algorithms

MDPI

Article

Ship Detection Algorithm Based on YOLOv5 Network Improved with Lightweight Convolution and Attention Mechanism

Langyu Wang [1,†], Yan Zhang [1,†], Yahong Lin [2], Shuai Yan [3], Yuanyuan Xu [1] and Bo Sun [3,*]

[1] Logistics Engineering College, Shanghai Maritime University, Shanghai 201306, China; 202130210136@stu.shmtu.edu.cn (L.W.); zhangyan@shmtu.edu.cn (Y.Z.); yyxu@shmtu.edu.cn (Y.X.)
[2] School of Electromechanical and Automotive Engineering, Yantai University, Yantai 264005, China; linyahong@ytu.edu.cn
[3] Shanghai Advanced Research Institute, Chinese Academy of Sciences, Shanghai 200204, China; yanshuai@sari.ac.cn
* Correspondence: sunb@sari.ac.cn
† These authors contributed equally to this work.

Abstract: Aiming at the problem of insufficient feature extraction, low precision, and recall in sea surface ship detection, a YOLOv5 algorithm based on lightweight convolution and attention mechanism is proposed. We combine the receptive field enhancement module (REF) with the spatial pyramid rapid pooling module to retain richer semantic information and expand the sensory field. The slim-neck module based on a lightweight convolution (GSConv) is added to the neck section, to achieve greater computational cost-effectiveness of the detector. And, to lift the model's performance and focus on positional information, we added the coordinate attention mechanism. Finally, the loss function CIoU is replaced by SIoU. Experimental results using the seaShips dataset show that compared with the original YOLOv5 algorithm, the improved YOLOv5 algorithm has certain improvements in model evaluation indexes, while the number of parameters in the model does not increase significantly, and the detection speed also meets the requirements of sea surface ship detection.

Keywords: ship detection; YOLOv5; attention mechanism; lightweight convolution

Citation: Wang, L.; Zhang, Y.; Lin, Y.; Yan, S.; Xu, Y.; Sun, B. Ship Detection Algorithm Based on YOLOv5 Network Improved with Lightweight Convolution and Attention Mechanism. *Algorithms* **2023**, *16*, 534. https://doi.org/10.3390/a16120534

Academic Editor: Laura Antonelli

Received: 28 September 2023
Revised: 2 November 2023
Accepted: 8 November 2023
Published: 22 November 2023

1. Introduction

With developments in artificial intelligence and deep learning technology, the use of image vision and neural network algorithms to realize marine vessel target detection has become an important research direction. Target detection algorithms based on deep learning can generally be divided into two categories. One approach involves a two-stage algorithm represented by R-CNN (regions with convolutional neural network) [1], which is characterized by high detection accuracy; however, the detection speed is slow and may not be able to adapt to the real-time requirements of ship detection. The other approach involves a two-stage algorithm represented by SSD (single shot multi-box detector) [2] and YOLO (you only look once) [3] as the representative of the one-stage algorithm. It is characterized by faster detection speeds but, at the same time, there are certain shortcomings in its accuracy. Lin et al. proposed Retina-Net [4], which uses focal loss to solve the sample imbalance problem, allowing one-stage networks to achieve two-stage accuracy. Shortly after, Zhang et al. [5] proposed the RefineDet network.

More researchers are turning their attention to deep learning, and are using deep learning algorithms to solve the problem of ship detection. Liu et al. [6] redesigned the ship anchor box size based on YOLOv3, introduced soft non-maximum suppression, and reconstructed mixed loss functions to improve the network's learning and expression abilities for ship features. Zhou et al. proposed a lightweight model called Lira-YOLO that combines

the advantages of YOLOv3 and Retina-Net, which greatly reduces the number of parameters and computational complexity while ensuring detection accuracy [7]. Hong et al. [8] used a residual network instead of continuous convolution operation in YOLOv4 to solve the problem of network degradation and gradient disappearance, and established a nonlinear target tracking model based on the UKF method, which improved the accuracy of ship detection. Wang et al. [9] used SSD512, Faster-RCNN, and other methods to obtain a self-constructed SAR image dataset with an 89.43% AP value; however, from the experimental results, some misdetections and omissions still occurred. FS Hass [10] et al. proposed a deep learning model based on YOLOv3 for distinguishing between icebergs and ships in 2023, which can be used to map marine objects prior to a journey. Ye et al. [11] proposed an enhanced attention mechanism YOLOv4 (EA-YOLOv4) algorithm, which can reduce the missed detection of overlapping ships without affecting its efficiency. Li et al. [12] proposed an improved YOLOv5 for the phenomena of complex backgrounds and dense ships, which reduces the missed detection rate in the SAR image dataset with complex backgrounds. Kong et al. [13] proposed a lightweight ship detection network based on YOLOx-Tiny, which can provide theoretical and technical support for platforms with limited computational resources, and has excellent performance in SAR remote sensing datasets. Krishna Patel et al. [14] combined a graph neural network (GNN) and YOLOv7 into an algorithm that can be used for automatic ship detection in high-resolution satellite image datasets.

Although the above algorithms have made some improvements when common target detection algorithms are transplanted to ship detection, there are still some degrees of shortcomings. Ship targets on the sea or river are generally large, but many studies aim to improve the detection ability of small targets. At the same time, the environments of seas, lakes, or rivers are more complicated, and the weather also easily interferes with the accuracy of target detection; therefore, it is necessary to retain richer semantic information, while retaining a larger sense of the field, in order to better localize large targets. In the process of detection, missed targets will have a great impact on the safety of the sea surface, so it is very important to improve the recall rate while satisfying the accuracy. In addition, current research is generally dominated by large models with many parameters and significant computation, which are not friendly enough for the equipment; hence, it is especially important to develop an algorithm that effectively improves the precision and recall under changes in parameter numbers and model size.

