*Mathematical Biosciences and Engineering*

http://www.aimspress.com/journal/mbe

MBE, 21(1): 1321–1341.
DOI: 10.3934/mbe.2024057
Received: 02 November 2023
Revised: 18 December 2023
Accepted: 19 December 2023
Published: 27 December 2023

*Research article*

# Fish sonar image recognition algorithm based on improved YOLOv5

**Bowen Xing[1], Min Sun[1], Minyang Ding[2] and Chuang Han[3,*]**

[1] College of Engineering Science and Technology, Shanghai Ocean University, Shanghai 201306, China

[2] Marine Design and Research Institute of China, Shanghai 200011, China

[3] Heilongjiang Province Key Laboratory of Laser Spectroscopy Technology and Application, Harbin University of Science and Technology, Harbin 150080, China

* **Correspondence:** Email: hanchuang@hrbust.edu.cn.

**Abstract:** Fish stock assessment is crucial for sustainable marine fisheries management in rangeland ecosystems. To address the challenges posed by the overfishing of offshore fish species and facilitate comprehensive deep-sea resource evaluation, this paper introduces an improved fish sonar image detection algorithm based on the you only look once algorithm, version 5 (YOLOv5). Sonar image noise often results in blurred targets and indistinct features, thereby reducing the precision of object detection. Thus, a C3N module is incorporated into the neck component, where depth-separable convolution and an inverse bottleneck layer structure are integrated to lessen feature information loss during downsampling and forward propagation. Furthermore, lowercase shallow feature layer is introduced in the network prediction layer to enhance feature extraction for pixels larger than $4 \times 4$. Additionally, normalized weighted distance based on a Gaussian distribution is combined with Intersection over Union (IoU) during gradient descent to improve small target detection and mitigate the IoU's scale sensitivity. Finally, traditional non-maximum suppression (NMS) is replaced with soft-NMS, reducing missed detections due to occlusion and overlapping fish targets that are common in sonar datasets. Experiments show that the improved model surpasses the original model and YOLOv3 with gains in precision, recall and mean average precision of 2.3%, 4.7% and 2.7%, respectively, and 2.5%, 6.3% and 6.7%, respectively. These findings confirm the method's effectiveness in raising sonar image detection accuracy, which is consistent with model comparisons. Given Unmanned Underwater Vehicle advancements, this method holds the potential to support fish culture decision-making and facilitate fish stock resource assessment.

**Keywords:** sonar; fish detection; YOLOv5; deep learning; algorithm optimization

## 1. Introduction

Fish is one of the world's foremost traded food commodities, substantially contributing to global human food security and nutritional provisioning. The sustainable growth of fish supplies for human consumption is mainly rooted in aquaculture [1]. Notably, marine fisheries, particularly mariculture, have exhibited rapid expansion in recent years. Consequently, farmed fish have become an increasingly vital source of sustainable dietary protein, contributing to a global production of 57.5 million tons in 2020 [2]. However, it is essential to acknowledge that since the 1960s, the annual average growth rate in global fish consumption, at 3.2%, has consistently surpassed both the population growth rate, which stands at 1.6%, and the rate of consumption growth for livestock and poultry products, at 2.8% [3]. Additionally, aquaculture production has exceeded capture production. Global fish stocks face significant pressure, with over one-third of them already depleted, primarily due to overfishing and marine pollution [4, 5]. In 2019, over one-third (35.4%) of the world's fish stocks were overexploited, marking a 1.2% increase since 2017. Unfortunately, the current state of fish stocks continues to deteriorate, contrasting with the United Nations Sustainable Development Goals, which aim to restore fish stocks to biologically sustainable levels [6]. The consequences of fish overfishing are far-reaching, negatively impacting food production, biodiversity and the ecological balance of marine ecosystems. Additionally, the expansion of offshore aquaculture populations is increasing the stress on the available aquaculture space, both onshore and offshore. In order to facilitate the sustainable management of fish stocks and safeguard marine ecosystems, expanding aquaculture operations to remote and deep seas is necessary. Central to this task is the assessment of marine resources, which plays a pivotal role in ensuring the stability and viability of offshore aquaculture. Foremost among these assessments is the accurate identification of species, which is the primary requisite for practical evaluation.

Traditional fishing sampling survey methods prove to be inefficient and environmentally destructive. Furthermore, optical cameras exhibit limitations in their range for underwater photography. In contrast, sonar imaging technology, which utilizes sound waves to detect underwater objects, offers valuable insights into underwater targets' location, shape, size, and movement. Compared to optical imaging and alternative underwater detection techniques, sonar imaging obtains an extended detection range [7] and a broader detection angle. Consequently, it is better suited for complex, deep-sea, and remote underwater environments, including deep waters, muddy waters, and low-light conditions. With the continuous development of sonar imaging technology, using high-resolution sonar image information for underwater object detection has become an essential means. In the field of object detection based on sonar images, although traditional methods have achieved object detection by utilizing the mathematical-statistical characteristics of sonar images, mathematical-morphological processing, and pixel differences between images, there are still limitations. Given that only one type of mathematical and statistical feature cannot fully reflect the complex background and textural features of sonar images [8], traditional image recognition methods need to determine which features to extract, which cannot guarantee the comprehensiveness and accuracy of feature extraction processes. In contrast, deep learning-based object detection excels in both accuracy and speed. Target detection algorithms can be categorized into two main types based on their processes: two-stage object detection algorithms, which rely on candidate frames, and single-stage object detection algorithms that are optimized for faster detection. Deep learning and computer vision evolved by introducing the R-CNN algorithm in 2014 [9]. Subsequently, several two-stage algorithms emerged, including the Fast R-CNN [10], Faster

R-CNN [11], SPP-Net [12] and Mask R-CNN [13]. Two-stage algorithms involve the initial generation of candidate frames, foreground identification and bounding box adjustments, resulting in slower detection speeds. In contrast, one-stage object detection algorithms eliminate the need for candidate frame generation and enable end-to-end detection. They directly predict target category probabilities and location coordinates within the network, achieving a balance between accuracy and speed. One of the most notable single-stage object detection algorithm series is the YOLO (you only look once) series [14–19]. Detecting small and dense objects in sonar images is a challenge. Tong et al. [20] comprehensively analyzed small object detection by exploring multi-scale feature learning, data augmentation, training strategy, context-based and detection based on the generative adversarial networks. They conducted experiments to assess classical detection methods on widely used datasets, aiming to evaluate the strengths and weaknesses of these detection algorithms. However, their analysis should have included the YOLO series of algorithms.

In recent years, due to the absence of a universally recognized large-scale sonar image dataset similar to the COCO dataset, there has been relatively less research on sonar image object detection algorithms. Moreover, previous studies have mainly focused on the detection of large targets such as shipwrecks [21–24], while there is a lack of research attention on small targets. Through an analysis of the requirements for small target detection in sonar images, this paper describes in-depth research on the small target scene of fish based on a self-made dataset and proposes resource assessment as a new application scenario. Furthermore, although deep learning-based sonar image detection has shown progress, it is still greatly affected by environmental factors, demands high-quality sonar images, and exhibits a notable occurrence of missed targets and false detection. In practical detection, it is essential to consider the associated complexities and uncertainties [25, 26], such as by optimizing the detection results in the face of noise and densely packed schools of fish. To address these problems, this paper introduces an enhanced YOLOv5-based algorithm for recognizing fish in sonar images. The main innovation of the improved algorithm lies in the proposal of an enhanced C3N module to replace the original C3 module in the network, increase in the depth of the network to enhance the model's expressive power and use of a reverse bottleneck layer to reduce the information loss caused by the conversion of information between different dimensional feature spaces. Next, to help the model perceive more details and spatial information and be more comprehensive in gradient descent, normalized weighted distance (NWD), which is insensitive to the scale of small targets, is introduced to combine with Intersection over Union (IoU) to compute the loss function, and different weights are assigned to it according to the proportion of small targets in the dataset. Finally, soft-NMS is introduced to replace the traditional non-maximum suppression (NMS) algorithm, and all bounding boxes are weighted and summed after the Gaussian distribution. The highest confidence bounding box is no longer used as the standard for the high overlapping targets to reduce the model's missed and false detection rates for the small-scale and overlapping fish targets.

