mathematics

MDPI

Article

A Kind of Water Surface Multi-Scale Object Detection Method Based on Improved YOLOv5 Network

Zhongli Ma, Yi Wan, Jiajia Liu *, Ruojin An and Lili Wu

College of Automation, Chengdu University of Information Technology, Chengdu 610103, China;
mazl@cuit.edu.cn (Z.M.); wy17062920@163.com (Y.W.); an1587601@163.com (R.A.); 15082386704@163.com (L.W.)
* Correspondence: liujj@cuit.edu.cn

Abstract: Visual-based object detection systems are essential components of intelligent equipment for water surface environments. The diversity of water surface target types, uneven distribution of sizes, and difficulties in dataset construction pose significant challenges for water surface object detection. This article proposes an improved YOLOv5 target detection method to address the characteristics of diverse types, large quantities, and multiple scales of actual water surface targets. The improved YOLOv5 model optimizes the extraction of bounding boxes using K-means++ to obtain a broader distribution of predefined bounding boxes, thereby enhancing the detection accuracy for multi-scale targets. We introduce the GAMAttention mechanism into the backbone network of the model to alleviate the significant performance difference between large and small targets caused by their multi-scale nature. The spatial pyramid pooling module in the backbone network is replaced to enhance the perception ability of the model in segmenting targets of different scales. Finally, the Focal loss classification loss function is incorporated to address the issues of overfitting and poor accuracy caused by imbalanced class distribution in the training data. We conduct comparative tests on a self-constructed dataset comprising ten categories of water surface targets using four algorithms: Faster R-CNN, YOLOv4, YOLOv5, and the proposed improved YOLOv5. The experimental results demonstrate that the improved model achieves the best detection accuracy, with an 8% improvement in mAP@0.5 compared to the original YOLOv5 in multi-scale water surface object detection.

Keywords: surface target detection; YOLOv5; multi-scale targets; spatial pyramid pooling; attention mechanism

MSC: 68T07

check for updates

Citation: Ma, Z.; Wan, Y.; Liu, J.; An, R.; Wu, L. A Kind of Water Surface Multi-Scale Object Detection Method Based on Improved YOLOv5 Network. *Mathematics* **2023**, *11*, 2936. https://doi.org/10.3390/math11132936

Academic Editors: Huawen Liu, Chengyuan Zhang, Weiren Yu and Chunwei Tian

Received: 2 June 2023
Revised: 26 June 2023
Accepted: 28 June 2023
Published: 30 June 2023

1. Introduction

Visual-based intelligent equipment for water surface object detection plays a crucial role in regulating water environments, ensuring maritime safety, executing military tasks, conducting marine resource exploration, and monitoring unmanned islands and reefs [1,2]. The diversity of water surface target types, uneven distribution of target sizes, and the difficulties in constructing comprehensive datasets significantly increase the challenges associated with water surface object detection. The current multi-scale object detection suffers from uneven detection accuracy, making it a challenging task to improve the overall performance of the current object detector [3].

In 2012, AlexNet [4] achieved breakthrough results in large-scale image classification with the adoption of Convolutional Neural Networks (CNNs) on the ImageNet [5] dataset, which ignited the popularity of deep learning techniques. In 2013, Zuo Jianjun et al. [6] used the background subtraction method for the detection of floating objects on water surfaces. The method involved first establishing a background image without any objects and then subtracting the background image from the input image to detect the objects. In 2016, Xu Peng [7] investigated various motion detection methods, including frame differencing, Gaussian mixture background model, and optical flow. They determined that combining Gaussian differencing with three-frame differencing and utilizing mathematical morphology achieved effective detection of moving objects on the water surface. Wang Fangchao [8] introduced a rapid detection method for water surface ships based on geometric features. They employed an improved Sobel algorithm to enhance the contrast of ships at a coarse resolution. In 2018, Tang Lidan [9] addressed the issue of low detection accuracy for small objects in the Faster R-CNN algorithm. They proposed a detection method that combined the ResNet and DenseNet fusion backbone networks with a recurrent feature pyramid, achieving effective detection of water surface targets. In 2019, Liu Hehe [10] designed a network structure for water surface object detection using the SSD algorithm and made improvements to it. However, the aforementioned methods generally suffer from slow detection speeds. In 2020, Liang Yuexiang [11] presented a deep convolutional neural network-based fine-grained detection method using YOLOv3-tiny. This method could assist ship operators in identifying water surface targets. In 2022, Wang [12] proposed the SPMYOLOv3 algorithm for detecting water surface debris. This algorithm addresses the challenges of varying object shapes and sizes, as well as the difficulty in distinguishing objects from the background in water surface debris detection. The improved algorithm achieves a detection accuracy of 73.32% on a water surface debris dataset. In 2022, Wang Zhiguo [13] proposed an improved YOLOv4 method for water surface object detection. They validated the algorithm on a self-constructed dataset, achieving a detection accuracy of 89.86%. However, due to the limited size of the dataset, further improvements are required to enhance the model's accuracy and robustness.

This study focuses on the demand for accurate multi-scale object detection on water surfaces and combines the strengths and weaknesses of current mainstream object detection algorithms. YOLOv5, which exhibits outstanding performance in terms of speed and detection accuracy, was chosen as the base algorithm [14–16]. However, it was observed that the basic YOLOv5 model suffered from issues such as missed detection and false positives for small objects, as well as overfitting and low accuracy due to imbalanced sample classification. To address these problems, this paper conducts research on the improvement of the YOLOv5 network architecture. In response to the demand for improved accuracy in multi-scale object detection on water surfaces, this study focuses on the research and improvement of the YOLOv5 network model structure. The specific optimization methods employed in this research are described as follows. Firstly, the optimization of object bounding boxes is performed using the K-means++ algorithm. This approach effectively addresses classification errors by refining the localization accuracy of detected objects. Secondly, attention mechanisms are integrated into the backbone network architecture. This inclusion enhances the learning capacity of the object detection network, enabling it to effectively detect and classify multi-scale targets present on water surfaces. Furthermore, the SPPF (Spatial Pyramid Pooling with FPN) layer in the backbone network is enhanced. This modification facilitates faster and stronger learning capabilities of the network for detecting multi-scale targets on water surfaces. To address the issue of imbalanced sample distribution, the loss function is optimized by introducing Focal loss [17]. This adaptation allows for better handling of the uneven distribution of positive and negative samples, thereby improving the model's ability to detect multi-scale water surface objects. By incorporating these enhancements, the proposed model exhibits significant improvements in detecting multi-scale targets on water surfaces. The optimizations contribute to increased detection accuracy, making the model better suited for real-world scenarios involving multi-scale water surface objects. Experimental results demonstrate that the improved YOLOv5 model significantly enhances the detection accuracy and recognition precision of multi-scale water surface targets.

2. Related Work

2.1. YOLOv5 Model Analysis

The model structure of YOLOv5 is shown in Figure 1.

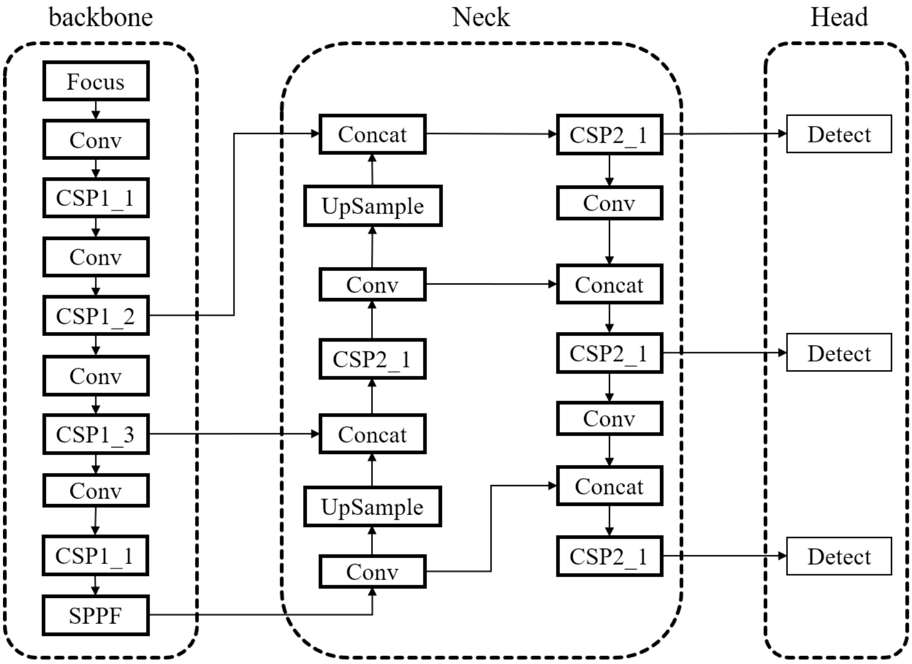


Figure 1. Network structure diagram of YOLO V5.

