




Article

Multi-Scale Object Detection Model for Autonomous Ship Navigation in Maritime Environment

Zeyuan Shao ¹, Hongguang Lyu ^{1,*}, Yong Yin ^{1,*}, Tao Cheng ², Xiaowei Gao ², Wenjun Zhang ¹, Qianfeng Jing ¹, Yanjie Zhao ^{3,4} and Lunping Zhang ^{3,4}

¹ Navigation College, Dalian Maritime University, Dalian 116026, China

² SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London (UCL), Gower Street, London WC1E 6BT, UK

³ China Ship Scientific Research Center, Wuxi 214082, China

⁴ Taihu Laboratory of Deepsea Technological Science, Wuxi 214082, China

* Correspondence: lhg@dlnu.edu.cn (H.L.); bushyiny@dlnu.edu.cn (Y.Y.)

Abstract: Accurate detection of sea-surface objects is vital for the safe navigation of autonomous ships. With the continuous development of artificial intelligence, electro-optical (EO) sensors such as video cameras are used to supplement marine radar to improve the detection of objects that produce weak radar signals and small sizes. In this study, we propose an enhanced convolutional neural network (CNN) named VarifocalNet * that improves object detection in harsh maritime environments. Specifically, the feature representation and learning ability of the VarifocalNet model are improved by using a deformable convolution module, redesigning the loss function, introducing a soft non-maximum suppression algorithm, and incorporating multi-scale prediction methods. These strategies improve the accuracy and reliability of our CNN-based detection results under complex sea conditions, such as in turbulent waves, sea fog, and water reflection. Experimental results under different maritime conditions show that our method significantly outperforms similar methods (such as SSD, YOLOv3, RetinaNet, Faster R-CNN, Cascade R-CNN) in terms of the detection accuracy and robustness for small objects. The maritime obstacle detection results were obtained under harsh imaging conditions to demonstrate the performance of our network model.

Keywords: autonomous ships; sea-surface; object detection; computer vision; convolutional neural network (CNN); VarifocalNet



Citation: Shao, Z.; Lyu, H.; Yin, Y.; Cheng, T.; Gao, X.; Zhang, W.; Jing, Q.; Zhao, Y.; Zhang, L. Multi-Scale Object Detection Model for Autonomous Ship Navigation in Maritime Environment. *J. Mar. Sci. Eng.* **2022**, *10*, 1783. <https://doi.org/10.3390/jmse10111783>

Academic Editor: Alessandro Ridolfi

Received: 20 October 2022

Accepted: 17 November 2022

Published: 19 November 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The shipping industry is gradually moving towards artificial intelligence (AI) navigation owing to the continuous development of AI technologies, communications tools, and computers [1,2]. For autonomous ships, accurate and robust sensing of sea-surface obstacles is critical for autonomous navigation. Marine radar can detect and track objects and is currently widely used for detecting objects on the sea [3,4]. However, it has various drawbacks, such as blind spots at short ranges and it is difficult to use it to detect targets that produce weak signals in cluttered environments [5]. As a good complement to marine radar, electro-optical (EO) sensors can obtain rich video and image feature information, and they are more suitable for use in image processing and computer vision [6,7].

Using EO sensors for detecting sea-surface objects in maritime conditions has several challenges. First, ocean events such as the incidents of waves and water surface reflection can affect computer vision algorithms [8,9]; second, the observation distance and angle will cause changes in the appearance of sea-surface objects [10]; thirdly, complex backgrounds comprising port buildings can affect the detection of foreground objects [11]. For sea-surface image recognition in complex maritime environments, two types of methods can be used: traditional and deep learning methods [12]. Traditional methods include horizon detection, background subtraction, and foreground segmentation. Ref. [13] conducted a

comprehensive review and evaluation of traditional methods based on sea-surface object detection and used the Singapore Maritime Dataset (SMD) to quantitatively analyze the traditional object detection methods. Ref. [14] tested 37 different background subtraction methods on the IPATCH dataset benchmark, and showed that the multi-feature method has the best subtraction performance. Ref. [15] considered the ship's shape and texture and obtained the ship's foreground object by eliminating the background of clouds, islands, and sea clutter. Arshad et al. [16] used morphological operations to process the background image of the ship, and then used the Sobel operator to detect the edge of the ship to distinguish it from the background; however, the performance of this method was not good for complex textures, and it had more noise. The complex feature engineering of traditional methods must be improved along with their real-time performance.

Deep learning methods chiefly employ different convolutional neural networks (CNN) to extract feature information from images. The representative models of CNN include LeNet [17], Inception [18–20], VGGNet [21–24], ResNet [25] and DenseNet [26]. The CNN used for object detection can be divided into two categories: (1) Two-stage methods, such as R-CNN [27], Faster R-CNN [28], and Cascade R-CNN [29]; (2) one-stage methods, including YOLOv3 [30], YOLOv4 [31], RetinaNet [32], and single shot multiBox detector (SSD) [33]. Among them, the two-stage detector represented by R-CNN has the problems of high computational cost and poor real-time performance. In contrast, one-stage detectors are faster in real-time, but less accurate. Anchor-free based methods, such as Fully Convolutional One-Stage Object Detection (FCOS) [34], VarifocalNet [35], etc., can eliminate the limitations brought by traditional anchor-based detection methods, they have recently achieved encouraging results in object-dense and complex scenes, and they ensure a good balance between detection efficiency and accuracy.

Ref. [36] first used CNN combined with horizon features for image-based sea-surface object detection tasks. Through experiments on the SeaShips dataset, it was found that the detection accuracy of large objects such as general cargo ships and container ships is high, while the recognition accuracy of small fishing boats is low. Liu et al. [37] used the YOLOv4 algorithm combined with reverse depthwise separable convolutional (RDSC) to detect objects such as ships and buoys on the sea, and found that the use of the RDSC module instead of traditional convolution reduced the amount of model computation, but produced partial accuracy loss. Guo et al. [38] used rotational libra R-CNN to detect sea-surface objects, and they proposed a method of balancing the pyramid, which can effectively improve the multi-scale object detection efficiency at sea. The feature pyramid networks (FPN) proposed in ref. [39] enhances the semantic information of the feature map by transferring deep semantic information from top to bottom to the underlying feature map. This method improved detection for variable object shapes and large-scale changes but fails to improve the detection accuracy small objects. Ghahremani et al. [40] used the cascade CNN method to achieve high-accuracy detection of distant sea-surface objects, but did not consider the detection in dense maritime scenes. Zhang et al. [35] proposed VarifocalNet on the basis of FCOS. Their method shows great potential in the task of object detection in dense and complex scenes by ranking many candidate detections.

In summary, most algorithms in complex maritime scenes can better detect large objects, while ignoring the accuracy of small-scale objects. In the field of computer vision, small objects are often difficult to detect. On the public dataset Microsoft COCO, the detection ability of VarifocalNet for small objects is poor, and the detection accuracy of small objects is much lower than that of medium and large objects. The detection accuracy of small, medium, and large objects is 26.7%, 47.3%, and 54.3%, respectively. [35]. For the sea-surface image dataset, Iancu et al. [41] proposed that the pixel size of sea-surface objects will affect the detection accuracy, conducted experiments on the ABOships dataset, and found that the detection accuracy will decrease as the object pixels decrease. Navigation experiments show that the ship's response time can be effectively improved when small-scale objects are detected early and accurately, thus enhancing the navigation safety of the ship. Moreover, it is necessary to detect large-scale obstacles quickly and accurately in the

case of narrow channels and formation navigation. Thus, obstacle object detection models for autonomous ships must exhibit good multi-scale detection performance, particularly for small-scale objects.

This study aims to investigate the accuracy and robustness of an efficient CNN-based multi-scale object detection method for the detection of sea-surface objects, and proposes a model named VarifocalNet * for robust detection results for complex and changeable sea-surface images. On the basis of VarifocalNet, VarifocalNet * further improves the detection performance of multi-scale objects on the sea, especially the detection accuracy of small objects. Firstly, the introduction of deformable convolutional networks (DCN) in the network strengthens the model's adaptability to the geometric transformation of sea-surface objects [42], enhancing object feature extraction. Then, a new loss function and inference algorithm are designed, combining the distance-intersection over union (*DIoU*) loss [43] and soft non-maximum suppression (SNMS) optimization algorithms [44] as well as multiscale forecasting methods to further enhance sea-surface object detection.

In conclusion, our learning-based maritime obstacle detection method significantly differs from previous methods in the following respects:

- We use DCN to optimize the backbone network by introducing a learnable offset to describe the feature orientation of the object, so that the receptive field of the network is not limited by a fixed range to more flexibly adapt to changes in the geometry of sea-surface objects.
- The *DIoU* loss function combined with the SNMS method is proposed to judge and screen the candidate boxes in the same grid multiple times to improve the reliability of the object detection box in dense maritime scenes.
- We provide useful training networks and data augmentation tricks and filter out some useless tricks for the task of object detection task on unmanned surface vehicle (USV)-captured scenarios.
- Experimental results in different complex maritime scenarios have demonstrated our superior sea-surface object detection performance in terms of accuracy and robustness.

The rest of this paper is structured as follows: Section 2 proposes the enhanced maritime obstacle detection network model based on VarifocalNet. Section 3 verifies the detection effects of different CNN models by analyzing and discussing the experimental results. Section 4 discusses the advantages and limitations of the model proposed in this paper, and research prospects and conclusions are given in Section 5.