## 2. Methods

### *2.1. YOLOv5 model*

The YOLOv5 architecture comprises four primary components: input, backbone, neck and prediction head. Its network structure closely resembles YOLOv4, with particular enhancements applied to the backbone and neck segments. Specifically, in this paper, YOLOv5–6.0 is selected. In contrast to

version 5.0, version 6.0 substitutes the Focus module in the backbone section with a $6 \times 6$ convolutional layer, enhancing GPU device efficiency. Furthermore, it replaces the spatial pyramid pooling (SPP) module in the neck section with SPPF [12] and introduces a new serial passing maxpool layer, resulting in accelerated computation.

### 2.1.1. Network structure

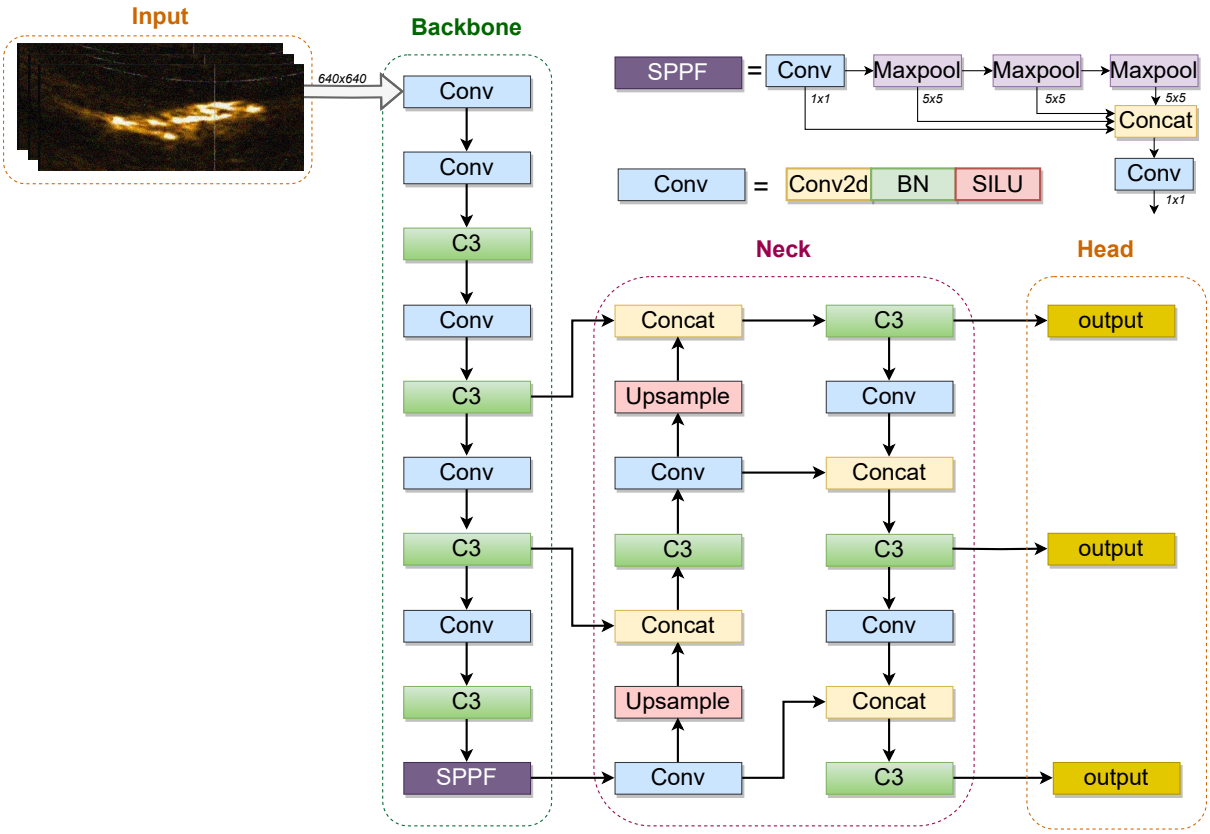



**Figure 1.** YOLOv5 network structure.

The network structure is shown in Figure 1. The input is enriched with background information by using Mosaic [17] data augmentation. Adaptive anchor frame computation determines the optimal anchor frame, while adaptive image scaling standardizes image dimensions. The backbone primarily comprises the fundamental convolutional module (Conv), the C3 module and an enhanced spatial pyramid pooling module SPPF. The Conv module conducts spatial information extraction through convolution operations, followed by feature transformation and extraction through integration of the Batch Normalization and activation function. The C3 module, composed of three Conv blocks, enhances network depth and receptive field, thereby improving feature extraction capabilities. The SPP module facilitates the integration of deep semantic information with shallow data, consequently enhancing the network's detection accuracy. The improved SPPF module further increases computational speed while maintaining accuracy. The neck section adopts the PANET [27] structure, which includes the FPN [28] layer, merging deep and shallow features with the feature pyramid incorporating both shallow and deep

features. This fusion of information from different network layers in the backbone enhances semantic and positional awareness. Three feature maps are outputted from the neck to the head component: $80 \times 80$, $40 \times 40$, and $20 \times 20$ feature maps. These dimensions cater to small, medium and large object detection, respectively.

#### 2.1.2. Loss function

In target detection algorithms, the loss function measures how well the model predicts results, helping to identify differences between the model and actual data. Therefore, the loss function is crucial during the model training process. Choosing a suitable loss function facilitates model improvement and helps to achieve quicker convergence during training. IoU shows the ratio of the intersection to the union of the ground truth box $B^{gt}$ and the prediction box $B^{p}$, as shown in Figure 2(a). The IoU loss function is equal to one minus the IoU. The formulas for IoU and the IoU loss function are respectively as follows:

$$IoU = \frac{B^{gt} \bigcap B^{p}}{B^{gt} \bigcup B^{p}} \tag{1}$$

$$LOSS_{IoU} = 1 - IoU \tag{2}$$

The YOLOv5 uses complete intersection over union (CIoU) loss as its bounding box loss function. It adds a correction factor to quantify the dissimilarity between the aspect ratios of the predicted box and the ground truth box, building upon the distance intersection over union (DIoU) loss. This adjustment excludes the case that the ground truth box contains the prediction box and the center distance is the same, which can not be distinguished. The specific formula for this function is presented below.

$$CIoU = IoU - \frac{\rho^2(b, B)}{C^2} - \alpha V \tag{3}$$

$$B = b^{gt} \tag{4}$$

(In Eq (3), the variables $b$ and $B$ denote the centers of the prediction box and the ground truth box, respectively. $\rho$ is the distance between the center points of the prediction box and the ground truth box. The variable $C$ represents the diagonal length of the minimum bounding rectangle of the prediction box and the ground truth box.)

$V$ is used to measure the consistency of the aspect ratio of the prediction box and ground truth box; it is expressed as

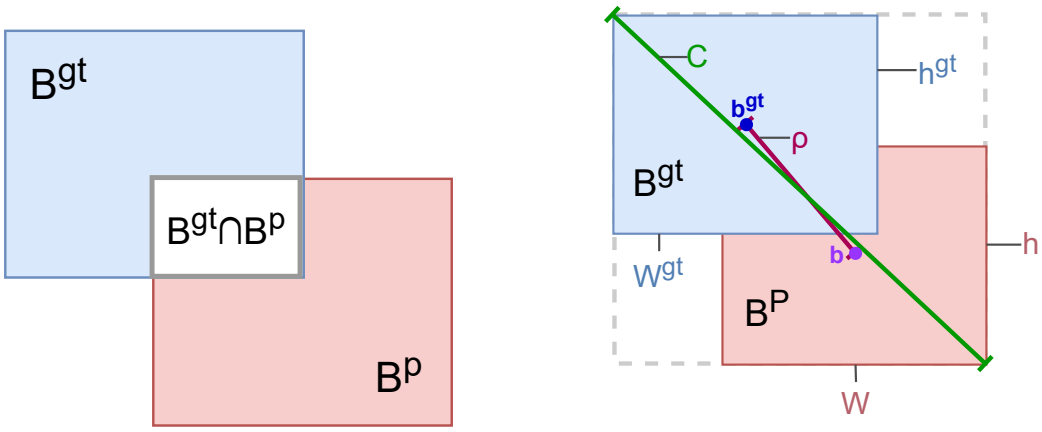
$$V = \frac{4}{\pi^2}\left(\arctan\frac{\omega^{gt}}{h^{gt}} - arctan\frac{\omega}{h}\right)^2 \tag{5}$$

$\alpha$ is the influence factor of $V$:

$$\alpha = \frac{V}{(1 - IoU) + V} \tag{6}$$

The final CIoU loss function is defined as follows, as in Figure 2(b):

$$LOSS_{CIoU} = 1 - IoU + \frac{\rho^2(b, B)}{C^2} + \alpha V \tag{7}$$




(a) Schematic diagram of IoU (b) Schematic diagram of the CIoU

**Figure 2.** Schematic diagram of the loss function.

## 2.2. *Improved YOLOv5 model*

In this paper, the YOLOv5 algorithm serves as the foundation.

