Guo *et al. Complex Eng Syst* 2023;3:16
DOI: 10.20517/ces.2023.19

Complex Engineering Systems

Research Article

Open Access

AWF-YOLO: enhanced underwater object detection with adaptive weighted feature pyramid network

Qianren Guo[1,2], Yuehang Wang[3,2], Yongji Zhang[3,2], Hongde Qin[4], Hong Qi[3,2], Yu Jiang[3,2]

[1]College of Software, Jilin University, Changchun 130012, Jilin, China.
[2]Key Laboratory of Symbolic Computation and Knowledge Engineering, Ministry of Education, Jilin University, Changchun 130012, Jilin, China.
[3]College of Computer Science and Technology, Jilin University, Changchun 130012, Jilin, China.
[4]Science and Technology on Underwater Vehicle Laboratory, Harbin Engineering University, Harbin 150001, Heilongjiang, China.

Correspondence to: Prof. Yu Jiang, College of Computer Science and Technology, Jilin University, No. 2699 Qianjin Street, Changchun 130012, China. E-mail: jiangyu2011@jlu.edu.cn; ORCID: 0000-0001-9025-3375

How to cite this article: Guo Q, Wang Y, Zhang Y, Qin H, Qi H, Jiang Y. AWF-YOLO: enhanced underwater object detection with adaptive weighted feature pyramid network. *Complex Eng Syst* 2023;3:16. http://dx.doi.org/10.20517/ces.2023.19

Received: 17 Jun 2023 **First Decision:** 27 Jun 2023 **Revised:** 13 Aug 2023 **Accepted:** 25 Aug 2023 **Published:** 19 Sep 2023

Academic Editor: Hamid Reza Karimi **Copy Editor:** Fanglin Lan **Production Editor:** Fanglin Lan

Abstract

Underwater scenarios are influenced by various factors such as light attenuation, scattering, and absorption, which degrade the quality of images and pose significant challenges for underwater object detection in marine research and ocean engineering. To address these challenges, we propose a novel adaptive-weight feature detection framework based on YOLOv8, called AWF-YOLO, designed to detect objects in turbid underwater scenarios accurately. AWF-YOLO incorporates several key components to improve detection performance. Firstly, a novel adaptive-weight feature pyramid network is introduced to facilitate the fusion of multi-scale feature semantics. In addition, an adaptive-weight feature extraction module is proposed to enhance underwater object detection by capturing relevant and discriminative information to enhance feature extraction further. We integrate a dedicated small object detection head into the detection network to overcome the challenges associated with detecting small objects in complex underwater scenarios. This component focuses on effectively identifying and localizing small objects, leading to improved overall detection accuracy. Extensive experiments conducted on the detection underwater objects dataset demonstrate that the proposed AWF-YOLO achieves significant performance improvements, thus making it highly suitable for complex and dynamic underwater scenarios.

Keywords: Object detection, underwater image, deep learning, multi-scale feature fusion, YOLO

1. INTRODUCTION

Amid increasing concern for deep-sea and marine resources, underwater intelligent exploration plays an essential role in understanding marine ecosystems, conducting surveys in unexplored regions, and fostering sustainable marine utilization [1–3]. This pivotal technology has emerged as a significant driving force behind contemporary marine scientific research and exploration.

Underwater object detection holds immense significance in the realm of intelligent underwater exploration, as it enables precise localization and recognition of submerged objects. However, this task presents distinctive challenges that distinguish it from other computer vision applications [4]. Figure 1 illustrates the influence of underwater environments on image quality, wherein factors such as light attenuation, scattering, and absorption result in blurry, noisy, and low-contrast underwater images. Moreover, the presence of numerous dense small objects further complicates matters, thereby undermining the accuracy and reliability of detection algorithms. Consequently, it becomes crucial to conduct further research aimed at effectively optimizing underwater object detection techniques. These advancements should specifically address real-world application scenarios encountered in underwater environments, with the ultimate objective of enhancing the accuracy and robustness of underwater object detection.

