Journal of
Marine Science
and Engineering

MDPI

Article

A Lightweight Sea Surface Object Detection Network for Unmanned Surface Vehicles

Zhangqi Yang, Ye Li, Bo Wang *, Shuoshuo Ding and Peng Jiang

National Key Laboratory of Science and Technology on Underwater Vehicle, Harbin Engineering University, Harbin 150001, China; yangzhangqi@hrbeu.edu.cn (Z.Y.); liye@hrbeu.edu.cn (Y.L.); dingshuoshuo@hrbeu.edu.cn (S.D.); jiangpeng@hrbeu.edu.cn (P.J.)
* Correspondence: wb@hrbeu.edu.cn

Abstract: For unmanned surface vehicles (USVs), perception and control are commonly performed in embedded devices with limited computing power. Sea surface object detection can provide sufficient information for USVs, while most algorithms have poor real-time performance on embedded devices. To achieve real-time object detection on the USV platform, this paper designs a lightweight object detection network based on YOLO v5. In our work, an improved ShuffleNet v2 based on the attention mechanism was adopted as a backbone network to extract features. The depth-wise separable convolution module was introduced to rebuild the neck network. Additionally, the fusion method was changed from Concat to Add to optimize the feature fusion module. Experiments show that the proposed method reached 32.64 frames per second (FPS) on the Nvidia Jetson AGX Xavier and achieved a mean average precision (mAP) of 93.1% and 93.9% on our dataset and Singapore Maritime Dataset, respectively. Moreover, the number of model parameters of the proposed network was only 25% of that of YOLO v5n. The proposed network achieves a better balance between speed and accuracy, which is more suitable for detecting sea surface objects for USVs.

Keywords: YOLO v5; lightweight object detection; ShuffleNet v2; unmanned surface vehicles

check for updates

Citation: Yang, Z.; Li, Y.; Wang, B.; Ding, S.; Jiang, P. A Lightweight Sea Surface Object Detection Network for Unmanned Surface Vehicles. *J. Mar. Sci. Eng.* **2022**, *10*, 965. https://doi.org/10.3390/jmse10070965

Academic Editors: Mai The Vu and Hyeung-Sik Choi

Received: 21 June 2022
Accepted: 12 July 2022
Published: 14 July 2022

1. Introduction

In recent years, unmanned surface vehicles (USVs) have been widely used to perform varieties of tasks due to the increase in human maritime activities. The advantages of USVs are commonly considered high speed, small size, and labor-saving. In addition, it has been proved that USVs can complete sophisticated tasks and are more adaptable and efficient. The environmental perception system of the USV is composed of various sensors, such as optical cameras and radar. Radar has relative stability in complex marine environments so that it can be used in various weather conditions [1]. However, optical cameras can obtain high-resolution images, which provide rich environmental information. Visual object detection plays an important role in obstacle avoidance and autonomous navigation of the USV. However, the computing power of embedded devices carried by USV is generally insufficient. Thus, it is significant to develop a lightweight sea-surface objects detection network that can easily deploy on embedded devices of USV.

Recently, with the development of computer hardware, many researchers are focusing on object detection based on deep learning [2], where neural networks can be used to extract the features of the detected object automatically. These algorithms can be divided into two main types. The first is two-stage object detection, which divides object detection into two-phase such as R-CNN [3], Fast R-CNN [4], and Cascade R-CNN [5]. Another is single-stage object detection, which directly predicts the position and category, such as SSD [6], YOLO [7–10], and RefineDet [11]. The performance of object detection algorithms has significantly improved, and some have been applied in practical scenarios, including self-driving cars [12], unmanned aerial vehicles [13], and more.

At present, to make USVs more intelligent, object detection methods based on deep learning have also been used to detect sea surface objects. However, most of the related studies focus on improving detection performance by increasing the model parameters or adding different components to the network. Although the recognition accuracy of most existing detection algorithms is high, the real-time performance of most algorithms is insufficient due to the high complexity and a large number of model parameters. To satisfy the requirements of USVs, it is essential to reduce the model parameters and improve the detection speed.

In this study, we design a lightweight sea surface object detection network based on the framework of YOLO v5 [14] to improve detection speed on an embedded device. YOLO v5 is an excellent object detection algorithm considering the balance between speed and accuracy. However, its network still has many redundant parameters, and further optimization and improvement are required. Consequently, we proposed a lightweight object detection network for USVs. Firstly, improved ShuffleNet v2 [15] based on squeeze operation and excitation (SE) attention mechanism [16] was used as the backbone network of YOLO v5. In addition, depth-wise separable convolution was introduced to rebuild the neck network. Meanwhile, the Add fusion method was applied to optimize the Pixel Aggregation Network (PAN) [17] used in YOLO v5. Experiments were carried out on our dataset, which was collected and labeled for this study, and the Singapore Maritime Dataset. The results show that the network greatly improved the detection speed, which is extremely suitable for deploying on USVs to detect sea surface objects. The primary contributions of the paper are as follows:

(1) An improved ShuffleNet v2 based on the SE attention mechanism was proposed as the backbone feature extraction network, which significantly reduces the number of model parameters.

(2) A combination of the depth-wise separable convolution and the ADD feature fusion method was adopted to rebuilt the neck network, which is conducive to reducing the complexity of computation.

(3) We provided a solution for deploying sea surface object detection algorithms on embedded devices carried by USVs. All experiments were tested on NVIDIA Jetson AGX XAVIER, and real-time performance was demonstrated.

The remainder of this paper is structured as follows. Section 2 introduces some of the most important related works. The visual perception system of USVs is introduced in Section 3. Section 4 describes the proposed object detection network. Section 5 describes the experimental results and analysis. Finally, a summary and prospects are provided in Section 6.

2. Related Work

Many researchers have achieved outstanding achievements in the field of sea surface object detection. Some object detection algorithms based on deep learning have been successfully employed in USVs to detect sea surface objects. Tao Liu et al. [18] proposed a sea surface object detection algorithm based on YOLO v4. They introduced the module of Reverse Depthwise Separable Convolution [19] to reduce the number of weights. A novel method [20] was proposed to fuse DenseNet [21] and YOLOV3 [9], which enhanced the environmental adaptability of the USV. Xiaoqiang Sun et al. [22] introduced a fast weighted feature fusion network into the USV object detection network, which achieved better performance under different lighting and weather conditions. A ship detection algorithm [23] was designed based on Discrete Cosine Transform. The approach can improve detection accuracy and enhance real-time performance. A visual detection algorithm based on improved YOLOv3 was proposed to detect sea surface targets [24]. The accuracy of YOLO v3 to detect sea-surface targets was improved by increasing the inference time.

Most studies have improved detection accuracy in complex marine environments. Improving the accuracy usually increases the number of model parameters. Due to the limited computing power of embedded devices, the application of object detection networks on embedded devices carried by USVs is seriously limited. To apply object detection methods based on deep learning to embedded devices, lightweight networks have been studied to reduce the number of model parameters and the complexity of computation.