In order to solve these problems, we propose an improved ship target detection algorithm based on YOLOv5. The main contributions of the proposed algorithm are as follows:

(1) Aiming at the problems of semantic information loss and small sensory fields, the RFE model is introduced in the backbone part, which increases the sensory field, reduces the model parameters, and also reduces the risk of potential overfitting. The module is divided into three branches, each with two standard convolutions, as well as one null convolution. A residual block is also introduced, which significantly enhances the detection capability for large ships.

(2) In order to improve the accuracy of ship type identification and the recall of ships, coordinate attention mechanism is embedded into the YOLOv5 structure. The coordinate attention mechanism considers a more efficient way to capture position and channel information, resulting in better performance of YOLOv5 in the face of overlapping targets and small targets. At the same time, the coordinate attention mechanism improves the recall of the model, which gives the improved YOLOv5 a greater advantage in marine surveillance. Coordinate attention mechanism is introduced to lift the model's performance and solve the problem of insufficient detection ability.

(3) A lightweight convolutional GSconv is used in the neck section to reduce the complexity of the model and maintain accuracy, resulting in a better balance between model accuracy and speed. Similarly, the one-time aggregation module VoV-GSCSP is used instead of the C3 module to form a slim-neck module with GSConv, which serves to maintain sufficient accuracy with little change in the number of parameters.

(4) Finally, in terms of the loss function, SIoU is used to replace the CIoU of YOLOv5. SIoU is used instead of CIoU to redefine the loss function utilizing the vector angle between the bounding box regressions, which effectively improves the detection accuracy and increases the convergence speed of the network.

The other sections in this study are organized as follows. In Section 2, we detail the ship detection method used in this study. In Section 3, the setting and evaluation indicators for this experiment are described. In Section 4, the experimental results are presented. Finally, Section 5 provides the conclusion of this study.

2. Methodology

2.1. Algorithm Overview

The network structure of YOLOv5 is mainly composed of the input, backbone, neck, and head parts. The input side includes the input image as well as preprocessing the image. The backbone part mainly utilizes the CSP structure to extract features from the input image, which allows the model to learn more features. The neck part uses the FPN (feature pyramid network) [15] plus PAN (path aggregation network) [16] structure, which can extract image features more fully and retain richer feature information. Finally, the head layer predicts the target features by the loss function and non-maximal value suppression, and makes position regression and judgments of the presence or absence of targets and classification.

The YOLOv5 algorithm, as a commonly used target detection algorithm, has been realized in many fields, but there are still some shortcomings in the scenario of marine vessel target detection.

Our algorithm firstly integrates the weight-sharing-based cavity convolution scale-aware RFE model (RFE) module into the SPPF module in the backbone after forming the SPPF–REF module, which enlarges the receptive field, improves the information extraction ability for multi-scale targets, and reduces the risk of overfitting. Secondly, the lightweight convolution (GSconv) replaces the neck part of the convolution module, which can serve to reduce the complexity of the model. Subsequently, the one-time aggregation module VoV-GSCSP is used instead of the C3 module, and the slim-neck module is formed with GSConv to improve the accuracy. A coordinate attention module is incorporated into the end of the neck as a way to improve the algorithm's detection ability. Finally, SIoU is used instead of CioU as the loss function to improve the accuracy again and accelerate the convergence speed as a way to better adapt to the sea surface environment. The improved network structure is shown in Figure 1.

2.2. Receptive Field Enhancement (RFE) Module

Dilated convolution, also called expansion convolution, is simply the process of expanding the convolution kernel by adding some spaces (zeros) between the elements of the convolution kernel as a way of enlarging the sensory field [17]. Yu et al. [18] proposed the receptive field enhancement module, which is divided into two parts, multi-branching based on the dilation convolution, and an aggregation weighting layer. It is characterized using 4 different scaled expansion convolution branches to capture multi-scale information and different dependency ranges, and the weights are shared among these branches; the only difference is the different acceptance domains, and its structure is shown in Figure 2.

We draw on the idea of study [18] to introduce the RFE module into the backbone, and constitute the SPPF–RFE module with SPPF, which effectively expands the receptive field and improves the information extraction ability for multi-scale targets. In this module 1, 2, and 3 are used as different expansion rates, and they all use a fixed 3×3 convolution kernel with a residual joining to prevent the problems of gradient explosion and gradient vanishing, while collecting information from different branches and weighting each branch of the feature. For the sea vessel scenario, different scales can be manipulated to fully apply each sample.

