applied sciences

MDPI

Article

A Remote Sensing Image Target Detection Algorithm Based on Improved YOLOv8

Haoyu Wang [1], Haitao Yang [2,*], Hang Chen [2,3], Jinyu Wang [1], Xixuan Zhou [1] and Yifan Xu [1]

[1] Graduate School, Space Engineering University, Beijing 101416, China; haoyu20324@163.com (H.W.); 2018205188@qdu.edu.cn (J.W.); zhouxixuan18@mails.ucas.ac.cn (X.Z.); yifan_xu@hgd.edu.cn (Y.X.)
[2] Space Engineering University, Beijing 101416, China; hitchenhang@foxmail.com
[3] University of Lorraine, CNRS, CRAN UMR 7039, 54000 Nancy, France
* Correspondence: yanghtt@126.com

Abstract: Aiming at the characteristics of remote sensing images such as a complex background, a large number of small targets, and various target scales, this paper presents a remote sensing image target detection algorithm based on improved YOLOv8. First, in order to extract more information about small targets in images, we add an extra detection layer for small targets in the backbone network; second, we propose a C2f-E structure based on the Efficient Multi-Scale Attention Module (EMA) to enhance the network's ability to detect targets of different sizes; and lastly, Wise-IoU is used to replace the CIoU loss function in the original algorithm to improve the robustness of the model. Using our improved algorithm for the detection of multiple target categories in the DOTAv1.0 dataset, the mAP@0.5 value is 82.7%, which is 1.3% higher than that of the original YOLOv8 algorithm. It is proven that the algorithm proposed in this paper can effectively improve target detection accuracy in remote sensing images.

Keywords: remote sensing image; target detection; YOLOv8; EMA; Wise-IoU

Citation: Wang, H.; Yang, H.; Chen, H.; Wang, J.; Zhou, X.; Xu, Y. A Remote Sensing Image Target Detection Algorithm Based on Improved YOLOv8. *Appl. Sci.* **2024**, *14*, 1557. https://doi.org/10.3390/app14041557

Academic Editor: Thomas Lindner

Received: 24 January 2024
Revised: 8 February 2024
Accepted: 13 February 2024
Published: 15 February 2024

1. Introduction

Remote sensing image target detection aims to identify specific target objects in remote sensing images, including the identification of feature objects and location determination. Remote sensing image target detection has a wide range of applications in the field of land resource management, national defense, and military and is a current research hotspot in the field of remote sensing image processing [1]. In recent years, with the continuous development of remote sensing sensors, the spatial resolution and spectral resolution of remote sensing images have continued to improve, and the information contained in images has become increasingly complex. Remote sensing images generally have a high level of background complexity, a large image size, a dense distribution of targets, and a large number of small targets, leading to an increase in the difficulty of target detection in remote sensing images.

Traditional target detection methods mostly realize the detection of image information by means of feature extraction, such as histograms of oriented gradients (HOGs) [2], scale-invariant feature transform (SIFT) [3], and local binary pattern histograms (LBPHs) [4]. These traditional methods have certain advantages in terms of extracting local features of the image and the design of classifiers, but they tend not to be well differentiated for the detection of targets in complex scenes, and they cannot effectively detect diverse geo-targets. In recent years, with the significant improvement of computer arithmetic power, deep learning has become a much-anticipated research direction in the field of artificial intelligence. Among deep learning algorithms, target detection algorithms based on convolutional neural networks (CNNs) have achieved faster operation speed and higher detection accuracy due to their weight sharing and invariance of translatability, gradually replacing traditional target detection algorithms and becoming the current

mainstream target detection method. Target detection algorithms based on deep learning can be mainly categorized into two types: two-stage detection and single-stage detection. Two-stage detection first extracts features from the image in the detection process and generates some candidate regions, then outputs the location and category information of the target after classifying and localizing the candidate regions. Typical two-stage detection algorithms include R-CNN [5], Fast R-CNN [6], and Faster R-CNN [7]. Single-stage detection uses a CNN for image feature extraction and passes the extracted features to multiple fully connected layers for target detection, directly outputting the location and category information of the target. Typical single-stage detection algorithms include the YOLO (you only look once) series [8–11] and SSD (single-shot multibox detector) [12,13]. The fact that a single-stage detection algorithm can complete the processing and analysis of the information in the image without pre-generating candidate regions eliminates the need for frequent data conversion and computation, considerably improving the efficiency of the detection. Therefore, single-stage detection algorithms are more widely used than two-stage detection algorithms for large-volume data processing.

Although single-stage detection algorithms represented by the YOLO series can achieve better performance than many target detection methods, remote sensing images are generally characterized by a high proportion of background information, a high degree of target aggregation, and a large number of small targets, making the detection accuracy of the YOLO series lower than that of two-stage detection algorithms. In order to enhance the detection performance of targets in remote sensing images, many researchers have made improvements to the YOLO series. Reference [14] introduces YOLO-SE, a novel YOLOv8-based network. The authors proposed the SEF module, an enhancement based on SEConv, to tackle multi-scale object detection. The authors of [15] developed an improved algorithm based on YOLOv8 called LAR-YOLOv8. They designed an attention-guided bidirectional feature pyramid network to generate more discriminative information by efficiently extracting features from the shallow network through a dynamic sparse attention mechanism and adding top–down paths to guide the subsequent network modules for feature fusion. Reference [16] proposes a BSS-YOLOv8 network model, which incorporates the BoTNet module into the backbone network of the YOLO-v8 model. The completeness of feature extraction was improved by connecting global and local features through its multi-head self-attention mechanism. The authors of reference [17] proposed an enhanced road defect detection algorithm, BL-YOLOv8s, based on YOLOv8s, which was optimized by integrating BiFPN concepts and reconstructing the neck structure of the YOLOv8s model. This optimization reduces the parameters, computational load, and overall size of the model.