At present, the research on lightweight neural networks can be divided into two aspects: one is network structure, and another is model compression. The former is to directly design lightweight networks, such as ShuffleNet(V1, V2) [15,25], MobileNet (V1, V2, V3) [26–28], and Xception [19]. The latter aims to compress models through pruning [29] or knowledge distillation [30], for example. The main idea of pruning is to find and discard several unimportant channels or kernels. The main idea of knowledge distillation is to distill knowledge from a large model to a small model. Nowadays, lightweight networks have become a popular research direction in object detection, such as PP-PicoDet [31], Nanodet [32], and YOLO-Fastest [33]. They have significantly reduced the number of model parameters and improved the detection speed, but the accuracy is comparatively low.

Although the research on lightweight networks has great engineering value, there is little research in the area. In the study of sea-sky-line detection, Lujing Yang [34] et al. proposed a lightweight network based on YOLO v5 to detect sea-sky-line. However, they did not consider other sea surface objects such as ships and buoys. In this work, we design a lightweight network to detect sea-surface objects well and quickly.

3. Perception System of USVs

Due to the influence of the wind and waves, USVs will roll, pitch, and heave in a marine environment. Relative motion between the detected object and the camera mounted on the USV will blur the image. To eliminate the influence of environmental disturbances, we use an optoelectronic pod with image stabilization to collect sea surface images in real-time. The perception system of the USV is shown in Figure 1. The optoelectronic pod collects sea surface images and transmits them to the embedded computing module through the digital switch. The inertial navigation system (INS) continuously transmits the position and orientation information of the USV. NVIDIA Jetson Xavier platform is adopted as an embedded computing device to process the collected images and implement object detection algorithms for feature extraction and detection. Then, the detection results are transmitted to the industrial personal computer (IPC) to provide information for path planning and motion control of USVs.

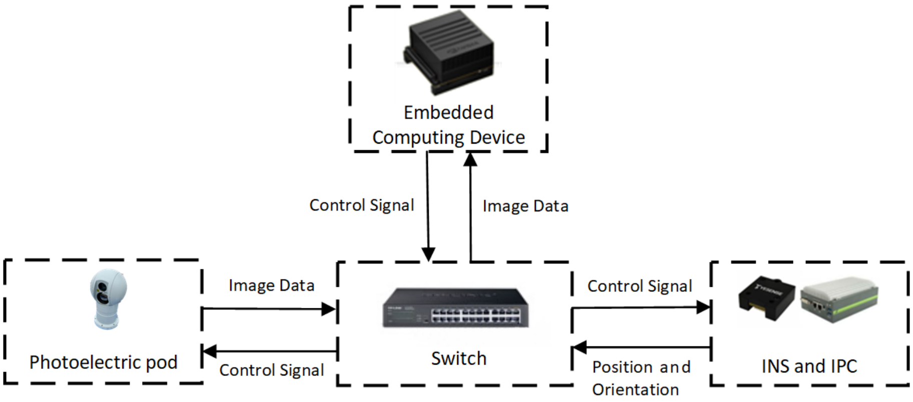

Figure 1. The visual perception system of USVs.

4. Method

The framework of YOLO v5 is mainly composed of three components: backbone, neck, and head. The backbone network is mainly composed of BottleneckCSP modules, which are used to extract feature information. The neck network adopts an improved feature pyramid network (FPN) structure, which is used to fuse extracted feature maps. The detection head is the final detection part of the model, which predicts the category and position of the detected object. To achieve the purpose of being lightweight, ShuffleNet v2 was finally chosen as the backbone network after comparing it to some other lightweight networks instead of the large backbone network used in YOLO v5. To enlarge the interest area, the SE attention mechanism was embedded into the ShuffleNet v2 structure. Additionally, the depth-wise separable convolution was used to rebuild the neck network, which can further reduce the computational complexity. Then, three feature maps were enhanced by the FPN and PAN structures for feature fusion. The Concat fusion method of PAN was replaced by the Add method to better fuse different scale feature maps and reduce the number of model parameters. The structure of our network is shown in Figure 2.

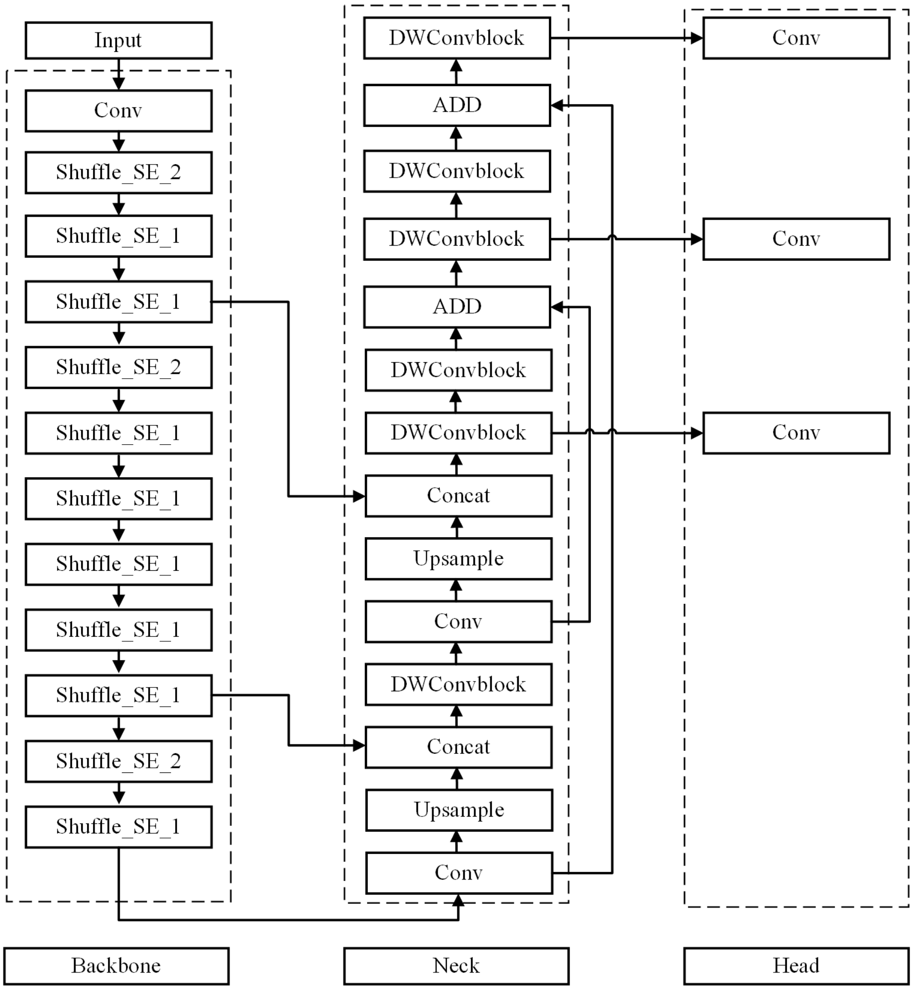

Figure 2. The proposed lightweight object detection network structure.