Model Key Features:

1. Mosaic Data Augmentation: In the training phase, the model utilizes Mosaic data augmentation, which greatly improves the training speed of the network [18,19].
2. Focus Structure and CSPNet-inspired Backbone: The model employs the Focus structure in the backbone network and draws inspiration from the CSPNet [20] architecture. It uses the ReLU activation function [21], which enhances the gradient flow within the network, improves computational speed, and facilitates better extraction of depth information.
3. Feature Fusion and Contextual Information: The model incorporates the "FPN+PAN" structure in the neck layer for feature fusion, allowing for repeated feature extraction. The SPPF module is utilized to effectively capture contextual image features [22,23].
4. GIOU loss and Anchor-based Detection: The YOLOv5 model employs the GIOU loss as the loss function at the output end. It utilizes anchor boxes to predict target boxes [24].

2.2. GAMAttention Attention Mechanism

Filtering out crucial features related to the target objects within the entire feature space is crucial for improving the accuracy of object detection [25]. Inspired by the Convolutional Block Attention Module (CBAM) [26], the GAMAttention attention mechanism [27] enhances the global interaction feature without significant information loss, thereby improving the detection accuracy of the model. The GAMAttention structure consists of two modules: channel attention and spatial attention, as shown in Figure 2.

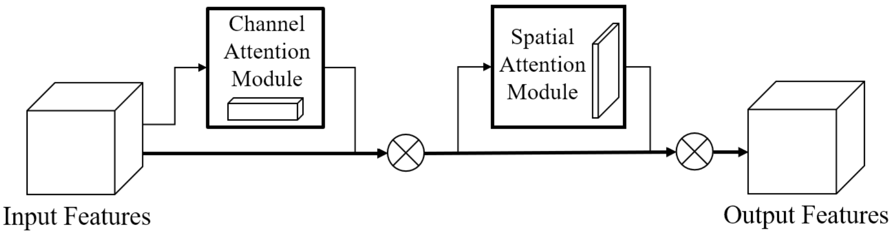


Figure 2. GAMAttention network structure diagram.

The GAM (Global Attention with Multi-scale attention) structure achieves global perception and multi-scale attention through two key components.

Firstly, the global perception module is used to capture the global contextual information of the image. It employs global average pooling and fully connected layers to enable global perception over the entire feature map, generating a global context description.

Secondly, the multi-scale attention module is designed to enhance the focus on objects at different scales. This module utilizes multiple parallel attention branches to independently apply attention weights to feature maps at different scales. Each branch adjusts the scale of the feature map through convolutional and pooling operations and calculates the corresponding attention weights. Subsequently, the feature maps at different scales are multiplied by their corresponding attention weights, resulting in weighted feature representations. This enables the model to better discriminate objects at different scales and enhance the detection capability for small objects.

2.3. SPPFCSPC Spatial Pyramid Pooling Module

In the YOLOv5-6.0 network model, the spatial pyramid pooling is implemented using the SPPF module. The SPPF module is an evolution of the Spatial Pyramid Pooling (SPP) module [28] and enables adaptive-sized output. The SPPF module, as depicted in Figure 3, allows for capturing features at multiple scales and achieving better contextual information representation.

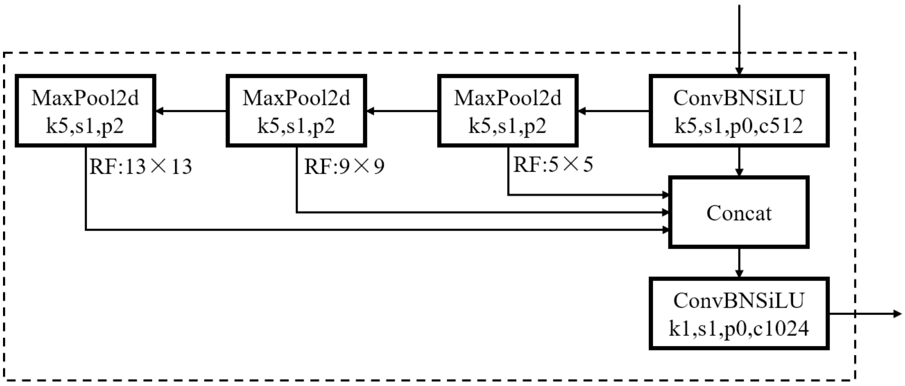


Figure 3. Network structure diagram of SPPF.

The Spatial Pyramid Pooling Fully Connected Spatial Pyramid Convolution (SPPFCSPC) module is designed based on the Spatial Pyramid Pooling Connected Spatial Pyramid Convolution (SPPCSPC) module and incorporates the design principles of the SPPF module. It is an improved approach that provides better performance. The structure of the SPPFCSPC module is illustrated in Figure 4.

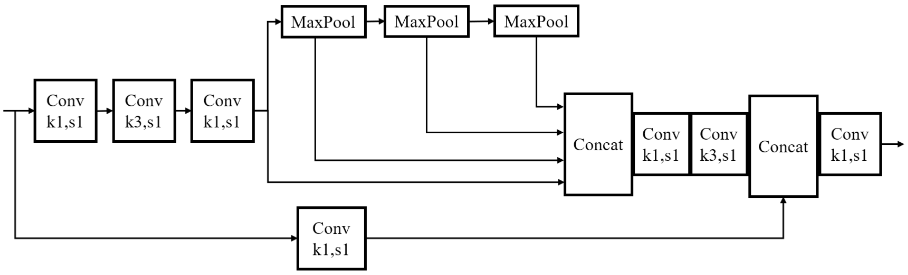


Figure 4. Network structure diagram of SPPFCSPC.

The SPPFCSPC module consists of two key techniques: Spatial Pyramid Pooling (SPP) and Fully Connected Spatial Pyramid Convolution (FCSPC).

First, the SPP component adopts the concept of spatial pyramid pooling, which involves pooling operations on the input feature map at different scales. This enables the capture of object information of varying sizes. The pyramid structure allows the network to process objects of different scales on a single fixed-size feature map, mitigating the impact of scale variations on object detection. The SPP generates fixed-length feature vectors that can serve as inputs to subsequent classifiers.

Secondly, the FCSPC component employs fully connected spatial pyramid convolution to operate on the feature vectors generated by the SPP. This integration and utilization of information at different scales aims to enhance the network's ability to handle multi-scale objects.

2.4. Focal Loss Classified Loss Function

In object detection tasks, the loss functions primarily consist of two categories: classification loss functions and regression loss functions [29]. Here, we focus on discussing the classification loss function. To address the challenges of difficult-to-classify samples and class imbalance, Dr. Kaiming He introduced the Focal loss [11] as a classification loss function. Its mathematical representation is as follows:

$$loss = -y(1-a)^{\gamma}\log y' - (1-y)a^{\gamma}\log(1-a) \tag{1}$$

In the equation, y represents the true label of the sample, a represents the predicted output values after applying the SoftMax function, and γ is a factor introduced on top of binary cross-entropy. If $\gamma > 0$, it indicates that the loss for easy samples will be reduced, allowing the network to pay more attention to difficult samples that are prone to misclassification and mitigate the impact of easy samples during network training.