2. The Proposed CNN-Enhanced Maritime Obstacle Object Detection Model

2.1. Network Architecture

The deep learning method is used to recognize the objects in the maritime image, chiefly using CNN to extract the shallow information and high-level information in the image. The flowchart of our maritime obstacle detection framework is shown in Figure 1. Input images are collected using the EO sensor on the USV. Efficient detection results are conducive to intelligent maritime supervision and safe navigation of autonomous ships. In the network model we built, ResNet50 [21] is used in the backbone network part, which is widely recognized as a high-precision deep backbone network; FPN [41] is used in the feature map part to improve the multi-scale detection capability of the network. Similar to traditional VarifocalNet, the concept of *IoU*-Aware Classification Score (IACS) was introduced in the head part of the network [39], which can simultaneously represent the confidence of object presence and localization accuracy in to produce more accurate detection ratings in object detectors. In addition, the corresponding Varifocal Loss and star bounding box feature representations are used to predict and estimate IACS. It can be seen from Figure 1 that the main part of VarifocalNet consists of two sub-networks, which are used for the regression and refinement of bounding boxes and the prediction of *IoU*-aware classification scores. The first sub-network is divided into two branches. One branch takes the feature map of the FPN layer as the input, first applies the three convolution layers activated by the ReLU function to generate 256 channel feature maps, and then

convolves the feature map again to generate a 4D distance vector at each spatial position, that is, the initial bounding box. The other branch applies the star bounding box feature representation to obtain the sampling points and the distance scale factor, and then the refined bounding box can be obtained by multiplying the distance transformation factor by the distance vector. The second sub-network is used to predict IACS, and it has a similar branching structure as the first subnetwork, with each spatial location outputting a vector consisting of C categories.

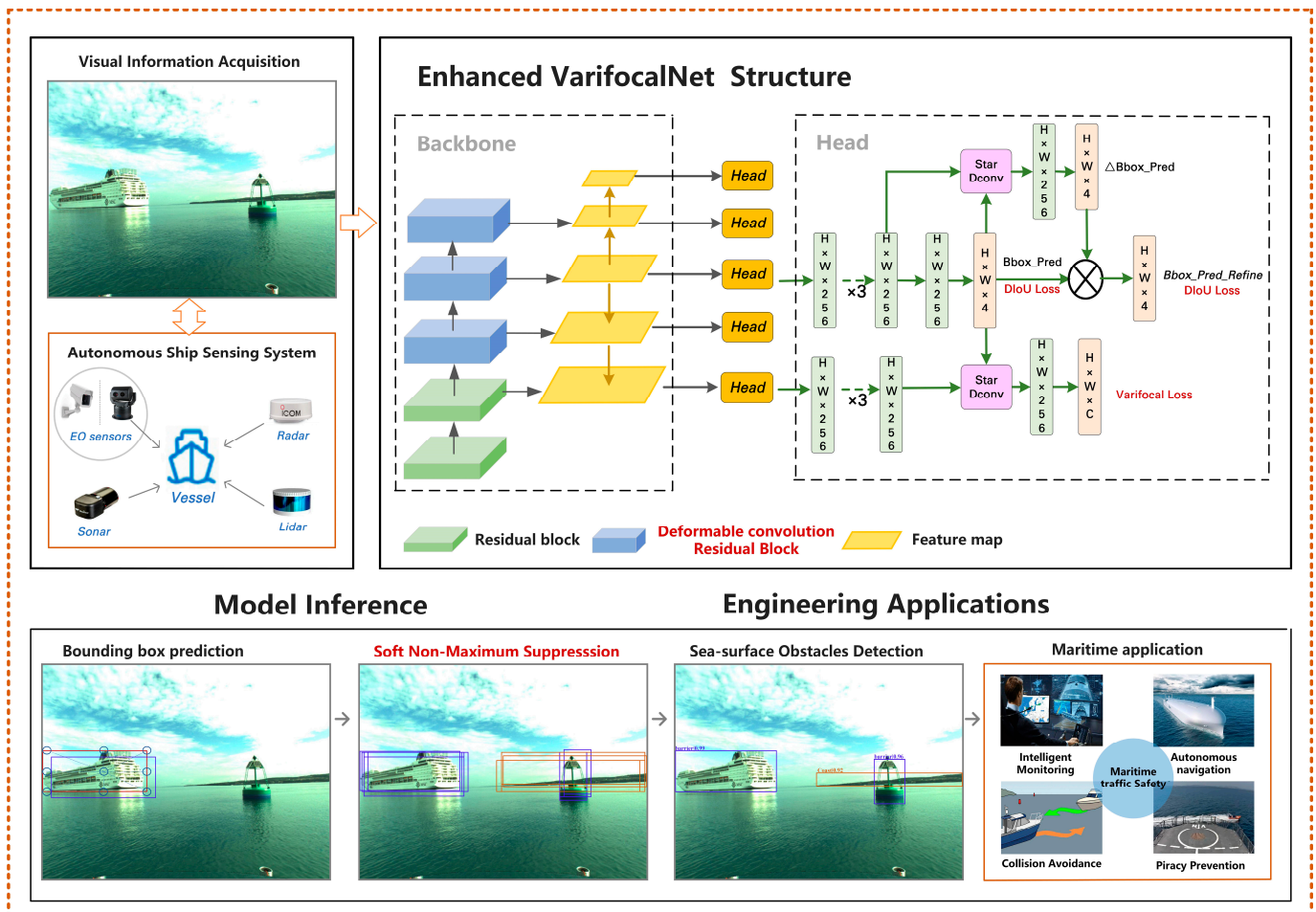


Figure 1. Flowchart of CNN-based maritime obstacle detection framework. To improve the object feature extraction ability, we use ResNet50 in the backbone part of the network, in which the last three traditional residual blocks are replaced by deformable convolution residual blocks. The feature map part uses feature pyramid network. In addition, the learning ability of the network is further improved by redesigning the loss function, introducing soft non-maximum suppression, and adopting multi-scale forecasting techniques.

In maritime images captured by EO sensors, the size of objects varies greatly; there are many small objects and the objects near the port are dense. To ensure high-quality detection results in complex maritime environments, we propose an enhanced CNN model. To be specific, we will introduce deformable convolutional module in Section 2.2 to enhance the feature extraction capability of sea-surface obstacles, and redesign the loss function in Section 2.3. In addition, the inference algorithm SNMS will be used in Section 2.4 to improve the detection accuracy of overlapping objects. In Section 2.5, we will use multi-scale techniques to further improve detection results. Benefiting from these strategies, our enhanced VarifocalNet has the capacity to efficiently detect sea-surface objects of various scales in real maritime scenes, especially the accurate detection of small objects on the sea.

2.2. Deformable Convolutional Module

For the same obstacle on the sea, there are unknown geometric transformations in different shooting angles or different scenes. Traditional convolution networks can only extract features within the matrix box, as shown in Figure 2a. The DCN [42] can break the constraint of the regular window and extract the object features in the image area more accurately, as shown in Figure 2b.

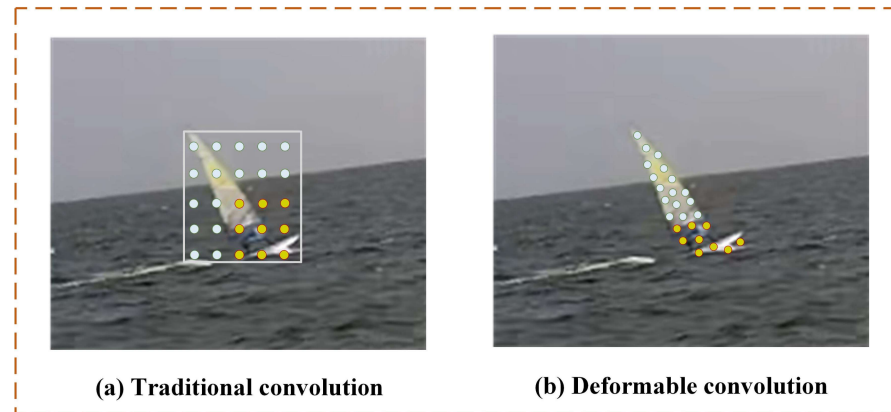


Figure 2. Schematic diagram of different convolution modules.

Taking the 3×3 convolution as an example, using $(-1, -1)$ to represent the upper left corner of the regular window and $(1, 1)$ to represent the lower right corner of the window, the regular window A can be defined as Formula (1).

$$A = \{(-1, -1), (-1, 0) \dots, (0, 1), (1, 1)\} \tag{1}$$

The steps of the traditional convolution are mainly divided into two steps, the first step is to use the regular window A to up-sample the input feature map x , and the second step is to weight the sampled values with w , where each output $y(p_0)$ needed to be sampled at the center position $x(p_0)$ of the regular window, as shown in Equation (2).

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n) \tag{2}$$

The deformable convolution is formulated as:

$$y(p_0) = \sum_{p_n \in R} w(p_n) \cdot x(p_0 + p_n + \Delta p_n) \tag{3}$$

Compared with the traditional convolution, the deformable convolution introduces an offset Δp_n , so that the sampling points can be diffused into a non-grid shape, thus better adapting to the geometric deformation of the target.

The backbone network ResNet50 is divided into five stages, as shown in Figure 3. The first stage is the image preprocessing, which first goes through the Conv, batch normalization, and activation function relu layer, and then gets the output in the maxpool layer. The last four stages are similar in structure, including ID modules that can be connected in series and convolution modules that cannot be connected in series. The second stage contains three modules, and the remaining three stages contain four, six, and three modules respectively.