4.1. Improvement of Backbone Network

The object detection algorithm for USVs not only needs to identify sea surface targets in a complex marine environment accurately but also needs to reduce computational complexity as much as possible. Later, the backbone network of YOLOv5 architecture was optimized.

The backbone network of YOLO v5 is the BottleneckCSP [35] module, which contains multiple convolutional layers. These convolutional layers can extract features from the image. A considerable number of parameters are included in the convolution kernel, which will increase the number of model parameters and slow down the inference speed. Therefore, the network of ShuffleNet v2 is used to replace the original backbone network. ShuffleNet v2 has the advantages of fewer model parameters and fast calculation speed. Figure 3 is the network structure of ShuffleNet v2, which introduces the operation of channel split to split the feature channels into two branches and calculate them separately. ShuffleNet v2 has a larger capacity due to the amount of network parameters, and calculations are reduced by the operation of channel split. At the same time, half of feature channels directly go through the module and join the next module without convolutional computation. The operation can be considered a kind of feature reuse, which has been used previously in DenseNet [21] and CondenseNet [36]. As their connection between adjacent layers is stronger than other layers, the feature reuse mode with dense connection will be redundant. The feature reuse mode in ShuffleNet v2 is more effective, which is very beneficial to our network.

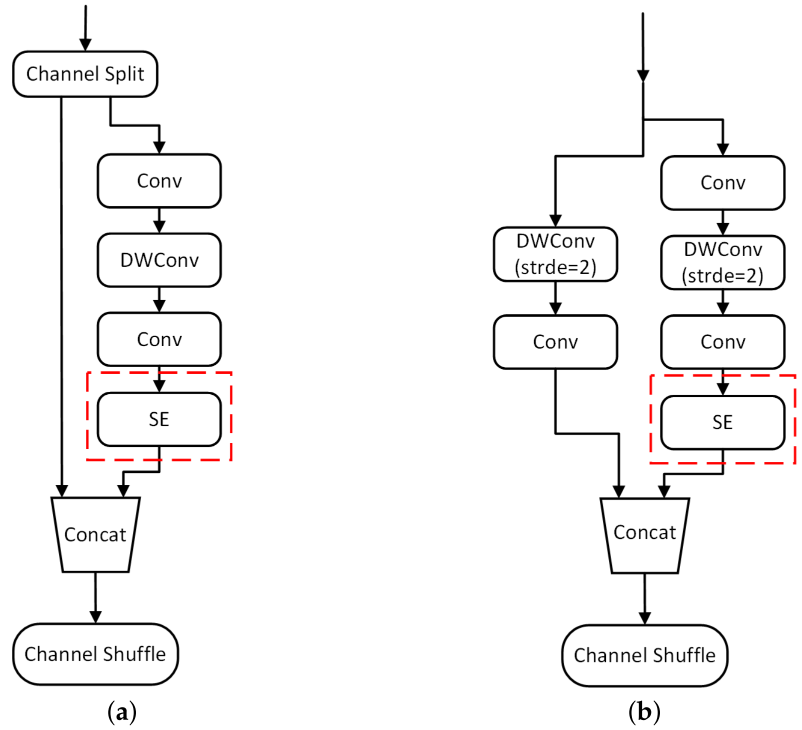

Figure 3. The improved ShuffleNet v2 unit is based on the SE attention mechanism. (**a**): the basic unit; (**b**): the unit for spatial down-sampling (2×).

In real marine environments, due to the wide view of the sea surface and the large sensing range of the unmanned surface vehicles, there are many small-sized targets. Moreover, the characteristics of different types of sea surface objects, such as ships, are relatively similar. These factors greatly increase the difficulty of sea surface object detection.

In recent years, attention mechanisms have been widely used in various computer vision tasks. SE is a lightweight attention mechanism. In this paper, we embed the SE module into ShuffleNet v2 to improve the weight of important channels. This greatly improves the ability to detect sea surface objects. As shown in Figure 4 , the SE module mainly consists of two steps: Squeeze and Excitation operations. The Squeeze operation compresses the feature map of each channel by global pooling, and the Excitation operation can obtain the weight of each channel. Then, the Scale operation multiplies these weights with the original feature map. The SE channel attention mechanism improves accuracy with little computational cost.

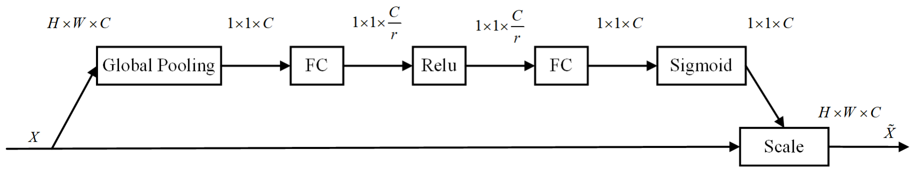

Figure 4. The structure of the SE attention mechanism.

4.2. Depth-Wise Separable Convolution

For the original YOLO v5 network, the BottleneckCSP module is in the neck network. Inspired by MobileNet, which uses depth-wise separable convolutions to replace the normal convolution, the module of depth-wise separable convolutions is used to replace the BottleneckCSP module in the neck network. As shown in Figure 5, the depth-wise separable convolution consists of a depth-wise convolution followed by a point-wise convolution. Depth-wise convolution adopts one filter on each channel. Then, a 1 × 1 convolution is applied by the point-wise convolution to generate novel feature maps. Depth-wise separable convolution can significantly reduce the computational complexity and the number of model parameters.

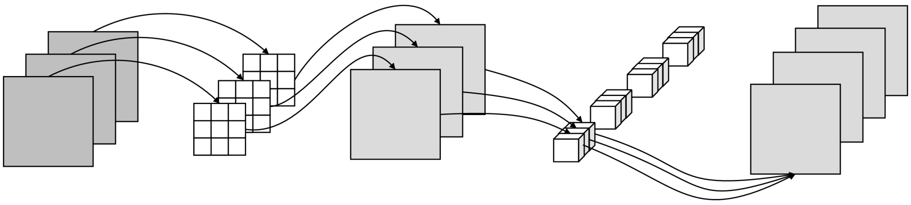

Figure 5. The process of depth-wise separable convolution.

4.3. Improvement of Feature Fusion Module

In object detection, the effective fusion of different scale feature maps is the key to improving the model performance. Following the arrow direction in Figure 6, the FPN enlarges the feature map size through the up-sampling operation and fuses it with the feature map from the backbone network to convey semantic information. The PAN is responsible for reducing the feature map size through the operation of down-sampling, and the feature maps in the PAN structure are fused with corresponding feature maps in the FPN structure to transfer strong positioning information. Feature maps of different scales are repeatedly fused to extract semantic information better and positioning information, which is helpful in detecting multiple-scale objects.

