electronics

MDPI

Article

YOLO-GCRS: A Remote Sensing Image Object Detection Algorithm Incorporating a Global Contextual Attention Mechanism

Huan Liao and Wenqiu Zhu *

School of Computer Science, Hunan University of Technology, Zhuzhou 412007, China; m21077500009@stu.hut.edu.cn
* Correspondence: zwq@hut.edu.cn

Abstract: With the significant advancements in deep learning technology, the domain of remote sensing image processing has witnessed a surge in attention, particularly in the field of object detection. The detection of targets in remotely sensed images is a challenging task, primarily due to the abundance of small-sized targets and their multi-scale distribution. These challenges often result in inaccurate object detection, leading to both missed detections and false positives. To overcome these issues, this paper presents a novel algorithm called YOLO-GCRS. This algorithm builds upon the original YOLOv5s algorithm by enhancing the feature capture capability of the backbone network. This enhancement is achieved by integrating a new module, the Global Context Block (GC-C3), with the C3 backbone network. Additionally, the algorithm incorporates a convoluted block known as CBM (Convolution + BatchNormalization + Mish) to enhance the network model's capability of extracting depth features. Moreover, a detection head, ECAHead, is proposed, which integrates an efficient attention channel (ECA) for extracting high-dimensional features from images. It achieves higher precision, recall, and mAP@0.5 values (98.3%, 94.7%, and 97.7%, respectively) on the publicly available RSOD dataset compared to the original YOLOv5s algorithm (improving by 5.3%, 0.8%, and 2.7%, respectively). Furthermore, when compared to mainstream detection algorithms like YOLOv7-tiny and YOLOv8s, the proposed algorithm exhibits improvements of 2.0% and 7.5%, respectively, in mAP@0.5. These results provide validation for the effectiveness of our YOLO-GCRS algorithm in addressing the challenges of missed and false detections in remote sensing object detection.

Keywords: deep learning; image processing; multi-scale distribution; remote sensing image; CBM; ECA; YOLOv5s

Citation: Liao, H.; Zhu, W. YOLO-GCRS: A Remote Sensing Image Object Detection Algorithm Incorporating a Global Contextual Attention Mechanism. *Electronics* **2023**, *12*, 4272. https://doi.org/10.3390/electronics12204272

Academic Editor: Duc Thanh Nguyen

Received: 16 September 2023
Revised: 3 October 2023
Accepted: 12 October 2023
Published: 16 October 2023

1. Introduction

Remote sensing image object detection is an essential technology with significant applications in various military and civilian applications, including search and rescue, reconnaissance [1], geological disasters [2], and everyday life [3]. The primary objective of this system is to identify and accurately locate and classify targets of interest, such as aircraft, vehicles, and ships, in optically complex remote sensing images. However, remote sensing images present challenges due to their complex background information, varying target scales, and abundance of small targets. These factors often result in difficulties in detection, leading to high numbers of false detections and missed detections. Traditional methods rely on human extraction of features, such as HoGDetector [4], DPM [5], and other classical algorithms. However, these approaches suffer from a high algorithmic complexity, low detection efficiency, and time-consuming processes.

Deep learning has made significant advancements in object detection in recent years [6]. By training deep neural networks on large datasets to learn target feature information, the proposed algorithm achieves a higher accuracy compared to traditional manual feature extraction methods while also being easier and more effective to implement. Currently,

there are two main classes of deep-learning-based object detection algorithms. The first category comprises traditional algorithms like R-CNN [7], Fast-RCNN [8], and Faster-RCNN [9]. These algorithms are known for their intricate design, resource-intensive nature, and relatively slower detection rates. The second type consists of single-stage regression-based algorithms like SSD [10], RetinaNet [11], and the series of YOLO methods [12–15]. Compared with the two-stage algorithm, the single-stage algorithms have simpler array designs, higher detection precision, and faster speeds.

In the field of remote sensing object detection, significant research efforts have yielded successful results. Wen et al. [16] presented an enhanced SSD algorithm that detects objects at multiple scales in remote sensing images. By incorporating advanced techniques for background modeling and employing efficient strategies for supervised learning, their approach significantly enhanced the detection of objects at multiple scales. The algorithm was evaluated using the COCO dataset and demonstrated an exceptional performance. In a separate investigation conducted by Qu et al. [17], a YOLOv3 model with an auxiliary network was proposed to enhance the detection of objects in remote sensing images. To optimize the detection performance, they incorporated the CBAM attention mechanism, which effectively suppressed irrelevant information and emphasized critical details. As a result of these enhancements, the object detection capability of their approach was significantly improved. Shen et al. [18] aimed to improve cross-scale detection in road object detection tasks through the utilization of the YOLOv3 model. To achieve this, they employed the K-means-GIoU algorithm to generate prior boxes and introduced a specialized detection branch specifically designed for detecting small targets. Furthermore, they incorporated channel and spatial attention modules to further enhance the overall performance of the method. These enhancements resulted in a higher mAP value and improved the detection accuracy, particularly for small-scale objects. Furthermore, Zhu et al. [19] proposed the addition of a prediction head to the YOLOv5 model for detecting objects of different scales, which proved to be effective for object detection. To summarize, deep learning techniques play a crucial role in remote sensing image object detection, offering significant value and the potential for various applications.

However, the aforementioned existing methods do not adequately address the challenges of missed detections and false positives that frequently arise in remote sensing image object detection. In order to tackle these issues, we devised an innovative algorithm for remote sensing object detection, named YOLO-GCRS. This algorithm builds upon the YOLOv5 framework and aims to provide effective solutions.

Our research makes significant contributions to the field in several aspects:

- We design the YOLO-GCRS by integrating the global-context-aware mechanism from YOLOv5 with the C3 architecture in version 6.1 of YOLOv5s. This integration enhances the network model's ability to capture the global features of an image. Additionally, we conducted an analysis to evaluate the detection performance when adding the mechanism at different positions within the backbone.
- We propose a new convolutional extraction module, CBM, to replace the CBS module in the original framework. This replacement significantly improves the model's detection accuracy when objecting objects.
- Lastly, we introduce a detection head called ECAHead, which incorporates the ECA attention mechanism. This design allows for the comprehensive extraction of high-dimensional channel features.