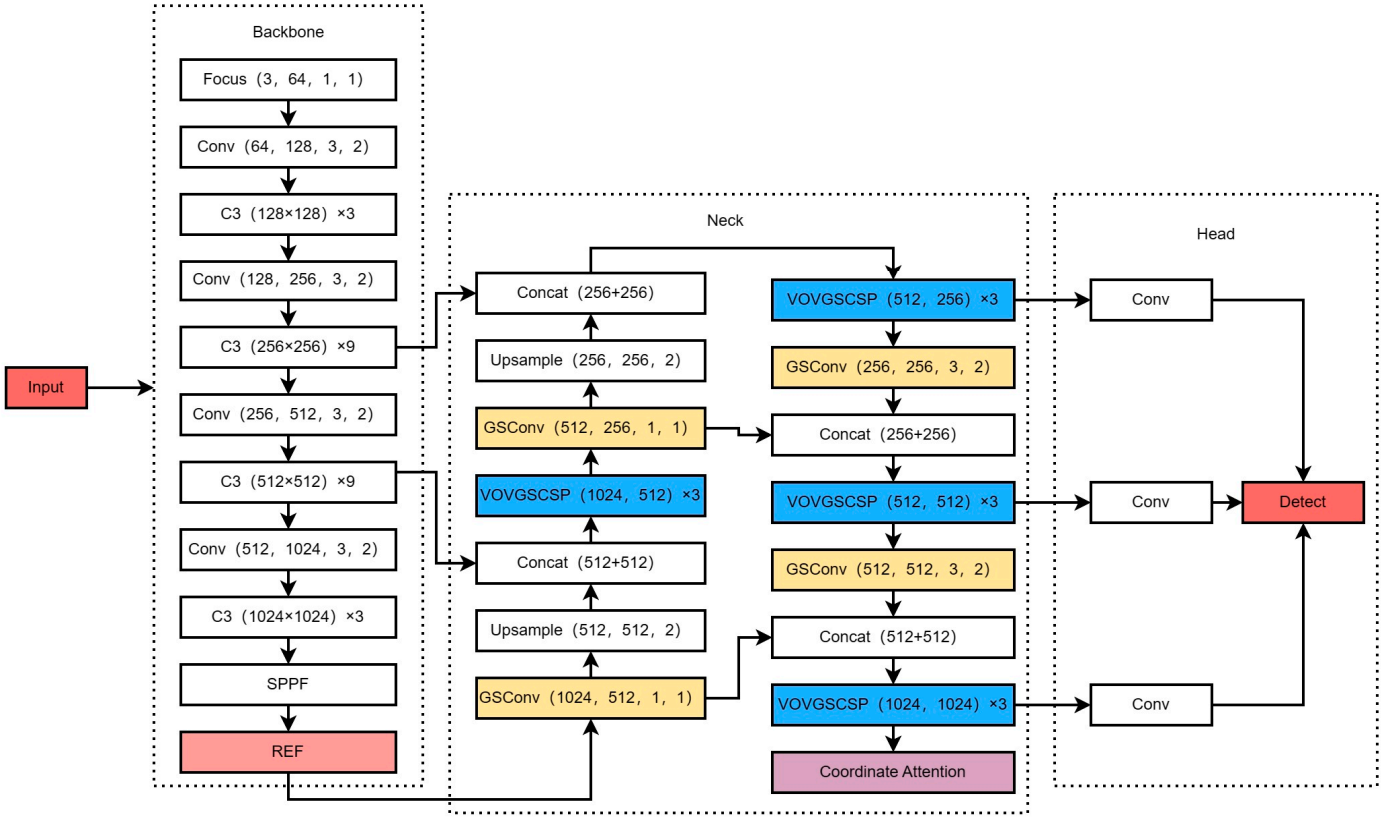

Figure 1. Improved network structure diagram.

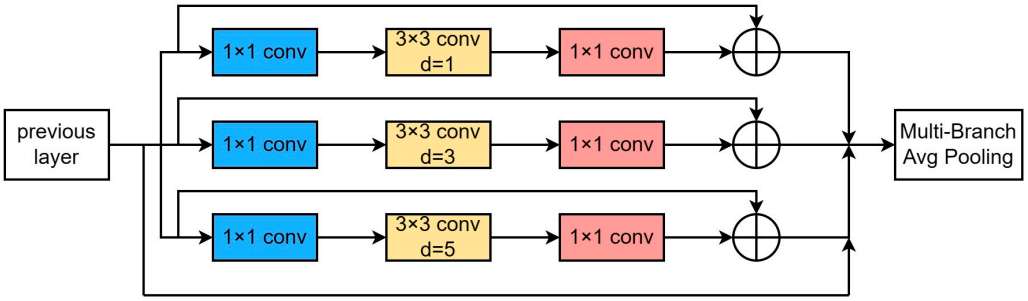

Figure 2. RFE module framework diagram.

2.3. Slim-Neck Structure Based on Lightweight Convolutional GSConv Modules

With the increase in accuracy, the number of layers of the network will increase, but the obvious effect is the decrease in the rate. In improved YOLOv5, with the addition of modules, there is an inevitable growth in the number of parameters and calculations. This situation increases the computational overhead, making it more difficult to deploy on mobile equipment. In order to appropriately balance the number of parameters and accuracy, we use a lightweight convolutional approach to improve the model. The lightweight models such as Xception, MobileNets, and ShuffleNets, greatly improve the speed of the detector in DSC operation, but have large flaws in their accuracy. Li et al. [19] pioneered the SC, the DSC and Shuffle together organically, using Shuffle to permeate the information generated by SC operation into various parts of the information generated by DSC.

This method mixes the information from SC completely into the output of DSC by exchanging the local feature information uniformly on different channels. The use of GSConv minimizes the negative impact of DSC defects on the model, and effectively exploits the advantages of DSC. Li et al. were inspired by the application of GSConv to automated driving, and introduced GSConv into ship detection algorithms, which reduces the amount of computation of YOLOv5, while ensuring that the output of the convolutional computation remains as constant as possible. However, if GSConv is used in all stages of the model, the network layer of the model will be deeper, which significantly increases the inference time. After experimental comparisons as well as theory from the literature [14], the choice was made to replace the Conv module in the neck part of YOLOv5 with a GSConv module, while the Conv module in the backbone part remains unchanged. This move can reduce some computation, while improving the accuracy. Table 1 shows a speed comparison between the original Conv and GSconv of YOLOv5 under RTX3090, where FPS is the number of transmitted frames per second, FLOPs is the number of floating-point operations, and Params is the number of parameters.

Table 1. Convolutional model data comparison.

Name	FPS	FLOPs	Params
Conv2D	491.43	38.789G	295.68K
GSConv2D	473.33	19.881G	151.42K

Li et al. also proposed a GSConv-based slim-neck module in study [19]. GS bottleneck was introduced on the basis of GSConv, after which VOV-GSCSP was designed using the one-time aggregation method. Under different algorithms, different datasets, and different application scenarios, GSConv, GS Bottleneck, and VOV-GSCSP should be applied flexibly.