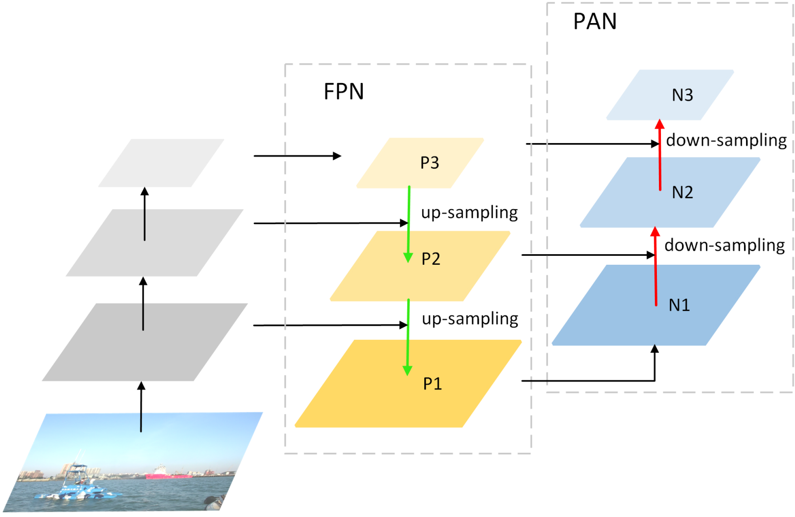

Figure 6. The structure of the feature fusion network.

Considering the requirement of being lightweight, we further improved the fusion method of the PAN structure to improve the detection speed. The Concat fusion method increases the number of channels and the amount of calculation. However, the Add fusion method increases the information of each channel without increasing the number of channels. As a result, the Add method is a better choice, as shown in Figure 7.

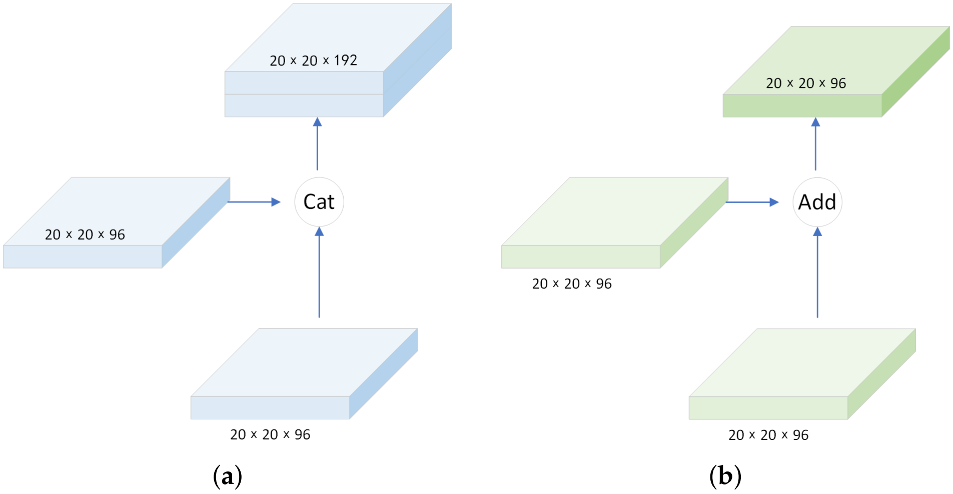

Figure 7. (**a**): The Concat fusion method; (**b**): The Add fusion method.

The two figures above describe the different characteristics of the Add fusion method and the Concat fusion method, while the formula below can more directly describe the difference between the two methods. Suppose the two input channels are $X_1, X_2, \ldots, X_C$ and $Y_1, Y_2, \ldots, Y_C$, respectively. The output channel of the Concat method is ($*$ denotes convolution and K represents the convolution kernel):

$$Z_{\text{concat}} = \sum_{i=1}^{c} X_i * K_i + \sum_{i=1}^{c} Y_i * K_{i+c} \tag{1}$$

The output channel for Add is:

$$Z_{add} = \sum_{i=1}^{c} (X_i + Y_i) * K_i = \sum_{i=1}^{c} X_i * K_i + \sum_{i=1}^{c} Y_i * K_i \tag{2}$$

It can be seen from the above formula that: (1) The Add method has more advantages for object detection because the amount of information is increased by increasing the information under each channel, while the number of channels does not increase. (2) The Concat method involves combing two input channels, which increases the number of channels while the information under each channel is not increased. (3) Therefore, the parameters of the Concat method are nearly twice as numerous as the Add method. When the two input channels have similar dimensions to the corresponding channel, the Add method can reduce the number of parameters compared with the Concat method.

5. Experiment

5.1. Introduction of Dataset

The datasets used in the experiments were our own and the Singapore Maritime Dataset. Both are randomly split into 80% for training and 20% for validation. The following is a detailed description of the two datasets.

The images in our dataset were collected in the real marine environment. The optoelectronic pod was installed horizontally on top of the USV for image acquisition. Figure 8 shows the "QZ" USV developed by Harbin Engineering University. To verify the effectiveness and robustness of the proposed lightweight object detection network, we collected images under different weather conditions, such as sunny, rainy, and foggy days. This experiment was carried out in the Zhanjiang sea area, Guangdong Province, China.

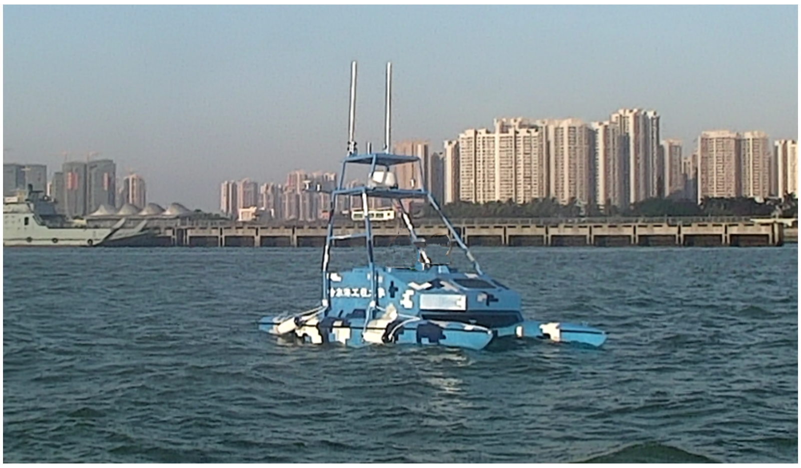

Figure 8. The "QZ" USV platform.

Then, we selected images from the collected videos and manually labeled them. A total of 5 categories were labeled, and the numbers of each category are shown in Table 1. Figure 9 shows some images of our dataset.

Table 1. The instances information statistics of our dataset.

