IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 9 July 2024, accepted 28 July 2024, date of publication 31 July 2024, date of current version 8 August 2024.

*Digital Object Identifier 10.1109/ACCESS.2024.3436073*

RESEARCH ARTICLE

# Underwater Target Detection Based on Improved YOLOv7 Algorithm With BiFusion Neck Structure and MPDIoU Loss Function

**JINYU OU AND YIJUN SHEN**
School of Mechanical and Electrical Engineering, Hainan University, Haikou 570228, China
Corresponding author: Yijun Shen (15393061293@163.com)

**ABSTRACT** Underwater target detection has developed greatly in recent years. However, the accuracy of underwater target detection is limited by the complex underwater environment. Based on YOLOv7, we propose an underwater object detection algorithm model to improve precision and confidence (BiFusion Neck module and a MPDIoU loss function). Compared to traditional networks module, the Bifusion Neck module preserves more features from the lower layers by utilizing the output of the P2 feature layer. Moreover, the loss function was improved on the basis of IoU introducing Minimum Point Distance. Finally, the LSKA attention mechanism is introduced to enhance the feature extraction of targets at different scales. The experimental results demonstrate that the BFD-YOLO model proposed in this study achieves an average detection accuracy(mAP50) of 84.8% on a customized dataset, surpassing the performance of the YOLOv7 algorithm by 11.5% and outperforming other tested algorithms. Furthermore, the BFD-YOLO algorithm exhibits strong performance on various datasets and demonstrates superior generalization capabilities.

**INDEX TERMS** Underwater object detection, YOLOv7, BiFusion, MPDIoU, BFD-YOLO, LSKA.

## I. INTRODUCTION

The deterioration of the ecological environment has led many developed countries to pay more attention to the vast oceans. The exploration and utilization of the ocean, especially the deep sea, involves a wide range of activities including resource extraction, deep-sea fishing, cultural heritage protection, and national defense security. These activities urgently require the support of advanced underwater optical and acoustic technologies, revealing the rich scientific research value contained within. However, traditional manual diving fishing methods are fraught with limitations, including high risk, low efficiency, and environmental damage to the seabed ecosystem. The development of a robotic system capable of autonomously completing underwater target recognition and intelligent fishing tasks becomes crucial. Such an automated solution can not only improve operational efficiency and reduce costs but also help protect the marine ecological environment and promote the sustainable development of marine fisheries.

One of the keys to underwater robot technology is its sensing and detection capabilities, which rely on sonar detection and optical imaging technologies. Although sonar detection technology is an indispensable method for the positioning and recognition of underwater targets, it faces challenges when dealing with complex underwater environments and external noise interference. In contrast, underwater optical imaging technology, with its intuitive, high-resolution, and cost-effective advantages, has become the preferred tool for close-range environmental detection. However, the unique physical characteristics of underwater environments, such as light attenuation, scattering, and absorption, along with complex ecological conditions, such as the significant size difference and dense distribution of aquatic organisms, pose additional challenges to underwater target detection [1], [2], [3].

Since the 1950s, the exploration of underwater target detection algorithms has been a focal point of research across the

The associate editor coordinating the review of this manuscript and approving it for publication was Turgay Celik.

globe, yielding a multitude of advancements. Initial endeavors predominantly leveraged traditional machine learning algorithms, necessitating processes such as region selection, feature extraction, and classifier design for operational functionality. For instance, Mathur and Goel [4] developed an underwater remotely operated vehicle (ROV) predicated on RGB and HSV color models to utilize the color attributes of objects for detection. Despite achieving basic detection capabilities, this method was susceptible to interference from similarly colored background objects. Zhu et al. [5] addressed light path considerations between the light source and camera, employing color compatibility for object detection to process images directly, thereby diminishing computational demands and enhancing real-time performance and stability. González-Sabbagh and Robles-Kelly [6] introduced a physical autonomous underwater vehicle equipped with a binocular vision system for image capture, focusing on shape feature extraction and center feature identification of objects, supplemented by an optical flow algorithm for motion speed detection within images. Mittal et al. [7] explored the classification of texture information in images through the use of multi-layer perceptron (MLP) neural networks and support vector machines (SVM), incorporating Canyon edge detection and Hough transform despite facing challenges with the latter's practical application. Moradi [8] delved into the analysis of color and texture characteristics in underwater fish videos to enable unrestricted environmental fish number calculation and detection. Moreover, Huang et al. [9] proposed an innovative automatic active contour detection method, capitalizing on color image information and overcoming detection hurdles posed by lighting conditions and object shadows through the integration of a visual attention mechanism and an optimal area selection strategy for initialization, thus expediting algorithm convergence. Xu et al. [10] proposed a vision-based underwater positioning technology, adopting a novel approach in the image preprocessing stage with weighted and correlation coefficients, alongside an image color segmentation method, to address the robustness issues associated with traditional template matching techniques under varying lighting conditions and affine transformations.

Despite the simplicity and minimal data requirements of traditional underwater target detection methods, they often falter in practical application due to two main drawbacks: the non-targeted nature and high computational complexity of the sliding window mechanism for area selection,leading to inefficiencies; and the poor robustness and low identification accuracy of artificially designed feature algorithms when contending with scale changes, occlusions, and other interference factors. Conversely, recent years have seen a paradigm shift towards deep learning solutions, which, by constructing deep machine learning models and utilizing extensive training datasets, can autonomously extract more critical and comprehensive feature representations. This shift has significantly improved the accuracy of classification and detection tasks in underwater environments. Notably, Han et al. [11] combined max-RGB and grayscale transformation techniques to enhance underwater image quality before employing an innovative deep convolutional neural network architecture for target detection, showcasing superior performance over models like YOLOv3 and Faster RCNN. Lin et al. [12] introduced the RoIMix image enhancement algorithm to better simulate target overlap, occlusion, and blur, thereby generating a detection model with enhanced generalizability across various scenes. Li et al. [13] designed a convolutional neural network tailored for underwater image enhancement to construct a high-quality underwater image training set, demonstrating robust performance across diverse underwater scenes. Ren et al. [14] applied the YOLOv5 detection algorithm to underwater target detection, innovatively integrating a Swin Transformer into the detection framework, which exhibited exceptional effectiveness in complex underwater scenarios. Islam et al. [15] employed a supervised adversarial training approach, using generative adversarial networks to improve the quality of underwater original images, thereby enhancing the subsequent target detection dataset's overall performance. Finally, Bochkovskiy et al. [16] developed a network structure—Sample-Weighted hyPEr Network (SWIPENet)—specifically for detecting small underwater sample targets. The network's novel weighted loss function, Invert Multi-Class Adaboost (IMA), significantly mitigated the impact of noise on SWIPENet's performance, as evidenced by experimental results on URPC2017 and URPC2018, two authoritative underwater robot competition datasets, underscoring SWIPENet and IMA's substantial superiority over existing underwater target detection algorithms.

In terms of above research background, this topic focuses on the underwater seafood in the underwater environment, including image blurred, overlapping target, small target number of realisticchallenges. We plan to propose and design a underwater target detection scheme based on YOLOV7algorithm. The innovation from several key technology links, include network structure design and loss function refinement optimization to build a real-time and accurate identification in underwater target model. Through this method, the existing technical bottlenecks are expected to be overcome, significantly enhancing the accuracy and efficiency of underwater target detection. This will provide robust technical support and practical value, advancing the modernization and intelligent processes of marine fisheries.

This study focuses on the core issue of improving the accuracy and efficiency of underwater object detection, employing deep convolutional neural networks as the foundational architecture. Through the introduction of the Bifusion Neck mechanism and a novel bounding box loss function, MPDIoU, the YOLOv7 object detection algorithm has been innovatively improved. The contributions of this research are primarily reflected in the following three aspects:

1. This study proposes an improved YOLOv7 algorithm for deep-sea fish detection.
2. The introduction of the Bifusion Neck mechanism enhances feature fusion efficiency, boosting the model's ability to recognize target features in complex underwater environments.
3. A novel MPDIoU loss function is proposed, which significantly improves the precision and speed of bounding box localization while maintaining computational efficiency.
4. A new attention mechanism, LSKA, is incorporated into the backbone network, enhancing feature extraction capabilities and improving detection accuracy.
5. Extensive experiments demonstrate that the improved YOLOv7 algorithm significantly outperforms the original algorithm in underwater object detection tasks, particularly in low-light conditions.