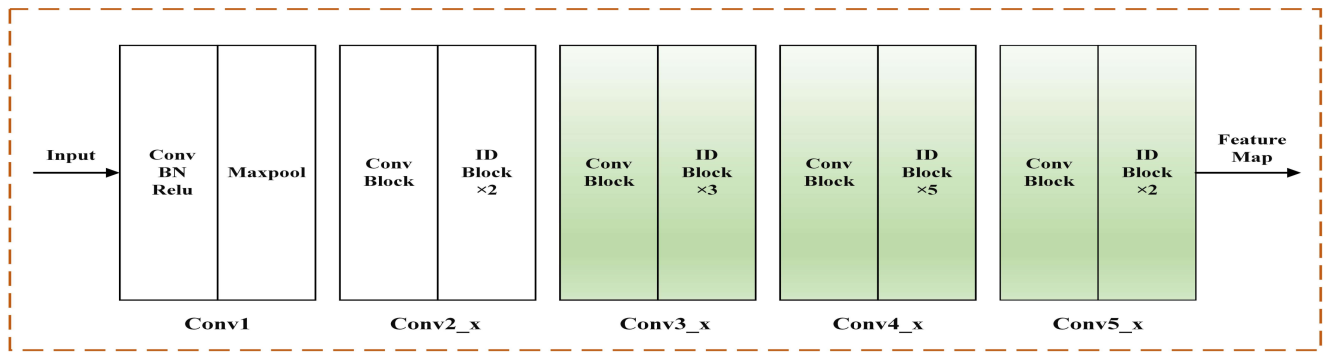


Figure 3. ResNet50 backbone network architecture.

In this paper, DCN is introduced in the last three stages of ResNet50 to better obtain the offset parameters of convolution sampling points, so as to extract features in more appropriate areas. The improved ResNet50 can better adapt to the irregular deformation of obstacles on the sea, which is conducive to the accurate detection of targets.

2.3. Loss Function

In the field of object detection, *IoU* is used as an index to evaluate the bounding box [45], that is, the intersection ratio of the predicted box and the ground truth box, as shown in Equation (4).

$$IoU = \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

The *IoU* loss can be expressed as:

$$IoULoss = -\ln(IoU) \tag{5}$$

The *IoU* loss directly takes *IoU* as the loss function, and the disadvantage is that the *IoU* value is 0 when the predicted box and the ground truth box are disjoint. The *GIoU* bounding box loss [46] used in VarifocalNet overcomes this shortcoming. For any two boxes *A* and *B*, find the smallest box *C* that can contain them, then calculate the difference $|C - (A \cup B)|$ between the areas of *C* and *A ∪ B*, and calculate the ratio of the difference to the area of box *C*. Finally, subtract the ratio from the *IoU* values of *A* and *B* to obtain *GIoU*, as shown in Formula (6).

$$GIoU = IoU - \frac{|C - (A \cup B)|}{C} \tag{6}$$

However, when the ground truth box completely contains the prediction box, *GIoU* cannot distinguish the relationship of its relative position. The *DIoU* loss [43] is used in this paper for this problem. *DIoU* makes up for the shortcomings of *GIoU* by considering the overlapping area and center distance between the target boxes. As shown in Formula (7):

$$L_{(DIoU)} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} \tag{7}$$

where *b* and *b^{gt}* represent the center points of the prediction box and the ground truth box, respectively, ρ represents the Euclidean distance, $\rho(b, b^{gt})$ represents the distance between the center points of the prediction box and the ground truth box, and *c* represents the diagonal distance of the smallest outer rectangle of the prediction box and the ground truth box as shown in Figure 4.

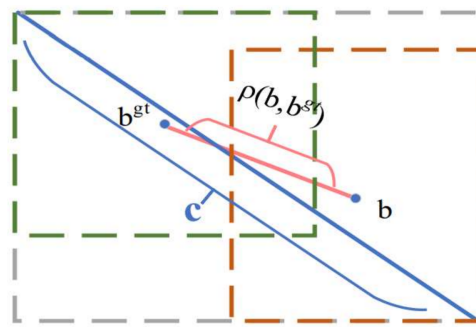


Figure 4. *DIoU* loss for bounding box regression. The green box represents the ground truth box; the orange box represents the prediction box; and the gray box is the minimum outer rectangle of both.

Compared with the *GIoU* method, the *DIoU* loss used in this paper has a faster convergence speed and can directly optimize the distance between the ground truth box and the prediction box, thereby improving the positioning accuracy of sea-surface object detection.

2.4. Inference Algorithm

For the redundant detection generated in the process of model inference, this paper uses the SNMS algorithm to further optimize the network. The SNMS algorithm is more robust than the traditional algorithm NMS [47].

When detecting a target, the network will produce several candidate boxes near the target, and each box will have a corresponding score. The core idea of NMS algorithm is to score the candidate boxes, and then select the detection box with the highest score as the final detection result, and then set the score of the detection box with low score and high overlap to 0 and remove it.

However, in the actual maritime environment, the distance between the ship and the sea-surface objects is relatively close, which may cause image overlap and false object detection. When detecting sea-surface objects, NMS will remove one of the two detection boxes with high overlap, resulting in missed detection.

To improve this problem, we use the improved SNMS algorithm [44] with Gaussian weighting, as shown in Equation (8):

$$S_i = \begin{cases} S_i & (IoU(M, b_i) < N_t) \\ 0 & (IoU(M, b_i) \geq N_t) \\ S_i e^{-\frac{IoU(M, b_i)^2}{\sigma}} & (\forall b_i \notin D) \end{cases} \quad (8)$$

In the formula, S_i is the score of the current detection box and N_t represents the threshold of *IoU*. M represents the detection box with the highest score, b_i represents the box generated at the time of detection, $IoU(M, b_i)$ represents the degree of overlap between the current detection box and the detection box with the highest score, D is the final set of detection results, and σ represents the variance of the Gaussian penalty function.

In Gaussian weighting, the closer to the center of the Gaussian distribution, the greater the penalty and the lower the weight. The SNMS algorithm avoids the problem that the score is zero through this weighted scoring method, which helps to improve the detection accuracy of the detection algorithm for overlapping objects.

2.5. Multi-Scale Technology

The image's size has a great influence on the performance of the object detection model. Because the basic part of the network usually generates a feature map which is tens of times smaller than the original map, the feature description of small objects is not easy to be captured by the detection network. The use of multi-scale technology is often one of the skills to improve the accuracy of sea-surface object detection. The technology includes feature pyramid and image pyramid. In this paper, the image pyramid method is used in

the training stage. Several images with different resolutions are sent to the network model, and one is randomly selected for training in each training generation, so as to improve the robustness of the detection model to the sea-surface images.

3. Experimental Results and Analysis

In order to compare objectively and fairly, the experiment uses a unified software environment and hardware platform. The software environment is Windows 10, python 3.8, pytorch 1.8.0, torchvision 0.9.0. The hardware platform used was an Inter (R) Core (TM) i7-10700KF CPU@3.8.0GHz processor, 32 GB of memory, and a Geforce RTX 3090 graphics card. The optimization algorithm and learning rate used in the training process are the same. The training optimization algorithm uses the stochastic gradient descent (SGD) method. The iteration of the training is 36,000, the batch size is 2, and the learning rate is 0.00125. The other parameters are set to the default values of the original VarifocalNet [35]. For more details regarding experimental results, the reader may refer to the accompanying video material at: <https://pan.baidu.com/s/17rRBQ-gI3kFsY-3wqLHcig?pwd=zeqf> (accessed on 19 October 2022).

3.1. Datasets Description and Data Augmentation

The quality of datasets and the richness of prior knowledge directly affect the effect of CNN training. Bovcon et al. [48] established the Marine Obstacle Detection Dataset v2.0 (MODD2) to solve the object detection of water surface obstacle images in complex maritime environment. The dataset was collected in the port of Koper, Slovenia. During the recording of the dataset, the USV was manually guided by experts to simulate maritime navigation scenarios, including situations where marine obstacles may cause danger to the USV. Through qualitative analysis of the characteristics of the dataset, the dataset mainly has the following characteristics: (1) strong sunlight and water reflection; (2) dense targets near the port, overlapping and shielding; (3) large differences in the size of targets on the sea; (4) many small objects near the horizon, as shown in Figure 5. These characteristics put forward higher requirements for the algorithm of computer vision, so the MODD2 is ideal for the field of marine obstacle object detection and can be used as a test benchmark for deep learning algorithms.

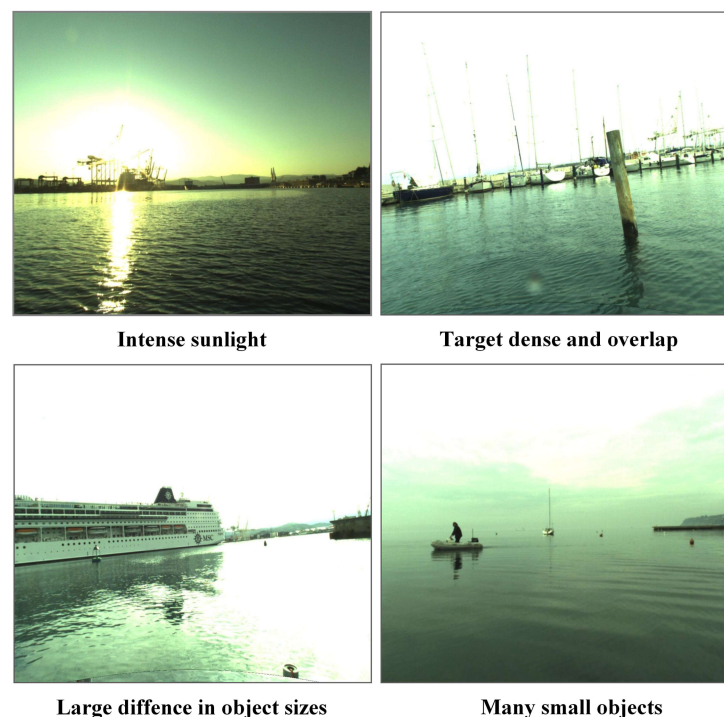


Figure 5. Characteristic analysis of MODD2 dataset. [48].

