinyao@cetc.com.cn

Abstract: In synthetic aperture radar (SAR) imaging, intelligent object detection methods are facing significant challenges in terms of model robustness and application security, which are posed by adversarial examples. The existing adversarial example generation methods for SAR object detection can be divided into two main types: global perturbation attacks and local perturbation attacks. Due to the dynamic changes and irregular spatial distribution of SAR coherent speckle backgrounds, the attack effectiveness of global perturbation attacks is significantly reduced by coherent speckle. In contrast, by focusing on the image objects, local perturbation attacks achieve targeted and effective advantages over global perturbations by minimizing interference from the SAR coherent speckle background. However, the adaptability of conventional local perturbations is limited because they employ a fixed size without considering the diverse sizes and shapes of SAR objects under various conditions. This paper presents a framework for region-adaptive local perturbations (RaLP) specifically designed for SAR object detection tasks. The framework consists of two modules. To address the issue of coherent speckle noise interference in SAR imagery, we develop a local perturbation generator (LPG) module. By filtering the original image, this module reduces the speckle features introduced during perturbation generation. It then superimposes adversarial perturbations in the form of local perturbations on areas of the object with weaker speckles, thereby reducing the mutual interference between coherent speckles and adversarial perturbation. To address the issue of insufficient adaptability in terms of the size variation in local adversarial perturbations, we propose an adaptive perturbation optimizer (APO) module. This optimizer adapts the size of the adversarial perturbations based on the size and shape of the object, effectively solving the problem of adaptive perturbation size and enhancing the universality of the attack. The experimental results show that RaLP reduces the detection accuracy of the YOLOv3 detector by 29.0%, 29.9%, and 32.3% on the SSDD, SAR-Ship, and AIR-SARShip datasets, respectively, and the model-to-model and dataset-to-dataset transferability of RaLP attacks are verified.

Keywords: synthetic aperture radar; object detection; deep neural network; adversarial example; local perturbation attack

Citation: Duan, J.; Qiu, L.; He, G.; Zhao, L.; Zhang, Z.; Li, H. A Region-Adaptive Local Perturbation-Based Method for Generating Adversarial Examples in Synthetic Aperture Radar Object Detection. *Remote Sens.* **2024**, *16*, 997. https://doi.org/10.3390/rs16060997

Academic Editor: Stefano Tebaldini

Received: 25 December 2023
Revised: 21 February 2024
Accepted: 8 March 2024
Published: 12 March 2024

1. Introduction

SAR technology can provide clear images of ground or maritime targets under all weather conditions and times, playing a crucial role in modern remote sensing. As an important application of SAR image object detection, ship detection is utilized mainly

in maritime surveillance, fishery management, and vessel traffic monitoring [1,2]. Traditional object detection techniques filter SAR images to distinguish between ships and sea backgrounds for object area localization via radar signal processing [3,4]. With the introduction of deep learning, object detection in SAR imagery has achieved significant breakthroughs, substantially improving detection accuracy and efficiency [5]. Additionally, the vulnerability of deep learning-based SAR imagery object detection models has become increasingly prominent. Research on adversarial examples [6,7] has become important for ensuring the security and robustness of SAR imagery object detection models.

The existing adversarial example generation methods aimed at SAR object detection can be divided into two main types: global perturbation attacks and local perturbation attacks. Global perturbation attacks [8–10] aim to deceive object detection systems by uniformly introducing minor perturbations across the entire image without significantly altering its visual appearance. These methods are easily influenced by coherent speckle backgrounds. On the one hand, global perturbation attacks do not distinguish between targets and backgrounds during generation, and the widespread distribution of SAR coherent speckle backgrounds [11–13] means that background information plays a more significant role in the generation of perturbations. The global nature of perturbations leads to the introduction of redundant coherent speckle features, which are not critical factors in model decision-making. Therefore, global perturbation attacks in SAR imagery are more likely to be disregarded by models as noise, thereby weakening their effectiveness. On the other hand, since global perturbation attacks indiscriminately cover the entire image, they can be confused with coherent speckle backgrounds. This causes deep learning models to classify the perturbations as regular scene variables during the recognition process, reducing the specificity and efficacy of the attack.

Local perturbation attacks [14–18] focus on pixel modifications in specific areas of an image, providing a method for precise perturbation of important object areas in object detection. The objective is to subtly disrupt key areas to mislead detection systems without significantly altering the overall image, which can somewhat alleviate the interference of coherent speckle backgrounds in SAR imagery. Although local perturbation attacks enhance the effectiveness of attacks by concentrating perturbations around the object area to reduce interference from the SAR coherent speckle background, this approach still faces challenges in terms of the adaptability of the perturbation size. Particularly when dealing with dynamic changes in speckle backgrounds in SAR imagery, fixed or preset perturbation sizes often fail to effectively adapt to changes in the object area under various conditions. Therefore, designing a method that can adaptively adjust the size of the perturbation to suit the object and background characteristics is crucial for enhancing the effectiveness of local perturbation attacks.

In summary, this paper presents a new solution framework to address interference from coherent speckle backgrounds and adaptability issues related to perturbation size: the region-adaptive local perturbations (RaLP) framework. This framework aims to resolve these challenges through two important modules: (1) the local perturbation generator (LPG), which reduces interference from coherent speckle backgrounds through image filtering preprocessing and concentrates perturbations in the object area to generate precise local adversarial perturbations; and (2) the adaptive perturbation optimizer (APO), which adjusts the size of adversarial perturbations using an adaptive strategy and enhances the visual naturalness of the perturbations by controlling the extent of variation. This approach not only improves the effectiveness of adversarial attacks, but is also applicable to objects of different sizes in SAR imagery, thereby achieving more accurate and efficient object interference. The main contributions of the RaLP framework can be summarized in three points:

1. We propose the region-adaptive local perturbation attack framework (RaLP), which innovatively addresses the key challenges of adversarial examples for SAR object detection: interference from a coherent speckle background and the adaptability of the perturbation size. By comprehensively considering the characteristics of SAR imagery and the require-

ments of the object detection task, this framework implements a novel adversarial attack strategy. It not only enhances the effectiveness and naturalness of adversarial perturbations, but also ensures the universality of the attack.

2. To overcome the interference of coherent speckle backgrounds and the issue of perturbation size adaptability in SAR imagery, we design two innovative modules: the local perturbation generator (LPG) and the adaptive perturbation optimizer (APO). The LPG module, by filtering the original image, effectively mitigates the interference of coherent speckle backgrounds and precisely superimposes adversarial perturbations within the object area, enhancing the specificity and effectiveness of the attack. The APO module introduces a size-adaptive strategy to adjust the size of perturbations, effectively solving the adaptability of perturbation size. It also uses a multiloss function to control the extent of changes in perturbation pixels, thereby increasing the naturalness of the adversarial perturbations.

3. We conducted experiments on multiple datasets, including SSDD, the SAR-Ship-Dataset, and AIR-SARShip-1.0 [19–21], to validate the offensive capabilities of the RaLP adversarial perturbations. The results of these experiments demonstrate the significant impact of RaLP's attack, which reduces the accuracy of the object detection models on these datasets by 29.0%, 29.9%, and 32.3%, respectively, markedly outperforming existing adversarial attack methods. Furthermore, in transferability experiments across multiple datasets and models, RaLP exhibited strong adversarial transferability, achieving the best attack effects at 21.8% and 24.7%, respectively. These results not only verify the effectiveness of the RaLP method in various SAR environments but also demonstrate its powerful ability to adapt to diverse targets and background conditions.

2. Related Work

Recent studies have shown that deep neural networks are susceptible to adversarial examples, posing significant challenges in terms of model robustness and application security for deep learning models. To explore and understand the potential vulnerabilities of deep learning models, Szegedy et al. [6] first proposed the concept of adversarial examples in 2013 and developed an optimization-based method for generating adversarial examples: L-BFGS. This approach introduces meticulously designed minor perturbations into the input data, causing the model to make incorrect predictions, although these perturbations are imperceptible to humans. Over time, the study of adversarial examples has gradually shifted from theoretical exploration to practical application, such as improving the robustness and security of deep learning models. Extending from computer vision to other areas, such as speech and natural language processing [22–26], research on adversarial examples now not only focuses on the effectiveness of attacks, but also includes exploring more covert and practically valuable attack methods [27–31]. These studies offer valuable insights and guidance for improving the security of deep learning models. The existing adversarial example generation algorithms can be categorized from multiple perspectives: according to the attack objective, they can be classified as targeted attacks or nontargeted attacks; based on the attacker's knowledge of the deep neural network, they can be categorized as white-box attacks, black-box attacks, or grey-box attacks. In this paper, adversarial example generation algorithms can also be classified based on the degree of perturbation to the input, such as global perturbation attacks and local perturbation attacks. This section introduces two types of attack methods based on the degree of input perturbation during the adversarial example generation process.

2.1. Global Perturbation Attack

A global perturbation attack is a type of method for adversarial example generation that is designed to mislead a model's prediction by applying subtle yet comprehensive changes to all the input data. The essence of this attack method is to influence the model's understanding of the entire dataset rather than targeting specific areas or features. Research on global perturbation attacks not only exposes the vulnerabilities of deep learning models, but also promotes continuous improvement in model robustness. The strategies for global

perturbation attacks can be broadly divided into several categories based on the principles and techniques used to execute the attack.