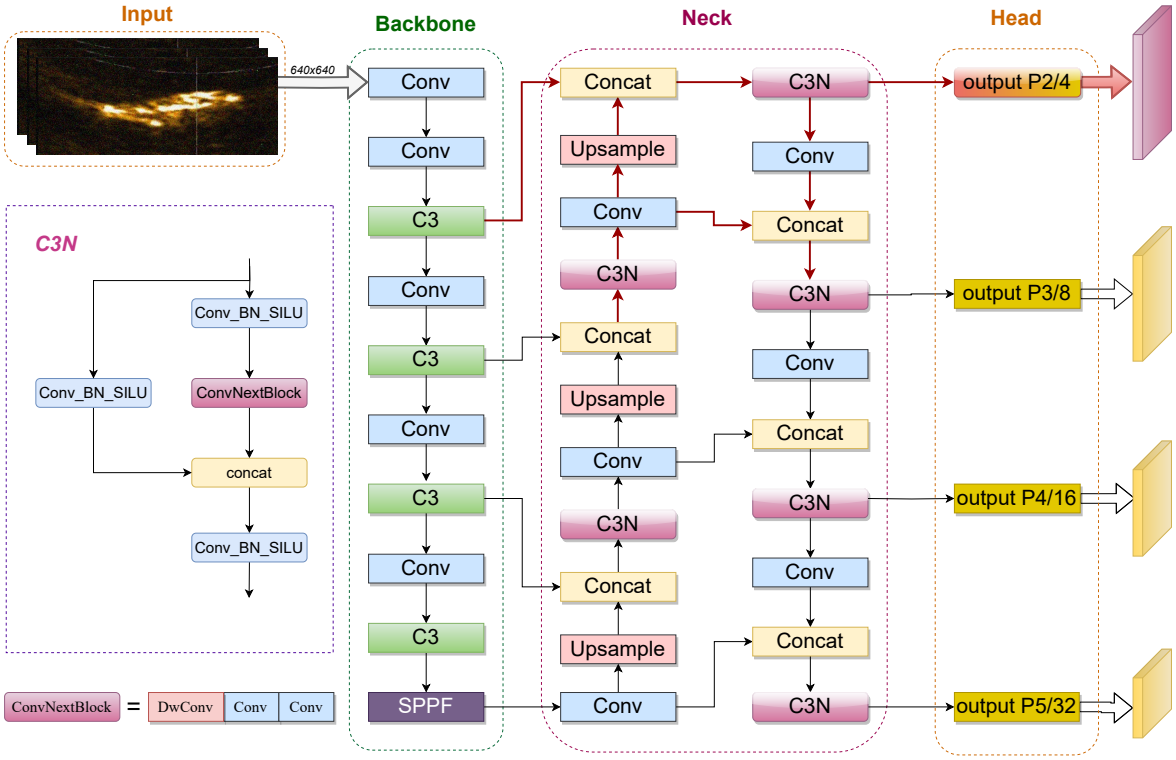


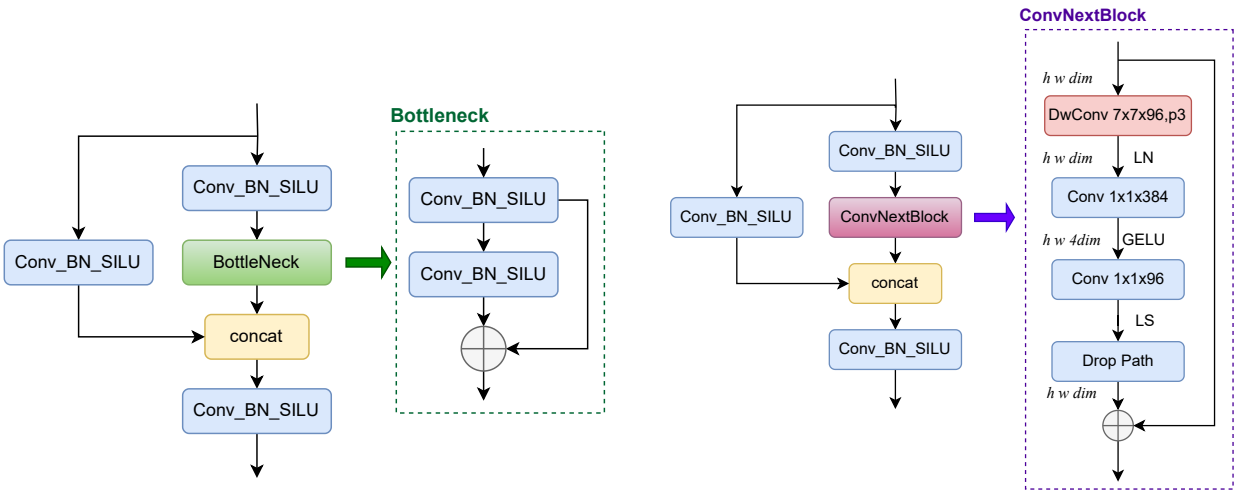

**Figure 3.** Overall network structure of the improved YOLOv5.

Essential modifications include replacing the C3 module in the neck section with the C3N module to enhance feature extraction capabilities. Additionally, a small object detection head is incorporated to improve the model's performance in terms of small object detection within sonar imagery. The

introduction of the NWD enhances the loss function index, thereby improving the model's small object detection capabilities in sonar imagery. Furthermore, enhancements to soft-NMS optimize the model's performance, particularly for scenarios involving dense objects. The network structure of the improved model is represented in Figure 3.

### 2.2.1. Improvement of C3N based on ConvNeXt

The ConvNeXt [29] network model is a new convolutional neural network (CNN) model obtained based on the ResNet [30] and Swin-Transformer [31], a model of a self-attention mechanism in the field of natural language processing. ConvNeXt combines various components inspired by the Transformer, including the training strategy, model architecture, inverted bottleneck and large kernel sizes. These enhancements, combined with optimizations in the base convolutional neural network stack and parameters, yield significantly improved inference speed. Consequently, the model attains remarkable precision levels when evaluated on ImageNet-22K.




(a) C3 (b) C3N

**Figure 4.** Comparison between C3 and C3N.

The BottleNeck module of the original YOLOv5 employs a basic convolutional neural network for feature extraction, but it falls short of capturing critical local features and contextual information. Consequently, this limitation results in a reduction in detection accuracy. In contrast, as shown in Figure 4, conventional convolution operations are discarded in the ConvNeXt module and replaced by deep separable convolution. Depthwise separable convolution offers the advantage of reduced parameters and computational workload compared to traditional convolution. Furthermore, following a comparison with MobileNetv2 [32], the ConvNeXt module implements a reverse bottleneck layer structure characterized by a wide central section and a narrow endpoint. This design empowers the ConvNeXt module to enhance its ability to capture feature correlations while efficiently mitigating information loss stemming from dimensionality compression during transformations within the feature space. The ConvNeXt model is a convolutional neural network renowned for its robust feature extraction capabilities, particularly excelling on large datasets. In this paper, the ConvNextBlock is integrated from the ConvNeXt model into the C3 module of the neck network structure of the YOLOv5m model, thus

combining both strengths. The resulting C3N module consists of three convolutional layers and a sequential module comprising multiple ConvNextBlock layers. In contrast to the C3 module, the C3N module has more convolutional and fully connected layers, enhancing network depth and complexity and ultimately augmenting the model's expressive capacity. Notably, the C3N module exhibits heightened feature extraction capabilities and improved information retention, making it well-suited for tackling the complexities inherent in sonar fish object detection. Consequently, it achieves superior performance on this specialized task.