Class	Instances	Percentage
Raft	482	22.22%
Ship	845	32.26%
Buoy	979	37.38%
USV	203	7.75%
Boat	110	4.2%

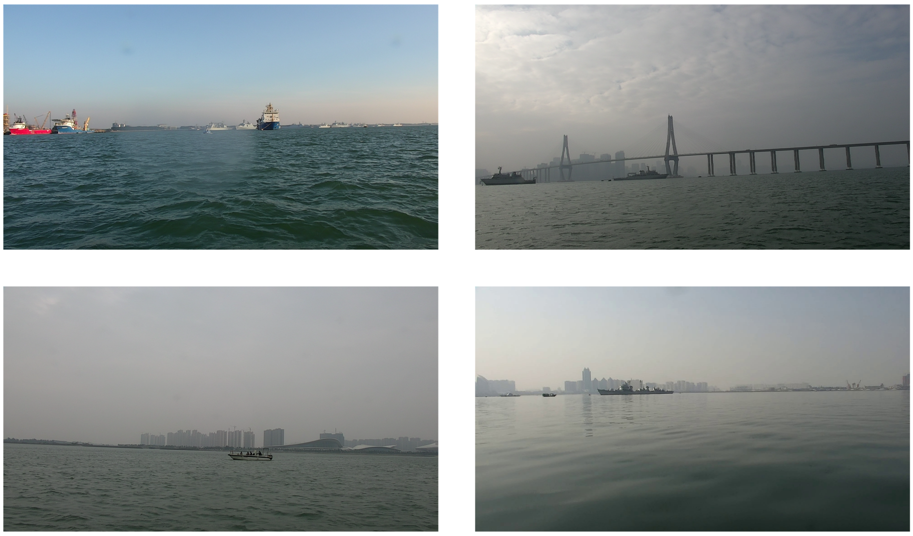

Figure 9. Training samples from our dataset.

The Singapore Maritime Dataset [37] is open source and contains videos collected from Singaporean waters with a high definition (1920 × 1080 pixels). The dataset contains various light and weather conditions. Images were obtained at an interval of three frames, with a total of 17,966 images being obtained. The experiment focused on eight categories: Ferry, Buoy, Vessel/ship, Speed boat, Boat, Kayak, Sailboat, and Other. Table 2 shows the numbers of each category. Figure 10 shows some images of the dataset.

Table 2. The instances information statistics of Singapore Maritime Dataset.

Class	Instances	Percentage
Ferry	8588	5.63%
Buoy	2973	2.17%
Vessel/ship	114,411	74.19%
Speed boat	7780	4.95%
Boat	1298	0.8%
Kayak	4308	2.7%
Sail boat	1926	1.18%
Other	12,551	9.54%

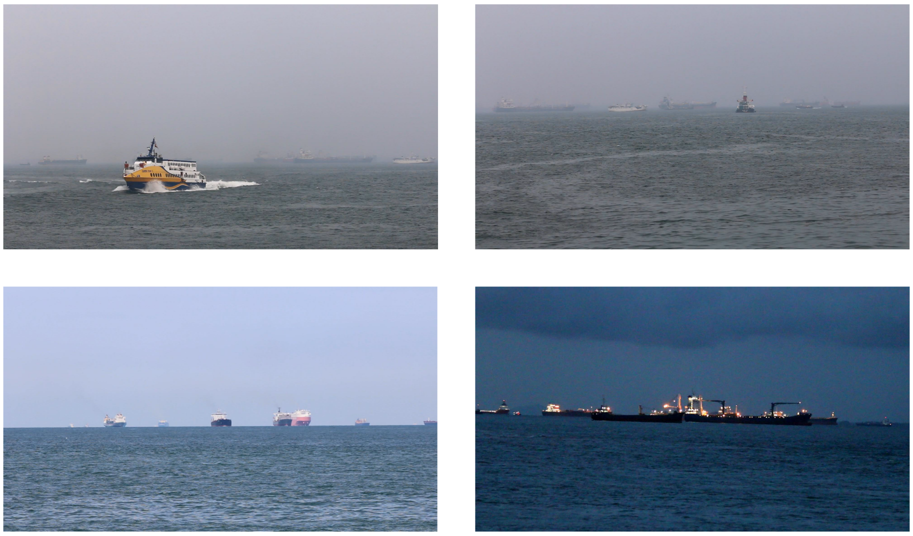

Figure 10. Training samples from Singapore Maritime Dataset.

5.2. Mosaic Augmentation

As shown in the above figures, our dataset and the Singapore Maritime Dataset contain some small-scale targets that are difficult to detect. To solve this problem, we proposed to do Mosaic image augmentation. The main idea is to crop four images randomly and then put them together into one image, as shown in the Figure 11. This enriched the background of the image and increased the number of small-sized objects. This significantly facilitated the robustness of the model and improved the performance when recognizing small targets.

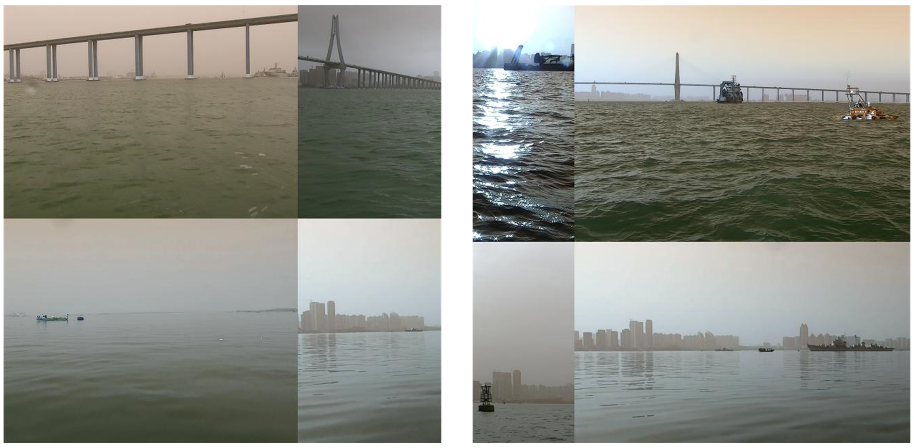

Figure 11. Mosaic image augmentation.

5.3. Training Details

Each experiment set the same initial training parameters to ensure fairness. The input image size was uniformly resized to 640 × 640. The networks are trained for 300 epochs with the optimizer of stochastic gradient descent (SGD). To speed up the training process, 4 Nvidia GeForce RTX 3090 GPUs are applied for parallel training. After training, the weight file of the model with the highest accuracy was saved; then, the validation set was utilized to evaluate the performance.

5.4. Evaluation Metrics

To better compare the performance of different networks, it was important to use appropriate evaluation metrics. Floating point operations (FLOPs) were used to evaluate the computational complexity. The number of frames per second (FPS) was used to evaluate the detection speed, and the mean average precision (*mAP*) was adopted to evaluate the accuracy. The *mAP* can be calculated as follows:

$$P = \frac{TP}{TP + FP} \tag{3}$$

$$R = \frac{TP}{TP + FN} \tag{4}$$

$$AP_i = \int_0^1 P(R)d(R) \tag{5}$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \tag{6}$$

TP is a true positive example; FP is a false positive example; FN is a false negative example; AP is the average accuracy of a certain category, and mAP is the average of APs in all categories.

