

## Article

# Detection Technique Tailored for Small Targets on Water Surfaces in Unmanned Vessel Scenarios

Jian Zhang <sup>1</sup>, Wenbin Huang <sup>2,3</sup>, Jiayuan Zhuang <sup>1,\*</sup>, Renran Zhang <sup>1</sup> and Xiang Du <sup>1</sup>

<sup>1</sup> Science and Technology on Underwater Vehicle Laboratory, Harbin Engineering University, Harbin 150001, China; 17604683416@163.com (J.Z.); zhangrenran@hrbeu.edu.cn (R.Z.); due@hrbeu.edu.cn (X.D.)

<sup>2</sup> School of Marine Science and Technology, Northwestern Polytechnical University, Xi'an 710072, China

<sup>3</sup> Marine Design & Research Institute of China, Shanghai 200011, China

\* Correspondence: zhuangjiayuan@hrbeu.edu.cn

**Abstract:** Lightweight detection methods are frequently utilized for unmanned system sensing; however, to tackle the challenge of low precision in detecting small targets on the water's surface by unmanned surface vessels, we present an enhanced method for ship target detection tailored specifically to this context. Building upon the mainstream single-stage YOLOv8 object detection model, our approach involves the integration of the Reparameterized Convolutional Spatial Oversampling Attention (RCSOSA) module, replacing the traditional Classic 2D Convolutional (C2f) module to bolster the network's feature extraction capabilities. Additionally, we introduce a downsampling module, Spatial to Depth Convolution (SPDConv), to amplify the extraction of features relevant to small targets, thereby enhancing detection accuracy. Finally, the Focal Modulation module, based on focal modulation, replaces the SPPF (Spatial Pyramid Pooling with FPN) module, leading to a reduction in channel count, parameter volume, and an augmentation of the network's feature representation. Experimental results demonstrate that the proposed model achieves a 3.6% increase in mAP@0.5 and a 2.1% improvement in mAP@0.5:0.95 compared to the original YOLOv8 model, while maintaining real-time processing capabilities. The research validates the higher accuracy and stronger generalization capabilities of the proposed improved ship target detection method in various complex water surface environments.



**Citation:** Zhang, J.; Huang, W.; Zhuang, J.; Zhang, R.; Du, X. Detection Technique Tailored for Small Targets on Water Surfaces in Unmanned Vessel Scenarios. *J. Mar. Sci. Eng.* **2024**, *12*, 379. <https://doi.org/10.3390/jmse12030379>

Academic Editors: Mohamed Benbouzid, Lei Zhang and Bing Huang

Received: 29 December 2023

Revised: 18 February 2024

Accepted: 21 February 2024

Published: 23 February 2024



**Copyright:** © 2024 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/>).

**Keywords:** unmanned surface vessels; ship detection; reparameterized convolution; focal modulation; YOLOv8; small target detection algorithm

## 1. Introduction

Following the resurgence characterized by machine vision, artificial intelligence has played a pivotal role in driving the development of machine vision and image recognition [1], concurrently advancing the field of unmanned systems technology. Unmanned surface vessels (USVs), equipped with the ability to autonomously or semi-autonomously execute various tasks, distinguish themselves with features such as low maintenance costs, minimal energy consumption, and extended operational durations compared to conventional maritime equipment. Furthermore, they present a viable alternative to human involvement in complex and hazardous tasks. Consequently, the exploration of unmanned vessel technology aligns seamlessly with the imperative of human exploration of the oceans [2]. Environmental perception and target recognition stand out as core technologies for USVs, demanding accurate comprehension of the vessel's surroundings—an ideal task for artificial intelligence [3].

In the current stage, unmanned autonomous perception technology primarily relies on electro-optical pods for acquiring optical image information or laser radar for obtaining point cloud data to identify target types. However, point cloud data obtained from laser radar has limited coverage and lacks information richness, posing high computational

demands for direct three-dimensional point cloud processing. Conversely, optical images from the water's surface offer advantages such as rich color and texture information, ease of acquisition, and mature processing methods. Therefore, utilizing electro-optical devices to capture external features of targets remains a crucial approach for environmental perception in USVs.

The majority of existing target detection algorithms are designed for medium to large objects with sparse distribution in natural scenes. Classical two-stage methods [4–6], employing techniques like selective search, edge detection, and region extraction networks, are utilized to generate a set of candidate regions likely to contain the target. Convolutional neural networks are then employed to extract features from these candidate regions, enabling precise estimation of target categories and regression of bounding box positions. While these methods offer high accuracy, their processing efficiency and inference speed are suboptimal, limiting their applicability to real-time detection tasks with high-efficiency demands.

USVs rely on perception sensors, often likened to their visual “eyes”, to navigate during operations. Proper calibration of these sensors is essential to enhance obstacle avoidance capabilities, improve operational efficiency, and ensure safety standards are met. Despite the numerous detection techniques mentioned earlier, USVs possess limited capabilities in small target detection, emphasizing the need for models specifically designed for this purpose. Small target detection models encounter several challenges: detailed information about small targets may be excessively lost after multiple layers of neural network processing, leading to a decrease in detection accuracy; the distribution of large, medium, and small targets is typically uneven; small targets, with lower resolution compared to large/medium targets, contain less discriminative information, making the extraction of distinctive features challenging. In practical scenarios, issues such as image blurring due to wind and waves, vessel motion, and reduced visibility caused by factors like backlighting, rain, and fog can impact both image quality and detection outcomes for USVs. Consequently, algorithmic performance may be adversely affected in complex environments. In order to overcome the aforementioned challenges, this study proposes an enhanced model recognition approach for detecting small targets on ships, with a specific focus on USVs. The approach integrates spatial-to-depth convolution augmentation to enhance the accuracy of small target detection algorithms. Leveraging the exceptional performance of the YOLOv8n model, this research adopts it as the baseline model to investigate the integration of more robust spatial features of water surface targets on a lightweight foundation. The primary contributions of this paper are summarized below:

- (1) The essay initiates the improvement process by integrating the RCSOSA reparameterized convolution module in place of the conventional C2f convolution module, thereby enhancing the network's feature extraction capabilities.
- (2) Additionally, the introduction of the SPDCConv spatial-to-depth convolution module in the design of a downsampling convolution module aims to augment the extraction of features pertinent to small targets, consequently raising overall detection accuracy.
- (3) In the final iteration, the SPPF module is replaced by the Focal Modulation module based on Focal Modulation (FM) for enhanced performance. To assess the model's generalization capability, we validate its detection performance using a bespoke dataset crafted for this purpose.

Small target detection provides numerous advantages for USVs, directly impacting their performance and efficiency across various tasks and environments. Firstly, it enhances the navigational and obstacle-avoidance capabilities of USVs. By promptly identifying and accurately locating small targets on the water surface, USVs can effectively maneuver to avoid obstacles, thereby reducing the likelihood of collisions and unforeseen incidents thus enhancing maritime safety.

Secondly, small target detection contributes to the enhancement of USVs' surveillance and reconnaissance capabilities. In tasks such as ocean monitoring, environmental surveillance, and border patrols, timely detection and tracking of small targets (such as life rafts, floating debris, and small vessels) are crucial for ensuring maritime safety, environmen-

tal protection, and resource conservation. Equipped with efficient small target detection systems, USVs can execute these tasks more effectively, providing timely and accurate information and data support.

Furthermore, small target detection improves the emergency response capabilities of USVs. In scenarios involving emergencies or search and rescue operations, rapid detection and localization of small targets on the water surface can significantly enhance the efficiency of rescue efforts, shorten response times, and minimize casualties and property losses.

In summary, the benefits of small target detection for USVs include enhancing navigational and obstacle avoidance capabilities, strengthening surveillance and reconnaissance capabilities, and improving emergency response capabilities. These advantages contribute to the broader and more effective application of USVs in maritime operations, providing robust support and assurance for tasks at sea.

The paper is structured as follows: Section 2 provides an overview of current relevant research. Section 3 analyzes the methods employed and offers a detailed explanation of the proposed techniques. Section 4 introduces the experimental preparations. Experimental results are presented in Sections 5 and 6. Section 6 describes the sea trials, while Section 7 serves as the conclusion and discussion of the paper.

## 2. Related Works

Detecting small targets in complex backgrounds is a crucial research direction in the field of image analysis. Small targets are commonly found in aerial or space images captured from a distance, as well as in video surveillance. Effectively analyzing and processing high-quality image data using computers to identify objects of different classes and annotate their positions constitute fundamental challenges in computer vision tasks. This capability is widely applied in scenarios such as smart urban transportation, logistics management, agriculture and forestry development, public safety, and disaster response, enabling precise localization and identification of targets while saving substantial human and time costs. Therefore, small target detection technology holds significant research significance and practical application value. However, compared to images in natural scenes, images captured from a distance pose challenges with high background complexity, small target sizes, and ambiguous appearances. Target detection accuracy for small targets using deep learning techniques is generally lower, with weak adaptability to targets in complex backgrounds. This often leads to occurrences of missed detections and false alarms, significantly impacting the reliability of detection results. Therefore, small target detection remains a challenging problem in computer vision tasks. In recent years, with the continuous development of deep learning theory and the increasing demand from the industry, research on small target detection technology has garnered increasingly widespread attention. However, there is currently limited research specifically dedicated to algorithms for small target detection. Existing small target detection algorithms typically propose improvements or optimization strategies based on general object detection methods. These enhancements involve deepening network layers, designing backbone networks capable of extracting richer features [7,8] and complicating the feature fusion process [9,10] to enhance the model's robustness to multi-scale targets, thereby improving the detection performance for small targets. Classic two-stage object detection algorithms like Faster R-CNN [6] and single-stage algorithms like SSD [11] have demonstrated good performance in terms of both accuracy and speed. Two-stage methods have introduced several improved algorithms addressing small target issues [12–23], while single-stage improvements [24–26] mainly focus on leveraging multi-scale feature fusion to fully utilize detailed information in low-level, high-resolution features. Additionally, techniques such as generative adversarial networks [27–29] and data augmentation [30,31] are employed to address small detection challenges.

