


Article

Multi-Scale Feature Enhancement Method for Underwater Object Detection

Mengpan Li ^{1,†}, Wenhao Liu ^{1,†}, Changbin Shao ^{1,*} , Bin Qin ¹, Ali Tian ² and Hualong Yu ¹ 

¹ School of Computer Science, Jiangsu University of Science and Technology, Zhenjiang 212100, China; limengpan@stu.just.edu.cn (M.L.); liuwenhao@stu.just.edu.cn (W.L.); qinbin_sd@126.com (B.Q.); yuhualong@just.edu.cn (H.Y.)

² School of Naval Architecture and Ocean Engineering, Jiangsu University of Science and Technology, Zhenjiang 212100, China; tianali@just.edu.cn

* Correspondence: shaocb@just.edu.cn

[†] These authors contributed equally to this work.

Abstract: With deep-learning-based object detection methods reaching industrial-level performance, underwater object detection has emerged as a significant application. However, it is often challenged by dense small instances and image blurring due to the water medium. In this paper, a Multi-Scale Feature Enhancement (MSFE) method is presented to address the challenges triggered by water bodies. In brief, MSFE attempts to achieve dual multi-scale information integration through the internal structural design of the basic C2F module in the Backbone network and the external global design of the feature pyramid network (FPN). For the internal multi-scale implementation, a LABNK module is constructed to address the vanishing or weakening phenomenon of fine-grained features during feature extraction. Specifically, it adopts a symmetrical structure to collaboratively capture two types of local receptive field information. Furthermore, to enhance the information integration ability between inter-layer features in FPN, a shallow feature branch is injected to supplement detailed features for the subsequent integration of multi-scale features. This operation is mainly supported by the fact that large-sized features from the shallow layer usually carry rich, fine-grained information. Taking the typical YOLOv8n as the benchmark model, extensive experimental comparisons on public underwater datasets (DUO and RUOD) demonstrated the effectiveness of the presented MSFE method. For example, taking the rigorous mAP (50:95) as an evaluation metric, it can achieve an accuracy improvement of about 2.8%.

Keywords: underwater object detection; convolutional neural network; YOLO detection model; multi-scale feature enhancement; local awareness operation; feature pyramid network



Academic Editor: Zhixun Su

Received: 15 November 2024

Revised: 28 December 2024

Accepted: 30 December 2024

Published: 2 January 2025

Citation: Li, M.; Liu, W.; Shao, C.; Qin, B.; Tian, A.; Yu, H. Multi-Scale Feature Enhancement Method for Underwater Object Detection.

Symmetry **2025**, *17*, 63. <https://doi.org/10.3390/sym17010063>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

The oceans are rich in diverse resources and have become a significant raw material provider for human activities and livelihoods [1]. In the context of the widespread adoption of artificial intelligence across various fields, object detection technology provides a feasible automation approach for underwater resource exploration. As a subfield of computer vision, object detection technology mainly serves to locate and classify specific object instances in image or video data. In brief, underwater exploration equipment with detection functions can greatly reduce reliance on manual diving operations, offering advantages in detection efficiency and safety while conserving both human and material resources.

Traditional object detection methods have adopted the two-stage sequence mode of feature extraction and object classification. The manual feature descriptor methods [2–4] dominate in acquiring the inherent features of object instances. In this context, early underwater object detection (UOD) techniques directly followed the traditional manner. However, traditional descriptors often struggle to capture accurate feature information of the underwater targets [5–10]. The main reason for this is that underwater images appear blurry due to the absorption and scattering effects of water on light. In response, numerous attempts [11–15] focus on how to integrate or fuse various feature descriptors. In summary, traditional UOD methods have achieved some results, but they have difficulty meeting the practical industrialization due to the disadvantage of feature extraction.

With the breakthroughs in deep learning across various tasks, Deep Learning-based Object Detection (DLOD) methods have gradually emerged as the mainstream approach for practical applications. In contrast to traditional feature descriptor methods, the DLOD methods usually achieve better accuracy owing to the superior feature extraction ability of Convolutional Neural Networks (CNNs). DLOD methods can be divided into two categories, namely two-stage and single-stage. The two-stage methods [16–19] directly inherit the two-stage implementation of feature extraction and classification. However, they usually achieve better accuracy but struggle with complex operations and low detection efficiency due to the time-consuming region-proposal operation. In contrast, single-stage methods, such as the YOLO series [20–26] adopt end-to-end detection implementation, with simple operation and high detection efficiency. Therefore, single-stage detectors have been the mainstream method, especially in real-time detection scenarios.

Although DLOD methods have made significant progress in real-life scenarios, the transferable applications of existing methods often achieve poor performance in UOD applications. Recently, some studies have summarized the main challenges and solutions [27,28]. Furthermore, the study [29] focused on deep-learning techniques and how these methods can improve accuracy in underwater detection. Besides, the survey [30] particularly emphasizes data availability and how datasets impact the development of underwater detection methods. In general, the implementation of UOD often cannot avoid conventional challenges in real-life scenes, such as sample scarcity, class imbalance, and noisy labels issue [27]. Furthermore, the negative impact of the water body on light and the distribution pattern of target instances give rise to the following additional difficulties.

- Image-blurring issue: The absorption and scattering of light by the water typically lead to a rapid reduction in light intensity. Under normal optical mechanisms, insufficient light can cause image darkening and color deviation. Additionally, the fluctuation and turbidity of a water body can further worsen the distortion of image features. For instance, unclear water will result in warped or blurry shapes of the target.
- Dense small target issue: The large-scale aggregation of some aquatic organisms often triggers the dense detection issue of small targets. For instance, the aggregation of small-size fishes can make it difficult for detection models to distinguish individual targets, resulting in false positive and false negative phenomena.

To address the challenges mentioned above, the attempts can be classified into two categories. The first solution is image enhancement methods, which focus on how to obtain clear, high-resolution images [31–34] for subsequent detection implementation. The other mainline directly focuses on how to enhance the detection ability of existing detectors, involving the structural design of model and loss construction of training implementation [35–39]. To be specific, the structural design attempts to improve the feature extraction capability of the model. For example, the design of the Backbone network using deformable convolution [35], dilated convolution [36] and attention modules [37]; the enhancement of the Feature Pyramid Network (FPN) using attention modules [36,38]; and

the auxiliary detection head mechanism. In addition, the construction of the loss function can also slightly improve the detection accuracy [35,37].

Overall, previous methods have achieved significant success. However, in terms of model structure, they did not address the challenges of underwater detection from the perspective of multi-scale feature fusion. In this case, this paper presents a Multi-Scale Feature Enhancement (MSFE) method to boost the ability of the detector. In brief, MSFE focuses on how to apply multi-scale feature fusion techniques to implement feature enhancement for blurry images. In other words, the idea of MSFE always revolves around how to integrate multi-scale feature information during the feature extraction process. Taking the YOLO model as a benchmark framework, the MSFE method can be described as follows.

- For the image blurring issue. It conducts a multi-scale local awareness operation to enhance the information integration ability of the basic C2F module in the Backbone network. To be specific, anchoring the BNK submodule in C2F, the multi-scale design adopts symmetric local awareness operations to overcome the feature vanishing or weakening phenomenon during the feature extraction. The ‘Stage Layer 3’ and ‘Stage Layer 4’ in Figure 1 denote its locations.
- For the dense small target issue. Following the multi-scale information integration mechanism of FPN between deep and shallow features. It extra injects a shallow branch into the FPN to supply large-scale features for subsequent prediction. This idea is mainly supported by the fact that shallow features usually carry rich, detailed information. The ‘Top Down Layer3’ and ‘- -’ in Figure 1 show this design. Formally, this idea can be viewed as an enhancement of normal FPN structure.

In summary, the MSFE method actually achieves a double multi-scale information integration through the internal operation design of the basic C2F module in the Backbone network and the external global design of FPN. Taking the single-stage YOLO as the benchmark framework, the implementation of MSFE is carried out in the lightweight YOLOv8n. Extensive comparisons in underwater datasets demonstrate its effectiveness in improving accuracy.