5.5. Ablation Experiments

To verify the effectiveness of different methods and submodules introduced into our network, ablation experiments were carried out in this study. Table 3 shows the results of the ablation experiments. In the listed models, Model1 could be regarded as a benchmark, which shows the performance of YOLO v5n without any modification. Model2 replaced the original backbone network with ShuffleNet v2. Model3 mixed the SE attention mechanism and ShuffleNet v2. Model4 rebuilt the neck network with depth-wise separable convolution. Model5 changed the Concat fusion method of PAN structure to the Add fusion method.

The results showed that Model1 achieved the best detection performance. Although the mAP of Model2 decreased by 3.5% compared with Model1, the model parameters were reduced by 69.95%. After adding the SE module, the mAP of Model 3 increased by 2.2%. The model parameters of Model4 were reduced by 23.21% compared with Model3, while the mAP only decreased by 1.0%. After replacing the fusion method of the PAN structure with the Add method, the mAP of Model5 increased by 0.7% compared with Model4. The results of ablation experiments show that ShuffleNet, DWConv, and Add introduced in this study can significantly reduce the number of model parameters and improve the detection speed. The SE attention mechanism improves the accuracy with a simple architecture and a small amount of computation. Eventually, Model5 achieves the best balance between accuracy and speed.

To demonstrate the effectiveness of the SE attention mechanism, the predicted heat maps are visualized in Figure 12. The brighter area means that it is important for the detection result. It can be seen that the heat map, after adding the attention mechanism, can better focus on the center area of the detected object, which suggests that the SE module can effectively improve the importance of the object area.

Table 3. Results of ablation experiments.

Model	ShuffleNet	SE	DWConv	Add	mAP	Parameters	FLOPs	FPS
Model1					95.7%	1.77 M	4.2 G	22.39
Model2	✓				91.2%	0.53 M	1.8 G	29.82
Model3	✓	✓			93.4%	0.59 M	1.9 G	28.61
Model4	✓	✓	✓		92.4%	0.45 M	1.4 G	31.35
Model5	✓	✓	✓	✓	93.1%	0.44 M	1.3 G	32.64

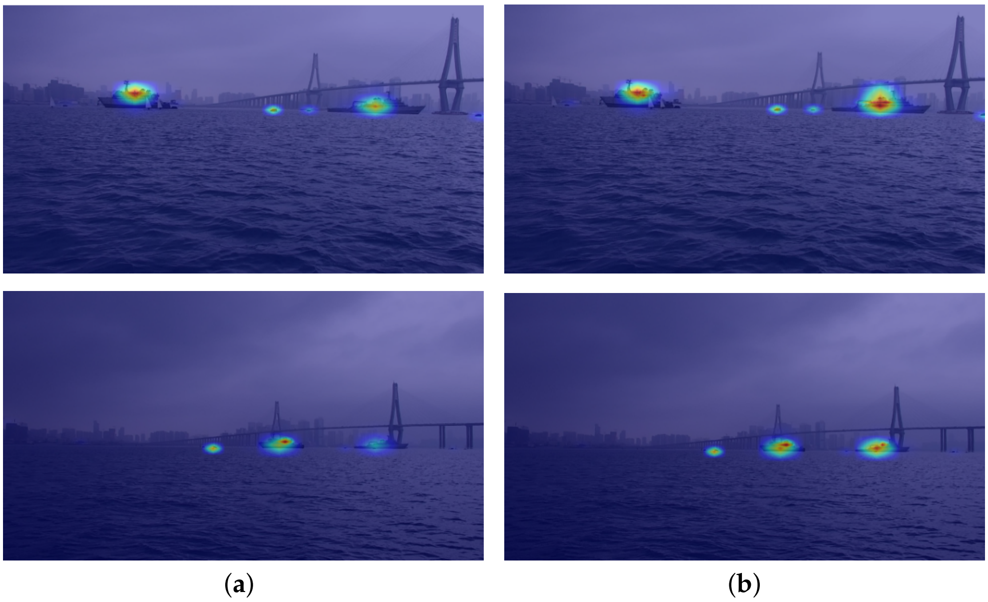
(**a**) (**b**)

Figure 12. Examples of predicted heat map. (**a**): Without attention mechanism; (**b**): With attention mechanism.

5.6. Comparison with Other Object-Detection Algorithms

The proposed network is based on the framework of YOLO v5. Five kinds of models with different sizes are designed in YOLO v5, which are YOLO v5x, YOLO v5l, YOLO v5m, YOLO v5s, and YOLO v5n. YOLO v5n is the smallest. Consequently, the results of our network are compared with YOLO v5n and some state-of-the-art lightweight object detection networks. All algorithms are evaluated on our dataset, rather than directly copying the results. The training and validation sets are identical in each experiment, and the evaluation metrics are exactly the same. This ensures that the experimental results are not affected by training details.

As shown in Table 4, YOLO v5n achieved the best performance in the experiments. Since our network was designed to improve the detection speed, the accuracy was slightly decreased compared with YOLO v5n, while the number of model parameters was reduced by 78.5%. The complexity of the network often reduced the inference speed. For USVs, considering both detection accuracy and speed, our proposed network was more suitable than YOLO v5n. The accuracy of our network is clearly better than that of YOLO v4-tiny [38], PP-PicoDet, Nanodet, and YOLO-Fastest.

As for the detection speed, the proposed network achieved 32.64 FPS when processing a video on Nvidia Jetson AGX Xavier. Our network significantly improved the detection speed compared with the YOLO v5n network, which provided a detection speed of 22.39 FPS. Compared with YOLO v4-tiny and PP-PicoDet, the detection speed was also higher. Although the YOLO-Fastest network achieved the fastest detection speed, its accuracy was only 82.3%. In summary, the proposed method can improve the inference speed to meet the requirements of USVs.

Table 4. Comparison of different object-detection algorithms.

Method	Parameters	mAP	FPS
YOLO v5n	1.77 M	95.7%	22.39
YOLO V4-tiny	6.05 M	91.0%	11.26
PP-PicoDet	0.78 M	89.8%	27.18
Nanodet	0.39 M	84.6%	34.27
YOLO-Fastest	0.25 M	82.3%	46.32
Ours	0.44 M	93.1%	32.64

Some detection results of the proposed network are shown in Figure 13. It can be seen from the results that almost all objects could be accurately recognized. Even small targets such as buoys could be accurately detected. After analyzing the detection results, we know that our lightweight object network performs well in the actual marine environment.