Researchers have made significant progress in underwater object detection by leveraging the power of deep learning. Convolutional neural networks (CNNs) [5] have demonstrated their capability to learn complex patterns and extract discriminative features from underwater imagery, improving detection performance. The ability of CNNs to automatically learn and adapt to different data distributions makes them particularly suitable for handling the challenges presented by the underwater environment. In recent years, numerous studies have focused on developing specialized architectures and techniques to enhance underwater object detection. The YOLO (You Only Look Once) [6] series of CNNs are widely used for object detection tasks. These models effectively balance the trade-off between accuracy and speed, making them suitable for underwater object detection tasks. To tackle the challenges in underwater object detection, researchers adjust the YOLO series of object detection algorithms to incorporate specific features of underwater objects, leading to noteworthy advancements. For example, Al Muksit *et al.* [7] developed the YOLO-Fish network based on YOLOv3 [8] by modifying the upsampling step and adding spatial pyramidal pooling to reduce the false detection of dense small objects. Zhang *et al.* [9] proposed a lightweight and efficient underwater object detection network based on YOLOv4 [10]. Additionally, they employed MobileNetv2 [11] as the backbone network for the lightweight feature extraction. Zhao *et al.* [12] enhanced the feature extraction of YOLOv4 by optimizing the backbone network and feature fusion module, thereby improving its ability to extract features from blurry images. Peng *et al.* [13] enhanced multi-scale feature fusion by introducing fast connections on the feature pyramid and proposed a piecewise focal loss function to alleviate the class imbalance. These modifications enable more convenient feature fusion and improve the overall performance of the network. Shang *et al.* [14] addressed the robust multi-scale coordination control issue against adversarial nodes in directed networks, proposing a local-information-based multi-scale filtering algorithm and establishing conditions for achieving multi-scale consensus and robustness. Dai *et al.* [15] introduced a multi-scale channel attention module (MS-CAM) to address issues arising from scale variations and small objects in underwater scenes. The MS-CAM module is designed to effectively highlight both large objects with broader distributions and small objects with more localized patterns. By incorporating the MS-CAM module, the detection system becomes more capable of accurately identifying objects at extreme scales. Building upon YOLOv5, Liu *et al.* [16] introduced the TC-YOLO, which incorporates transformer and coordinate attention mechanisms into the feature extraction network and feature fusion network, respectively. This novel architecture aims to enhance feature extraction and alleviate the issue of inconspicuous underwater image features. Despite these advancements, these methods still face challenges in accurately detecting objects in complex underwater environments. Difficulties persist in effectively addressing issues such as image feature blurring and the presence of dense small objects in underwater scenes, resulting in suboptimal accuracy. Lei *et al.* [17] introduced an enhanced YOLOv5 algorithm based on

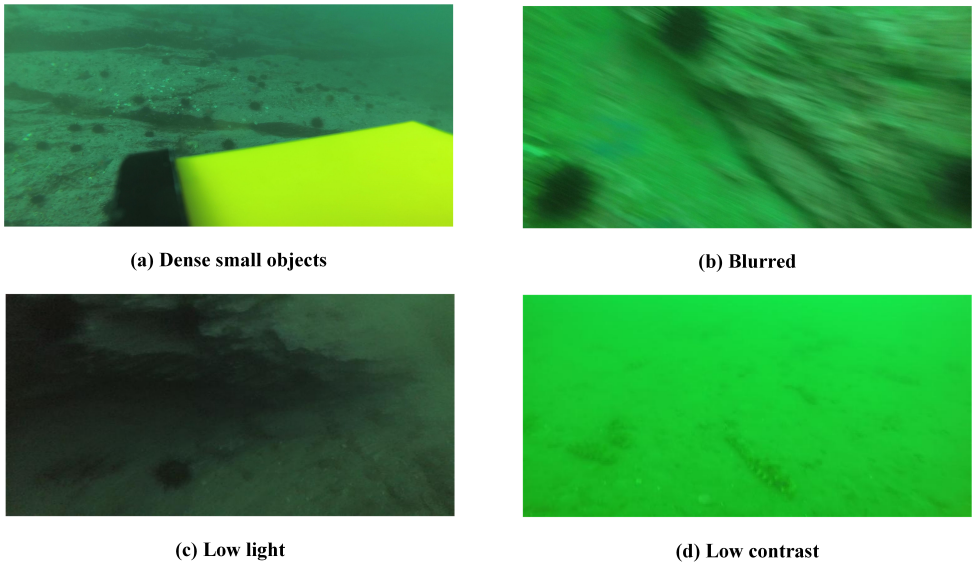

Figure 1. Some examples of challenging underwater detection scenarios.

the Swin Transformer backbone network, designed for images of underwater blurry objects. Li *et al.*[18] proposed an underwater object detection algorithm based on an improved YOLOv4. This algorithm utilizes the K-means++ algorithm to recluster anchor boxes in underwater scenes and employs the MSRCR method for underwater image enhancement. Ge *et al.*[19] introduced a single-stage underwater object detection method based on a dual-optimization network with feature anchor boxes. Qi *et al.*[20] proposed a novel underwater object detection network called UTD-Net, which is based on a hyperspectral unmixing method and a depth estimation model. This network introduces a novel joint anomaly detector and explores an autoencoder based on the depth estimation model to unmix target-water mixed pixels. In the domain of underwater image prior enhancement, Li *et al.*[21] established an Underwater Image Enhancement Benchmark (UIEB), consisting of 950 real-world underwater images, out of which 890 images have corresponding reference images. Additionally, Li proposed an underwater image enhancement network trained on this benchmark as a baseline, offering a reference point for future research in underwater image enhancement. In a parallel vein, Zhuang *et al.*[22] put forth a novel Retinex variational model guided by hyper-laplacian reflectance priors. This model simplifies a complex underwater image enhancement problem into a more manageable subproblem, allowing simultaneous estimation of reflection and illumination in a Retinex variational framework to enhance underwater images. Despite the impressive accuracy achieved by image prior enhancement, the integration of prior methods often results in a substantial investment of time in the enhancement process, leading to prolonged overall inference times. This issue is particularly pertinent for underwater object detection tasks with stringent real-time requirements. Hence, the design of an end-to-end model becomes exceedingly important to ensure swift inference times and meet the demands of real-time underwater object detection tasks.

