Journal of *Marine Science and Engineering*

MDPI

Article

Improved Ship Object Detection in Low-Illumination Environments Using RetinaMFANet

Ruixin Ma [1,2], Kexin Bao [2] and Yong Yin [1,*]

[1] Key Laboratory of Marine Simulation and Control, Dalian Maritime University, Dalian 116026, China
[2] Tianjin Research Institute for Water Transport Engineering, M.O.T., Tianjin 300456, China
* Correspondence: bushyin@dlmu.edu.cn

Abstract: Video-based ship object detection has long been a popular research issue that has received attention in the water transportation industry. However, in low-illumination environments, such as at night or in fog, the water environment has a complex variety of light sources, video surveillance images are often accompanied by noise, and information on the details of objects in images is worsened. These problems cause high rates of false detection and missed detection when performing object detection for ships in low-illumination environments. Thus, this paper takes the detection of ship objects in low-illumination environments at night as the research object. The technical difficulties faced by object detection algorithms in low-illumination environments are analyzed, and a dataset of ship images is constructed by collecting images of ships (in the Nanjing section of Yangtze River in China) in low-illumination environments. In view of the outstanding performance of the RetinaNet model in general object detection, a new multiscale feature fusion network structure for a feature extraction module is proposed based on the same network architecture, in such a way that the extraction of more potential feature information from low-illumination images can be realized. In line with the feature detection network, the regression and classification detection network for anchor boxes is improved by means of the attention mechanism, guiding the network structure in the detection of object features. Moreover, the design and optimization of the augmentation of multiple random images and prior bounding boxes in the training process are also carried out. Finally, on the basis of experimental validation analysis, the optimized detection model was able to improve ship detection accuracy by 3.7% with a limited decrease in FPS (frames per second), and has better results in application.

Keywords: deep learning; computer vision; ship object detection; RetinaNet; low-illumination environment

Citation: Ma, R.; Bao, K.; Yin, Y. Improved Ship Object Detection in Low-Illumination Environments Using RetinaMFANet. *J. Mar. Sci. Eng.* **2022**, *10*, 1996. https://doi.org/10.3390/jmse10121996

Academic Editors: Baran Yeter and Yordan Garbatov

Received: 25 October 2022
Accepted: 6 December 2022
Published: 15 December 2022

1. Introduction

In recent years, the water transport industry has primarily supervised ships via AIS (automatic identification system), but this depends on the construction of shipborne terminals and shore-based base stations, and cannot play a regulatory role in the case of ships that have not installed AIS, or that have intentionally turned off AIS in order to carry out illegal activities, and, therefore, evidence cannot be obtained on the spot. In light of these regulatory issues, video surveillance has emerged as a necessity due to its possessing the characteristics of intuitiveness and readability. However, traditional video surveillance can only be used for observation and to store evidence; it is not able to intelligently analyze it. Therefore, ship target detection when using video surveillance is critical. Many scholars are currently conducting research on video-based ship object detection technology, and object detection under low illumination is one of the most difficult tasks, representing a hotspot and a source of major difficulty in current research.

1.1. Object Detection Methods

Methods for video-based object detection can be broadly categorized into two types: traditional object detection algorithms and deep learning detection approaches.

In traditional object detection algorithms, the general technical framework consists of horizon detection, background subtraction, and foreground segmentation. In particular, horizon detection methods can be classified as follows: methods based on line features [1], methods based on area modeling [2], and hybrid methods incorporating both methods [3]. Background subtraction methods can be classified into three categories: the first determines whether each pixel belongs to the foreground or the background based on statistical information from a single image; the second calculates the probability that pixels belong to the foreground and background using GMM (Gaussian mixture model); and the third extracts image features from each pixel or small area and calculates the probability that it belongs to the foreground or background [4]. Foreground segmentation employs morphological methods to generate detection results based on the results of background subtraction. The specific detection algorithm generally employs the HOG (histogram of oriented gradients) method in conjunction with SVM (support vector machine) [5]. In addition, the majority of such methods use the method of artificial feature extraction. However, this is difficult to apply in cases of object detection at multiple scales in complex environments, such as at low illumination, due to its high subjectivity and single feature.

In recent years, the performance of deep learning-based object detection algorithms has surpassed that of traditional algorithms in many fields with increases in data volume and computing power. Deep learning is a data-driven method that trains the model on a large number of datasets in order to achieve the desired level of cognitive ability [6]. This method does not require artificial design features and has a greater capacity for generalization. Deep learning-based object detection algorithms can be divided into two-stage and one-stage algorithms. The former first generate candidate object frames, which are then classified and regressed, and include R-CNN (region-based convolutional neural network) [7], Fast R-CNN [8], and Faster R-CNN [9]. The latter do not generate object candidate boxes and instead rely solely on a neural network to complete classification and regression, with these including SSD (single-shot Multibox detector) [10], YOLO (you only look once) [11–13], FCOS (fully convolutional one-stage object detection) [14], and RetinaNet [15]. Scholars have improved and perfected ship object detection algorithms in recent years using deep learning models. Zou et al. [16] tested Faster R-CNN and SSD on SMD datasets [17], obtaining mAP (mean average precision) values of 84.33% and 80.23%, respectively, and used ResNet [18] to replace the infrastructure VGG (visual geometry group), obtaining an mAP of 88.08%. Scholler et al. [19] used a long-wavelength infrared image dataset to test Faster R-CNN, RetinaNet, and YOLOv3, and obtained mAP values of 81%, 86%, and 90%, respectively. Significantly, RetinaNet's mAP was improved to more than 90% after image preprocessing. The attention mechanism [20] was added to the classification and detection network in the RetinaNet model to enhance the network feature extraction ability, and the improved RetinaNet model demonstrates good detection performance.

Ship object detection is similar to general object detection. Furthermore, the accuracy of the two-stage detection network is greater than that of the one-stage detection network, but the detection speed is slower. Thus, this paper used RetinaNet as the basic detection algorithm in order to achieve real-time edge detection, and improved its network architecture based on the technical characteristics prevailing in low-illumination environments, in order to improve object detection accuracy without significantly reducing detection speed.