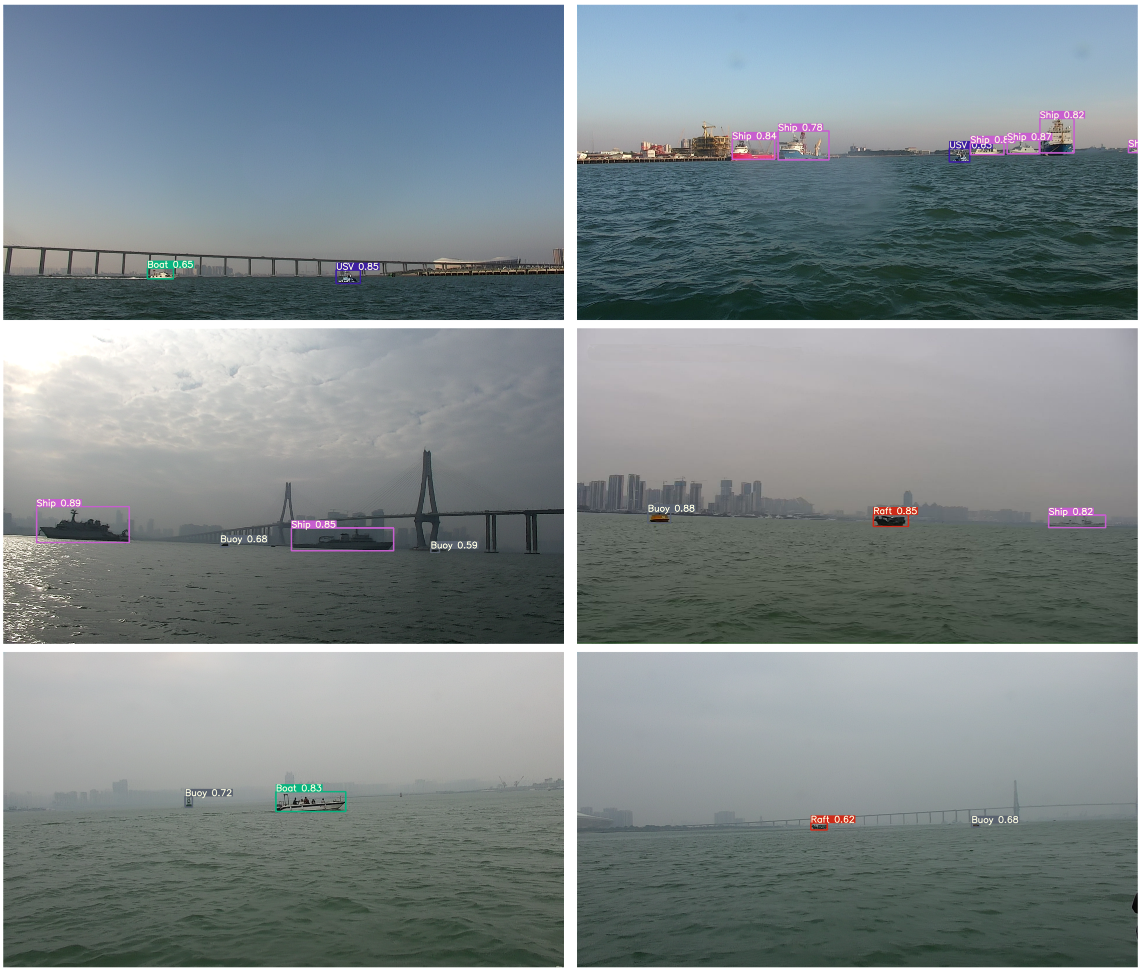

Figure 13. Visualization detection results on our dataset.

5.7. Detection Results on Singapore Maritime Dataset

To further verify the performance of our network in different marine environments, experiments were conducted on the Singapore Maritime Dataset. It can be concluded from the above experiments that YOLO v5n has the best performance and YOLO-Fastest has the fastest detection speed. Therefore, these two methods are compared with our method on the Singapore Maritime Dataset. As can be seen from the Table 5, the mAP of our network achieved 93.9%, which was 12.5% higher than YOLO-Fastest and 4.1% lower than YOLO v5n. The proposed method exhibited a better balance between accuracy and speed on the Singapore Maritime Dataset.

Table 5. Detection results on the Singapore Maritime Dataset.

Method	Ferry	Buoy	Vessel/Ship	Speed Boat	Boat	Kayak	Sail Boat	Other	mAP	FPS
YOLO v5n	97.5%	99.5%	98.9%	93.0%	87.7%	97.4%	99.4%	98.9%	96.5%	22.06
YOLO-Fastest	82.8%	92.5%	95.7%	68.9%	87.8%	38.4%	99.5%	82.3%	81.0%	45.98
Our	94.6%	98.7%	98.9%	87.5%	91.3%	86.0%	99.5%	94.8%	93.9%	32.11

The detection results on the Singapore Maritime Dataset are shown in Figure 14. The dataset contains different lighting and weather conditions. The proposed network had excellent performance in different conditions, such as sunny, foggy, and night, and could meet the requirements of USVs to perform tasks in different marine environments.

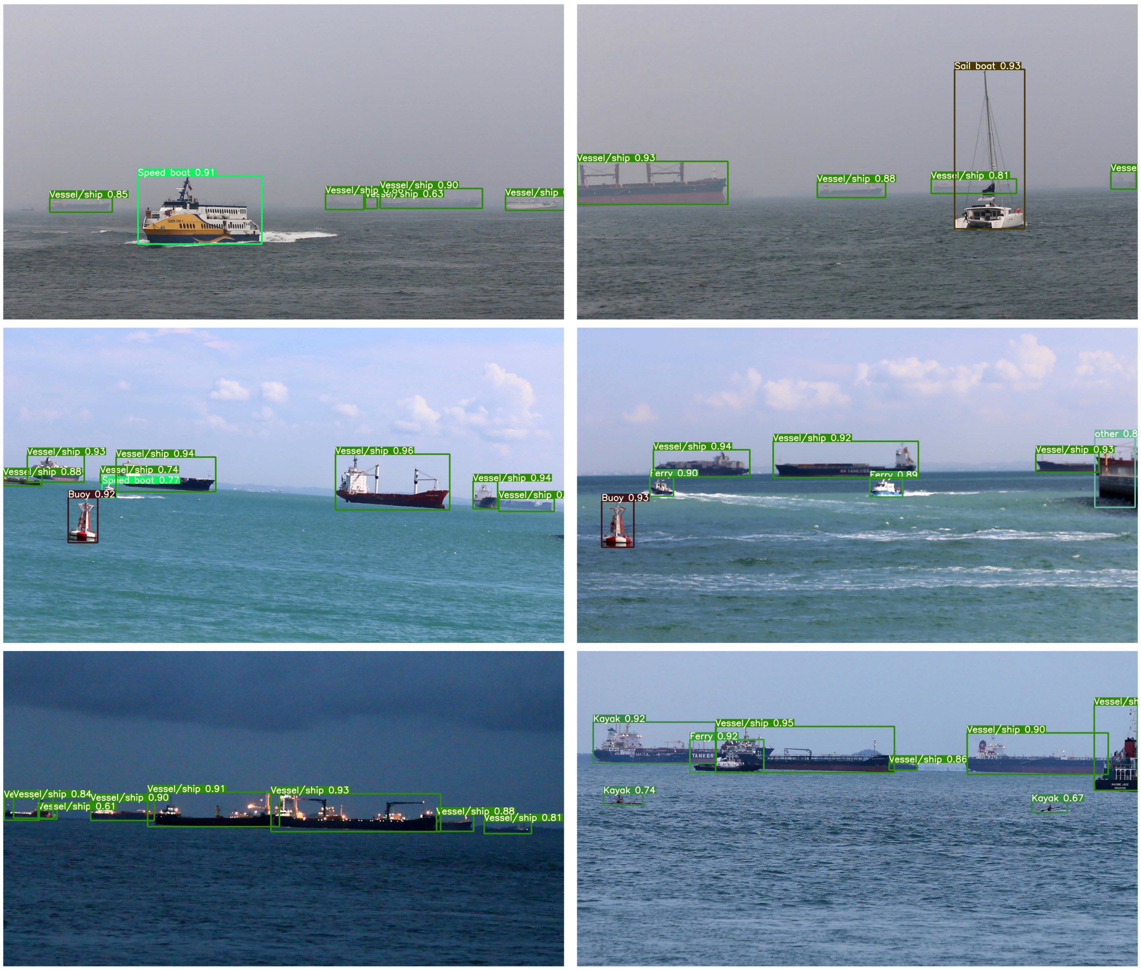

Figure 14. Visualization detection results on the Singapore Maritime Dataset.