Review articles on small target detection are scarce, and the application of small target detection is typically restricted to specific scenarios. Thus, the current research on small target detection lacks generalizability, and the network framework's extensive

parameter count makes it unsuitable for the complex conditions encountered in unmanned maritime operations.

### 3. Methods

#### 3.1. The YOLOv8 Algorithm Principles

YOLOv8 is an advanced object detection algorithm conceptualized as an improvement over YOLOv7. The core principles of the algorithm encompass several key aspects, including network architecture, loss functions, and NMS (non-maximum suppression). In terms of network architecture, YOLOv8 introduces an innovative backbone network and detection head design to better capture features within images. The design of the loss function takes into account the characteristics of object detection tasks, enhancing model robustness by balancing classification and localization losses. Additionally, non-maximum suppression, as a post-processing technique, removes redundant bounding boxes from the detection results, further improving the algorithm's performance.

YOLO is an object detection system that predicts based on global information within images. Since the introduction of the initial model by Joseph Redmon, Ali Farhadi, and others in 2015, researchers in the field have iteratively updated YOLO, leading to progressively enhanced model performance. Recently, YOLOv8 has been officially released.

Developed and maintained by the small startup Ultralytics, YOLOv8 follows the footsteps of its predecessor, YOLOv5, which was also created by the same company. Positioned as a SOTA (State-of-the-Art) model, YOLOv8 builds upon the success of previous YOLO versions, introducing new features and improvements to further boost performance and flexibility. Key innovations include a novel backbone network, a new Anchor-Free detection head, and a fresh loss function designed to operate seamlessly across a spectrum of hardware platforms, ranging from CPUs to GPUs. The network architecture is illustrated in Figure 1, and module structures are detailed in Figure 2.

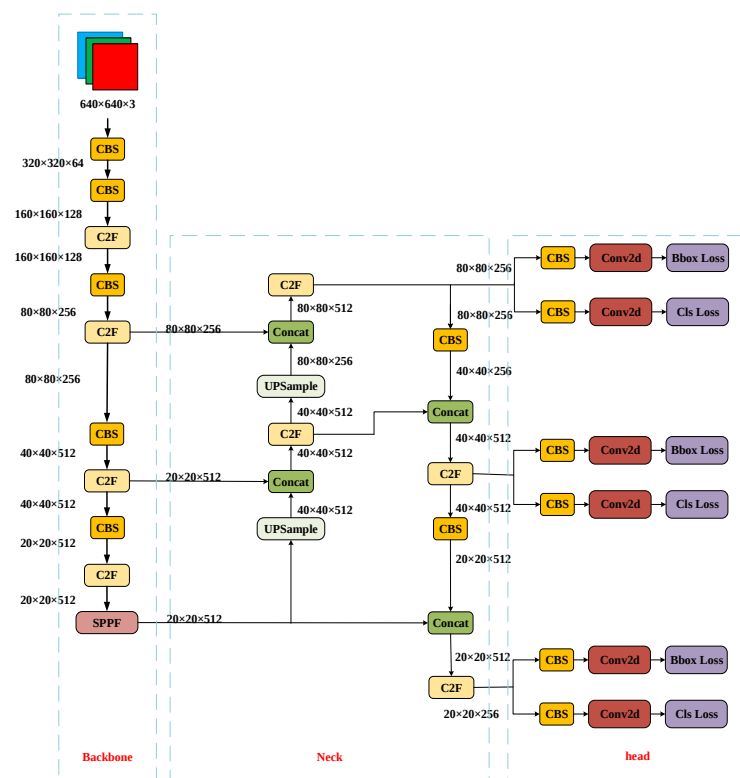


Figure 1. Structure diagram of the YOLOv8.

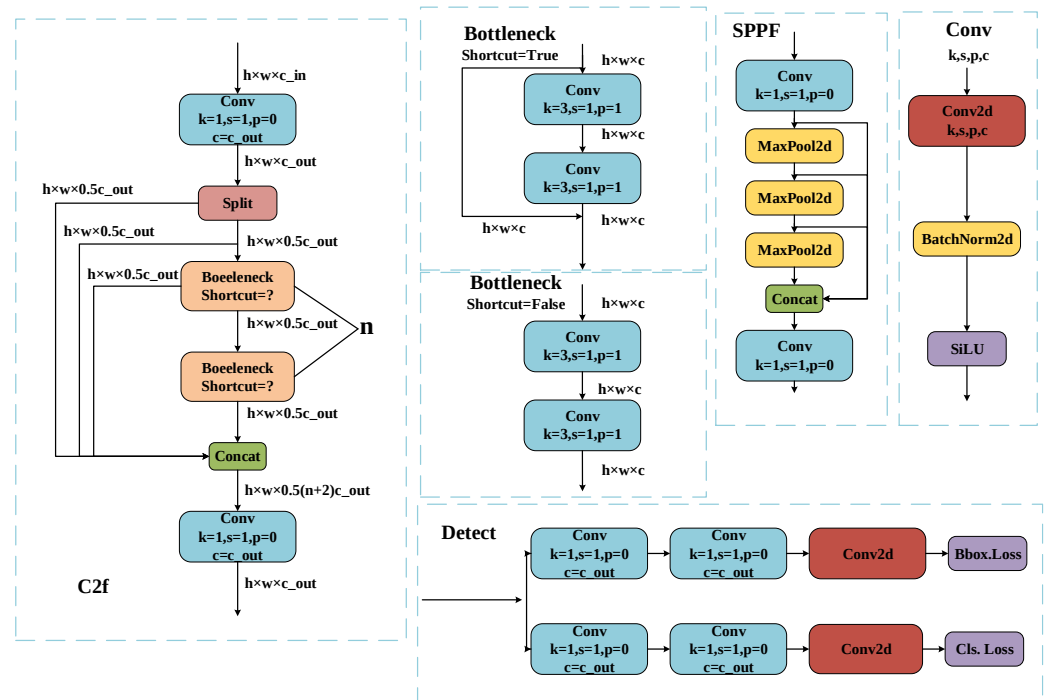


Figure 2. Detailed view of the YOLOv8 modules.

### 3.1.1. Backbone and Neck Configuration

YOLOv8 has undergone modifications in the Neck module by eliminating 2 convolutional connection layers. Notably, the kernel of the first convolutional layer has been reduced from  $6 \times 6$  to  $3 \times 3$ , and all C3 modules have been substituted with C2f. The revised structure introduces additional skip connections and incorporates supplementary Split operations. In the backbone, the number of C2f blocks has transitioned from 3-6-9-3 to 3-6-6-3. A closer examination of models of various sizes, such as N/S/M/L/X, reveals that the N/S and L/X model groups differ primarily in scaling factors. However, the channel configurations for backbone networks like S/M/L deviate, lacking adherence to a standardized set of scaling factors. Notably, the YOLOv7 network design did not adhere to a uniform application of scaling factors across all models.

### 3.1.2. Head Configuration

The Head segment experiences significant alterations, transitioning from the previously coupled structure to a decoupled arrangement. Furthermore, there is a shift from YOLOv5's Anchor-Based methodology to an Anchor-Free framework.

### 3.1.3. LOSS Configuration

The loss calculation process comprises two key components: the positive and negative sample allocation strategy and the loss computation. In the realm of modern object detection, significant attention is placed on refining the positive and negative sample allocation strategy. YOLOv5 adheres to a static allocation strategy while acknowledging the merits of dynamic allocation; YOLOv8 incorporates the TaskAlignedAssigner directly from TOOD (Task-relevant Objectness and Offset Design).

**TaskAlignedAssigner Matching Strategy:** The matching strategy employed by TaskAlignedAssigner can be succinctly described as follows: Positive samples are chosen based on the weighted score derived from classification and regression, denoted as  $t = \alpha s + \mu \beta$ . Here,  $s$  represents the predicted score corresponding to the annotated category, and  $\mu$  signifies the Intersection over Union (IoU) between the predicted box and the Ground Truth (GT) box. The product of these values quantifies the alignment degree. For each GT, all predicted boxes are assessed based on the classification score corresponding to the GT

category, and the IoU of the predicted box with GT is weighted to yield an alignment score (alignment\_metricsalignment\_metrics). Subsequently, the top K highest scores are selected as positive samples.

**Loss Calculation Components:** The loss calculation encompasses two branches: the classification branch and the regression branch, with the omission of the former objectness branch. The classification branch retains the Binary Cross-Entropy (BCE) Loss, while the regression branch integrates the Distribution Focal Loss, which utilizes the integral form proposed in the Distribution Focal Loss. Additionally, CIoU Loss is employed. These three loss components are weighted with specific proportions.

### 3.2. Improving YOLOv8 in This Paper

#### 3.2.1. Reduced Channel Spatial Object Attention Module

This section introduces the RCS-OSA (Reduced Channel Spatial Object Attention Module), denoted as RCS-OSA. The primary purpose of the RCS-OSA module is to efficiently reduce channel dimensions while enhancing spatial object attention. Strategically designed for optimal feature extraction and improved computational efficiency, the module operates by segmenting the input into two components. One segment undergoes direct propagation, while the other is subjected to processing through a stack of RCS modules. The resulting features from both direct passage and post-processing are subsequently merged after channel shuffling. This architectural element, integral to the RCS-YOLO framework, is intended to augment the model's capacity for feature handling through one-shot aggregation, maintaining computational efficiency concurrently.

The fundamental principles of RCS-OSA (Reduced Channel Spatial Object Attention) are elucidated as follows:

1. **RCS (Reparameterized Convolution based on Channel Shuffle):** This approach combines channel shuffling and reparameterized convolution to augment the network's feature extraction capabilities.
2. **RCS Module:** During the training phase, a multi-branch structure is employed to acquire diverse feature representations. In the inference phase, it transforms into a singular branch through structured reparameterization, reducing memory consumption.
3. **OSA (One-Shot Aggregation):** This methodology involves aggregating multiple features in a single iteration, diminishing the computational burden on the network and improving computational efficiency.
4. **Feature Concatenation:** The RCS-OSA module, by stacking RCS modules, ensures the reuse of features and enhances the flow of information between different layers.