Gradient-based attack methods: This type of method generates adversarial perturbations by computing the gradient of the model's loss function with respect to the input, and then adds these adversarial perturbations to the original samples to create adversarial examples. In 2014, Goodfellow et al. proposed the fast gradient sign method (FGSM) algorithm [7], which was the first global perturbation attack method based on gradient attacks. It generates adversarial examples by modifying the image according to the gradient through the backwards propagation of the loss function. Soon after, the FGSM was introduced into adversarial attacks on SAR image samples [32–36] by adding perturbations in the direction of the greatest gradient change within the network model to rapidly increase the loss function, ultimately leading to incorrect classification by the model. The original FGSM algorithm required only a single gradient update to produce adversarial examples; however, as a single-step attack with relatively high perturbation intensity that is only applicable to linear target functions, this approach results in a lower attack success rate. To address this issue, Kurakin et al. extended the FGSM algorithm and proposed the basic iterative method (BIM) [37] algorithm, which perturbs the image through multiple iterations and adjusts the calculation direction after each iteration, solving the problem of the low attack success rate of the FGSM algorithm. Huang et al. [38] designed a variant of BIM that moved from untargeted to targeted attacks by replacing the ground-truth label in the loss function with a target label. However, adversarial examples generated by BIM-like methods can easily become trapped in local maxima due to the limitation of the learning rate step size, thereby affecting the transferability of adversarial examples. To address this, Yin et al. improved upon the BIM algorithm and proposed the momentum iterative fast gradient sign method (MIFGSM) [39] algorithm, which uses momentum to make the direction of the gradient updates more stable, solving the problem of the BIM algorithm becoming trapped in local minima during the generation of adversarial examples. Furthermore, to increase the success rate of attacks, Madry et al. extended the BIM approach and proposed a variant of BIM, the projected gradient descent (PGD) algorithm [40]. This method generates adversarial examples by performing multiple perturbations on the input samples along the sign direction of the gradient using projected gradient descent. The PGD method has been widely applied in adversarial attacks in SAR imagery [33,36,41], enhancing the attack effectiveness by increasing the number of iterations and incorporating a layer of randomization. As an alternative to the aforementioned FGSM method and its variants, Dong et al. proposed a translation-invariant attack method (TIM) [42], in which the recognition areas of the attacked white-box model are less sensitive and the generated adversarial examples have better transferability. This algorithm can be extended to any gradient-based attack method. It uses a convolution operation before applying the gradient to the original image. Compared to those of the FGSM algorithm, the perturbations generated by TIM are smoother.

Optimization-based attack methods: The process of generating adversarial examples can be viewed as finding the optimal perturbation to produce effective adversarial examples. Therefore, adversarial example generation algorithms can be described as solving constrained optimization problems to implement adversarial attacks. The C&W [43] algorithm is a classic optimization-based global perturbation attack method. It is based on iterative optimization strategies using infinity, the 0-norm, and the 2-norm. By adjusting the parameters of the objective function, the algorithm significantly increases the solution space, thereby greatly enhancing the success rate of the adversarial examples. The C&W algorithm has been widely applied in adversarial attacks on SAR imagery [32,33]. However, this approach is limited by drawbacks such as slow training speed and poor transferability. Since the adversarial perturbation for each test sample must be optimized iteratively over a long period, this approach is not suitable for adversarial attack tasks that require an immediate response. Chen et al. proposed the elastic-net attacks to DNNs (EAD) [44] method for generating attack perturbations, which can be viewed as an extension of the

C&W method to the L_1 distance norm. By applying elastic-net regularization, the approach addresses the issue of high-dimensional feature selection in the L_1 norm, thereby finding more effective adversarial perturbations and significantly improving the transferability of global adversarial perturbations. In the SAR domain, to address the slow training speed of SAR image adversarial example generation by the C&W algorithm, Du et al. proposed the fast C&W [45] adversarial example generation algorithm. This algorithm builds a deep encoder network to learn the forwards mapping from the original SAR image space to the adversarial example space. Through this method, adversarial perturbations can be generated more quickly during an attack through fast forwards mapping.

Decision-based attack methods: These methods utilize the principle of hyperplane classification, determining the size of perturbations by calculating the minimum distance between the decision boundary of the original sample and its adversarial counterpart. After obtaining the perturbation vector, it is added to the original sample to generate the adversarial example. The DeepFool [46] algorithm and the HSJA [47] algorithm are classic decision-based global adversarial attack methods. The DeepFool algorithm iteratively generates a perturbation vector pointing towards the nearest decision boundary by iterating over the loss function until the generated adversarial example crosses the decision boundary. In contrast, the HSJA algorithm repeatedly performs gradient direction estimation, geometric series search steps, and binary searches of the estimated decision boundary to generate global adversarial examples. To improve the global generalization ability of the adversarial examples, Moosavi et al. proposed a method called universal adversarial perturbations (UAP) [48], which creates widely applicable adversarial perturbations by computing the shortest distance from the original sample to the classification boundaries of multiple target models.

2.2. Local Perturbation Attacks

Local perturbation attacks are mainly those in which the attacker perturbs only specific areas or parts of the input sample to generate adversarial examples. Unlike global perturbation attacks, they focus on making minor modifications to specific areas of the input data, reducing the overall interference of background speckles while maintaining the effectiveness of the attack. The key to local perturbation attack methods lies in precisely controlling the perturbation area to effectively deceive the model while maintaining the naturalness of the adversarial perturbation. Local perturbation attacks can be categorized into several types based on the perturbation optimization strategy, with each method employing different principles and techniques to execute the attack.

Gradient-based attack methods: Similar to global perturbation attacks, a series of methods for local perturbation attacks exist that optimize perturbations based on gradient strategies. Papernot et al. proposed the Jacobian-based saliency map attack (JSMA) [49] algorithm for generating adversarial examples targeting specific objectives. This algorithm utilizes the Jacobian matrix and saliency map matrix to identify the two pixels with the most influence on the model's classification results within the entire input area. It then modifies those pixels to generate adversarial examples. Dong et al. proposed the superpixel-guided attention (SGA) [50] algorithm, which adds perturbations to similar areas of an image through superpixel segmentation and then converts the global problem into a local problem using class activation mapping information. Additionally, Lu et al. introduced the DFool method, which adds perturbations to 'stop' signs and facial images to mislead the corresponding detectors. This was the first paper to propose the generation of adversarial examples in the field of object detection. DAG [10] is a classic method for adversarial object detection attacks; it yields effective results in actual attacks but is time-consuming due to the need for iterative attacks on each candidate box. Li et al. proposed the RAP attack [51] for two-stage networks, designing a loss function that combines classification and location losses. Compared to the DAG method, Li's method utilizes the location box information in object detection for the attack; however, its actual attack performance is moderate, and its transferability to RPN attacks is poor. In the field of SAR, researchers

have drawn inspiration from gradient-based attack concepts and designed a series of local perturbation attack methods. For instance, the SAR sticker [52] creates perturbations in specific areas of SAR images, maintaining the effectiveness of the attack while enhancing its stealth. Peng et al. proposed the speckle variant attack (SVA) [53], a method for adversarial attacks on SAR remote sensing images. This method consists of two main modules: a gradient-based perturbation generator and an object area extractor. The perturbation generator is used to implement transformations of the background SAR coherent speckle, disrupting the original noise pattern and continuously reconstructing the speckle noise during each iteration. This prevents the generated adversarial examples from overfitting to noise features, which achieves robust transferability. The object area extractor ensures the feasibility of adding adversarial perturbations in real-world scenarios by restricting the area of the perturbations.

Optimization-based attack methods: Su et al. proposed the one-pixel attack algorithm [54], which manipulates a single pixel that can alter the classification of the entire image, thereby deceiving the classification model into mislabelling the image as a specified tag, to some extent achieving a targeted attack. Xu et al. [32] drew inspiration from the one-pixel algorithm to generate local perturbation attacks in SAR images, transforming the creation of adversarial examples into a constrained optimization problem. This method only needs to identify the pixel location to be modified and then use a differential evolution optimization algorithm to perturb that pixel value for a successful attack. Compared to gradient-based adversarial example generation methods, the perturbations created by optimization-based methods are smaller in magnitude and more precise. Furthermore, inspired by sparse adversarial perturbation methods, Meng et al. [55] proposed the TRPG method for local adversarial perturbations when generating adversarial examples in SAR remote sensing images. This method involves extracting the object mask position in SAR images through segmentation, thereby aggregating the perturbations of the SAR adversarial examples into the object area. Finally, adversarial examples that are more consistent with SAR image characteristics are generated via the optimization-based C&W method. Moreover, considering the practicality of local perturbations in the physical world, a category of local perturbation attack methods known as 'adversarial patches' emerged. These local perturbations are applied within a certain area of the input image to attack network models and can be effectively applied in real-world scenarios. The earliest concept of adversarial patches was proposed by Brown et al. [18] in 2017, and the locally generated perturbations could achieve general and targeted attacks on real-world objects. Subsequently, Karmon et al. introduced the localized and visible adversarial noise (LaVAN) [56] method, which focuses more on exploiting the model's vulnerabilities to cause misclassification, with perturbation sizes far smaller than those designed by Brown. Moreover, in the field of object detection, a series of adversarial patch attack methods have been developed. The Dpatch method [16] generates local perturbations and uses them as detection boxes to interfere with detectors, while the Obj-hinder method [15] disrupts detectors by minimizing their class loss. Wang et al. [57] proposed an object detection black-box attack based on particle swarm optimization named EA, which guides the generation of perturbations in appropriate positions using natural optimization algorithms; however, this method is time-consuming.

By deeply exploring the diverse methods and techniques of global and local perturbation attacks, we can more clearly see that adversarial examples have made significant progress in multiple fields. However, when transitioning to SAR image processing, these methods face significant limitations. First, the coherent speckle background in SAR images dynamically changes, which makes maintaining the stability and effectiveness of adversarial examples under different conditions more challenging. Second, the size adaptability of local perturbations is a problem, especially when effective attacks on objects of different sizes and shapes are needed. These challenges require further optimization and adjustment of current adversarial example methods to suit the characteristics and diversity of the scenarios in SAR imagery.

3. Methods

3.1. Problem Formulation

To perform local adversarial attacks on SAR object detection, we first establish an SAR detection dataset X, where each SAR image is denoted as $x \in \mathbb{R}^{C \times H \times W}$. F represents the object detection network, which outputs a predicted label $F(x)$ for each image x, C, H, and W denote the number of channels, height, and width of the SAR image, respectively. The specific expressions are as follows:

$$F(x^*) \neq F(x) \text{ for most } x \in X \tag{1}$$

$$x^* = (1 - M_p) \odot x + M_p \odot p \tag{2}$$

In Equation (1), x and x^* represent the original sample and the adversarial sample, respectively; p and $\odot$ stand for the adversarial perturbation and the Hadamard product, respectively. The matrix $M_\mathbf{p}$ serves as a mask matrix, which is used to limit the location, size, and shape of adversarial perturbations, integrating the perturbations into the object area to perform localized adversarial attacks. To ensure that the mathematical formulas and experimental design presented in this paper can be accurately understood by readers, Table 1 provides clear definitions for all the key symbols and their corresponding terms. In the following sections, we explain in detail how to perform region-adaptive local perturbation (RaLP) for SAR object detection.

Table 1. Symbols and their corresponding terms.