To address the issue of imbalanced distribution of positive and negative samples, a balancing factor β is introduced to Equation (1), resulting in the Focal loss function that can mitigate class imbalance and explore difficult-to-classify samples. The modified Focal loss function is represented as Equation (2):

$$\text{loss} = -\beta y(1-a)^{\gamma}\log y' - (1-\beta)(1-y)a^{\gamma}\log(1-a) \tag{2}$$

3. Improvement of YOLOv5 Network Model

3.1. Basic Idea of Improving YOLOv5 Model

In multi-scale object detection on water surfaces, the features of small targets tend to become increasingly weak as the network deepens, leading to issues of missed and false detection. To address this issue, incorporating attention mechanisms into the YOLOv5 network can enhance its expressive power by focusing on crucial features in the feature maps while suppressing irrelevant features that contribute less to network training, thereby effectively reducing missed and false detection. The dual attention mechanism considers not only the varying importance of pixels across different feature channels but also the varying importance of pixels at different positions within the same feature channel. Therefore, the GAMAttention mechanism is attempted to be integrated into the YOLOv5 network model. In addition to the aforementioned challenges, water surface object detection is also affected by issues such as low contrast and lighting variations. To mitigate these challenges, image enhancement techniques are applied as a preprocessing step on the dataset. Image enhancement techniques, such as histogram equalization, contrast stretching, and adaptive histogram equalization, are employed to improve the contrast of the images and enhance the visibility of the objects in the water surface scenes.

By incorporating these techniques, the visibility and detectability of water surface objects in challenging lighting conditions and low-contrast scenarios can be significantly improved.

3.2. Optimization of Target Box Based on K-Means++

The traditional YOLO algorithm utilizes the K-means clustering algorithm [30] to obtain anchor boxes for object detection. This algorithm is simple to implement and efficient, but it is sensitive to the initial point selection, outliers, and isolated points. In comparison to K-means, K-means++ [31] improves the selection of initial points and has been shown to effectively enhance the classification error and improve the detection accuracy of multi-scale objects based on testing experiments conducted on public datasets such as PASCAL VOC. It is particularly suitable for scenarios with significant variations in object sizes.

The basic steps of K-means++ for generating anchor boxes are as follows:

1. Initialize the cluster centers by randomly selecting the bounding box regions of certain samples.
2. Calculate the distances from the initial cluster centers to each data sample.
3. Compute the probability for each sample to become the next cluster center using Equation (3).

$$P(x) = \frac{D(x)}{\sum D(x)^2} \tag{3}$$

4. Compare the probabilities and select the next cluster center using the roulette wheel selection method.

Repeat the above steps until the size of the obtained anchor box no longer change.

3.3. Embedding GAMAttention and SPPFCSPC Modules

(1) Embedding of GAMAttention module

The embedding of GAMAttention attention mechanism can introduce complexity and increase computational overhead to the network model. Therefore, it is considered to be added only at one position in the YOLOv5 model.

To incorporate the GAMAttention module into the backbone network for testing, the embedding process is depicted in Figure 5. The algorithm that embeds the GAMAttention module into the backbone network is named YWFA_B.

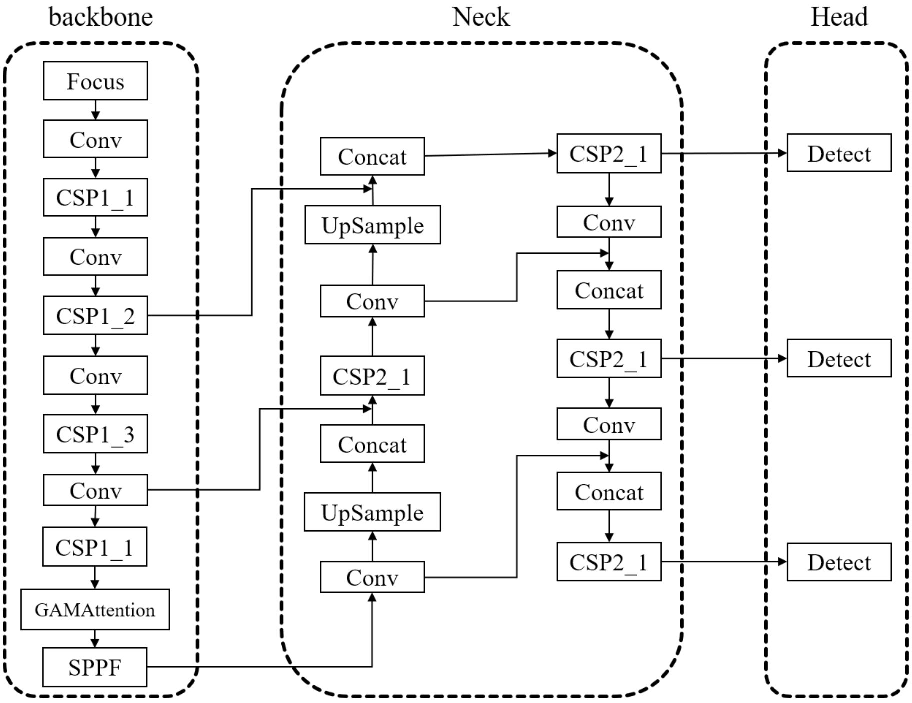


Figure 5. YWFA_B algorithm structure.

(2) Embedding of SPPFCSPC module

The SPPF module in the YOLOv5 backbone network is replaced with the SPPFCSPC module, which is named YWFS_B. The improved model with this modification is illustrated in Figure 6.

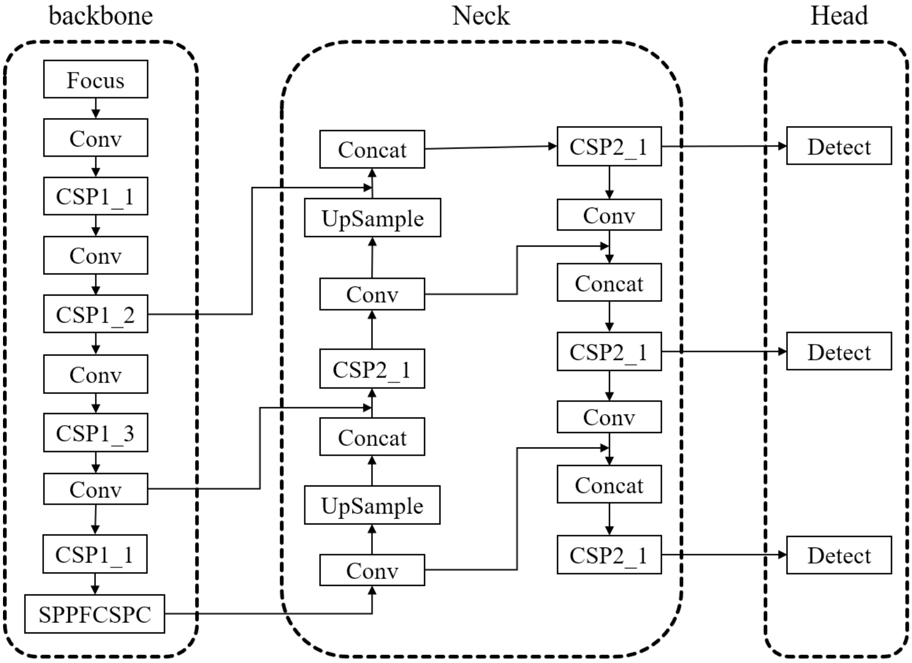


Figure 6. YWFS_B algorithm structure.

(3) GAMAttention and SPPFCSPC module are embedded at the same time

After conducting separate studies on the effects of different embedding approaches for the GAMAttention and SPPFCSPC modules in the network, it is being considered to simultaneously embed both the GAMAttention and SPPFCSPC modules into the network. The specific embedding approach will be chosen based on the results obtained from testing.

3.4. Loss Function Optimization

Due to the existence of class imbalance in the self-built dataset used for water surface object detection, the Mosaic data augmentation technique is applied to the input of YOLOv5. This technique involves randomly scaling, cropping, and placing four images before concatenating them (as shown in Figure 7). This approach helps mitigate the overfitting and low accuracy issues caused by the imbalanced distribution of samples among different classes.