Within the RCS (Reparameterized Convolution based on Channel Shuffle) module, the structure utilizes multiple branches during the training phase, encompassing  $1 \times 1$  and  $3 \times 3$  convolutions, along with a direct connection for the acquisition of rich feature representations. During the inference phase, the structure is reparameterized into a single  $3 \times 3$  convolution, reducing computational complexity and memory usage while retaining the feature expression capabilities acquired during the training phase. This design philosophy aligns closely with RCS, striving to improve computational efficiency without compromising performance.

The figure above serves as an exposition of the RCS structure, divided into the training phase (Part a) and the inference phase (Part b). During training, the input undergoes channel separation, with one portion processed through the RepVGG block (without any branching structure, using only  $3 \times 3$  convolutions, and employing ReLU as the sole activation function, as shown in Figure 3), while the other portion remains unchanged. Subsequently, the output of the RepVGG block and the unaltered input undergoes  $1 \times 1$  and  $3 \times 3$  convolutions, followed by channel shuffling and concatenation. In the inference phase, the initial multi-branch structure is streamlined into a singular  $3 \times 3$  RepConv block (the main idea of RepConv is to use multi-branch convolution layers during training and then reparameterize the parameters of the branches onto the main branch during inference, thereby reducing computational cost and memory consumption, as illustrated in Figure 3).

This design facilitates the acquisition of intricate features during training while minimizing computational complexity during inference.

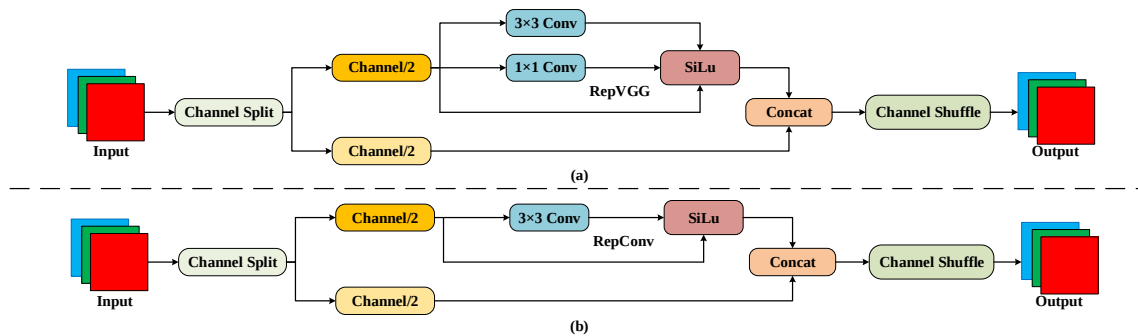


**Figure 3.** Diagram of RCS structure.

Within the RCS-OSA module (as shown in Figure 4), the input undergoes a partition into two segments: one segment experiences direct propagation, while the other undergoes processing through a stacked configuration of RCS modules. Post-processing, the features derived from the two segments are merged after channel shuffling. This design, integral to the RCS-YOLO architecture, plays a pivotal role in enhancing the model's efficiency in feature extraction and utilization through the strategy of one-shot aggregation, all while maintaining computational efficiency. The primary objective is to elevate the network's effectiveness in handling dense connections. The OSA module addresses this by representing diverse features with multiple receptive fields and aggregating all features only once in the final feature map, effectively overcoming the efficiency challenges associated with dense connections in DenseNet.

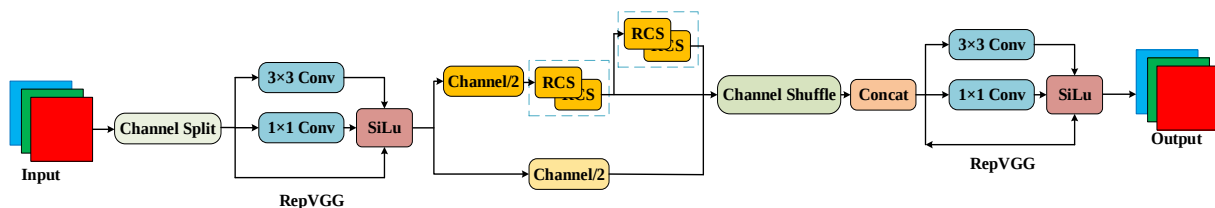


**Figure 4.** The diagram delineating the RCS-OSA structure.

### 3.2.2. Introducing Split-Processing for 3D Sparse Convolution

In this section, we present the concept of split-processing for 3D sparse convolution. The fundamental idea is to improve the efficiency of convolutional operations in a three-dimensional context, with a specific focus on sparse data. The strategy involves dividing the convolutional process into distinct stages, each designed to handle different aspects of the sparse input data. This approach not only optimizes computational resources but also capitalizes on the inherent sparsity of the data, resulting in more efficient and streamlined convolutional operations. Subsequent subsections provide a detailed exploration of the split-processing strategy, elucidating its components and the rationale underlying its application in the realm of 3D sparse convolution. SPD-Conv (Spatial-to-Depth Convolution): Improving CNN Performance for Small Targets and Low-Resolution Images.

The essential concept of SPD-Conv revolves around enhancing the performance of conventional Convolutional Neural Networks State-of-the-Art CNNs when confronted with small targets and low-resolution images. This improvement is realized through a series of fundamental steps:

**Replacement of Strided Convolution and Pooling Layers:** SPD-Conv is meticulously designed to take the place of strided convolution layers and pooling layers in traditional CNN architectures. The utilization of strided convolution and pooling layers often results in the loss of fine-grained information, particularly when processing low-resolution images or small targets.

**SPD (Spatial-to-Depth) Layer:** The SPD layer serves a critical role in downsampling the channel dimension of the feature map while preserving vital information. This strategic implementation mitigates the risk of information loss commonly encountered in traditional methods.

**Non-Strided Convolution Layer:** Following the SPD layer, SPD-Conv employs a non-strided convolution layer (i.e., with a stride of 1). This facilitates the processing of features with learnable parameters while concurrently reducing the channel count.

The structure of SPD-Conv is illustrated in Figure 5.

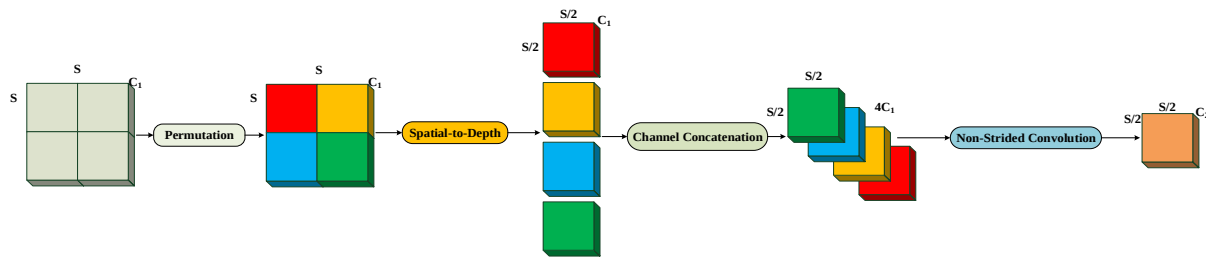


**Figure 5.** Diagram depicting SPD-Conv structure.

**Feature Map:** Traditional feature map with a channel number of  $C$ , height, and width.

**Spatial-to-Depth Transformation:** Through spatial-to-depth operations, spatial blocks of pixels are rearranged into the depth/channel dimension, increasing the channel number to 4 while reducing spatial dimensions by half.

**Channel Concatenation:** Different channel groups are merged along the channel dimension.

**Addition Operation:** The merged feature map may undergo addition operations with other processed feature maps (not detailed in the figure).

**Non-Strided Convolution:** A convolution with a stride of 1 is applied to the resulting feature map, reducing the channel dimension to  $D$  while maintaining spatial resolution, which remains half of the original size.

The SPD-Conv building block is designed to replace strided convolution and pooling layers in traditional CNNs. Strided convolution and pooling layers often result in information loss when dealing with low-resolution images and small targets. SPD-Conv utilizes an SPD layer, transforming the spatial dimensions of the feature map into depth dimensions by increasing the channel number to preserve more information. Following is a non-strided convolution layer that maintains spatial dimensions while reducing the channel count. This alternative approach avoids information loss, allowing the network to capture finer features and, consequently, enhancing performance in complex tasks.

### 3.2.3. Focal Modulation: Leveraging Focus Adjustments

This section introduces focal modulation, a technique centered around the adjustment of focus. Focal modulation is instrumental in refining the attention mechanism of the model, particularly in scenarios requiring intensified focus on specific regions. By dynamically adjusting focus during information aggregation, the model can adaptively prioritize pertinent details, maintaining sensitivity to local intricacies while concurrently augmenting awareness of global structures. Focal modulation presents a nuanced strategy for balancing the integration of contextual information and preserving sensitivity to fine-grained features. Subsequent subsections provide a detailed exploration of the foundational principles, implementation intricacies, and empirical evaluations of the proposed focal modulation methodology.

The underlying principle of FocalNets (Focal Modulation Networks [32]) revolves around the substitution of self-attention modules with a focal modulation mechanism. This strategic choice aims to effectively capture long-distance dependencies and contextual information within images. The Figure 6 offers a visual comparison between the traditional self-attention method and the innovative focal modulation approach.