Symbol	Corresponding Term	Symbol	Corresponding Term
X	Raw SAR detection dataset	x	Original Image
$x*$	Adversarial Example	F	Object Detection Model
M_p	Mask matrix	p	Adversarial perturbation
x_{new}	Mask function	C_p	Centre coordinates
$T(\cdot)$	Transition function	$\odot$	Hadamard product

3.2. Region-Adaptive Local Perturbation (RaLP) Framework

Adversarial attacks for SAR object detection face a challenge due to interference from the SAR coherent speckle background. Global adversarial perturbations add unnecessary background information to the feature space and can be hidden by background speckles in the image space. This approach significantly reduces the attack transferability effectiveness. In comparison to global adversarial attacks, local adversarial attacks concentrate more on altering the object area. This approach can reduce the interference from SAR coherent speckle backgrounds to some extent. However, due to the variability in object size and the dynamic changes in the SAR coherent speckle background, the attack effect of local adversarial perturbations is unstable.

To enhance the effectiveness and transferability of perturbations in SAR object detection and achieve target invisibility in the detector's field of view, we propose the region-adaptive local adversarial perturbation (RaLP) framework. The framework consists of two parts: the local perturbation generator (LPG) module and the adaptive perturbation optimizer (APO) module. The overall structure of the framework is illustrated in Figure 1. The LPG module processes the SAR coherent speckle background and adds local adversarial perturbations to the object region. On the other hand, the APO module adjusts the local perturbation size using label information, limits local perturbation pixel changes through a multiloss function, and optimizes the attacking effect of the perturbations through iterative processes. In summary, the LPG provides the RaLP with the necessary perturbations to be optimized. Moreover, the APO ensures that the perturbations applied within these areas are appropriately sized and maximize the attack effect. Together, these methods enable the RaLP framework to effectively generate targeted adversarial examples in various SAR im-

agery environments, significantly improving the adaptability of adversarial perturbations to attacks.

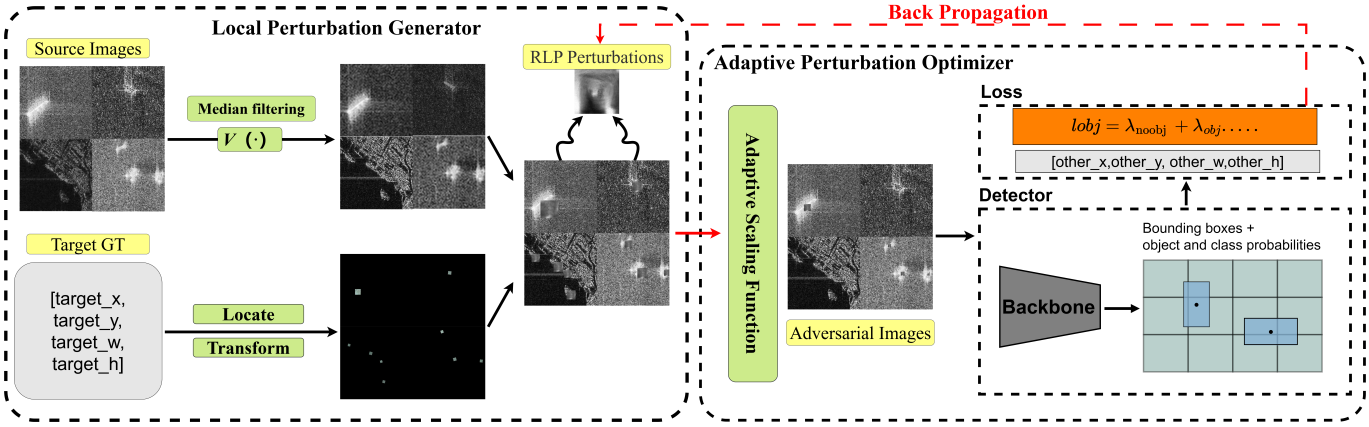


Figure 1. Diagram of the region-adaptive local adversarial perturbation (RaLP) framework.

3.2.1. Local Perturbation Generator

Considering the dynamic changes and spatial irregularities of SAR coherent speckle noise, indiscriminately using global SAR image pixels to generate adversarial perturbations and adding these perturbations to the entire image can result in the perturbations being interfered with and obscured by the coherent speckles, leading to a decrease in the adversarial transferability of the adversarial examples. To address this issue, we introduce the local perturbation generator (LPG) module, which is the first component of the RaLP framework. The LPG module first applies a noise suppression function $M(x)$ to reduce coherent speckle noise in the image while preserving important structural details, aiming to diminish the interference of speckle noise on the perturbations. Subsequently, using a feature enhancement function $E(x)$, the characteristics of the foreground object are enhanced to make them more prominent. Ultimately, a new dataset x_{new} is constructed with reduced background noise characteristics and enhanced foreground object features.

$$x_{\text{new}} = M(x) \odot E(x) \tag{3}$$

Once the image is preprocessed, the module proceeds with perturbation area localization, which includes setting the perturbation size and position. By concentrating the local perturbations on key areas rather than distributing them widely across the entire image, the specificity and effectiveness of the attack are enhanced. To achieve this goal, we require the object annotation information (x_1, x_2, y_1, y_2) to calculate the centre coordinate of the adversarial perturbation, which is denoted as C_p:

$$C_p = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2} \right) \tag{4}$$

The specific expression for constructing the mask matrix M_p is shown in Equation (5), where the transformation function is $T(\cdot)$. The goal is to input the existing object annotations and local adversarial perturbation information to build a mask matrix for merging the adversarial perturbations with the image.

$$M_p = T(p, C_p) \tag{5}$$

Adversarial examples x^* are generated as follows:

$$x^* = (1 - M_p) \odot x_{\text{new}} + M_p \odot p \tag{6}$$

3.2.2. Adaptive Perturbation Optimizer

In the RaLP framework, we design an adaptive perturbation optimization module to perform local adversarial perturbation attacks on multisized objects while controlling the

magnitude of the change in the perturbation pixel. First, the perturbation optimization module adaptively scales the local adversarial perturbations, enabling them to adapt to objects of varying sizes. Considering the diversity of target sizes in SAR imagery, we propose a scale-adaptive strategy to adjust the length and width of the adversarial perturbations (patch_w, patch_h) so that they maintain a proper ratio (w_p, h_p) with the length and width of the target box (sar_w, sar_h), where γ is the scaling factor. The specific expressions are as follows:

$$w_p = \frac{\left(\text{sar}_w * \frac{1}{4}\right)^{\gamma}}{\text{patch}_w} \tag{7}$$

$$h_p = \frac{\left(\text{sar}_h * \frac{1}{4}\right)^{\gamma}}{\text{patch}_h} \tag{8}$$

Incorporating the adaptive sizing strategy makes it necessary to readjust the mask matrix and recombine the perturbation with the imagery to generate new adversarial examples.

$$M_p = T\left(p, C_p, w_p, h_p\right) \tag{9}$$

$$x^* = \left(1 - T\left(p, C_p, w_p, h_p\right)\right) \odot x_{\text{new}} + T\left(p, C_p, w_p, h_p\right) \odot p \tag{10}$$

Subsequently, local perturbation attacks are implemented, and pixel variations in the perturbations are controlled through the constraints of multiple loss functions. The specific objective function has two parts:

1. Adversarial Loss-L_{Adv}: The object confidence is a crucial metric in the object detector, as it indicates the presence or absence of an object in the detection box. Therefore, to perform disappearance attacks on object detectors through local perturbations, we aim to minimize the object confidence of the detection boxes by setting it as the optimization objective. L_{Adv} has two components. The first is the loss of confidence in the detection box containing the object. The second component is the loss of confidence in the detection box where the object is absent. The specific expression is as follows:

$$\begin{aligned} L_{obj} = & \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[\widehat{C}_i^j \log\left(C_i^j\right) + \left(1 - \widehat{C}_i^j\right) \log\left(1 - C_i^j\right)\right] - \\ & \lambda_{\text{noobj}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{\text{noobj}} \left[\widehat{C}_i^j \log\left(C_i^j\right) + \left(1 - \widehat{C}_i^j\right) \log\left(1 - C_i^j\right)\right] \end{aligned} \tag{11}$$

In Equation (11), l_{ij}^{obj}, I_{ij}^{noobj}, and $\widehat{C}_i^j$ are constructed based on the object label information, where l_{ij}^{obj} represents whether the j-th bounding box predictor in the i-th grid contains an object; 1 is the output if it contains a target, and 0 is the output otherwise. I_{ij}^{noobj} gives the opposite outputs. $\widehat{C}_i^j$ represents the ground truth, which is closely related to the value of l_{ij}^{obj}. It outputs 1 if the bounding box contains an object and outputs 0 otherwise. C_i^j represents the probability that the model predicts whether the j-th bounding box in the i-th grid contains an object. λ_{noobj} is a weighting factor used to control the impact of negative samples on the loss function. In summary, to optimize the local adversarial perturbation with L_{Adv} and thereby reduce the confidence of detection boxes containing objects, we intuitively consider converting l_{ij}^{obj} to I_{ij}^{noobj}, which causes the original object detector to lose its ability to detect objects of a specified category. To achieve this, we need to process the input object labels, filter out the label information of the specified category, and only input the label information of the other categories for the loss calculation.

2. Variation Loss-L_{vara}: We incorporate variation loss into the objective loss function to control the pixel changes in perturbations and ensure that adversarial perturbations have

smooth patterns and colour transitions. This loss function calculates the difference between the pixels of adversarial perturbations. The formula for the variation loss is as follows:

$$L_{\text{vara}} = \sum_{i,j} \sqrt{\left(p_{i+1,j} - p_{i,j}\right)^2 + \left(p_{i,j+1} - p_{i,j}\right)^2} \tag{12}$$

$p_{i,j}$ refers to the pixel value of the ith row and jth column of the adversarial perturbation. To summarize, our objective loss function can be divided into two parts: the confidence loss and the variation loss. The expression for the total loss function is as follows, which includes a hyperparameter α that adjusts the weight between the two loss values.

$$L = L_{\text{obj}} + \alpha \cdot L_{\text{vara}} \tag{13}$$

4. Experimental Results and Analysis

4.1. Datasets

The SSDD dataset [19], SAR-Ship-Dataset [21], and AIR-SARShip-1.0 dataset [20] were used to validate the experiments on high-resolution SAR imagery. The images in the SSDD dataset were captured by the RadarSat-2, TerraSAR-X, and Sentinel-1 satellites and included 1160 images with 2456 ship objects. These objects have different scales, orientations, and shapes, and each image is approximately 500 × 500 in size. The SSDD dataset contains four polarization modes, namely, HH, HV, VV, and VH, with a resolution of 1-1-5 m.