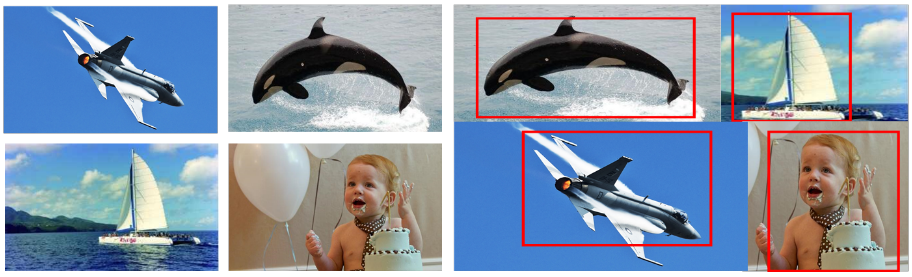


Figure 7. Mosaic Data Enhancement.

In this study, a modified version of YOLOv5 named YOLOv5_F is proposed, which incorporates the Focal loss as part of the classification loss function. This modification further addresses the issue of class imbalance. The improved YOLOv5 model is illustrated in Figure 8.

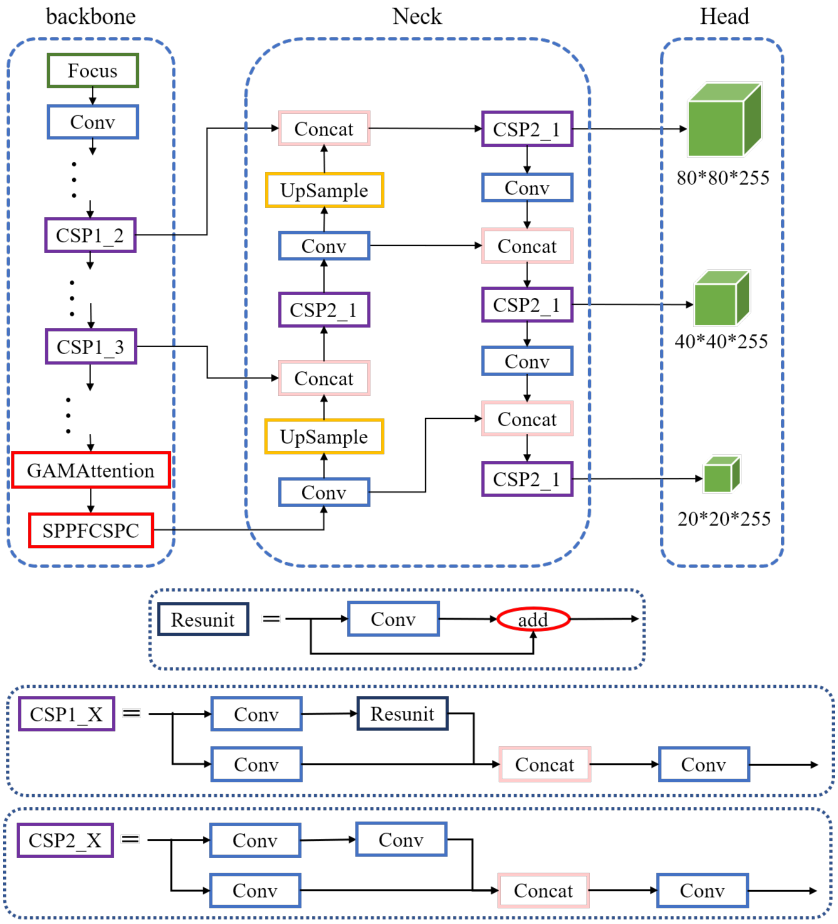
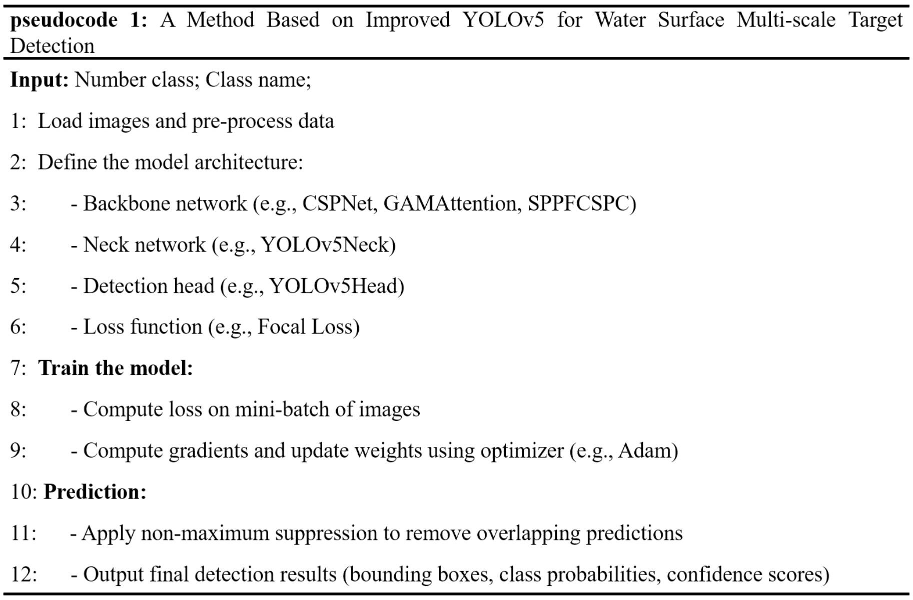


Figure 8. The improved YOLOv5 algorithm model in this paper.

3.5. Pseudocode of the Improved Model

The pseudocode of the improved model is shown in Figure 9.



```
pseudocode 1: A Method Based on Improved YOLOv5 for Water Surface Multi-scale Target Detection
Input: Number class; Class name;
Load images and pre-process data
Define the model architecture:
    - Backbone network (e.g., CSPNet, GAMAttention, SPPFCSPC)
    - Neck network (e.g., YOLOv5Neck)
    - Detection head (e.g., YOLOv5Head)
    - Loss function (e.g., Focal Loss)
Train the model:
    - Compute loss on mini-batch of images
    - Compute gradients and update weights using optimizer (e.g., Adam)
Prediction:
    - Apply non-maximum suppression to remove overlapping predictions
    - Output final detection results (bounding boxes, class probabilities, confidence scores)
```

Figure 9. Pseudo code diagram.

The proposed improvements in this study primarily focused on the backbone network and the loss function of the object detection algorithm. The GAMAttention attention mechanism is incorporated into the backbone network, enhancing its ability to capture relevant features. Additionally, the SPPF pooling module is modified to improve the network's capability to handle objects of various scales on the water surface. Furthermore, the loss function is optimized to better balance the uneven distribution of positive and negative samples.

4. Experimental Research and Result Analysis

The experiments were conducted under the following conditions: GeForce RTX 3080 12G GPU model, Windows 10 operating system, CUDA 11.3, and PyTorch 1.10.1 as the deep learning environment and framework. The self-built water surface multi-scale object dataset was used for evaluation and testing.

4.1. Establishment of Water Surface Target Data Set

Currently, there are many open-source datasets available with abundant object categories, such as the ImageNet dataset [5], the PASCAL VOC dataset [32,33], and COCO dataset [34]. However, there is currently no open-source dataset specifically designed for water surface object detection and recognition. Considering that the water surface visual system primarily focuses on monitoring the condition of the current water area, the key target objects for detection and recognition include ships, low-flying airplanes and birds, people on boats, and large fish leaping out of the water. Ships can be further classified into categories such as warships, cruise ships, cargo ships, sailboats, yachts, and other types of boats.

Based on the above considerations, this study has developed a self-built dataset comprising 5 major categories and 10 subcategories. The category labels are as follows: person, Cargoship, yacht, youlun, warship, bird, fish, fanchuan, other boat, and airplane. Some sample images from each category of the dataset are shown in Figure 10.