From the above researchers' work, it can be seen that the YOLO series has been widely used in many kinds of target detection tasks and that there exists a lot of room for improvement for detection jobs in specific domains. This work provides a remote sensing image target detection method based on the enhanced YOLOv8 algorithm. We achieved good detection results on the DOTA remote sensing dataset. The primary contributions of this paper are summarized as follows:

(1) The addition of an extra detection layer on the basis of the original multi-scale feature fusion network structure to additionally generate a feature map of larger size, improving the network's ability to learn feature information about small targets.
(2) A C2f-E structure based on the Efficient Multi-Scale Attention Module (EMA) [18] is proposed, which enhances the network's detection ability for targets of different sizes through a cross-space learning approach.
(3) The use of Wise-IoU [19] to replace CIoU in the original network, making full use of the wise gradient gain assignment strategy to improve the generalization ability of the model based on the improvement of the overall performance of the detector.

2. Introduction of YOLOv8 Detection Network

YOLOv8 is a deep neural network based on a single-stage target detection algorithm proposed by Ultralytics, which is the latest version of the current YOLO series; its network architecture mainly consists of two parts: a backbone and a head. The backbone is based on the DarkNet53 network, and it completes three downsampling operations of the input image through stacked residual modules and generates three feature maps with different scales and resolutions for input to the first half of head. The first half of head structure uses a feature pyramid network and path aggregation network for feature fusion in order to extract richer target information. The second half of the head part uses a decoupled head structure consisting of classifiers and detectors to classify and regress the position of the feature map. The specific network structure of YOLOv8 is shown in Figure 1.

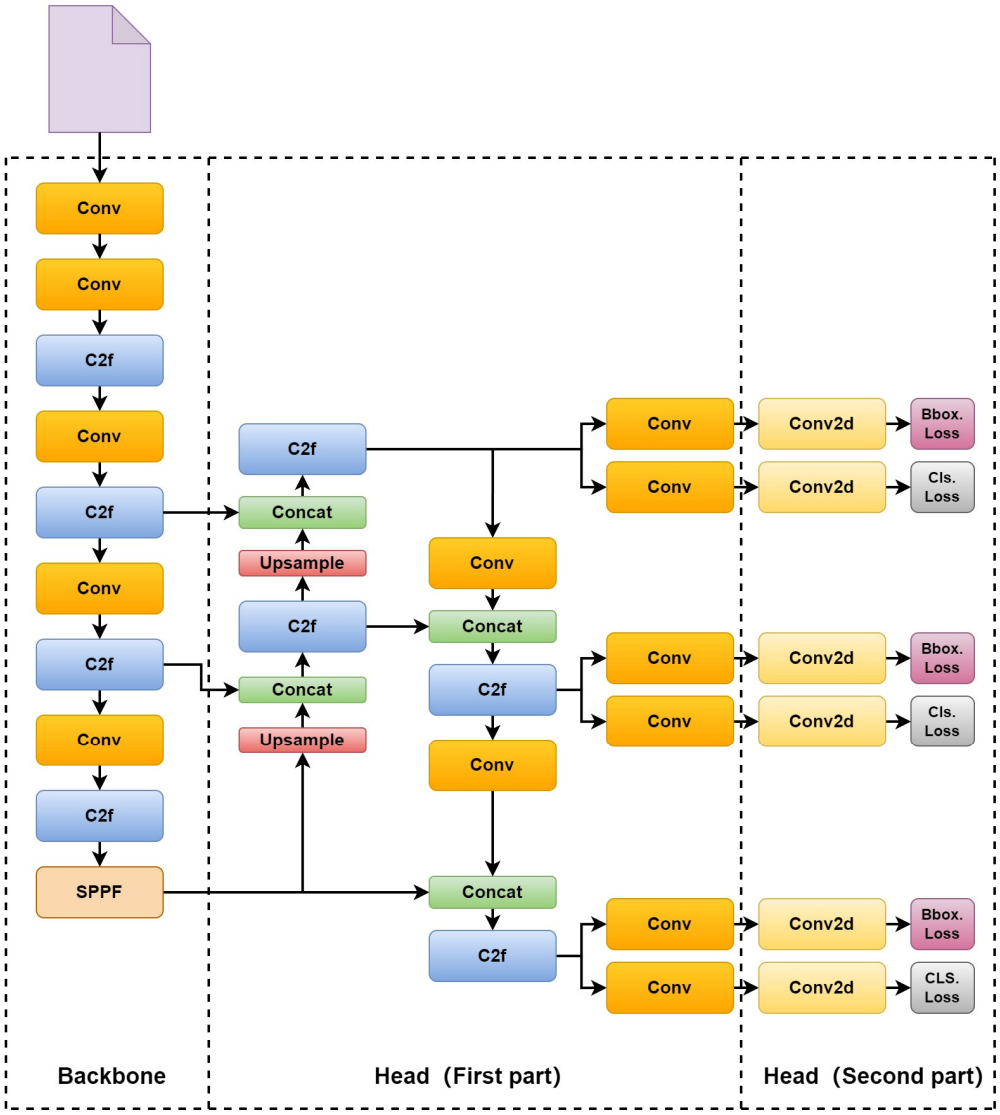

Figure 1. The structure of the YOLOv8 network.

3. Improvement of YOLOv8

3.1. Additional Detection Layer for Small Targets

Due to the small size of the small targets in the remote sensing images, their main feature information is mostly distributed in the shallow feature map; the original YOLOv8 network downsampling multiplier is larger, making the feature maps generated by the initial downsampling small in size. After the completion of three downsamplings, the

feature information for the small targets is difficult to fully retain. For this reason, we add an additional detection layer applicable to small targets on the basis of the original YOLOv8 network. The specific improved network structure is shown in Figure 2. After the FPN module in the neck structure generates an 80 × 80 feature map by upsampling, an additional upsampling operation is added to obtain a feature map with a larger size of 160 × 160, which is fused with the 160 × 160 feature map generated by the first C2f module in the backbone structure, then input into the head structure for classification and detection to achieve the extraction of feature information of the small targets in the shallower feature map.