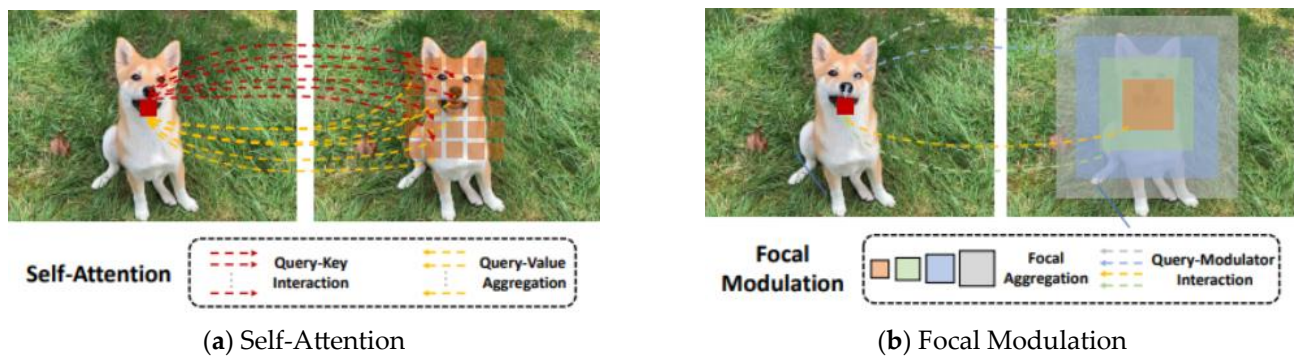


**Figure 6.** Contrastive visualization depicting Focal Modulation versus Self-Attention.

Self-attention necessitates intricate query-key interactions and query-value aggregations for each query token, which can be computationally demanding. In contrast, focal modulation streamlines these operations by initially aggregating spatial context at different granularities into modulators. Subsequently, these modulators are injected into query tokens in a query-dependent manner. This streamlined approach improves computational efficiency. In the diagram, the self-attention section is represented by red dashed lines indicating query-key interactions and yellow dashed lines indicating query-value aggregations. The focal modulation section is illustrated in blue for modulator aggregation and yellow for query-modulator interactions.

The Focal Modulation model is implemented through the following steps:

1. Focus Contextualization: Utilizing stacked depth convolutional layers to encode visual context at various scales.
2. Gate Aggregation: Employing a gating mechanism to selectively aggregate contextual information into modulators for each query token.
3. Element-Wise Affine Transformation: Injecting the aggregated modulators into each query token through element-wise affine transformations.

Focal Contextualization is an integral component of FocalNets. Focal Contextualization utilizes a series of depth-wise convolutional layers to encode visual contextual information across different scales. These layers capture visual features from proximity to distance, enabling the network to comprehend image content at various levels. Consequently, the network maintains sensitivity to local details while enhancing its understanding of global structures during the aggregation of contextual information.

A gating mechanism, commonly employed in deep learning to control information flow, is used to determine which information should be passed and which should be blocked. In recurrent neural networks, especially in LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) architectures, gating mechanisms regulate the flow of information in time-series data. In FocalNets, the purpose of gate aggregation is to selectively aggregate contextual information for each query token, i.e., the data unit under processing. This means that the network can decide which specific contextual information is crucial for the current query token, focusing on the most relevant details.

Through gate aggregation, FocalNets efficiently concentrate on the most critical information for the current task. This approach enhances model efficiency and performance by reducing the processing of unnecessary information while strengthening attention to key features. In visual tasks, this could result in improved object detection and image classification performance, particularly in complex or variable visual environments.

Via element-wise affine transformation, the modulators obtained through gate aggregation are injected into each query token. Affine transformation is a linear operation used for scaling, rotating, translating, and skewing data. In deep learning, element-wise affine transformation typically refers to a linear transformation applied to each element, described as  $y = ax + b$ , where  $x$  is the input,  $y$  is the output, and  $a$  and  $b$  are transformation parameters.

Through element-wise affine transformation, the model can finely adjust the features of each query token based on contextual information, enhancing or suppressing certain features. This precise adjustment mechanism allows the network to better adapt to complex visual scenarios, improving its ability to capture details and thereby enhancing performance in various visual tasks such as object detection and image classification.

### 3.2.4. Improved Structure of Small Target Detection Network for Vessels

The modified YOLOv8-STD network is illustrated in the figure below:

The portions with modifications are indicated by red boxes in the Figure 7. Firstly, all CBS (Cross-Stage Partial Network) modules in the network backbone, except for the input as the first CBS module, are replaced with SPDConv modules, and the C2f module in both the network backbone and feature pyramid is substituted with RCSOSA. In the last layer of the backbone, the SPPF structure is enhanced to the Focal Modulation module, resulting in the final modified YOLOv8-STD model.

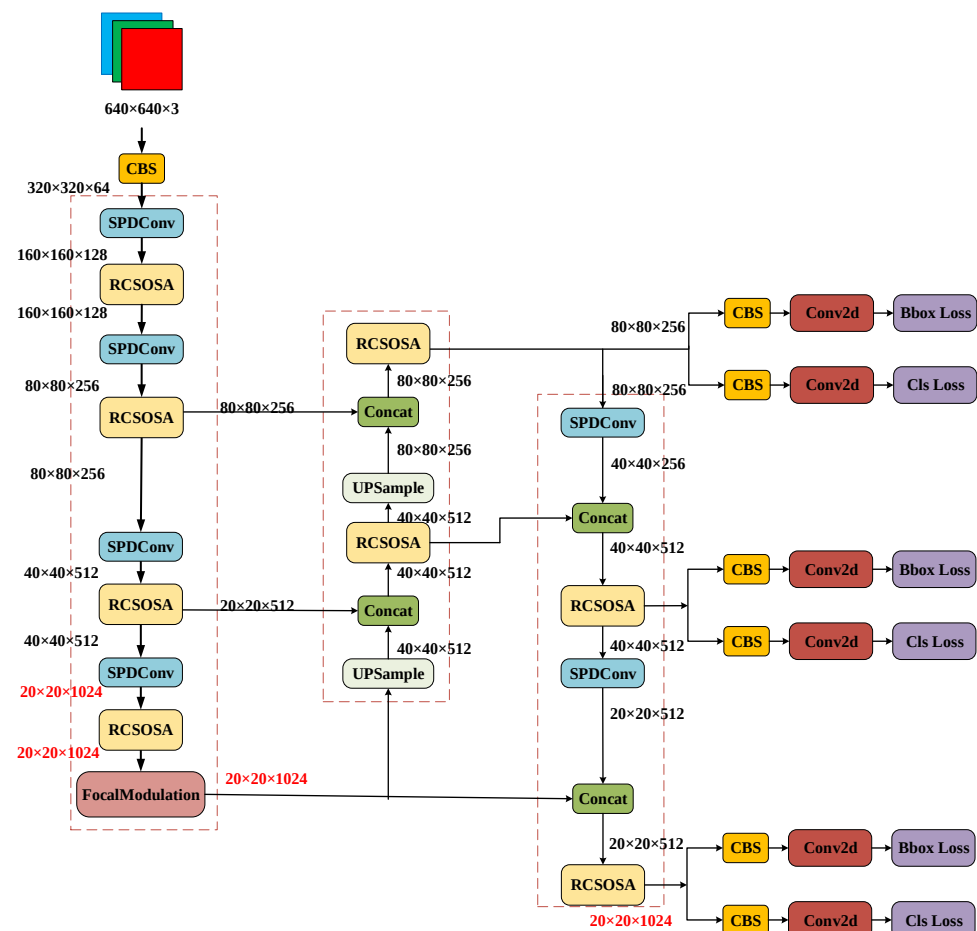


Figure 7. Structure diagram of the improved network.

## 4. Experiments

In order to affirm the efficacy and superiority of the proposed model in a challenging water surface detection environment, the platform and experimental parameters have been set up as detailed below.

### 4.1. Experimental Environment and Parameter Setting

The platform of this experiment is as follows in Table 1:

Table 1. Experimental configuration.

| Items                   | Type                          |
|-------------------------|-------------------------------|
| operating system        | ubuntu 22.04                  |
| programming language    | python 3.9                    |
| deep learning framework | torch 2.0, cuda11.8, cudnn8.8 |
| CPU                     | Intel i5 9600K 3.7 GHz 6 core |
| GPU                     | RTX2080 8G                    |

The experimental parameters are set as shown in the following Table 2:

Table 2. Experimental Parameter Setting.

| Parameter     | Configuration |
|---------------|---------------|
| learning rate | 0.01          |
| momentum      | 0.937         |
| weight decay  | 0.0005        |
| batch size    | 16            |
| optimizer     | SGD           |
| image size    | 640 × 640     |
| epochs        | 100           |

In order to enhance the generalization capability of the object detection model, it is common practice to apply data augmentation techniques before training neural networks. Commonly employed methods include scaling, translation, rotation, and color transformation. Given the intricate working environment of unmanned vessels, this study incorporates scaling, translation, color variation, and the mosaic data augmentation technique [20] to introduce diversity to the dataset. To avoid potential decreases in recognition accuracy resulting from symmetrical relationships among digits, this study abstains from utilizing flip-based data augmentation. The mosaic data augmentation method involves the random cropping and scaling of four images within the dataset, followed by a random arrangement and concatenation to create a single image. In cases where small images or blank areas are present, they are complemented with gray color to ensure consistency with the original input size. The visual representation of this effect is presented in Figure 8.

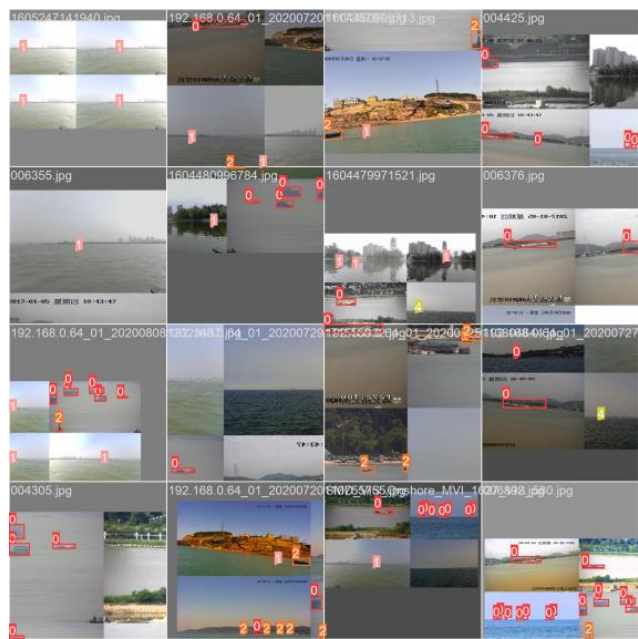


Figure 8. Mosaic enhancement diagram.