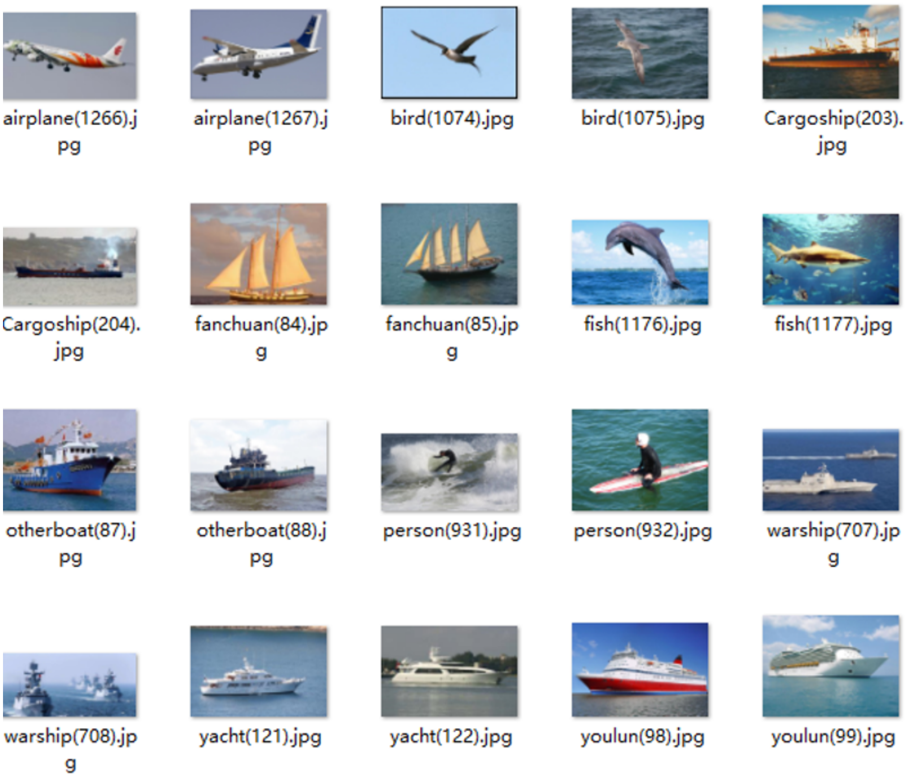


Figure 10. Part of the image of the data set established.

The dataset is divided into training, validation, and testing sets in a ratio of 8:1:1. Following the definition of absolute targets, the dataset is analyzed to determine the distribution of object scales in terms of large, medium, and small objects. Objects occupying less than 0.12% of the total image area are classified as small objects, those occupying more than 0.12% but less than 0.38% are classified as medium objects, and those occupying more than 0.38% are classified as large objects. The distribution of objects based on these criteria is illustrated in Figure 11.

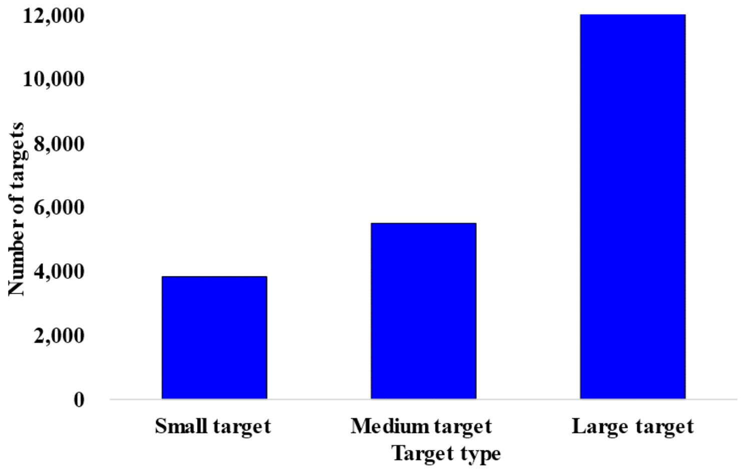


Figure 11. Distribution of large, medium, and small targets.

Due to the varying quantity and quality of images available online, the constructed multi-class object dataset may suffer from class imbalance, where certain classes have more samples than others. This can potentially lead to lower accuracy in algorithms. Figure 11 illustrates the distribution of object categories in the self-built dataset.

From Figure 12, it can be observed that the most frequent category in the self-built dataset is “person,” while the least frequent category is “yacht.” This indicates a certain degree of class imbalance issue among the samples in the dataset.

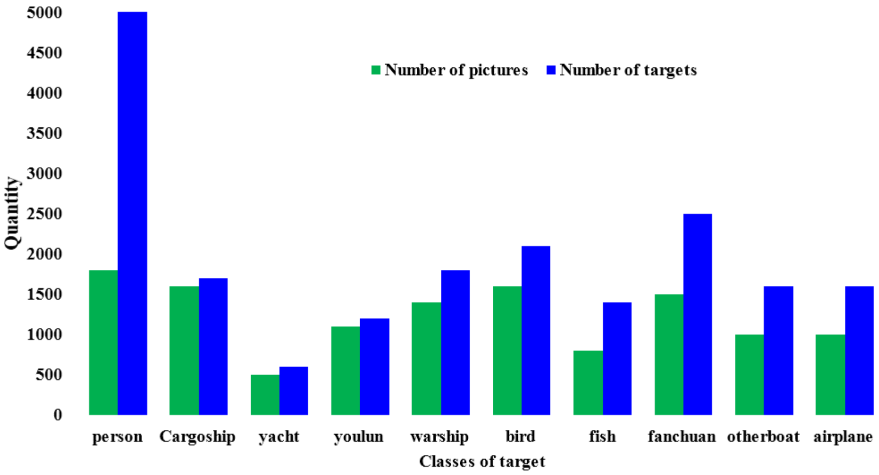


Figure 12. Distribution of target species.

4.2. Target Detection Evaluation Index

The problem of object detection involves locating and classifying objects in an image. Object detection models are typically trained on a fixed set of classes, so the model can only locate and classify those specific classes in the image. Furthermore, the location of the objects is usually represented by bounding boxes. Therefore, object detection requires both the localization of objects in the image and their classification. The accuracy of object detection is commonly measured using Mean Average Precision (mAP). Although mAP is not an absolute measure of the model’s output, it is a useful relative measure. When computing this metric on a dataset, it allows for easy comparison between different object detection methods. which is calculated as follows:

$$mAP = \frac{\sum\limits_{i=1}^{c} AP_i}{c} \tag{4}$$

where $AP = \int_0^1 P(R)dR$ represents the average precision for a specific class in the dataset. The average precision is calculated by plotting the precision-recall curve and calculating the area under the curve (AUC). The calculation of precision (P) and recall (R) involves the following formulas:

$$P = \frac{X_{TP}}{X_{TP} + X_{FP}} \tag{5}$$

$$R = \frac{X_{TP}}{X_{TP} + X_{FN}} \tag{6}$$

where X_{TP} represents true positive (correctly classified positive samples), X_{TN} represents true negative (incorrectly classified negative samples), X_{FP} represents false positive (incorrectly classified positive samples), and X_{FN} represents false negative (incorrectly classified negative samples).

In the experiments, mAP@0.5 is used as the performance evaluation metric for the network. mAP@0.5 calculates the average precision by considering the intersection over union (IOU) between the predicted bounding boxes and the ground truth boxes, with an IOU threshold of 0.5. The AP for each class is calculated, and then the average of all class APs is taken, resulting in mAP@0.5.

4.3. Classification and Comparison Experiment

The algorithm in this paper is trained for a total of 100 epochs. The specific parameter settings used in the training process are listed in Table 1.

Table 1. List of Hyperparameters.

Argument	Value
batch_size	−1
Imgsz	640
epochs	100
loss	Focal_loss
classes	10

(1) Comparative experiment of target box optimization

Here is a comparison of the object detection accuracy between YOLOv5 and YOLOv5 with optimized bounding boxes using the K-means++ algorithm (named YKmeans++). In the experiments, the Anchor box cluster centers obtained using the K-means++ algorithm are as follows: (13, 24), (21, 57), (38, 106), (50, 45), (69, 176), (99, 98), (129, 270), (219, 168), (364, 333). The training epochs for both models are set to 100, and the IOU threshold is set to 0.5. The comparison of object detection accuracy before and after bounding box optimization is shown in Table 2.