The MODD2 consists of 28 video sequences, each of which has a resolution of 1278×958 pixels, corresponding to 11,675 images by framing. The original paper used MATLAB to manually annotate each frame of the dataset and defined large and small object by whether the bounding box of the object exceeded the horizon. In the field of computer vi-sion, the Microsoft COCO challenge dataset defines the absolute size of large (pixel area $> 96^2$), medium ($32^2 < \text{pixel area} < 96^2$), and small objects (pixel area $< 32^2$). In order to better train and evaluate the mainstream deep learning network model, this paper re-labels and counts the MODD2 according to the format of COCO dataset and its definition of small, medium, and large object. A total of 6916 marine obstacle images were obtained by screening and removing the non-object images. Among them, there are 6225 training sets and 691 test sets. The statistical results of dataset information are shown in Table 1.

Table 1. Statistics of the number of large, medium, and small objects.

	Barrier	Coast	Number of Labels	Proportion (%)
Large objects	5290	6585	11,875	40.9
Medium objects	3736	1936	5672	19.6
Small objects	11,439	13	11,452	39.5
Total	20,465	8534	28,999	100.0

Among them, the labels are divided into two categories: barrier and coast. A total of 20,465 labels are barriers and 8534 labels are coast. The number of labels of each category is as shown in Figure 6. According to the scale defined by the COCO dataset, the number of large objects is 11,875, accounting for 40.9% of the total; the number of medium objects is 5672, accounting for 19.6% of the total; and the number of small objects is 11,452, accounting for 39.5%, as shown in Figure 7. In data enhancement, the main methods used are random flipping, brightness, and rotation transformation, as shown in Figure 8. In the data preprocessing stage, these lossless data enhancement techniques are used to generate more images to expand the training set, thereby improving the generalization performance of the network model.

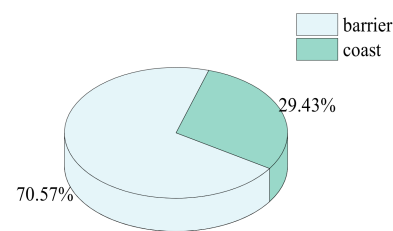


Figure 6. Proportion of various labels.

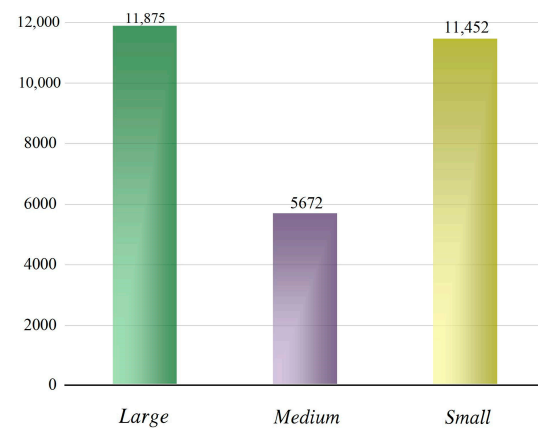


Figure 7. Statistics of the number of different scale objects.

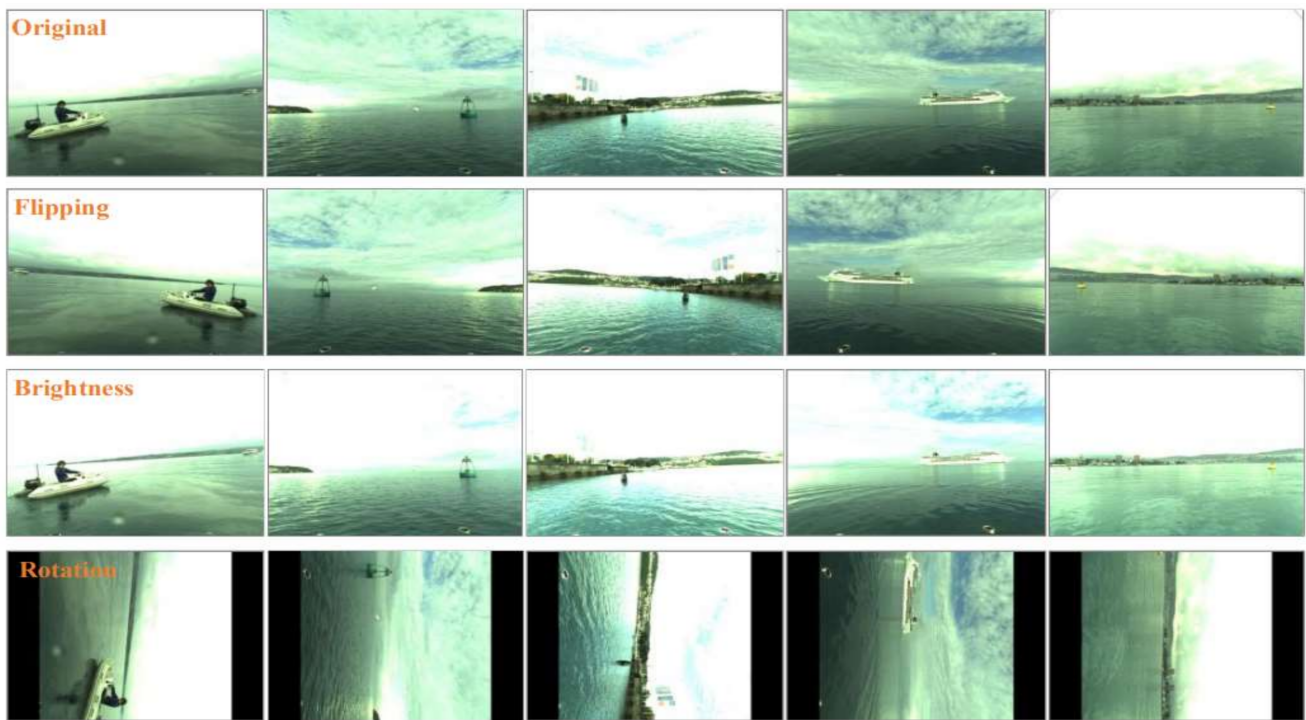


Figure 8. Data enhancement methods.

3.2. Evaluation Metrics

We evaluated the performance of each CNN model using the training loss metrics, precision metrics, and speed metrics.

The cross-entropy loss function is mainly used to evaluate the loss in the process of model training, validation, and testing, as shown in Formula (9):

$$E = -\sum_i^n t_i \log(y_i) \tag{9}$$

Precision rate P and recall rate R are used as precision indicators. Wherein the calculation formula of P can be expressed by the formula:

$$P = TP / (TP + FP) \tag{10}$$

In the formula, TP refers to the positive sample predicted by the model as a positive class; FP refers to the negative sample predicted by the model as a positive class. $P_{0.5}$ is the precision at a threshold of 0.5, and $P_{[0.5:0.95]}$ is the average precision at 10 different threshold values between 0.5 and 0.95 (step = 0.05). P_l , P_m , and P_s are the average precision of large, medium and small objects for 10 different thresholds between 0.5 and 0.95, respectively.

The recall R can be expressed by the Formula (11).

$$R = TP / (TP + FN) \tag{11}$$

FN represents the positive sample predicted by the model as a negative class. $R_{0.5}$ represents the recall when the threshold is 0.5, and $R_{[0.5:0.95]}$ represents the average recall of 10 different values between 0.5 and 0.95. R_l , R_m , R_s represent the average recall of large, medium, and small objects for 10 different thresholds between 0.5 and 0.95, respectively.

The speed of the model is measured by the number of parameters, and the frame rate. The number of parameters is the self-learning parameters of the model. The frame rate is the number of images processed by the model per second or the time it takes to process one image.

3.3. Training Loss

The loss curve of the enhanced network model is shown in Figure 9. *Iter* represents the number of training iterations and *Loss* represents the training loss. After about 36,000 network training iterations, the classification loss is 0.32, the bounding box loss is 0.13, the bounding box refinement loss is 0.1, and the total loss value is 0.6. The loss function of the network model can achieve good convergence effect.

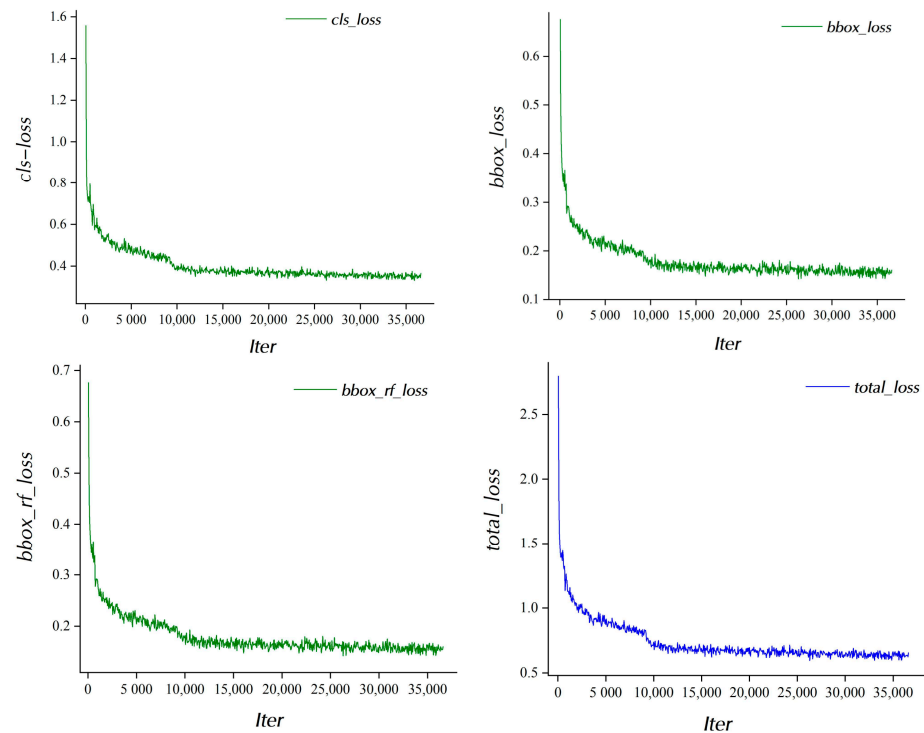


Figure 9. Loss curve of sea-surface obstacle dataset.