The paper is composed of five sections. The first section introduces the background and current research status of marine fish detection. The second section presents the structure of YOLOv7. The third section discusses the improved methods used in this study and the issues they address. The fourth section conducts extensive experiments to demonstrate the effectiveness and advancement of the proposed improvement algorithms. The fifth section provides a summary of the content and offers further prospects for future research.

## II. PREPARATIONYOLOV7 THE DETECTION ALGORITHM AND ITS IMPROVEMENTS

As the latest iteration of the YOLO series algorithms, YOLOv7 has exhibited unprecedented efficiency and precision in object detection under conventional environments. However, when addressing the unique challenges of underwater environments, such as complex target characteristics, light scattering, and image distortion, the original detection performance of YOLOv7 is limited and cannot achieve the desired recognition effect. To solve this problem, this study specifically investigates underwater object detection and makes key improvements to the YOLOv7 model. Innovatively, it introduces a Bifusion mechanism that integrates multi-level and multi-dimensional feature information, effectively enhancing the model's ability to capture subtle features of underwater targets. It also utilizes an improved multi-point IoU (mpDIoU) bounding box loss function, which can more finely assess the overlap between predicted and real bounding boxes. This is particularly suitable for handling underwater targets of varying sizes and shapes, thereby significantly improving the model's accuracy and stability in locating underwater targets. These improvements make YOLOv7 more suited to the needs of underwater object detection in terms of network structure and optimization strategies, and are expected to significantly enhance the performance of underwater object detection.

### A. THE YOLOV7 DETECTION ALGORITHM

YOLOv7 is a basic network model in the YOLO series, with its most notable feature being its high-speed detection capability. To enhance the accuracy and robustness of object detection, YOLOv7 incorporates new feature fusion and context information capturing techniques. These techniques enable YOLOv7 to demonstrate an outstanding performance balance across various detection metrics. Compared to other known detection networks, YOLOv7 shows higher levels of speed and accuracy within the range of 5-160 frames per second, making it widely applicable in different scenarios. Additionally, YOLOv7 supports deployment on various hardware platforms, including edge GPUs, standard GPUs, and cloud GPUs, allowing users to select different network models based on their needs (YOLOv7-tiny, YOLOv7, and YOLOv7-w6). The detection approach of YOLOv7 is similar to that of YOLOv4 and YOLOv5. The YOLOv7 network structure is illustrated in Figure 1. The network is primarily composed of three parts: the Backbone (main network), Neck, and Head.

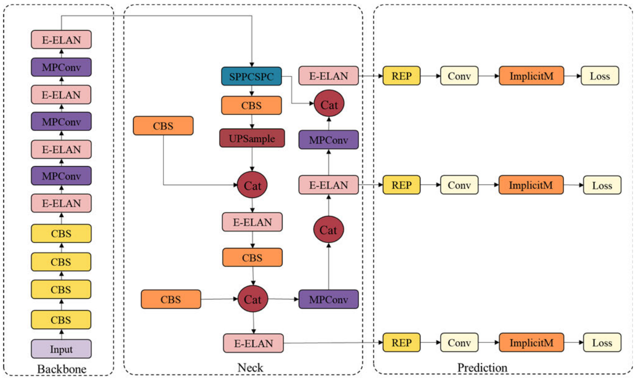


**FIGURE 1.** YOLOv7 structure diagram of the network.

Backbone: The backbone of YOLOv7 is responsible for extracting features from input images, and its design significantly influences the model's performance and efficiency. YOLOv7 utilizes CSPDarknet53 as its backbone, comprising CBS, MPconv, and ELAN structures. The CBS module combines CONV, BN, and Silu activation function to extract feature information from the image through convolution. ELAN integrates the design principles of VoVNet and CSPNet to enhance gradient lengths by optimizing the stacking structure of computational blocks. Its purpose is to optimize feature extraction, gradient flow, and improve object detection performance. MPconv primarily reduces channel numbers of feature maps while preserving resolution, effectively reducing computational cost and enhancing network efficiency.

Neck: The neck component of YOLOv7 primarily consists of SPPCSP modules and optimized PAN modules. These modules further integrate and process the features extracted by the backbone, facilitating more accurate object detection in the subsequent Head section. The SPPCSP module enhances feature information by incorporating the concat operation with the feature maps before the SPP module, based on the SPP (Spatial Pyramid Pooling) module. The optimized PAN module manages information flow between feature maps of different scales, leading to more efficient feature fusion and improved utilization efficiency, ultimately enhancing the accuracy of object detection.

Head: The head section of YOLOv7 serves the crucial role of generating the final prediction output by processing the features forwarded from the backbone and neck components to determine class and position information for the targets.

## B. YOLOV7 IMPROVEMENT

### 1) WHOLE FRAME

This section presents several improvement strategies implemented in this study. An object detection algorithm named BFD-YOLO is proposed based on YOLOv7. BFD-YOLO enhances the evaluation of detection accuracy for small targets by utilizing MPDIOU, and it also conducts a more comprehensive assessment of detection accuracy for objects of different sizes. By effectively handling objects of various sizes, MPDIOU improves the model's ability to generalize in practical applications. Moreover, the problem of imbalanced feature hierarchy is addressed through the incorporation of bifusion neck, which enhances the representation capability of features. In addition, the LSKA attention mechanism is added after SPPCSPC to enhance the feature extraction capability. The structure of BFD-YOLO is depicted in Figure 2.

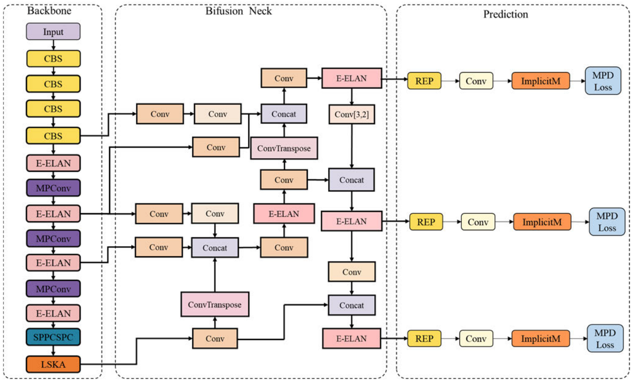


**FIGURE 2.** **BFD-YOLO structure diagram of the network.**

### 2) IMPROVEMENT OF THE LOSS FUNCTION

The purpose of border regression is to approach the true detection window by fine-tuning the detection window of the detector output. IoU (Intersection over Union) has become the mainstream standard for evaluating the loss of the predictive box in the detection field since it was proposed. The IoU formula is shown in Equation (1).

$$IoU(pbb, gbb) = \frac{\text{Area}(pbb \cap gbb)}{\text{Area}(pbb \cup gbb)} \tag{1}$$

In formula 1, *pbb* represents the predicted bounding box, and *pbb* denotes the ground truth bounding box.

The IOU function, however, exhibits several issues that affect its convergence speed and accuracy. Firstly, the IOU loss function can yield identical values for different predicted results, thereby diminishing the convergence speed and accuracy of bounding box regression.Different predicted bounding boxes (i.e., prediction results) can achieve the same IOU (Intersection Over Union) value, meaning these distinct predictions are treated as equivalent during the optimization process. Since different prediction results receive the same loss value, the model cannot differentiate between these predictions, which hinders the optimization process from correctly adjusting the model parameters. More precise localization: Secondly, it may result in unfair treatment towards objects of different sizes, as the function solely considers the overlapping region and disregards the size of the objects. Furthermore, a gradient vanishing problem may arise in certain scenarios, particularly when there is no overlap between the predicted box and the ground truth box. Lastly, the computational complexity of the IOU loss function becomes notably high when handling multiple overlapping objects or densely packed scenes.