#### 4.2. Introduction to USV and Datasets

At present, there is a lack of extensive publicly available datasets suitable for the detection of small targets in maritime environments. Since individual datasets cover limited scenarios, the effectiveness of training may not adequately reflect the model's capacity for generalization. This study addresses this limitation by integrating frames and eliminating duplicates from a self-compiled dataset, incorporating the publicly available WSODD (Water Surface Object Detection Dataset) [33] and the FloW dataset. The dataset encompasses a diverse range of realistic and complex scenarios, including backlit conditions, fog, waves, target clustering, and background interference.

The evaluation of the performance of data-driven deep neural network algorithms is typically conducted on large public datasets. However, there is currently no reliable large dataset available for validation, with the only available small-scale datasets sourced from fixed cameras installed in ports. In comparison to fixed cameras, unmanned vessel cameras have lower video perspectives, are prone to image instability during high-speed motion, and are more susceptible to adverse weather conditions such as rain and snow. Consequently, these datasets are not directly applicable in the field of unmanned vessels. To obtain more realistic and reliable data, this study utilized the electro-optical payload equipment carried by unmanned vessels to collect data, simulating real-world scenarios in maritime areas such as the South China Sea, Bohai Sea, and Weihai.

Figure 9 illustrates the composition of the dataset, which includes images sourced from the FloW dataset, the WSODD [33], and authentic images collected through an optical-electric payload mounted on unmanned vessels. After a rigorous manual filtering process to eliminate duplicates and subpar images, the curated dataset comprises 3958 images featuring annotations for 9250 targets. To assess the model's robustness, the dataset encompasses various target vessels, each characterized by distinctive markings at different positions and diverse visual attributes.

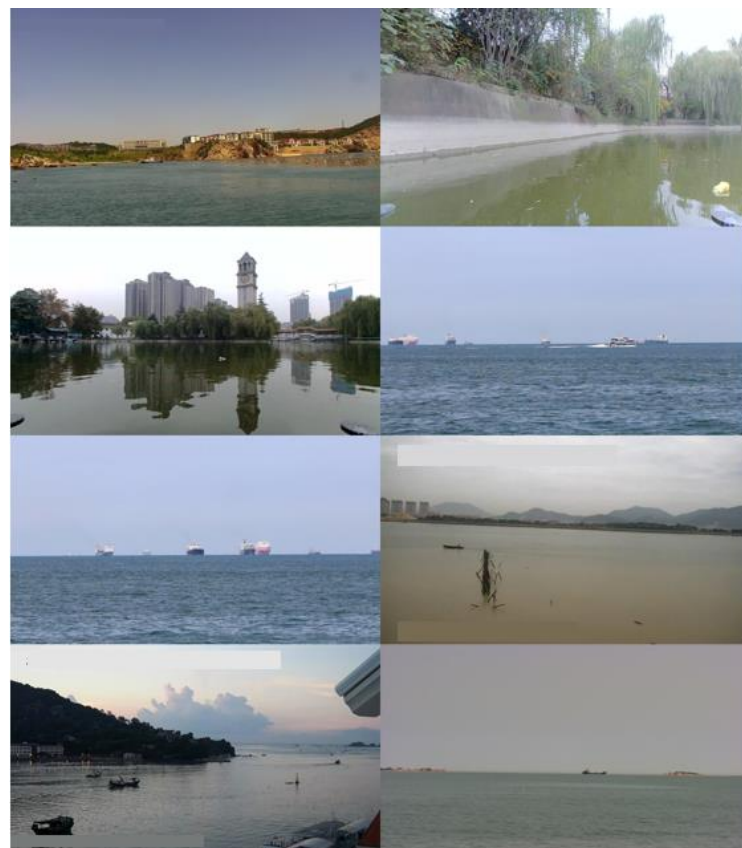


Figure 9. Illustration showcasing partial dataset examples.

### 4.3. Evaluation Metrics

Precision ( $P$ ) is stipulated as the ratio of accurately detected positive samples to the total number of positive samples detected simultaneously. A higher precision signifies a decreased likelihood of false target detection, hence the term “accuracy”. Recall ( $R$ ) is expressed as the ratio of correctly detected positive samples to the total number of positive samples. The formulas for accuracy and recall are provided in Equations (1) and (2), respectively:

$$P = \frac{TP}{TP + FP} \quad (1)$$

$$R = \frac{TP}{TP + FN} \quad (2)$$

where  $P$  denotes the accuracy rate and  $R$  denotes the recall rate. The above formulae can be used to obtain the values of accuracy and recall at different thresholds, and the P-R curve is plotted. The area enclosed by the P-R curve and the coordinate axis is the average accuracy ( $AP$ ), and its calculation formula is shown in Equation (3):

$$AP = \int_0^1 P(R)d(R) \quad (3)$$

In real-world applications, the utilization of integration for obtaining average accuracy introduces additional complexities. Consequently, the interpolation sampling method is often preferred to compute the average value due to its convenience, as outlined in Equation (4):

$$AP = \frac{1}{11} \sum P(R), R \in 0, 0.1, 0.2, \dots, 0.95 \quad (4)$$

To gauge the lightweight nature of the model, the experiments will analyze the number of parameters in the network model and the GFLOPs (the quantity of floating-point operations). These metrics exhibit a negative correlation with the model’s lightweight characteristics. A model with lower values for these parameters is deemed more suitable for efficient deployment on USVs.

mAP, short for “Mean Average Precision”, typically refers to the average of the Average Precision (AP) values for all categories within all images. It is commonly used to evaluate the performance of algorithms in object detection tasks. The term “IoU = 0.5@mAP” specifically indicates the mean average precision value at an IoU threshold of 0.5.

## 5. Experimental Results and Analysis

The training results and statistics for mAP (Mean Average Precision) are summarized in Table 3. It is evident from the table that the method introduced in this paper registers an increase in mAP for each target category in comparison to the baseline YOLOv8 model.

**Table 3.** Comparison of mAP before and after improvement.

| Class     | mAP@.5 Ours/Baseline% | mAP@.5:.95 Ours/Baseline% |
|-----------|-----------------------|---------------------------|
| all       | 77.6/74.0             | 41.1%/39.0%               |
| ship      | 90.4/90.2             | 57.4/56.6                 |
| Floater   | 88.9/87.6             | 50.3/47.0                 |
| boat      | 62.0/60.8             | 29.3/28.7                 |
| buoy      | 78.3/79.5             | 38.4/40.3                 |
| speedboat | 71.5/52.1             | 30.6/22.5                 |

The comparison of precision-recall curves between the methodology presented in this paper and the baseline model is illustrated in Figures 10–13. The discernible larger coverage area of the curve for the method in this paper, as opposed to the baseline model, suggests a heightened level of accuracy associated with the approach presented in this study.