Our proposed network model is compared with the baseline model, as shown in Figure 10. The yellow curve is the loss value of the baseline model, and the blue curve our_loss represents the loss value of our proposed model. Observing the two curves in the figure, the improved method proposed in this study does not significantly affect the training process of the object detection algorithm.

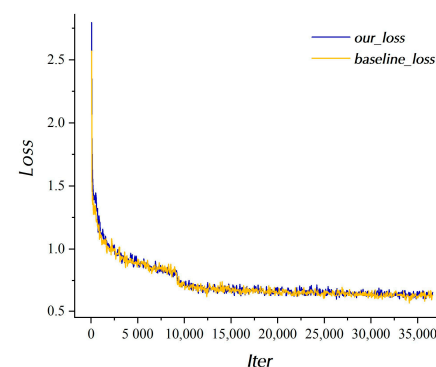


Figure 10. Comparison of losses during training.

3.4. Ablation Experiment

In the context of deep learning complex neural networks, ablation studies are often used to describe the process of removing or replacing certain modules to analyze the impact of specific modules on the network model. In this paper, based on ResNet50, FPN and

VarifocalNet, four improved methods are used to conduct ablation experiments. Table 2 shows the results.

Table 2. Combine experiment of four improved method, including soft non-maximum suppression (SNMS), deformable convolutional networks (DCN), distance-intersection over union (*DIoU*), and multi-scale (MS).

Methods	SNMS	DCN	<i>DIoU</i>	MS	$P_{0.5}$	$P_{[0.5:0.95]}$	P_s	P_m	P_l	$R_{[0.5:0.95]}$	R_s	R_m	R_l
VarifocalNet	×	×	×	×	0.979	0.771	0.525	0.738	0.893	0.763	0.576	0.809	0.919
Our methods	✓	×	×	×	0.979	0.779	0.536	0.746	0.901	0.819	0.603	0.819	0.936
	✓	✓	×	×	0.980	0.781	0.545	0.746	0.903	0.822	0.605	0.824	0.936
	✓	✓	✓	×	0.981	0.783	0.548	0.749	0.903	0.823	0.613	0.826	0.937
	✓	✓	✓	✓	0.986	0.789	0.563	0.752	0.903	0.837	0.650	0.829	0.940

Note: ✓: used; ×: not used.

If the four methods of DCN, SNMS, *DIoU* loss function, and multi-scale forecasting are added to the benchmark model, the detection effect can be improved. When all the four methods were added, the detection effect was the best. The average precision rate is 78.9% and the average recall rate is 83.7%. When the threshold is 0.5, the average precision rate is 98.6%. In addition, the average precision rate is increased by 2.3% and the average recall rate is increased by 9.7% compared with baseline. The detection effect of small objects is improved the most, and the precision rate is increased by 7.2%. The recall rate increased by 12.8%.

3.5. Comparisons with Other Detection Methods

In order to further verify the performance of the proposed network model, we compare it with SSD network, YOLOv3, Faster R-CNN, RetinaNet, Cascade R-CNN, and our method. SSD, YOLOv3, and RetinaNet are widely used one-stage object detection networks, Faster R-CNN and Cascade R-CNN are classical two-stage networks.

In the comparison experiment, the backbone network of SSD is VGG16, the backbone network of YOLOv3 is Darknet53, and the backbone networks of Faster R-CNN, RetinaNet, and Cascade R-CNN are ResNet50. The SSD network receives an input image of 300 × 300 pixels. The YOLOv3 network receives an input image of 608 × 608 pixels. The image pixels received by Faster R-CNN, RetinaNet, Cascade R-CNN, VarifocalNet are 1333 × 800. In additional, VarifocalNet * is the detection network proposed in this study. Table 3 shows the comparison results.

Table 3. Comparison of precision of each network model.

Detection Network	$P_{0.5}$	$P_{[0.5:0.95]}$	P_s	P_m	P_l	$R_{[0.5:0.95]}$	R_s	R_m	R_l
SSD	0.863	0.493	0.259	0.616	0.639	0.575	0.380	0.676	0.702
YOLOv3	0.937	0.528	0.494	0.692	0.603	0.639	0.595	0.747	0.681
Faster R-CNN	0.870	0.689	0.429	0.718	0.860	0.722	0.451	0.777	0.894
RetinaNet	0.962	0.704	0.386	0.723	0.833	0.764	0.577	0.777	0.872
Cascade R-CNN	0.910	0.742	0.431	0.741	0.891	0.768	0.463	0.792	0.921
VarifocalNet	0.979	0.771	0.525	0.738	0.893	0.763	0.576	0.809	0.919
VarifocalNet *	0.986	0.789	0.563	0.752	0.903	0.837	0.650	0.829	0.940

Compared with the other five mainstream deep learning networks, the VarifocalNet * proposed in this paper has a greater improvement in accuracy. In terms of SSD Network, the VarifocalNet * can improve the average precision rate $P_{[0.5:0.95]}$ by 60%. The improvement of P_s, P_m, P_l is 117%, 22% and 41%, and the improvement of $R_{[0.5:0.95]}, R_s, R_m,$ and R_l is 46%, 71%, 23%, and 34%, respectively. Compared with the YOLOv3, the VarifocalNet * can improve the average precision rate $P_{[0.5:0.95]}$ by 49%. The improvement of P_s, P_m, P_l

is 14%, 9%, and 50%, and the improvement of $R_{[0.5:0.95]}$, R_s , R_m , and R_l is 31%, 9%, 11%, and 38%, respectively. Compared with the Faster R-CNN, the VarifocalNet * can improve the average precision rate $P_{[0.5:0.95]}$ by 13%, in which the precision rate P_s of small objects is improved by 31%, the precision rate P_m of medium objects is improved by 5%, and the precision rate P_l of large objects is increased by 5%. The improvement of recall is more obvious, and the improvement of $R_{[0.5:0.95]}$, R_s , R_m , and R_l is 16%, 44%, 7%, and 5%. Compared with the RetinaNet model, the proposed network model can improve the overall precision rate by about 12% and the overall recall rate by 10%. Compared with Cascade R-CNN, VarifocalNet * can also show a greater advantage, with the overall precision rate $P_{[0.5:0.95]}$ increased by about 6%, and the small object increased by 31%. Figure 11 more intuitively shows the pros and cons of our model and SSD, YOLOv3, Faster R-CNN, RetinaNet, Cascade R-CNN in object detection performance at different scales. It can be seen that VarifocalNet * performs the best in the detection accuracy of large, medium, and small objects, and the detection accuracy of small objects is significantly higher than other competing methods.

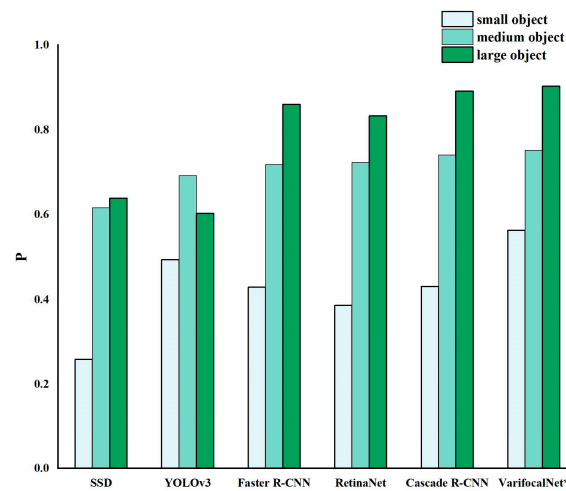


Figure 11. Comparison of different competing methods in detection accuracy of large, medium, and small objects.

In addition, Table 4 lists the impact of different modules proposed in this paper on the speed performance of the algorithm.

Table 4. Analysis of the impact of the improved module on the detection speed performance of the algorithm.

Methods	SNMS	<i>DIoU</i>	DCN	MS	Model Parameters/M	Inference Time/ms
VarifocalNet	×	×	×	×	32.49	45
Our methods	✓	×	×	×	32.49	45
	✓	✓	×	×	33.07	48
	✓	✓	✓	×	33.07	48
	✓	✓	✓	✓	33.07	91

Note: ✓: used; ×: not used.

From the analysis of Table 4, it can be concluded that the SNMS and *DIoU* modules do not increase the parameters of the model and affect the inference speed of the algorithm. The DCN module adds 0.58 M parameters to the network model, and the time to infer a single image increases by 3 ms. Therefore, the DCN module introduced in this paper has no significant effect on the real-time performance of the network model. Then, the multi-scale prediction technology achieves a satisfactory gain in accuracy, but also has an

impact on the real-time performance. The inference time for a single image increases by 43 ms compared to the baseline.