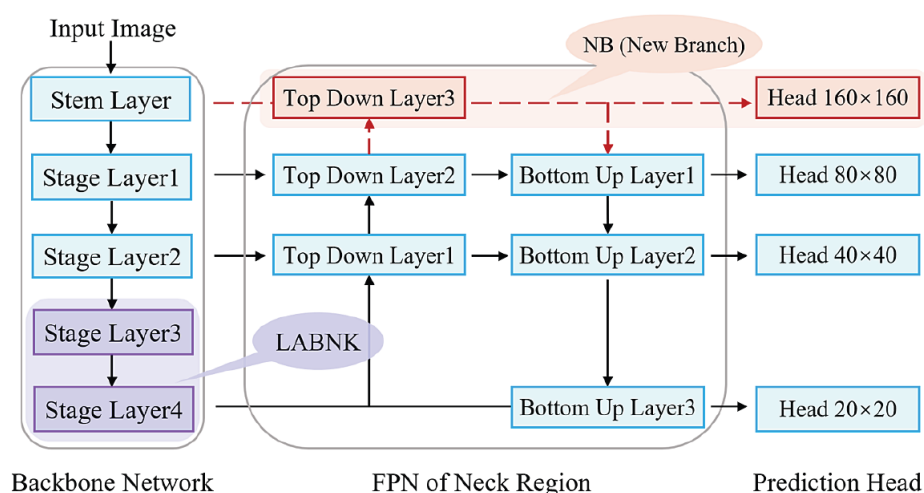


Figure 1. The diagram of the MSFE method based on YOLOv8 detector. The entire figure displays the architecture of YOLOv8. The blue regions represent the original components in the initial framework. To achieve feature enhancement for the detector, the MSFE achieves two structural designs, namely LABNK and NB. For the Backbone network, the LABNK conducts a local awareness operation for the internal BNK in the basic C2F. The purple region indicates its location. In contrast, the NB injects a new feature branch into the FPN in the Neck region. The red region highlights its location.

2. Related Work

2.1. Deep-Learning-Based Object Detection (DLOD) Methods

With the rise of deep-learning technology, DLOD methods have gradually become the mainstream detection approach in practice. In contrast to the hand-crafted feature descriptors used in traditional methods, the DLOD methods achieve superior performance using learnable deep features. In this context, deep detection methods have not only achieved significant success in detection accuracy but also demonstrated strong robustness in various noisy environments. According to the workflow, the DLOD methods are generally divided into two-stage and single-stage approaches.

Two-stage Method. These methods directly follow the separable implementation pattern of feature extraction and object classification in traditional detection approaches. The most competitive models are the R-CNN series [16–19,40], such as Fast R-CNN [17] and Faster R-CNN [18]. The evolution of two-stage object detectors begins with R-CNN [16], which employs proposal generation techniques to generate candidate bounding box regions from the input image. However, it requires feature extraction for each candidate region, resulting in high computational costs. To address this issue, Fast R-CNN turns to obtain candidate regions on the deep feature, allowing the features of the entire input image to be computed only once. Building upon Fast R-CNN, Faster R-CNN introduces the Region-Proposal Network (RPN) that enables nearly cost-free region proposals. Moreover, the RPN actually integrates region generation with the detection process. This integration enables end-to-end training, marking an important milestone in DLOD technology.

Single-stage Method. The YOLO series [20–22,41,42], have emerged as convenient implementations in object detection. These methods directly inherit the end-to-end nature of deep learning. In contrast to two-stage methods, YOLOv1 [20] achieves greater efficiency by replacing the region-proposal process with the grid division strategy on the input. In this grid division setting, the network divides the image into regions and predicts bounding boxes and probabilities for each region simultaneously. Although YOLO significantly improves detection speed, it suffers from a decline in localization accuracy compared to two-stage detectors, particularly for small objects. Subsequent versions of YOLO [23–26] and the later SSD [43] have focused more on addressing this issue.

2.2. Underwater Object Detection Algorithms

Early UOD technology directly followed traditional object detection methods, employing a two-stage sequence mode of feature extraction and object classification. Among them, manual feature descriptor methods like SIFT, HOG, and LBP were employed to acquire local features (e.g., the edge, texture, or keypoint information). However, these methods struggle in underwater scenes due to blurry images caused by light absorption and scattering. In response, the adaptive color matching method [11] attempts to overcome the difficulty of weakened light. Subsequent underwater target recognition algorithms utilize fused descriptors to integrate texture, edge, or color information of underwater targets [11–15]. Despite some success, traditional UOD methods depend heavily on manual features, hindering their adaptation to diverse underwater environment [36] and making them difficult to industrialize implementation.

DLOD algorithms provide significant advantages over traditional UOD methods, especially in their ability to automatically learn and extract features related to the object region in underwater images. In recent years, numerous studies [35–39] have focused on modifying existing model architectures and loss functions to improve the performance of object detection in challenging environments. For example, the Backbone network was designed using space-to-depth convolution [44], the new network architecture [45–47], and the attention module [44]; the FPN was enhanced through spatial convolution down-

sampling [44]. Moreover, model performance could be further improved by designs of loss function [44–48]. In addition to the above improvements, the use of the Transformer architecture has also significantly enhanced the detection performance [49,50].

In addition to the above innovations, researchers have also made significant advancements in image enhancement techniques. For instance, the co-training algorithm combining image enhancement and object detection [32] mitigated background interference from image degradation using an underwater image enhancement module before the detection network. The underwater image processing and object detection algorithm based on the deep CNN method [33] improved underwater detection by preprocessing images with Max-RGB and grayscale methods. The MSEMNCN algorithm [34] improved accuracy by integrating the MobileCenterNet model with image enhancement techniques, effectively enhancing the quality of input images and enabling more precise detection.

3. Methodology

3.1. Overview of the MSFE Method

Given the efficiency advantages of the YOLO models in the detection task, this paper adopts the YOLOv8 detector as the benchmark architecture. In line with the end-to-end design of the YOLO series, the YOLOv8 retains the sequential construction of three modules: the Backbone network, Neck region, and Head region. Figure 1 illustrates its overall architecture. As the foundation of feature extraction, the Backbone is responsible for extracting feature information from the input image. In contrast, the Neck region indicates the FPN, which provides feature information for subsequent prediction through the multi-scale information fusion of deep and shallow features. Finally, the Head outputs the class and localization of the object instances.

To address the image blurring and dense small target issue in UOD, two structural designs are proposed to improve the detection capability of the model in this paper. The core idea focuses on how to integrate multi-scale feature information during the feature-extraction process. Two key points are summarized as follows.

(1) The structural design of the basic C2F module in the Backbone network. To enhance the information acquisition ability of basic C2F, a multi-scale perception operation is applied to the internal submodule BottleNeck (BNK) in C2F. This design aims to increase its sensitivity to blurred object instances. Finally, the reconstructive C2F is applied to the last two modules of the Backbone. Intuitively, the “Stage Layer 3” and “Stage Layer 4” in Figure 1 locate the application of this design. More details are presented in Section 3.2.

(2) The reconstruction of the FPN module. The main function of FPN is to integrate the deep and shallow features to exact effective information. In this setting, this paper further adds a shallow branch into the FPN framework. We expect that the extra branch can supply more large-scale features for subsequent prediction. The red box and ‘- -’ in the top region locate this design in Figure 1. A further explanation is presented in Section 3.3.

3.2. The Structural Design of C2F Module

As an evolution version of normal DarkNet in YOLO detection models, the CSPDarkNet has been the mainstream structure of the Backbone network from the YOLOv5 model. Taking the CSPDarkNet as the benchmark structure, the YOLOv8 network further enhances its feature-extraction ability. The main design concept is that it adopts the C2F (CSPDarkNet to 2-Stage FPN) module to replace the C3 module. This module can enrich gradient flow information by multi-branch structure while maintaining a lightweight structure. In the Backbone region, the C2F module is applied after the convolution layers to further enhance and refine the extracted feature. In the Neck region, the C2F module is applied before prediction to optimize the feature maps and reduce redundant information. To give

an intuitive diagram, the overall structure of the C2F has been illustrated in Figure 2a. Furthermore, the core component, namely BottleNeckK (BNK), has been shown in Figure 2b.