The SAR-Ship-Dataset is created using Gaofen-3 and Sentinel-1 SAR data. The dataset comprises a total of 102 views of Gaofen-3 and 108 views of Sentinel-1 SAR images, which are used to construct the high-resolution SAR ship object dataset. The dataset consists of 43,819 images, each with a size of 256 × 256 pixels. It also contains 59,535 ship objects with a resolution range of 3–25 m.

The AIR-SARShip-1.0 dataset was constructed based on Gaofen-3 satellite data. It contains 31 images with dimensions of 3000 × 3000 pixels, featuring a single polarization and resolutions of 1 m and 3 m. The scene types include ports, reefs, and sea surfaces under different sea conditions. The objects cover more than ten types among thousands of ships, including transport ships, oil tankers, and fishing boats. To facilitate the training of the detectors, this paper divides the 3000 × 3000 pixel images in the AIR-SARShip-1.0 dataset into 500 × 500 pixel slices. For the SSDD and SAR-Ship datasets, we directly use the images at their original sizes. Figure 2 shows the diversity of ship objects in the three SAR datasets mentioned in this paper. Finally, we divide the three SAR ship object detection datasets into training and validation sets at a ratio of 8:2, which are used for model training and performance validation, respectively.

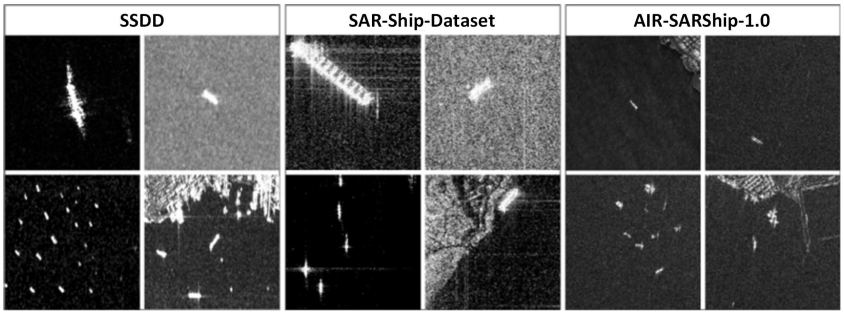


Figure 2. Example image for SAR dataset comparison. Different types of ships show distinct differences in the imagery; transport ships and oil tankers have much larger pixel sizes in the images than fishing boats and other types of ships. The arrangement of ship objects in coastal areas is denser.

4.2. Metrics

1. Quantitative standards: To quantitatively evaluate the effectiveness of the local perturbation attack algorithm, this paper utilizes the mean average precision (mAP) as the performance metric. The methods for calculating the precision and recall are outlined in Equation (14). In these formulas, true positives (TP) refer to detection boxes that have an overlap with the ground truth greater than a certain threshold. False positives (FP) represent two types of detection boxes: those that have an overlap with the ground truth less than the threshold and redundant detection boxes that have an overlap above the threshold but with low category confidence. False negatives (FN) refer to ground-truth boxes that are not successfully detected; the sum of FNand TP equals the total number of real ground-truth labels. Finally, we analyse the results using the precision-recall (PR) curve. The average precision (AP) is the area under the PR curve. The mean average precision (mAP) is calculated by averaging the APs across all categories.

$$\begin{aligned} \text{Precision} &= \frac{TP}{TP + FP} \\ \text{Recall} &= \frac{TP}{TP + FN} \end{aligned} \tag{14}$$

2. Qualitative standards: After generating the adversarial examples, we utilize the interpretable method Grad-CAM [58] for visual analysis, which intuitively demonstrates the effectiveness of the region-adaptive local adversarial perturbation (RaLP) attack method. Grad-CAM calculates the importance weights of each channel's features for target recognition at the last convolutional layer. These importance weights are subsequently used to perform a weighted summation of the feature maps at the last convolutional layer to produce a heatmap. Finally, the heatmap is upscaled through upsampling to the size of the original image and is fused with the original image for display. Equation (15) presents the calculation method for the importance weights in Grad-CAM, where Z denotes the number of pixels in the feature map. y^c represents the predicted score for class c; A^k_{ij} represents the data at position (i, j) in feature layer A for channel k. W^c_k represents the weight corresponding to A^k. Since the probability scores for each point on the heatmap need to be positive, those with negative scores are adjusted using the ReLU function, as shown in Equation (16).

$$w^c_k = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A^k_{ij}} \tag{15}$$

$$L^c_{\text{Grad-CAM}} = \text{ReLU}\left(\sum_k w^c_k A^k \right) \tag{16}$$

4.3. Experimental Setting

4.3.1. Detectors

In this study, YOLOv3 [59] is chosen as the primary detector for SAR object detection tasks, mainly due to its efficiency and wide application in the field of real-time object detection. YOLOv3 not only excels in speed and accuracy, but also has gained widespread recognition in both academic and industrial circles. Moreover, a large number of studies have used YOLOv3 as a benchmark to evaluate the effectiveness of their attack methods, which facilitates comparison with the results of other related research. In addition, we selected the FCOS [60] and Faster RCNN [61] models as experimental models. Faster RCNN, a classic two-stage detection method, is known for its accuracy and robustness; FCOS, as another single-stage detection method, simplifies the model architecture because it is anchor free. By comparing these three models, we comprehensively evaluate the effectiveness of the RaLP attack method on different object detection architectures, demonstrating the broad applicability and effectiveness of the RaLP attack method.

4.3.2. Experimental Setup and Baseline Model Evaluation

To verify the effectiveness of the RaLP method, this study establishes an experimental baseline that includes YOLOv3, FCOS, and Faster R-CNN and covers both single-stage and two-stage object detection methods. Moreover, two approaches, Obj-hinder [15] and DPatch [16], were adopted to conduct a comprehensive test on the robustness of these detection models against the characteristics of SAR images and to compare them with our designed RaLP method. Obj-hinder focuses on assessing the model's sensitivity to real labels through local occlusions, while DPatch explores how to mislead the model by adding local perturbations to the image without changing the characteristics of the target itself.

During the experiment, we trained the YOLOv3, FCOS, and Faster Rcnn detectors on the SSDD dataset, SAR-Ship-Dataset, and AIR-SARShip-1.0 dataset. Subsequently, we used the RaLP, Obj-hinder, and DPatch methods to attack the detectors trained on the above three types of datasets and compared the effectiveness of these three attack methods. Next, we conduct transfer experiments from dataset to dataset, model to model, and across dataset-model combinations to verify the transferability of the RaLP method's attacks. Moreover, we use an adaptive mechanism to adjust the size of the adversarial perturbations and validate the optimization effect of this mechanism by comparing different sizes of adversarial perturbations. Finally, we use deep learning interpretability methods to visualize the detection results and analyse the reasons why the RaLP method is effective.

4.3.3. Parameter Settings

In terms of the experimental details, we first processed the dataset with a background noise treatment, setting a module size of 3 for median filtering of the original images to construct a dataset with more pronounced target features. Subsequently, a series of data augmentation methods were used to transform the adversarial perturbations, including flipping, cropping, warping, brightness transformations, and adding noise. Regarding the hyperparameter training, we set the batch size to 5, the number of epochs to 100, and the initial learning rate to 0.03; the learning rate was adjusted every 50 epochs using a monitor. All the experiments were implemented in PyTorch with Nvidia Tesla V100 GPUs as the computing devices. For fairness, all the experiments were conducted under the same conditions.

4.4. Attack Results

Table 2 presents the experimental results. The YOLOv3 mAP accuracies when trained on the SSDD dataset, SAR-Ship-Dataset, and AIR-SARShip-1.0 dataset were 89.9%, 87.8%, and 88.9%, respectively. After applying the RaLP attack algorithm proposed in this paper, the YOLOv3 mAP detection accuracies decreased to 60.9%, 57.9%, and 56.6%, with decreases of 29.0% for SSDD, 29.9% for the SAR-Ship-Dataset, and 32.3% for AIR-SARShip-1.0, respectively. In a horizontal comparison with other attack algorithms, the RaLP algorithm showed greater reductions—17.3%, 17.8%, and 15.2%—on the three datasets compared to the Dpatch algorithm and reductions of 2.4%, 2.9%, and 1.6% compared to the Obj-hinder algorithm, demonstrating a more effective attack performance than the other two algorithms. Furthermore, we quantitatively assessed the training time for the RaLP perturbations, and the required number of floating-point operations (FLOPs) reached 48.3 G.

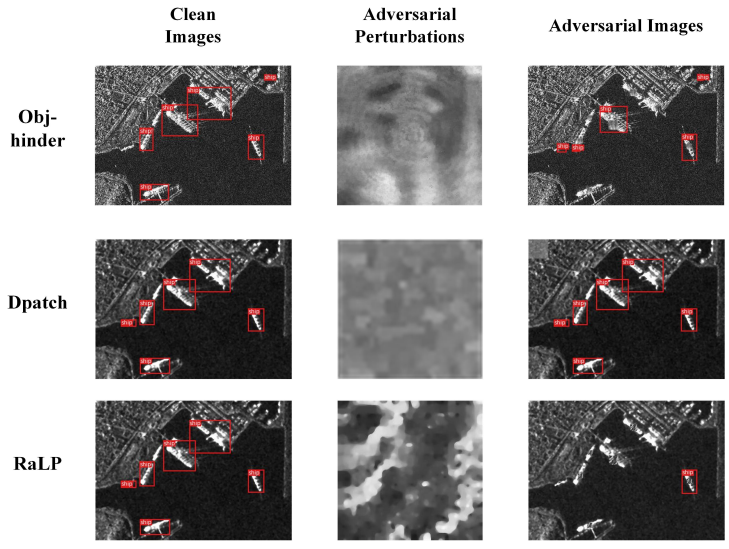
Table 2 shows the superiority of the RaLP attack algorithm through the experimental results, and we analyse them from multiple perspectives: 1. The Dpatch algorithm places perturbations in the corners of the image, making the success of adversarial examples in attacking the object detector highly dependent on the size of the perturbations. Moreover, this method does not apply filtering to SAR coherent speckle backgrounds; thus, it is unable to effectively adapt to dynamic changes in SAR coherent speckle backgrounds, making it challenging to maintain the effectiveness of the attack perturbations. 2. The main limitation of the Obj-hinder attack algorithm is the fixed size of its perturbations, which makes it unable to adapt to the diversity and complexity of SAR targets, impacting the effectiveness of adversarial examples. 3. The RaLP algorithm incorporates speckle filtering and size

adaptability mechanisms to mitigate the impact of the SAR coherent speckle background and perturbation size, demonstrating a more significant attack advantage.