Subsequently, the detection results of obstacles in several different maritime scenarios are visually displayed in Figures 12–14. We compare the proposed VarifocalNet * method with 5 typical methods, i.e., SSD, YOLOv3, Faster R-CNN, RetinaNet, and Cascade R-CNN. Near ports, maritime obstacles tend to overlap densely, as shown in Figure 12. In Figure 13, obstacles of different scales appear at the same time, and there is strong sunlight and water surface reflection interference. In Figure 14, there are many small objects concentrated near the horizon, accompanied by dense fog.

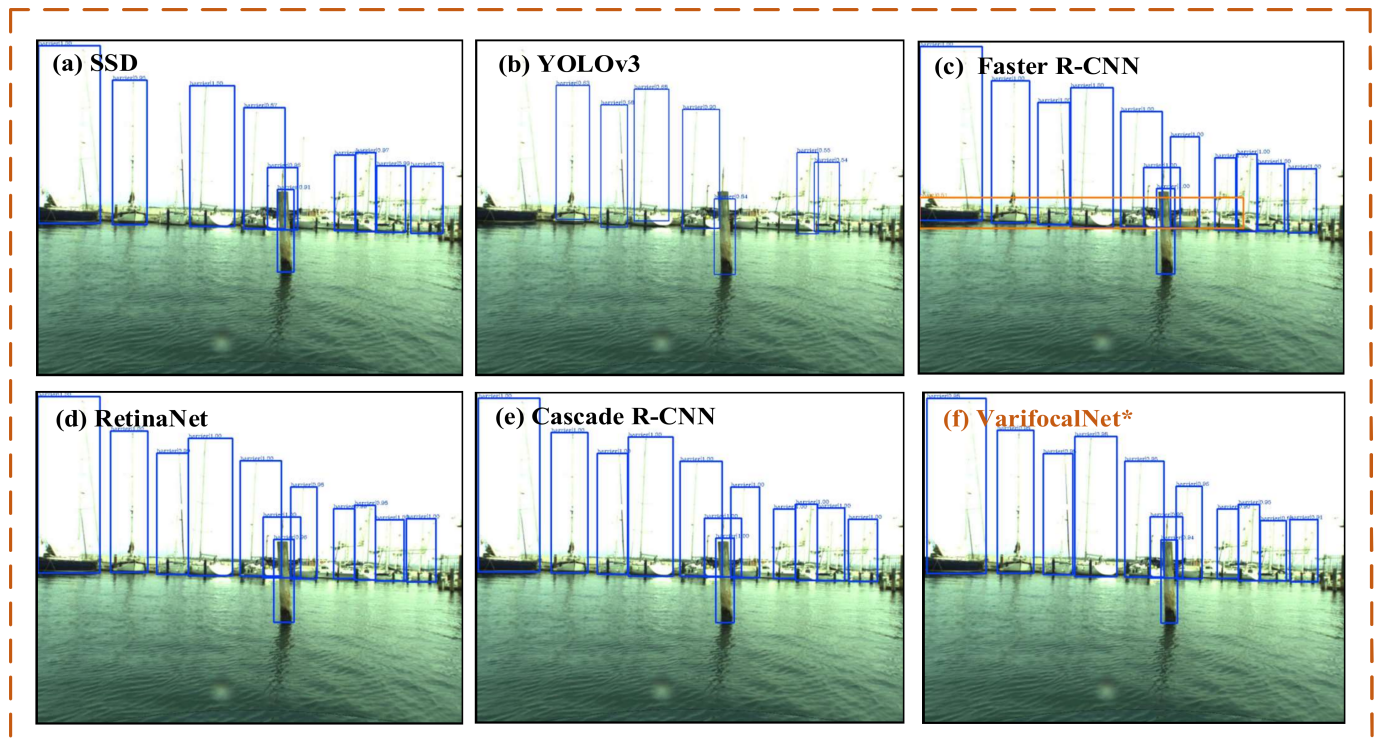


Figure 12. The visual comparisons of different competing methods for detection of sea-surface obstacles under intensive occlusion scene. All ships cannot be detected using SSD and YOLOv3 methods, while there are false detections using Faster R-CNN, as shown in (a–c). In (d–f), RetinaNet, Cascade R-CNN and VarifocalNet * methods have no missed targets.

For example, the acquired sea-surface images are often disturbed by wave splashing droplets and water surface reflections. Therefore, Faster R-CNN and RetinaNet are sensitive to unstable imaging backgrounds, resulting in erroneous detection results, as shown in Figures 12 and 13. In Figure 12, limited by the ability to extract high-level features, the detection performance of SSD and YOLOv3 for large objects is poor. In Figure 14, cascade R-CNN is lacking in the ability to detect tiny objects. In contrast, our method can efficiently and accurately detect sea-surface obstacles in different maritime scenarios, especially in the detection ability of small objects.

Finally, we use VarifocalNet * to test the generalization performance on another water surface challenge dataset [49] as shown in Figure 15. There are dense small fishing boats and floating objects in sub-image (a), there is sea fog on the water-surface in sub-image (b), sub-image (c) is a dawn scene at sea with poor visibility, and sub-image (d) is at sea night scene, light reflections present on the water surface.

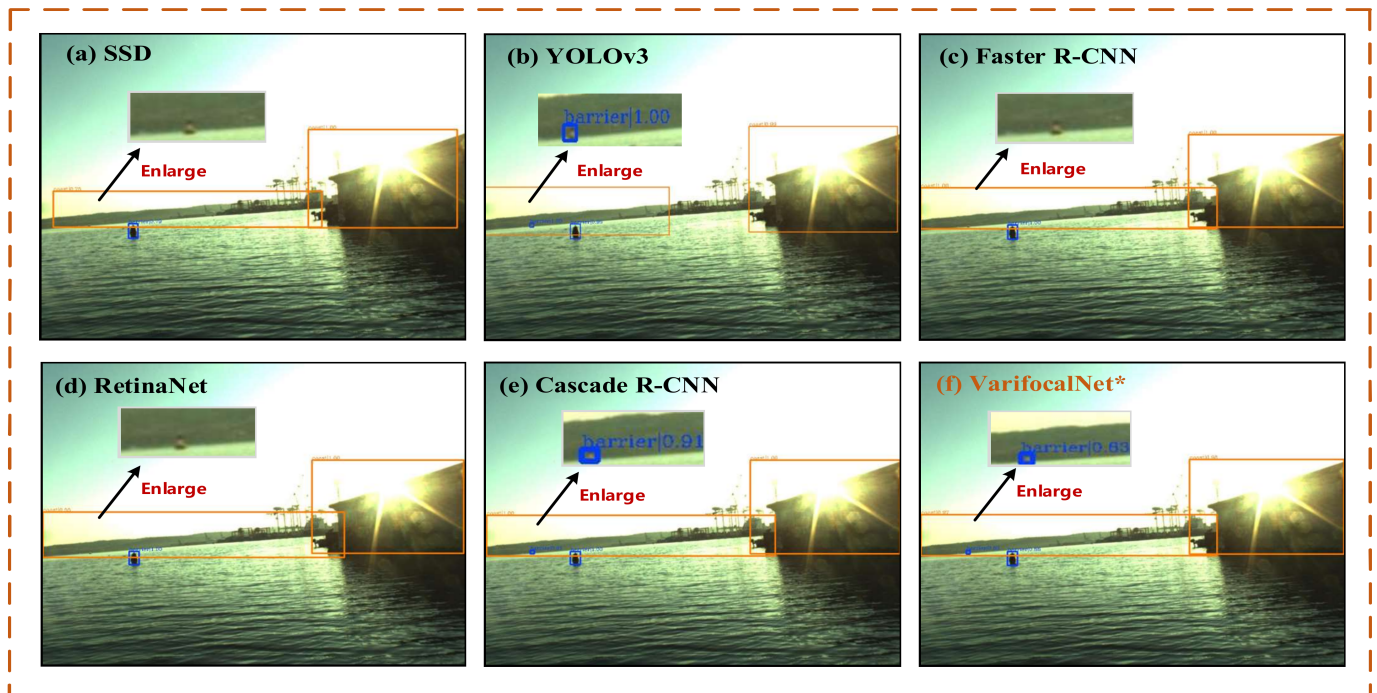


Figure 13. The visual comparisons of different competing methods for detection of sea-surface obstacles under strong sunlight interference. In (a,c,d), Small scale objects are not detected using SSD, Faster R-CNN, RetinaNet methods. In (b), incomplete coasts are detected using YOLOv3 method. Cascade R-CNN and VarifocalNet * methods can better adapt to strong light scenes at sea, as shown in (e,f).