The slim-neck module composed of GSConv and VOV-GSCSP was chosen for this improvement, and the Conv module in the neck part of YOLOv5 is replaced by GSConv, and the C3 module in the neck part of YOLOv5 is replaced by VOV-GSCSP. After experimental comparisons, the improved YOLOv5 algorithm improves in accuracy compared to the original version.

2.4. Coordinate Attention

The core logic of the attention mechanism is to focus on the key issues, meanwhile some less important information would be overlooked. Qibin Hou [20] et al. analyzed previous excellent attention modules such as SE (squeeze-and-excitation attention) and CBAM (convolutional block attention module), and came to the conclusion that their spatial location information is lost in the process of modeling channel relations. However, other attention modules without this problem are also effective, but the number of parameters is too large to be applied to the network of mobile devices; thus, this coordinate attention mechanism is also proposed.

For coordinate attention mechanism, in order to obtain the attention on the width and height of the image and encode the precise location information, the input feature map is firstly divided into width and height directions for global average pooling, respectively, to obtain the feature maps in the width and height directions, respectively.

$$z_c^h(h) = \frac{1}{w} \sum_{0 \le i \le w} x_c(h, i) \tag{1}$$

$$z_c^h(w) = \frac{1}{H} \sum_{0 \le i \le H} x_c(j, w) \tag{2}$$

Based on the two features generated above, the two feature maps are further subjected to a combining operation, followed by a transform operation using a 1×1 convolution, as follows:

$$f = \delta\left(F_1\left[z^h, z^w\right]\right) \tag{3}$$

where F_1 is the 1×1 convolutional transform function, square brackets denote the combining operation along the spatial dimension, and δ is the nonlinear activation function *h*-Swish. The intermediate feature mapping f is decomposed into two separate tensors $f^h \in R^{C/r \times H}$ and $f^w \in R^{C/r \times W}$, and r is the module size reduction rate. Transforming f^h and f^w into tensors with the same number of channels and undergoing sigmoid activation, the resulting g^h and g^w, respectively, are the following:

$$g^h = \sigma\left(F_h\left(f^h\right)\right) \tag{4}$$

$$g^w = \sigma(F_w(f^w)) \tag{5}$$

Finally, the output of the attention module is obtained as follows:

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(i) \tag{6}$$

In this research, the coordinate attention mechanism was added to the last layer of the neck part to strengthen the feature extraction ability and improve the network accuracy.

2.5. Border Loss Function Improvement

We use SIoU as the bounding box loss function to achieve a more accurate loss calculation between the predicted frame and the real frame in the traffic sign detection task. IoU (intersection over union) is the intersection over union ratio of the predicted frame and the real frame, which is used to measure the accuracy of the predicted frame. As in Equation (7), the closer its value is to 1, this indicates that the model is more effective and closer to the true value.

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{7}$$

In YOLOv5-7.0 version, CIoU is used as the loss function by default. It is compared with GIoU and DIoU, which consider the overlap area, center point distance, and aspect ratio at the same time. And SIoU further considers the vector angle between the real frame and the predicted frame, and redefines the related loss function, which contains four parts: angle loss, distance cost, shape cost, and IoU cost [21]. SIoU is defined as follows:

$$LOSS_{SIOU} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{8}$$

where Δ is the distance loss, and Ω is the shape loss.

SIoU speeds up the convergence of the network because it aids the calculation of the distance between the two frames by introducing the concept of angle between the real and predicted frames. In the case of comparable computation, SIoU can better accelerate the network convergence and achieve better results.

The flowchart of the improved YOLOv5 method is shown in Figure 3. The improved YOLOv5 can perform mosaic data enhancement on the input images and compute the adaptive anchor frame, and use *k*-means clustering to obtain *n* anchor frames. Then, the image is sent to the focus structure for a concat operation. Then, a series of convolution and pooling are carried out, and the obtained results are sent into the REF + SPPF module to enlarge the receptive field. Next, the slim-neck module is introduced to better balance the accuracy and speed of the model. Finally, coordinate attention mechanism is used to capture location information to improve the accuracy of the model.

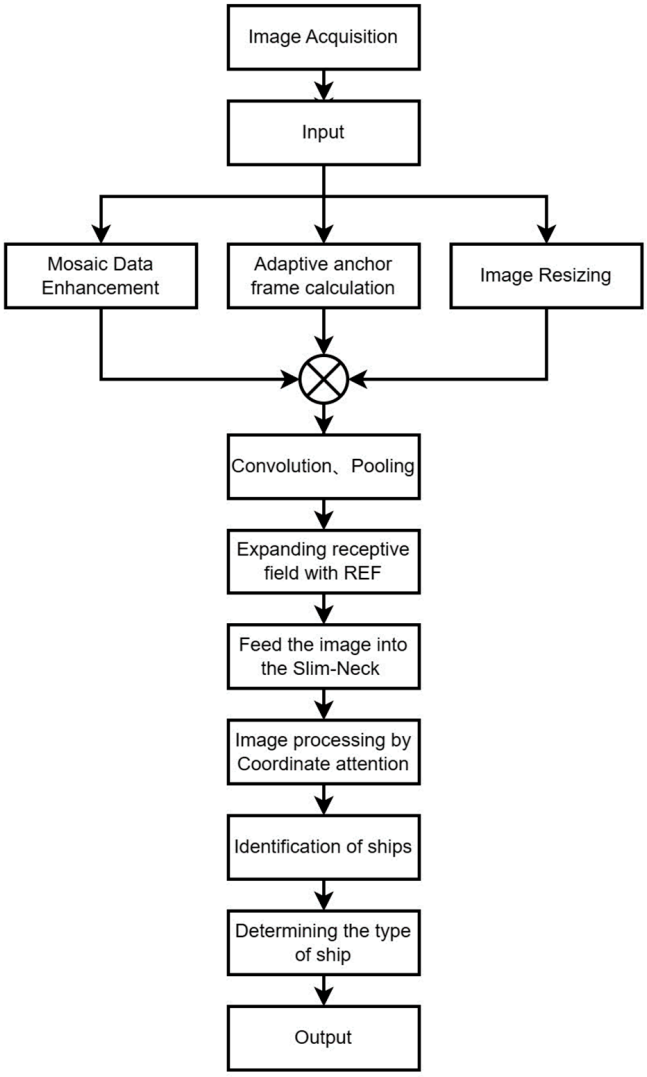

Figure 3. Flowchart of ship detection.