1.2. Object Detection Methods in Low-Illumination Environments

Object detection based on deep learning in settings with low illumination generally can be viewed as a specific improvement over the standard object detection algorithm. LV [21] proposed a low-illumination image object detection algorithm based on improved SSD, and the original low-illumination image was enhanced using an image enhancement algorithm based on Retinex theory. In this method, a dual-branch SSD structure

is designed, and the ResNet50 network is used to replace the original VGG16 feature extraction network. The dual-branch structure incorporates a differential feature fusion module (DFF) to improve the model's extraction effect on complementary features and the algorithm's detection accuracy on low illumination image objects. LI [22] proposed an algorithm for the fusion of infrared and visible images under different visual angles based on saliency detection, using a Mask R-CNN network to extract the object saliency regions in the infrared image, locally fusing each object region with the visible image according to the field of view conversion model point, and combining the main infrared image with a clear visible background. The fused image can help with object detection in low-light situations. To address the issue of low accuracy of multi-scale pedestrian object detection in low-illumination environments, CHEN [23] proposed a pedestrian detection method based on YOLOv5s and infrared and visible image fusion. The generation countermeasure network is utilized to generate the visible light and infrared fusion image dataset. The SENet (squeeze and excitation networks) channel attention module is incorporated into YOLOv5s so that the network can pay more attention to the highlighted objective, thus enhancing the mAP of pedestrian detection. A network model of radar and camera feature fusion was proposed by CHANG [24] for the impact of bad scenes, such as low light, rain, and fog, on the detection capability of intelligent driving vision systems. The radar attention mechanism feature module RCBAM (radar of convolutional block attention module) was built using millimeter wave radar information and an attention model. Because radar data are not easily impacted by weather or light, the addition of the RCBAM module's feature fusion network can considerably enhance the object detection network's robustness and anti-interference capability. Based on the single-stage object detection algorithm RetinaNet in deep learning, LIU [25] combined the characteristics of the SAR (synthetic aperture radar) image with less feature information, adopted the idea of multi-feature layer fusion, and proposed a more appropriate loss function calculation method. Based on the traditional YOLOv3, an enhanced YOLOv3 algorithm for ship detection was proposed by NIE [26]. The prediction box uncertain regression, the negative logarithm likelihood function, and the improved binary cross-entropy function were used to redesign the loss function, and the non-maximum suppression (NMS) algorithm with Gaussian soft threshold function was used to post-process the prediction boxes. Their experimental results indicate that the proposed method is capable of delivering enhanced ship detection in adverse weather and environmental conditions, such as fog and low-light settings, as well as in navigation environments with complicated backdrops.

According to pertinent research and analysis, the specific application scenario of ship object detection in low-illumination environments determines that the technical challenges encountered by the general object detection algorithm are exacerbated, thereby increasing the performance requirements for the model. The existing algorithms suffer from the following technical issues:

(1) In low-illumination environments, such as at night, the light source is complex, background light interferes, the ship image is underexposed, contrast and brightness are low, and image quality is poor, resulting in less available information and trouble extracting sufficient effective features. Currently, the general object detection algorithm has limited ship object extraction capability and insufficient accuracy.

(2) In the ship monitoring image, both small ships, such as fishing boats and tugboats, and large ships, such as cruise ships and container ships, can be seen in the same frame. This means that the scales of the image data can be very different from one goal to the next. Smaller objects in the CNN model will correspond to smaller and smaller areas on the feature map, or they may even disappear during the convolution and pooling process. Additionally, if the network's receptive field is much bigger than the object's size in the deeper layer, it will be hard for the object to show up in the feature map.

(3) The ITU-R M.1371-5 proposal divides ships into dozens of types, such as passenger ships, freighters, oil tankers, tugs, sailboats, and yachts. Different subdivided ships look very different in shape, texture, and other ways. In real life, the camera's view angle

is usually not fixed, and ship images taken from different angles show ships that are various sizes, in different places, and with different attitudes. Furthermore, the water environment is more complicated than the land environment, and some general object detection algorithms can mistake ship reflections, water waves, and shore buildings for other things.

According to the above analysis, there are still numerous challenges to be tackled in ship object detection in low-illumination environments. In deep learning object identification, the low-level feature maps contain rich high-resolution features and narrow receptive fields, which facilitates the detection of image targets with low-illumination. However, high-level feature maps have semantic information with low resolution, which is essential for detecting massive objects. For the challenges, such as the diversity of ship object scales, the lack of available information in images, and the diversity of ship object shapes in locations with poor illumination environments. The effectiveness of a new object identification approach based on the RetinaMFANet (multiscale feature and attention network), which incorporates the multiscale feature fusion network and attention feature detection network, is evaluated through a design experiment.

2. The Framework of RetinaMFANet

This RetinaMFANet method combines the features of low-light images with less information, improves the lower image feature layer, fuses multiscale basic network feature information, and then adds an attention mechanism to the features extraction network to focus on the area of interest, which improves the accuracy of detection.

2.1. Feature Extraction Network

The feature extraction network of RetinaMFANet employs a deep residual network (ResNet) [18]. In general, the deeper the neural network, the more information it can obtain, and the more features it has, the better it works. Simply increasing the depth will result in gradient dispersion or gradient explosion, which will degrade network performance. However, the ResNet method introduces skip link lines and employs the concept of cross-layer connection to directly transmit the previous layer's output to the rear via identity mapping. Even if the network depth is increased, the network model will not degrade (as shown in Figure 1).

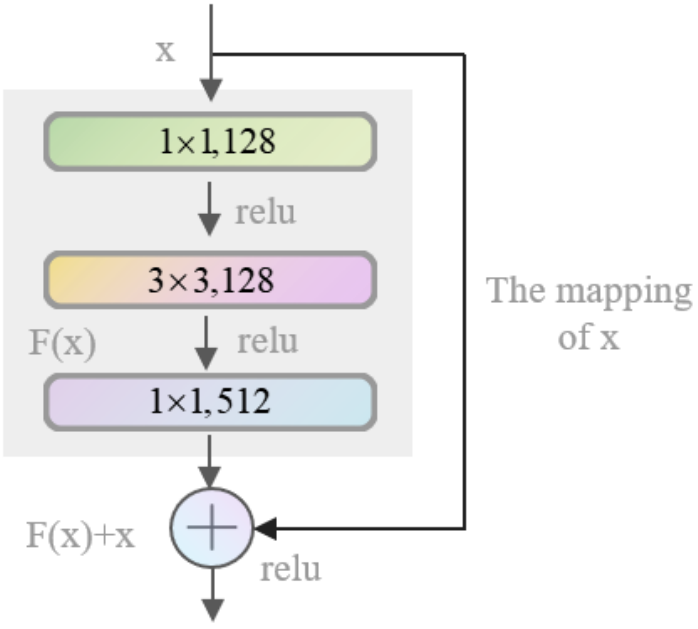


Figure 1. Residual learning module of ResNet.

RetinaMFANet is interesting because it uses the ResNet50 architecture and has five blocks with three convolutional layers each. In Table 1, the network structure parameters are shown.

Table 1. Feature extraction network (ResNet50) parameter.