In this paper, we propose an adaptive-weight underwater object detection framework called AWF-YOLO. Our proposed method addresses the challenges associated with complex and dynamic underwater scenes, specifically tackling issues such as image blurring caused by low contrast and the reduced detection accuracy resulting from dense small objects. By employing AWF-YOLO, we achieve improved accuracy in underwater object detection compared to existing approaches. To overcome the challenge of inconspicuous underwater image features, we propose three key modules: an Adaptive Weighted Feature Pyramid Network (AWFPN), an Adaptive Weighted Feature Extraction (AWFE) module, and a dedicated detection head. The AWFPN performs a weighted multi-scale fusion between deep feature maps with strong semantic information and shallow feature maps with rich spatial information but weaker semantic information. This fusion process ensures that each

feature map at different scales retains both strong semantic and spatial information, resulting in an optimal balance. Furthermore, this fusion approach is designed to maintain computational efficiency, ensuring that the network operates efficiently without compromising its accuracy. In addition, the AWFE module addresses the issue by assigning weights to the output feature maps of each branch within the module. These weights are dynamically learned, allowing the network to adaptively determine the importance of each feature map. By updating the weights, the module effectively integrates the feature maps, enhancing underwater object detection by capturing relevant and discriminative information. To specifically address the challenge of detecting dense small objects in underwater environments, we introduce a dedicated detection head designed to effectively locate and detect dense small objects underwater, leading to improved overall detection accuracy. Our main contributions are as follows:

(1) A highly accurate framework for underwater object detection is proposed to effectively address the challenges associated with complex and dynamic underwater scenes.

(2) A novel AWFPN is proposed to facilitate the fusion of multi-scale feature semantics. An AWFE module is proposed to enhance underwater object detection by capturing relevant and discriminative information. Furthermore, we integrate a dedicated small object detection head specifically designed for detecting dense small objects.

(3) The proposed AWF-YOLO is evaluated through numerous experiments on the DUO dataset, consistently demonstrating significant performance improvements. These results establish its suitability for handling complex underwater scenarios effectively.

2. METHODOLOGY

In this paper, we propose a highly accurate framework for underwater object detection to effectively address the challenges posed by image blur and dense small objects. Firstly, we introduce an AWFPN, which empowers each feature map at different scales with strong semantic and spatial information. These improvements aim to enhance the accuracy and robustness of underwater object detection by addressing the unique challenges posed by underwater environments. Secondly, we propose an AWFE module that enables the network to dynamically learn the importance of each feature map and subsequently fuse them together. Lastly, we enhance the original YOLOv8 architecture by introducing a dedicated small detection head specifically designed for detecting dense small objects in underwater environments.

2.1. YOLOv8

YOLOv8 is an advanced real-time object detection algorithm based on deep learning. It employs a single-stage detection strategy, treating object detection as an end-to-end regression problem. The network follows a specific flow, where the input image undergoes feature extraction in the backbone network. Subsequently, the extracted features are fused in the neck region to capture low-level and high-level representations. Finally, the fused features are used for object regression prediction in the detection heads. This sequential process allows the network to extract and utilize features for accurate object detection. In our proposed method, we leverage the latest version of YOLOv8. To strike a balance between accuracy and lightweight design, we opt for the YOLOv8-s variant. The network architecture of YOLOv8 is illustrated in Figure 2.

The network architecture of YOLOv8 mainly consists of the backbone network, neck, and head. Its architecture components and features are as follows

Backbone: A lightweight CNN is employed as the backbone to extract image features. The backbone network commonly utilizes architectures similar to EfficientNet [23] or CSPDarknet53 [10]. These architectures consist of