### 2.2.2. Addition of the P2 shallow feature layer

The YOLOv5 model's backbone network acquires three scales of object detection layers, denoted as P3, P4 and P5, through three rounds of downsampling. These layers correspond to different sizes of feature maps in the neck section, where Pi represents the resolution corresponding to 1/2i of the original image. The model's head section conducts object detection on the detection head derived from the three levels of feature maps.

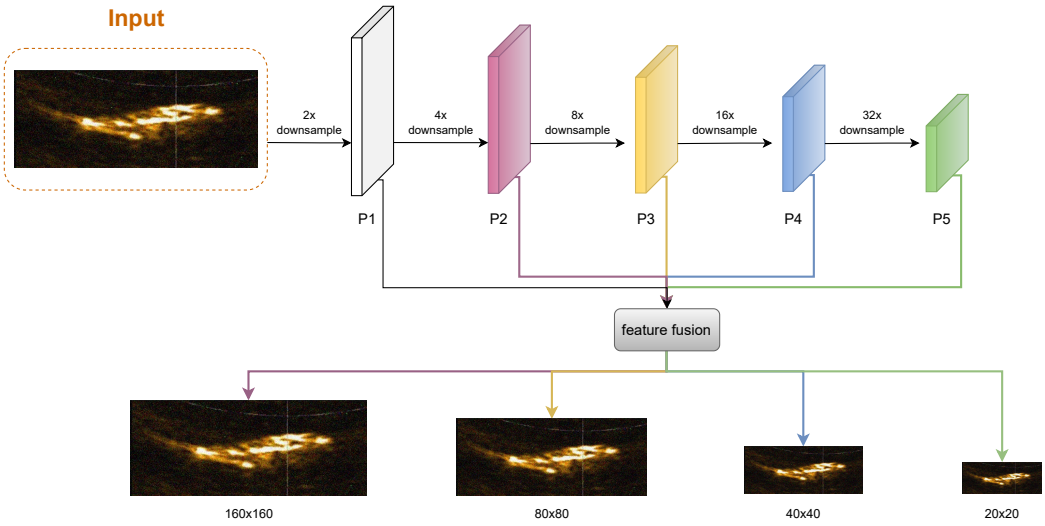



**Figure 5.** Module with the addition of P2 detection header.

When the input image size is 640×640 pixels, the respective sizes of the detection feature maps are 80×80, 40×40 and 20×20, designed for object detection of sizes 8×8, 16×16 and 32×32, respectively. The grid is divided based on the feature map's size, and each image element is predicted and assigned to a target by using the prediction box. Finally, the model stores target position and classification information in the output feature map. Nevertheless, smaller fish targets are frequently encountered in sonar object detection due to the extended sonar detection range. Much of the feature information from these objects is lost during multiple downsampling stages, and detecting these objects remains challenging despite processing them with a higher-resolution P3-layer detection head. To enhance the accuracy of smaller target detection, as depicted in Figure 5, this paper introduces an additional shallow detection layer, denoted as P2, which results in 160 × 160 detection feature maps aimed at detecting targets larger than 4×4 pixels. Following this enhancement, despite increased computational and memory demands, the detection accuracy for small objects greatly improves.

2.2.3. Improvement of NWD loss function

Considering the significant variation in sensitivity displayed by the IoU metric across objects of different sizes and the prevalent presence of small target clusters in sonar datasets, the traditional IoU calculation method tends to constrain the detection capability for small and weak targets. Although the CIoU considers the overlapping area, centroid distance and aspect ratio, the IoU-based metric is highly sensitive to position deviation in the case of small targets. Due to the fewer pixels occupied by the target, even slight position deviation can lead to a sharp change in the IoU, and the CIoU is also unsuitable for small target detection tasks. To solve this problem, a new approach is introduced to enhance the model's precision [33]. This approach utilizes the Wasserstein distance, i.e., the NWD, to assess the similarity between bounding boxes. It replaces the conventional IoU and is subsequently validated by using the Faster R-CNN architecture. The NWD utilizes a two-dimensional Gaussian distribution bounding box to quantify the similarity between Gaussian distributions. This approach is less sensitive to scale differences, making it particularly well-suited for small object detection tasks. It is worth mentioning that, owing to the NWD's capacity to apprehend finer intricacies and spatial data as compared to the IoU, it remains unaltered in its ability to assess similarity, even when confronted with extensively overlapping targets. Consequently, it is better suited for employment in underwater fish object detection within aquaculture.

First, consider the bounding boxes $A$ and $B$:

$$A = (cx_a, cy_a, \omega_a, h_a) \tag{8}$$

$$B = (cx_b, cy_b, \omega_b, h_b) \tag{9}$$

The second-order Wasserstein distance between the two-dimensional Gaussian distributions $N_a$ and $N_b$ of the bounding boxes is expressed as follows:

$$W_2^2(N_a, N_b) = \left\| ([cx_a, cy_a, \frac{\omega_a}{2}, \frac{h_a}{2}]^T, [cx_b, cy_b, \frac{\omega_b}{2}, \frac{h_b}{2}]^T) \right\|_2^2 \tag{10}$$

(The symbol T in the formula denotes the transpose operation of a matrix.)

Since $W_2^2(N_a, N_b)$ is a distance measure and cannot be directly used for similarity assessment, the NWD is obtained through exponential normalization. It quantifies the similarity between two Gaussian distributions by calculating the distance between them:

$$NWD(N_a, N_b) = exp(-\frac{\sqrt{W_2^2(N_a, N_b)}}{C}) \tag{11}$$

(C is a constant related to the data set. $NWD(N_a, N_b)$ represents the Wasserstein distance between the normal distributions of variables.)

The NWD outperforms the conventional IoU model in terms of the scale, localization and similarity assessment of occluded objects in small object detection. However, adopting the NWD approach, which entails modeling the bounding box as a two-dimensional Gaussian distribution and assessing similarity, leads to a decelerated network convergence, prolonging the training process. Consequently, substituting the IoU directly with the NWD is not the optimal course of action. Within this paper,

the amalgamation of two metrics, i.e., the IoU and NWD, affords a more comprehensive evaluation of bounding box similarity, concurrently mitigating the increase in time expenditure. Through the fine-tuning of the weight distribution between the IoU and NWD, the enhanced loss function is derived via weighted averaging.

$$LOSS_{NI} = r_{NWD} * (1 - NWD) + 1 - rNWD * (1 - IoU) \tag{12}$$

($r_{NWD}$ is the weight of the NWD and 1- $r_{NWD}$ is the weight of the IoU; the sum of the two is 1.)

### 2.2.4. Soft-NMS

Underwater fish frequently often cluster together. To enhance the precision of occluded object detection, it is necessary to enhance the NMS algorithm to lower the rate of false positives. NMS is a prevalent technique to eliminate redundant bounding boxes in object detection. In the standard NMS approach, the box with the lower score is discarded if the IoU between two bounding boxes exceeds a predefined threshold. Typically, this algorithm designates the bounding box with the highest confidence as the reference and retains boxes with minimal overlap, considering the overlapping region. Consequently, this method may fail to detect additional objects when multiple objects overlap. The NMS algorithm also needs a pairwise comparison of each bounding box, resulting in computational overhead and slower processing. It is challenging for datasets on underwater fish schooling sonar scenes, where the presence of dense objects can cause the conventional NMS algorithm to misidentify them as redundant frames, resulting in missed detections. The traditional NMS algorithm is shown in Eq (13).

$$\begin{cases} S_i = S_i & iou(M, bi) < N_t) \\ S_i = 0 & iou(M, bi) \geq N_t) \end{cases} \tag{13}$$

($S_i$ denotes the score of the current detection frame and $i$ is the ordinal number of the remaining frames except for the $M$, the frame with the most significant score sorted from highest to lowest score. $N_t$ is the specified threshold and $b_i$ is the frame to be processed.)