Layer Name	Output Size	50-Layer
Conv1	112×112	$[7 \times 7, 64]$, stride 2
		$[3 \times 3]$ max pool, stride 2
Conv2_x	56×56	$\begin{bmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{bmatrix} \times 3$
Conv3_x	28×28	$\begin{bmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{bmatrix} \times 4$
Conv4_x	14×14	$\begin{bmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{bmatrix} \times 6$
Conv5_x	7×7	$\begin{bmatrix} 1 \times 1, 512 \\ 3 \times 3, 512 \\ 1 \times 1, 2048 \end{bmatrix} \times 3$
	1×1	Average pool, 1000-d fc softmax

2.2. New Multiscale Feature Fusion Network

Each level of a CNN model's features has a different receptive field, as do the sizes of objects that can be found. A high-resolution feature map are generated with more information, while deep features are produced by using neural networks to down-sample shallow features many times. Even though the receptive field is large and the global semantic information is complete, the loss of detailed information is significant. When small objects are in the deep feature image, a bigger receptive field makes the matching area on the feature image smaller. RetinaNet builds multi-scale feature maps using FPN (feature pyramid networks) [27] and finds objects on multi-scale feature maps with the suitable object scales and receptive fields. Figure 2 shows how it was built.

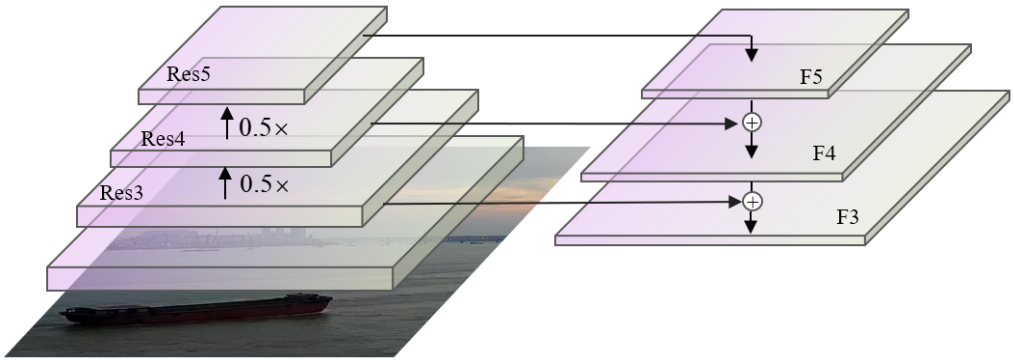
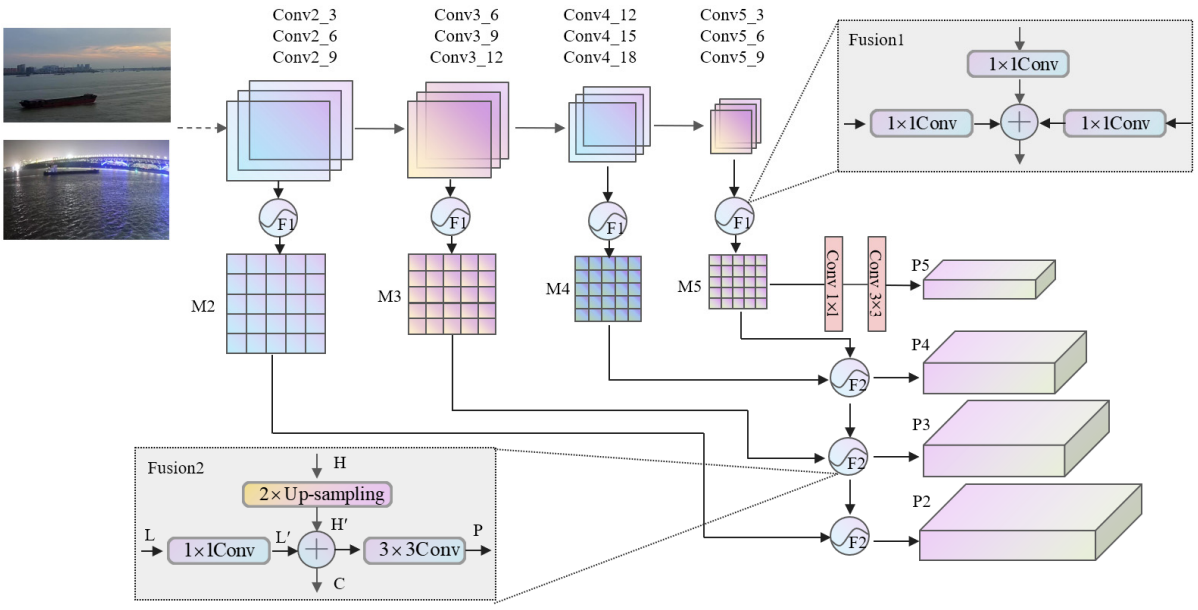


Figure 2. FPN structure in ResNet.

FPN principally utilizes ResNet feature maps at various levels to construct a multi-scale feature pyramid. Each feature layer is integrated with the deeper layer of high semantic information to improve the prediction effect of each layer. Furthermore, the shallow layer of high-resolution features is better suited for detecting small objects. Regarding the detection effect, shallow details are better for finding where an object is, while deep semantic information is better for recognizing what it is. However, during the combination process, FPN only adds deep-layer information sampled at the same latitude as shallow-layer information. This combination method ignores the structural error introduced by

feature map bilinear interpolation. Simultaneously, there is less valuable information in the low-light environment, which is more susceptible to the impact of error.

Thus, based on the issues of the current network, this paper proposes a new feature fusion network. The structure is shown in Figure 3. Fusion1 combines three layers with varying representation capabilities, extracts more potential information from images, and generates a more detailed feature pyramid. To establish a coupling relationship between deep and shallow features, Fusion2 uses the FPN-like approach to fuse the feature maps of different scales adjacent to the pyramid, establishing a connection from deep features to shallow features. Fusion1 convolves the outputs of the last three convolutional layers of Conv2, Conv3, Conv4, and Conv5 at the same dimension of 1×1, respectively. Additionally, then they are paralleled together and fused again by 1×1 convolution of the valid information to generate the fused M2, M3, M4, and M5. For example, Table 1 (Section 2.1) displays the network (ResNet50) structure. For the fusion operation of Conv2, Conv2_3, Conv2_6, and Conv2_9 are the last three layers of Conv2. They will be fed into the Fusion1 unit procedure, which will build the fused feature map M2. The same method is used to generate M3, M4, and M5, in turn. As the receptive field grows and the fine granularity of features decreases, the semantic information in these four feature maps become more abundant. M2 is a feature map fused from Conv2 including extra object-specific information. It is more sensitive to small items and locates objects more effectively. Conv5 has more semantic information and is more valuable for object categorization; hence, it is used to fuse the M5 feature map. On the fused feature structure, Fusion2 has built a connection between shallow features and deep features. L' is obtained by convolving the shallow feature L by 1×1 convolution, and H' is obtained by bilinear difference up-sampling of deep feature H. To eliminate the aliasing effect, we set $C = L' + H'$. It can reduce the structural error of the feature map via 3×3 convolution. Thus, this improved network outputs layer enables to extract more shallow feature information than the original one. The fused P2, P3, P4, and P5 are the final detection layers.