3. Experiments

3.1. Experimental Environment and Dataset

The experimental environment is Windows 11, AMD Ryzen 7 5800H, with Radeon Graphics@3.20GHz, 16 GB RAM, NVIDIA Ge Force RTX 3060, and 6 GB of video memory. The compiled language is Python3.10.9. The deep learning framework is Pytorch as 2.0.0, and IDE (integrated drive electronics) for Pycharm2022.3 community version. The benchmark model is YOLOv5s.

The dataset is the SeaShips dataset proposed by Shao et al. [22]. It is derived from visible light monitoring images of ships at sea, and consists of 6 types of ships, mining ships, general cargo ships, bulk carriers, container ships, fishing ships, and passenger ships. The dataset was carefully selected to mostly cover all possible imaging variations, for example, different scales, hull parts, illumination, viewpoints, backgrounds, and occlusions. All of the images are annotated with ship-type labels and high-precision bounding boxes. This dataset was also used in both study [6] and study [11].

Mosaic data enhancement was used in the training process, i.e., randomly selecting four images to splice and fuse, which could enrich the extracted ship target. The results after stitching the spliced images are shown in Figure 4. The numbers represent the predicted ship types, 0 for an ore carrier, 1 for a general cargo ship, 2 for a bulk cargo carrier, 3 for a container ship, and 4 for a fishing boat.

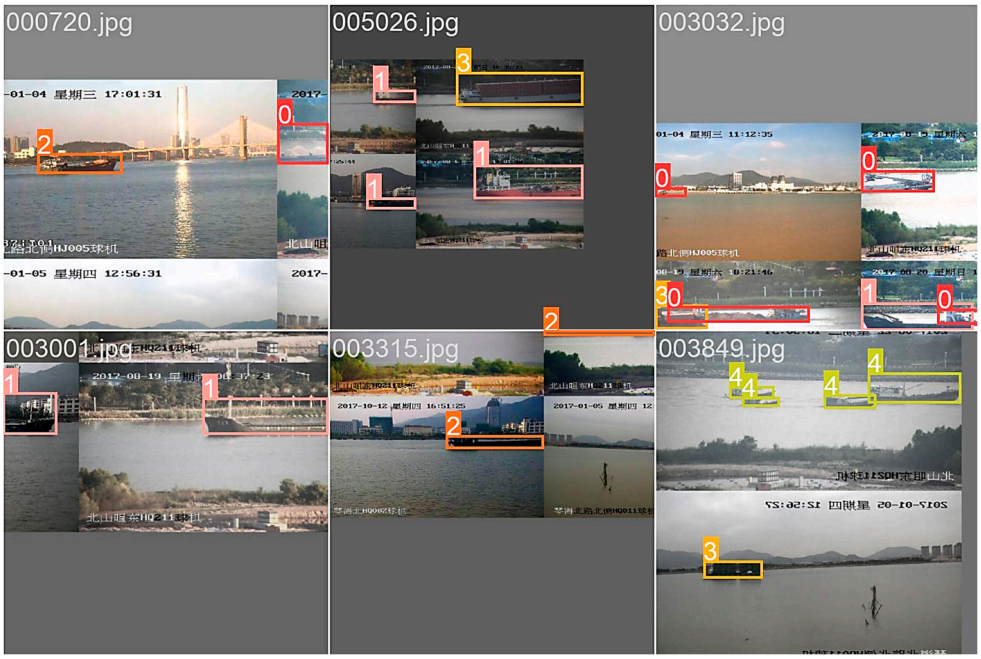

Figure 4. Mosaic data enhancement.

The original dataset contains more than 30,000 images, in which 7000 images from the open source were chosen for this experiment. Since the distribution of the categories in those images is not balanced, the majority are mining ships and bulk carriers, while container ships and passenger ships are relatively fewer. The images with adjacent l numbers may be taken from the adjacent moment, such as the previous and next second in a video, so that those images may have tiny differences. For the above reasons, if the dataset is randomly divided, this easily leads to overfitting; for example, the small number of fishing boats and container ships may have poor training results. Therefore, in this experiment, the 7000 images are firstly divided into 6 subsets according to the types of ships, and then the 6 subsets are divided into a training set, validation set, and test set according to the ratio of 8:1:1; then, they are summarized to become the training set, validation set and test set of the whole dataset. There are 5600 training sets, 700 validation sets, and 700 test sets. Its overall distribution histogram is shown in Figure 5, where blue represents a bulk carrier, orange is a container ship, green is a fishing ship, red is a bulk carrier, purple is an ore carrier, and brown is a passenger ship.