The soft-NMS algorithm [34] has evolved beyond the simple exclusion of bounding boxes with high overlap with the reference box when calculating bounding box overlap; it now employs a weighting mechanism. This algorithm applies the Gaussian exponent to the computed IoU values and subsequently arranges the overlapping areas based on transformed real numbers from 0 to 1. Following this, all bounding boxes receive weights and are summed according to these weights, ultimately identifying the bounding box with the highest weighted sum, as shown in Eq (14).

$$S_i = S_i e^{-\frac{iou(M,b_i)^2}{\sigma}}, \forall b_i \notin D \tag{14}$$

($D$ denotes the filtered candidate frame, and $\sigma$ is a Gaussian weight function.)

Compared to the conventional NMS algorithm, soft-NMS retains the same computational complexity as the original NMS while seamlessly integrating into the YOLOv5 framework, eliminating the need to retrain the model. This paper introduces the soft-NMS algorithm based on a Gaussian reset mechanism to replace the NMS algorithm within the YOLOv5 framework. Without increasing computational time, this substitution improves the algorithm's ability to effectively handle dense bounding boxes, consequently increasing the average detection precision and recall rates.

## 3. Data acquisition and enhancement

### *3.1. Data acquisition*

The images comprising our homemade dataset were acquired from two distinct locations: the *Luchao Diversion River Lake* and the pool adjacent to the Second Teaching Building at Shanghai Ocean University. These images were obtained by using the sonar equipment called *Oculus*, as illustrated in Figure 6. Captured videos can be saved in the *.oculus* format and converted to the *.mp4* format.

(a) Collection site (b) Sonar equipment

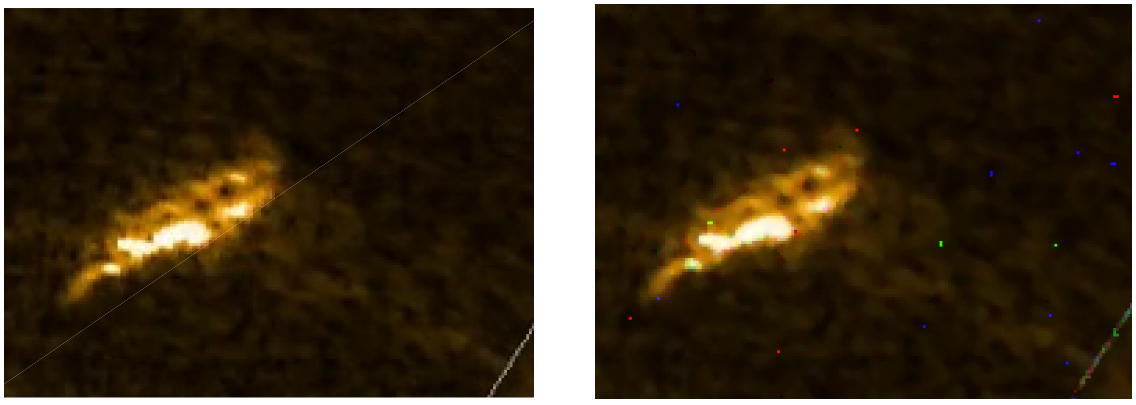
**Figure 6.** Data acquisition.

To obtain a diverse range of fish states, we conducted collections in various environments and screened the collected images in two rounds. We collected data during different times (2:00 p.m. to 5:00 p.m. and 8:00 p.m. to 11:00 p.m.) to account for the impact of lighting conditions on both sunny and rainy days to assess the influence of weather, and under conditions of with and without bait to account for the impact of bait variables on fish density and swimming speed. The captured video was processed by using an algorithm to extract frames, producing approximately 9000 images. To address issues arising from variations in video quality and the high similarity among images obtained via frame extraction, which could significantly impact fish detection accuracy, manual screening of the acquired images was necessary. First, we removed images marred by excessive frame noise from suboptimal shooting parameters. Second, images devoid of discernible fish presence were excluded, followed by the elimination of images lacking distinct fish features, such as silhouettes and swimming postures. In the first screening phase, we selected 1172 images that exhibited clear fish presence, distinct fish features, and minimal noise from a pool of 9875 images. In the second screening process, our objective was to diversify the dataset by screening various features, such as fish occlusion, densely packed schools of small and large objects, incomplete fish body parts in the photographs, and fish exhibiting differing speeds (both fast and slow). The aforementioned diversified data acquisition strategy is to obtain a more representative dataset.

### 3.2. *Image preprocessing*

Sonar detects by transmitting high-frequency sound waves, which return when they meet obstacles, and the acoustic array receives these echo signals and composes the sonar image. However, due to the complexity of the marine environment and the specificity of the sonar detection mode, a large noise interference will be generated during the detection process, which makes it challenging to perform feature extraction for object detection in sonar images. Compared to optical images, the accuracy of underwater object detection in sonar images is usually relatively low. On the one hand, the complexity of underwater acoustic signals leads to a large amount of speckle noise in imaging, which can be easily confused with the detection target. On the other hand, fish cause water ripples when swimming. When their speed becomes faster, the water ripples unfold in all directions, which affects the signal reception of the underwater sonar and makes the feature edges of the weakly textured targets in the water environment difficult to discern. In addition, since sonar relies on echo signal imaging, the imaging of the objects in the more distant positions may be blurred.

Sonar image preprocessing addresses noise issues, separates the background, mitigates interference signals and enhances the sonar image's signal-to-noise ratio, resulting in clearer images that facilitate object feature extraction. Image preprocessing is vital for sonar image detection, with denoising as a pivotal step in this process. Spatial domain denoising methods in the algorithm encompass median and mean filtering, among others. Mean filtering is easy to implement but tends to blur local features, which hinders sonar image object detection. In the context of the sonar system, both reverberation and ocean noise contribute as additive components, typically exhibiting impulse characteristics that can be effectively mitigated through median filtering. According to the literature, median filtering can not only successfully suppress impulse interference, it can retain edge details, enhancing the image's signal-to-noise ratio and reducing image noise to a certain degree [35]. Through actual verification, we found that median filtering is very suitable for sonar images. Therefore, this paper introduces the median filtering algorithm to the application of sonar image processing. In Figure 7, isolated noise points are eliminated by replacing the pixel values of a point in the adjacent image region with the median value of pixel values at that point, thus enhancing the preservation of object edge details.



(a) original figure (b) after median filtering

**Figure 7.** Comparison before and after median filtering.

### 3.3. Dataset expansion

Deep learning networks require a substantial amount of training data to effectively capture object features, as an insufficient dataset may result in network overfitting and diminished generalization capability. Considering the various factors influencing the side-scan sonar imaging process, such as fluctuating echo amplitudes and variations in imaging equipment, the images undergo random rotations, cropping, pretzel noise adding, mirror flipping and contrast adjustment to enhance the data. This dataset augmentation enhances dataset diversity and algorithm robustness. Empirical evidence has shown that if the overall data are expanded first, it may result in multiple images being produced from the same image after the expansion in both the training and validation sets in the process of dividing the training and validation sets. This may cause the network to process the same images as in the training set during the prediction process, thus inflating the prediction results. Therefore, this paper adopts the strategy of first dividing the dataset and then applying the expansion. Specifically, the entire dataset was divided into a training set and a validation set at a ratio of 8:2. When expanding the images in the validation set, only cropping, rotating, translating and other processes that do not change the images themselves were performed. After expansion, the dataset included 1580 images, and the LabelImg software was used to create the dataset labels.

## 4. Results and discussion

In this paper, an experimental platform for deep learning was built by using Ubuntu 20.04.1 as the operating system, Intel$^R$ Xeon$^R$ Platinum 8255C @2.50GHz as the host CPU, NVIDIA GeForce RTX 3080 as the GPU and 10 GB as the display RAM. This study entailed the use of the Pytorch Deep Learning framework, version 1.9.0, CUDA version 11.1 and Python 3.8.10.