Note:Conv2_3, Conv2_6, Conv2_9 are the last three layers of Conv2. They will be fed into the Fusion1 unit procedure, which will build the fused feature map M2.

Figure 3. Improved feature fusion structure.

2.3. Attention Feature Detection Network

Before being sent to the feature detection sub-network for regression and classification of the previous bounding box, the detection feature maps P2, P3, P4, and P5 go through four 3×3 convolution layers. The RetinaNet algorithm performs border regression and classification separately on the feature map to ensure that the two different losses of

regression and classification have no influence between each other. Furthermore, all detection layers share a detection head, but each set of parameters is unique, which improves feature expression ability.

This paper focuses on ship object detection in low-light conditions. It enhances the detection approach of the RetinaNet feature detection subnetwork by introducing an attention mechanism to concentrate on the object detection region. CBAM (convolutional block attention module) [28] combines space and channel attention and uses maximum and average pooling to guide the neural network to extract object features more precisely. Figure 4 shows the network structure.

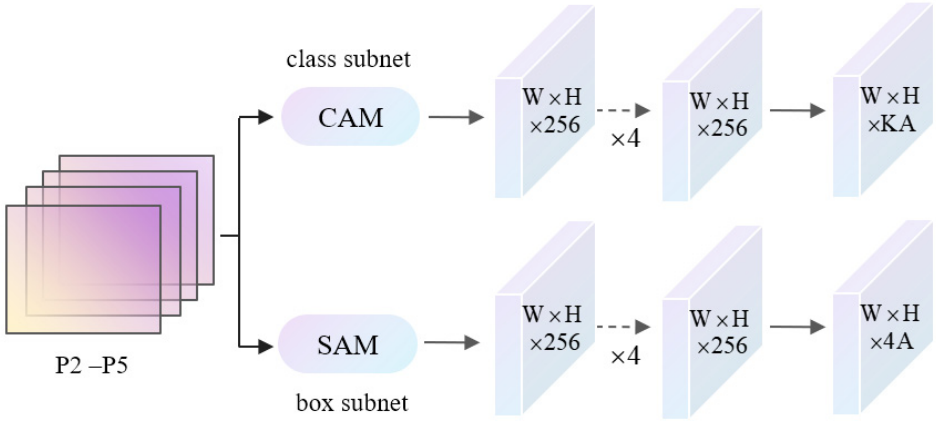


Figure 4. Subnetwork structure of feature detection.

The CAM (channel attention module) is concerned with implementing which channel features are more significant. The core notion of CAM is to construct a vector with a length equal to the number of channels in the network, where each element corresponds to the weight of each channel in the feature map. The various channels of the convolutional feature map encode various object attributes. Learning is used to continuously update the weights of each channel, informing the network which image attributes to prioritize. This module focuses on directing the network's attention to the image's foreground, hence increasing the network's focus on significant features. It can improve the collection of valuable information about objects in low-illumination and make feature extraction for classification tasks more effective. Figure 5 depicts its structure. The input feature maps are pooled using maximum and average pooling to aggregate the spatial information of the feature map. The output features are added to the main elements using the MLP (multilayer perceptron) sharing network, and the weight coefficients between them are obtained using sigmoid function scaling. To obtain the final feature map, the weight coefficients are multiplied by the input feature map.

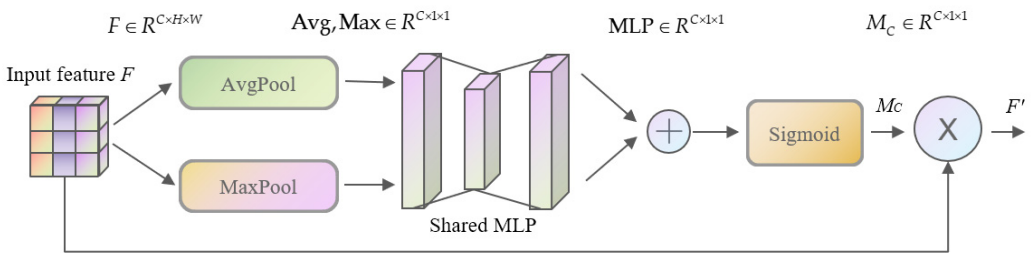


Figure 5. Structure of the channel attention module.

The channel attention module can be expressed as:

$$\mathrm{M_c}(F) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))) \quad (1)$$

$$F' = \mathrm{M_C}(F) \otimes F \quad (2)$$

where F denotes the input feature map, F' presents the output of this module, M_C indicates the channel attention operation, AvgPool means the use of average pooling for input features, MaxPool denotes the use of maximum pooling for input features, MLP denotes multi-layer perceptron operation, σ which denotes sigmoid function, and symbol $\otimes$ denotes element by element multiplication.

The SAM (spatial attention module) focuses on which parts of the space features are significant and can determine which areas of the image's information warrant attention. The SAM generates a mask of the same size as the original feature map, and the value of each element in the mask represents the pixel weight at the corresponding location in the feature map. After learning, the individual weights are regularly modified, which informs the network of the places that require attention. The primary objective of the module is to improve the target localization effect and emphasize the target scoring weights that must be localized in space. It can enhance the acquisition of ship target location information, making it more applicable to location extraction regression tasks. Figure 6 depicts its structure. To obtain two feature maps, the input feature map ($C \times H \times W$) is pooled to the maximum and average of a channel dimension ($1 \times H \times W$). The two feature maps are spliced, and the dimension is reduced to one feature map via a convolution layer ($1 \times H \times W$). The spatial weight coefficients are then generated using the sigmoid function, and the final feature map is obtained by multiplying the input features of the module with the spatial weight coefficients.

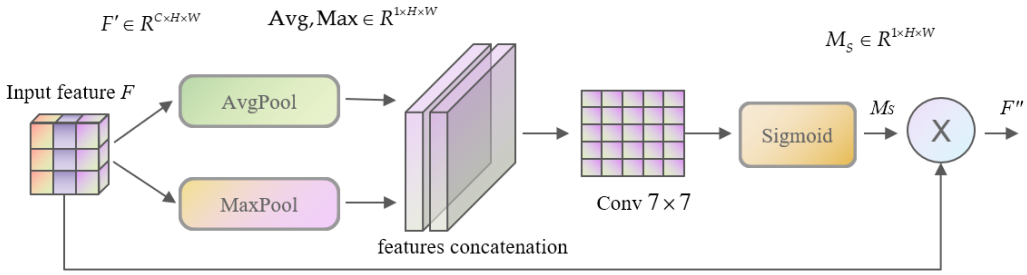


Figure 6. Structure of the spatial attention model.