To address the aforementioned issues, MPDIoU is used as the loss function for YOLOv7. MPDIoU Is improved by the IOU. The MPDIoU loss function aims to further optimize the quality of the bounding box regression by introducing the concept of minimum point distance, especially when the object has a complex geometry. MPDIoU All four corners of the bounding box are considered, and the distance of the furthest point pair between the prediction box and the real box is calculated, and included in the design of the loss function to facilitate the refined adjustment of the model to the edge position of the bounding box during training. MPDIoU The calculation is shown in equation (2):

$$\begin{aligned} d_1^2 &= \left(x_1^B - x_1^A\right)^2 + \left(y_1^B - y_1^A\right)^2 \\ d_2^2 &= \left(x_2^B - x_2^A\right)^2 + \left(y_2^B - y_2^A\right)^2 \\ \text{MPDIOU} &= \frac{A \cap B}{A \cup B} - \frac{d_1^2}{w^2 + h^2} - \frac{d_2^2}{w^2 + h^2} \end{aligned} \tag{2}$$

In the formula 2 A,B—Represents two arbitrary convex shapes, respectively w, $h$—Enter the width and height of the image; $\left(x_1^A, y_1^A\right)$, $\left(x_2^A, y_2^A\right)$—Represents the upper left and lower right point coordinates of A, respectively; $d_1^2, d_2^2$—The square of the Euclidean distance between the upper left and the distance between the points of A and right of B, respectively.

The MPDIoU loss function simplifies the similarity comparison between two bounding boxes by using more geometric constraint information, and helps to adjust the position and size of the bounding boxes with less training samples, thus improving the accuracy of target detection.

In summary, IoU (Intersection Over Union) calculation requires finding the intersection and union areas of two bounding boxes, which involves complex geometric computations, especially when the bounding boxes partially overlap. MPDIoU (Minimum Pairwise Distance Intersection Over Union) simplifies this process by calculating the distance between the center points of the bounding boxes, eliminating the need to directly compute the intersection and union areas. This significantly reduces the computational load, particularly when dealing with a large number of bounding boxes. MPDIoU introduces the pairwise distance between the center points of the bounding boxes, which more accurately reflects the relative positional relationship between the two boxes. This approach offers higher detection accuracy in scenarios where the bounding boxes are very close in position but do not completely overlap in shape. Traditional IoU has a higher computational complexity, especially in high-resolution images and scenes with many objects. By using simpler distance calculations and weighting methods, MPDIoU lowers the overall computational complexity and improves processing speed.Traditional IoU is sensitive to cases where bounding boxes partially overlap but differ significantly in area, which can lead to substantial errors. MPDIoU considers the shapes and positions of the bounding boxes during its calculation, reducing this type of error to some extent and enhancing overall robustness.

### 3) NECK IMPROVEMENTS

Object detection methods based on deep learning have made significant progress, with detection networks becoming increasingly powerful in terms of architecture design and training strategies. However, most research still relies on superior backbone designs, leading to insufficient information exchange between high-level and low-level features. Therefore, a design paradigm based on a light backbone and a heavy neck holds equal importance in detection tasks.

In recent years, the primary research direction for the Neck has been the utilization of pyramid strategies, including image pyramids and feature pyramids. The image pyramid strategy detects instances by scaling images. Unlike the image pyramid method, the feature pyramid method integrates pyramid representations of different scales and semantic information layers. The Feature Pyramid Network (FPN) was proposed to aggregate high-level semantic features and low-level features through a top-down pathway, providing more accurate localization, but such networks tend to lose underlying positional information. Subsequently, to enhance the capability of hierarchical feature representation, new works emerged on bidirectional FPNs, such as Wang et al. [17] Wang added an additional bottom-up pathway on top of FPN to enhance the feature hierarchy at the top of the feature pyramid network, shorten the information path between low-level and top-level features, and help propagate accurate signals from low-level features. However, Zhang's bidirectional fusion is relatively simple and has some disadvantages. Zhang et al. [18] introduced a novel feature fusion method by repetitively applying weighted bidirectional feature pyramid networks, using bidirectional pathways for multiple extractions of the same layer's features to achieve higher-level integration and introducing learnable weights for different input features, simplifying Zhang to achieve better performance and efficiency. Zhu [19] was proposed to preserve high-quality features for accurate localization through a parallel FPN structure with bidirectional fusion and related improvements. Moreover, Luo et al. [20] utilizes neural architecture search to explore the topology of feature pyramid networks. Different from the aforementioned Necks, Slim-Neck, as a new lightweight design paradigm, introduces GSConv to reduce model complexity while maintaining accuracy, using a one-time aggregation method to design cross-level part networks (GSCSP) modules VoV-GSCSP, which lowers the computation and structural complexity while maintaining sufficient accuracy. Learning scale features to identify targets is key to localizing objects. To exchange multi-scale information effectively and fully, a new structure that enhances Zhang, the Bifusion Neck, is used. It integrates features from three adjacent layers output by the backbone network using a Bidirectional Cascading (BiC) module. This process can preserve more precise localization information, which is highly beneficial to the detection process as more accurate localization information leads to improved detection. The node diagrams for FPN, Zhang, Wang, and Bifusion Neck are illustrated in Figure 3.

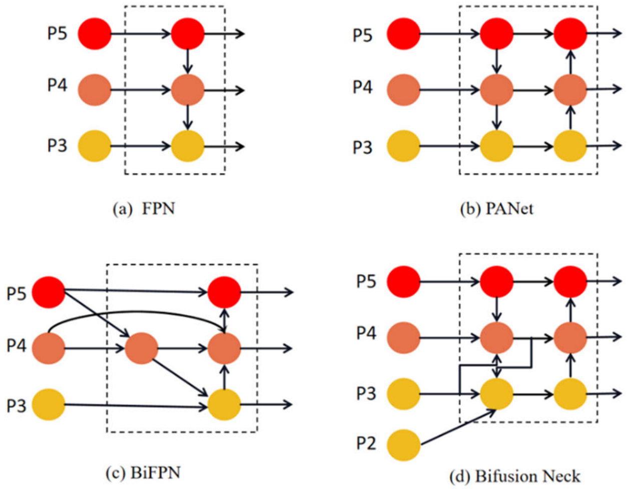


**FIGURE 3.** Node diagram.

The architecture of FPN+PAN introduces potential drawbacks as it involves multiple upsampling and downsampling operations on the feature maps, which may result in the loss of original feature information. Consequently, the detection success rate of the model can be compromised. Additionally, lower-level feature maps, characterized by higher resolution and richer detailed information, prove advantageous for detecting small targets. Nonetheless, the fusion of these lower-level feature maps in FPN+PAN contributes to a significant increase in parameters, leading to model redundancy and a decrease in detection speed. To mitigate these issues, it is crucial to incorporate even lower-level features without incurring additional parameter costs.

The structure of four types of Necks presented in the form of a node diagram makes it easier to discern the structure of each Neck. In the diagram, P3, P4, and P5 represent three different feature layers of the backbone network, with nodes indicating the feature maps output from each feature layer. The Bifusion Neck, compared to the other three Necks, includes an additional output from the P2 feature layer. As a lower-layer feature layer, P2 retains more shallow feature information. Taking the second output node of P3 as an example, it merges information from the P2, P3, and P4 feature layers, achieving a thorough fusion of low-level features with high-level features, which enhances the accuracy of localization information.

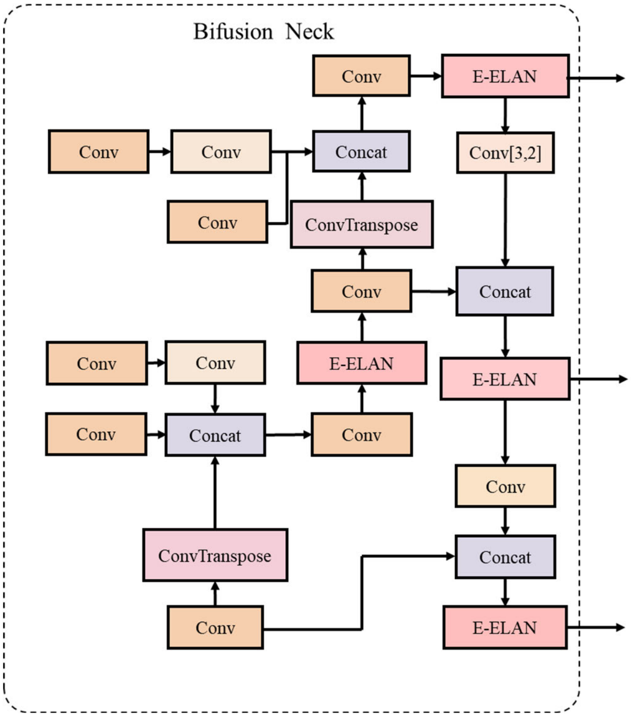


FIGURE 4. The Bifusion neck structure diagram.