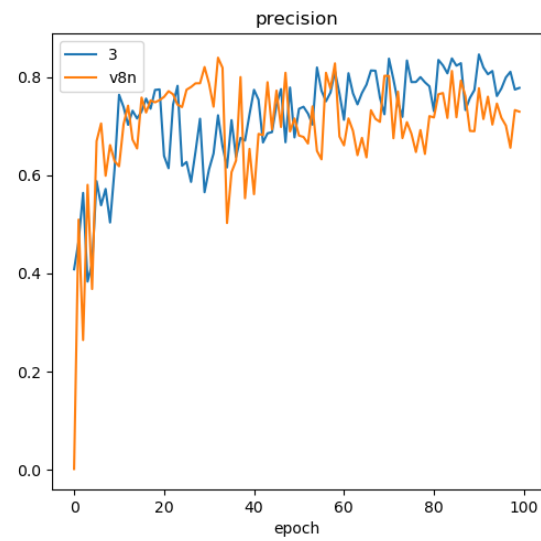


**Figure 10.** Comparison of the Precision curves.

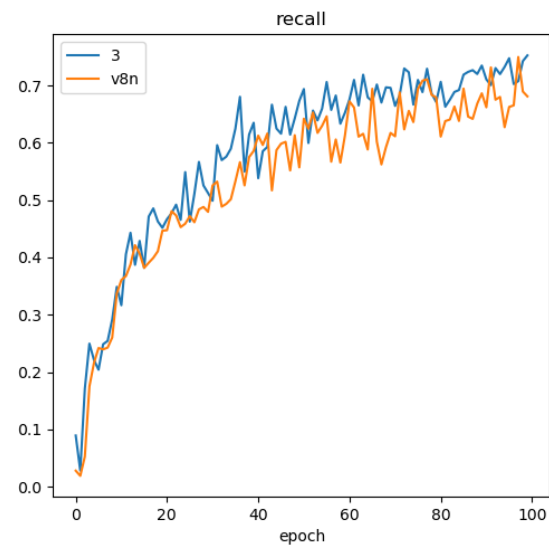


**Figure 11.** Comparison of the Recall curves.

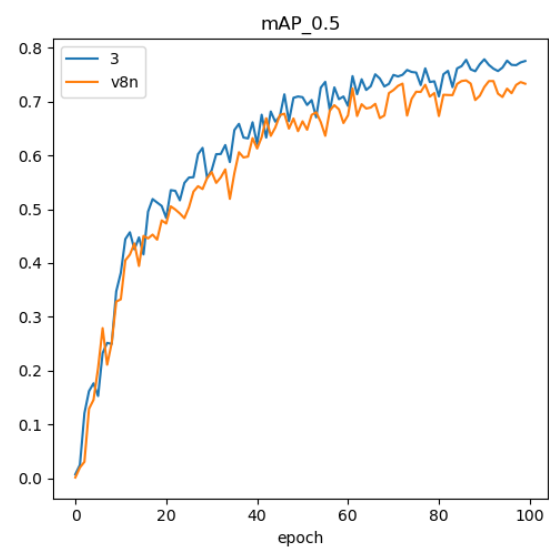


**Figure 12.** Comparison of the mAP@.5 curves.

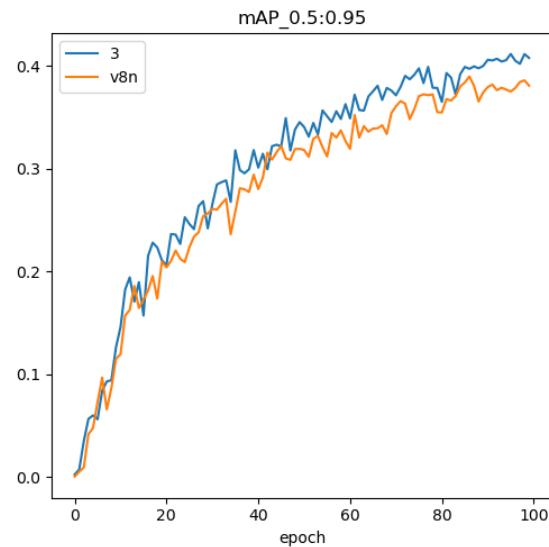


**Figure 13.** Comparison of the mAP@.5:.95 curves.

Figure 14 delineates the trajectory of training results, indicating a more pronounced reduction in loss over the course of 100 training epochs. Importantly, this approach significantly amplifies the recall rate. Consequently, our methodology not only advances accuracy but also assimilates the characteristics of surface targets with greater efficacy, thereby diminishing the prospect of oversight and enabling the discernment of additional targets in intricate aquatic environments.

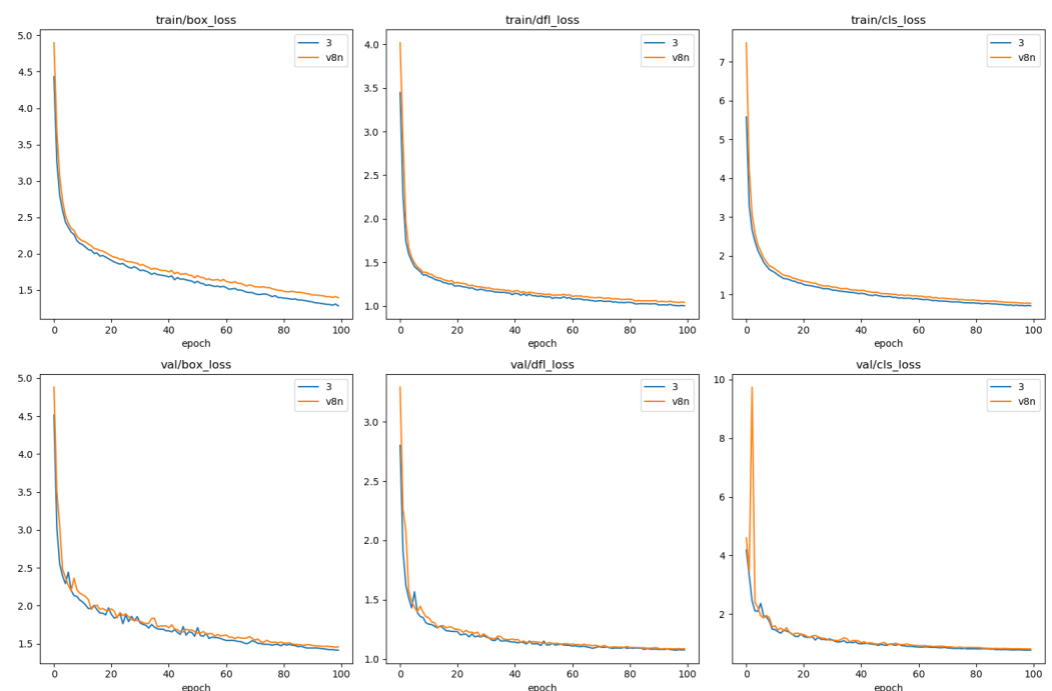


**Figure 14.** Comparison of loss function.

Leveraging a proprietary dataset, we assess the viability of the proposed network enhancements and their effectiveness in detecting surface vessels. Precision and recall curves are depicted in Figures 10 and 11, while Figure 12 showcases the convergence curve for the mAP metric [22]. Figure 14 tracks the evolution of loss values during the training process. Specific precision values post-iteration are detailed in Table 3. Examination of Figure 14 reveals that our algorithm achieves faster convergence and lower loss during

training compared to the original algorithm. Figure 10 demonstrates a higher precision for our algorithm, and Figure 11 exhibits a significantly improved recall, indicating lower omission rates and increased sensitivity to small-sized targets. According to Figure 13, our proposed method outperforms the original YOLO algorithm in terms of detection accuracy. Quantitatively, Table 4 emphasizes the enhancements of our method compared to the original approach: a 3.6% increase in precision, a 10.1% boost in recall, and a 2.1% enhancement in the overall mAP metric [22].

**Table 4.** Comparison of hull number recognition by original and improved algorithms.

| Method     | Precision | Recall | mAP@.5:.95 |
|------------|-----------|--------|------------|
| YOLOv8     | 0.74      | 0.646  | 0.39       |
| YOLOv8-STD | 0.776     | 0.747  | 0.411      |

In addition to this, the study includes a thorough investigation of the model's detection time. The average detection time per frame is recorded at 2.6 ms. With a video stream input of 30 frames per second from the unmanned vessel's electro-optical pod, the model achieves real-time detection capabilities.

## 6. Verification on Actual Vessels

To validate the feasibility of the proposed algorithm on USV platforms, it was deployed on the USV platform depicted in Figure 9. The interface of the perception system's operational environment is illustrated in Figure 13. To quantitatively demonstrate the system's capabilities, this paper conducted on-site tests in the Weihai Sea area. Under sea condition level three, the tested vessel approached the USVs at a cruising speed of 10 knots from a distance of 4 km. The USV ceased its operation upon the successful identification of the target buoy.

The USV platform features a 7.5 m workboat with protective fenders floating externally (as shown in Figure 15), ensuring excellent navigability and safety. The use of a deep-V hull design meets strict requirements for key performance factors such as speed, stability (steady maneuvering), navigability, and safety. The seamless integration of hardware components serves as a fundamental assurance for the smooth validation of the unmanned platform. Based on the existing spatial structure of the unmanned platform and requirements for remote control devices, autonomous equipment, power supply for mission payloads, and connectivity, the spatial layout planning of hardware devices and circuit connections for the unmanned platform was completed. Equipped with optical, radar, sonar, and other devices, the platform possesses autonomous driving capabilities, as shown in Figure 16.



**Figure 15.** 7.5 m USVs Platform Diagram.

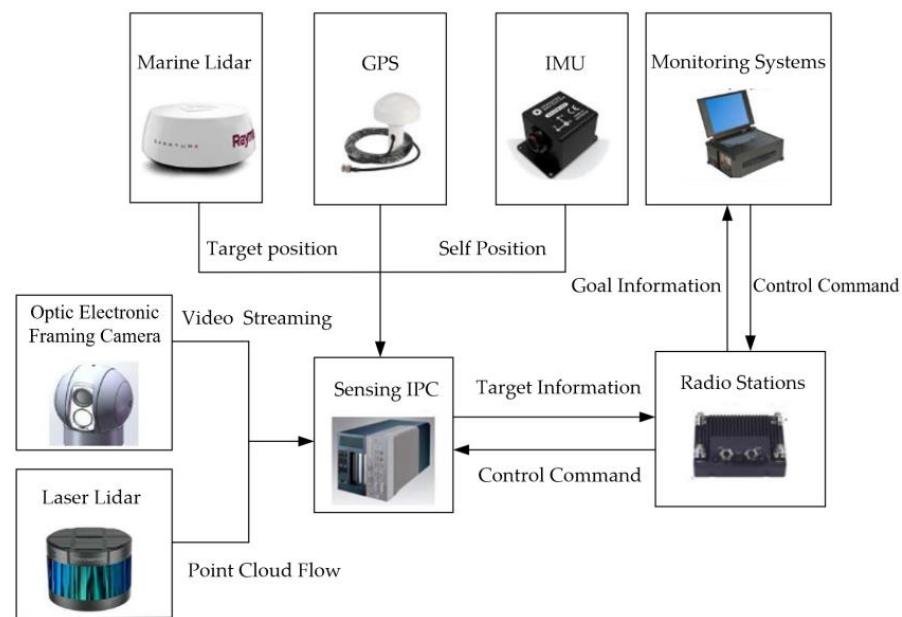


**Figure 16.** Hardware structure diagram of the sensing system of the USV platform.

The radar captures azimuthal data, encompassing direction angles and distances, and this data is subsequently fed into the perception industrial control computer to derive latitude and longitude information, as shown in Figure 17. Figure 18 is a radar interface diagram.