The spatial attention module can be expressed as:

$$M_s(F) = \sigma(f^{7\times7}([\text{AvgPool}(F); \text{MaxPool}(F)])) \tag{3}$$

$$F'' = M_s(F) \otimes F \tag{4}$$

where F denotes the input feature map, F'' means the output of this module, M_S which indicates the channel attention operation, AvgPool denotes the use of average pooling for input features, MaxPool denotes the use of maximum pooling for input features, $f^{7\times7}$ presents a 7×7 convolution operation, σ which denotes sigmoid operation, and symbol $\otimes$ denotes element by element multiplication.

2.4. Loss Function

The loss function used in this paper's algorithm is divided into position loss and classification loss. The position loss function is as follows:

$$L_{\text{loc}}(x, l, g) = \sum_{i\in\text{Pos}}^{N} \sum_{m\in\{cx,cy,w,h\}} x_{ij}^{k} \text{smooth}_{\text{L1}}(l_i^m - \hat{g}_j^m) \tag{5}$$

where, x_{ij}^k represents the intersection and merge ratio (IoU) between the j-th real frame of the i-th bounding box of class k. l_i^m and $\hat{g}_j^m$ represent the four location parameters of the

bounding box and the four parameters of the real box, respectively. L1 loss is shown in Equation (6).

$$\text{smooth}_{\text{L1}}(x) = \begin{cases} 0.5x^2, |x| < 1 \\ |x| - 0.5, \text{others} \end{cases} \tag{6}$$

Considering the imbalance problem of positive and negative samples, RetinaNet adopts focal loss as the classification loss. Focal loss is improved in the cross-entropy loss function (CE loss), and the CE loss is shown in Equation (7).

$$\text{CE}(y) = -\frac{1}{n}\sum_{t=1}^{n}\left[\text{I}\{y = y'_t\}\log(p_t)\right] \tag{7}$$

where n represents the total number of bounding boxes, and y'_t is the correct category corresponding to the t-th bounding box. p_t is the predicted category of the t-th bounding box. I is the symbolic function, and the judgment condition is in curly brackets.

The optimized cross-entropy loss function focal loss expression is shown in Equation (8).

$$\text{FL}(p_t) = -\alpha(1 - p_t)^{\gamma}\ln(p_t) \tag{8}$$

where the expression for p_t is shown in Equation (9).

$$p_t = \begin{cases} p, y = 1 \\ 1 - p, others \end{cases} \tag{9}$$

α is the weighting factor, which is used to control the weight of positive and negative samples on the total loss. Its smaller value can reduce the weight of negative samples, generally taken as 0.25. γ is the modulation coefficient, the purpose of which is to reduce the weight of the easily classified samples, so that the model training can focus more on the hard-to-classify samples, generally taken as 2. From the formula, we can see that when the positive and negative samples are not uniform, the loss caused by negative samples decreases significantly. Additionally, when the samples are homogeneous, the loss is only appropriately reduced, thus attenuating the classification error caused by uneven samples.

3. Experimental Verification and Result Analysis

3.1. Evaluation Index of Detection Performance

The most commonly used indicator to measure detection accuracy in the field of object detection is IoU (intersection over union). Furthermore, IoU can be regarded as the ratio of the area of the intersection area of bounding box D and labeled bounding box G to the area of the union area. It is shown in Equation (10).

$$\text{IoU}(D, G) = \frac{|D \cap G|}{|D \cup G|} \tag{10}$$

By setting a threshold value for the cross-merge ratio, detections above this threshold are considered *TP* (true-positive). Detection results below this threshold are considered as *FP* (false-positive). The combined labeled enclosing frame data are *FN* (false-negative). The accuracy rate P and recall rate R of the model can be calculated in Equations (11) and (12), respectively.

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

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

By plotting the precision-recall curve, the area enclosed by this curve and the two axes of precision and recall is calculated. That is, AP (average precision) is the area under the PR curve, and its expression is shown in Equation (13).

$$AP = \int_0^1 P(R)dR \tag{13}$$

If the dataset contains multiple categories, mAP is the average of the AP values of the different object categories. AP and mAP avoid the impact of the unequal confidence in different models on the evaluation. Thus, these two methods are suitable for most models in the field of object detection.

3.2. Dataset Construction

High-quality datasets are essential prerequisites for training excellent performance models in the object detection algorithm based on deep learning. There are currently two types of ship object images [17,29–33]. The first is a monitoring image obtained from cameras installed on land and the ship. The majority of monitoring images come from VIS (visible light) or infrared sensors. The second type of image is remote sensing imagery obtained from satellite sensors, which are typically collected by radar. SAR (synthetic aperture radar) is a popular remote sensing technology. Table 2 summarizes some publicly available ship object detection datasets.

Table 2. Public datasets for ship object detection.

Dataset	Data Type	Sensor	Label Categories	Data Amount
MarDCT	video/land-based	VIS/IR	DCT	12
SMD	video/mixed	VIS/NI	DT/7 + horizon	36/12,604
SeaShips	image/land-based	VIS	D/6	168/31,455
Buoy	video/buoy	VIS	horizon	10/998
MODD2	video/USV	VIS	D/2 *	28/11,675
SEAGULL	video/UAV	VIS/IR/NI	DT/5	19/151,753

Note: In the data type, USV means unmanned boat, UAV means unmanned aircraft. In sensors, VIS denotes visible light, IR denotes infrared sensor, NI denotes near-infrared sensor. In the label category, D denotes object detection, C denotes image classification, T denotes object tracking. The number after the slash indicates the number of vessel categories. * indicates that it is divided into two categories: large obstacle and small obstacle. The number before/after the slash in the data volume indicates the number of videos/frames.

This work focused on the detection of ship objects in low-light environments. There are few images of ship data available, and the publicly available ship dataset cannot be used. Thus, using a shore-based camera in the Nanjing section of the Yangtze River Channel, the experiment acquired a large number of ship image materials, including 1258 images, and generated a ship image dataset in low illumination conditions. The dataset was then annotated with "LabelMe" annotation software and converted to MS COCO (Microsoft common objects in context) dataset format. To guarantee the validity of the experimental results, the dataset was separated into training and test sets in a 7:3 ratio, and the images were dispersed equitably according to the object scale and density. Figure 7 illustrates a portion of the dataset's ship object photos.