The subsequent sections of this paper are structured as follows: Section 2 presents a comprehensive overview of the methodology utilized in YOLOv5s. In Section 3, we introduce our proposed method. The experimental results and analysis are presented in Section 4. Lastly, Section 5 provides a summary of the paper and discusses potential avenues for future research.

2. Background

The YOLOv5s model primarily comprises input data (input), a backbone network (backbone), neck feature fusion (neck), and a multilevel detection head (head). Figure 1 illustrates the complete structure of YOLOv5s.

In YOLOv5s, the image input module employs a range of data augmentation techniques, including Mosaic4, automatic image cropping, splicing, and scaling, to preprocess the input images. Furthermore, this module automatically determines the optimal anchor frame for the model by considering the target size in the training samples. These enhancements contribute to an improved performance in terms of a higher detection accuracy and faster processing speeds for the single-stage algorithms.

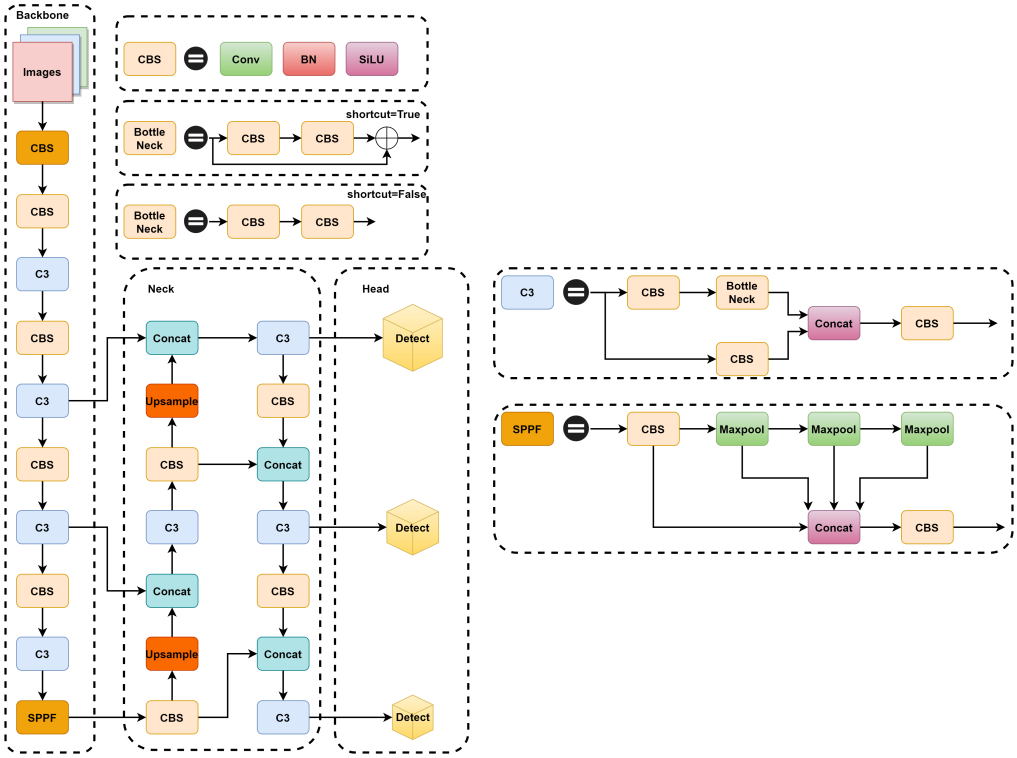

Figure 1. Design of YOLOv5s.

The YOLOv5s backbone network comprises the C3, CBS, and SPPF structures. The C3 structure divides input features into two parts, inspired by the cross-stage network CSPNet [20]. The main part progressively extracts features through convolution, normalization, and activation functions. Meanwhile, the branches adjust the channels using convolutional layers, effectively eliminating redundant gradient information. On the other hand, the structure of CBS incorporates a Conv layer, a BatchNorm layer with normalization techniques [21], and the SiLU activation function [22] to enhance feature extraction in multi-scale object detection. This combination enhances the network's ability to characterize the input. Lastly, the SPPF architecture utilizes multiple 5×5 maximal pooling layers to enhance the sensory diversity at various levels of the network, resulting in the incorporation of more comprehensive hierarchical features. The YOLOv5s backbone network offers substantial improvements in the design simplicity, accuracy, and speed compared to the two-stage algorithm by incorporating these structures.

The feature fusion module incorporates two crucial modules: the feature pyramid network (FPN) [23] and the path aggregation network (PAN) [24]. This integration enhances the overall performance of multi-scale object detection by effectively combining the strengths of both FPN and PAN.

The forecasting module in YOLOv5s incorporates three detection layers with varying scales: 80 × 80, 40 × 40, and 20 × 20. These layers serve a vital function by estimating the categories and positions of targets of different sizes, including small, medium, and large objects.

The loss functions used in YOLOv5s include the classification loss, target loss, and confidence loss. To handle the classification loss and confidence loss, the BCEWithLogitsLoss function is utilized, as illustrated in Formula (1). Conversely, the target loss is calculated using the Intersection over Union (IoU) function [25], and YOLOv5s version 6.1 introduces the CIoU, as demonstrated in Formula (2).

$$C = -\frac{1}{n}\sum_{x}[ylna + (1-y)ln(1-a)] \tag{1}$$

The sample is denoted by x, the label by y, the predicted output by a, and the total number of samples by n.

$$CIOU = IOU - \frac{\rho^2(b, b^{gt})}{c^2} - \alpha v \tag{2}$$

where IOU represents the intersection over the union ratio between the predicted and actual frames, and b and b^{gt} denote the center coordinates of the prediction frame and the rear frame. ρ represents the Euclidean distance between the centroids of the true frame and the predicted frame. c represents the diagonal distance between the predicted frame and the minimum outer join matrix of the true frame. α is a positive trade-off parameter that influences the evaluation. v is a measure of the aspect ratio consistency parameter.