### 4.1. Evaluation metrics

To objectively and comprehensively evaluate the effectiveness of the proposed method for fish detection in sonar images, we selected precision, recall and mean average precision (mAP) as the evaluation metrics. The mAP_0.5 metric refers to the mAP when the IoU threshold is 0.5. Precision assesses the model's ability to identify positive samples as positive during object detection. It quantifies the proportion of correctly predicted positive samples among all predicted positive samples. Recall quantifies the model's ability to identify all positive samples during object detection. It measures the proportion of correctly predicted positive samples among all positive samples. Equations (15) and (16) provide the formulas for the precision and recall metrics, respectively. In these equations, True and False represent the model's predictions of true and false samples, respectively, while Positives and Negatives indicate the actual true and false status during annotation, respectively. In addition, gigaflops per second (GFLOPs) were used to measure the parameter size of the model.

$$Presicion = \frac{True\ Positives}{True\ Positives + False\ Positives} \tag{15}$$

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives} \tag{16}$$

### *4.2. Experimental results of the improved model*

To ensure the performance of the model, the following hyperparameters were utilized consistently throughout this study's experiments: the total number of training rounds or epochs was set to 200, the number of batch samples, or batch size, is 8, the initial learning rate is 0.01, the number of rounds of the warm-up learning parameter is 3, the momentum of the warm-up learning rate is 0.8 and the training method is stochastic gradient descent.

#### 4.2.1. Baseline model selection

YOLOv5 mainly has four official base models, i.e., YOLOv5s, YOLOv5m, YOLOv5l and YOLOv5x, and the only difference between these models is that the depth and width of their networks are gradually increasing. The larger the depth and width, the stronger the feature extraction ability, and the number of parameters and floating-point operations are also increased accordingly. The self-constructed sonar fish dataset was trained by using the above official base models, and the results of the indicators are shown in Table 1.

**Table 1.** Performance comparison for different official YOLOv5 versions.

| Model | mAP_0.5/% | Recall/% | Parameters/M | GFLOPs |
|---|---|---|---|---|
| YOLOv5s | 75.1 | 73.9 | 7.01 | 15.8 |
| YOLOv5m | 78.1 | 75.3 | 20.85 | 47.9 |
| YOLOv5l | 74.8 | 72.5 | 46.10 | 107.6 |
| YOLOv5x | 75.9 | 68.1 | 86.17 | 203.8 |

Since sonar images are more blurred and have more noise interference than optical video images, deep learning object detection in sonar images is relatively more complex, and the overall index is relatively low. Therefore, to improve the precision of object detection in sonar images, a relatively deeper network should be selected to enhance the ability of model feature extraction. While the YOLOv5s model has the shallowest depth, its precision metrics are notably low. Conversely, YOLOv5l and YOLOv5x possess deeper model architectures, yet, their metrics, including mAP_0.5 and recall, do not surpass those of YOLOv5m. Although the YOLOv5m model exhibits a relative increase in GFLOPs, it achieves commendable overall metrics. Despite the GFLOPs increment in YOLOv5m, its mAP_0.5 exceeds that of YOLOv5s by 3%, and the number of parameters and GFLOPs remained within acceptable bounds. After comprehensive table data analysis, we selected YOLOv5m as the baseline model for subsequent algorithm enhancements.

#### 4.2.2. Optimized algorithm results

The improved model achieved a precision of 76.4%, a recall of 80.0% and a mAP_0.5 of 80.8%. These results represent improvements of 2.3%, 4.7% and 2.7%, respectively, relative to the pre-improvement YOLOv5m model, and 4.3%, 6.1% and 5.4% improvements relative to the YOLOv5s model. The loss function curve in Figure 8 also illustrates that the model begins to converge around 150 rounds, with the curve gradually flattening and the loss function gently approaching the X axis.

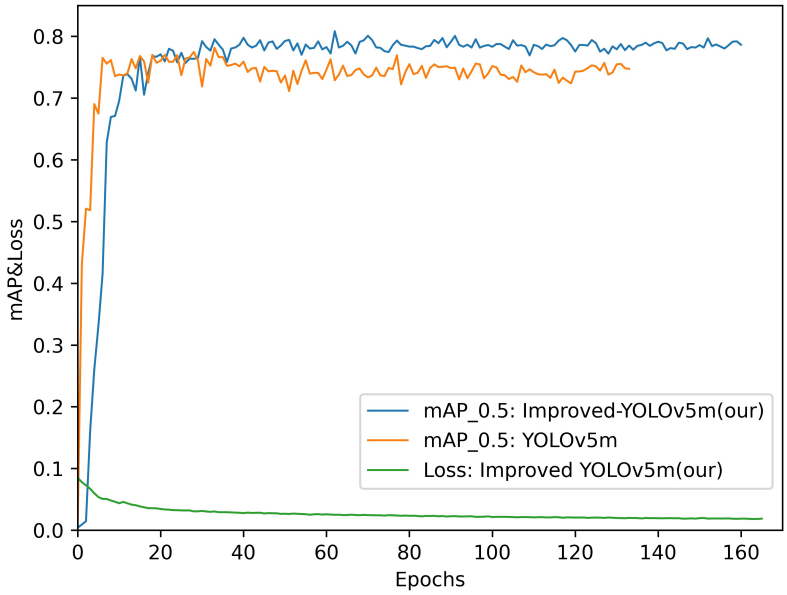
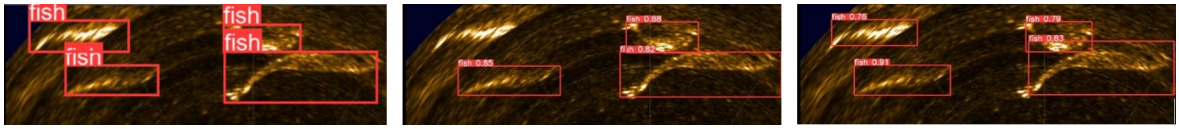
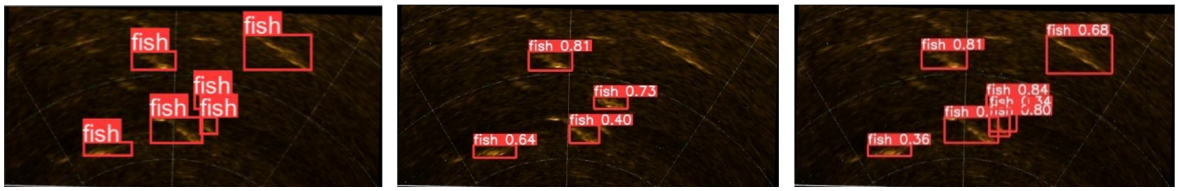



**Figure 8.** Comparison of algorithm improvement results.

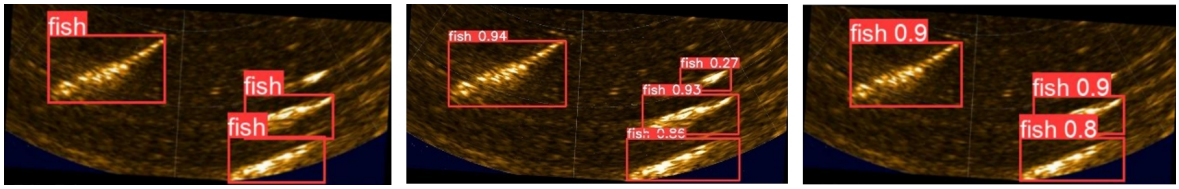
In order to verify the effectiveness of the improved algorithm and exclude the possibility of accuracy enhancement due to an increase in the false detection rate, three pictures were randomly selected from the dataset to be detected.



(a) Reduction of missed detections



(b) Reduction of missed detections (small targets)



(c) Reduction of false positives

**Figure 9.** Comparison of detection results before and after improvement. (The left-middle-right parts of the comparison of detection results in (a)–(c) are the actual fish object, the pre-improvement detection results and the post-improvement detection results, respectively.)