Table 2. Comparison of RaLP attack with other attack algorithms.RaLP demonstrates superior attack performance, which is indicated through the use of bold data representation.

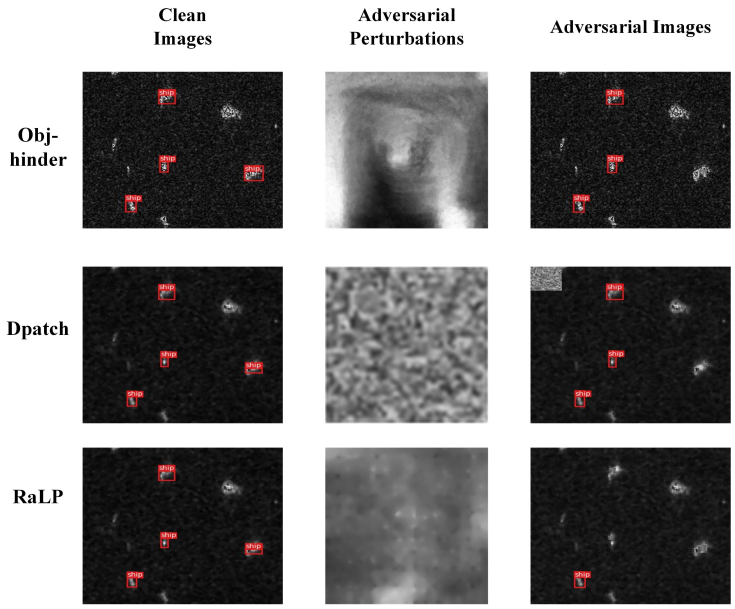
Datasets	Method	mAP		Reduce (↓)
		Clean	Adversarial	
SSDD	Dpatch	89.9%	78.2%	11.7%
	Obj-hinder		63.3%	26.6%
	RaLP		**60.9%**	**29.0%**
SAR-Ship-Dataset	Dpatch	87.8%	75.7%	12.1%
	Obj-hinder		60.8%	27.0%
	RaLP		**57.9%**	**29.9%**
AIR-SARShip-1.0	Dpatch	88.9%	71.8%	17.1%3
	Obj-hinder		58.2%	30.7%
	RaLP		**56.6%**	**32.3%**

In Figure 3a–c, we visualize the adversarial perturbations generated by the three types of attack algorithms on different datasets, organized as one attack algorithm per row, and display the detection results before and after the integration of adversarial perturbations. As shown in Figure 3, regardless of the dataset, the incorporation of our designed RaLP perturbation into the images significantly impacts the object detector, leading to a substantial decrease in the number of detected ship objects. An analysis of the RaLP perturbation compared with the other two types reveals that although the perturbations generated by the three algorithms have similar shapes and sizes, the RaLP perturbation is smoother than the other perturbations. In contrast, the Dpatch perturbation closely resembles random noise.

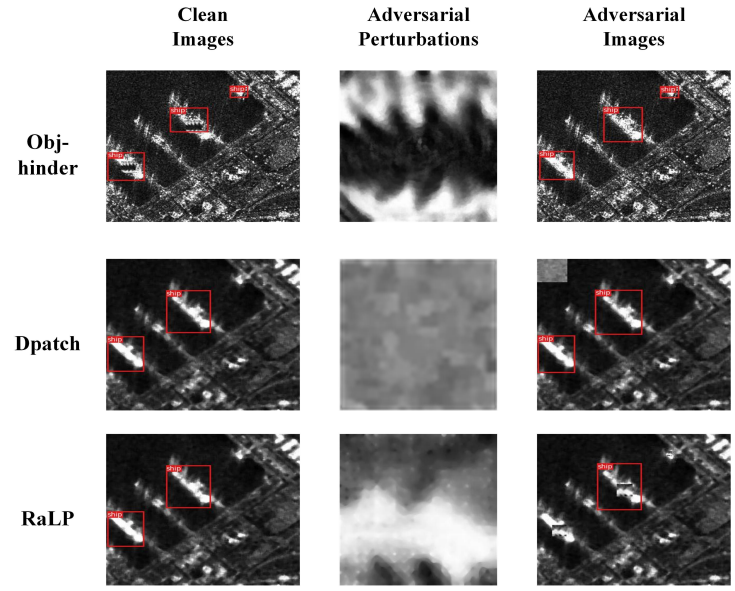


(**a**) SSDD

Figure 3. *Cont.*



(**b**) SAR-Ship-Dataset



(**c**) AIR-SARShip-1.0

Figure 3. The visualization of adversarial perturbations across multiple datasets and their corresponding attack effects. Each subplot corresponds to a specific dataset, and each row within the figure shows the adversarial perturbations generated by a single attack method, as well as the actual images resulting from these perturbations during the attack process.

4.5. Attack Transferability

To conduct a more comprehensive evaluation of the RaLP attack algorithm, we designed experiments in three scenarios to verify the transferability of region-adaptive perturbations: from dataset to dataset (Scenario 1), from model to model (Scenario 2), and across datasets and models (Scenario 3). First, for Scenario 1, we chose YOLOv3 as the sole object detector, using RaLP adversarial perturbations generated on one dataset to attack YOLOv3 detectors trained on two other datasets. Next, for Scenario 2, we trained YOLOv3, FCOS, and FasterRcnn detectors on the same dataset and then used RaLP adversarial perturbations generated on YOLOv3 to attack the FCOS and FasterRcnn detectors. Finally, for Scenario 3, we trained the YOLOv3 detector on one dataset and generated RaLP

adversarial perturbations to attack the FCOS and FasterRcnn detectors trained on two other datasets.

In the experiments, we evaluated the transferability of the RaLP attack algorithm across different datasets using the mAP as the primary metric. For specific scenarios, we meticulously designed three sets of experiments, and the results are comprehensively presented in Tables 3–5. In **Scenario 1**, the experimental results are subdivided into three categories based on the target dataset. Within each category, when the source dataset is consistent with the target dataset, the results reflect the initial efficacy of the attack; in other cases, these results reveal the transferability of RaLP adversarial perturbations across different datasets. As shown in Table 3, the adversarial perturbations generated by the RaLP method demonstrated good attack transferability among the three datasets. When the RaLP adversarial perturbations generated on the SAR-Ship-Dataset and AIR-SARShip-1.0 datasets were transferred to the SSDD dataset, the mAP accuracy of the YOLOv3 detector decreased to 70.4% and 68.1%, respectively, which is a decrease of 19.5% and 21.8% from the original detection results and a reduction of 9.5% and 7.2% compared to the direct attack on the SSDD dataset. Similarly, when transferring the RaLP adversarial perturbations generated on the SSDD and AIR-SARShip-1.0 datasets to the SAR-Ship-Dataset, the mAP accuracy of the YOLOv3 detector decreased to 68.1% and 68.2%, respectively, which is a decrease of 19.7% and 19.6% from the original detection results and a reduction of 10.2% and 10.3% compared to the direct attack on the SAR-Ship-Dataset. When the adversarial perturbations generated on the SSDD and SAR-Ship-Dataset were transferred to the AIR-SARShip-1.0 dataset, the mAP accuracy of the YOLOv3 detector decreased to 68.6% and 73.1%, respectively, which is a decrease of 20.3% and 15.8% from the original detection results and a reduction of 12% and 16.5% compared to the direct attack on the AIR-SARShip-1.0 dataset. By comparing the three sets of results, we found that the RaLP attack perturbations generated from data with the same source exhibit different attack transferability on different target datasets. Notably, the adversarial perturbations obtained on the AIR-SARShip-1.0 dataset showed the best attack transferability on the SSDD dataset. It is important to note that attack transferability does not directly correlate linearly with attack efficacy on the original dataset. The RaLP attack perturbations generated on the AIR-SARShip-1.0 dataset had the strongest attack effect, but their transferability was not the best. The experiments in Scenario 1 demonstrate that RaLP adversarial perturbations trained on one dataset have good attack transferability to other datasets when using the same target detector. Although there are differences in the transferability of RaLP adversarial perturbations generated from different datasets, these differences are minimal.

Scenario 2: We continue to divide the experimental results into three groups according to the dataset. Each group of results demonstrates the effects of adversarial perturbations generated by RaLP on specific datasets using the YOLOv3 object detector, as well as their transferability to black-box detection models. As shown in Table 4, RaLP adversarial perturbations exhibit significant transferability between different detection models. On the SSDD dataset, the adversarial perturbations created by YOLOv3 effectively reduced the mAP accuracy of the two black-box detection models by 22.4% and 24.4%, respectively. Similarly, on the SAR-Ship-Dataset, the perturbations led to a decrease in the mAP accuracy of 24.3% and 24.6% for the two models. On the AIR-SARShip-1.0 dataset, the perturbations effectively reduced the mAP accuracy of the black-box models by 24.7% and 24.0%. Additionally, we found that the impacts of RaLP adversarial perturbations were almost identical on Faster R-CNN and FCOS, with average reductions in mAP of 23.8% and 24.3%, respectively. The experiments in Scenario 2 indicate that the transferability of adversarial perturbations is not affected by the different design philosophies of the anchor-based and anchor-free detection models.

Table 3. Scenario 1: dataset-to-dataset attack transferability.

Source Dataset	Target Dataset	Clean	Adversarial	Reduce (↓)
SSDD	SSDD	89.9%	60.9%	29.0%
SAR-Ship-Dataset			70.4%	19.5%
AIR-SARShip-1.0			68.1%	21.8%
SSDD	SAR-Ship-Dataset	87.8%	68.1%	19.7%
SAR-Ship-Dataset			57.9%	29.9 %
AIR-SARShip-1.0			68.2%	19.6%
SSDD	AIR-SARShip-1.0	88.9%	68.6%	20.3%
SAR-Ship-Dataset			73.1%	15.8%
AIR-SARShip-1.0			56.6%	32.3%

Table 4. Scenario 2: model-to-model attack transferability.