YOLOv5s utilizes the NMS [26] post-processing technique to eliminate redundant candidate frames. The process is as follows:

- Group all rectangular boxes based on their category labels and sort the groups in descending order of confidence scores.
- Start by identifying the rectangular box with the most reliable confidence score in step 1. Proceed by sequentially evaluating the remaining rectangular boxes. Compute the IOU between every bounding box and the currently chosen box that has the highest score. Remove any boxes that surpass a predefined IOU threshold, ensuring that only the most relevant boxes are retained for further analysis.
- Repeat step 2 for the remaining rectangular boxes obtained from step 2 until all boxes have been processed.

3. Proposed Method

Figure 2 visually represents the overarching structure of our proposed YOLO-GCRS. The procedure commences by extracting features from the input remote sensing images employing the GC-C3 and CBM architectures. These structures play a crucial role in capturing relevant information. Subsequently, the features are fused at multiple scales through the neck structure, enabling a comprehensive understanding of the image. Finally, the ECAHead is employed to extract high-dimensional channel information, and the prediction information is outputted.

3.1. Global Context Block

In object detection, convolutional deep neural networks are commonly used to extract image feature information. These networks primarily focus on local pixel locations. However, to capture long-range dependencies, it is necessary to stack convolutional layers multiple times. Unfortunately, directly stacking these layers repetitively can lead to computational inefficiency and difficulties in optimizing the model. This is due to the problematic transfer of information over large distances.

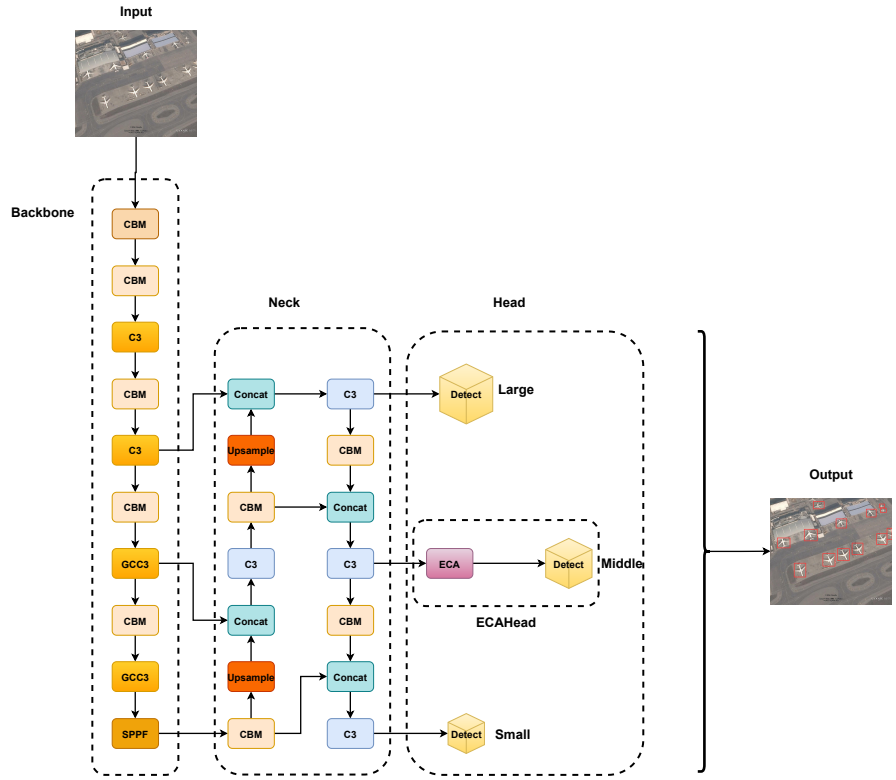

Figure 2. Design of YOLO-GCRS.

The Global Context Block, as illustrated in Figure 3, consists of two modules: Context Modeling and Transform. The model's capacity to capture contextual information is significantly improved by these modules. The Global Context Block and the SE (Squeeze-and-Excitation) [27] block differ in their approaches to the Context Modeling module. The SE utilizes global average pooling, whereas the Global Context Block employs a 1×1 convolution and the SoftMax function.

The height and width of the feature map in the figure are denoted by H and W, respectively, while the number of channels in the feature map is represented by C. In the Context Modeling section, the main branch first reduces the feature map using a conv 1×1 convolution and reshapes it to $1 \times W \times H$. Then, the Softmax operation is applied to obtain the attention weights. The auxiliary branch multiplies the $C \times H \times W$ feature map by the attention weights. Finally, the values within each channel of the feature map are summed to obtain the global relationship of size $C \times 1 \times 1$. The Transform structure incorporates two conv 1×1 operations to minimize the parameter count. Additionally, LayerNorm is utilized to address model optimization issues. The next step involves combining the global information of $H \times W \times C$ and $C \times 1 \times 1$ through the broadcast mechanism. This allows for obtaining the output of crucial global information from the augmented image. The calculation of the Global Context Block is demonstrated in Equations (3)–(5).

$$\alpha_j = \frac{e^{W_{kx_j}}}{\sum_m e^{W_{kx_m}}} \tag{3}$$

$$\delta(\cdot) = W_{v2} ReLU(LN(W_{v1}(\cdot))) \tag{4}$$

$$z_i = x_i + W_{v2} ReLU(LN(W_{v1} \sum_{j=1}^{N_p} \frac{e^{W_{kx_j}}}{\sum_{m=1}^{N_p} e^{W_{kx_m}}} x_j)) \tag{5}$$

where x_i and x_j are an instance of the input sample, α_j is the weight of the global attention pooling, $\delta(\cdot)$ represents the Transform structure, and W_{v_2} and W_k represent linear transformations.