In Figure 9, the leftmost side is all of the actual fish objects, and the three pictures have four, six and

three objects, respectively. The middle picture and the rightmost picture are the YOLOv5 detection results before and after the improvement, respectively. It can be seen that the baseline model misses one fish object in Figure 9(a), misses two smaller fish objects in Figure 9(b) and misjudges one fish object in Figure 9(c), and these are all improved in the improved model. The results show that the improved algorithm decreases the rates of missed detections and false positives while enhancing the precision and mAP.

### *4.3. Model validation*

#### 4.3.1. Comparison of C3N performance in different positions

We performed replacement experiments to assess the performance advantages and disadvantages of integrating the C3N module into various positions within the improved YOLOv5m network, substituting the module in different locations. As presented in Table 2, the results reveal that when the C3 module in the neck position is replaced with the C3N module, improvements are observed in the recall and mAP_0.5 relative to the models with other replacements. It is important to note that a trade-off exists between precision and recall. Substituting the neck position enhances recall while marginally sacrificing precision , as compared to replacing the backbone position. Nonetheless, compared to the baseline, precision still sees a 2.3% enhancement, accompanied by improvements in the recall and mAP_0.5 to 80%, which is significant. This outcome affirms that the enhanced C3N module, derived from the ConvNeXt model, augments the model's expressiveness and feature extraction capabilities by augmenting network depth and complexity, thus enhancing average precision at a small computational cost. These findings suggest the efficacy of introducing the C3N enhancement at the neck location within our algorithm, enabling more efficient feature extraction and information retention mechanisms for superior performance.

**Table 2.** Comparison of C3N replacement positions.

| C3N replacement position | Precision/% | Recall/% | mAP_0.5/% |
|---|---|---|---|
| Improved model (without C3N) | 77.3 | 74.7 | 78.3 |
| C3N_Backbne | **78.5** | 74.7 | 77.9 |
| C3N_Neck(our) | 76.4 | **80.0** | **80.8** |
| C3N_all | 69.4 | 63 | 65.3 |

#### 4.3.2. Ablation experiments

To enhance the accuracy and efficiency of fish detection, we developed an improved network model based on the original YOLOv5m model. The improvement involves several essential modifications: replacing the C3 module in the neck section with an improved C3N module based on ConvNeXt, incorporating a shallow detection layer, introducing a new metric called the NWD combined with the IoU to compute the loss function and upgrading the NMS to soft-NMS. To assess the impact of these enhancements, we conducted four sets of ablation experiments (Experiments 1-4), the results of which are presented in Table 3.

As can be seen in Table 3, comparing the results of Experiments 1–3 with the improved results, it is found that the mAP of the model was improved by 2.5% after the introduction of the C3N module in the neck section, which indicates that the model with the addition of the C3N can more accurately detect

the fish information in the sonar images. The recall increased by 5.3%, indicating that fish objects can be detected more comprehensively. With the addition of a shallow detection layer, the model can learn more information about small object features, which improves small object detection accuracy. After introducing the NWD fusion IoU structure, the model recall increased by 2.4%, the mAP increased by 2% and the precision increased by 0.9%. A comparison between the results of Experiment 4 and the pre-improvement outcomes revealed a notable 2.7% increase in precision after substituting the original NMS in the YOLOv5m model with the enhanced soft-NMS. This substitution effectively addresses the challenge of detecting fish features obscured within densely packed bounding boxes. The improved model proposed in this paper achieved a mAP_0.5 of 80.8%, marking a 1.9% improvement over the original model. Furthermore, the recall achieved a substantial 4.7% increase, and precision experienced a notable 2.3% gain. The data comparison above demonstrates that our model surpasses YOLOv5m in terms of feature extraction capability, multi-scale fusion performance for sonar fish detection and overall performance.

**Table 3.** Ablation experiments.

| Experiment | C3N | P2 | NWD | Soft-NMS | Precision/% | Recall/% | mAP_0.5/% |
|---|---|---|---|---|---|---|---|
| Ours | ✓ | ✓ | ✓ | ✓ | 76.4 | **80.0** | **80.8** |
| Experiment 1 | – | ✓ | ✓ | ✓ | **77.3** | 74.7 | 78.3 |
| Experiment 2 | ✓ | – | ✓ | ✓ | 75.8 | 79.2 | 79.7 |
| Experiment 3 | ✓ | ✓ | – | ✓ | 75.5 | 77.6 | 78.8 |
| YOLOv5m | – | – | – | – | 74.1 | 75.3 | 78.1 |
| Experiment 4 | – | – | – | ✓ | 76.8 | 75.2 | 77.1 |

Note: ✓ means structure added, – means no structure added.

#### 4.3.3. Comparison experiments with other models

To validate the effectiveness of the proposed model, the improved YOLOv5m model was tested in comparison with other mainstream detection models. All models underwent testing by using identical datasets and training devices, following the principle of controlling variables. The results are presented in Table 4 and Figure 10.

**Table 4.** Algorithm comparison results.

| Model | Precision/% | Recall/% | mAP_0.5/% | GFLOPS |
|---|---|---|---|---|
| YOLOv3 | 73.9 | 73.7 | 74.1 | 154.5 |
| YOLOv5m | 74.1 | 75.3 | 78.1 | 47.9 |
| YOLOv6 | 75.1 | 75.1 | 79.5 | 45.17 |
| YOLOv7 | 74.9 | 76.5 | 79.7 | 103.2 |
| YOLOv8 | 70.6 | 74.7 | 75.8 | 28.4 |
| Ours | 76.4 | 80.0 | 80.8 | 53.1 |

As can be seen in Table 4, the improved YOLOv5m model demonstrated the best precision in sonar fish object detection. As compared to YOLOv3, YOLOv6, YOLOv7 and YOLOv8 models, the improved model achieved increases in mAP_0.5 of 6.7%, 1.3%, 1.1% and 5%, respectively, positioning it as the leader in both precision and recall. Acknowledging the complexity of sonar images, our

model incorporates a small object detection head and a C3N module with a more intricate convolutional structure. This leads to a slightly larger number of GFLOPs than some models. Nevertheless, our focus on enhancing precision in sonar fish detection justifies this minor increase in complexity, making it feasible for practical applications. Despite sacrificing some speed as a result of increased convolutional layers and model depth, the improved model outperformed YOLOv8 across all accuracy metrics, enabling the prediction of more fish objects. Consequently, our model ensures the highest mAP and optimal overall performance while balancing precision and recall.

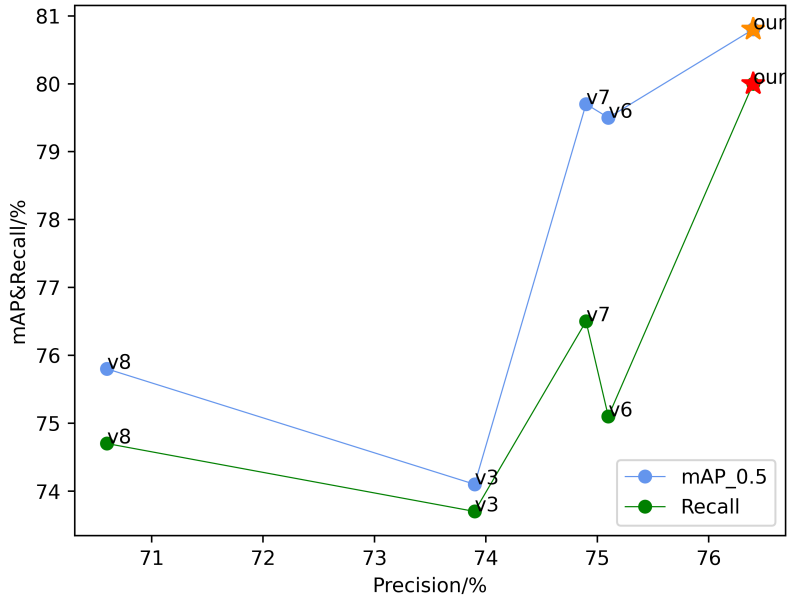



**Figure 10.** Comparison of algorithms.

## 5. Conclusions

In this paper, we have improved the fish object detection algorithm, building on YOLOv5m. Our focus has been on addressing issues related to the poor recognition of smaller and occluded fish within sonar images. Our improvements encompass three key aspects: feature information, loss function and dense fish recognition.