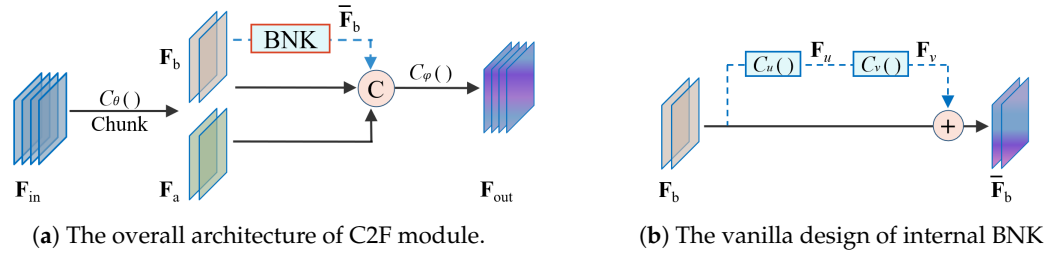


Figure 2. The diagram of the C2F module and the initial design of the internal BNK (BottleNeckK). (a) The overall architecture of the C2F module. (b) The vanilla structure of the internal BNK.

3.2.1. Introduction of the Initial C2F Module

(1) The overall architecture of the C2F module

According to the diagram in Figure 2a, the workflow of the C2F can be briefly described as follows. Given an input feature $F_{in} \in \mathbb{R}^{H \times W \times C}$, the first step is to transform it as $F_a \in \mathbb{R}^{H \times W \times C'/2}$ and $F_b \in \mathbb{R}^{H \times W \times C'/2}$, through a $C_\theta(\cdot)$ -based feature extraction and subsequent channel segmentation operation (chunk), in which the $C_\theta(\cdot)$ is a convolution layer. Second, the feature branch F_b is utilized to generate a new feature \bar{F}_b through the forward computation of the BottleNeckK (BNK) submodule. Third, the features F_a , F_b , \bar{F}_b are concatenated (Concat) as the combination $[F_a, F_b, \bar{F}_b]$ in terms of channel. Finally, the feature group will be further learned as the output feature F_{out} via a channel compression of the convolutional layer $C_\varphi(\cdot)$. The entire process is formalized as follows.

$$F_{in} \xrightarrow[\text{Chunk}]{\theta} \begin{cases} F_b \xrightarrow{\text{BNK}} \bar{F}_b \\ F_a, F_b \end{cases} \xrightarrow{\text{Concat}} [F_a, F_b, \bar{F}_b] \xrightarrow{\varphi} F_{out} \quad (1)$$

In line with Equation (1), it is clear that the terminal feature F_{out} is actually produced from three types of features, which include the initial transformation results F_a and F_b of input F_{in} , as well as the evolution form \bar{F}_b of the F_b .

In summary, the C2F improves the feature diversity by multi-branch construction of intermediate features. In contrast to the initial F_a and F_b , we find that it turns to supply a new feature \bar{F}_b for the terminal output. In view of parameter updates, this multi-branch structure can enrich the gradient flow information from loss functions during training.

(2) The initial BNK submodule in C2F

For the internal operation of BNK, Figure 2b illustrates its workflow. In which a residual architecture is used to generate the additional feature \bar{F}_b for C2F. In brief, taking the F_b as input feature, the \bar{F}_b is produced with the combination (sum) of the F_b and its evolution version F_v , where the evolution feature F_v is acquired through a serial CNN parameter set (u, v) in a residual branch. Formally, it can be formalized as follows.

$$\begin{cases} F_b \xrightarrow{u} F_u \xrightarrow{v} F_v \\ F_b \rightarrow F_b \end{cases} \rightarrow F_b + F_v \rightarrow \bar{F}_b \quad (2)$$

According to Equation (2), we can find that the BNK actually uses a residual branch to conduct a simple repair of the original F_b . In this setting, the branch becomes the only support or reliance to increase feature diversity. This is obviously unfavorable to extract rich feature information. For example, the initial BNK can play a positive role in real-life scenes. However, when encountering image blurring issues, the simple residual construction often

struggles to capture detailed information, such as the high-frequency edge details. This drawback may trigger a low adaptability of detectors in underwater environments.

3.2.2. The Design of the Local Awareness BNK (LABNK)

In this subsection, we present a local awareness (LA) design for the initial BNK, namely LABNK. In brief, taking the initial BNK as a benchmark framework, it no longer directly reuses the input feature but extracts new features through the LA operation. This avoids the single dependence of terminal output on the residual branch. The key LA operation is inspired by the HCF-Net [51] in the infrared object detection task in which one of the attention modules is constructed with the multi-scale element-wise multiplication operation to enhance the blurry information in the infrared images.

(1) The overall architecture of the LABNK module

Figure 3a illustrates the internal structure of the LABNK module. In terms of framework, the LABNK completely follows the residual construction of the initial BNK.

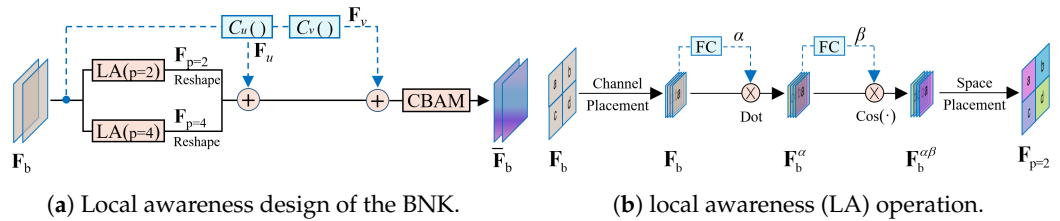


Figure 3. The schematic diagram of the local awareness enhanced BottleNeck submodule (LABNK) and the mechanism of the internal local awareness (LA) module. (a) The workflow of LABNK. Formally, it conducts the symmetrical awareness operation with $p = 2, 4$. (b) The principle of the internal LA module. It adopts two weighting operations to highlight the fine-grained information.

To be specific, instead of the simple copy of input feature F_b , the LABNK employs a symmetrical LA operation to obtain critical information from F_b . In other words, it conducts a patch-wise (2×2 and 4×4) feature enhancement to the critical information in F_b , namely LA ($p = 2$) and LA ($p = 4$). Subsequently, the values from different branches are combined through the summation, i.e., $\tilde{F}_b = F_{p=2} + F_{p=4} + F_u + F_v$. Finally, the output feature \bar{F}_b is obtained through the spatial channel attention module (CBAM). Formally, its overall workflow is represented by the following mapping expression.

$$\begin{cases} F_b \xrightarrow{u} F_u \xrightarrow{v} F_v \\ F_b \xrightarrow{LA(p=2)} F_{p=2} \\ F_b \xrightarrow{LA(p=4)} F_{p=4} \end{cases} \rightarrow F_u + F_v + F_{p=2} + F_{p=4} \xrightarrow{CBAM} \bar{F}_b \quad (3)$$

In summary, unlike the initial BNK, the LABNK does not reuse the input F_b for numerical supplementation. Instead, it employs the local awareness (LA) operation to complete a precise feature extraction through a local-aware enhancement operation. In addition, LABNK utilizes the intermediate feature F_u in the residual branch as the information source for the output and employs CBAM [52] to integrate multiple feature types. Since the CBAM is commonly used to integrate various types of feature information, its working principle is not detailed here. The following section provides a detailed discussion of the local awareness operation, which serves as the core component of the LABNK.

(2) The internal operation of the local awareness (LA) module

As shown in Figure 3b, the LA module consists of three main steps: channel placement, feature weighting, and space placement. The workflow is described as follows.

(i) Channel Placement: The unfold operation is employed to transform the input feature F_b from a high-dimensional spatial feature into a lower-dimensional channel feature.

(ii) Feature Weighting: The aim of this operation is to conduct the enhancement of critical information in F_b . It performs weighting operations on internal component of F_b at the channel level, following the algorithmic flow: $F_b \xrightarrow{FC^a} \alpha \cdot F_b \xrightarrow{FC^b} \cos(\beta, \alpha \cdot F_b)$. The two FC modules are weight generation modules similar to the SE channel attention mechanism [53], which are used to generate weight coefficient vectors α and β . During the weight generation, the dot product and cosine similarity are employed, respectively.

(iii) Space Placement: This operation performs a conversion from channel-level features to spatial-level features. Its goal is to restore the low-dimensional features to high-dimensional ones in order to match the initial dimension.