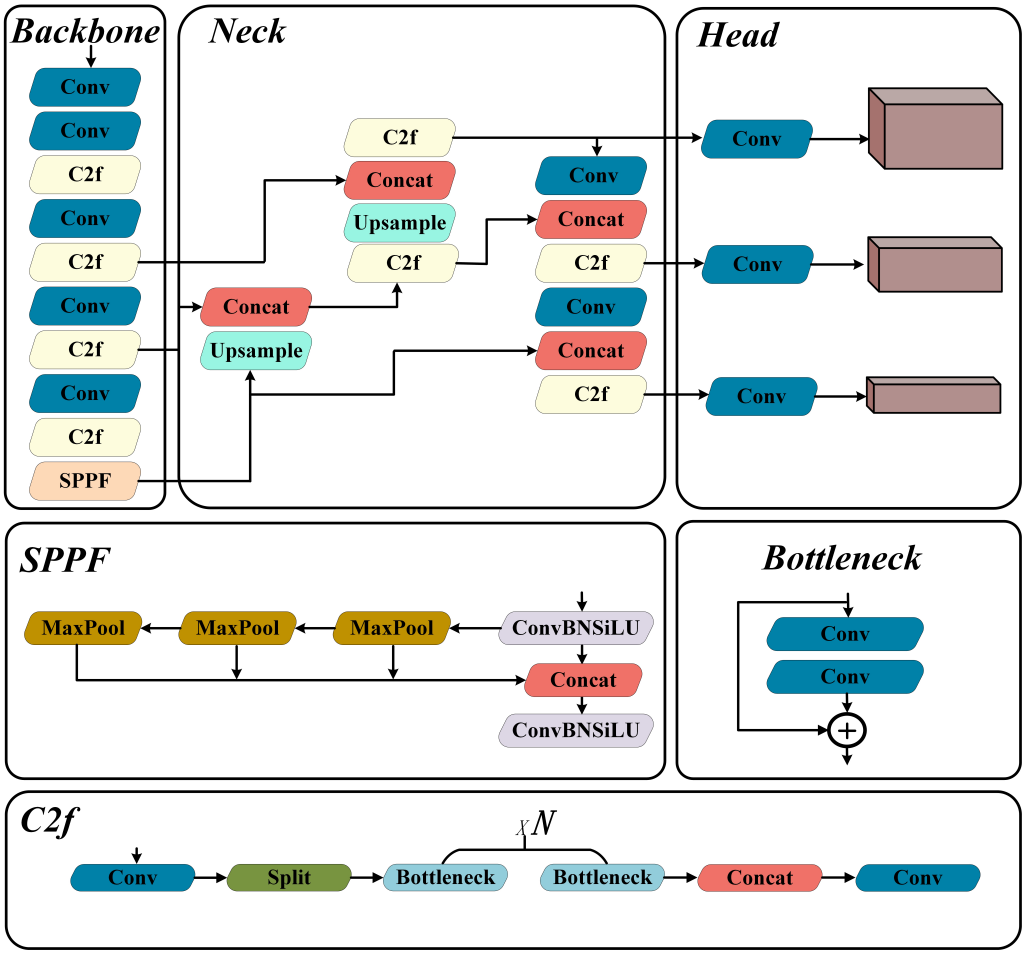

Figure 2. YOLOv8 architecture.

multiple convolutional layers and pooling layers that progressively extract and represent the semantic features of the image. This enables the network to capture informative and discriminative features essential for accurate object detection.

Neck: To enhance the detection performance, YOLOv8 introduces a feature fusion module known as the "Neck". This module plays a crucial role in merging feature maps from different layers, providing valuable multi-scale background information. Two commonly used "Neck" architectures are the Feature Pyramid Network (FPN)[24] and the Path Aggregation Network (PAN)[25]. These architectures achieve feature fusion through inter-layer connections and upsampling operations, effectively enhancing the feature information available for object detection.

Head: The detection head is a critical component of YOLOv8 responsible for object detection and localization on the feature map. It comprises a sequence of convolutional and fully-connected layers designed to extract information about object location, class, and confidence from the feature map. YOLOv8 employs multiple detection heads, each specialized in detecting objects at different scales. This multi-scale approach significantly improves the detection performance for both small objects and long-range objects. By incorporating these multiple heads, YOLOv8 enhances its ability to accurately detect objects of various sizes within the input image.

By employing effective feature extraction, multi-scale fusion, and precise object detection techniques, YOLOv8 demonstrates high detection accuracy while maintaining efficient performance for various real-time object

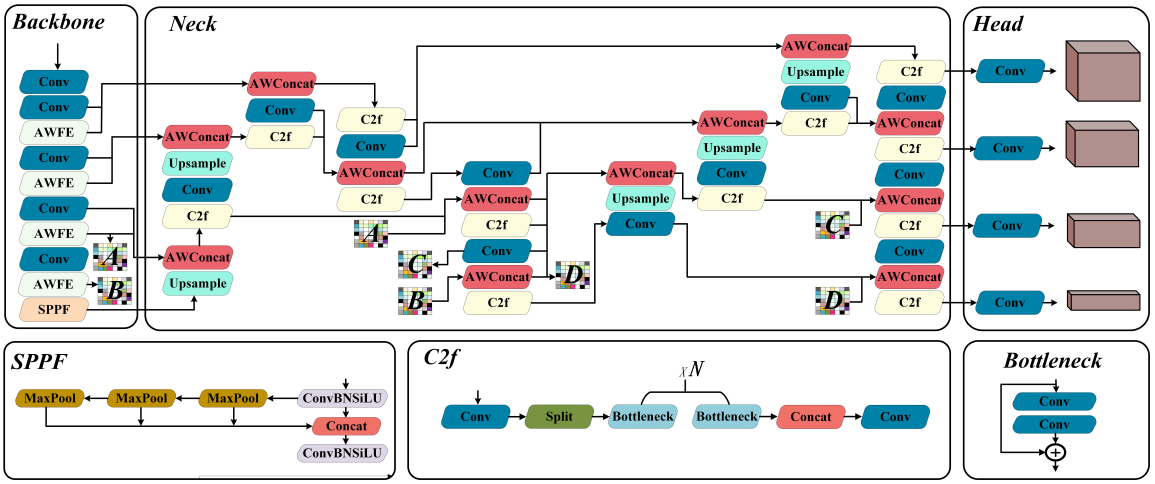

Figure 3. The proposed AWF-YOLO architecture.

detection tasks. However, considering the complexities of the underwater environment, further improvements are required to adapt the model for underwater object detection and enhance the accuracy of YOLOv8 in such scenarios. It is essential to address challenges specific to underwater imaging, such as image blur, color distortion, and the presence of dense small objects.

2.2. AWF-YOLO

We propose a highly accurate framework for underwater object detection to effectively address the challenges posed by image blur and dense small objects, depicted in Figure 3.