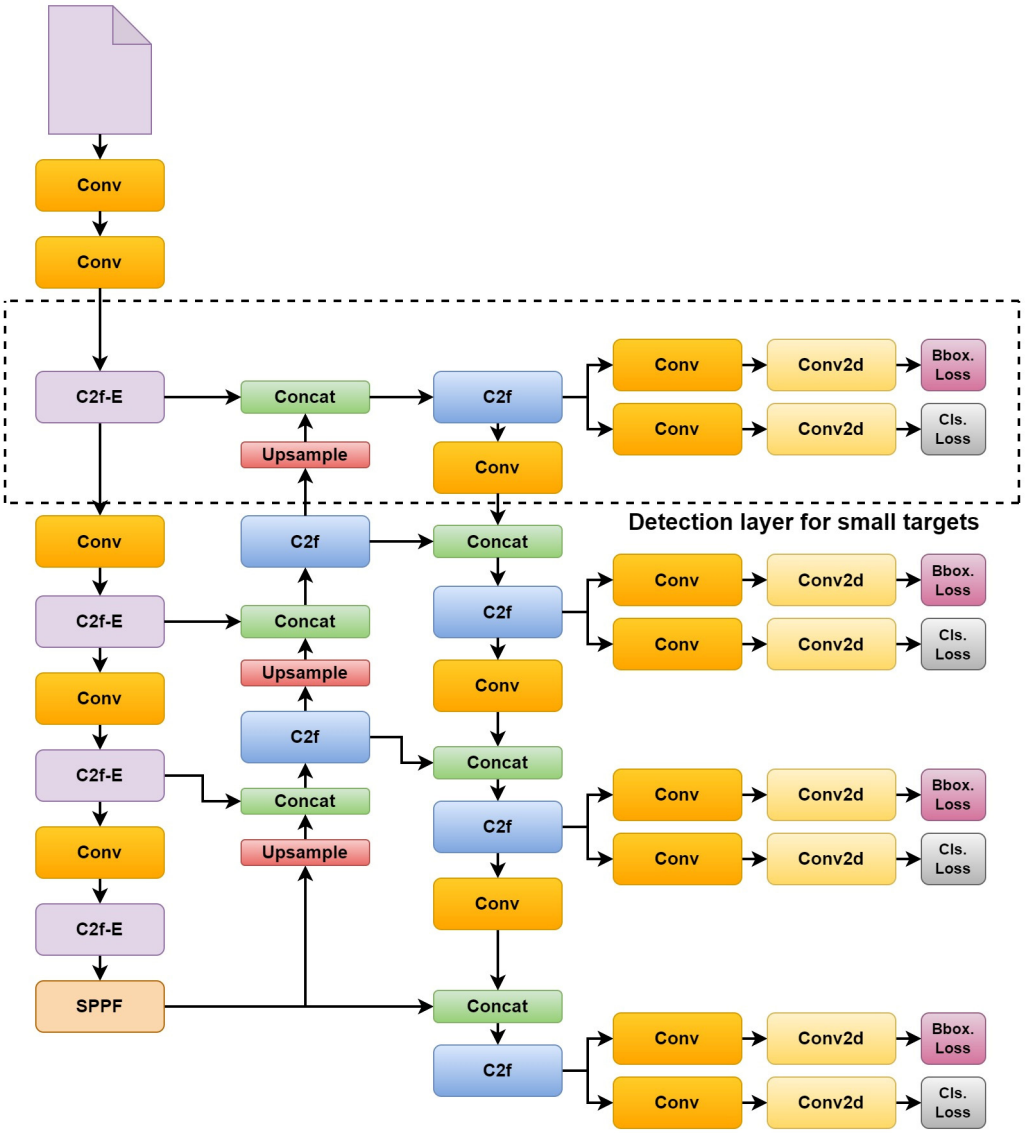

Figure 2. The structure of our network.

3.2. C2f-E Module

In computer vision detection tasks, the effectiveness of feature extraction can have an impact on the final detection accuracy. The introduction of a spatial attention mechanism can generate more recognizable features, which can significantly improve the accuracy of detection. The C2f module is one of the basic modules in the structure of the YOLOv8 network and can provide richer gradient flow information while keeping the network lightweight. The algorithm proposed in this paper is based on the good characteristics of the C2f module and embeds an efficient multi-scale attention module (EMA) based

on cross-spatial learning, which constitutes the C2f-E module and further enhances the network's ability to extract the target features and ensure localization accuracy with a complex background.

The structure of the C2f-E module is shown in Figure 3. After the feature map is convolved and input to the split function, it enters the EMA attention module, which divides the given feature map into N sub-feature maps along the direction of the channel dimension in order to learn different feature information. EMA extracts the attention weights of the sub-feature maps by three parallel routes, where two parallel paths are on the 1 × 1 Conv branch and the third path is on the 3 × 3 Conv branch. In order to capture the dependencies between all the channels and to reduce the computational cost, after encoding the channels, two averaging pooling operations were performed along the X and Y spatial directions, respectively, in the 1 × 1 branch, while only one 3 × 3 convolutional kernel was stacked in the 3 × 3 branch. The module connects the feature information extracted in the two spatial directions of the 1 × 1 branch and inputs it to a 1 × 1 convolution layer, then splits the output into two vectors and utilizes two sigmoid functions to fit the two-dimensional binomial distributions; finally, the obtained results are input into the cross-spatial learning module together with the feature information extracted from the 3 × 3 branch. The cross-spatial learning module establishes the interdependence between channels and spatial locations; its structure is shown in Figure 4. The outputs of the 3 × 3 branch and 1 × 1 branch are first globally average-pooled, and the softmax function is used to fit at the output of the average pooling. The respective spatial attention feature maps on the two branches are obtained by multiplying the output of each branch by the matrix dot product operation. Finally, the output feature information of the two branches is mapped to the set of two spatial attention weight values, which are fed into the subsequent modules by the sigmoid function.