The structure of the Bifusion Neck is shown in Figure 4. Its fusion principle involves using 1 × 1 convolutions to reduce the dimensionality of feature maps of the same scale; for larger scale feature maps, first applying 1 × 1 convolution to reduce dimensionality, followed by 3 × 3 convolution with a stride of 2 for downsampling; and for smaller scale feature maps, using 2 × 2 transposed convolutions for upsampling. Then, the feature maps obtained from these three parts are concatenated and further reduced in dimensionality through 1 × 1 convolution. The result after concatenation is sent into a dashed-line box, whose structure is the same as that of Zhang, which first goes through an upsampling pathway and then a downsampling pathway, delivering the three output feature maps to the detection head. The node diagram form of presenting the structure of four types of Necks facilitates an easier understanding of each Neck's structure. In the diagram, P3, P4, and P5 represent three different feature layers of the backbone network, with nodes indicating the feature maps output from each feature layer. The Bifusion Neck, in comparison to the other three Necks, includes an additional output from the P2 feature layer, which, being a lower-layer feature layer, retains more shallow feature information. For instance, the second output node of P3 merges information from the P2, P3, and P4 feature layers, enabling a comprehensive fusion of low-level features with high-level features for more precise localization information. The "Bifusion" fusion method is specifically designed to address potential issues with single-scale feature fusion, such as insufficient cross-layer feature interaction and inadequate multi-scale feature fusion. This improvement likely employs bi-directional feature fusion technology, not only fusing high-level abstract features and low-level detail features from the bottom up but also possibly integrating high-level global information with low-level local information from the top down, to construct a more rich and comprehensive feature representation.

### 4) LARGE SEPARABLE KERNEL ATTENTION

Attention mechanisms have emerged as one of the ways to enhance neural representations, particularly with the continuous development of deep learning. These mechanisms encompass various types, including channel attention mechanisms (e.g., SE), spatial attention mechanisms (e.g., GeNet, GcNet, and SGE), selection attention mechanisms, and hybrid attention mechanisms (e.g., CBAM and BAM). Selection attention technique, in particular, proves effective in improving the ability to focus on contextual regions. For instance, Condconv and Dynamic convolution employ parallel adaptive kernels to aggregate feature information from multiple convolution kernels. Meanwhile, SKNet introduces diverse convolution kernels and aggregates the information from these kernels along the channel dimension.

This study investigates the challenge of large-scale variations in fish species within the ocean and proposes a methodology called Large Separable Kernel Attention (LSKA). LSKA is an extension of LKA, and their structures are depicted in Figure 5. By breaking down the 2D convolution kernel into cascaded horizontal and vertical 1-D kernels, LSKA overcomes the issue of quadratic growth observed in deep convolution layers of traditional LKA modules. In contrast to LKA, LSKA's decomposition enables the direct utilization of deep convolution layers with larger kernels within the attention module, eliminating the need for additional modules. Decomposing the large

2D convolution kernel into smaller 1D convolution kernels considerably reduces the number of parameters and computational requirements, thereby enhancing the model's efficiency in processing large-sized inputs. The design of LSKA emphasizes shape recognition over texture in visual tasks, a distinction particularly evident with increasing kernel size, further enhancing the model's resilience to shape variations.

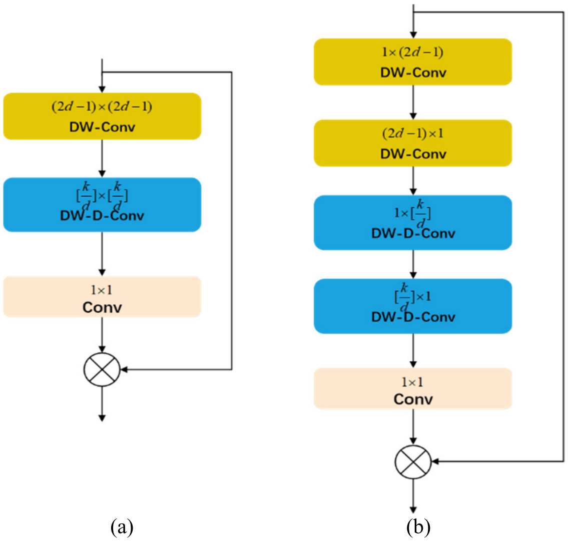


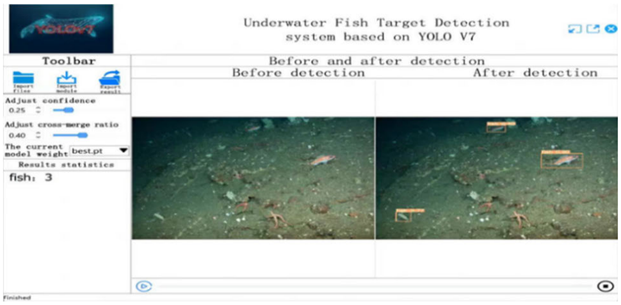
**FIGURE 5.** Structures of LKA and LSKA. (a): Diagram of LKA structure; (b): Diagram of LSKA structure.

## III. DESIGN OF THE EXPERIMENTS AND ANALYSIS OF THE RESULTS

### A. DESIGN OF THE TEST INTERFACE

First, determine the main functions and layout of the interface, including toolbars, before and after detection images, and results statistics.

Based on the requirements analysis, select the appropriate development tools and framework, such as PySide6. Using the selected development tools and framework, create a new project and define the main window with its size and title. Add a horizontal layout to the main window to place the toolbar and detect the two vertical layouts of the front and rear screens. In the vertical layout of the toolbar, add buttons or sliders such as import files, import modules, export results, adjust confidence, adjust intersection and ratio, and current model weights. In the vertical layout of the screen before and after detection, two image views are added to display the images before and after detection respectively. In the results statistics area, add a text editor or other control to display the statistics of the test results. Finally, the corresponding event processing function is added to each button or slider to implement their functionality. The detection interface map is shown in Figure 6.


Underwater Fish Target Detection
system based on YOLO V7

Adjust confidence 0.25

Adjust cross-merge ratio 0.40

The current model weight best.pt

Results statistics

fish: 3

Finished

**FIGURE 6.** Test interface diagram.

### B. EXPERIMENTAL PROCESS FLOW

#### 1) INTRODUCTION AND PROCESSING OF DATA SETS

To curate the necessary dataset for our model, we employed internet search techniques and web crawlers to extensively gather a diverse collection of images encompassing both deep-sea and shallow-water fish species. Our primary objective was to amass samples captured across a range of water depths and lighting conditions, including natural illumination, bioluminescence emitted by deep-sea organisms, and artificial light sources. By replicating complex and varied marine ecological environments, we aimed to enhance the generalization capabilities of our constructed model. Ultimately, a total of 5780 images were meticulously collected, serving as a comprehensive resource for model training and validation purposes.

By using the open source image annotation tool LabelImg, we manually annotate each fish picture collected in detail, including multi-dimensional information such as fish species, position and posture, thus ensuring the high accuracy and completeness of the data set. The sample plots in the dataset (Figure 7) clearly show carefully labeled images of various types of fish, reflecting our rigorous and meticulous work during the data preparation phase.