2.2.1. Adaptive weighted feature pyramids network

In our proposed method, we propose the AWFPN to handle the multi-scale feature fusion task, replacing the neck component of YOLOv8. The AWFPN significantly enhances the representational capacity of feature maps. It achieves this by establishing connections between feature maps at different resolutions through a combination of upsampling and downsampling operations. To provide further clarification, the AWFPN employs a hierarchical structure. It starts by taking the low-level feature maps from the backbone network, which have a relatively high resolution. These feature maps are then downsampled through bottom-up downsampling operations, resulting in feature maps with reduced resolution but higher-level semantic information. Simultaneously, the AWFPN performs top-down upsampling operations to gradually increase the resolution of feature maps. This process involves upsampling the lower-resolution feature maps and merging them with the higher-resolution feature maps from the backbone network. This fusion of multi-scale information enables the AWFPN to capture both fine-grained details and high-level contextual information, facilitating more accurate object detection.

To better utilize semantic information across various scales, we propose a dual PANet architecture that facilitates comprehensive feature aggregation between different layers through both top-down and bottom-up pathways. Our approach incorporates the adaptive weighted fusion concept from AWFE (see Section 2), enabling the differentiated fusion of different input features and learning the importance of each input feature. Unlike previous feature pyramid fusion methods, we believe that existing feature fusion modules have not fully exploited the semantic information available in the same-scale feature maps from the preceding layer. Consequently, the features from the previous layer remain underutilized. To address this limitation, we introduce a selective aggregation process within each feature aggregation module. We carefully choose a subset of same-scale feature maps from the previous layer for aggregation and apply adaptive weighted processing to enhance the representational information of the aggregated feature maps.

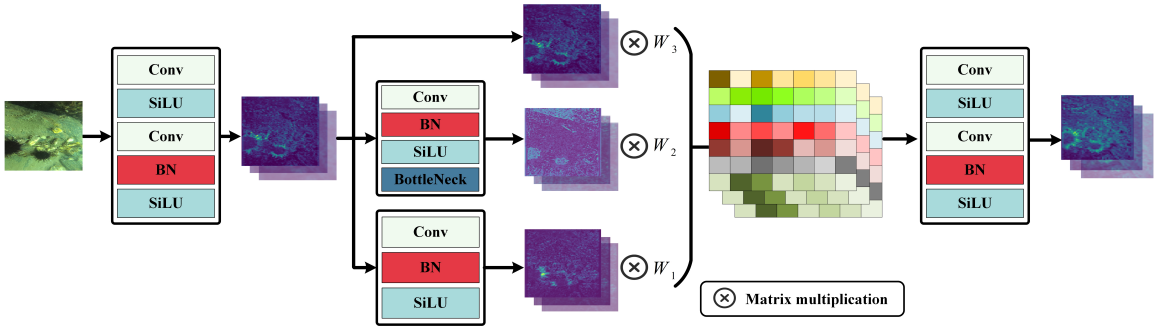

Figure 4. The detailed structure of the Adaptive Weighted Feature Extraction (AWFE) module.

Each feature aggregation module in AWFPN can be described as follows: Given the input feature maps F_{in}, we apply the adaptive weighting fusion process to the pre-layer feature maps F_{low} at the same scale as F_{in}. We calculate each F_{low} adaptive weight w_i by normalizing the weights with a small constant ϵ to avoid instability. The adaptive weights indicate the importance of each feature map in the fusion process. Assuming the feature maps of F_{low} and F_{in} have sizes of M rows and N columns, we have $(M, N) \in \{(20, 20), (40, 40), (80, 80), (160, 160), (320, 320)\}$. The adaptive weighted fusion is performed as follows:

$$F_{agg} = \left[\frac{w_1}{\epsilon + \sum_{j=1}^{n+1} w_j} \cdot F_{in}, \frac{w_2}{\epsilon + \sum_{j=1}^{n+1} w_j} \cdot F_{low}^1, \frac{w_3}{\epsilon + \sum_{j=1}^{n+1} w_j} \cdot F_{low}^2, \ldots, \frac{w_{n+1}}{\epsilon + \sum_{j=1}^{n+1} w_j} \cdot F_{low}^n \right], \tag{1}$$

where n represents the size of a subset of the same-scale feature maps in the previous layer. Finally, we obtain the aggregated feature map, denoted as F_{agg}. By incorporating an adaptive weighting fusion process into each feature aggregation module, the AWFPN effectively exploits the semantic information present in each layer of the feature map. This process aims to enhance the feature representation specifically for underwater object detection.

2.2.2. Adaptive weighted feature extraction module

We propose the AWFE module, which serves as a replacement for all C2f modules (as depicted in Figure 2) in the YOLOv8 backbone network. Due to the inconsistent nature of the output feature information from each branch, their contributions to the final output are unequal. To address this concern, we present the AWFE module, which aims to handle the output feature maps from each branch. This module enables the network to learn the significance of each feature map and subsequently merge them in the dimension. Additionally, we introduce an extra branch that directly connects to the output features of the other two branches using weighted connections. This approach promotes feature reuse and mitigates the loss of image details. Figure 4 provides a visual representation of the AWFE module.