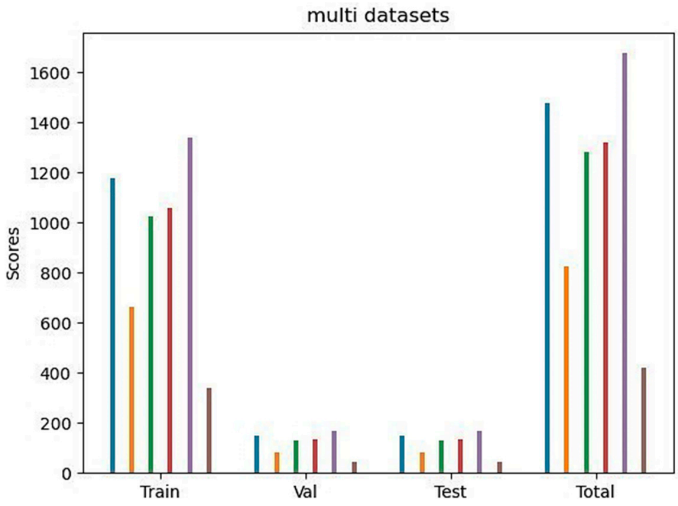

Figure 5. Histogram of the overall distribution of the dataset.

It can be seen that after the division, there is a certain proportion of various ships in the training set, validation set, and test set, and the proportions of the three subsets are the same. The division can effectively avoid overfitting and improve the training effect. Some samples are shown in Figure 6. Figure 6a,b represent images of the ship collected at different moments in time and at different locations, respectively.

(**a**) (**b**)

Figure 6. Image samples from the training set.

3.2. Training Parameter Settings

After experimental comparison, suitable training parameters were selected to train the improved YOLOv5 network. The specific parameters are shown in Table 2.

Table 2. YOLOv5 training parameters.

Parameters	Value
Epochs	300
Batch-Size	16
Optimizer	Adam
Learning Rate	0.01
Mosaic	1.0
Momentum	0.937
Weight-Decay	0.0005

3.3. Model Evaluation Indicators

These parameters, precision, recall, mean average precision (mAP), and FPS parameters are used as evaluation indexes, where mAP includes mAP50 and mAP50:95 [23].

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{11}$$

In the above equation, *TP* (true positive) represents the number of detected frames that satisfy the IoU (intersection over union) ratio of predicted and labeled frames greater than 0.5. *FP* (false positive) represents the number of detected frames that satisfy the IoU $\leq$ 0.5. *FN* (false negative) represents the number of undetected labeled frames. *AP* represents the area of the precision–recall curve calculated by interpolation. *AP* represents the area of the precision–recall curve calculated by interpolation. mAP50 is the *m AP* of all of the images when the IoU threshold is 0.5, and mAP50:95 is the *m AP* of all of the images under different IoU thresholds (IoU thresholds ranging from 0.5 to 0.95 with a step size of 0.05).

4. Results and Discussion

We compared the different roles of the various improved modules through comparative experiments. In the training, due to improvements in the backbone part, all of the training processes didn't used the official pre-training model. The size of the input images is 640 × 640. A total of 300 rounds of training were used. The training would be stopped if the continuous 50 rounds were verified without any improvement. On the basis of this experiment, the optimal model for comparison was selected. According to the experiment, the optimal model comparison is shown in Table 3.

Table 3. Comparison of evaluation indexes of different training models.

Model	Precision	Recall	mAP50	mAP50:95	Params	FPS
YOLOv5	0.943	0.929	0.966	0.679	7M	52.08
YOLOv5 + SIoU	0.941	0.943	0.968	0.671	7M	55.25
YOLOv5 + SIoU+RFE	0.942	0.943	0.972	0.681	7.8M	50.51
YOLOv5 + SIoU + RFE + GSConv	0.944	0.945	0.972	0.689	7.4M	52.63
YOLOv5 + SIoU + RFE + GSConv + VOV-GSCSP	0.948	0.949	0.971	0.7	7.9M	50.00
YOLOv5 + SIoU + RFE + GSConv + VOV-GSCSP + CA	**0.957**	**0.955**	**0.978**	**0.71**	7.9M	49.50

From the table, it can be seen that when the edge loss function is changed from CIoU to SIoU, the recall and mAP50 are improved, which is due to the fact that compared to the previous design, the angle between the two frames is taken into account more in addition to the overlapping region, distance, and length and width. With almost no change in the number of parameters (Params), SIoU obtains an increase in precision and recall, which meets the demand of high precision and high recall for surface vessel detection. Meanwhile, the introduction of angle accelerates the convergence of the concept of distance, thus making the convergence faster and preventing the network from hovering around the optimal point, so all subsequent experiments used SIoU as the loss function.

Compared with only using SIoU after adding the RFE module, the mAP50 is further improved to 0.972, mAP50:95 to 0.681, and the recall to 0.943. The main reason is that the repeated use of pooling layer in the SPPF module loses detailed information, and it is easy to misidentify the target as background. After the introduction of the RFE module, more details can be retained while reinforcing the semantic information, which improves the performance of the model, resulting in an increase in both precision and recall.

After adding the lightweight convolutional GSConv, a more significant reduction in the number of parameters can be seen, while the accuracy, recall, and mAP50:95 continue to improve. After GSConv is combined with VOV-GSCSP to form the slim-neck module, the number of parameters is also comparable to that before it was added, while the accuracy, recall, and mAP50:95 still improve. The reason is that the lightweight convolutional GSConv can better balance the accuracy and speed of the model, and after forming the slim-neck module, this makes it possible to achieve higher computational cost benefits without a significant increase in the number of parameters.

After the final addition of coordinate attention mechanism, all of the indexes were greatly improved. The introduction of coordinate attention mechanism further enhanced the important features in the network, and comprehensively improved the model detection ability of the network, which helps to detect more targets and localize them more accurately. The improved YOLOv5s has faster convergence speeds and better accuracy. A comparison of the training process between YOLOv5 and the improved YOLOv5 algorithm is shown in Figure 7.