Utilizing information output by the perception industrial control computer, the optical-electric system calculates the deviation from the target. Following this, the optical-electric camera tracks the target in real time, employing the YOLO network framework to detect and identify objects within the input video stream, as shown in Figure 19.

```

【2023-11-14 20:54:54:581】FC FD 54 21 9B 3B 71 22 A9 1D 00 00 A2 1A 00 00
DA 00 00 00 C4 40 CC 01 C4 40 CC 01 00 01 00 00 2E 00 00 00 00 00 99 CA
A8 5F 00 00 00 00 00 00 00 00 03 31 31 32
【2023-11-14 20:54:54:581】FC FD 5D 21 9B 3B 31 3C 0B 0E 00 00 83 05 00 00
94 F4 FF FF CE 18 2C 01 CE 18 2C 01 00 01 00 00 1F 00 00 00 00 00 99 CA
A8 5F 00 00 00 00 00 00 00 00 03 31 31 32
【2023-11-14 20:54:54:582】FC FD 5F 21 9B 3B DC 64 3C 92 00 00 6A 7F FF FF
F5 E4 FF FF 36 42 98 03 36 42 98 03 00 01 00 00 8B 00 00 00 00 00 99 CA
A8 5F 00 00 00 00 00 00 00 00 03 31 31 32
【2023-11-14 20:54:54:582】FC FD 4F 21 9B 3B E4 12 79 0F 00 00 64 0A 00 00
4C 09 00 00 46 73 C8 00 46 73 C8 00 00 01 00 00 41 00 00 00 00 00 99 CA
A8 5F 00 00 00 00 00 00 00 00 03 31 31 32
【2023-11-14 20:54:54:582】FC FD 61 21 9B 3B F5 72 18 2B 00 00 B8 DC FF FF
09 10 00 00 E8 49 E0 01 E8 49 E0 01 00 01 00 00 3A 00 00 00 00 00 99 CA
A8 5F 00 00 00 00 00 00 00 00 03 31 31 32
【2023-11-14 20:54:54:582】FC FD 45 21 9B 3B FB 63 25 89 00 00 7B 88 FF FF
E9 E1 FF FF 1C 43 F8 02 1C 43 F8 02 00 01 00 00 49 00 00 00 00 00 99 CA
A8 5F 00 00 00 00 00 00 00 00 03 31 31 32
【2023-11-14 20:54:54:582】FC FD 5E 21 9B 3B F3 4F 93 08 00 00 C9 FC FF FF
F4 F8 FF FF B8 42 FA 00 B8 42 FA 00 00 01 00 00 2D 00 00 00 00 00 99 CA
A8 5F 00 00 00 00 00 00 00 00 03 31 31 32
【2023-11-14 20:54:54:582】FC FD 4E 21 9B 3B BF 65 9C A1 00 00 D3 70 FF FF
D0 E7 FF FF 2A 44 CA 03 2A 44 CA 03 00 01 00 00 41 00 00 00 00 00 99 CA
A8 5F 00 00 00 00 00 00 00 00 03 31 31 32

```

**Figure 17.** Radar Message Diagram.

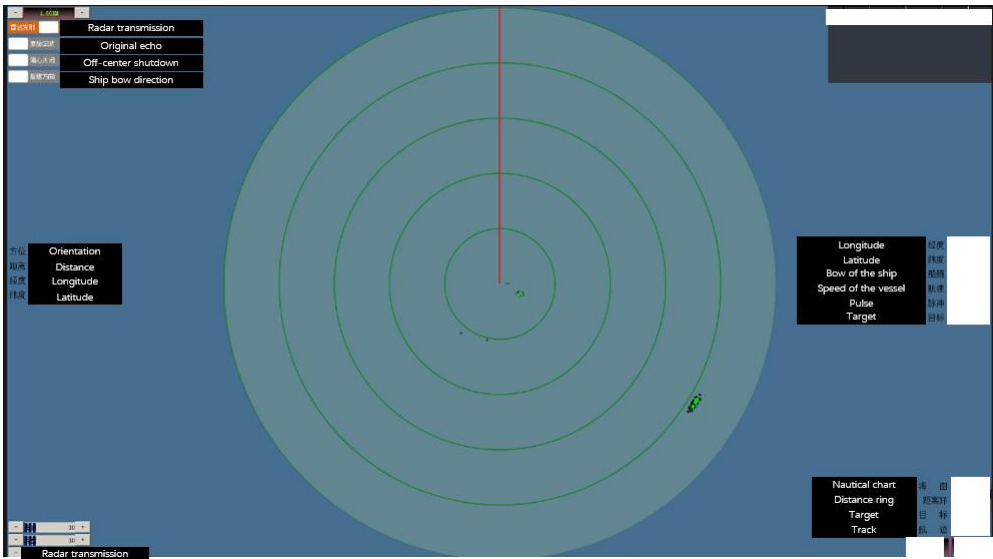


Figure 18. Radar Interface Diagram.



Figure 19. Video Stream Detection Image.

The interface of the perception industrial control computer is shown in Figure 20.

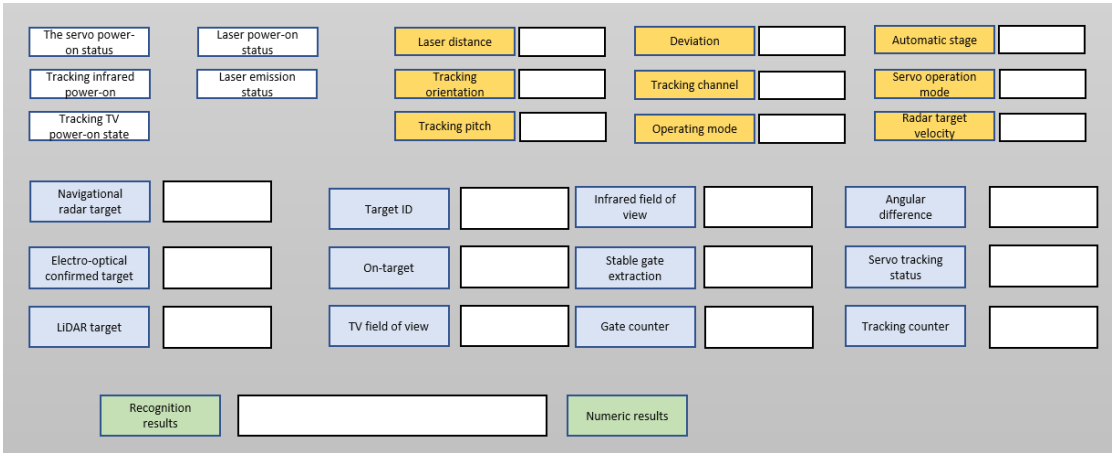


Figure 20. The interface of the perception industrial control computer.

The Figure 21 displays a stationary fishing vessel serving as a fixed target, with the unmanned surface vehicle executing straight-line motion at a speed not less than 30 knots. The diagram reveals a distance surpassing 4 km between the target and the USV. The laser's return values indicate accurate and stable gate positioning derived from image recognition, ensuring real-time performance in line with the requirements for optical-electric tracking.



**Figure 21.** Target Localization Test Results.

## 7. Conclusions and Discussion

With the rapid advancement of deep learning technology, the field of water surface target detection is garnering increasing attention in both academic and industrial spheres. This study delves into the fundamental characteristics of small target detection in maritime environments and proposes an innovative approach utilizing a one-stage object detection model for real-time detection of small targets on water surfaces. In this approach, we comprehensively improve the one-stage object detection model, covering optimizations in network architecture, reparameterization of convolution modules, enhancements in downsampling architecture, and the application of spatial pyramid pooling. Experimental validation reveals that our proposed improvements lead to a 2.1% increase in mean average precision compared to the original model, effectively addressing the challenge of small target recognition under complex unmanned vessel perspectives.

Further on-site validation demonstrates the method's capability to accurately identify and localize buoys with a diameter of 1.5 m within a 0.5 km radius, as well as precisely recognize moving vessels with a length of 10 m within a 3 km range. It is noteworthy that the method exhibits outstanding robustness, showing insensitivity to the complexities of the water surface environment and demonstrating excellent adaptability.

The enhancement of small target detection confers manifold benefits upon unmanned surface vessels, directly impacting their performance and efficiency across diverse tasks and environments. Firstly, the improvement in small target detection significantly enhances the navigation and obstacle avoidance capabilities of unmanned surface vessels. By promptly detecting and accurately locating small targets on the water surface, unmanned vessels can more effectively evade obstacles, reducing the likelihood of collisions and unforeseen incidents, thus enhancing navigational safety.

Secondly, the enhancement in small target detection contributes to bolstering the monitoring and surveillance capabilities of unmanned surface vessels. In tasks such as

maritime surveillance, environmental monitoring, and border patrols, timely detection and tracking of small targets (such as floating objects, lifeboats, and small vessels) play a crucial role in safeguarding maritime security, environmental protection, and resource conservation. Equipped with efficient small target detection systems, unmanned surface vessels can execute these tasks more effectively, providing timely and accurate information and data support.

Furthermore, the improvement in small target detection also elevates the emergency response capabilities of unmanned surface vessels. In situations involving emergency events and search and rescue operations, rapid detection and localization of small targets on the water surface can greatly enhance the efficiency of rescue efforts, shorten response times, and minimize casualties and property losses to the greatest extent possible.

In summary, the enhancement of small target detection confers benefits upon unmanned surface vessels, including strengthened navigation and obstacle avoidance capabilities, reinforced monitoring and surveillance capabilities, and enhanced emergency response capabilities. These benefits broaden and enhance the application of unmanned surface vessels in the maritime domain, providing robust support and safeguarding for maritime tasks.

This study emphasizes the feasibility and practicality of our proposed single-model approach for small target detection in unmanned surface vessel scenarios. Additionally, the significance of this research extends to various domains, including maritime accident retracing, emergency response, and target statistics. Moreover, the method holds critical importance in both military and civilian domains. Future research directions encompass increasing the intra-class diversity of datasets, expanding dataset size, and exploring lightweight methods to reduce network architecture parameters under various constraints, aiming to address the limitations of computing resources and device capabilities.

Further research into advancing small target detection focuses on enhancing the effectiveness and versatility of USVs across various operational scenarios. To achieve this, optimizing detection accuracy, particularly in challenging environmental conditions such as adverse weather, varying lighting, and complex backgrounds, requires refining algorithms and employing advanced sensor fusion techniques and machine learning integration for real-time adaptation to dynamic environmental factors [34–37]. Future studies should delve into developing robust and versatile hardware and sensor systems specifically tailored for small target detection on USVs. This involves designing and implementing multi-modal sensor arrays capable of capturing various data types, including visual, thermal, and acoustic signals, to enhance target discrimination and localization capabilities in diverse operating environments.