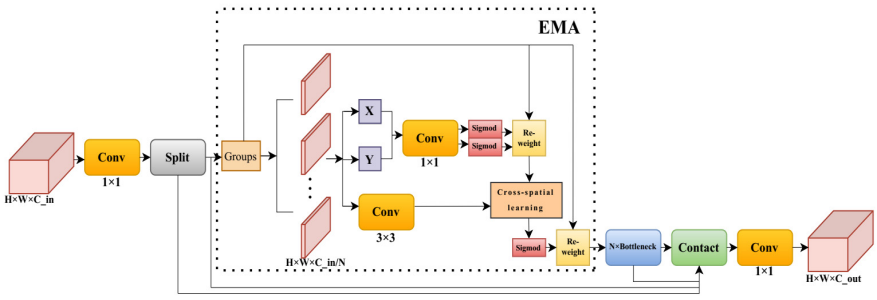

Figure 3. The structure of the C2f-E module.

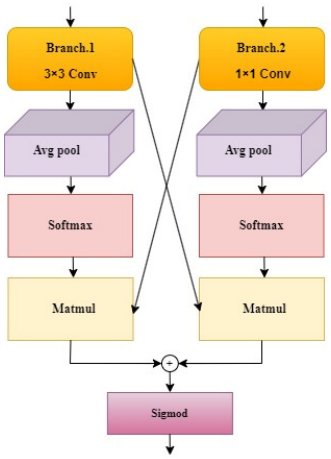

Figure 4. The structure of the cross-spatial learning module.

3.3. Improvement of the Loss Function

In the target detection task, the detection performance depends on the design of the loss function. The IoU loss function is used to measure the overlap between the prediction box derived from the detection and the ground-truth box. Its value ranges between 0 and 1; the closer to 1, the greater the difference between the prediction result and the ground-truth box, and vice-versa.

The classical IoU loss function is schematized in Figure 5 and is calculated as follows:

$$\mathrm{IoU} = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|} \tag{1}$$

$$\mathrm{L_{IoU}} = 1 - \mathrm{IoU} \tag{2}$$

where A is the prediction-box size, and B is the ground-truth box size.

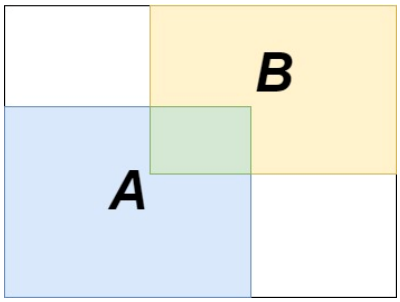

Figure 5. Sketch map of IoU.

In order to continuously improve the localization effect of the prediction box, in recent years, on the basis of the traditional IoU loss function, many new types of loss functions have been produced, such as GIoU, DIoU, CIoU, SIoU, and EIoU. The YOLOv8 network adopts CIoU, which, in contrast with the other loss functions mentioned above, considers the aspect ratio into the computation so that the localization effect of the prediction box is improved. The schematic of CIoU is shown in Figure 6.

$$\mathrm{L_{CIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2(\mathrm{A},\mathrm{B})}{c^2} + \alpha \mathrm{V} \tag{3}$$

$$\mathrm{V} = \frac{4}{\pi^2}\left(\arctan\frac{w_b}{h_b} - \arctan\frac{w_a}{h_a}\right)^2 \tag{4}$$

$$\alpha = \frac{\mathrm{V}}{(1 - \mathrm{I_{IoU}}) + \mathrm{V}} \tag{5}$$

where $\frac{\rho^2(\mathrm{A},\mathrm{B})}{c^2} + \alpha \mathrm{V}$ is a penalty term to minimize the overfitting problem that occurs during model training, where c is the distance between the geometric centers of the two boxes, α is the weight function, and V is the aspect ratio similarity metric function.

However, since the training data will inevitably contain some low-quality examples that are not accurately labeled, the CIoU in the original YOLOv8 network, which only considers geometric factors such as the center spacing of the two boxes and the aspect ratio, inevitably increases the penalty term's punishment for these low-quality examples, thus reducing the overall generalization performance of the model [20]. Therefore, a good loss function is needed to replace the CIoU to achieve the weakening of the influence brought about by geometric factors, then reduce its intervention in the training process when the prediction box matches well with the ground-truth box so that the model can obtain better generalization ability.

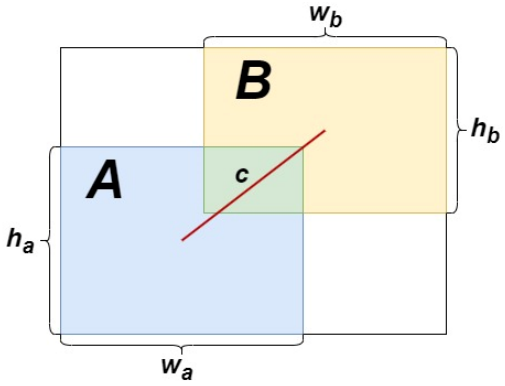

Figure 6. Sketch map of CIoU.

In this paper, Wise-IoU is used to replace the CIoU in the original YOLOv8 network. Wise-IoU provides a more reasonable gradient gain assignment strategy, which reduces the competitiveness of high-quality examples in the gain assignment and, at the same time, reduces the deleterious gradient produced by low-quality examples so that Wise-IoU can act on the ordinary-quality examples in a more focused way, thus improving the generalization ability of the model [21]. Wise-IoU has three versions: Wise-IoUv1, Wise-IoUv2, and Wise-IoUv3. The schematic of WIoU is shown in Figure 7.

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{6}$$

$$R_{WIoU} = \exp\left(\frac{(x_b - x_a)^2 + (y_b - y_a)^2}{W^2 + H^2}\right) \tag{7}$$

where R_{WIoU} takes values between 1 and e, which amplify L_{IoU} for ordinary-quality anchor boxes, while L_{IoU} takes values between 0 and 1, which reduce R_{WIoU} for high-quality anchor boxes in such a way that accelerates the rate of convergence of the loss function without introducing new computational metrics such as aspect ratios.