In summary, the LA module involves applying two sequential weighting operations to feature components in low-dimensional channel mode. As shown in Figure 3a, this paper employs the $p = 2$ and $p = 4$ settings to perform the local (patch-wise) weighting operation. This module enhances the model's ability to capture detailed information by selective emphasis on key information or suppression of irrelevant information. This method is crucial for addressing the issue of weakened feature information of small targets and blurry target boundaries in underwater images.

3.3. Small Receptive Field Detection Layer

The superiority of DLOD methods mainly stems from the robust feature extraction ability of the Backbone network, which provides strong support for subsequent detection implementation. In brief, owing to extensive training data, the CNNs are able to integrate the semantic information of the input image. However, the dimension compression of features in the Backbone network often leads to the vanishing or weakening phenomenon of feature information. This has been a critical bottleneck to detection performance.

To address the above issue, the FPN structure [54–57] combines shallow and deep features to integrate detail and semantic information, significantly improving detection performance. The multi-scale feature joint framework has become the mainstream approach in detection models. For example, the YOLO series generally adopts a three-branch FPN structure, which generates three types of detection heads, namely the 20×20 , 40×40 , and 80×80 size, as shown in Figure 1. However, this usual three-branch mode often struggles with extremely small targets. In this case, some additional branch setting may serve as a feasible solution, such as the four-branch attempts in the drone detection field [58–60].

As the stated image blurring and dense small target issue in the introduction, the normal FPN setting is no longer sufficient to handle the worsening conditions under the water detection scene. This motivates us to inject an additional shallow feature branch into the normal FPN structure. Taking the YOLOv8 as a benchmark framework, the top region in Figure 1 shows this idea. Formally, the red 'Top Down Layer3' and '-' in Figure 1 illustrate the forward propagation of the features extracted from the Stem Layer.

As shown in Figure 1, the new 160×160 feature from the Stem Layer has been injected into the Neck region, providing richer detailed information by combining with the deeper features. For the operational detail, this branch still follows the symmetrical construction form of the FPN structure. To be specific, it still receives assistance from small-scale features from deeper layers and provides large-scale features of the current layer for subsequent detection branches. This symmetrical addition plays a crucial role in improving the model's ability to detect dense small targets, and the effectiveness of this auxiliary branch is validated through the ablation study in the experimental section.

3.4. Loss Function

For model training, the bounding box regression loss L_{reg} and category loss L_{cls} are utilized for parameter update. The overall loss is presented in Equation (4).

$$L_{\text{total}} = L_{\text{reg}} + L_{\text{cls}} \quad (4)$$

where the regression loss L_{reg} consists of CIoU loss and DFL loss. The two parts of the loss are represented by Equation (5) and Equation (6), respectively.

$$\text{CIoU} = \text{IoU} - \frac{\rho^2(b, b^{st})}{c^2} - \alpha v \quad (5)$$

where IoU represents the intersection ratio of the two bounding boxes; c denotes the diagonal length of the smallest closed box covering both bounding boxes; $\rho(\cdot)$ refers to the Euclidean distance; b, b^{st} represent the coordinates of the centers of the predicted and actual boxes, respectively; $\alpha = \frac{v}{(1-\text{IoU})+v}$ is the weight coefficient; and $v = \frac{4}{\pi^2} \left(\arctan \frac{w^{st}}{h^{st}} - \arctan \frac{w}{h} \right)^2$ measures the difference in the shapes of the predicted and actual boxes.

$$\text{DFL}(\mathcal{S}_i, \mathcal{S}_{i+1}) = -((y_{i+1} - y) \log(\mathcal{S}_i) + (y - y_i) \log(\mathcal{S}_{i+1})) \quad (6)$$

where \mathcal{S}_i and \mathcal{S}_{i+1} represent the predicted values of the network output and the proximity region, respectively. Similarly, y, y_i , and y_{i+1} denote the true label value, the integral value of the label, and the integral value of the proximity label.

The category loss, denoted as L_{cls} , is calculated using the binary cross entropy loss (BCE Loss), as described in Equation (7).

$$L_{\text{cls}} = -\frac{1}{n} \sum_{i=1}^n [y_i \cdot \log p(\hat{y}_i) + (1 - y_i) \cdot \log(1 - p(\hat{y}_i))] \quad (7)$$

where y_i represents the true label, and \hat{y}_i denotes the predicted value of the sample.

3.5. Effect Validation of the Two Structural Components

This section validates the effect of different methods on the public DUO dataset. Taking the YOLOv8n as the benchmark model, Figure 4 illustrates the test performance trends under various settings. To be specific, the brown forked (\times) curve draws the test accuracy of the YOLOv8n. The blue triangle (\triangle) curve represents the accuracy trend with the LABNK-based YOLOv8n. The yellow dot (\circ) curve illustrates the effect of the additional shallow branch (New Branch, NB). Finally, the red square (\square) curve illustrates the collaborative effect of both methods (LABNK + NB).

- The individual validation of LABNK and New Branch: For the LABNK, its positive effect on performance improvement becomes gradually apparent after 150 iterations. In contrast, the additional shallow branch yields noticeable improvements as early as the 50th epoch. In summary, the trend of test performance trend in Figure 4 indicates that both operations can significantly enhance the detection accuracy of the model.
- The collaborative validation of LABNK and New Branch: In contrast to the baseline method, the combination of the two operations (LABNK + NB) shows a more positive effect from the 80th epoch. This indicates that there is no conflict between the two operations. Furthermore, we can see that their functional overlap is not significant. For example, taking the NB method as the baseline, the combined setting provides a greater accuracy advantage across the entire training process. One main reason is that both operation items perform multi-scale feature enhancement at different levels. The LABNK focuses more on fine-grained operations within the convolutional module. In contrast, the NB emphasizes the feature fusion mechanism across different layers.

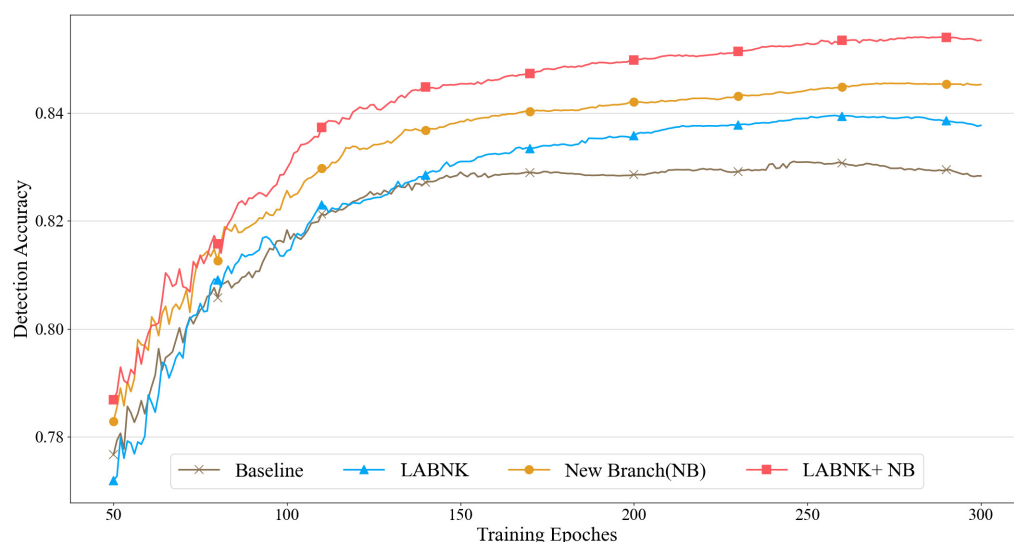


Figure 4. The test performance trends of different methods during training on the DUO dataset.

4. Experiments

4.1. Underwater Object Detection Datasets

This section evaluates the effectiveness of the proposed method on two publicly available datasets: DUO [61] and RUOD [62]. The DUO dataset includes 7782 labeled underwater images with four identities: echinus, starfish, scallop, and holothurian. The pre-set training data with 6671 images is used to learn the parameters of the model. In addition, the test set with 1111 images is used to conduct model evaluation. Similarly, the RUOD dataset contains 14,000 labeled underwater images with ten identities. The training and test sets consist of 9800 and 4200 images, respectively.