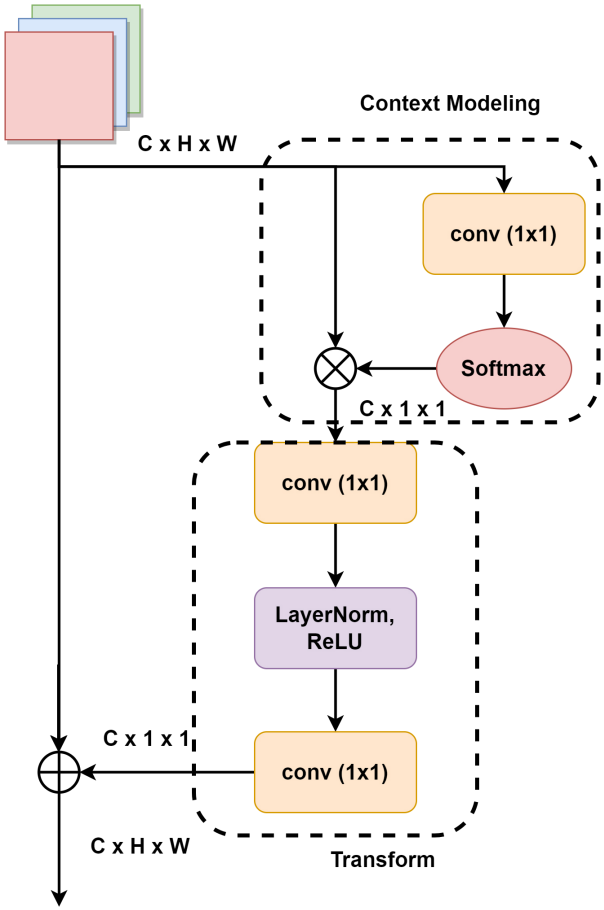

Figure 3. Design of the Global Context Block.

In order to enhance the feature extraction capabilities of YOLOv5s, a novel network feature module named GC-C3 block is introduced. This module combines the Global Context Block and the C3 block, as depicted in Figure 4. The GC-C3 block structure incorporates the GC block into both branches of the C3 block. By integrating these components, the module enables the extraction of long-range contextual feature information, ultimately enhancing the network's performance.

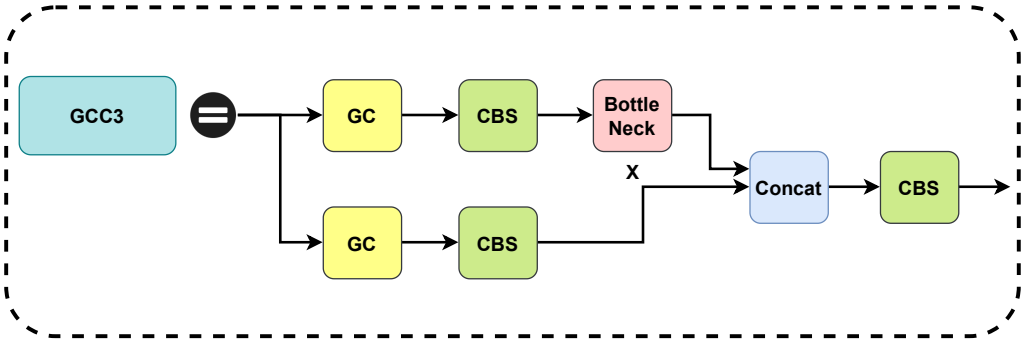

Figure 4. Structure of GC-C3 block.

3.2. CBM Model

The Mish activation function, as shown in Figure 5, does not have an upper bound or a lower bound. This characteristic helps to prevent saturation caused by capping. Unlike the ReLU activation function, which has hard zero bounds, Mish has a slight allowance for negative values. This adaptability facilitates improved gradient movement within the neural network. Additionally, the Mish activation function, as proposed by Misra et al. [28], provides a smoother activation function compared to Swish. This smoothness allows for better information propagation throughout the neural network, resulting in improved accuracy and generalization.

Equation (6) is utilized to compute the Mish activation function.

$$Mish(x) = x \cdot tanh(ln(1 + e^x)) \tag{6}$$

where x is the input sample and ln, $tanh$, and exp are all common math functions.

YOLOv5s utilizes the Convolution + BatchNormalization + SiLU (CBS) convolution block to extract feature information. The SiLU activation function, which is implemented by the Swish activation function, is used within this block. A new module called Convolution + BatchNormalization + Mish (CBM) is introduced to improve the network's ability to extract deep-level features and improve the accuracy.The CBM convolutional block's structure is shown in Figure 6.

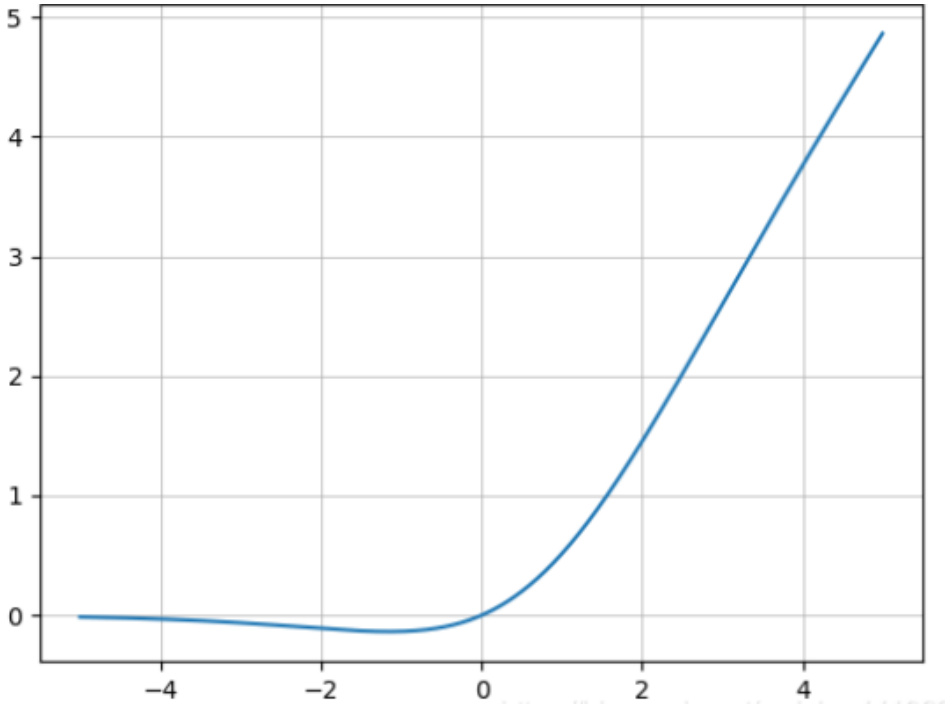

Figure 5. Mish activation function graph.