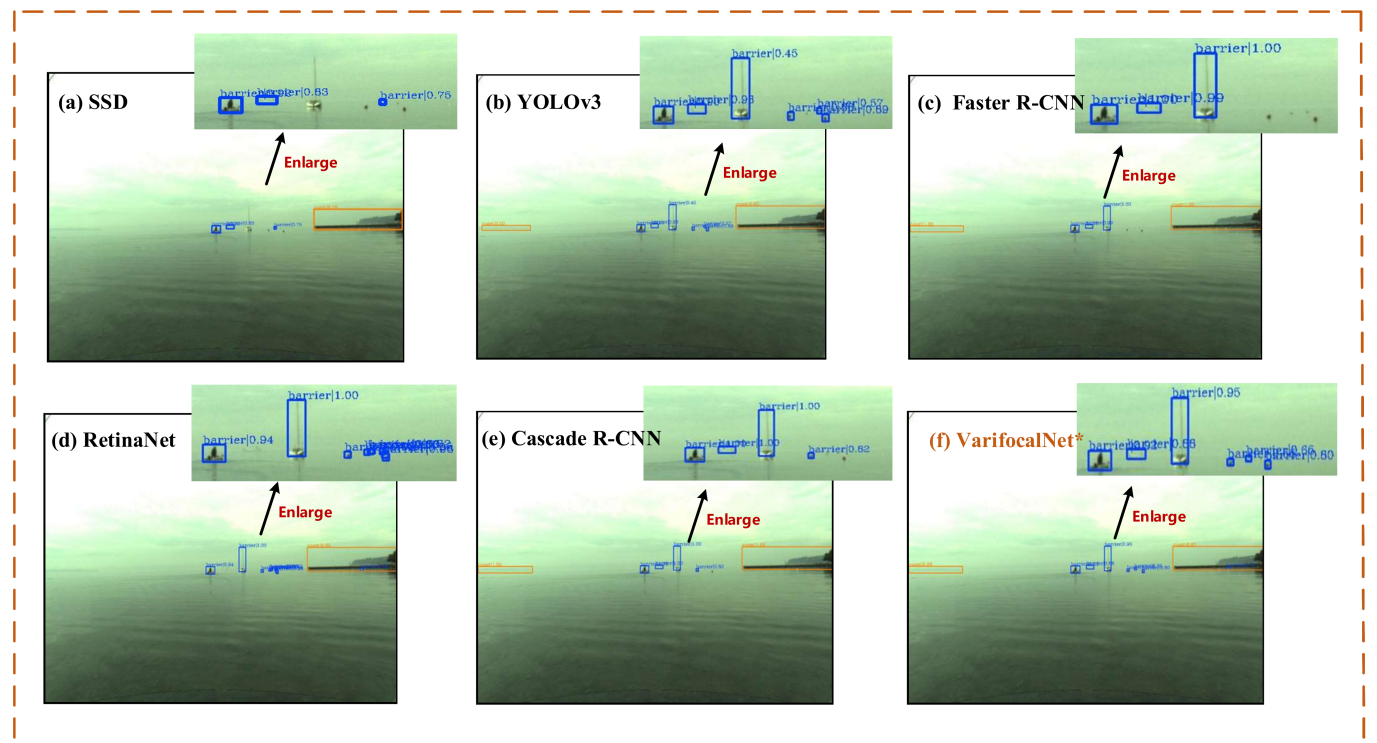


Figure 14. The visual comparisons of different competing methods for detection of sea-surface obstacles under sea fog environment. In (a–c,e), the SSD, YOLOv3, Faster R-CNN, and Cascade R-CNN have varying degrees of missed detection, and the RetinaNet generates many false detection results, as shown in (d). In contrast, our method is able to achieve more satisfactory detection performance, as shown in (f).

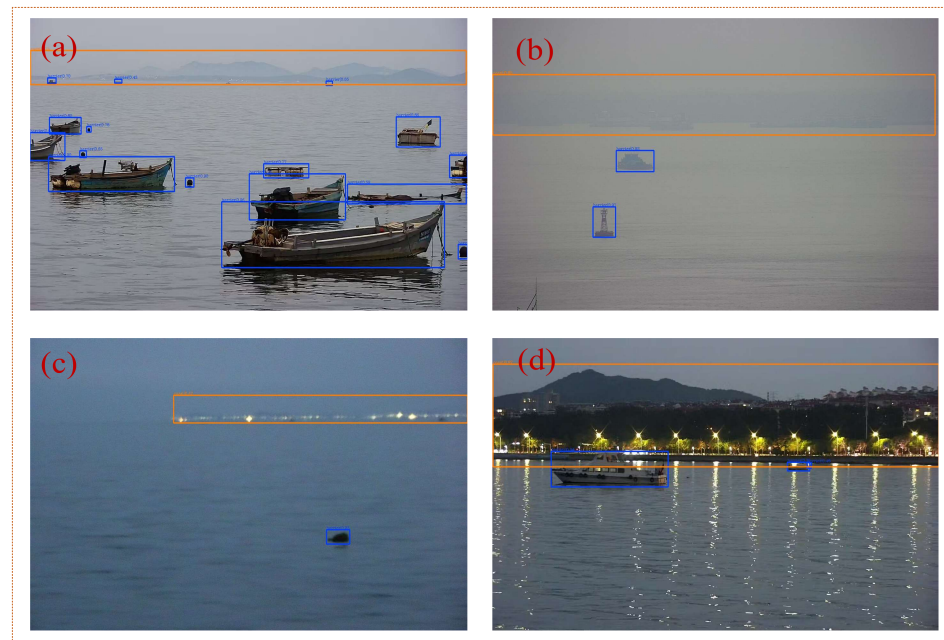


Figure 15. Detection results in real maritime scenarios. (a) Dense-target scene. (b) Sea fog scene. (c) dawn scene. (d) night scene.

By analyzing the detection results in Figure 15, Subgraph (a) shows that our algorithm has good dense and small object detection ability. In addition, in scenes with poor visibility, such as fog, dawn, night, etc. VarifocalNet * can also robustly detect sea-surface objects (as shown in Figure 15b–d).

Owing to our proposed effective enhanced methods (DCN, *DIoU*, SNMS, MS), it is able to efficiently and robustly detect obstacles on the sea even the observed visual quality has been noticeably degraded. Reliable detection of weak and small objects, e.g., wooden fishing boat, tiny buoy is a good supplement to marine radar, thereby further enhancing the navigation safety of autonomous ships.

4. Discussion

4.1. Advantages of the Model

Compared with other scenes, the maritime environment is a complex and dynamic system with incident waves, sea surface reflection, sea fog interference and so on. When an autonomous ship sails at sea, the suspension of splashing droplets caused by sea water may also lead to the failure of optical recognition. Therefore, it is very important to accumulate a large amount of data and continuously update the architecture of the iterative AI algorithm based on the real sea test site.

In view of the continuous evolution of AI algorithms, we applied a new deep learning framework based on VarifocalNet to the recognition of sea-surface objects. VarifocalNet * has the following advantages over current leading technologies:

(1) Adaption to complex maritime environment

In traditional image-based sea-surface object detection, global information such as the sky, horizon, and ocean have a critical auxiliary role in sea-surface object recognition, and most scholars rely on the characteristics of the horizon to detect sea-surface objects. Most deep learning algorithms are not designed for the maritime environment. Through careful analysis of the characteristics of the maritime environment, this study builds an object detection model suitable for this environment. Firstly, the loss function varifocal loss designed in the VarifocalNet * network alleviates the imbalance between the background and foreground classes in the training of the target detector and can better adapt to the interference of the coastal background. The SNMS algorithm used in the inference process

of the network model can further improve the detection performance of the network for overlapping targets near the port by selecting anchor boxes more suitable for the shape of sea-surface objects. Compared with the traditional convolution module, DCN can better extract the characteristics of obstacles when the geometric shape of the target changes owing to the camera shooting angle.

(2) Good performance for detecting multi-scale objects

When an autonomous ship sails at sea, the size of obstacles in the image obtained by the EO sensors vary significantly, as shown in Figure 7, mainly for large and small objects. In this study, FPN was used to continuously up-sample the feature map output from the last layer of the backbone network and to perform channel splicing with the intermediate feature map at the same scale to improve the multi-scale detection performance of the network. Table 2 shows that the average precision of large, medium, and small objects can reach 90.3%, 75.2%, and 56.3%, respectively. It can be shown that the proposed network model can achieve excellent performance in the detection of objects at various size scales.

(3) High capability for detecting small objects

Improving the detection performance for sea-surface objects in images is very important, especially the detection precision of small objects, but this remains difficult. Based on the FCOS+ATSS network, the benchmark model VarifocalNet incorporates three new components: varifocal loss, star bounding box feature representation, and bounding box refinement, and shows good robustness for detecting dense and small objects in the COCO dataset. Based on the benchmark model, DCN was used to replace the traditional convolution module, and the *GIoU* of the bounding box loss was improved. Combining this with multi-scale forecasting inspired a significant breakthrough in the detection of small sea-surface objects, and the detection precision has been significantly improved.

4.2. Limitations

The network architecture proposed in this study has better detection results for different marine scenes in the MODD2 dataset. However, current research on the mechanism of VarifocalNet * is limited and fine-tuning the CNN hyperparameters requires further research. In future studies, the accuracy may be further improved by continuously optimizing the architecture of the CNN model and completing more experiments.

The high precision performance of the VarifocalNet * network model in the test set depends on the training of a large amount of labeled data. However, based on the supervised method, the cost of manual annotation is very high and adaptability to new scenarios is not sufficient and must be improved. For this problem, methods based on weak supervised or unsupervised learning [50] are considered, such as the use of generative adversarial network (GAN) to augment the maritime dataset [51].

5. Summary and Future Work

This study proposes a high precision network model constructed using ResNet50, FPN, and VarifocalNet for multi-scale object detection for autonomous ships in maritime environment, and the network is improved according to the characteristics of the maritime environment. The improved network model has promising results for the MODD2 dataset, with an average accuracy of 78.9% and 98.6% when the *IoU* threshold is 0.5. Compared with the benchmark model based on VarifocalNet, the proposed model further improves the detection precision, especially for small-scale obstacles, for which the detection accuracy is improved by 7.2%. In terms of real-time performance, the proposed method does not significantly affect the reasoning speed and can meet engineering application requirements.

In the future, we will consider engineering applications, convert the model into TensorRT format, and apply the algorithm to autonomous ships. To ensure accuracy, the complexity of the model will be reduced to further enhance real-time performance. In addition, videos containing a series of images can be regarded as spatiotemporal data, and successive frame images usually change smoothly and exhibit high temporal dependence.

These forms of data offer a promising direction for the reduction of the false detection rate of sea-surface objects.