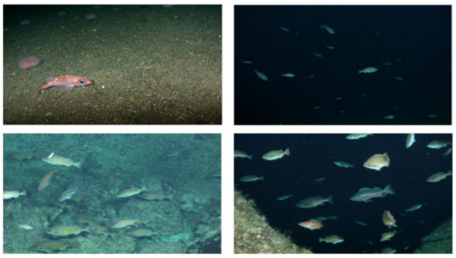


**FIGURE 7.** Data set sample display Fig.

To further optimize the model training process and evaluate its performance, a 7:2:1 ratio is carefully divided for training, validation, and test set. The table of data quantity allocation for each part is shown in Table 1 which helps to effectively monitor the fitting degree during model training and obtain reliable and representative model performance evaluation results in the final stage.

**TABLE 1. Allocation of data quantity for each part.**

| training set | validation set | test set |
|---|---|---|
| 4040 | 1150 | 590 |

### 2) CONFIGURATION AND TRAINING OF EXPERIMENTA-L PARAMETERS

All experiments in this study were conducted within the Pycharm environment and under the Pytorch 1.11.0 framework, with the compiler version being Python 3.8 and the operating system being Windows 10. The hardware configuration for the experiments included an AMD Ryzen 5 5600 6-Core Processor and an NVIDIA GeForce RTX 2080Ti GPU, the latter being used for computational acceleration. During the experimental process, we ensured that all involved training hyperparameters remained unchanged. Specifically, the input size of the images was set to 640 × 640, and the initial learning rate was set to 0.001. We used Adam as the optimizer, and the entire training process included 200 epochs. The batch size was set to 32. In addition, to further enhance the data, we also employed the mosaic data augmentation technique. The specific experimental environment and hyperparameters are shown in Table 2.

**TABLE 2. Experimental environment and hyperparameters.**

| | | |
|---|---|---|
| Hardware environment | CPU | AMD Ryzen 5 |
| | GPU | NVIDIA GeForce RTX 2080Ti |
| | RAM | 32G |
| Software environment | programming language | Python3.8 |
| | Deep learning frameworks | Pytorch1.11.0 |
| | CUDA | 11.3 |
| Experimental hyperparameters | Image_size | 640x640 |
| | Lr | 0.001 |
| | epoch | 200 |
| | optimizer | Adam |
| | Batch_size | 32 |

## C. ABLATION EXPERIMENTS

This study is based on the YOLOv7 algorithm and incorporates several improvement strategies. These strategies include using MPDIoU as the loss function and adopting the BIfusion neck as a novel neck structure. By employing the Bifusion mechanism, the study effectively integrates multi-level and multi-dimensional feature information, significantly enhancing the model's ability to capture subtle characteristics of underwater targets. Furthermore, an improved multi-point IoU (mpDIoU) bounding box loss function is introduced to enable a more precise evaluation of the overlap between predicted and ground truth bounding boxes. This approach proves to be particularly suitable for handling underwater targets of different sizes and shapes, resulting in a significant improvement in the accuracy and stability of the model for underwater target localization. To validate the specific impact of each improvement strategy on the algorithm, the study conducted training while keeping the parameters unchanged. Ablation experiments were conducted using YOLOv7 as the baseline model, and the specific experimental results are presented in Table 3.

**TABLE 3. Allocation of data quantity for each part.**

| Bifusion neck | MPDiou | LSKA | Para | Gflops | mAP 50(%) | F1-sore | FPS |
|---|---|---|---|---|---|---|---|
| | | | 37.19 | 105.1 | 73.3 | 0.72 | 112 |
| | √ | | 37.19 | 105.1 | 82.3 | 0.76 | 120 |
| √ | | | 37.10 | 105.3 | 82.8 | 0.77 | 118 |
| | | √ | 37.46 | 105.3 | 82.1 | 0.77 | 105 |
| √ | √ | | 37.10 | 105.3 | 83.7 | 0.78 | 124 |
| | √ | √ | 37.46 | 105.3 | 82.7 | 0.77 | 116 |
| √ | | √ | 37.37 | 105.5 | 83.0 | 0.77 | 115 |
| √ | √ | √ | 37.37 | 105.5 | 84.8 | 0.79 | 117 |

The symbol "√" denotes the adoption of specific strategies. According to the results presented in Table 3, when MPDiou is utilized, the parameters and computational complexity of YOLOv7 remain unchanged. However, there is a significant improvement in the average detection accuracy, with mAP50 reaching 82.3%. This represents a 9% enhancement compared to the baseline YOLOv7 model, and the detection speed has been improved, with FPS up to 120. When employing the BIfusion neck as the neck structure, the model experiences a slight reduction in parameter size to 37.1M, while the computational complexity slightly increases to 105.3Gflops. Concurrently, the model achieves an average detection accuracy of 82.8%, demonstrating a 9.5% improvement compared to the baseline YOLOv7 model. When the LSKA attention mechanism is added to the model, the number of parameters and computation are slightly improved, but the detection accuracy and F1-sore are significantly improved, reaching 82.1% and 0.77, respectively.When using MPDiou and LSKA, the model's parameters and computational cost slightly increase to 37.46M and 105.3 GFLOPs, respectively. Despite this, detection accuracy significantly improves, with mAP50 reaching 82.7%. When BIfusionneck and LSKA are used together, the model's parameter count is 37.37M, and the average detection accuracy (mAP50) rises to 83.0%.The highest average detection accuracy, with a mAP50 of 84.8%, was achieved when BIfusion neck, MPDIou and LSKA attention were used simultaneously. In addition, the F1 score reached 0.77, an improvement of 11.5% and 0.07 compared to the YOLOv7 model, respectively. And the detection speed is improved compared to the original model, with an FPS of 117, an increase of 5. compared to YOLOv7.In conclusion, the application of any of these improvement strategies enhances the model's detection accuracy and F1-score, thereby establishing the efficacy of these strategies in improving the model's detection accuracy and stability.

### D. COMPARISON OF DIFFERENT LOSS FUNCTIONS

In order to validate the superior performance of the MPDIou employed in this study, comparative experiments were conducted, contrasting it with the mainstream loss functions, namely DIOU, GIOU, and CIOU. The specific experimental results are presented in Table 4. All experiments were executed using identical hyperparameters on the same dataset.

**TABLE 4. Comparison of different loss functions.**

| LOSS | mAP50(%) | mAP50-95(%) |
|---|---|---|
| CIOU | 73.3 | 40.4 |
| GIOU | 72.4 | 40.2 |
| DIOU | 71.2 | 38.2 |
| MPDIOU | 82.3 | 50.1 |

As shown in Table 4, when employing MPDIOU, the algorithm attains the highest mAP50 and mAP50-95, reaching 82.3% and 50.1% respectively. Conversely, utilizing DIOU as the loss function results in the model exhibiting the lowest average detection accuracy of 71.2%, which is lower than when CIOU is employed. To visually illustrate the superior performance of MPDIOU, the losses and mAP for different IOUs have been visualized and are depicted in Figure 8. As shown in Figure 8, using MPDIoU results in a faster decrease in model loss, leading to an accelerated convergence rate. Additionally, the model's detection accuracy improves more rapidly compared to other loss functions, and it also demonstrates superior detection accuracy overall.

As depicted in Figure 8, the loss curve of MPDIou consistently remains lower than that of other IOUs, suggesting a faster convergence rate. Additionally, the average detection accuracy curve for MPDIou surpasses all other IOUs while demonstrating reduced fluctuations, indicating enhanced stability. In conclusion, MPDIou exhibits superior performance compared to other loss functions, outperforming them in this particular scenario.

### E. COMPARISON OF DIFFERENT ALGORITHMS

To validate the superior performance of the proposed model in this study, comparative experiments were conducted using the same dataset to compare the BFD-YOLO model with various mainstream models. These models include two-stage models like Faster-RCNN, one-stage models like SSD and Centernet, as well as YOLOv3, YOLOv5, YOLOv6, and YOLOv8 models. The experimental results are displayed in Table 5.