Figure 6. CBM Model.

3.3. ECAHead

Hu et al. were the first to introduce the SE attention mechanism, which initially compresses the input feature maps, but this compression and dimensionality reduction can hinder the learning of inter-channel dependencies. In contrast, the ECA attention mechanism [29] avoids dimensionality reduction and effectively captures local cross-channel interactions by utilizing one-dimensional convolutions. This allows for the extraction of inter-channel dependencies. Figure 7 illustrates the architecture of the ECA attention mechanism. The specific implementation steps of the ECA attention mechanism are as follows:

- The global average pooling operation is applied to the input feature map.
- Subsequently, a one-dimensional convolution procedure with a convolution kernel size of k is executed. The Sigmoid activation function, as demonstrated in Equation (7), is used to calculate the weights, represented by ω, for each channel.

$$\omega = \delta(C1D_k(y)) \tag{7}$$

  where $C1D$ denotes one-dimensional convolution, δ denotes the Sigmoid function, and ω denotes the weights obtained after computation.
- The original input feature map's elements are then given weights, leading to the production of the ultimate output feature map. This multiplication process ensures that each element of the input feature map is appropriately weighted to contribute to the final representation.

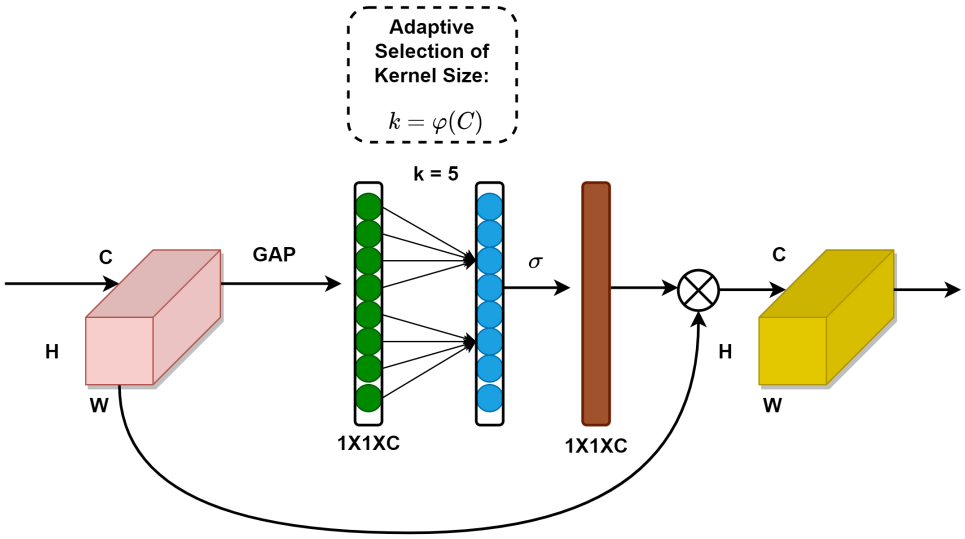

Figure 7. Structure of the ECA attention mechanism.

The ECA mechanism falls under the category of channel attention mechanisms, and YOLOv5s exhibits high-dimensional channel features at the Head layer. To efficiently extract and utilize these high-dimensional channel features, we propose a new module called ECAHead. This module integrates the ECA attention mechanism with the Head layer of YOLOv5s.

4. Experiments and Results Analysis

4.1. Experimental Environment

This paper presents the experimental environment, which is detailed in Table 1. The table provides an overview of the specific conditions under which the experiments were conducted.

Table 1. Configuration of the experimental environment.

Options	Configuration
Operating System	Ubuntu
CPU	E5-2680 v4
GPU	GeForce RTX 3060
Memory	14 GB
CUDA	11.1
Pytorch version	1.10.0

4.2. Datasets

For this experiment, the RSOD dataset [30] was employed. The RSOD dataset is specifically designed for remote sensing applications and has been made publicly available by Wuhan University. It comprises four distinct categories, namely aircraft, playground, overpass, and oil tank. The utilization of this dataset enables the evaluation and analysis of object detection algorithms in the context of remote sensing scenarios.

Table 2 provides a comprehensive breakdown of the types and quantities of datasets.

Table 2. Distribution of datasets.

Labeling of the Dataset	Number of Images
aircraft	446
playground	189
overpass	176
oil tank	165

Additionally, Figures 8 and 9 provide a thorough representation of the sample RSOD dataset and its fundamental features. The visual representation reveals the presence of numerous small targets within the dataset, accompanied by intricate background information. Additionally, the dataset exhibits a multi-scale distribution, emphasizing the diverse range of object sizes present.

Figure 8. Illustrative representation of the RSOD dataset.

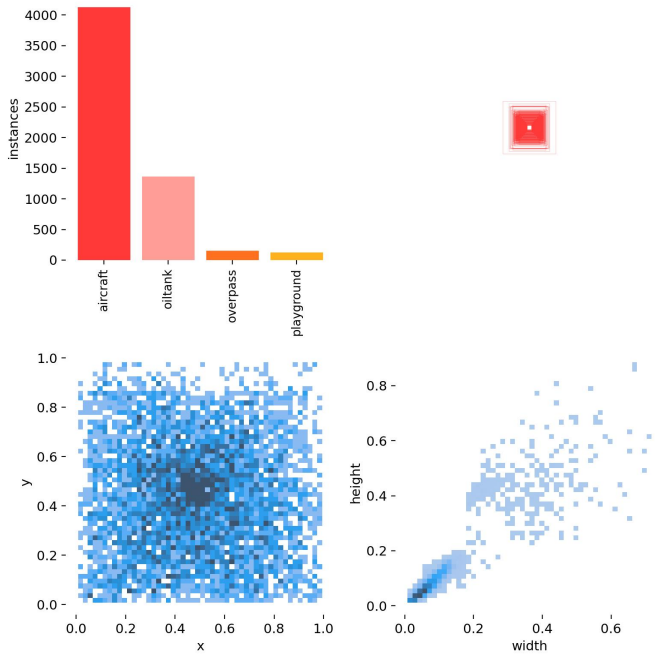

Figure 9. Labeling information distribution.