4.2. Experimental Setting

(1) Environment configuration: The hardware platform and parameters of the environment used in the experimental training phase are shown in Table 1.

Table 1. Parameters of Environment and Hardware Platform.

Environment	Version
Operating System	Ubuntu 20.04
CUDA Version	11.7
CPU	Intel(R) Core(TM)i7-12700KF@3.6GHz
GPU	GeForce RTX 3090
Display Memory	24G
Python Version	Python3.9.17
Deep-Learning Framework	Pytorch 2.0.1

(2) Training Settings: None of the experiments in this paper used pretrained weights. The specific parameters for training are detailed in the left half of Table 2. To prevent overfitting, the training process employed the model's default data augmentation methods. The settings used for data augmentation are also shown in the right half of Table 2.

Table 2. Training Setting and Data Augmentation Method in the YOLOv8n.

Training Setting	Value	Data Augmentation	Value
optimizer	SGD (weight decay = 5×10^{-4})	(hsv_h, hsv_s, hsv_v)	(0.015, 0.7, 0.4)
batch size	16	translate	0.1
epoch	300	scale	0.5
initial training rate	0.01	fliplr	0.5
image size	640×640	mosaic	1.0

(3) Evaluation metrics: In this paper, we evaluate the model's performance using five metrics: mAP(50), mAP(50:95), Params, Detection Time, and FLOPs. mAP(50) represents the mean Average Precision at an Intersection over Union (IoU) threshold of 0.5, while mAP(50:95) denotes the mean Average Precision averaged over IoU thresholds ranging from 0.5 to 0.95 in increments of 0.05. Additionally, Params measures the model's complexity, and Detection Time assesses its real-time performance. FLOPs are also utilized to assess the model's computational efficiency. The formula for mAP is given in Equation (8).

$$\text{mAP} = \frac{1}{K} \sum_{k=1}^K \text{AP}(k) \quad (8)$$

where K is the total number of identities in the dataset. AP represents the value of the area under the precision (P) and recall (R) curves for a single identity. P focuses on prediction accuracy, measuring how many of the predicted positive samples are actually correct. R focuses on detection comprehensiveness, indicating how many of the true positive samples are correctly identified by the model. The formula for P and R is given in Equation (9).

$$\begin{cases} P = TP / (TP + FP) \\ R = TP / (TP + FN) \end{cases} \quad (9)$$

where the above equation is the precision equation based on the confusion matrix, TP indicates that the predicted value is a positive sample and the true value is also a positive sample; FP indicates that the predicted value is a positive sample, but the true value is a negative sample; FN indicates that the predicted value is a negative sample, the true value is also a negative sample.

4.3. Ablation Study on the DUO Dataset

To validate the contributions of the two components of the MSFE method to detection accuracy, this section sequentially evaluates the performance improvements by the LABNK module and the New Branch (NB) measure, using YOLOv8n as the baseline model. Specifically, the model is trained for 300 epochs on the DUO training set, following the training setup described in Section 4.2.

- In terms of AP for each identity: The individual use of both components, compared to the baseline method, yields a positive effect, with the NB component achieving better results in detecting holothurian. Additionally, the combination of two components also demonstrates a slight advantage in accuracy for some identities, except for holothurian. This suggests that the combined approach may not provide a universal improvement across all identities.
- In terms of comprehensive performance: Based on mAP(50) and mAP(50:95), the individual use of the two components has resulted in improving performance compared to the baseline model, as shown in Table 3. Furthermore, compared to using LABNK and NB as individual benchmarks, the combination of both components

achieves the greatest improvement. This is primarily due to the synergistic effect of the two multi-scale feature enhancement operations.

Table 3. The Comparison of Ablation Experiments with Different Setups.

No.	LABNK	NB	AP				mAP (50)	mAP (50:95)
			Holothurian	Echinus	Scallop	Starfish		
1	×	×	83.1%	91.9%	65.3%	91.9%	83.0%	63.8%
2	✓	×	84.1%	92.1%	68.8%	92.4%	84.3%	64.2%
3	×	✓	85.3%	92.2%	68.2%	92.5%	84.5%	65.5%
4	✓	✓	84.7%	92.6%	70.8%	93.3%	85.4%	66.6%

4.4. Visualization Validation

To provide a more intuitive validation of the MSFE method, this section employs the visualization of attention heatmaps to verify its effectiveness. As described in the Grad-CAM [63] study, attention heatmaps effectively reflect the regions of interest in the input space, therefore facilitating the understanding of the model's decision basis.

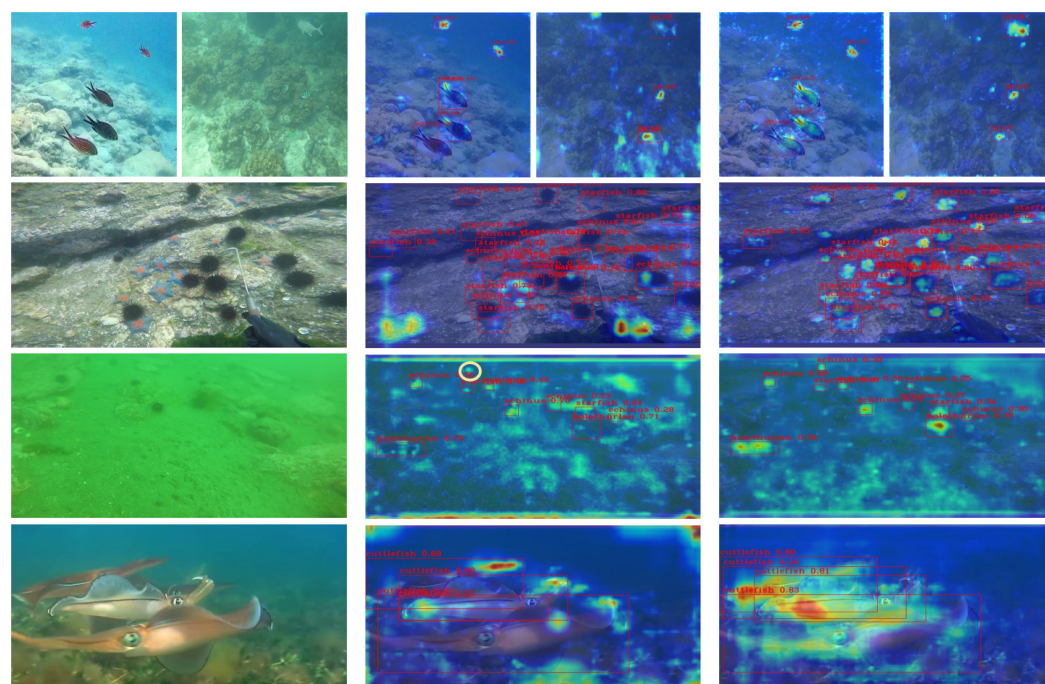
For the implementation, the RUOD dataset is used as the benchmark. We separately obtain the baseline detector (YOLOv8n) and the MSFE-based detector following the training setup in Section 4.2. Subsequently, several representative examples from the test set are selected to compare the differences in attention region between the two detectors.

4.4.1. Comparative Validation in Blurry Scenario

For the image blurring issue, this subsection selects five underwater images from the test set under low-light conditions to compare the attention difference between the two detectors. The original images and their corresponding attention heatmaps are shown in Figure 5, in which the depth of the color in the heatmap indicates the intensity of the attention. For instance, deeper colors (red and yellow) indicate stronger attention in that region. The breadth of the heatmap indicates the range of attention to the target region.

According to the comparison in Figure 5, the following conclusions can be drawn.

- **Small-sized targets:** (i) Normal background: Taking the three images in the first and second rows as references, the baseline model (YOLOv8n) often exhibits insufficient attention to the target, which clearly increases the risk of missed detections. In contrast, with the support of LABNK and NB, the MSFE method demonstrates a stronger attention intensity on small-sized targets. In addition, it reduces unnecessary attention on non-target regions in terms of attention breadth. (ii) Weak background-target differentiation: In the third row, both models exhibit fuzzy attention regions. However, we can find that the MSFE method provides a clearer distinction between the background and target regions. For instance, the YOLOv8n fails to detect the sea urchin target (highlighted by the yellow circular annotation in the middle image).
- **Large-sized targets:** For the large-sized targets with occlusion, taking the cuttlefish instance in the fourth row as the reference, the baseline model shows insufficient attention to the target region, which may lead to difficulty in clearly identifying the boundary information of different target instances. In contrast, the MSFE provides a more comprehensive assessment of the target region. In this case, the sufficient attention supply is beneficial for acquiring the target's boundary information.