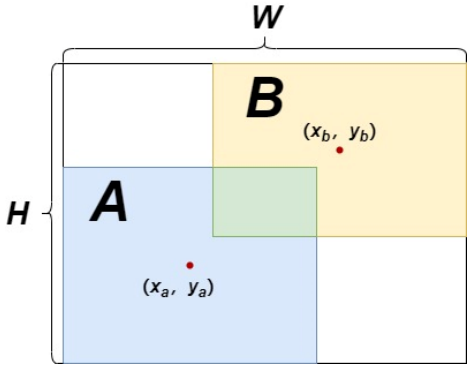

Figure 7. Sketch map of WIoU.

Wise-IoUv2 with monotonic focusing ability can be obtained by introducing the monotonic focusing mechanism for cross-entropy on the basis of Wise-IoUv1.

$$L_{WIoUv2} = \left(\frac{L^*_{IoU}}{\overline{L_{IoU}}}\right)^{\gamma}_{L_{WIoUv1}}, \gamma > 0 \tag{8}$$

where L^*_{IoU} is the monotonic focusing factor of Wise-IoUv1, and $\overline{L_{IoU}}$ is the running mean with momentum. Wise-IoUv2 with monotonic focusing capability can enhance its feature information extraction ability for complex examples, thus improving the classification performance.

By introducing non-monotonic focusing coefficients on the basis of Wise-IoUv1, Wise-IoUv3 with a dynamic non-monotonic focusing mechanism can be obtained.

$$L_{WIoUv3} = \frac{\beta}{\delta \alpha^{\beta-\delta}} L_{WIoUv1} \tag{9}$$

$$\beta = \frac{L_{IoU}}{\overline{L_{IoU}}} \tag{10}$$

where β is the outlier degree of the anchor box. The smaller the value of β, the smaller the IoU loss of the example corresponding to the anchor box and the higher the quality of this anchor box. α and δ are two adjustable parameters that control the mapping relationship between the non-monotonic focusing coefficient and β. Since $\overline{L_{IoU}}$ is not a fixed value, the criteria for determining the quality of the anchor boxes are also dynamic, which allows Wise-IoUv3 to adopt appropriate gradient assignment strategies based on anchor boxes of different qualities.

4. Experimentation and Analysis

4.1. Dataset and Experimental Environment

In this paper, we choose to use Dataset for Object Detection in Aerial Images (DOTA) [22] to train and evaluate our improved model. DOTA is a large image dataset for target detection in aerial images released by Wuhan University in 2017, including three versions: DOTAv1.0, DOTAv1.5, and DOTAv2.0. The DOTAv1.0 version is used in this experiment, which contains a total of 2806 aerial images with resolutions ranging from 800 pixels × 800 pixels to 4000 pixels × 4000 pixels, including 188,282 annotations for 15 types of remote sensing feature targets: plane, ship, storage tank, baseball diamond, tennis court, basketball court, ground track field, harbor, bridge, large vehicle, small vehicle, helicopter, roundabout, soccer ball field, and swimming pool. Compared with other publicly available remote sensing datasets, DOTA v1.0 contains more remote sensing images with small and multi-sized targets, and there are enough samples in each category available for training. Therefore, we selected the DOTA v1.0 dataset to train our network so that the model can achieve improved effectiveness on the remote sensing target detection task. In this experiment, six categories of targets with the highest numbers of labels in the DOTAv1.0 dataset, i.e., small vehicle, large vehicle, plane, ship, storage tank, and harbor, are selected as the targets to be detected; the number and size of each category is shown in Figure 8. These six categories of targets are distributed in different remote sensing backgrounds, and many of them account for a small proportion of the pixels in the image, which can fully reflect the effect of the improved algorithm proposed in this paper on the detection of targets with different levels of background complexity, multiple sizes, and small sizes.

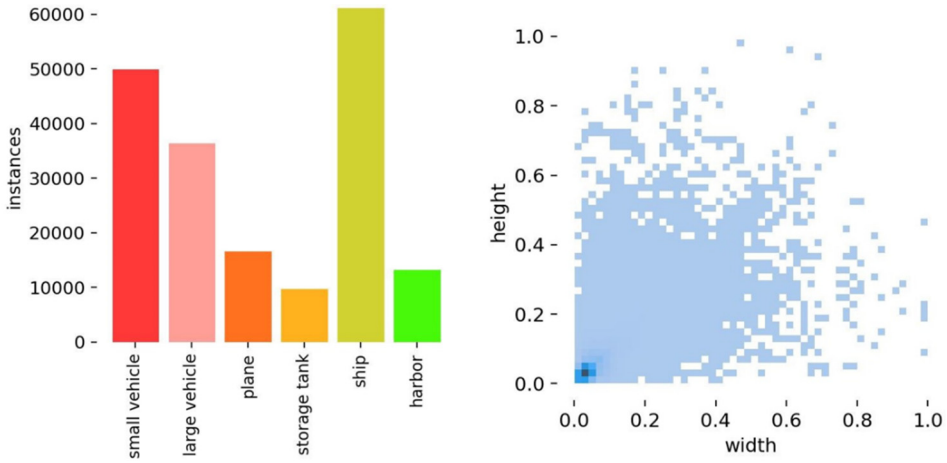

Figure 8. The number and size distribution for each category.

The original image size of the DOTA dataset is large, and using the images directly for training can lead to an excessive amount of computational parameters. Therefore, before training, the original image size is uniformly reduced to 640 pixels × 640 pixels by pixel transformation. In addition to this, we also use mosaic image enhancement on the input images. The mosaic algorithm randomly scales, crops, arranges, and splices four images into a new image for model training, which is equivalent to a lossless improvement of the training epoch and reduces the consumption of graphics card memory while improving the training effect. At the same time, the large targets in the dataset are also randomly scaled to small targets, and the size of the small targets in the original image are scaled to smaller sizes, which increases the number of small targets, realizes the improvement of the diversity of the training data, and improves the model's robustness.