Table 2. object Detection Accuracy Comparison (before and after bounding box optimization.

Network Model	P/IOU0.5	R/IOU0.5	mAP@0.5
YOLOv5	89.7%	81.6%	86.5%
YKmeans++	90.60%	82.42%	87.37%

From Table 1, it can be visually observed that the YOLOv5 network with optimized bounding boxes (YKmeans++) shows an improvement of 0.87% in mAP@0.5 compared to the network without optimization. This indicates that optimizing the bounding boxes has a certain effect on improving object detection accuracy.

(2) Comparative experiment of GAMAttention embedded in backbone network

The YWFA_B algorithm obtained by embedding the GAMAttention mechanism module into the backbone network of the YOLOv5 network model is tested and compared with the original YOLOv5 algorithm. The results are shown in Table 3.

Table 3. Comparison of Embedded GAMAttention.

Network Model	P/IOU0.5	R/IOU0.5	mAP@0.5
YOLOv5	89.7%	81.6%	86.5%
YWFA_B	83.2%	92.5%	92.2%

From Table 3, it can be observed that when the GAMAttention module is embedded in the backbone of the YOLOv5 network, there is a significant improvement in mAP@0.5, which increased by 5.7% compared to the original network. This suggests that incorporating the GAMAttention attention mechanism enhances the network's ability to train and perform better on the given dataset.

(3) Comparative experiment of replacing SPPFCSPC module with backbone network

The model algorithm YWFS_B (direct replacement of backbone network) formed by replacing SPPF in YOLOv5 network structure with SPPFCSPC module is compared and analyzed with YOLOv5 original network, and the results are shown in Table 4.

Table 4. Comparison of Embedded SPPFCSPC.

Network Model	P/IOU0.5	R/IOU0.5	mAP@0.5
YOLOv5	89.7%	81.6%	86.5%
YWFS_B	90.4%	85.5%	93.4%

From Table 4, it can be observed that replacing the SPPF module with the SPPFCSPC module in the YOLOv5 network (YWFS_B algorithm) improves the detection accuracy on the self-built test dataset. The improved model achieves a 6.9% increase in mAP@0.5 for water surface object detection compared to the original YOLOv5 algorithm.

(4) Comparative experiment before and after adding the loss function Focal loss

Comparing the algorithms before and after the YOLOv5 network is added to the classification loss function Focal loss, training, and testing are carried out on the self-built data set in this paper, and the results are shown in Table 5.

Table 5. Comparison of Focal loss before and after Addition.

Network Model	P/IOU0.5	R/IOU0.5	mAP@0.5
YOLOv5	89.7%	81.6%	86.5%
YWFS_BF	92.3%	88.2%	93.7%

From Table 5, it can be observed that adding the Focal loss as a classification loss function in the YOLOv5 network (YWFS_BF algorithm) helps balance the impact of imbalanced class distribution on the detection accuracy. The addition of Focal loss results in a 7.2% increase in mAP@0.5 for water surface object detection compared to the original YOLOv5 algorithm.

4.4. Comparative Experiment of Improved YOLOv5 Model

Based on the analysis of the comparative experiments in (2) and (3) of Section 4.3, it can be concluded that embedding the GAMAttention attention mechanism in the backbone network and replacing the SPPF module with the SPPFCSPC module in the backbone network can significantly improve the detection accuracy of the YOLOv5 algorithm for water surface objects. By combining these two improvements, a modified YOLOv5 algorithm is proposed, named YOLOv5_WFT.

In YOLOv5_WFT, the classification loss function Focal loss is incorporated, and the optimized bounding boxes obtained through K-means++ are used during network training. This modification results in an improved model referred to as YOLOv5_WFT_KF.

The two modified YOLOv5 models and the original YOLOv5 model are tested on the test set, and the results are shown in Table 6. From Table 6, it can be observed that the improved YOLOv5 models can significantly enhance the accuracy of object detection. Specifically, YOLOv5_WFT_KF achieves an 8.1% improvement in mAP@0.5 compared to YOLOv5.

Table 6. Comparison of target detection quality before and after YOLOv5 improvement.

Network Model	P/IOU0.5	R/IOU0.5	mAP@0.5
YOLOv5	89.7%	81.6%	86.5%
YOLOv5_WFT	93.6%	86.4%	94%
YOLOv5_WFT_KF	93.8%	87.2%	94.6%

Figure 13 lists the training results of YOLOv5 and YOLOv5_WFT_KF for this water surface target data set.

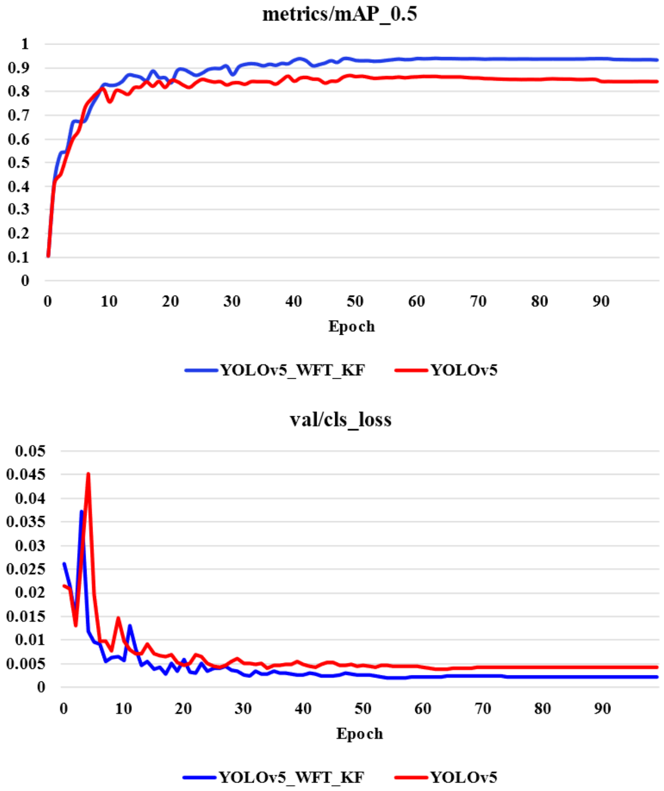


Figure 13. The training results of YOLOv5 model and YOLOv5_WFT_KF model.

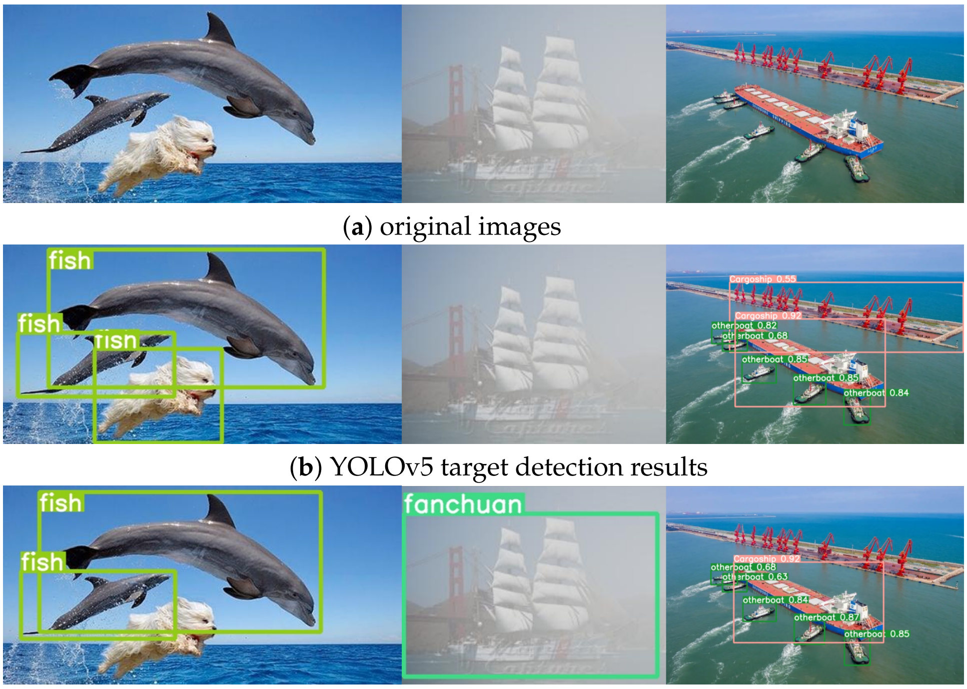
From Figure 13, it can be observed that the proposed YOLOv5_WFT_KF model demonstrates good training convergence. The training loss curve shows that the model's training process reaches convergence around epoch = 55. After this point, there is no evidence of overfitting or failure to converge, indicating that the model has effectively learned the underlying patterns and features of the multi-scale water surface objects during training. This indicates the stability and effectiveness of the proposed model in achieving convergence and avoiding training issues.