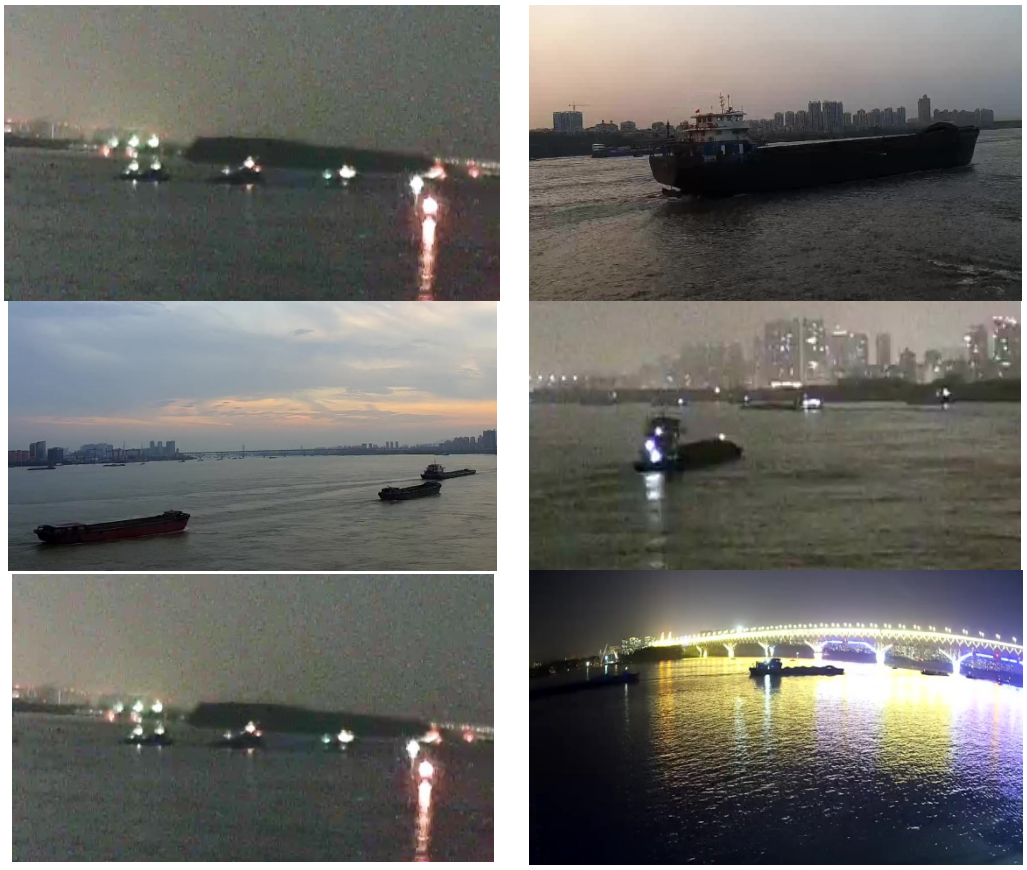


Figure 7. Partial images of ship object in a low-illumination environment.

3.3. Image Multi-Random Augmentation

The dataset presented in this research was collected primarily for ship object detection in low-light situations. The overall number of photos is very modest, which can be increased through image multi-random augmentation. This paper increased the expressiveness of the dataset and the model's robustness by randomly dithering the image's brightness, contrast, saturation, and hue, as well as translating, rotating, cropping, and zooming, to simulate complex environment changes, such as the ship sinking and floating, and uneven light brightness.

The luminance dithering of this method was achieved by RGB (red, green, and blue) images with random.uniform (−32,32). Contrast dithering was achieved by RGB image *random.uniform (0.5,1.5). Saturation dithering was achieved by the S channel of the HSV (hue, saturation, and value) image *random.uniform (0.5,1.5). The hue dithering is achieved by the H channel of HSV image with random.randint (−18,18). This method was designed to include three layers of random meanings for the multiple random increments of the image. From the two fixed routes designed, one is chosen randomly with 50% probability. Each dithering operation was executed with 50% probability. Additionally, the parameters in each dithering operation are randomly generated. In short, the image multi-random authentication architecture is shown in Figure 8.

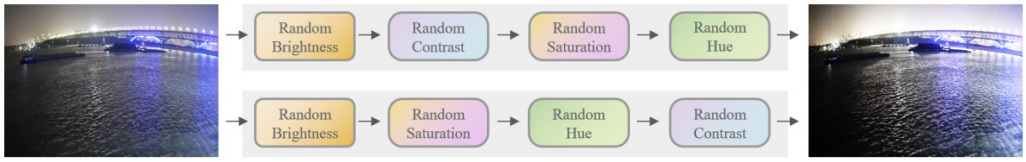


Figure 8. Image multi-random augmentation architecture.

3.4. The Design of Prior Bounding Box

The ship annotation data show that the ship object is generally slender, and the width of the bounding box is significantly greater than the height scale. The original prior bounding box size of RetinaNet is primarily used for general object detection in nature, and is not fully applicable to the unique needs of ship object detection. According to NIE [26] and LIU [34], the design size of YOLOv3's prior bounding box is improved, which improves the detection algorithm's mAP value.

The K-means++clustering algorithm was used in this paper to cluster ship object tags in the training set, and a preset box suitable for the ship object detection task was set based on the clustering results, with the aspect ratio of this algorithm's preset box set as [1.0,2.0,3.0]. The method of clustering real box dimensions can bring the anchor size closer to the true value, divide the space of scale and aspect ratio into several corresponding subspaces, and allow the model to better fit the real position of the object, lowering the training difficulty and making the model easier to learn.

3.5. Analysis of Experimental Results

The experiment was conducted on an Ubuntu 16.04 system with the Pytorch 1.9 framework, and CUDA 10.2 and cuDNN 7.6 were used to accelerate the training. The threshold value of the prediction probability and NMS were both set to 0.5. In the SGD (stochastic gradient descent) optimizer, the batch size was set to 16 and the initial learning rate was set to 0.002. There were a total of 50,000 iterations. Total_loss basically converges after 50,000 training sessions. Figure 9 depicts the entire training procedure.

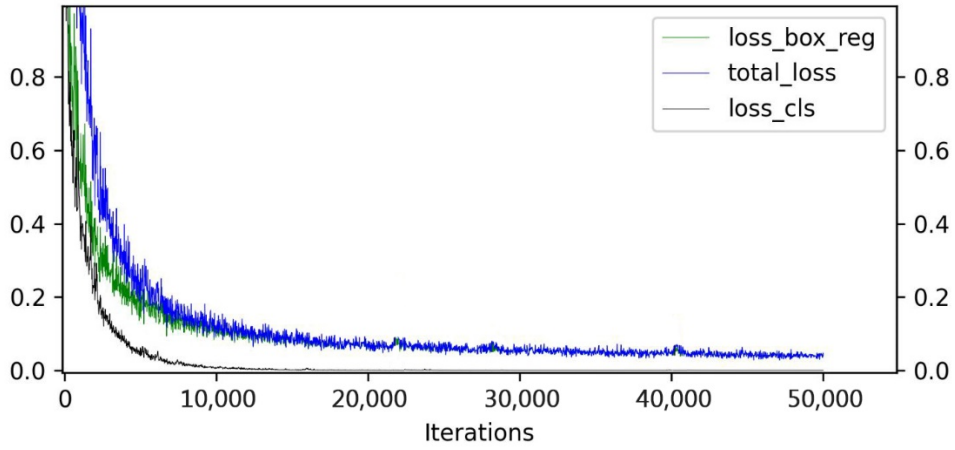


Figure 9. The process of model training on training set.