Author Contributions: Data curation, Z.S.; formal analysis, Z.S. and H.L.; methodology, Y.Y. and H.L.; resources, X.G. and Q.J.; supervision, T.C.; writing—original draft preparation, Z.S.; funding acquisition, W.Z., H.L. and Q.J.; validation, Y.Z. and L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the National Key R&D Program of China (Grant No. 2019YFB1600602); the National Natural Science Foundation of China (No. 52071049); Liaoning Provincial Science and Technology Plan (Key) project (No. 2022JH1/10800096); the Natural Science Foundation of Liaoning Province (No. 2020-BS-070); the Fundamental Research Funds for the Central Universities (Grant No. 3132022131).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

1. Thombre, S.; Zhao, Z.; Ramm-Schmidt, H.; Garcia, J.M.V.; Malkamaki, T.; Nikolskiy, S.; Hammarberg, T.; Nuortie, H.; Bhuiyan, M.Z.H.; Sarkka, S.; et al. Sensors and AI Techniques for Situational Awareness in Autonomous Ships: A Review. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 64–83. [\[CrossRef\]](#)
2. Lyu, H.; Shao, Z.; Cheng, T.; Yin, Y.; Gao, X. Sea-Surface Object Detection Based on Electro-Optical Sensors: A Review. *IEEE Intell. Transp. Syst. Mag.* **2022**, 2–27. [\[CrossRef\]](#)
3. Vicen-Bueno, R.; Carrasco-Alvarez, R.; Jarabo-Amores, M.; Nieto-Borge, J.; Rosa-Zurera, M. Ship detection by different data selection templates and multilayer perceptrons from incoherent maritime radar data. *IET Radar Sonar Navig.* **2011**, *5*, 144. [\[CrossRef\]](#)
4. Zhuang, J.-Y.; Zhang, L.; Zhao, S.-Q.; Cao, J.; Wang, B.; Sun, H.-B. Radar-based collision avoidance for unmanned surface vehicles. *China Ocean Eng.* **2016**, *30*, 867–883. [\[CrossRef\]](#)
5. Szpak, Z.; Tapamo, J.-R. Maritime surveillance: Tracking ships inside a dynamic background using a fast level-set. *Expert Syst. Appl.* **2011**, *38*, 6669–6680. [\[CrossRef\]](#)
6. Bloisi, D.D.; Previtali, F.; Pennisi, A.; Nardi, D.; Fiorini, M. Enhancing Automatic Maritime Surveillance Systems with Visual Information. *IEEE Intell. Transp. Syst.* **2017**, *18*, 824–833. [\[CrossRef\]](#)
7. Prasad, D.K.; Prasath, C.K.; Rajan, D.; Rachmawati, L.; Rajabally, E.; Quek, C. Object Detection in a Maritime Environment: Performance Evaluation of Background Subtraction Methods. *IEEE Intell. Transp. Syst.* **2019**, *20*, 1787–1802. [\[CrossRef\]](#)
8. Liu, K.K.; Wang, J.H. A Method of Detecting Wave Grade Based on Visual Image Taken by USV. *Appl. Mech. Mater.* **2013**, 291–294, 2437–2441. [\[CrossRef\]](#)
9. 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. [\[CrossRef\]](#)
10. Muhovic, J.; Mandeljic, R.; Bovcon, B.; Kristan, M.; Pers, J. Obstacle Tracking for Unmanned Surface Vessels Using 3-D Point Cloud. *IEEE J. Ocean. Eng.* **2020**, *45*, 786–798. [\[CrossRef\]](#)
11. Shao, Z.; Wu, W.; Wang, Z.; Du, W.; Li, C. SeaShips: A Large-Scale Precisely Annotated Dataset for Ship Detection. *IEEE Trans. Multimedia* **2018**, *20*, 2593–2604. [\[CrossRef\]](#)
12. Chen, X.; Yang, Y.; Wang, S.; Wu, H.; Tang, J.; Zhao, J.; Wang, Z. Ship Type Recognition via a Coarse-to-Fine Cascaded Convolution Neural Network. *J. Navig.* **2020**, *73*, 813–832. [\[CrossRef\]](#)
13. 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. [\[CrossRef\]](#)
14. Chan, Y.-T. Comprehensive comparative evaluation of background subtraction algorithms in open sea environments. *Comput. Vis. Image Underst.* **2021**, *202*, 103101. [\[CrossRef\]](#)
15. Zhu, C.; Zhou, H.; Wang, R.; Guo, J. A Novel Hierarchical Method of Ship Detection from Spaceborne Optical Image Based on Shape and Texture Features. *IEEE Trans. Geosci. Remote Sens.* **2010**, *48*, 3446–3456. [\[CrossRef\]](#)
16. Arshad, N.; Moon, K.-S.; Kim, J.-N. Multiple Ship Detection and Tracking Using Background Registration and Morphological Operations. In *Signal Processing and Multimedia*; Kim, T., Pal, S.K., Grosky, W.I., Pissinou, N., Shih, T.K., Ślęzak, D., Eds.; Springer: Berlin/Heidelberg, Germany, 2010; Volume 123, pp. 121–126. [\[CrossRef\]](#)
17. Lecun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **1998**, *86*, 2278–2324. [\[CrossRef\]](#)

18. Ioffe, S.; Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv* **2015**, arXiv:1502.03167.
19. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the Inception Architecture for Computer Vision. *arXiv* **2015**, arXiv:1512.00567.
20. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 1–9 June 2015; pp. 1–9. [[CrossRef](#)]
21. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [[CrossRef](#)]
22. Long, J.; Shelhamer, E.; Darrell, T. Fully Convolutional Networks for Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *4*, 640–651.
23. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *arXiv* **2015**, arXiv:1505.04597.
24. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2015**, arXiv:1409.1556.
25. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
26. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2261–2269. [[CrossRef](#)]
27. Girshick, R.; Donahue, J.; Darrell, T.; 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; pp. 580–587. [[CrossRef](#)]
28. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [[CrossRef](#)]
29. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving into High Quality Object Detection. *arXiv* **2017**, arXiv:1712.00726.
30. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
31. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-J.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
32. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 318–327. [[CrossRef](#)]
33. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *Computer Vision—ECCV 2016*; Leibe, B., Matas, J., Sebe, N., et al., Eds.; Springer International Publishing: Cham, Switzerland, 2016; Volume 9905, pp. 21–37. [[CrossRef](#)]
34. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully Convolutional One-Stage Object Detection. *arXiv* **2019**, arXiv:1904.01355.
35. Zhang, H.; Wang, Y.; Dayoub, F.; Sünderhauf, N. VarifocalNet: An *IoU*-aware Dense Object Detector. *arXiv* **2021**, arXiv:2008.13367.
36. Shao, Z.; Wang, L.; Wang, Z.; Du, W.; Wu, W. Saliency-Aware Convolution Neural Network for Ship Detection in Surveillance Video. *IEEE Trans. Circuits Syst. Video Technol.* **2020**, *30*, 781–794. [[CrossRef](#)]
37. Liu, T.; Pang, B.; Zhang, L.; Yang, W.; Sun, X. Sea Surface Object Detection Algorithm Based on YOLO v4 Fused with Reverse Depthwise Separable Convolution (RDSC) for USV. *J. Mar. Sci. Eng.* **2021**, *9*, 753. [[CrossRef](#)]
38. Guo, H.; Yang, X.; Wang, N.; Song, B.; Gao, X. A Rotational Libra R-CNN Method for Ship Detection. *IEEE Trans. Geosci. Remote Sens.* **2020**, *58*, 5772–5781. [[CrossRef](#)]
39. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944. [[CrossRef](#)]
40. Ghahremani, A.; Bondarev, E.; De With, P.H. Cascaded CNN Method for Far Object Detection in Outdoor Surveillance. In Proceedings of the 2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS), Las Palmas de Gran Canaria, Spain, 26–29 November 2018; pp. 40–47. [[CrossRef](#)]
41. Iancu, B.; Soloviev, V.; Zelioli, L.; Lilius, J. ABOships—An Inshore and Offshore Maritime Vessel Detection Dataset with Precise Annotations. *Remote Sens.* **2021**, *13*, 988. [[CrossRef](#)]
42. Dai, J.; Qi, H.; Xiong, Y.; Li, Y.; Zhang, G.; Hu, H.; Wei, Y. Deformable Convolutional Networks. *arXiv* **2017**, arXiv:1703.06211.
43. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-*IoU* Loss: Faster and Better Learning for Bounding Box Regression. *arXiv* **2019**, arXiv:1911.08287. [[CrossRef](#)]
44. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS—Improving Object Detection with One Line of Code. *arXiv* **2017**, arXiv:1704.04503.
45. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. UnitBox: An Advanced Object Detection Network. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 516–520. [[CrossRef](#)]
46. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. *arXiv* **2019**, arXiv:1902.09630.

47. Neubeck, A.; Van Gool, L. Efficient Non-Maximum Suppression. In Proceedings of the 18th International Conference on Pattern Recognition (ICPR'06), Hong Kong, China, 20–24 August 2006; pp. 850–855. [[CrossRef](#)]
48. 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](#)]
49. Zhou, Z.; Sun, J.; Yu, J.; Liu, K.; Duan, J.; Chen, L.; Chen, C.L.P. An Image-Based Benchmark Dataset and a Novel Object Detector for Water Surface Object Detection. *Front. Neurobot.* **2021**, *15*, 723336. [[CrossRef](#)]
50. Wang, S.; He, Z. A prediction model of vessel trajectory based on generative adversarial network. *J. Navig.* **2021**, *74*, 1161–1171. [[CrossRef](#)]
51. Chen, Z.; Chen, D.; Zhang, Y.; Cheng, X.; Zhang, M.; Wu, C. Deep learning for autonomous ship-oriented small ship detection. *Saf. Sci.* **2020**, *130*, 104812. [[CrossRef](#)]