4.3. Evaluation Metrics

4.3.1. Precision

The precision of a prediction is determined by dividing the number of correct predictions by the total number of results predicted for positive samples. It gauges the precision of the model in terms of accurately recognizing positive occurrences.

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

True Positives (TP) refer to cases where the prediction and the label value both correspond to positive examples. On the other hand, False Positives (FP) denote situations where the forecast is categorized as a positive instance, yet the label value is, in fact, a negative instance.

4.3.2. Recall

The likelihood of a positive sample being correctly identified among all predicted positive outcomes is denoted by the recall. This measures the effectiveness of the model in capturing all relevant positive instances.

$$Recall = \frac{TP}{TP + FN} \tag{9}$$

False Negatives (FN) are situations where the forecast is labeled as a negative result, yet the actual result is a positive one. They signify instances where the model fails to identify a positive instance correctly.

4.3.3. Mean Average Precision

The mAP is a measure that determines the mean precision across all categories. It provides a comprehensive evaluation of the model's performance.

$$mAP = \frac{1}{N} \sum AP_i \tag{10}$$

The total number of categories is denoted by N, and the average precision in category i is represented by AP_i. When evaluating the mAP@0.5, we refer to the average accuracy value obtained when using an IoU threshold of 0.5. The level of overlap between the predicted bounding box and the ground truth bounding box is determined by this threshold in order for it to be deemed a correct detection.

4.3.4. FLOPs

FLOPs, also known as floating point operations, represent the computation amount and can be utilized as a metric to measure the algorithm complexity. They are characterized as the quantity of floating point operations carried out. FLOPs are expressed in G(billion).

4.3.5. FPS

The concept of Frames Per Second (FPS) denotes the rate at which frames are transmitted within a single second in a picture or video. It is measured in $frame/s$. For this experiment, the GPU FPS was selected as the standard for evaluating the system performance and speed.

$$FPS = \frac{Frames}{Time} \tag{11}$$

4.4. Parameter Setting and Network Training

4.4.1. Parameter Setting

Table 3 presents the training parameter settings outlined in this research paper.

Table 3. Configuration of the experimental parameters.

Parameters	Value
weights	yolov5s.pt
division ratio	7:2:1 (train:val:test)
optimizer	SGD
batch size	16
epochs	100

The pre-training weights of yolov5s.pt were obtained from the migration of ImageNet. The division ratio indicates the ratio at which the dataset is split.

4.4.2. Network Training

The network training results are visually represented by the loss function curve. This research paper introduces a loss function consisting of three primary components: classification loss, target loss, and confidence loss.

$$L_{loss} = L_{cls} + L_{obj} + L_{box} \tag{12}$$

L_{cls}, L_{obj}, and L_{box} correspond to the classification loss, confidence loss, and target loss, respectively.

An evaluation of the network training effectiveness can be performed by analyzing the loss function curves. Figure 10 illustrates the loss function curves for the YOLOv5s and YOLO-GCRS models.

The visualization results suggest that the YOLO-GCRS model reduces the loss as the number of iterations increases. After around 80 iterations, the loss value stabilizes and converges towards zero, signifying the attainment of optimal effectiveness in model training.

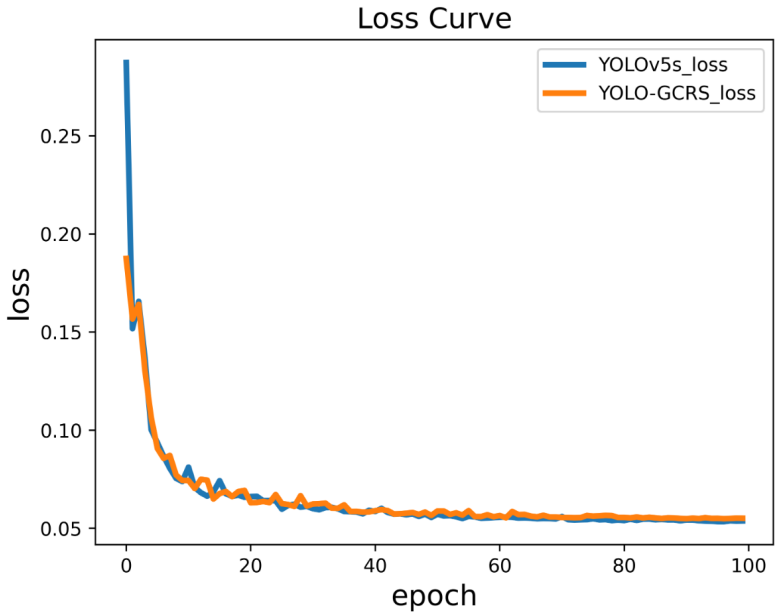

Figure 10. Loss Curve.

4.4.3. Ablation Experiments

This research paper extensively showcases the designed module's effectiveness and sophistication through the conduction of ablation experiments. This study conducted ablation experiments using the validation set as the dataset.

Initially, we assessed the possible placements of the GC-C3 module in the backbone and scrutinized their corresponding impacts. Table 4 provides a summary of these findings.

Table 4. Comparison of the GC-C3 effects at different positions.

Method	Precision	Recall	mAP@0.5	FLOPs
YOLOv5s	0.930	0.939	0.950	15.8
+front-2	0.980	0.934	0.970	15.9
+behind-2	0.968	0.950	0.965	16.4
+backbone-4	0.968	0.939	0.956	16.4

We identified three scenarios for the placement of the GC-C3 module in the YOLOv5s backbone. The first scenario, front-2, involves replacing the initial two C3 structures with GC-C3. The second scenario, behind-2, entails replacing the last two C3 structures. Lastly, the backbone-4 scenario involves replacing all C3 structures with GC-C3.