The improved RetinaMFANet was compared to the traditional RetinaNet, Faster R-CNN, SSD-improved [21], YOLOv3-enhanced [26] algorithms. The algorithm performance indicators were AP and FPS. Table 3 displays the comparison results. According to the test results, the standard two-stage detection algorithm Faster R-CNN had a higher detection accuracy than the one-stage detection method RetinaNet, despite having a slower detection speed. Because SSD-improved is oriented to general targets and YOLOV3-enhanced is oriented to ship targets for the object detection algorithm in low-illumination environments, the detection accuracy of YOLOV3-enhanced is slightly higher. Due to the incorporation of a multiscale fusion network into the original RetinaNet, the improved detection technique has raised the model's computational cost. YOLOv3-enhanced achieved the fastest computing speed due to the use of DarkNet53. Faster R-CNN is the two-stage detection algorithm that has the slowest detection speed. The SSD-improved algorithm is also slower due to the addition of a differential modal perception fusion module. The RetinaMFANet model outperforms the other four algorithms in terms of detection accuracy, achieving a 3.7% improvement over the initial detection accuracy, and this approach was only 2.07 FPS

slower than the fundamental algorithm. In general, the method suggested in this research is superior for ship object detection in situations with low illumination.

Table 3. Algorithm performance comparison table.

Method	Backbone	*AP*/%	FPS
Faster R-CNN	ResNet50	77.8	19.23
RetinaNet	ResNet50	75.2	25.53
SSD-improved	ResNet50	76.9	21.88
YOLOv3-enhanced	Darknet53	78.0	26.13
RetinaMFANet	ResNet50	78.9	23.46

The multiscale feature fusion network and attention feature detection network, as well as the training process's image multi-random augmentation and previous bounding box design modules, were upgraded as part of this paper's work to improve RetinaMFANet, which is based on RetinaNet. To more accurately assess each module's contribution to the enhanced algorithm, ablation experiments were conducted on each module in this research. The findings are displayed in Table 4.

Table 4. Ablation experimental results of each module of improved algorithm.

Multiscale Feature Fusion Network	Attention Feature Detection Network	Image Multi-Random Augmentation	Priori Anchor Design	*AP*/%
×	×	×	×	75.2
√	×	×	×	77.5
√	√	×	×	77.7
√	√	√	×	78.4
√	√	√	√	78.9

Table 4 shows that the four critical modules improve the detection accuracy of the model, with the multi-scale feature fusion network contributing the most to the algorithm's detection accuracy improvement, while RetinaNet employs the conventional FPN module. This is because the fusion network makes extensive use of the underlying feature information and fuses the deep and shallow image features more efficiently. Additionally, the attention method can assist the model in focusing on the object outline, and the attention feature detection network allows the model to concentrate more on the important aspects of the image and increases object detection accuracy. We also verified the effects of parallel and serial usage of CAM and SAM in low illumination are comparable, while the effects of parallel (CAM for Class and SAM for Box) are marginally superior to those of serial use. The model's ability for generalization is enhanced by the image multi-random augmentation. The model's detection accuracy was enhanced by the building of a prior bounding box, which enables the model to explicitly look for objects that fit the width-to-height ratio of the ship. Finally, these four modules are utilized comprehensively. When compared to the traditional RetinaNet, the RetinaMFANet method improved the *AP* by 3.7% and validated the effectiveness of each module's integration.

Finally, visual comparisons between the RetinaMFANet's detection results and those from the original RetinaNet, SSD-improved, and YOLOv3-enhanced methods were made. The RetinaMFANet algorithm performed better than the other three algorithms in terms of accuracy and missed detection rate when there are many interference elements and small objects present. The detection outcomes for the identical set of photos are shown in Figure 10. They are RetinaNet, SSD-improved, YOLOv3-enhanced, and RetinaMFANet, in that order. RetinaMFANet, which is suitable for ship target detection in various complex scenarios, can also accurately detect the ship in situations where there is a high likelihood of ship target overlap and when the ship target is very similar to the navigation background.

Figure 10. Visual comparison of detection results.

The detection results of Figure 10 demonstrate that the RetinaMFANet algorithm can accurately locate and identify ship targets, in contrast to the other three algorithms' serious miss detection, multiple detections, and ship misidentification issues. This indicates that our algorithm has better precision and robustness for ship target detection in low-illumination environments.

Furthermore, the RetinaMFANet algorithm has been evaluated in a high-visibility environment, and Figure 11 shows the results of the detection of ship objects, particularly small objects. RetinaMFANet uses a multiscale feature fusion network, which helps it detect the characteristics of small targets more effectively. As a result, it outperforms the RetinaNet method in terms of missed and false detections for small ships.

Figure 11. Visual detection result.

4. Conclusions

We have completed a lot of improvement work in order to pursue higher *AP*. Compared with general moving objects, the speed of ship movement is relatively slow. Due to the practical application effect, the precision and recall of ship target detection are more important for ship safety supervision. Meanwhile, our algorithm can also meet the requirements of real-time detection. The improved RetinaMFANet method primarily addresses the issues of ship object missing detection and false detection caused by the diversity of ship objective scale, the lack of image available information, and the variability of ship objective morphology in low-illumination environments. In order to extract more potential feature information from low-illumination photos, this method first suggests a new multiscale feature fusion network structure for the feature extraction module based on the RetinaNet model's remarkable performance in general object detection. Moreover, in accordance with the feature detection network, the anchor box regression and classification detection network is improved by employing the attention mechanism, so that the network structure can be guided to extract object features. Additionally, it is worth mentioning that the design and optimization of image multi-random augmentation and prior bounding box in the training process are also carried out. Finally, the experiments are used to validate the improved

model's effectiveness on the dataset of low-light environments. The results demonstrated that, in comparison to the standard RetinaNet technique, the enhanced RetinaMFANet approach enhances ship recognition accuracy and has a superior application effect under the restriction of FPS drop.

The next step will concentrate on the demand for real-time detection at the terminal, improving object detection speed and accuracy in low-light environments with increased background light interference, and achieving the goal of covering multiple environmental scenes and multiple ship objects dynamically.