Source Datasets	Model	Clean	Adversarial	Reduce (↓)
SSDD	Faster R-CNN	82.9%	60.5%	22.4%
SSDD	FCOS	80.8%	56.4%	24.4%
SAR-Ship-Dataset	Faster R-CNN	87.4%	63.1%	24.3%
SAR-Ship-Dataset	FCOS	89.9%	65.3%	24.6%
AIR-SARShip-1.0	Faster R-CNN	78.4%	53.7%	24.7%
AIR-SARShip-1.0	FCOS	81.2%	57.2%	24.0%

Table 5. Scenario 3: cross-dataset and model attack transferability.

Source Dataset	Target Dataset	Model	Clean	Adversarial	Reduce (↓)
SSDD	SAR-Ship-Dataset	Faster R-CNN	94.0%	85.1%	8.9%
		FCOS	95.5%	88.5%	7.0%
	AIR-SARShip-1.0	Faster R-CNN	95.9%	73.1%	22.8%
		FCOS	59.8%	44.6%	15.2%
SAR-Ship-Dataset	SSDD	Faster R-CNN	98.0%	82.0%	16.0%
		FCOS	91.1%	82.3%	8.8%
	AIR-SARShip-1.0	Faster R-CNN	95.9%	82.6%	13.3%
		FCOS	59.8%	44.3%	15.5%
AIR-SARShip-1.0	SAR-Ship-Dataset	Faster R-CNN	94.0%	79.9%	14.1%
		FCOS	95.5%	85.0%	10.5%
	SSDD	Faster R-CNN	98.0%	83.5%	14.5%
		FCOS	91.1%	80.1%	11.0%

Scenario 3: We investigated the combined effect of RaLP adversarial perturbations on attack transferability across different datasets and models. The experiments followed the setup of Scenario 2, categorizing the results into three groups based on the dataset, generating RaLP adversarial perturbations from a source dataset, and testing them on black-box detection models trained on other datasets. The results presented in Table 5 indicate that the RaLP perturbations exhibit a certain level of attack transferability between black-box models trained on various datasets. For instance, when the SSDD dataset was used as the source, the YOLOv3-generated adversarial perturbations caused 8.9% and 7% decreases in mAP for models trained on the SAR-Ship-Dataset and 22.8% and 15.2% decreases for models trained on the AIR-SARShip-1.0 dataset. When using the SAR-Ship-Dataset as the source, the adversarial perturbations led to 16% and 8.8% reductions in mAP for models trained on the SSDD dataset and 13.3% and 15.5% reductions for models trained on the AIR-SARShip-1.0 dataset. With AIR-SARShip-1.0 as the source, the perturbations caused a 14.1% and 10.5% decrease in mAP for models trained on the SAR-Ship-Dataset and a 14.5% and 11.0% decrease for SSDD-trained models. The findings from Scenario 3

not only further confirm that the design philosophy of different models has a minor impact on the transferability of adversarial perturbations but also highlight the significant effect of dataset variation on the effectiveness of the attacks.

4.6. Attack Effectiveness of Region-Adaptive Local Adversarial Perturbations

The magnitude and distribution of adversarial perturbations profoundly impact the effectiveness of adversarial attacks. Previous methods of generating local perturbations involved designing perturbations of a fixed size to be combined with image targets. However, in SAR ship detection datasets, the target sizes are diverse, ranging from ocean-going tankers to coastal fishing boats. Employing a one-size-fits-all perturbation can lead to extreme cases, such as small targets in images being completely obscured by large perturbations or large targets remaining unaffected by small perturbations. Therefore, in this paper, we employ an adaptive scaling strategy to process adversarial perturbations to accommodate targets of varying sizes, as demonstrated in Figure 4.

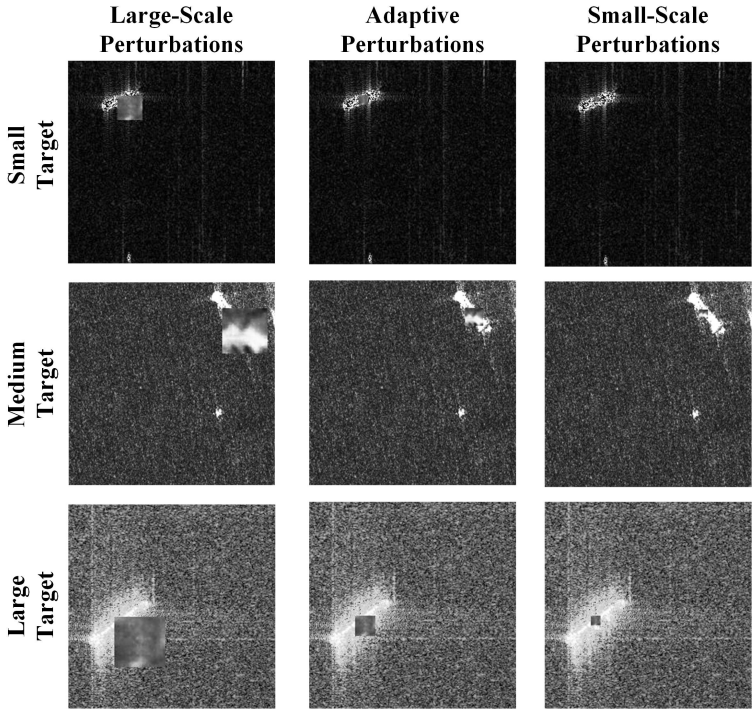


Figure 4. Illustrative examples of adversarial examples created by combining targets of different sizes with perturbations of various sizes. Each row represents adversarial example images formed by combining a target of a specific size with three types of perturbations of different sizes. The combination of adversarial perturbations generated using an adaptive strategy for the targets results in a superior visual effect.

While the adaptive strategy effectively mitigates the extreme cases caused by size discrepancies, the initial size of the region-adaptive local perturbation (RaLP) still significantly influences the potency of the attack. Therefore, this study aims to identify the optimal initial size for RaLP. As shown in Table 6, we tested adversarial perturbations across a spectrum of sizes ranging from 10×10 to 120×120. These findings suggest an upwards trend in the efficacy of RaLP perturbations with increasing size; however, the relationship is not linear. Notably, the size of 100×100 exhibited a peak in adversarial effectiveness. Below this threshold, the impact of the perturbations progressively increased, while sizes beyond this point resulted in a gradual decrease in the effect.

Table 6. Effectiveness of adversarial perturbations of different sizes.

Size	mAP
10 × 10	73.9%
50 × 50	71.7%
80 × 80	67.3%
100 × 100	57.9%
120 × 120	60.9%

Figure 5 presents a visual comparison of the effect of RaLP across varying sizes. The visualization illustrates that as the size of the effect increases, the texture and the informational content become more discernible, leading to an enhanced visual representation and a consistent pattern of textural features. Smaller-sized adversarial perturbations, which possess limited semantic information with overly simplistic patterns, tend to be less effective in terms of disruptive capacity. Conversely, larger perturbations, with their rich pattern details, can more effectively interfere with the detector's ability to capture object features, thereby improving their disruptive impact. However, the adversarial effect of the 120 × 120 RaLP perturbations is less potent than that of the 100 × 100 perturbations. This could be attributed to convergence towards a fixed pattern of texture in the perturbations, where further size increases merely introduce additional noise, which paradoxically weakens the perturbation's intended adversarial effect.

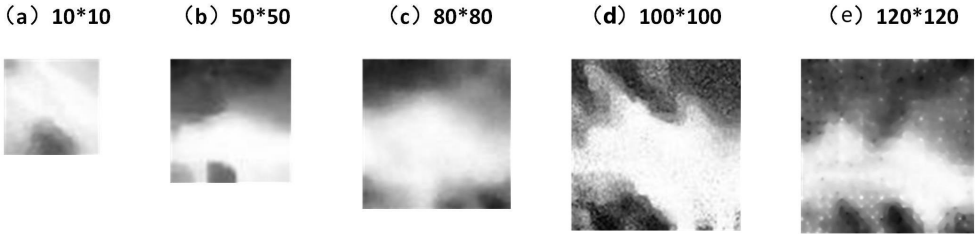


Figure 5. Illustrative diagram of adversarial perturbations of different sizes. From left to right, the sizes of the adversarial perturbations are 10 × 10, 50 × 50, 80 × 80, 100 × 100, and 120 × 120.

In Sections 4.4 and 4.5, we employed the mean average precision (mAP) as a metric for a preliminary evaluation of the attack effectiveness of adversarial examples. While the mAP reflects the model's performance across the entire dataset, it may not reveal the model's vulnerabilities in individual instances in detail. To more comprehensively assess the attack effectiveness of our region-adaptive local perturbation (RaLP) algorithm, we conducted more stringent experimental tests.

During the specific experimental process, we first identified the total number of objects in the images. We then input the original images into the detection model and counted the objects the model could accurately predict and locate. Furthermore, we introduced local perturbations to the object areas based on the annotation information to generate adversarial examples and retested the model's detection capability to observe its detection performance on these perturbed images.

Table 7 shows a significant decrease in the model's detection success rate upon introducing adversarial perturbations across the SSDD dataset, SAR-Ship-Dataset, and AIR-SARShip-1.0 dataset. On the SSDD dataset, the true detection success rate decreased by 38.6%, that on the SAR-Ship-Dataset decreased by 31.0%, and that on the AIR-SARShip-1.0 dataset decreased by 43.0%. These results further confirm the ability of the RaLP algorithm to significantly impair the model's detection capabilities and highlight the importance of considering the model's sensitivity to specific objects when designing adversarial attacks.

Table 7. The impact of the RaLP algorithm on the detection success rate of SAR object detection models.

Datasets	Sample Status	Detection Success Rate	Detection Miss Rate
SSDD	Original Samples	91.3%	8.7%
	Adversarial Samples	52.7%	47.3%
SAR-Ship-Dataset	Original Samples	81.7%	18.3%
	Adversarial Samples	50.7%	49.3%
AIR-SARShip-1.0	Original Samples	87.6%	12.4%
	Adversarial Samples	44.6%	55.4%

4.7. Visual Analysis

Our study demonstrates that RaLP perturbations can effectively diminish the confidence of detection boxes, thereby misguiding the object detector's outcomes. The underlying mechanisms by which RaLP perturbations mislead the detector remain unclear, prompting further investigation.