The experiments are based on the PyTorch deep learning framework, and a single NVIDIA GeForce GPU 3090 graphics card is used for model training. The specific configuration of the experimental environment is shown in Table 1.

Table 1. Experimental environment configuration.

Item	Name
Operating system	Windows11
CPU	Intel(R) Core(TM) i9-9820X
GPU	NVIDIA GeForce RTX 3090
RAM	32 G
Deep learning framework	PyTorch (1.13.1)
Interpreter	Python (3.10)
CUDA version	CUDA (11.7)

4.2. Evaluation Indicators

In target detection tasks, P (precision), R (recall), and mAP (mean average precision) are usually chosen to evaluate the performance of target detection algorithms. The specific formulas for P and the R are presented as follows:

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

where TP is the number of positive samples correctly detected by the algorithm, FP is the number of negative samples incorrectly detected as positive samples, and FN is the number of positive samples incorrectly detected as negative samples by the algorithm. A P–R curve can be plotted according to the values of P and R, and integration of the curve can obtain the value of AP, that is, the detection accuracy of a single category in the dataset. The specific formula is shown below:

$$AP = \int_0^1 PRdR \tag{13}$$

For target detection with multiple categories, mAP is obtained by averaging the sum of the AP of each category:

$$mAP = \frac{\sum_{k=1}^{n} AP_k}{n} \tag{14}$$

4.3. Comparison of Loss Functions

In this experiment, the convergences of Wise-IoUv1, Wise-IoUv2, and Wise-IoUv3, CIoU, EIoU, SIoU, and DIoU are verified under the experimental premise of the same network model with equal hyperparameters. The curves of each loss function with the number of training epochs are shown in Figure 9.

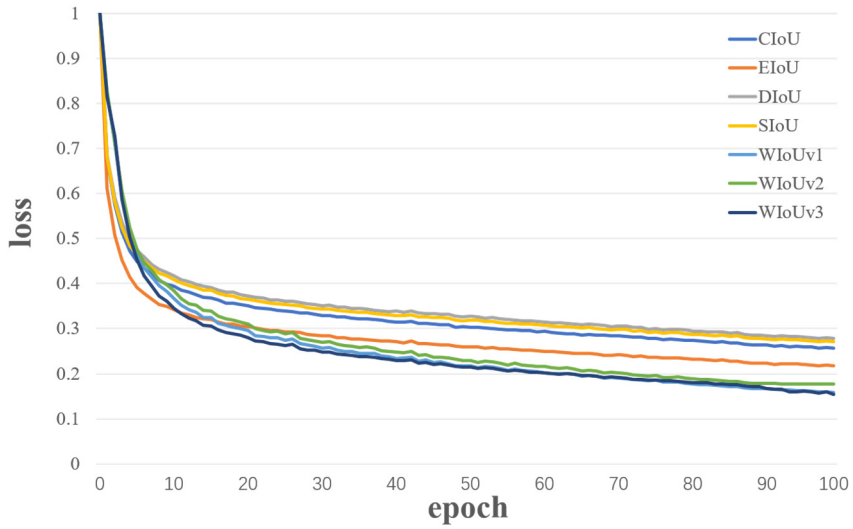

Figure 9. Comparison of different loss functions.

From the experimental results shown in Figure 9, it can be seen that the value of each loss function decreases and converges with an increase in the number of epochs, but Wise-IoUv3 converges faster and has smaller loss values compared to the other functions. Therefore, Wise-IoUv3 is chosen as the bounding-box loss function for the improved network proposed in this paper, which is effective in improving detection performance for remote sensing image targets.

4.4. Attention Module Comparison Test

In order to verify the effectiveness of the C2f structure combined with the EMA module, three typical attention modules, i.e., CBAM [23], GAM [24], and ECA [25], are selected to be embedded in the same position in the C2f structure for a side-by-side comparison of the improved algorithm proposed in this paper; the experimental results are shown in Table 2.

Table 2. Comparisons of different attention modules.

Attention Module	P/%	R/%	mAP@0.5%	mAP@0.5:0.95%
CBAM	83.9	77.2	82.5	54.9
GAM	84.3	77.2	**82.7**	54.8
ECA	83.8	77.1	82.1	54.5
EMA	**84.5**	**77.3**	**82.7**	**55.1**

Based on the results shown in Table 2, it can be seen that the P, R, and mAP@0.5 values obtained by training with the module combining C2f with EMA are superior to those obtained by training using C2f combined with CBAM or ECA. The mAP@0.5:0.95 value is improved by 0.3%, with an equal mAP@0.5 obtained by training of the embedded GAM attention module, which proves that the improvement proposed in this paper is more effective for remote sensing target detection.

4.5. Ablation Study

In order to verify the effectiveness of the three improvement modules proposed in this paper, we use YOLOv8s as a benchmark model for comparison purposes and train on the DOTAv1.0 dataset under the same experimental environment to obtain the test data for each improvement module, the results of which are shown in Table 3, where "Base" denotes the original YOLOv8s model, "$\surd$" denotes the added module, and "$\times$" denotes that the module was not added.

Table 3. Ablation study results.