We represent the calculation process of adaptive weighted fusion as follows: Let O denote the current output feature map, I_i denote the feature map of the i-th layer, n denote the number of feature maps, and w_i represent the learnable weights indicating the importance of the i-th feature map. To ensure stability, we introduce a small parameter ϵ set to 0.0001. The fusion operation between two feature maps X and Y, along the channel dimension, is denoted as $[X, Y]$. Therefore, the calculation can be summarized as follows:

Normalize the weights:

$$w_i = \frac{\exp(w_i)}{\sum_{j=1}^{n} \exp(w_j) + \epsilon}. \tag{2}$$

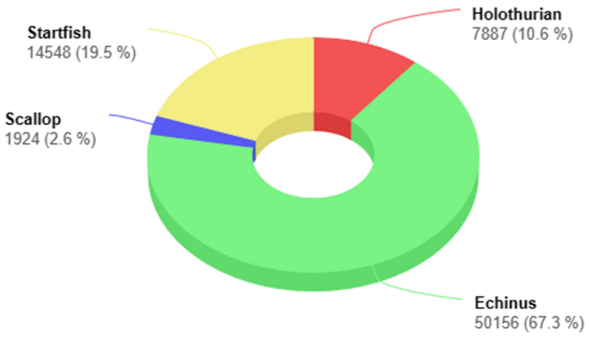

Figure 5. The distribution of objects in the DUO.

Compute the weighted feature maps:

$$I_i' = w_i \cdot I_i. \tag{3}$$

Perform fusion of the weighted feature maps:

$$O = [I_1', I_2', \ldots, I_n']. \tag{4}$$

Through the utilization of the proposed adaptive weighted fusion scheme, we successfully integrate information from multiple feature maps, taking into account their respective importance weights. This process enhances the overall representation and ultimately enhances the performance of underwater object detection.

For the AWFE module, the computation process of adaptive weighted fusion of feature maps of each branch can be expressed as:

$$F_{out} = [\frac{w_1}{\epsilon + \sum_{j=1}^{3} w_j} \cdot f_1(F_{in}), \frac{w_2}{\epsilon + \sum_{j=1}^{3} w_j} \cdot f_2(F_{in}), \frac{w_3}{\epsilon + \sum_{j=1}^{3} w_j} \cdot F_{in}], \tag{5}$$

where F_{out} represents the output features, F_{in} represents the input features, and f_1 and f_2 represent two non-linear transformations that efficiently extract features and enhance the non-linearity of the features.

2.2.3. A dedicated small object detection head

To tackle the challenge of detecting dense small objects underwater, we have extended the YOLOv8 architecture by incorporating a dedicated detection head with a scale size of 160×160. This supplementary detection head possesses lower semantic information but higher resolution when compared to other network layers. It is specifically tailored to address the detection of dense small objects in underwater environments.

Moreover, the inclusion of this high-resolution detection head in the feature fusion process enriches the semantic and spatial information within each feature layer. This integration successfully mitigates the adverse effects of low contrast often encountered in underwater images, thereby greatly enhancing the capability of the network to accurately detect small object features.

The incorporation of this dedicated detection head in the proposed AWF-YOLO represents a significant enhancement to tackle the specific challenges associated with underwater object detection, and it has yielded substantial improvements in the precision and robustness of detecting dense small objects in underwater scenes.

3. RESULTS

In this section, we conduct comparison experiments and ablation experiments to thoroughly assess the performance of our proposed AWF-YOLO. To evaluate accuracy and robustness, we utilize several evaluation metrics, including mAP@0.5:0.95. This metric calculates the average precision (AP) across 10 Intersection over Union (IoU) thresholds ranging from 0.5 to 0.95. Additionally, we report mAP@0.5 and mAP@0.75, which represent the AP at IoU thresholds of 0.5 and 0.75, respectively. These metrics provide a comprehensive evaluation of the performance in accurately and robustly detecting underwater objects.

3.1. Dataset

In our experiments, we select the DUO dataset, which is a publicly available annotated underwater object dataset introduced by Liu *et al.*[26]. The DUO dataset is created by gathering and re-annotating various underwater datasets from the URPC series. After filtering out highly similar images, we obtain a total of 7,782 images. Among these, 6,671 images are used for training, and the remaining 1,111 images are allocated for testing purposes. The dataset exhibits a retention rate of 95%, indicating that only a small number of similar images are retained in the new dataset.

The DUO dataset includes annotations for four types of objects: Holothurian, Echinus, Scallop, and Starfish. These annotations encompass a total of 74,515 objects, comprising 7,887 Holothurian instances, 50,156 Echinus instances, 1,924 Scallop instances, and 14,548 Starfish instances. Figure 5 illustrates the distribution of object counts for each category in the DUO dataset.

3.2. Experimental environment

To validate the effectiveness of the proposed network model, we conduct experiments using the following setup. The experiments are performed on a machine running Ubuntu 18.04 with an Intel i9-12900KF CPU and an NVIDIA GeForce RTX 3090 GPU with 24GB of VRAM. The system had 32GB of RAM. The experiment environment is set up with Python 3.8, and the deep learning framework is PyTorch 1.13.1. We initialize the learning rate to 0.001 and use a decay coefficient of 0.0005 for weight decay. The momentum parameter is set to 0.937. The batch size for training is set to 16. These experimental settings are chosen to ensure a reliable and efficient evaluation of the performance of the model.