Figure 6 shows the features extracted by YOLOv3 from both the original and adversarial examples using Grad-CAM [58]. A comparison of the class activation maps of the original samples reveals that the detector focuses on the object regions, scoring them highly. However, when RaLP perturbations are introduced, there is a noticeable shift and attenuation in the focus on these features. This suggests that RaLP perturbations reduce the precision of object detectors by flexibly adjusting the size of the perturbations. This is achieved without obscuring the object; instead, it amplifies the perturbations' interference with the detector's ability to extract object features. As a result, there is a loss of contextual semantic information about the object, leading to false detections. Additionally, the application of speckle filtering techniques to images effectively mitigates interference from the SAR coherent speckle background, enhancing the attack stability of the RaLP adversarial perturbations.

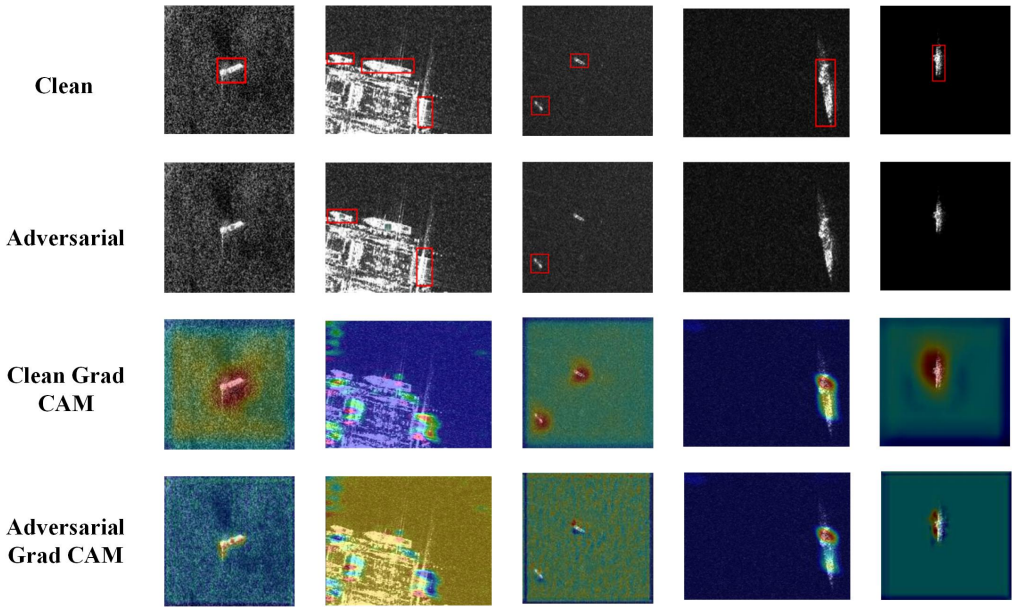


Figure 6. Grad-CAM visualization schematic. Each column in the figure represents the detection results and class activation visualization for the original and adversarial examples.

5. Conclusions and Future Work

In this paper, we introduce an adversarial attack method named region-adaptive local perturbation (RaLP) and provide a thorough explanation and analysis of the framework, algorithm, and experimental results of RaLP. The RaLP method takes into account the

unique properties of SAR imagery by processing the original images through filtering and masking operations to ensure that perturbations are concentrated in the object areas. This effectively reduces interference from coherent speckle background noise, maintaining the efficacy and stability of the adversarial perturbations. During the phase of training adversarial perturbations, we employ various data transformation techniques to enhance the robustness of adversarial perturbations across different operational environments. However, these operations increase the time costs and reduce the training efficiency, so the balance between the effectiveness of adversarial attacks and training efficiency must be considered. Additionally, considering the specificity of SAR object detection tasks and the diversity of objects, RaLP employs an adaptive optimization strategy based on confidence loss. The size of the perturbation is dynamically adjusted according to the object size, significantly enhancing the effectiveness of the attack. The experimental results show that RaLP demonstrates a stronger attack capability than do classic methods such as DPatch and Obj-hinder across multiple datasets, and it exhibits good transferability. A visual analysis further reveals the mechanism behind the successful attacks of RaLP.

The RaLP method exhibits a robust attack capability, averaging more than 10% higher mAP values than the baseline. However, future research needs to address several key issues: (1) Enhanced attack transferability: In light of the potential for insufficient transferability of adversarial examples trained on a single model, we will draw upon pertinent research to explore multimodel training approaches, aiming to improve the transferability of adversarial examples across different models [62,63]. (2) Physical attack design: The current experimental methods involve purely digital exploration and cannot be realized in the physical world. We will further investigate how to generate SAR adversarial examples under real physical constraints, in particular, dynamic and unstable environments. (3) Integration of self-supervised learning and graph neural networks: The effectiveness of self-supervised learning has been substantiated in numerous studies, offering innovative approaches for the design of enhanced adversarial examples [64–66]. Moreover, by integrating graph neural networks [67,68], self-supervised learning facilitates a profound understanding of model vulnerabilities [69], enabling the creation of more elusive adversarial examples. (4) Defence mechanism research: To balance adversarial attacks and defence, exploring effective defence strategies is necessary to enhance the robustness and security of deep neural networks.

Author Contributions: Conceptualization, J.D.; methodology, J.D.; software, J.D.; validation, J.D., L.Z., H.L., L.Q., G.H. and Z.Z.; formal analysis, J.D.; investigation, J.D.; resources, H.L. and L.Z.; data curation, J.D.; writing—original draft preparation, J.D.; writing—review and editing, H.L. and L.Z.; visualization, J.D.; supervision, H.L., L.Z., L.Q., G.H. and Z.Z.; project administration, H.L.; funding acquisition, H.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Major Program Project of Xiangjiang Laboratory under Grant 22XJ01010, the National Natural Science Foundation of China under Grant No. 41801291, and in part by the High-Performance Computing Center of Central South University.

Data Availability Statement: The data associated with this research are available online. The SSDD dataset is available at https://github.com/TianwenZhang0825/Official-SSDD (accessed on 12 March 2023). The SAR-Ship-Dataset is available at https://github.com/CAESAR-Radi/SAR-Ship-Dataset. (accessed on 20 March 2023). The AIR-SARShip-1.0 dataset is available at https://eod-grss-ieee.com/dataset-detail/RE40QTJhTVhjWnVQMWtDRVdPUkRIUT09 (accessed on 6 April 2023).

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

References

1. Migliaccio, M.; Ferrara, G.; Gambardella, A.; Nunziata, F.; Sorrentino, A. A physically consistent speckle model for marine SLC SAR images. *IEEE J. Ocean. Eng.* **2007**, *32*, 839–847. [CrossRef]
2. Adil, M.; Nunziata, F.; Buono, A.; Velotto, D.; Migliaccio, M. Polarimetric scattering by a vessel at different incidence angles. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 561–566. [CrossRef]

3. Gambardella, A.; Nunziata, F.; Migliaccio, M. A physical full-resolution SAR ship detection filter. *IEEE Geosci. Remote Sens. Lett.* **2008**, *5*, 760–763. [CrossRef]
4. Marino, A. A notch filter for ship detection with polarimetric SAR data. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2013**, *6*, 1219–1232. [CrossRef]
5. Chang, Y.L.; Anagaw, A.; Chang, L.; Wang, Y.C.; Hsiao, C.Y.; Lee, W.H. Ship detection based on YOLOv2 for SAR imagery. *Remote Sens.* **2019**, *11*, 786. [CrossRef]
6. Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; Fergus, R. Intriguing properties of neural networks. *arXiv* ***2013***, arXiv:1312.6199.
7. Goodfellow, I.J.; Shlens, J.; Szegedy, C. Explaining and harnessing adversarial examples. *arXiv* ***2014***, arXiv:1412.6572.
8. Song, D.; Eykholt, K.; Evtimov, I.; Fernandes, E.; Li, B.; Rahmati, A.; Tramer, F.; Prakash, A.; Kohno, T. Physical adversarial examples for object detectors. In Proceedings of the 12th USENIX Workshop on Offensive Technologies (WOOT 18), Baltimore, MD, USA, 13–14 August 2018.
9. Chow, K.H.; Liu, L.; Loper, M.; Bae, J.; Gursoy, M.E.; Truex, S.; Wei, W.; Wu, Y. Adversarial objectness gradient attacks in real-time object detection systems. In Proceedings of the 2020 Second IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA), Atlanta, GA, USA, 28–31 October 2020; pp. 263–272.
10. Xie, C.; Wang, J.; Zhang, Z.; Zhou, Y.; Xie, L.; Yuille, A. Adversarial examples for semantic segmentation and object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, October 2017; pp. 1369–1378.
11. Gao, G. Statistical modeling of SAR images: A survey. *Sensors* **2010**, *10*, 775–795. [CrossRef]
12. Tsokas, A.; Rysz, M.; Pardalos, P.M.; Dipple, K. SAR data applications in earth observation: An overview. *Expert Syst. Appl.* **2022**, *205*, 117342. [CrossRef]
13. Singh, P.; Diwakar, M.; Shankar, A.; Shree, R.; Kumar, M. A Review on SAR Image and its Despeckling. *Arch. Comput. Methods Eng.* **2021**, *28*, 4633–4653. [CrossRef]
14. Saha, A.; Subramanya, A.; Patil, K.; Pirsiavash, H. Role of spatial context in adversarial robustness for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 784–785.
15. Thys, S.; Van Ranst, W.; Goedemé, T. Fooling automated surveillance cameras: Adversarial patches to attack person detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 16–20 June 2019; pp. 16–17.
16. Liu, X.; Yang, H.; Liu, Z.; Song, L.; Li, H.; Chen, Y. Dpatch: An adversarial patch attack on object detectors. *arXiv* ***2018***, arXiv:1806.02299.
17. Xu, K.; Zhang, G.; Liu, S.; Fan, Q.; Sun, M.; Chen, H.; Chen, P.Y.; Wang, Y.; Lin, X. Adversarial t-shirt! evading person detectors in a physical world. In Proceedings of the Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; Proceedings, Part V 16; Springer: Heidelberg, Germany, 2020; pp. 665–681.
18. Brown, T.B.; Mané, D.; Roy, A.; Abadi, M.; Gilmer, J. Adversarial patch. *arXiv* ***2017***, arXiv:1712.09665.
19. Zhang, T.; Zhang, X.; Li, J.; Xu, X.; Wang, B.; Zhan, X.; Xu, Y.; Ke, X.; Zeng, T.; Su, H.; et al. SAR ship detection dataset (SSDD): Official release and comprehensive data analysis. *Remote Sens.* **2021**, *13*, 3690. [CrossRef]
20. Wang, Y.; Wang, C.; Zhang, H.; Dong, Y.; Wei, S. A SAR dataset of ship detection for deep learning under complex backgrounds. *Remote Sens.* **2019**, *11*, 765. [CrossRef]
21. Xian, S.; Zhirui, W.; Yuanrui, S.; Wenhui, D.; Yue, Z.; Kun, F. AIR-SARShip-1.0: High-resolution SAR ship detection dataset. *J. Radars* **2019**, *8*, 852–863.
22. Samanta, S.; Mehta, S. Towards crafting text adversarial samples. *arXiv* ***2017***, arXiv:1707.02812.
23. Ebrahimi, J.; Rao, A.; Lowd, D.; Dou, D. Hotflip: White-box adversarial examples for text classification. *arXiv* ***2017***, arXiv:1712.06751.
24. Gao, J.; Lanchantin, J.; Soffa, M.L.; Qi, Y. Black-box generation of adversarial text sequences to evade deep learning classifiers. In Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 24 May 2018; IEEE: Toulouse, France; pp. 50–56.
25. Carlini, N.; Wagner, D. Audio adversarial examples: Targeted attacks on speech-to-text. In Proceedings of the 2018 IEEE Security and Privacy Workshops (SPW), San Francisco, CA, USA, 24 May 2018; pp. 1–7.
26. Kuleshov, V.; Thakoor, S.L.T.E.S. Adversarial examples for natural language classification problems. In Proceedings of the 6th International Conference on Learning Representations (ICLR 2018), Vancouver, BC, Canada, 30 April–3 May 2018.
27. Athalye, A.; Engstrom, L.; Ilyas, A.; Kwok, K. Synthesizing robust adversarial examples. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 284–293.
28. Yakura, H.; Sakuma, J. Robust audio adversarial example for a physical attack. *arXiv* ***2018***, arXiv:1810.11793.
29. Hendrycks, D.; Zhao, K.; Basart, S.; Steinhardt, J.; Song, D. Natural adversarial examples. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 15262–15271.
30. Qin, Y.; Carlini, N.; Cottrell, G.; Goodfellow, I.; Raffel, C. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In Proceedings of the 36th International Conference on Machine Learning (PMLR), Long Beach, CA, USA, 10–15 June 2019; pp. 5231–5240.