**TABLE 5. Comparison of different loss functions.**

| Model | P(%) | R(%) | mAP50(%) | mAP50-95(%) | F1-score |
|---|---|---|---|---|---|
| Faster-Rcnn[21] | 45.6 | 81.5 | 74.8 | 33.0 | 0.59 |
| SSD[22] | 89.3 | 54.9 | 73.5 | 35.1 | 0.68 |
| Centernet[23] | 97.6 | 34.4 | 73.1 | 42.0 | 0.51 |
| YOLOv7[24] | 79.6 | 65.4 | 73.3 | 41.8 | 0.72 |
| YOLOv6[25] | 75.9 | 62.4 | 74.3 | 42.3 | 0.69 |
| YOLOv3[26] | 79.7 | 63.8 | 74.2 | 41.1 | 0.71 |
| YOLOv5 | 74.6 | 69.1 | 76.3 | 43.6 | 0.72 |
| YOLOv8 | 74.0 | 75.9 | 76.4 | 44.5 | 0.75 |
| BFD-YOLO | 80.8 | 76.5 | 84.8 | 51.8 | 0.79 |

Based on the findings presented in Table 5, the proposed BFD-YOLO algorithm in this study attains the highest mAP50 value (84.8%) and F1-score (0.79) among other mainstream algorithms on the dataset. Faster-Rcnn, as a two-stage detection algorithm, exhibits superior detection accuracy, surpassing all tested models except for BFD-YOLO, YOLOv5, and YOLOv8, while achieving the highest recall value (81.5%). Within the one-stage detection algorithms, SSD and Centernet demonstrate remarkable precision, with Centernet achieving the highest precision (97.6%). However, Centernet exhibits a lower recall rate, which is the lowest among all tested models at 34.4%. Moreover, SSD, Centernet, and Faster-Rcnn algorithms display reduced stability with lower F1-scores compared to the YOLO series algorithms. The YOLO series algorithms perform admirably in this scenario, particularly YOLOv8, which demonstrates accuracy and F1-score slightly below that of the proposed BFD-YOLO algorithm in this study. While the precision of BFD-YOLO is slightly lower than that of SSD and Centernet algorithms, and the recall is slightly lower than that of Faster-Rcnn, it maintains an overall precision of 80.8% and recall of 76.5%, achieving a better balance between the two and resulting in an overall more stable model. Furthermore, compared to Faster-Rcnn, SSD, Centernet, YOLOv7, YOLOv6, YOLOv3, YOLOv5, and YOLOv8 algorithms, the proposed BFD-YOLO algorithm achieves higher accuracy, with improvements of 10.0%, 11.3%, 11.7%, 11.5%, 10.5%, 10.6%, 8.5%, and 8.4% respectively. In conclusion, the proposed BFD-YOLO algorithm outperforms the other tested models in this scenario.

#### 1) ALGORITHMSCOMPARISON OF DIFFERENT ALGORITHMS ACROSS DIFFERENT DATASETS

The algorithm proposed in this study is validated to exhibit strong generalization ability and versatility through comparative experiments with different algorithms on datasets from diverse domains. For evaluation purposes, the PASCAL VOC 2007 dataset is chosen in this section. This dataset is widely recognized as a standard in computer vision, comprising 9,963 images and encompassing twenty categories, including person, bus, and train. The dataset is partitioned into training and validation sets in an 8:2 ratio. Detailed experimental results are provided in Table 6.

According to Table 6, the proposed BFD-YOLO algorithm demonstrates the highest detection accuracy, achieving an mAP50 of 76.3%. It surpasses other tested models, such as YOLOv5s, YOLOv7, YOLOv8n, YOLOv8s, YOLOX, and MOD-YOLOs, by margins of 15.7%, 12.7%, 28.3%, 14.5%, 8.9%, and 5.5%, respectively. Moreover, the BFD-YOLO

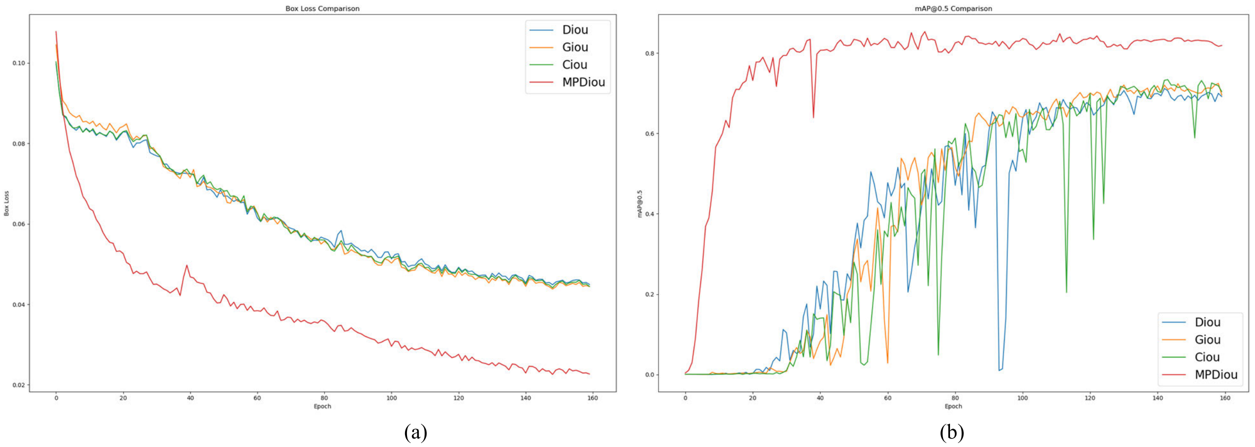


**FIGURE 8.** Loss curves and mAP50 curves for different IOUs. (a): Loss curves; (b): mAP50 curves.

**TABLE 6.** Comparison of different algorithms for VOC dataset.

| Model | YOLOv5s | YOLOv7 | YOLOv8n | YOLOv8s | YOLOX[27] | MOD-YOLOs[28] | BFD-YOLO |
|---|---|---|---|---|---|---|---|
| | AP50(%) | AP50(%) | AP50(%) | AP50(%) | AP50(%) | AP50(%) | AP50(%) |
| aeroplane | 65.2 | 70.2 | 51.6 | 67.9 | 71.8 | 76.9 | 88.1 |
| bicycle | 75.5 | 79.4 | 59.9 | 75.9 | 82.0 | 83.3 | 80.9 |
| bird | 53.3 | 52.0 | 31.1 | 50.6 | 59.4 | 66.6 | 72.9 |
| boat | 49.4 | 49.9 | 34.1 | 50.1 | 57.9 | 62.9 | 67.4 |
| bottle | 44.2 | 40.3 | 27.2 | 36.0 | 47.0 | 45.5 | 67.6 |
| bus | 64.5 | 70.8 | 65.6 | 73.2 | 78.3 | 81.5 | 72.8 |
| car | 85.5 | 84.8 | 76.7 | 85.4 | 85.9 | 87.6 | 86.7 |
| cat | 55.0 | 65.1 | 43.1 | 56.1 | 65.6 | 73.5 | 78.5 |
| Chair | 49.3 | 48.4 | 37.1 | 46.5 | 53.0 | 53.8 | 64.9 |
| Cow | 59.1 | 66.1 | 34.6 | 57.6 | 55.5 | 70.8 | 77.5 |
| diningtable | 44.6 | 48.2 | 48.3 | 59.7 | 62.7 | 67.1 | 71.3 |
| dog | 51.4 | 61.9 | 34.6 | 57.6 | 62.5 | 71.9 | 76.1 |
| horse | 69.6 | 71.5 | 63.7 | 74.0 | 79.2 | 79.1 | 87.9 |
| motorbike | 74.5 | 76.8 | 61.2 | 74.4 | 82.7 | 79.7 | 88.5 |
| person | 79.7 | 79.9 | 76.3 | 82.8 | 84.0 | 83.8 | 86.9 |
| pottedplant | 36.1 | 41.7 | 23.1 | 34.2 | 41.8 | 49.1 | 58.0 |
| sheep | 59.5 | 63.2 | 33.4 | 55.6 | 60.2 | 63.0 | 68.3 |
| sofa | 56.3 | 60.0 | 53.6 | 62.8 | 67.6 | 69.3 | 69.0 |
| train | 74.1 | 77.6 | 64.3 | 76.6 | 81.0 | 81.8 | 85.8 |
| tvmonitor | 64.1 | 62.1 | 50.6 | 62.5 | 69.1 | 68.4 | 74.3 |
| Average | 60.6 | 63.6 | 48.5 | 61.8 | 67.4 | 70.8 | 76.3 |

algorithm exhibits the highest precision in 16 out of the 20 categories. Overall, the performance of the BFD-YOLO algorithm on the VOC2007 dataset outperforms that of the other tested models.