Dataset	Base	Layer for Small Target	Wise-IoUv3	EMA	Small Vehicle	Large Vehicle	Plane	Storage Tank	Ship	Harbor	mAP@0.5%
DOTA	√	×	×	×	64.6	86.4	91.8	71.5	89.2	**84.9**	81.4
	√	√	×	×	67.1	86.3	91.9	75.4	89.3	81.1	81.8
	√	×	√	×	69.4	86.4	91.7	72.7	88.8	84.1	82.2
	√	√	√	×	69.0	86.6	92.2	75.8	88.9	82.3	82.5
	√	√	√	√	**71.2**	**87.3**	**92.7**	**76.3**	**89.4**	79.2	**82.7**

As seen from the ablation experiment results in Table 3, each of the improvement modules proposed in this paper has a role in improving the detection accuracy. For DOTA v1.0, the mAP@0.5% is improved by 0.4% over the original algorithm when the small target detection layer is added to YOLOv8s alone, and its accuracy is improved by 0.8% when Wise-IoUv3 is introduced alone. When both are introduced together, the mAP@0.5% is improved by 1.1% over the original algorithm, and the detection accuracy is improved for all categories of small targets except plane. When all three modules are added, the mAP@0.5% of our algorithm is improved by 1.3% compared with that of the original YOLOv8s, indicating that the improved algorithm helps to enhance the model's ability to extract the feature details of the remote sensing image targets, which, in turn, improves the overall detection accuracy of the model.

The curves of the experimental results comparing the original algorithm and the improved algorithm on the DOTAv1.0 are shown in Figure 10. The figure shows that the mAP@0.5 of the improved algorithm gradually increases relative to that of the original YOLOv8 algorithm as the number of training epochs increases.

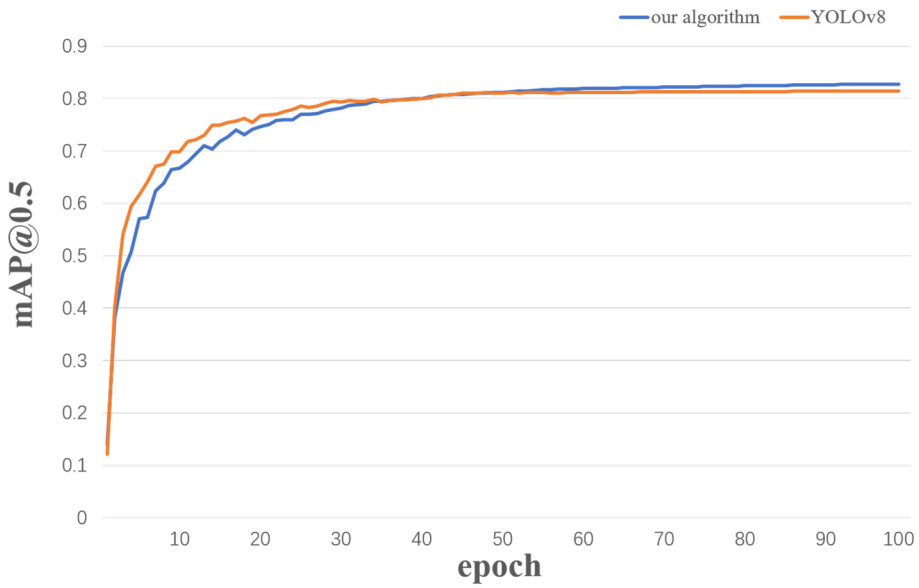

Figure 10. Comparison of mAP@0.5 curves before and after improvement.

In addition, we use the original YOLOv8 algorithm and the improved algorithm proposed in this paper for target detection of four types of remote sensing images. The first type contains dense targets, the second type contains small targets, the third type contains multi-scale targets, and the last type contains targets in complex scenes. The specific results are shown in Figure 11. For dense target images, the original algorithm misses one, and the improved algorithm detects all; for small target images, the original algorithm misses two, and the improved algorithm detects all; for multi-scale target images, the original algorithm misses three, and the improved algorithm detects all; and for targets in complex scenes, the original algorithm misses one, and the improved algorithm detects all. Neither

of the algorithms incorrectly detects targets in the any of the four types of images. This proves that the improved algorithm proposed in this paper offers obvious improvement over the original YOLOv8 algorithm in detecting targets in remote sensing images with different characteristics.

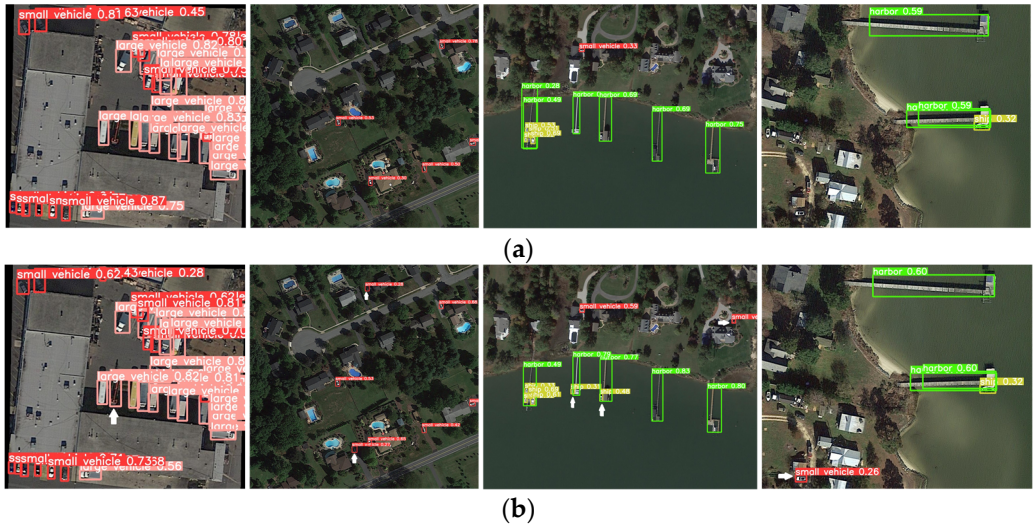

Figure 11. Comparison of detection results between the original YOLOv8 and the improved YOLOv8. The detection results of YOLOv8 are shown on the top, and the results of our algorithm are shown on the bottom.