Moreover, research efforts should explore integrating small target detection capabilities with autonomous navigation and decision-making systems on USVs. This integration with advanced artificial intelligence and machine learning enables USVs to autonomously detect and respond to small targets in real time, enhancing navigation efficiency and adaptive mission execution in complex maritime environments.

Additionally, there is a need for further studies to validate and evaluate the performance of small target detection systems in real-world operational scenarios. This includes conducting field trials and validation tests in diverse maritime environments to assess the robustness, reliability, and effectiveness of small target detection algorithms and systems under practical conditions.

Furthermore, research can focus on developing standard evaluation metrics and benchmark datasets to assess the performance of small target detection algorithms and systems. This will facilitate comparative studies and enable researchers to objectively evaluate and compare the efficacy of different detection methods and technologies.

**Author Contributions:** Conceptualization, J.Z. (Jian Zhang) and J.Z. (Jiayuan Zhuang); methodology, J.Z. (Jian Zhang); software, J.Z. (Jian Zhang); validation, J.Z. (Jiayuan Zhuang), W.H., R.Z. and X.D.; formal analysis, J.Z. (Jian Zhang) and J.Z. (Jiayuan Zhuang); investigation, J.Z. (Jian Zhang); resources, J.Z. (Jiayuan Zhuang); data curation, W.H.; writing—original draft preparation, J.Z. (Jian Zhang);

writing—review and editing, J.Z.; visualization, J.Z. (Jian Zhang); supervision, J.Z. (Jiayuan Zhuang) and W.H.; project administration, J.Z. (Jiayuan Zhuang); funding acquisition, J.Z. (Jiayuan Zhuang) All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the Heilongjiang Provincial Excellent Youth Fund, grant number YQ2021E013, and The National Key Research and Development Program of China, grant number 2021YFC2803400.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data is unavailable due to privacy restrictions.

**Conflicts of Interest:** The authors declare no conflicts of interest.

## References

1. Liu, Z.; Zhang, Y.; Yu, X.; Yuan, C. Unmanned surface vehicles: An overview of developments and challenges. *Annu. Rev. Control* **2016**, *41*, 71–93. [[CrossRef](#)]
2. Campbell, S.; Naeem, W.; Irwin, G.W. A review on improving the autonomy of unmanned surface vehicles through intelligent collision avoidance manoeuvres. *Annu. Rev. Control* **2012**, *36*, 267–283. [[CrossRef](#)]
3. Huang, B.; Zhou, B.; Zhang, S.; Zhu, C. Adaptive prescribed performance tracking control for underactuated autonomous underwater vehicles with input quantization. *Ocean. Eng.* **2021**, *221*, 108549. [[CrossRef](#)]
4. 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; IEEE Press: Piscataway, NJ, USA, 2014; pp. 580–587.
5. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE Press: Piscataway, NJ, USA, 2015; pp. 1440–1448.
6. 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](#)] [[PubMed](#)]
7. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE Press: Piscataway, NJ, USA, 2016; pp. 770–778.
8. 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 (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE Press: Piscataway, NJ, USA, 2017; pp. 2261–2269.
9. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE Press: Piscataway, NJ, USA, 2017; pp. 936–944.
10. Fu, C.Y.; Liu, W.; Ranga, A.; Tyagi, A.; Berg, A.C. DSSD: Deconvolutional single shot detector. *arXiv* **2017**, arXiv:1701.06659.
11. 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*; Springer international Publishing: Cham, Switzerland, 2016; pp. 21–37.
12. Bell, S.; Zitnick, C.L.; Bala, K.; Girshick, R. Inside-outsidenet: Detecting objects in context with skip pooling and recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE Press: Piscataway, NJ, USA, 2016; pp. 2874–2883.
13. Chen, C.; Liu, M.Y.; Tuzel, O.; Xiao, J. R-CNN for small object detection. In *Computer Vision-ACCY 2016*; Springer International Publishing: Cham, Switzerland, 2017; pp. 214–230.
14. Cai, Z.; Fan, Q.; Feris, R.S.; Vasconcelos, N. A unified multiscale deep convolutional neural network for fast object detection. In *Computer Vision-ECCV 2016*; Springer International Publishing: Cham, Switzerland, 2016.
15. Krishna, H.; Jawahar, C.V. Improving small object detection. In Proceedings of the 2017 4th IAPR Asian Conference on Pattern Recognition (ACPR), Nanjing, China, 26–29 November 2017; IEEE Press: Piscataway, NJ, USA, 2017; pp. 340–345.
16. Wang, J.; Chen, K.; Yang, S.; Loy, C.C.; Lin, D. Region proposal by guided anchoring. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE Press: Piscataway, NJ, USA, 2019; pp. 2960–2969.
17. Singh, B.; Davis, I.S. An analysis of scale invariance in object detection-SNIP. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE Press: Piscataway, NJ, USA, 2018; pp. 3578–3587.
18. Hu, X.; Xu, X.; Xiao, Y.; Chen, H.; He, S.; Qin, J.; Heng, P.A. SLNet: A scale-in-sensitive convolutional neural network for fast vehicle detection. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 1010–1019. [[CrossRef](#)]
19. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE Press: Piscataway, NJ, USA, 2017; pp. 2980–2988.

20. Kong, T.; Yao, A.; Chen, Y.; Sun, F. Hyper Net: Towards accurate region proposal generation and joint object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; IEEE Press: Piscataway, NJ, USA, 2016; pp. 845–853.
21. Ghodrati, A.; Diba, A.; Pedersoli, M.; Tuytelaars, T.; Van Gool, L. Deep Proposal: Hunting objects by cascading deep convolutional. In Proceedings of the IEEE International Conference computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; IEEE Press: Piscataway, NJ, USA, 2015; pp. 2578–2586.
22. Shao, Z.Y.; Liu, H.; Yin, Y.; Cheng, T.; Gao, X.W.; Zhang, W.J.; Jing, Q.F.; Zhao, Y.J.; Zhang, L.P. Multi-Scale Object Detection Model for Autonomous Ship Navigation in Maritime Environment. *J. Mar. Sci. Eng.* **2022**, *10*, 1783. [\[CrossRef\]](#)
23. Cai, Z.W.; Vasconcelos, N. Cascade R-CNN: Delving into high quality object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; IEEE Press: Piscataway, NJ, USA, 2018; pp. 6151–6162.
24. Cui, L.; Ma, R.; Lv, P.; Jiang, X.; Gao, Z.; Zhou, B.; Xu, M. MDSSD: Multi-scale deconvolutional single shot detector for small objects. *Sci. China Inf. Sci.* **2020**, *63*, 120113. [\[CrossRef\]](#)
25. Rekavandi, A.M.; Rashidi, S.; Boussaid, F.; Hoefs, S.; Akbas, E. Transformers in Small Object Detection: A Benchmark and Survey of State-of-the-Art. *arXiv* **2023**, arXiv:2309.04902.
26. Li, Z.; Zhou, F. FSSD: Feature fusion single shot Multi-box detector. *arXiv* **2018**, arXiv:1712.00960.
27. Li, J.; Liang, X.; Wei, Y.; Xu, T.; Feng, J.; Yan, S. Perceptual generative adversarial networks for small object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; IEEE Press: Piscataway, NJ, USA, 2017; pp. 1951–1959.
28. Bai, Y.; Zhang, Y.; Ding, M.; Ghanem, B. SOD-MT-GAN: Small object detection vin multi-task generative adversarial network. In *Computer Vision-ECCV 2018*; Springer: Berlin/Heidelberg, Germany, 2018.
29. Noh, J.; Bae, W.; Lee, W.; Seo, J.; Kim, G. Better to follow, follow to be better: Towards precise supervision of feature super-resolution for small object detection. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; IEEE Press: Piscataway, NJ, USA, 2019; pp. 9724–9733.
30. Kisantal, M.; Wojna, Z.; Murawski, J.; Naruniec, J.; Cho, K. Augmentation for small object detection. In Proceedings of the 9th International and information Conference on Advances in Computing and Information Technology (ACITY 2019), Sydney, Australia, 21–22 December 2019.
31. Chen, C.; Zhang, Y.; Lv, Q.; Wei, S.; Wang, X.; Sun, X.; Dong, J. RRNet: A hybrid detector for object detection in drone-captured images. In Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), Seoul, Republic of Korea, 27–28 October 2019; IEEE Press: Piscataway, NJ, USA, 2019; pp. 100–108.
32. Yang, J.; Li, C.; Dai, X.; Gao, J. Focal Modulation Networks. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 4203–4217. [\[CrossRef\]](#)
33. Zhou, Z.G.; Sun, J.E.; Yu, J.B.; Liu, K.Y.; Duan, J.W.; Chen, L.; Chen, C.L.P. An Image-Based Benchmark Dataset and a Novel Object Detector for Water Surface Object Detection. *Front. Neurobotics* **2021**, *15*, 723336. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Huang, B.; Song, S.; Zhu, C.; Li, J.; Zhou, B. Finite-time distributed formation control for multiple unmanned surface vehicles with input saturation. *Ocean. Eng.* **2021**, *233*, 109158. [\[CrossRef\]](#)
35. Zhou, B.; Huang, B.; Su, Y.; Wang, W.; Zhang, E. Two-layer leader-follower optimal affine formation maneuver control for networked unmanned surface vessels with input saturations. *Int. J. Robust. Nonlinear Control.* **2023**, *34*, 3631–3655. [\[CrossRef\]](#)
36. Huang, B.; Zhang, S.; He, Y.; Wang, B.; Deng, Z. Finite-time anti-saturation control for Euler–Lagrange systems with actuator failures. *ISA Trans.* **2022**, *124*, 468–477. [\[CrossRef\]](#) [\[PubMed\]](#)
37. Zhou, B.; Huang, B.; Su, Y.; Zhu, C. Interleaved periodic event-triggered communications based distributed formation control for cooperative unmanned surface vessels. *IEEE Trans. Neural Netw. Learn. Syst.* **2024**, *10*, 123–135. [\[CrossRef\]](#) [\[PubMed\]](#)

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.