In order to more intuitively reflect the difference in the performance of the above detection algorithms, a portion of the detected images are selected for demonstration, as shown in Figure 8.

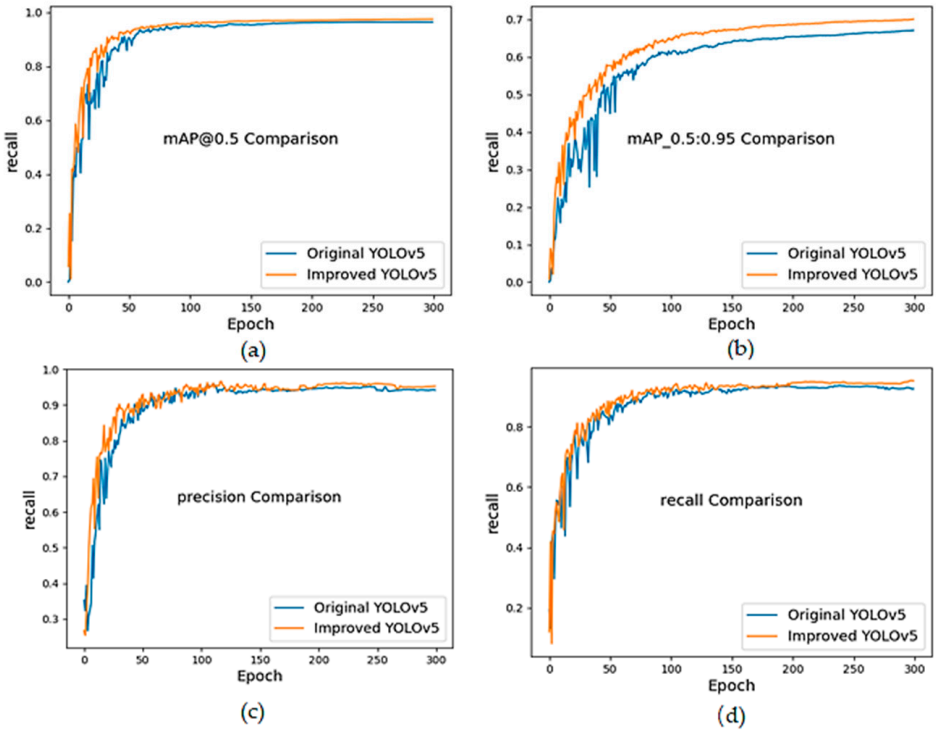

Figure 7. Comparison of YOLOv5 and improved YOLOv5; (**a**) mAP0.5; (**b**)mAP0.5:0.95; (**c**) precision; (**d**) recall.

As can be seen from the comparison chart, Figure 7a shows that for the same target, the confidence degree of the original YOLOv5 has been greatly improved, from 0.40 to 0.73. In Figure 7b, the original YOLOv5 incorrectly recognizes the container ship as a general cargo ship, while the improved algorithm recognizes it correctly. Compared to the original algorithm, the improved YOLOv5 has improved in accuracy. In Figure 7c, there is a missed detection when there are multiple targets in the image, and some of them are small. The original YOLOv5 misses the fishing boat while the improved YOLOv5 recognizes the fishing boat accurately. The improved YOLOv5 shows improved recall compared to the original algorithm. In Figure 7d, there is a situation where there is more overlap between the two ships. When facing overlapping targets, the original YOLOv5 is lower than the improved YOLOv5 in confidence. The confi-dence of original YOLOv5 are 0.56 and 0.83, and the confidence of the improved YOLOv5 are increased to 0.73 and 0.85. The improved YOLOv5 possesses stronger recognition ability when facing overlapping targets.

Based on the above situation, the improved YOLOv5 algorithm's ability to detect overlapping targets and various types of ships was significantly improved, and the error rate was reduced. The improved YOLOv5 algorithm has a better recognition effect for false identification, missed identification, and overlapping targets.

The experiments show that each module improves the performance of YOLOv5 under this dataset in terms of precision, recall, mAP50, and mAP50:95. Finally, in terms of precision, the improved YOLOv5 improves the performance over the original YOLOv5 by 1.4%, and in terms of recall by 2.6%, in terms of mAP50 by 1.2%, and in terms of mAP50:95 by 3.1%, with smaller model parameters and faster detection speeds, which can meet the needs of sea vessel detection.

We used a public dataset in this experiment. Due to limitations in the dataset, there are few images relating to bad weather in it. However, we believe that the receptive field enhancement module and the coordinate attention mechanism can improve the detection ability of YOLOv5 under bad weather.

Figure 8. Comparison of partially detected images; (**a**) confidence improvement; (**b**) false identification improvement; (**c**) missed identification improvement; (**d**) multi-target overlapping.

5. Conclusions

In this study, to address the problems of insufficient precision and low recall arising from ship detection, an improved YOLOv5 algorithm for ship target detection was proposed by improving the backbone part, neck part, and loss function, using YOLOv5s as the baseline model. The experimental results show that the algorithm, with no large increase in the number of parameters and detection speeds that meet real-time requirements, is effective in improving the performance of the algorithm by improving the sensing field, using lightweight convolution, and improving the loss function; thus, the performance of the algorithm is effectively improved.

Restrictions in the current model exist in the recognition of targets from complex backgrounds. Currently, due to limitations in the public dataset used, this study was limited to ship detection in mostly good weather. Although the sample is more than adequate, the background is mostly simple, with the sea and harbor as the main focus. Future research will further enrich the maritime traffic elements in the ship dataset to ensure that ships can be accurately identified, even in more complex maritime traffic environments. In practical applications, such as facing rain, snow, haze, and other bad weather, image enhancement methods such as panning, flipping, contrast enhancement, and other image enhancement methods can be additionally added, or algorithmic fusion such as dark channel de-fogging algorithm can be used as a means of enhancing the robustness and

generalization of the model; these measures will provide a technical basis for the next step in carrying out complex tasks such as ship tracking under complicated conditions. At the same time, while ensuring that the algorithm has high detection accuracy, strong robustness, and good model performance, efforts will be made to improve the detection speed of the model, appropriately delete the redundant structure of the algorithm, lighten the model, and enhance the portability of the algorithm.