3.3. Evaluation criteria

The classification of real and predicted scenarios is presented in Table 1. True Positive (TP) corresponds to cases where both the predicted and real scenarios are positive. False Positive (FP) refers to cases where the predicted scenario is positive, but the real scenario is negative. False Negative (FN) indicates cases where the predicted scenario is negative while the real scenario is positive. Lastly, True Negative (TN) represents cases where both the predicted and real scenarios are negative. The evaluation metrics employed in this paper encompass AP and mean AP (mAP). mAP serves as the primary performance measure for object detection algorithms and is widely used to evaluate the effectiveness of object detection models. A higher mAP value indicates superior detection performance on the given dataset. The precision-recall (P-R) curve is constructed by plotting precision on the vertical axis and recall on the horizontal axis. AP corresponds to the area under the P-R curve, while mAP is the average of the AP values across all categories. IoU quantifies the overlap between predicted and ground truth bounding boxes. Performance metrics follow the standard COCO[27] style metrics.mAP@0.5 signifies the mAP value computed with an IoU threshold of 0.5, whereas mAP@0.75 represents the mAP value with an IoU threshold of 0.75. mAP@0.5:0.95 denotes the average mAP value across different IoU thresholds ranging from 0.5 to 0.95, with a step size of 0.05.

According to the confusion matrix provided in Table 1, the metric for precision (P) can be defined as $\frac{TP}{TP+FP}$, and the metric for recall (R) can be defined as $\frac{TP}{TP+FN}$. To calculate the AP, we integrate the precision values $P(r)$ over the range of 0 to 1. Furthermore, the mAP can be represented as follows:

Table 1. Confusion matrix

		Predicted	
		Positive	**Negative**
Actual	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 2. Ablation analysis on the DUO dataset

Baseline(YOLOv5)	Small head	AWFE	AWFPN	mAP@0.5(%)	mAP@0.75(%)	mAP@0.5:0.95(%)
✓	✗	✗	✗	82.1	71.1	62.6
✓	✓	✗	✗	85.4	73.9	65.5
✓	✓	✓	✗	86.0	74.9	66.5
✓	✓	✓	✓	**86.2**	**75.0**	**67.0**

Table 3. Ablation analysis on the DUO dataset

Baseline(YOLOv8)	Small head	AWFE	AWFPN	mAP@0.5(%)	mAP@0.75(%)	mAP@0.5:0.95(%)
✓	✗	✗	✗	85.7	75.7	67.9
✓	✓	✗	✗	85.8	76.3	68.5
✓	✓	✓	✗	86.8	77.5	69.8
✓	✓	✓	✓	**87.2**	**77.8**	**70.1**

$$mAP = \frac{\sum_{i=1}^{n} AP_i}{n}, \tag{6}$$

where n represents the number of classes.

We have explicitly outlined the inclusion of time indicators, spatial metrics, and floating-point operations measurements, denoted as "Speed", "Parameters", and "GFLOPs", correspondingly. In this context, "Speed" signifies the inference speed of the model, quantified in milliseconds (ms). "Parameters" reflects the tally of model parameters, measured in units of individual parameters. Additionally, "GFLOPs" represents the count of billion floating-point operations executed per second (operations/s).

3.4. Ablation experiments on DUO

The proposed AWF-YOLO model presents substantial improvements over the original YOLOv8 model by integrating advancements in feature extraction, feature fusion, and detection heads. To evaluate the influence of different enhanced modules and their combinations on model performance, we conducted a series of comprehensive ablation experiments. Furthermore, to verify the effectiveness and general applicability of our proposed method, we performed experiments using YOLOv5 and YOLOv8 as baselines. The outcomes of these experiments are summarized in Table 2 and Table 3, respectively.

The results in Table 2 demonstrate the substantial improvements in performance metrics for YOLOv5 upon integrating the small detection head. Notably, the model achieved increases of 3.3%, 2.8%, and 2.9% in mAP@0.5, mAP@0.75, and mAP@0.5:0.95, respectively. Moreover, the introduction of the AWFE architecture in the backbone network further elevated the model performance, with gains of 0.6%, 1.0%, and 1.0% in mAP@0.5, mAP@0.75, and mAP@0.5:0.95, respectively. Additionally, modifying the neck network to AWFPN led to improvements of 0.2%, 0.1%, and 0.5% in mAP@0.5, mAP@0.75, and mAP@0.5:0.95, respectively.

Based on YOLOv8, Table 3 reveals significant improvements in performance metrics after incorporating the