Author Contributions: Conceptualization, R.M.; methodology, R.M.; software, R.M.; validation, R.M.; formal analysis, K.B.; investigation, K.B.; resources, R.M.; data curation, R.M.; writing—original draft preparation, R.M.; writing—review and editing, K.B.; visualization, R.M.; supervision, Y.Y.; project administration, Y.Y.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by [Fundamental Research Funds for the Guangxi science and technology agency] grant number [No. 2021AB07045 and No. 2021AB05087]. Additionally, The APC was funded by [Basic Research Fund of Central-Level Nonprofit Scientific Research Institutes, No. TKS20210301].

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed or generated in this study are available from the corresponding author upon reasonable request.

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

References

1. Fefilatyev, S.; Goldgof, D.; Shreve, M.; Lembke, C. Detection and tracking of ships in open sea with rapidly moving buoy-mounted camera system. *Ocean Eng.* **2012**, *54*, 1–12. [CrossRef]
2. Chen, Z.H.; Yang, J.X.; Kang, Z. Moving ship detection algorithm based on gaussian mixture model. In Proceedings of the 3rd International Conference on Modelling, Simulation and Applied Mathematics (MSAM 2018), Shanghai, China, 22–23 July 2018; Atlantis Press: Paris, France, 2018; pp. 197–201. [CrossRef]
3. Prasad, D.K.; Rajan, D.; Rachmawati, L.; Rajabally, E.; Quek, C. Video processing from electro-optical sensors for object detection and tracking in a maritime environment: A survey. *IEEE Trans. Intell. Transp. Syst.* **2017**, *18*, 1993–2016. [CrossRef]
4. Ma, R.; Yin, Y.; Li, Z.; Chen, J.; Bao, K. Research on Active Intelligent Perception Technology of Vessel Situation Based on Multisensor Fusion. *Math. Probl. Eng.* **2020**, *2020*, 9146727. [CrossRef]
5. Ye, C.; Lu, T.; Xiao, Y. Maritime surveillance videos based ships detection algorithms: A survey. *J. Image Graph.* **2022**, *27*, 2078–2093.
6. Ma, R.; Yin, Y.; Bao, K. Ship Detection Based on LiDAR and Visual Information Fusion. In Proceedings of the 2022 Conference on Lasers and Electro-Optics (CLEO) IEEE, San Jose, CA, USA, 15–20 May 2022.
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.B. 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.Q.; He, K.M.; Girshick, R.B.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 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 European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
11. Redmon, J.; Divvala, S.; Girshick, R.B.; 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. 1063–1069.
12. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Los Alamitos, CA, USA, 21–26 July 2017; pp. 6517–6525.
13. Redmon, J.; Farhadi, A. YOLOv3: An incremental improvement. *arXiv* **2018**.
14. Tian, Z.; Shen, C.; Chen, H. FCOS: Fully Convolutional One-Stage Object Detection. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019.

15. Lin, T.Y.; Goyal, P.; Girshick, R.B.; 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.
16. Zou, J.J.; Yuan, W.; Yu, M.H. Maritime objective detection of intelligent ship based on faster R-CNN. In Proceedings of the 2019 Chinese Automation Congress (CAC), Hangzhou, China, 22–24 November 2019; pp. 4113–4117. [CrossRef]
17. Schöller, F.E.T.; Plenge-Feidenhans, L.M.K.; Stets, J.D.; Blanke, M. Assessing deep-learning methods for object detection at sea from LWIR images. *IFAC-PapersOnLine* **2019**, *52*, 64–71. [CrossRef]
18. Yue, B.; Chen, L.; Shi, H. Ship Detection in SAR Images Based on Improved RetinaNet. *J. Signal Process.* **2022**, *38*, 128–136.
19. Lv, D.; Zhang, X. Low Illumination Objective Detection Algorithm Based on Improved SSD. *Autom. Instrum.* **2022**, *37*, 53–58, 69.
20. Li, Y.-B.; Wang, Y.-L.; Yan, Y. Infrared and visible images fusion from different views based on saliency detection. *Laser Infrared* **2021**, *51*, 465–470.
21. Chen, S.; Wang, C.; Zhou, Y. A Pedestrian Detection Method Based on YOLOv5s and Image Fusion. *Electron. Opt. Control.* **2022**, *29*, 96–101, 131.
22. Chang, L.; Bai, J.; Huang, L. Multi-Objective Detection Based on Camera and Radar Feature Fusion Networks. *Trans. Beijing Inst. Technol.* **2022**, *42*, 318–323.
23. Liu, J.; Zhao, T.; Liu, M. Ship Objective Detection in SAR Image Based on RetinaNet. *J. Hunan Univ. (Nat. Sci.)* **2020**, *47*, 85–91.
24. He, K.M.; Zhang, X.Y.; Ren, S.Q.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; 770–778.
25. Lin, T.Y.; Dollar, P.; Girshick, R.B.; 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. 936–944.
26. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convo1utional b1ock attention module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
27. Bloisi, D.D.; Iocchi, L.; Pennisi, A.; Tombolini, L. ARGOS-Venice boat classification. In Proceedings of the 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Karlsruhe, Germany, 25–28 August 2015; pp. 1–6. [CrossRef]
28. Prasad, D.K.; Rajan, D.; Prasath, C.K.; Rachmawati, L.; Rajabally, E.; Quek, C. MSCM-LiFe: Multi-scale cross modal linear feature for horizon detection in maritime images. In Proceedings of the 2016 IEEE Region 10 Conference (TENCON), Singapore, Singapore, 22–25 November 2016; pp. 1366–1370. [CrossRef]
29. Shao, Z.F.; Wu, W.J.; Wang, Z.Y.; Du, W.; Li, C.Y. SeaShips: A large-scale precisely annotated dataset for ship detection. *IEEE Trans. Multimed.* **2018**, *20*, 2593–2604. [CrossRef]
30. Fefilatyev, S.; Smarodzinava, V.; Hall, L.O.; Goldgof, D.B. Horizon detection using machine learning techniques. In Proceedings of the 5th International Conference on Machine Learning and Applications (ICMLA'06), Orlando, FL, USA, 14–16 December 2006; pp. 17–21. [CrossRef]
31. Bovcon, B.; Mandeljc, R.; Perš, J.; Kristan, M. Stereo obstacle detection for unmanned surface vehicles by IMU-assisted semantic segmentation. *Robot. Auton. Syst.* **2018**, *104*, 1–13. [CrossRef]
32. Ribeiro, R.; Cruz, G.; Matos, J.; Bernardino, A. A data set for airborne maritime surveillance environments. *IEEE Trans. Circuits Syst. Video Technol.* **2019**, *29*, 2720–2732. [CrossRef]
33. Nie, X.; Liu, W.; Wu, W. Ship detection based on enhanced YOLOv3 under complex environments. *J. Comput. Appl.* **2020**, *40*, 2561–2570.
34. Liu, G.; Zheng, Y.; Zhao, M. Vehicle information detection based on improved RetinaNet. *J. Comput. Appl.* **2020**, *40*, 854–858.