To conduct a comparative experiment using the publicly available OZFish dataset, which is part of Australia's Research Data Sharing Data Discovery Program, aimed at advancing machine learning research for automatic fish detection from videos. Approximately 80,000 labeled fish samples were extracted from videos, covering over 500 species, 200 genera, and 70 families. A unique feature of the OzFish dataset is that each image contains numerous instances of various fish species and shapes. On average, each captured frame in OzFish contains 25 fish objects, with many frames containing up to 80-120 fish objects. The dataset is divided into 80% training and 20% testing sets, randomly stratified from each habitat. Figure 9 shows some labeled samples from the OZFish dataset.

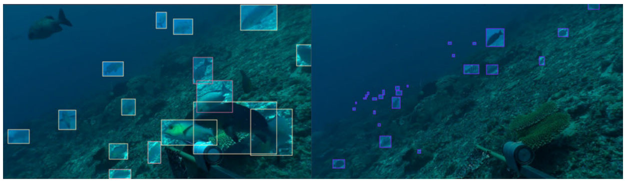
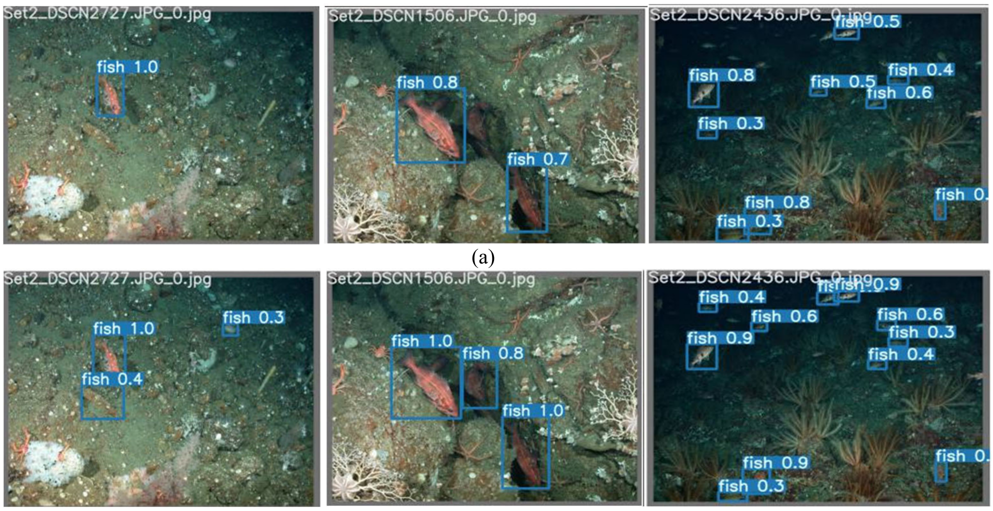


**FIGURE 9.** Some labeled samples in the OZfish dataset.

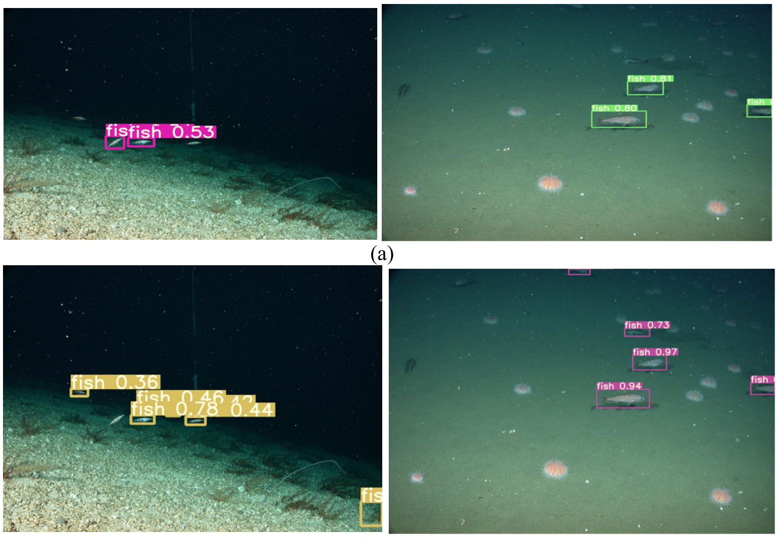
To further verify the superior performance of the BFD-YOLO algorithm, we compared its performance with different models on the OZFish dataset, as shown in Table 7. According to Table 7, the proposed BFD-YOLO algorithm achieves the highest detection accuracy with an mAP50 of 76.3%. Compared to other object detection models,



(a)

(b)

**FIGURE 10.** Experimental results diagram.(a): YOLOv7 detection (b):BFD-YOLO detection.



(a)

(b)

**FIGURE 11.** Comparison diagram of the effect before and after improvement under low illumination. (a): YOLOv7 detection (b):BFD-YOLO detection.

YOLOv3, YOLO-fish, TOOD, ATSS, and YOLOV7, the mAP50 is improved by 7.2%, 6%, 4.5%, 5.1%, and 3.2%, respectively. Additionally, BFD-YOLO has the highest precision and recall, at 84.3% and 74.5%, respectively. In summary, the BFD-YOLO algorithm outperforms other tested algorithms on the OZFish dataset.

### 2) DETECTION VISUALIZATION ANALYSIS

To assess the efficacy of the introduced enhancements, a set of random images were deliberately chosen for testing purposes. The resulting experimental outcomes are presented in Figure 10. The original YOLOv7 model exhibited a substantial number of missed detections and false positives in relation to small-scale and low-illumination targets. In sharp contrast, the refined model, as proposed in this paper, demonstrated commendable performance in both scenarios, displaying an almost negligible occurrence of missed detections. Moreover, the improved model exhibited enhanced precision in target identification, as demonstrated by bounding boxes that closely conformed to the actual objects, along with higher confidence scores in comparison to the original YOLOv7 model. In summary, the model proposed in this paper showcases exceptional performance in the specific domain of fish detection.

Further analysis reveals that when employing models before and after training for inference testing, the differ-

**TABLE 7.** Comparison of different algorithms on OZfish data set.

| Model | P(%) | R(%) | mAP50 (%) | mAP50-95(%) | F1-sore |
|---|---|---|---|---|---|
| YOLOv3 | 80.1 | 65.3 | 69.1 | 34.7 | 0.72 |
| TOOD[29] | 81.5 | 67.3 | 71.8 | 36.2 | 0.73 |
| YOLO-fish[30] | 82.3 | 64.4 | 70.3 | 35.2 | 0.73 |
| ATSS[31] | 78.7 | 70.6 | 71.2 | 35.9 | 0.73 |
| YOLOv7 | 80.8 | 71.2 | 73.1 | 37.1 | 0.73 |
| BFD-YOLO | 84.3 | 74.5 | 76.3 | 38.2 | 0.74 |

ence in target detection capabilities across models can be visually observed through images. Particularly in low-light environments, the improved detection model demonstrates a significant enhancement in performance, especially for relatively small targets. Specifically, the detection confidence for the same target has also been significantly increased. These experimental results fully demonstrate that the improved YOLOv7 algorithm exhibits outstanding performance in the detection of underwater fish species, especially in terms of low-light conditions and the detection of small targets, where its capability has been effectively enhanced. Moreover, the model's generalization ability has also been further strengthened. For a comparison of the model's performance before and after improvements under low-light conditions, refer to Figure 11.