Upon analyzing the results, we observed that the behind-2 experiment yielded the best outcomes. Specifically, compared to the original YOLOv5s, it showed improvements of 3.8% in precision, 1.1% in recall, and 1.5% in mAP@0.5, with only a marginal 0.6 increase in FLOPs. Although front-2 exhibited a higher mAP@0.5 metric, a comprehensive evaluation of precision, recall, and mAP@0.5 revealed that behind-2 outperformed it.

Moving on, we proceed to discuss the proposed base extraction convolutional block (CBM), as presented in Table 5.

Table 5. Experimental comparison of different loss functions.

Method	Precision	Recall	mAP@0.5	FLOPs
YOLOv5s(CBS)	0.930	0.939	0.950	15.8
+CBR	0.959	0.937	0.958	15.8
+CBM	0.953	0.948	0.967	15.8

The experiment compares the base extracted convolutional block CBS (Convolution + BatchNormalization + SiLU) in the original YOLOv5s with two modified versions. CBR (Convolution + BatchNormalization + ReLU) replaces the activation function with ReLU, while CBM is a proposed novel base extraction convolution block.

We compared the effectiveness of various activation functions by running experiments with these blocks. Notably, our proposed CBM convolution block consistently outperformed the others, achieving optimal results in terms of mAP@0.5.

Moving forward, we delve into the effectiveness of ECAHead in detecting objects at various positions. Table 6 illustrates the different detection effects of ECAHead at different locations.

Table 6. Evaluating the detection effect of ECAHead in various positions.

Method	Precision	Recall	mAP@0.5	FLOPs
YOLOv5s	0.930	0.939	0.950	15.8
+ECAHead-s	0.964	0.924	0.941	15.8
+ECAHead-m	0.946	0.973	0.973	15.8
+ECAHead-l	0.954	0.942	0.969	15.8
+ECAHead-a	0.944	0.938	0.957	15.8

The detection heads in the original model, ECAHead s, m, l, and a, respectively replace the small, middle, large, and all detection targets. After evaluating their performance levels, we discovered that ECAHead-middle yielded the most favorable metric results.

To further highlight the superiority of ECAHead, we integrated it with mainstream attention mechanisms in the head and conducted comparative experiments. Table 7 illustrates the experimental outcomes of different attentional mechanisms for head detection.

Table 7. Comparison of the experimental effects of different attentional mechanisms for detecting heads.

Method	Precision	Recall	mAP@0.5	FLOPs
YOLOv5s	0.930	0.939	0.950	15.8
+ECAHead	0.946	0.973	0.973	15.8
+SEHead	0.987	0.947	0.964	15.8
+CBAMHead	0.957	0.952	0.962	15.8
+SAHead	0.954	0.931	0.956	15.8

In our comparisons, we evaluated ECAHead against other mainstream attention mechanisms, including SE, CBAM [31], and SA (ShuffleAttention) [32]. The results unequivocally demonstrate that ECAHead consistently outperformed the alternatives, providing strong evidence for its effectiveness.

We conducted ablation experiments on YOLO-GCRS to demonstrate the module's direct impact visually. Table 8 displays the experimental outcomes for each module.

The results of the ablation tests conducted on each module showed marked improvements in the precision, recall, and mAP@0.5 when compared to the initial YOLOv5s. Specifically, our proposed YOLO-GCRS demonstrated improvements of 5.3% in precision, 0.8% in recall, and 2.7% in mAP@0.5 over the original YOLOv5s. Furthermore, the increase in FLOPs was merely 0.6, ensuring that the real-time detection capabilities with FPS were maintained within an acceptable range. The findings of the research provide strong evidence for the effectiveness of YOLO-GCRS.

In addition, to showcase the sophistication of YOLO-GCRS, we compared it with the most advanced algorithms in its category. Table 9 shows the experimental results for various mainstream algorithms.

Table 8. Ablation experiment.

Method	Precision	Recall	mAP@0.5	FLOPs	FPS/(frame/s)
YOLOv5s	0.930	0.939	0.950	15.8	76.8
+GC-C3	0.968	0.950	0.965	16.4	43.5
+CBM	0.953	0.948	0.967	15.8	69.1
+ECAHead	0.946	0.973	0.973	15.8	74.7
YOLO-GCRS	0.983	0.947	0.977	16.4	42.6

Table 9. Examining and contrasting mainstream algorithms through experimentation.

Method	Precision	Recall	mAP@0.5
YOLOv5s	0.930	0.939	0.950
YOLOv7-tiny	0.953	0.957	0.957
YOLOv8s	0.871	0.864	0.902
YOLO-GCRS	0.983	0.947	0.977

Our designed YOLO-GCRS algorithm has been proven to achieve superior results over state-of-the-art peer algorithms on experimental test datasets.

To evaluate the YOLO-GCRS model's ability to generalize, we conducted comparative experiments on the NWPU VHR-10 remote sensing dataset. The parameter settings for these experiments were identical. The results of these experiments are shown in Table 10.

Table 10. Evaluating the performance levels of diverse models on the NWPU VHR-10 dataset.

Method	Precision	Recall	mAP@0.5
YOLOv5s	0.935	0.927	0.942
YOLO-GCRS	0.950	0.933	0.955

The YOLO-GCRS model in the NWPU VHR-10 dataset showed remarkable improvements compared to the initial YOLOv5s. Significantly, it achieved a 1.5% rise in precision (P), a 0.6% rise in recall (R), and a 1.3% rise in mAP@0.5. These findings serve as strong evidence for the remarkable generalization ability of the YOLO-GCRS model.

4.4.4. Visualization Experiments

In order to gain a more thorough understanding of how YOLO-GCRS addresses detection issues in remote sensing datasets, we conducted visual and comparative experiments in various contexts. These experiments aim to illustrate the efficiency of YOLO-GCRS in resolving these concerns in a detailed manner.

To begin with, Figure 11 illustrates the challenges linked to the identification of minor targets, such as missed detections and false detections. By employing the YOLO-GCRS methodology, we can successfully mitigate the occurrence of false detections and missed detections in aircraft, particularly when it comes to small targets.