(a) Test input images.

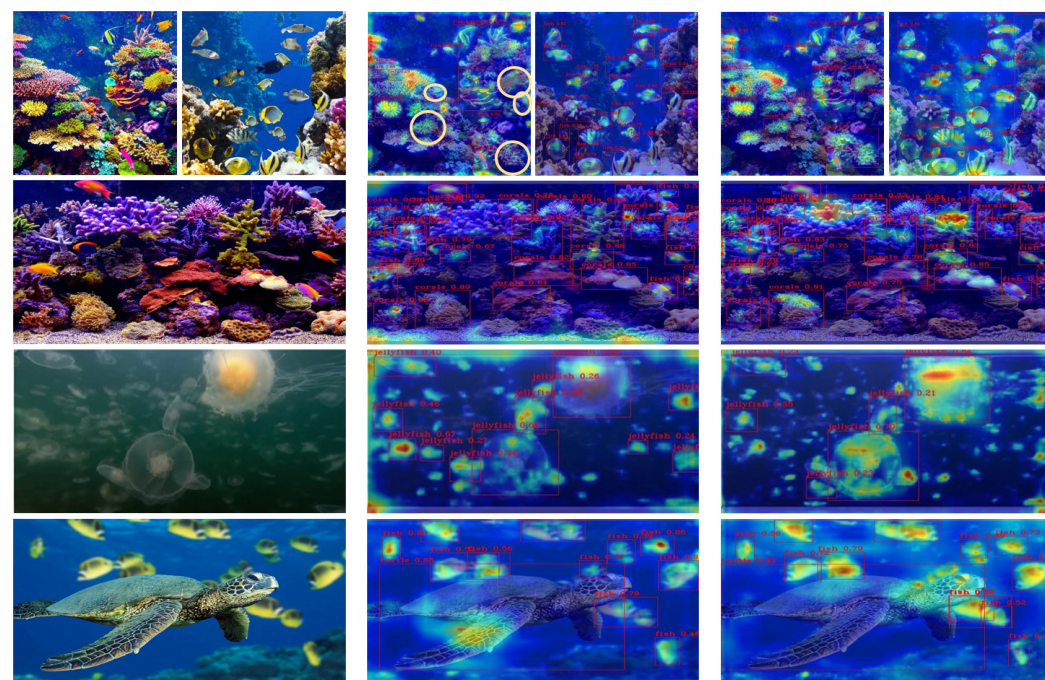
(b) Heatmaps from YOLOv8n.

(c) Heatmaps from MSFE.

Figure 5. The comparative validation with the visualization of attention heatmap in blurry scenes. (a) Original test images. (b) The visualized results from the initial YOLOv8n model. (c) The visualized results from the MSFE-based YOLOv8n.

4.4.2. The Comparative Validation in Dense Scenario

For the dense small target issue, five images from the test set are used to verify the differences in the attention regions between the two models. Figure 6 visualizes the heatmap and detection results from two models.



(a) Test input images.

(b) Heatmaps from YOLOv8n.

(c) Heatmaps from MSFE.

Figure 6. The comparative validation with the visualization of attention heatmap in dense scenes. (a) Original test images. (b) The visualized results from the initial YOLOv8n. (c) The visualized results from the MSFE-based YOLOv8n.

As shown in Figure 6, both models show significant improvements in distinguishing between targets and backgrounds under sufficient lighting conditions. However, regarding the multi-scale target problem, the following conclusions can be drawn.

- Non-multi-scale situation: Referring to the images in the first two rows, the initial YOLOv8n shows weak attention on the target, similar to the situation shown in Figure 5. To confirm this case, the missed and false detection targets are specifically annotated in the third image of the first row (highlighted by the yellow circular annotations). In contrast, the MSFE method shows much stronger attention on the target.
- Multi-scale situation: Taking the images in the third and fourth rows as examples, when there are large scale (size) differences between different targets, the MSFE clearly demonstrates a more global capture of the feature information of large-sized targets. For instance, the MSFE performs a more comprehensive feature assessment for the key head region of a large-sized jellyfish target, which evidently enhances the model's generalization capability when dealing with multi-scale target instances.

4.5. General Comparison on the DUO Dataset

To provide a comprehensive performance evaluation, this subsection compares different methods on the public DUO dataset. Specifically, this section compares the MSFE method with recent underwater object detection methods. In addition, we list the test results of some regular detection models. All comparison results are listed in Table 4.

Table 4. The General Comparison of Mainstream Models on the DUO Dataset.

Model	mAP (50)	mAP (50:95)	Params/M	FLOPs/G	Time/ms
Faster R-CNN [37]	83.0%	63.5%	41.4	210.3	125.0
Cascade R-CNN [37]	83.5%	64.8%	69.2	236.0	62.5
RetinaNet [37]	81.7%	61.9%	36.4	207.0	58.8
YOLOv5n	81.2%	60.2%	1.8	4.1	2.1
YOLOv6n [23]	80.8%	60.5%	4.7	11.4	3.4
YOLOX-nano [22]	71.1%	46.0%	3.8	2.6	2.9
YOLOv7-tiny [24]	81.9%	59.9%	6.0	13.1	2.7
YOLOv8n [25]	83.0%	63.8%	2.8	8.1	3.3
YOLOv9t [26]	84.0%	64.5%	2.6	10.7	6.3
CIM-DAIM [64,65]	77.0%	—	26.1	118.9	7.6
YOLOv7-CHS [66]	84.6%	65.5%	32.0	40.3	31.3
Deformable-DETR [37]	84.4%	63.7%	40.0	200.7	52.6
MSFE-YOLOv8n (Ours)	85.4%	66.6%	3.1	12.2	5.7

- In terms of accuracy metrics: The MSFE exhibits a more notable advantage over other methods in both mAP(50) and mAP(50:95) metrics. In contrast to the baseline method (YOLOv8n), it achieves an accuracy improvement of 2.4%. Compared to the recent underwater detection method (YOLOv7-CHS) in 2023, it also achieves an accuracy gain of approximately 1.0%. It is worth mentioning that the report of the MSFE method is built upon the lightweight YOLOv8n model. Therefore, it has the potential to achieve higher accuracy when applied to other large models.
- In terms of other metrics: The MSFE method shows a slight trade-off trouble between accuracy and efficiency. To be specific, for the benchmark YOLOv8n model, the auxiliary modification results in an increase of 0.3M in parameters and an additional computational cost of 4G in FLOPs. In this case, other conventional small models without a specific design for underwater object detection demonstrate better detection efficiency, such as the YOLOv5n, YOLOv6n, YOLOX-nano, and YOLOv9t.

4.6. Comprehensive Performance Comparison Based on RUOD Dataset

This subsection further evaluates the MSFE method on the RUOD dataset. Similar to the previous comparison, the MSFE is compared with existing underwater object detection methods and several conventional small YOLO models. Table 5 lists the comparison results. According to the reports, a comparative conclusion is reached as follows.

Table 5. The General Comparison of Mainstream Models on the RUOD Dataset.