4.6. Comparison Test

In order to further verify the effectiveness of the proposed algorithm, the improved algorithm is compared with typical algorithms in the field of target detection in remote sensing images in recent years on the DOTAv1.0 dataset; the results of the experiments are shown in Table 4. It can be seen that with input images of the same size and the same number of training epochs, the improved algorithm has an improved performance for all selected performance indicators compared with the SSD, YOLOv5, YOLOv7, and YOLOX algorithms, as well as the original YOLOv8 algorithm, in each of the selected performance indexes, which proves that the improved algorithm enhances the feature extraction ability of different sizes of targets and small targets with complex backgrounds by increasing the small target detection layer, introducing the EMA module, and improving the loss function so as to obtain a more accurate detection effect, which reduces misdetection and omission to a certain extent.

Table 4. Comparisons with different detection algorithms.

Algorithm	P/%	R/%	mAP@0.5%
SSD	79.8	76.4	79.3
YOLOv5	84.2	74.4	80.1
YOLOv7 [26]	81.5	74.5	79.3
YOLOX [27]	84.4	75.7	80.6
YOLOv8	84.0	76.6	81.4
Our algorithm	**84.5**	**77.3**	**82.7**

5. Conclusions

The remote sensing target detection task is characterized by high detection difficulty due to basic features such as a large proportion of image background information, dense targets, and a large number of small targets. To address the above problems, this paper proposes an improved YOLOv8 remote sensing image target detection algorithm based on the YOLOv8 target detection algorithm by adding a detection layer for small target feature

information extraction to its multi-scale feature fusion network structure; introducing the EMA attention mechanism in the C2f module; and, finally, using Wise-IoU to replace the CIoU loss function in the original network. The ablation and control experiments are designed based on the DOTA dataset, and the experimental results show that the detection effect of the algorithm proposed in this paper is better than that of the original algorithm and other typical target detection algorithms, and it achieves improved detection performance for remote sensing targets. However, due to the addition of various new modules, the computational volume and complexity of the algorithm increase slightly. In the future, we plan to consider introducing more advanced convolution modules into the model to reduce the number of parameters and improve the detection speed of the model.

Author Contributions: Conceptualization, H.W.; Methodology, H.W.; Software, H.W.; Validation, H.Y., H.C., J.W. and X.Z.; Resources, H.Y.; Data curation, H.W.; Writing—original draft, H.W.; Writing—review & editing, H.Y. and H.C.; Visualization, H.W.; Supervision, H.Y., H.C., J.W., X.Z. and Y.X. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

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

References

1. Niu, R.; Zhi, X.; Jiang, S.; Gong, J.; Zhang, W.; Yu, L. Aircraft Target Detection in Low Signal-to-Noise Ratio Visible Remote Sensing Images. *Remote Sens.* **2023**, *15*, 1971. [CrossRef]
2. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR′05), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
3. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
4. Budiman, A.; Fabian; Yaputera, R.A.; Achmad, S.; Kurniawan, A. Student attendance with face recognition (LBPH or CNN): Systematic literature review. *Procedia Comput. Sci.* **2023**, *216*, 31–38. [CrossRef]
5. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
6. Girshick, R. Fast R-CNN. *arXiv* **2015**. [CrossRef]
7. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. In Proceedings of the 28th International Conference on Neural Information Processing Systems, Montreal, QC, Canada, 7–12 December 2015; MIT Press: Cambridge, MA, USA, 2015; pp. 21–37.
8. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; IEEE Computer Society: Washington DC, USA, 2016; pp. 779–788.
9. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
10. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**. [CrossRef]
11. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**. [CrossRef]
12. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the 14th European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37.
13. Li, Z.; Zhou, F. FSSD: Feature fusion single shot multibox detector. *arXiv* **2017**. [CrossRef]
14. Wu, T.; Dong, Y. YOLO-SE: Improved YOLOv8 for Remote Sensing Object Detection and Recognition. *Appl. Sci.* **2023**, *13*, 12977. [CrossRef]
15. Yi, H.; Liu, B.; Zhao, B.; Liu, E. Small Object Detection Algorithm Based on Improved YOLOv8 for Remote Sensing. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2023**, *17*, 1734–1747. [CrossRef]
16. Wang, S.; Cao, X.; Wu, M.; Yi, C.; Zhang, Z.; Fei, H.; Zheng, H.; Jiang, H.; Jiang, Y.; Zhao, X.; et al. Detection of Pine Wilt Disease Using Drone Remote Sensing Imagery and Improved YOLOv8 Algorithm: A Case Study in Weihai, China. *Forests* **2023**, *14*, 2052. [CrossRef]

17. Wang, X.; Gao, H.; Jia, Z.; Li, Z. BL-YOLOv8: An Improved Road Defect Detection Model Based on YOLOv8. *Sensors* **2023**, *23*, 8361. [CrossRef] [PubMed]
18. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J.; Huang, Z. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In Proceedings of the ICASSP 2023—2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
19. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**. [CrossRef]
20. Liu, Z.; Ye, K. YOLO-IMF: An improved YOLOv8 algorithm for surface defect detection in industrial manufacturing field. In Proceedings of the International Conference on Metaverse, Honolulu, HI, USA, 23–26 September 2023; Springer Nature: Cham, Switzerland, 2023; pp. 15–28.
21. Zhu, Q.; Ma, K.; Wang, Z.; Shi, P. YOLOv7-CSAW for maritime target detection. *Front. Neurorobotics* **2023**, *17*, 1210470. [CrossRef] [PubMed]
22. Xia, G.-S.; Bai, X.; Ding, J.; Zhu, Z.; Belongie, S.; Luo, J.; Datcu, M.; Pelillo, M.; Zhang, L. DOTA: A Large-scale Dataset for Object Detection in Aerial Images. In Proceedings of the IEEE 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018. [CrossRef]
23. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam:Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
24. Liu, Y.C.; Shao, Z.R.; Hoffmann, N. Global attention mechanism: Retain information to enhance Channel-spatial interactions. *arXiv* **2021**. [CrossRef]
25. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
26. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**. [CrossRef]
27. Li, S.; Fu, X.; Dong, J. Improved Ship DetectionAlgorithm Based on YOLOX for SAR Outline Enhancement Image. *Remote Sens.* **2022**, *14*, 4070. [CrossRef]