Additionally, Figure 12 illustrates the detection outcomes of targets across various scales. The complexity of the background information portrayed in Figure 12 is matched by the multi-scale distribution of the aircraft types. Through the implementation of YOLO-GCRS, we can effectively reduce the occurrence of missed detections.

Finally, Figure 13 demonstrates the identification of extensive targets with intricate background information. YOLO-GCRS demonstrates its ability to effectively detect large-scale targets and reduce the occurrence of false detections in such scenarios.

To conclude, the YOLO-GCRS algorithm tackles the issues of missed detections and false detections due to the abundance of small targets, multi-scale distribution, and intricate background data.

Figure 11. Missed and false detections of small targets in complex backgrounds. The (**left**) side displays the identification outcomes of YOLOv5s, while the (**right**) side exhibits the outcomes of YOLO-GCRS.

Figure 12. Multi-scale small target missed detections in complex environments. The (**left**) side displays the identification outcomes of YOLOv5s, while the (**right**) side exhibits the outcomes of YOLO-GCRS.

Figure 13. False detections of large targets in intricate environments. The (**left**) side displays the identification outcomes of YOLOv5s, while the (**right**) side exhibits the outcomes of YOLO-GCRS.

5. Conclusions

In this paper, we proposed a remote sensing image detection algorithm called YOLO-GCRS, which incorporates a global contextual attention mechanism. The key innovations of YOLO-GCRS include the GC-C3 structure, the CBM convolution module, and the ECAHead structure. The GC-C3 structure enables the model to establish long-range contextual information, facilitating the extraction of global features. The CBM convolution module can effectively improve the depth feature extraction ability of the model. Lastly, the ECAHead structure enhances the extraction of high-dimensional channel features and improves the focus on features of interest by incorporating attention mechanisms.

YOLO-GCRS attains a mAP@0.5 of 97.7%, surpassing the original YOLOv5s by 2.7%. In comparison to mainstream detection algorithms such as YOLOv7-tiny and YOLOv8s, YOLO-GCRS showcases enhancements in the mAP@0.5 metrics, varying between 2% and 7.5%. Our proposed algorithm efficiently resolves the problems of missed detections and false detections in remote sensing images. Moving forward, we will focus on improving the lightweight design and enhancing the speed of our models.

Author Contributions: Writing—original draft, H.L.; Writing—review and editing, W.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the following funding sources: the National Key Research and Development Program (NKRDP) projects (funding number: 2019QY1604); the Hunan Provincial Self-Science Foundation (funding number: 2022JJ50051); the Natural Science Foundation of Hunan Province (funding number: 2021JJ50058); the Open Platform Innovation Foundation of the Education Department of Hunan (funding number: 20K046).

Data Availability Statement: This research employed publicly available datasets for its experimental studies.

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

References

1. Liu, F.; Zhu, J.; Wang, W.; Kuang, M. Surface-to-air missile sites detection agent with remote sensing images. *Sci. China Inf. Sci.* **2021**, *64*, 1–3. [CrossRef]
2. Zhang, Y.; Ning, G.; Chen, S.; Yang, Y. Impact of rapid urban sprawl on the local meteorological observational environment based on remote sensing images and GIS technology. *Remote Sens.* **2021**, *13*, 2624. [CrossRef]
3. Luo, S.; Yu, J.; Xi, Y.; Liao, X. Aircraft target detection in remote sensing images based on improved YOLOv5. *IEEE Access* **2022**, *10*, 5184–5192. [CrossRef]
4. 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.
5. Felzenszwalb, P.F.; Girshick, R.B.; Mcallester, D.A. Cascade object detection with deformable part models. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010.
6. Lecun, Y.; Bengio, Y.; Hinton, G. Deep learning. *Nature* **2015**, *521*, 436. [CrossRef] [PubMed]
7. 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.
8. Girshick, R. Fast r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
9. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, 1137–1149 . [CrossRef] [PubMed]
10. 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 Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
11. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
12. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
13. 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.
14. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.

15. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
16. Wen, G.; Cao, P.; Wang, H.; Chen, H.; Liu, X.; Xu, J.; Zaiane, O. MS-SSD: Multi-scale single shot detector for ship detection in remote sensing images. *Appl. Intell.* **2023**, *53*, 1586–1604. [CrossRef]
17. Qu, Z.; Zhu, F.; Qi, C. Remote sensing image target detection: Improvement of the YOLOv3 model with auxiliary networks. *Remote Sens.* **2021**, *13*, 3908. [CrossRef]
18. Shen, L.; Tao, H.; Ni, Y.; Wang, Y.; Stojanovic, V. Improved YOLOv3 model with feature map cropping for multi-scale road object detection. *Meas. Sci. Technol.* **2023**, *34*, 045406. [CrossRef]
19. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 based on transformer prediction head for object detection on drone-captured scenarios. In Proceedings of the IEEE/CVF INTERNATIONAL conference on Computer Vision, Montreal, BC, Canada, 11–17 October 2021; pp. 2778–2788.
20. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 13–19 June 2020; pp. 390–391.
21. Ioffe, S.; Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proc. Mach. Learn. PMLR* **2015**, *37*, 448–456.
22. Elfwing, S.; Uchibe, E.; Doya, K. Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Netw.* **2018**, *107*, 3–11. [CrossRef] [PubMed]
23. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
24. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
25. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. Unitbox: An advanced object detection network. In Proceedings of the 24th ACM International Conference on MULTIMEDIA, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520.
26. Neubeck, A.; Van Gool, L. Efficient non-maximum suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, 20–24 August 2006; Volume 3, pp. 850–855.
27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
28. Misra, D. Mish: A self regularized non-monotonic activation function. *arXiv* **2019**, arXiv:1908.08681.
29. 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 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 11534–11542.
30. Long, Y.; Gong, Y.; Xiao, Z.; Liu, Q. Accurate object localization in remote sensing images based on convolutional neural networks. *IEEE Trans. Geosci. Remote Sens.* **2017**, *55*, 2486–2498. [CrossRef]
31. 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.
32. Zhang, Q.L.; Yang, Y.B. Sa-net: Shuffle attention for deep convolutional neural networks. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 2235–2239.