This paper presents a comparative analysis of detection quality for classification targets among Faster R-CNN, YOLOv4, YOLOv5, and the proposed YOLOv5_WFT_KF algorithm. As shown in Table 7, the YOLOv5_WFT_KF algorithm, which is an improved version of YOLOv5, significantly improves the accuracy of multi-scale target detection on water surface compared to the other three algorithms.

Table 7. FasterR-CNN YOLOv4 YOLOv5 and YOLOv5_WFT_KF Classification target detection quality.

Algorithm Model	Airplane (%)	Bird (%)	Cargoship (%)	Fanchuan (%)	Fish (%)
Faster R-CNN	81.16	88.99	81.09	80.66	60.38
YOLOv4	94.34	80.31	83.10	88.92	35.03
YOLOv5	99.1	95.7	93.9	94.7	85.4
YOLOv5_WFT_KF	95.2	94.4	93.8	98.4	93.9
Otherboat (%)	**Person (%)**	**Youlun (%)**	**Warship (%)**	**Yacht (%)**	**mAP@0.5 (%)**
57.51	52.70	90.61	80.84	71.13	74.51
81.79	65.81	85.14	95.99	95.24	80.57
59.3	66.6	91.8	91.4	86.8	86.5
89.9	83.7	99.5	94.3	97.3	94.6

Figure 14 illustrates a comparison of detection performance on three randomly selected images from a self-built water surface target dataset before and after the improvement of the YOLOv5 model. Figure 14a shows the original images, while Figure 14b,c show the detection results of the original YOLOv5 model and the improved YOLOv5_WFT_KF model, respectively.



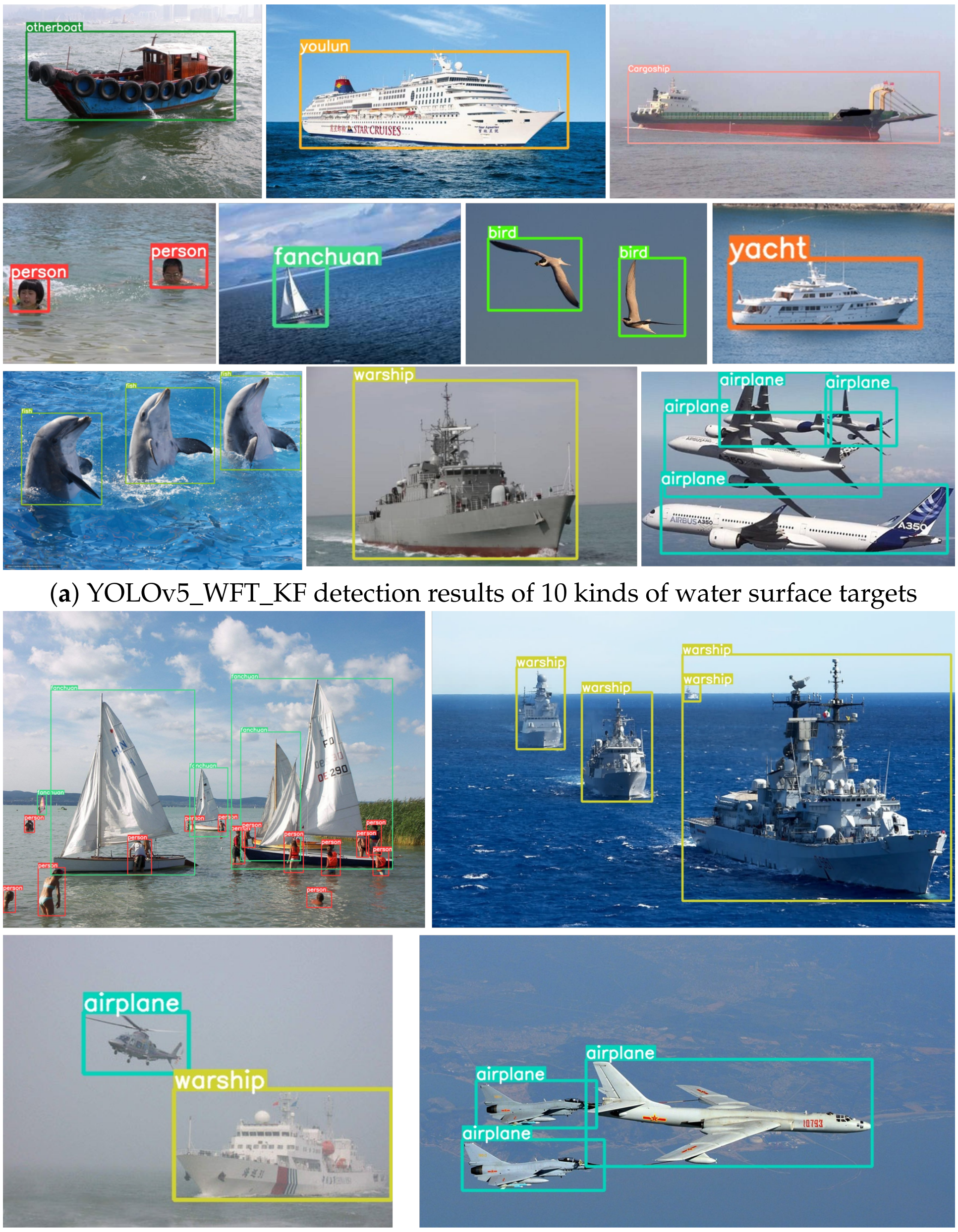
(**a**) original images

(**b**) YOLOv5 target detection results

(**c**) YOLOv5_WFT_KF target detection results

Figure 14. Comparison of target detection results between YOLOv5 and its improved model.

From Figure 15a, it can be observed that the proposed improved YOLOv5 algorithm effectively detects each category in the dataset used in the study without any instances of missed detections or false detections. Figure 15b showcases the detection results for multi-scale objects. It can be seen from the figure that the proposed improved algorithm accurately detects objects of different sizes and categories in an image, including large, medium, and small objects, with high confidence scores. This effectively demonstrates the feasibility of the proposed algorithm improvements presented in this study.



(**a**) YOLOv5_WFT_KF detection results of 10 kinds of water surface targets

(**b**) YOLOv5_WFT_KF detection results of multi-scale water surface targets

Figure 15. YOLOv5_WFT_KF target detection results.

From Figures 14 and 15, it can be observed that the improved YOLOv5 model, YOLOv5_WFT_KF, can recognize the ten classes of targets in the self-built dataset as well as multi-scale targets. Compared to the YOLOv5 model, the improved model can provide more accurate bounding boxes for small and occluded targets with high confidence.

In addition to comparing the accuracy of the network models for object detection, this paper also conducted experiments to compare the number of layers, parameters, FLOPs, and FPS between the improved and original YOLOv5 models when inputting an RGB three-channel 640 × 640 color image. The results of the comparison are summarized in Table 8.

Table 8. Comparison of parameters before and after YOLOV5 improvement.