Model	mAP (50)	mAP (50:95)	Params/M	FLOPs/G	Time/ms
Faster R-CNN [67]	81.8%	52.8%	41.4	246.0	114.3
Cascade R-CNN [67]	81.1%	54.8%	69.2	271.2	90.9
RetinaNet [67]	79.3%	50.7%	36.4	273.4	68.2
YOLOv5n	80.6%	53.9%	1.8	4.2	2.5
YOLOv6n [23]	80.2%	56.6%	4.6	11.5	3.3
YOLOX-nano [22]	70.2%	42.8%	3.8	2.9	3.5
YOLOv7-tiny [24]	81.5%	55.1%	6.1	13.1	2.6
YOLOv8n [25]	82.9%	58.5%	2.8	8.1	3.4
YOLOv9t [26]	83.3%	59.3%	2.6	11.7	6.7
DETR [68]	82.6%	54.7%	15.9	62.0	25.0
MarineYOLO [69]	80.5%	44.6%	20.6	10.9	13.9
MSFE-YOLOv8n (Ours)	84.8%	61.3%	3.1	12.3	5.8

- In terms of accuracy metrics: In contrast to reports on the DUO dataset, all methods show a decline in test accuracy on the RUOD dataset. One possible reason is that the RUOD dataset contains 10 identities, making it more challenging than the DUO dataset, which has only 4 identities. To check the effect of MSFE, we observe that it gains similar improvements for the YOLOv8n. Moreover, it achieves an accuracy gain of about 1.5–2.0% over the best YOLOv9t in terms of mAP(50) and mAP(50:95). This further highlights its positive effect in handling underwater blurry images.
- In terms of other metrics: Taking small YOLO models as the baseline, the MSFE method exhibits similar disadvantages to the situation on the DUO dataset in terms of parameter count, computational cost, and detection time. The main reason is that the MSFE injects an additional LA module and a new feature branch into the initial BNK module and FPN, respectively. This drawback further gives rise to the disadvantages in terms of FLOPs and detection time. Fortunately, the MSFE method still maintains a certain efficiency advantage over other underwater detection methods. This is due to the fact that earlier methods inherently employed larger models.

5. Conclusions

In this paper, we present a Multi-Scale Feature Enhancement (MSFE) method for underwater object detection. It attempts to address the challenges of dense small instances and image blurring from the perspective of multi-scale feature fusion. For the Backbone network's internal structure, a multi-scale local awareness operation was applied to enhance the information integration capability of the basic C2F module. To facilitate the information integration of FPN between deep and shallow features, an additional shallow branch was introduced to provide large-scale features for subsequent predictions. In summary, the two structural designs, namely LABNK and NB, focused on internal fine-grained operations within the convolutional layers and external inter-layer collaboration, respectively.

Extensive experiments on two public datasets have demonstrated the effectiveness of the presented MSFE method. The attention heatmap visualizations provided supportive

evidence for its rationale. Furthermore, an ablation study was provided for individual validation and collaborative validation of both operation items.

Although the MSFE method has made some progress in underwater object detection, it does not involve other challenges such as class imbalance, noisy labels, and color distortion. Moreover, the parameter increments introduced by two structural designs have led to a decrease in detection efficiency. Overall, the main trend in underwater object detection methods should be to improve detection accuracy while ensuring low computational cost.

Author Contributions: Conceptualization, M.L., C.S. and W.L.; methodology, M.L., C.S., B.Q., A.T. and H.Y.; software, M.L., W.L. and A.T.; validation, B.Q., A.T. and H.Y.; formal analysis, B.Q. and A.T.; investigation, M.L., W.L. and A.T.; resources, C.S., B.Q. and H.Y.; data curation, M.L. and W.L.; writing—original draft preparation, M.L., C.S. and W.L.; writing—review and editing, C.S. and A.T.; visualization, W.L.; supervision, H.Y.; project administration, C.S. and H.Y.; funding acquisition, B.Q. and H.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62176107, and the National Natural Science Foundation of China, grant number 62376109.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author. The code is available at <https://github.com/mint1072/MSFE.git>, accessed on 1 November 2024).

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

References

1. Zhao, M.; Zhou, H.; Li, X. YOLOv7-SN: Underwater Target Detection Algorithm Based on Improved YOLOv7. *Symmetry* **2024**, *16*, 514. [\[CrossRef\]](#)
2. Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2001, Kauai, HI, USA, 8–14 December 2001; IEEE: Piscataway, NJ, USA, 2001; Volume 1, p. I.
3. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; IEEE: Piscataway, NJ, USA, 2005; Volume 1, pp. 886–893.
4. Felzenszwalb, P.; McAllester, D.; Ramanan, D. A discriminatively trained, multiscale, deformable part model. In Proceedings of the 2008 IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; IEEE: Piscataway, NJ, USA, 2008; pp. 1–8.
5. Chen, X.; Chen, H. A novel color edge detection algorithm in RGB color space. In Proceedings of the IEEE 10th International Conference On Signal Processing Proceedings, Beijing, China, 24–28 October 2010; IEEE: Piscataway, NJ, USA, 2010; pp. 793–796.
6. Beijbom, O.; Edmunds, P.J.; Kline, D.I.; Mitchell, B.G.; Kriegman, D. Automated annotation of coral reef survey images. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 1170–1177.
7. Nagaraja, S.; Prabhakar, C.; Kumar, P.P. Extraction of texture based features of underwater images using RLBP descriptor. In Proceedings of the 3rd International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA) 2014, Bhubaneswar, India, 14–15 November 2014; Springer: Berlin/Heidelberg, Germany, 2015; Volume 2, pp. 263–272.
8. Fatan, M.; Daliri, M.R.; Shahri, A.M. Underwater cable detection in the images using edge classification based on texture information. *Measurement* **2016**, *91*, 309–317. [\[CrossRef\]](#)
9. Srividhya, K.; Ramya, M. Accurate object recognition in the underwater images using learning algorithms and texture features. *Multimed. Tools Appl.* **2017**, *76*, 25679–25695. [\[CrossRef\]](#)
10. Shi, X.; Huang, H.; Wang, B.; Pang, S.; Qin, H. Underwater cage boundary detection based on GLCM features by using SVM classifier. In Proceedings of the 2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), Hong Kong, China, 8–12 July 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1169–1174.
11. Bazeille, S.; Quidu, I.; Jaulin, L. Color-based underwater object recognition using water light attenuation. *Intell. Serv. Robot.* **2012**, *5*, 109–118. [\[CrossRef\]](#)
12. Hou, G.J.; Luan, X.; Song, D.L.; Ma, X.Y. Underwater man-made object recognition on the basis of color and shape features. *J. Coast. Res.* **2016**, *32*, 1135–1141. [\[CrossRef\]](#)