31. Duan, R.; Ma, X.; Wang, Y.; Bailey, J.; Qin, A.K.; Yang, Y. Adversarial camouflage: Hiding physical-world attacks with natural styles. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 1000–1008.
32. Zhang, Z.; Liu, S.; Gao, X.; Diao, Y. An Empirical Study Towards SAR Adversarial Examples. In Proceedings of the 2022 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML), Xi'an, China, 28–30 October 2022; IEEE: Toulouse, France, 2022; pp. 127–132.
33. Sun, H.; Xu, Y.; Kuang, G.; Chen, J. Adversarial robustness evaluation of deep convolutional neural network based SAR ATR algorithm. In Proceedings of the 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS, Brussels, Belgium, 11–16 July 2021; pp. 5263–5266.
34. Pang, L.; Wang, L.; Zhang, Y.; Li, H. Adversarial Examples of SAR Images for Deep Learning based Automatic Target Recognition. In Proceedings of the 2021 IEEE 6th International Conference on Signal and Image Processing (ICSIP), Nanjing, China, 22–24 October 2021; IEEE: Toulouse, France, 2021; pp. 24–27.
35. 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]
36. Zhou, J.; Peng, B.; Peng, B. Adversarial Attacks on Radar Target Recognition Based on Deep Learning. In Proceedings of the IGARSS 2022–2022 IEEE International Geoscience and Remote Sensing Symposium, Kuala Lumpur, Malaysia, 17–22 July 2022; IEEE: Toulouse, France, 2022; pp. 2646–2649.
37. 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.
38. Huang, T.; Chen, Y.; Yao, B.; Yang, B.; Wang, X.; Li, Y. Adversarial attacks on deep-learning-based radar range profile target recognition. *Inf. Sci.* **2020**, *531*, 159–176. [CrossRef]
39. Dong, Y.; Liao, F.; Pang, T.; Su, H.; Zhu, J.; Hu, X.; Li, J. Boosting adversarial attacks with momentum. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 9185–9193.
40. Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv* ***2017***, arXiv:1706.06083.
41. Inkawhich, N.; Davis, E.; Majumder, U.; Capraro, C.; Chen, Y. Advanced techniques for robust sar atr: Mitigating noise and phase errors. In Proceedings of the 2020 IEEE International Radar Conference (RADAR), Washington, DC, USA, 28–30 April 2020; IEEE: Toulouse, France, 2020, pp. 844–849.
42. Dong, Y.; Pang, T.; Su, H.; Zhu, J. Evading defenses to transferable adversarial examples by translation-invariant attacks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 16–20 June 2019; pp. 4312–4321.
43. Carlini, N.; Wagner, D. Towards evaluating the robustness of neural networks. In Proceedings of the 2017 IEEE Symposium on Security and Privacy (sp), San Jose, CA, USA, 25 May 2017; pp. 39–57.
44. Chen, P.Y.; Sharma, Y.; Zhang, H.; Yi, J.; Hsieh, C.J. Ead: Elastic-net attacks to deep neural networks via adversarial examples. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, LA, USA, 2–7 February 2018; Volume 32.
45. 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.* **2021**, *19*, 1–5.
46. Moosavi-Dezfooli, S.M.; Fawzi, A.; Frossard, P. Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 2574–2582.
47. 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), San Francisco, CA, USA, 18–21 May 2020; pp. 1277–1294.
48. Moosavi-Dezfooli, S.M.; Fawzi, A.; Fawzi, O.; Frossard, P. Universal adversarial perturbations. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1765–1773.
49. Papernot, N.; McDaniel, P.; Jha, S.; Fredrikson, M.; Celik, Z.B.; Swami, A. The limitations of deep learning in adversarial settings. In Proceedings of the 2016 IEEE European Symposium on Security and Privacy (EuroS&P), Saarbrücken, Germany, 21–24 March 2016; pp. 372–387.
50. Dong, X.; Han, J.; Chen, D.; Liu, J.; Bian, H.; Ma, Z.; Li, H.; Wang, X.; Zhang, W.; Yu, N. Robust superpixel-guided attentional adversarial attack. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 12895–12904.
51. Li, Y.; Tian, D.; Chang, M.C.; Bian, X.; Lyu, S. Robust adversarial perturbation on deep proposal-based models. *arXiv* **2018**, arXiv:1809.05962.
52. Yu, Y.; Zou, H.; Zhang, F. SAR Sticker: An Adversarial Image Patch that can Deceive SAR ATR Deep Model. In Proceedings of the IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium, Pasadena, CA, USA, 16–21 July 2023; IEEE: Toulouse, France, 2023; pp. 7050–7053.
53. Peng, B.; Peng, B.; Zhou, J.; Xia, J.; Liu, L. Speckle-variant attack: Toward transferable adversarial attack to SAR target recognition. *IEEE Geosci. Remote Sens. Lett.* **2022**, *19*, 1–5. [CrossRef]

54. Su, J.; Vargas, D.V.; Sakurai, K. One pixel attack for fooling deep neural networks. *IEEE Trans. Evolut. Comput.* **2019**, *23*, 828–841. [CrossRef]
55. Meng, T.; Zhang, F.; Ma, F. A Target-region-based SAR ATR Adversarial Deception Method. In Proceedings of the 2022 7th International Conference on Signal and Image Processing (ICSIP), Suzhou, China, 20–22 July 2022; pp. 142–146.
56. Karmon, D.; Zoran, D.; Goldberg, Y. Lavan: Localized and visible adversarial noise. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018; pp. 2507–2515.
57. Wang, Y.; Tan, Y.a.; Zhang, W.; Zhao, Y.; Kuang, X. An adversarial attack on DNN-based black-box object detectors. *J. Netw. Comput. Appl.* **2020**, *161*, 102634. [CrossRef]
58. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
59. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
60. Tian, Z.; Shen, C.; Chen, H.; He, T. Fcos: Fully convolutional one-stage object detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636.
61. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *Adv. Neural Inf. Process. Syst.* **2015**, *28*. [CrossRef]
62. Peng, J.; Ye, D.; Tang, B.; Lei, Y.; Liu, Y.; Li, H. Lifelong learning with cycle memory networks. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–14. [CrossRef]
63. Peng, J.; Tang, B.; Jiang, H.; Li, Z.; Lei, Y.; Lin, T.; Li, H. Overcoming long-term catastrophic forgetting through adversarial neural pruning and synaptic consolidation. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 4243–4256. [CrossRef]
64. Zhang, Z.; Ren, Z.; Tao, C.; Zhang, Y.; Peng, C.; Li, H. GraSS: Contrastive Learning With Gradient-Guided Sampling Strategy for Remote Sensing Image Semantic Segmentation. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 1–14. [CrossRef]
65. Li, H.; Cao, J.; Zhu, J.; Luo, Q.; He, S.; Wang, X. Augmentation-Free Graph Contrastive Learning of Invariant-Discriminative Representations. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, 1–11. [CrossRef] [PubMed]
66. Tao, C.; Qi, J.; Guo, M.; Zhu, Q.; Li, H. Self-supervised remote sensing feature learning: Learning paradigms, challenges, and future works. *IEEE Trans. Geosci. Remote Sens.* **2023**, *61*, 5610426. [CrossRef]
67. Zhu, J.; Han, X.; Deng, H.; Tao, C.; Zhao, L.; Wang, P.; Lin, T.; Li, H. KST-GCN: A knowledge-driven spatial-temporal graph convolutional network for traffic forecasting. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 15055–15065. [CrossRef]
68. Li, H.; Cao, J.; Zhu, J.; Liu, Y.; Zhu, Q.; Wu, G. Curvature graph neural network. *Inf. Sci.* **2022**, *592*, 50–66. [CrossRef]
69. Tao, C.; Qi, J.; Zhang, G.; Zhu, Q.; Lu, W.; Li, H. TOV: The original vision model for optical remote sensing image understanding via self-supervised learning. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *33*, 4916–4930. [CrossRef]