6. Conclusions

In this study, we designed a lightweight sea surface object detection network to meet the requirements of USVs better. To improve the detection speed, an improved ShuffleNet v2 based on the SE attention mechanism was used as the backbone network and the depth-wise separable convolution module was adopted to rebuild the neck network, which significantly improved the detection speed. Then, the Add fusion method was introduced to the PAN structure to reduce the number of model parameters and improve accuracy. Experimental results showed that the model size of our network was reduced by 76.3% compared with YOLO v5, while the mAP was reduced by less than 3% on the two datasets. The proposed model achieves a better balance between speed and accuracy. However, due to the diversity of samples, there are still deficiencies in some areas that need further study and improvement.

In the future, the proposed network will be further optimized to reduce the detection speed and improve the accuracy. In addition, we hope to achieve real-time object detection on embedded devices with less computing power, such as the Raspberry Pi, which is more suitable for deployment on USVs.

Author Contributions: Data curation, Z.Y.; formal analysis, Y.L.; methodology, B.W.; resources, P.J.; supervision, B.W.; writing—original draft preparation, Z.Y.; writing—review and editing, Z.Y.; validation, S.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Natural Science Foundation of Heilongjiang Province, grant number JJ2020LH1535, by the Natural Science Foundation of Advanced Research, grant number 80912010101, by the Foundation of National Key Laboratory, grant number JCKYS2022SXJQR-06 and by the Foundation of National Key Laboratory, grant number 2021JCJQ-SYSJJ-LB06912.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Stateczny, A.; Kazimierski, W.; Gronska-Sledz, D.; Motyl, W. The empirical application of automotive 3D radar sensor for target detection for an autonomous surface vehicle's navigation. *Remote Sens.* **2019**, *11*, 1156.
2. Zhao, Z.Q.; Zheng, P.; Xu, S.t.; Wu, X. Object detection with deep learning: A review. *IEEE Trans. Neural Netw. Learn Syst.* **2019**, *30*, 3212–3232.
3. 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.
4. 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.* **2015**, *28*.
5. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
6. 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; pp. 21–37.
7. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
8. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
9. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
10. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
11. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
12. Gupta, A.; Anpalagan, A.; Guan, L.; Khwaja, A.S. Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues. *Array* **2021**, *10*, 100057.
13. Du, D.; Qi, Y.; Yu, H.; Yang, Y.; Duan, K.; Li, G.; Zhang, W.; Huang, Q.; Tian, Q. The unmanned aerial vehicle benchmark: Object detection and tracking. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 370–386.
14. YOLO v5. Available online: https://doi.org/10.5281/zenodo.5563715 (accessed on 12 October 2021).
15. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. Shufflenet v2: Practical guidelines for efficient cnn architecture design. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 116–131.
16. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
17. Wang, W.; Xie, E.; Song, X.; Zang, Y.; Wang, W.; Lu, T.; Yu, G.; Shen, C. Efficient and accurate arbitrary-shaped text detection with pixel aggregation network. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 8440–8449.
18. 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.

19. Chollet, F. Xception: Deep learning with depthwise separable convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1251–1258.
20. Li, Y.; Guo, J.; Guo, X.; Liu, K.; Zhao, W.; Luo, Y.; Wang, Z. A novel target detection method of the unmanned surface vehicle under all-weather conditions with an improved YOLOV3. *Sensors* **2020**, *20*, 4885.
21. Huang, G.; Liu, Z.; Van Der Maaten, L.; Weinberger, K.Q. Densely connected convolutional networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 4700–4708.
22. Sun, X.; Liu, T.; Yu, X.; Pang, B. Unmanned Surface Vessel Visual Object Detection Under All-Weather Conditions with Optimized Feature Fusion Network in YOLOv4. *J. Intell. Robot. Syst.* **2021**, *103*, 1–16.
23. Zhang, Y.; Li, Q.Z.; Zang, F.N. Ship detection for visual maritime surveillance from non-stationary platforms. *Ocean Eng.* **2017**, *141*, 53–63.
24. Liu, T.; Pang, B.; Ai, S.; Sun, X. Study on visual detection algorithm of sea surface targets based on improved YOLOv3. *Sensors* **2020**, *20*, 7263.
25. Zhang, X.; Zhou, X.; Lin, M.; Sun, J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
26. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv* **2017**, arXiv:1704.04861.
27. Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; Chen, L.C. Mobilenetv2: Inverted residuals and linear bottlenecks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
28. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for mobilenetv3. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 1314–1324.
29. Li, H.; Kadav, A.; Durdanovic, I.; Samet, H.; Graf, H.P. Pruning filters for efficient convnets. *arXiv* **2016**, arXiv:1608.08710.
30. Wang, T.; Yuan, L.; Zhang, X.; Feng, J. Distilling object detectors with fine-grained feature imitation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 4933–4942.
31. Yu, G.; Chang, Q.; Lv, W.; Xu, C.; Cui, C.; Ji, W.; Dang, Q.; Deng, K.; Wang, G.; Du, Y.; et al. PP-PicoDet: A Better Real-Time Object Detector on Mobile Devices. *arXiv* **2021**, arXiv:2111.00902.
32. NanoDet-Plus. Available online: https://github.com/dog-qiuqiu/Yolo-FastestV2 (accessed on 12 August 2021).
33. YOLO-Fastest. Available online: https://github.com/RangiLyu/nanodet (accessed on 26 December 2021).
34. Yang, L.; Zhang, P.; Huang, L.; Wu, L. Sea-sky-line Detection Based on Improved YOLOv5 Algorithm. In Proceedings of the 2021 IEEE 2nd International Conference on Information Technology, Big Data and Artificial Intelligence (ICIBA), Chongqing, China, 17–19 December 2021; pp. 688–694.
35. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020.
36. Huang, G.; Liu, S.; Van der Maaten, L.; Weinberger, K.Q. Condensenet: An efficient densenet using learned group convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 2752–2761.
37. 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.
38. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. Scaled-yolov4: Scaling cross stage partial network. In Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13029–13038.