Network Model	Layers	Parameters	Flops	FPS
YOLOv5	213	7 M	15.8 G	203.415
YOLOv5_WFT_KF	239	15.2 M	22.4 G	172.015

According to the experimental results in Table 8, the improved algorithm in this paper achieves an average precision increase of 8.1% compared to YOLOv5 when using the same input resolution. This is due to the addition of attention mechanisms and replacement of the network spatial pyramid module, which leads to an increase in the number of layers and parameters of the network model, with the parameter amount being twice that of the original network. However, this also results in a decrease in the FPS during inference.

5. Conclusions

To address the demand for multi-scale and small target detection on water surfaces, this paper proposes an improved YOLOv5 model based on the characteristics of the deep learning YOLOv5 model. The improved model mainly improves the target box setting, embeds the attention mechanism module in the backbone network, replaces the original spatial pyramid SPPF module, and uses the Focal loss classification loss function to alleviate overfitting and detection accuracy problems caused by multi-scale targets and imbalanced class samples in water surface targets. Tests on a self-built water surface target dataset show that the improved YOLOv5 model has a certain degree of improvement in detection accuracy for most classes of targets and is suitable for multi-scale target detection on water surfaces. However, there are relatively few images of occluded targets in the dataset, and the detection performance of the model for occluded targets still needs to be verified.

Furthermore, the addition of multiple modules increases the computational complexity and the model's complexity, which will affect the real-time performance of the algorithm. Therefore, in future research, the model's lightweight design should be considered to reduce the consumption of unnecessary resources while ensuring a certain level of detection accuracy, thus improving the real-time performance and efficiency of network detection.

Author Contributions: Conceptualization, J.L.; Methodology, Z.M.; Software, Y.W.; Validation, Y.W.; Formal analysis, J.L.; Investigation, R.A.; Data curation, L.W. All authors have read and agreed to the published version of the manuscript.

Funding: The authors are grateful to College of Automation, Chengdu University of Information Technology, this paper is supported by International Cooperation Project of Science and Technology Bureau of Chengdu (No. 2019-GH02-00051-HZ), Sichuan unmanned system and intelligent perception Engineering Laboratory Open Fund, and Research Fund of Chengdu University of information engineering, under Grant (No. WRXT2020-001, No. WRXT2020-002, No. WRXT2021-002 and No. KYTZ202142), and the Sichuan Science and Technology Program China, under Grant (No. 2022YFS0565); This paper is also supported by the Key R&D project of Science and Technology Department of Sichuan Province, under Grant (2023YFG0196 and 2023YFN0077), Science and Technology achievements transformation Project of Science and Technology Department of Sichuan Province, under Grant (2023JDZH0023), Sichuan Provincial Science and Technology Department, Youth Fund project, under Grant (2023NSFSC1429).

Data Availability Statement: Data sharing not applicable. Since this article uses a self-built dataset, the dataset is not shared.

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

References

1. Li, H. Research on Multi-Target Recognition and Tracking Technology of Ship Vision System for Sea-Air Targets. Master's Thesis, Harbin Engineering University, Harbin, China, 2019.
2. Yin, K.; Wang, X.; Wu, Y.; Qin, M.; Zhang, J.; Chu, Z. Water Surface Garbage Detection Based on YOLOv5. *Comput. Knowl. Technol.* **2022**, *18*, 28–30.
3. Chen, K.; Zhu, Z.; Deng, X.; Ma, C.; Wang, H. A Survey on Deep Learning for Multi-Scale Object Detection. *J. Softw.* **2021**, *32*, 1201–1227.
4. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [CrossRef]
5. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; et al. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.* **2015**, *115*, 211–252. [CrossRef]
6. Zuo, J.; Wu, Y. Intelligent Monitoring Technology of Water Floating Objects. *Softw. Guide* **2013**, *12*, 150–152.
7. Xu, P. Research on Dynamic Obstacle Detection Technology of Water Surface Unmanned Vessels. Master's Thesis, Shenyang Li Gong University, Shenyang, China, 2016.
8. Wang, F.; Zhang, M.; Gong, L. A Rapid Detection Algorithm for Surface Ships Based on Geometric Features. *J. Nav. Univ. Eng.* **2016**, *28*, 57–63.
9. Tang, L. Research on Image-Based Object Detection of Unmanned Surface Vehicles. Master's Thesis, Harbin Institute of Technology, Harbin, China, 2018.
10. Liu, H. Research on Detection of Water Surface Target Images Based on SSD Algorithm. Master's Thesis, Dalian Maritime University, Dalian, China, 2019.
11. Liang, Y.; Feng, H.; Xu, H. Fine-grained Detection of Ship Visible Light Images Based on YOLOv3-tiny. *J. Wuhan Univ. Technol. (Transp. Sci. Eng.)* **2020**, *44*, 1041–1045+1051.
12. Wang, Y.L.; Ma, J.; Luo, X.;Wang, S. Surface garbage target detection based on SPMYOLOv3. *Comput. Syst. Appl.* **2023**, *32*, 163–170.
13. Wang, Z.; Wang, Y.; Tan, X.; Zhang, H. Rapid Detection of Water Surface Targets Based on YOLOv4. *Ship Electron. Eng.* **2022**, *42*, 110–113.
14. Li, Y.; Fan, Y.; Wang, S.; Bai, J.; Li, K. Application of YOLOv5 Based on Attention Mechanism and Receptive Field in Identifying Defects of Thangka Images. *IEEE Access* **2022**, *10*, 81597–81611. [CrossRef]
15. Wang, J.; Chen, Y.; Dong, Z.; Gao, M. Improved YOLOv5 network for real-time multi-scale traffic sign detection. *Neural Comput. Appl.* **2023**, *35*, 7853–7865. [CrossRef]
16. Zhu, X.; Lyu, S.; Wang, X.; Zhao, Q. TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-captured Scenarios. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, 24 October 2022; IEEE: Piscataway, NJ, USA, 2021; pp. 2778–2788.
17. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2999–3007.
18. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y. Cutmix: Regularization strategy to train strong classifiers with localizable features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6023–6032.
19. Dulal, R.; Zheng, L.; Kabir, M.A.; McGrath, S.; Medway, J.; Swain, D.; Swain, W. Automatic Cattle Identification using YOLOv5 and Mosaic Augmentation: A Comparative Analysis. *arXiv* **2022**, arXiv:2210.11939.
20. Wang, C.Y.; Liao, H.Y.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. *arXiv* **2019**, arXiv:1911.11929.
21. Glorot, X.; Bordes, A.; Bengio, Y. Deep Sparse Rectifier Neural Networks. In Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics, Lauderdale, FL, USA, 14 June 2011.
22. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. *arXiv* **2017**, arXiv:1612.03144.
23. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. *arXiv* **2018**, arXiv:1803.01534.
24. Rezatofighi, H.; Tsoi, N.; Gwak, J. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. *arXiv* **2019**, arXiv:1902.09630.
25. Zhang, X.; Sun, C.; Han, H.; Wang, H.; Sun, H.; Zheng, N. Object-fabrication Targeted Attack for Object Detection. *arXiv* **2022**, arXiv:2212.06431.
26. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
27. Liu, Y.; Shao, Z.; Hoffmann, N. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. *arXiv* **2021**, arXiv:2112.05561.
28. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *37*, 1904–1916. [CrossRef]
29. Yang, Q.; Zhang, C.; He, Q.; Wang, H. Research Progress on Loss Functions for Object Detection. *Comput. Sci. Appl.* **2021**, *11*, 2836–2844.

30. MacQueen, J. Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Los Angeles, CA, USA, 21 June 1967; Volume 1, pp. 281–297.
31. Arthur, D. Vassilvitskii S. k-means++: The advantages of careful seeding. In Proceedings of the Eighteenth annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans, LA, USA, 7 January 2007.
32. Everingham, M.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [CrossRef]
33. Everingham, M.; Eslami, S.A.; Van Gool, L.; Williams, C.K.; Winn, J.; Zisserman, A. The pascal visual object classes challenge: A retrospective. *Int. J. Comput. Vis.* **2015**, *111*, 98–136. [CrossRef]
34. Lin, T.Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; Zitnick, C.L. Microsoft coco: Common objects in context. In Proceedings of the Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, 6–12 September 2014; Springer: Cham, Switzerland, 2014; pp. 740–755.

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.