Additionally, the next step will be to try to apply the improved algorithm to other goals. We designed the algorithm for ships because there is a practical need for it. We believe that the algorithm can also be effective in detecting other targets, especially objects that are similar in shape to ships (e.g., cars), because the receptive field enhancement module and lightweight module are beneficial for target detection. In the future, we will continue to try to explore the generalizability of the algorithm.

Author Contributions: Conceptualization, L.W. and Y.Z.; methodology, L.W.; software, L.W.; validation, Y.L. and Y.X.; investigation, L.W. and Y.Z.; resources, Y.Z. and B.S.; data curation, L.W. and S.Y.; writing—original draft preparation, L.W.; writing—review and editing, Y.Z. and B.S.; supervision, Y.Z.; project administration, Y.Z. and B.S.; funding acquisition, Y.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Shanghai Science and Technology Plan Project, grant number 20040501200.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: No new data were created in this study. The data analyzed in this study are from dataset SeaShips (accessed on 1 June 2023), which could be downloaded from the website, https://github.com/jiaming-wang/SeaShips.

Acknowledgments: The authors thank the staff from the Experimental Auxiliary System and the Information Center of Shanghai Synchrotron Radiation Facility (SSRF) for on-site assistance with the GPU computing system.

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

References

1. Ross, G.; Jeff, D.; Trevor, D.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; IEEE: New York, NY, USA, 2014; pp. 580–587. [CrossRef]
2. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y. SSD: Single shot multi-box 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. [CrossRef]
3. Joseph, R.; Santosh, D.; Ross, G.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE: New York, NY, USA, 2016; pp. 779–788. [CrossRef]
4. Tsung, L.; Priya, G.; Ross, G.; He, K.; Dollár, P. Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [CrossRef]
5. Shi, Z.; Long, W.; Xiao, B.; Lei, Z.; Li, S.Z. Single-shot refinement neural network for object detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 4203–4212. [CrossRef]
6. Liu, R.W.; Yuan, W.; Chen, X.; Lu, Y. An enhanced CNN-enabled learning method for promoting ship detection in maritime surveillance system. *Ocean Eng.* **2021**, *235*, 109435. [CrossRef]
7. Zhou, L.; Wei, S.; Cui, Z.; Fang, J.; Yang, X.; Wei, D. Lira-YOLO: A lightweight model for ship detection in radar images. *J. Syst. Eng. Electron.* **2020**, *31*, 950–956. [CrossRef]
8. Hong, X.; Cui, B.; Chen, W.; Rao, Y.; Chen, Y. Research on multi-ship target detection and tracking method based on camera in complex scenes. *J. Mar. Sci. Eng.* **2022**, *10*, 978. [CrossRef]
9. 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]
10. Frederik, H.; Jamal, A. Deep learning for detecting and classifying ocean objects: Application of YoloV3 for Iceberg–ship discrimination. *Int. J. Geo-Inf.* **2020**, *9*, 758. [CrossRef]

11. Ye, Y.; Zhen, R.; Shao, Z.; Pan, J.; Lin, Y. A novel intelligent ship detection method mased on attention mechanism feature enhancement. *J. Mar. Sci. Eng.* **2023**, *11*, 625. [CrossRef]
12. Li, Y.; Zhu, W.; Li, C.; Zeng, C. SAR image near-shore ship target detection method in complex background. *Int. J. Remote Sens.* **2023**, *44*, 924–952. [CrossRef]
13. Kong, W.; Liu, S.; Xu, M.; Yasir, M.; Wang, D.; Liu, W. Lightweight algorithm for multi-scale ship detection based on high-resolution SAR images. *Int. J. Remote Sens.* **2023**, *44*, 1390–1415. [CrossRef]
14. Patel, K.; Bhatt, C.; Mazzeo, P.L. Improved ship detection algorithm from satellite images using YOLOv7 and graph neural network. *Algorithms* **2022**, *15*, 473. [CrossRef]
15. 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. [CrossRef]
16. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018. [CrossRef]
17. Fisher, Y.; Vladlen, K. Multi-scale context aggregation by dilated convolutions. *arXiv* **2016**, arXiv:1511.07122. [CrossRef]
18. Yu, Z.; Huang, H.; Chen, W.; Su, Y.; Liu, Y.; Wang, X. YOLO-FaceV2: A scale and occlusion aware face detector. *arXiv* **2022**, arXiv:2208.02019. [CrossRef]
19. Li, H.; Li, J.; Wei, H.; Liu, Z.; Zhan, Z.; Ren, Q. Slim-neck by GSConv: A better design paradigm of detector architectures for autonomous vehicles. *arXiv* **2022**, arXiv:2206.02424. [CrossRef]
20. Qi, H.; Da, Z.; Jia, F. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021. [CrossRef]
21. Zhora, G. SIoU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740. [CrossRef]
22. Shao, Z.; Wu, W.; Wang, Z.; Du, W.; Li, C. SeaShips: A large-scale precisely annotated dataset for ship detection. *IEEE Trans. Multimed.* **2018**, *20*, 2593–2604. [CrossRef]
23. Fang, X.; Bao, L.; Ying, L. Research on the coordinate attention mechanism fuse in a YOLOv5 deep learning detector for the SAR ship detection task. *Sensors* **2022**, *22*, 3370. [CrossRef]