## IV. CONCLUSION

In this study, through in-depth exploration of underwater fish detection projects, we implemented innovative improvements to the YOLOv7 model by introducing the Bifusion mechanism and the MPDIoU bounding box loss function, significantly enhancing the model's performance. The integration of the Bifusion mechanism has improved the model's ability to capture minute details and the overall form of fish, greatly enhancing the precision in differentiating between various species of fish in complex underwater environments. Simultaneously, the application of the MPDIoU loss function optimized the precise localization of fish by considering the overlap area of bounding boxes, the distance between centers, and the aspect ratio, thereby improving the precision of predicted bounding boxes in matching the actual ones. These improvements have elevated the model's mAP50 value from 73.3% to 84.8%, fully validating the effectiveness and feasibility of these measures. Additionally, experimental validation on diverse datasets demonstrated that the BFD-YOLO algorithm outperforms other tested algorithms on the VOC2007 dataset, highlighting its exceptional generalization capability.

Despite significant progress, there remains much to be explored and improved in the field of underwater fish detection. Future research should focus on optimizing the model structure, multimodal fusion, data augmentation, and adaptive learning, as well as improving detection performance for rare fish species or under extreme conditions. This work will not only advance underwater fish detection technology but also introduce new ideas and methods for related fields of study.

## REFERENCES

[1] C.-H. Yeh, C.-H. Lin, L.-W. Kang, C.-H. Huang, M.-H. Lin, C.-Y. Chang, and C.-C. Wang, "Lightweight deep neural network for joint learning of underwater object detection and color conversion," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 11, pp. 6129–6143, Nov. 2022.

[2] P. Liu, W. Qian, and Y. Wang, "YWnet: A convolutional block attention-based fusion deep learning method for complex underwater small target detection," *Ecol. Informat.*, vol. 79, Mar. 2024, Art. no. 102401.

[3] Y. Shen, C. Zhao, Y. Liu, S. Wang, and F. Huang, "Underwater optical imaging: Key technologies and applications review," *IEEE Access*, vol. 9, pp. 85500–85514, 2021.

[4] M. Mathur and N. Goel, "Enhancement of nonuniformly illuminated underwater images," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 35, no. 3, Mar. 2021, Art. no. 2154008.

[5] Z. Zhu, X. Li, J. Zhai, and H. Hu, "PODB: A learning-based polarimetric object detection benchmark for road scenes in adverse weather conditions," *Inf. Fusion*, vol. 108, Aug. 2024, Art. no. 102385.

[6] S. P. González-Sabbagh and A. Robles-Kelly, "A survey on underwater computer vision," *ACM Comput. Surveys*, vol. 55, no. 13s, pp. 1–39, Dec. 2023.

[7] A. Mittal, S. Dhalla, S. Gupta, and A. Gupta, "Automated analysis of blood smear images for leukemia detection: A comprehensive review," *ACM Comput. Surveys*, vol. 54, no. 11, pp. 1–37, Jan. 2022.

[8] A. Calantropio, F. Chiabrando, and R. Auriemma, "Photogrammetric underwater and uas surveys of archaeological sites: The case study of the Roman shipwreck of Torre Santa sabina," *Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 1, pp. 643–650, Jun. 2021.

[9] C.-J. Huang, H.-W. Cheng, Y.-H. Lien, and M.-E. Jian, "A survey on video streaming for next-generation vehicular networks," *Electronics*, vol. 13, no. 3, p. 649, Feb. 2024.

[10] S. Xu, M. Zhang, W. Song, H. Mei, Q. He, and A. Liotta, "A systematic review and analysis of deep learning-based underwater object detection," *Neurocomputing*, vol. 527, pp. 204–232, Mar. 2023.

[11] F. Han, J. Yao, H. Zhu, and C. Wang, "Underwater image processing and object detection based on deep CNN method," *J. Sensors*, vol. 2020, pp. 1–20, May 2020.

[12] W.-H. Lin, J.-X. Zhong, S. Liu, T. Li, and G. Li, "ROIMIX: Proposal-fusion among multiple images for underwater object detection," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 2588–2592.

[13] C. Li, S. Anwar, and F. Porikli, "Underwater scene prior inspired deep underwater image and video enhancement," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107038.

[14] F. Lei, F. Tang, and S. Li, "Underwater target detection algorithm based on improved YOLOv5," *J. Mar. Sci. Eng.*, vol. 10, no. 3, p. 310, Feb. 2022.

[15] M. J. Islam, Y. Xia, and J. Sattar, "Fast underwater image enhancement for improved visual perception," *IEEE Robot. Autom. Lett.*, vol. 5, no. 2, pp. 3227–3234, Apr. 2020.

[16] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[17] Y. Wang, Y. Han, C. Wang, S. Song, Q. Tian, and G. Huang, "Computation-efficient deep learning for computer vision: A survey," *Cybern. Intell.*, vol. 1, pp. 1–24, Jul. 2024.

[18] H. Zhang, Q. Du, Q. Qi, J. Zhang, F. Wang, and M. Gao, "A recursive attention-enhanced bidirectional feature pyramid network for small object detection," *Multimedia Tools Appl.*, vol. 82, no. 9, pp. 13999–14018, Apr. 2023.

[19] M. Zhu, "Dynamic feature pyramid networks for object detection," in *Proc. 15th Int. Conf. Signal Process. Syst.*, 2023, pp. 1–26.

[20] J. Luo, Y. Li, W. Zhou, Z. Gong, Z. Zhang, and W. Yao, "An improved data-driven topology optimization method using feature pyramid networks with physical constraints," *Comput. Model. Eng. Sci.*, vol. 128, no. 3, pp. 823–848, 2021.

[21] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[22] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "SSD: Single shot MultiBox detector," in *Computer Vision—ECCV*. Cham, Switzerland: Springer, 2016, pp. 21–37.

[23] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "CenterNet: Keypoint triplets for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 6568–6577.

[24] C.-Y. Wang, A. Bochkovskiy, and H.-Y.-M. Liao, "YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7464–7475.

[25] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, Y. Li, B. Zhang, Y. Liang, L. Zhou, X. Xu, X. Chu, X. Wei, and X. Wei, "YOLOv6: A single-stage object detection frame-work for industrial applications," 2022, *arXiv:2209.02976*.

[26] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.

[27] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "YOLOx: Exceeding YOLO series in 2021," 2021, *arXiv:2107.08430*.

[28] P. Su, H. Han, M. Liu, T. Yang, and S. Liu, "MOD-YOLO: Rethinking the YOLO architecture at the level of feature information and applying it to crack detection," *Expert Syst. Appl.*, vol. 237, Mar. 2024, Art. no. 121346.

[29] C. Feng, Y. Zhong, Y. Gao, M. R. Scott, and W. Huang, "TOOD: Task-aligned one-stage object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 3490–3499.

[30] A. A. Muksit, F. Hasan, M. F. Hasan Bhuiyan Emon, M. R. Haque, A. R. Anwary, and S. Shatabda, "YOLO-Fish: A robust fish detection model to detect fish in realistic underwater environment," *Ecol. Informat.*, vol. 72, Dec. 2022, Art. no. 101847.

[31] S. Zhang, C. Chi, Y. Yao, Z. Lei, and S. Z. Li, "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9756–9765.

**JINYU OU** was born in Haikou, Hainan, China, in 2002. She received the degree in transportation from Taiyuan University of Science and Technology, Taiyuan, Shanxi, China, and the master's degree from the School of Mechanical and Electrical Engineering, Hainan University, Haikou, in 2024.

Her research interests include intelligent manufacturing technology, underwater object detection based on deep learning, robot control, and artificial intelligence. She has won national scholarships and university scholarships for many times.

**YIJUN SHEN** was born in Wuhan, Hubei, China. He received the Ph.D. degree from the University of Southampton, in 2003.

He was a Postdoctoral and an Associate Researcher at Cranfield University, a Researcher at Brunell University, a Senior Consulting Expert of the world famous marine energy design consulting company, wood, and a deep-sea energy technology expert. He is currently the Deputy Director of the State Key Laboratory of South China Sea Marine Resources Utilization, Hainan University. Mainly engaged in the development and utilization of Marine resources such as deep-sea oil and gas and renewable energy. He is a member of the British Academy of Marine Engineering and Science and Technology and Technical Committee of the International Society of Marine and Polar Engineering (1 SOPE). He is the Chief Editor of the *Marine Resource* and *Ocean Science* journal. He also serves as the Chairperson for underwater systems, risers, umbilical cables, and submarine pipelines (SURF).