1) First, we addressed the challenges of target blurring caused by underwater noise and the inadequate target feature information inherent in sonar image methods. Median filtering denoising was applied to our self-constructed dataset to enrich data diversity. Furthermore, the enhanced C3N module was introduced into the neck network to minimize feature information loss during downsampling and forward propagation. We substantiated the effectiveness of this approach through comparative experiments, showcasing its superiority over alternative network positions. The original YOLOv5m network was expanded by adding a P2 shallow feature layer to enable fish object detection at four scales.

2) To calculate the model loss function such that it is more suitable for small objects, the NWD was combined with the original IoU to help the model capture more details and spatial information.

3) Finally, the traditional NMS was replaced with soft-NMS to reduce missed detection rates when fish objects overlap, enhancing the model's accuracy. Ablation experiments demonstrate that the improved YOLOv5m model exhibits more excellent enhancements in precision, mAP and recall com-

pared to other YOLO model versions.

Nevertheless, there is potential for further optimization of the enhanced model. Unfortunately, this study did not include a sonar image dataset from the distant sea due to the limited experimental conditions. While this limitation does not impact the current advancements and validation of the fish detection algorithm for sonar images, it highlights the need for further exploration within this application domain. Additionally, the inherent challenges in sonar image object detection, particularly for small fish targets, as compared to optical images, necessitate the adoption of a deeper network in this study. Although this deepening is accompanied by enhancements to the C3N module to improve feature extraction, it does come at the cost of reduced detection speed. Hence, there remains room for improvement in both accuracy and speed. In future research, we intend to optimize the network model and explore network pruning methods to streamline its architecture.

## Use of AI tools declaration

The authors declare that they have not used Artificial Intelligence tools in the creation of this article.

## Acknowledgments

This work was supported by Shanghai Science and Technology Committee (STCSM) Local Universities Capacity-Building Project (No. 22010502200).

## Conflict of interest

The authors declare that there is no conflict of interest.

## References

1. FAO, The state of world fisheries and aquaculture 2016: Opportunities and challenges, Rome: Food and Agriculture Organization of the United Nations, (2007).
2. FAO, The state of world fisheries and aquaculture 2022: Towards blue transformation, Food and Agriculture Organization of the United Nations, (2022).
3. FAO, The State of World Fisheries and Aquaculture, Food and Agriculture Organization of the United Nations, (2018).
4. J. Álvarez, J. M. F. Real, F. Guarner, M. Gueimonde, J. M. Rodríguez, M. S. de Pipaon, et al., Microbiota intestinal y salud, *Gastroenterología y Hepatología*, **44** (2021), 519–535. https://doi.org/10.1016/j.gastrohep.2021.01.009
5. R. Lulijwa, E. J. Rupia, A. C. Alfaro, Antibiotic use in aquaculture, policies and regulation, health and environmental risks: A review of the top 15 major producers, *Rev. Aquacult.*, **12** (2020), 640–663. https://doi.org/10.1111/raq.12344
6. J. D. Sachs, C. Kroll, G. Lafortune, G. Fuller, F. Woelm, *Sustainable Development Report 2022*, Cambridge University Press, 2022. https://doi.org/10.1017/9781009210058
7. National oceanic and atmospheric administration,*Natl. Weather Serv.*, (2012), 1950–2011.

8. F. Yang, Z. Du, Z. Wu, Object recognizing on sonar image based on histogram and geometric feature, *Mar. Sci. Bull. Tianjin*, **25** (2006), 64.
9. R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2014), 580–587. https://doi.org/10.1109/CVPR.2014.81
10. R. Girshick, Fast R-CNN, in *Proceedings of the IEEE International Conference on Computer Vision*, (2015), 1440–1448. https://doi.org/10.1109/ICCV.2015.169
11. S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, *Adv. Neural Inf. Process. Syst.*, **28** (2015).
12. K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, *IEEE Trans. Pattern Anal. Machine Intell.*, **37** (2015), 1904–1916. https://doi.org/10.1109/TPAMI.2015.2389824
13. K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in *Proceedings of the IEEE International Conference on Computer Vision*, (2017), 2961–2969.
14. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 779–788. https://doi.org/10.1109/CVPR.2016.91
15. J. Redmon, A. Farhadi, Yolo9000: Better, faster, stronger, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 7263–7271.
16. J. Redmon, A. Farhadi, Yolov3: An incremental improvement, preprint, arXiv: 1804.02767.
17. A. Bochkovskiy, C. Y. Wang, H. Y. M. Liao, Yolov4: Optimal speed and accuracy of object detection, preprint, arXiv: 2004.10934.
18. C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, et al., Yolov6: A single-stage object detection framework for industrial applications, preprint, arXiv: 2209.02976.
19. C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao, Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, (2023), 7464–7475.
20. K. Tong, Y. Wu, F. Zhou, Recent advances in small object detection based on deep learning: A review, *Image Vision Comput.*, **97** (2020), 103910. https://doi.org/10.1016/j.imavis.2020.103910
21. I. Karoui, I. Quidu, M. Legris, Automatic sea-surface obstacle detection and tracking in forward-looking sonar image sequences, *IEEE Trans. Geosci. Remote Sens.*, **53** (2015), 4661–4669. https://doi.org/10.1109/TGRS.2015.2405672
22. X. Wang, Q. Li, J. Yin, X. Han, W. Hao, An adaptive denoising and detection approach for underwater sonar image, *Remote Sens.*, **11** (2019), 396. https://doi.org/10.3390/rs11040396
23. T. Yulin, S. Jin, G. Bian, Y. Zhang, Shipwreck target recognition in side-scan sonar images by improved yolov3 model based on transfer learning, *IEEE Access*, **8** (2020), 173450–173460. https://doi.org/10.1109/ACCESS.2020.3024813
24. Y. Yu, J. Zhao, Q. Gong, C. Huang, G. Zheng, J. Ma, Real-time underwater maritime object detection in side-scan sonar images based on transformer-yolov5, *Remote Sens.*, **13** (2021), 3555. https://doi.org/10.3390/rs13183555

25. T. Jin, X. Yang, Monotonicity theorem for the uncertain fractional differential equation and application to uncertain financial market, *Math. Comput. Simul.*, **190** (2021), 203–221.https://doi.org/10.1016/j.matcom.2021.05.018

26. J. Yang, Y. Zhang, T. Jin, Z. Lei, Y. Todo, S. Gao, Maximum lyapunov exponent-based multiple chaotic slime mold algorithm for real-world optimization, *Sci. Rep.*, **13** (2023), 12744. https://doi.org/10.1038/s41598-023-40080-1

27. S. Liu, L. Qi, H. Qin, J. Shi, J. Jia, Path aggregation network for instance segmentation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 8759–8768. https://doi.org/10.1109/CVPR.2018.00913

28. T. Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 2117–2125.

29. Z. Liu, H. Mao, C. Y. Wu, C. Feichtenhofer, T. Darrell, S. Xie, A convnet for the 2020s, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (2022), 11976–11986. https://doi.org/10.1109/CVPR52688.2022.01167

30. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778.

31. Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, et al., Swin transformer: Hierarchical vision transformer using shifted windows, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, 10012–10022. https://doi.org/10.1109/ICCV48922.2021.00986

32. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L. C. Chen, Mobilenetv2: Inverted residuals and linear bottlenecks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2018), 4510–4520. https://doi.org/10.1109/CVPR.2018.00474

33. J. Wang, C. Xu, W. Yang, L. Yu, A normalized gaussian wasserstein distance for tiny object detection, preprint, arXiv:2110.13389.

34. N. Bodla, B. Singh, R. Chellappa, L. S. Davis, Soft-nms–improving object detection with one line of code, in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, 5561–5569. https://doi.org/10.1109/ICCV.2017.593

35. A. Kumar, S. S. Sodhi, Comparative analysis of gaussian filter, median filter and denoise autoenocoder, in *2020 7th International Conference on Computing for Sustainable Global Development*, (2020), 45–51. https://doi.org/10.23919/INDIACom49435.2020.9083712


AIMS Press