13. Cheng, E.; Lin, X.; Chen, Y.; Yuan, F.; Yang, W. GLCM Based No-Reference Perceptual Blur Metric For Underwater Blur Image. *Int. J. Circuits Syst. Signal Process.* **2016**, *10*, 291–296.
14. Chen, Z.; Zhang, Z.; Dai, F.; Bu, Y.; Wang, H. Monocular vision-based underwater object detection. *Sensors* **2017**, *17*, 1784. [\[CrossRef\]](#)
15. Vasamsetti, S.; Setia, S.; Mittal, N.; Sardana, H.K.; Babbar, G. Automatic underwater moving object detection using multi-feature integration framework in complex backgrounds. *IET Comput. Vis.* **2018**, *12*, 770–778. [\[CrossRef\]](#)
16. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
17. Girshick, R. Fast r-cnn. *arXiv* **2015**, arXiv:1504.08083.
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [\[CrossRef\]](#)
19. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask r-cnn. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969.
20. Redmon, J. You only look once: Unified, real-time object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
21. Redmon, J.; Farhadi, A. YOLO9000: Better, faster, stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.
22. Ge, Z. Yolox: Exceeding yolo series in 2021. *arXiv* **2021**, arXiv:2107.08430.
23. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
24. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. 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, Vancouver, BC, Canada, 17–24 June 2023; pp. 7464–7475.
25. Ultralytics. YOLOv8. 2023. Available online: <https://github.com/ultralytics/ultralytics> (accessed on 1 November 2024).
26. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. Yolov9: Learning what you want to learn using programmable gradient information. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; Springer: Berlin/Heidelberg, Germany, 2024; pp. 1–21.
27. Chen, L.; Huang, Y.; Dong, J.; Xu, Q.; Kwong, S.; Lu, H.; Lu, H.; Li, C. Underwater Object Detection in the Era of Artificial Intelligence: Current, Challenge, and Future. *arXiv* **2024**, arXiv:2410.05577.
28. Cong, X.; Zhao, Y.; Gui, J.; Hou, J.; Tao, D. A Comprehensive Survey on Underwater Image Enhancement Based on Deep Learning. *arXiv* **2024**, arXiv:2405.19684.
29. Xu, S.; Zhang, M.; Song, W.; Mei, H.; He, Q.; Liotta, A. A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing* **2023**, *527*, 204–232. [\[CrossRef\]](#)
30. Jian, M.; Yang, N.; Tao, C.; Zhi, H.; Luo, H. Underwater object detection and datasets: A survey. *Intell. Mar. Technol. Syst.* **2024**, *2*, 9. [\[CrossRef\]](#)
31. Liu, C.; Shu, X.; Xu, D.; Shi, J. GCCF: A lightweight and scalable network for underwater image enhancement. *Eng. Appl. Artif. Intell.* **2024**, *128*, 107462. [\[CrossRef\]](#)
32. Zhang, X.; Fang, X.; Pan, M.; Yuan, L.; Zhang, Y.; Yuan, M.; Lv, S.; Yu, H. A marine organism detection framework based on the joint optimization of image enhancement and object detection. *Sensors* **2021**, *21*, 7205. [\[CrossRef\]](#)
33. Han, F.; Yao, J.; Zhu, H.; Wang, C. Underwater image processing and object detection based on deep CNN method. *J. Sens.* **2020**, *2020*, 6707328. [\[CrossRef\]](#)
34. Ji, W.; Peng, J.; Xu, B.; Zhang, T. Real-time detection of underwater river crab based on multi-scale pyramid fusion image enhancement and MobileCenterNet model. *Comput. Electron. Agric.* **2023**, *204*, 107522. [\[CrossRef\]](#)
35. Liu, Q.; Huang, W.; Duan, X.; Wei, J.; Hu, T.; Yu, J.; Huang, J. DSW-YOLOv8n: A new underwater target detection algorithm based on improved YOLOv8n. *Electronics* **2023**, *12*, 3892. [\[CrossRef\]](#)
36. Zhao, S.; Zheng, J.; Sun, S.; Zhang, L. An improved YOLO algorithm for fast and accurate underwater object detection. *Symmetry* **2022**, *14*, 1669. [\[CrossRef\]](#)
37. Feng, J.; Jin, T. CEH-YOLO: A composite enhanced YOLO-based model for underwater object detection. *Ecol. Inform.* **2024**, *82*, 102758. [\[CrossRef\]](#)
38. Shen, X.; Sun, X.; Wang, H.; Fu, X. Multi-dimensional, multi-functional and multi-level attention in YOLO for underwater object detection. *Neural Comput. Appl.* **2023**, *35*, 19935–19960. [\[CrossRef\]](#)
39. Zhou, Z.; Hu, Y.; Yang, X.; Yang, J. YOLO-based marine organism detection using two-terminal attention mechanism and difficult-sample resampling. *Appl. Soft Comput.* **2024**, *153*, 111291. [\[CrossRef\]](#)

40. Cai, Z.; Vasconcelos, N. Cascade r-cnn: Delving into high quality object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
41. Farhadi, A.; Redmon, J. Yolov3: An incremental improvement. In Proceedings of the Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; Springer: Berlin/Heidelberg, Germany, 2018; Volume 1804, pp. 1–6.
42. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. Yolov4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934.
43. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; Proceedings, Part I 14; Springer: Berlin/Heidelberg, Germany, 2016; pp. 21–37.
44. Cheng, S.; Wang, Z.; Liu, S.; Han, Y.; Sun, P.; Li, J. Attention-Based Lightweight YOLOv8 Underwater Target Recognition Algorithm. *Sensors* **2024**, *24*, 7640. [\[CrossRef\]](#)
45. Song, P.; Li, P.; Dai, L.; Wang, T.; Chen, Z. Boosting R-CNN: Reweighting R-CNN samples by RPN’s error for underwater object detection. *Neurocomputing* **2023**, *530*, 150–164. [\[CrossRef\]](#)
46. Wang, H.; Xiao, N. Underwater object detection method based on improved faster RCNN. *Appl. Sci.* **2023**, *13*, 2746. [\[CrossRef\]](#)
47. Chen, L.; Liu, Z.; Tong, L.; Jiang, Z.; Wang, S.; Dong, J.; Zhou, H. Underwater object detection using Invert Multi-Class Adaboost with deep learning. In Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN), Glasgow, UK, 19–24 July 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1–8.
48. Liu, H.; Song, P.; Ding, R. Towards domain generalization in underwater object detection. In Proceedings of the 2020 IEEE International Conference on Image Processing (ICIP), Abu Dhabi, United Arab Emirates, 25–28 October 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 1971–1975.
49. Gao, J.; Zhang, Y.; Geng, X.; Tang, H.; Bhatti, U.A. PE-Transformer: Path enhanced transformer for improving underwater object detection. *Expert Syst. Appl.* **2024**, *246*, 123253. [\[CrossRef\]](#)
50. Ji, X.; Chen, S.; Hao, L.Y.; Zhou, J.; Chen, L. FBDPN: CNN-Transformer hybrid feature boosting and differential pyramid network for underwater object detection. *Expert Syst. Appl.* **2024**, *256*, 124978. [\[CrossRef\]](#)
51. Xu, S.; Zheng, S.; Xu, W.; Xu, R.; Wang, C.; Zhang, J.; Teng, X.; Li, A.; Guo, L. HCF-Net: Hierarchical Context Fusion Network for Infrared Small Object Detection. In Proceedings of the 2024 IEEE International Conference on Multimedia and Expo (ICME), Niagara Falls, ON, Canada, 15–19 July 2024.
52. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Tel Aviv, Israel, 23–27 October 2018; pp. 3–19.
53. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
54. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
55. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
56. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
57. Ghiasi, G.; Lin, T.Y.; Le, Q.V. Nas-fpn: Learning scalable feature pyramid architecture for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 7036–7045.
58. Tang, S.; Zhang, S.; Fang, Y. HIC-YOLOv5: Improved YOLOv5 for small object detection. In Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 13–17 May 2024; IEEE: Piscataway, NJ, USA, 2024; pp. 6614–6619.
59. Liu, H.; Duan, X.; Lou, H.; Gu, J.; Chen, H.; Bi, L. Improved GBS-YOLOv5 algorithm based on YOLOv5 applied to UAV intelligent traffic. *Sci. Rep.* **2023**, *13*, 9577. [\[CrossRef\]](#)
60. Shang, J.; Wang, J.; Liu, S.; Wang, C.; Zheng, B. Small target detection algorithm for UAV aerial photography based on improved YOLOv5s. *Electronics* **2023**, *12*, 2434. [\[CrossRef\]](#)
61. Liu, C.; Li, H.; Wang, S.; Zhu, M.; Wang, D.; Fan, X.; Wang, Z. A dataset and benchmark of underwater object detection for robot picking. In Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shenzhen, China, 5–9 July 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 1–6.
62. Fu, C.; Liu, R.; Fan, X.; Chen, P.; Fu, H.; Yuan, W.; Zhu, M.; Luo, Z. Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing* **2023**, *517*, 243–256. [\[CrossRef\]](#)
63. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 618–626.
64. Yuan, J.; Hu, Y.; Sun, Y.; Yin, B. A multi-scale feature representation and interaction network for underwater object detection. *IET Comput. Vis.* **2023**, *17*, 265–281. [\[CrossRef\]](#)

-
65. Zeng, B.; Zhou, Y.; He, D.; Zhou, Z.; Hao, S.; Yi, K.; Li, Z.; Zhang, W.; Xie, Y. Research on Lightweight Method of Insulator Target Detection Based on Improved SSD. *Sensors* **2024**, *24*, 5910. [[CrossRef](#)]
 66. Zhao, L.; Yun, Q.; Yuan, F.; Ren, X.; Jin, J.; Zhu, X. YOLOv7-CHS: An Emerging Model for Underwater Object Detection. *J. Mar. Sci. Eng.* **2023**, *11*, 1949. [[CrossRef](#)]
 67. Gao, Z.; Shi, Y.; Li, S. Self-attention and long-range relationship capture network for underwater object detection. *J. King Saud-Univ.-Comput. Inf. Sci.* **2024**, *36*, 101971. [[CrossRef](#)]
 68. Lin, X.; Huang, X.; Wang, L. Underwater object detection method based on learnable query recall mechanism and lightweight adapter. *PLoS ONE* **2024**, *19*, e0298739. [[CrossRef](#)]
 69. Liu, L.; Chu, C.; Chen, C.; Huang, S. MarineYOLO: Innovative deep learning method for small target detection in underwater environments. *Alex. Eng. J.* **2024**, *104*, 423–433. [[CrossRef](#)]

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