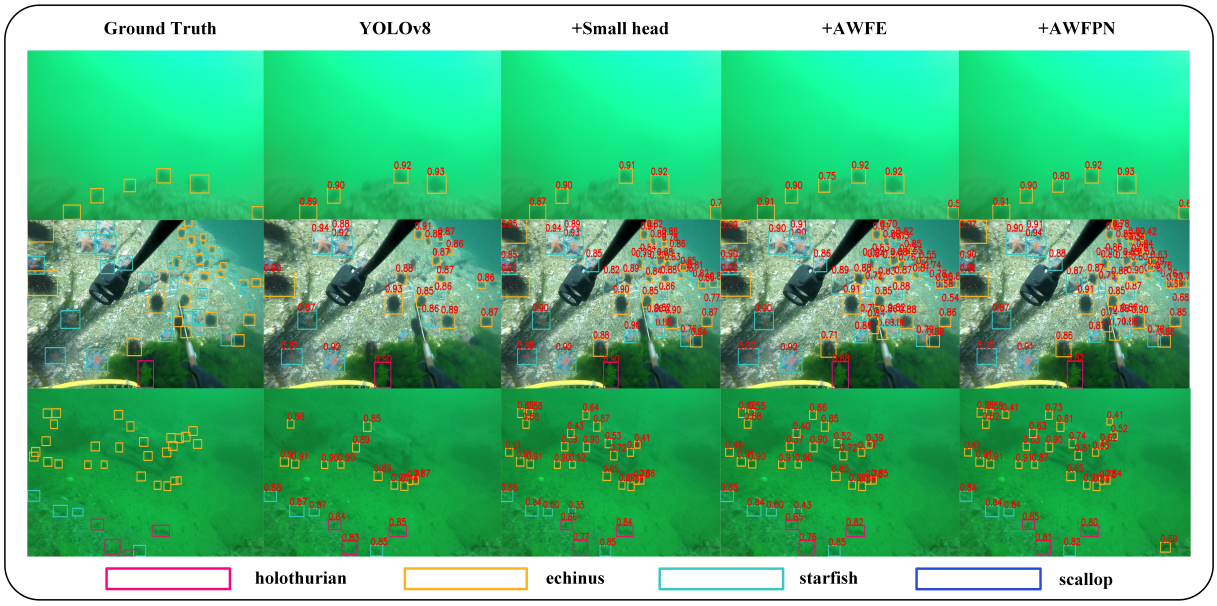

Figure 6. Visualization of ablation effect.

small detection head, with increases of 0.1%, 0.6%, and 0.6% in mAP@0.5, mAP@0.75, and mAP@0.5:0.95, respectively. Introducing the AWFE architecture in the backbone network further enhances the model performance, resulting in improvements of 1.0%, 1.2%, and 1.3% in mAP@0.5, mAP@0.75, and mAP@0.5:0.95, respectively. Moreover, modifying the neck network to AWFPN leads to performance improvements of 0.4%, 0.3%, and 0.3% in mAP@0.5, mAP@0.75, and mAP@0.5:0.95, respectively.

From Table 2 and Table 3, it is evident that the incorporation of the small detection head, AWFE architecture, and AWFPN architecture in both YOLOv5 and YOLOv8 frameworks leads to significant improvements in performance metrics. These results demonstrate the effectiveness of each module in our approach.

To provide a more intuitive demonstration of the significance of each module, we conducted tests in three distinct scenarios: low-contrast scenes, dense small object scenes, and low-contrast dense small object scenes. As our baseline, we employed YOLOv8 and subsequently incrementally integrated modules while applying visual processing. The experimental findings are showcased in Figure 6.

In low-contrast scenes, the addition of the AWFE and AWFPN modules significantly improves the model's detection performance, successfully identifying all objects with indistinct features. In dense small object scenes, the inclusion of the small detection head notably enhances the model's ability to detect small objects, successfully identifying a large number of small targets. In low-contrast dense small object scenes, as we progressively incorporate the small detection head, AWFE module, and AWFPN module, the model's detection capabilities improve significantly. These results demonstrate the effectiveness of each module in our approach.

3.5. Quantitative evaluation

Table 4 displays the comparative experimental results on the DUO dataset, which comprises complex underwater scenes and objects of various scales, posing significant challenges to object detection algorithms. As depicted in Table 3, our method exhibits the most remarkable performance in terms of mAP@0.5, mAP@0.75, and mAP@0.5:0.95, outperforming all other single-stage and two-stage object detection models. To be specific, in mAP@0.5, mAP@0.75, and mAP@0.5:0.95, our method surpasses Faster-RCNN[28] by 10.4%, 14.7%, and 13.9%, respectively. In the same metrics, it also outperforms Cascade RCNN[29] by 7.1%, 9.3%, and 10%.

Table 4. Comparison of object detection networks for underwater object detection

	Method	mAP@0.5(%)	mAP@0.75(%)	mAP@0.5:0.95(%)
Multi-stage	Fast RCNN [32]	75.6	62.3	55.7
	Faster RCNN [28]	76.8	63.1	56.2
	Grid RCNN [33]	74.5	64	56.3
	Cascade RCNN [29]	80.1	68.5	60.1
One-stage	SSD [34]	70.3	57.2	49.5
	RetinaNet [35]	71.2	58.4	51.7
	FCOS [36]	77.5	60.3	53.0
	YOLOv3 [8]	83.6	69.7	62.1
	YOLOv5 [37]	82.1	71.1	62.6
	YOLOX [38]	82.4	69.0	62.8
	YOLOv6 [30]	84.8	73.5	65.6
	YOLOv7 [31]	85.4	74.6	66.8
	RT-DETR [39]	77.6	55.2	49.0
	YOLOv8 [40]	85.7	75.7	67.9
	AWF-YOLO(Ours)	**87.2**	**77.8**	**70.1**

Table 5. Comparing across spatial and temporal dimensions

Method	Speed	parameters	GFLOPs
YOLOv5	45.1	**9,111,923**	**23.8**
YOLOv6	39.3	16,297,619	44.0
YOLOv8	**38.4**	11,127,132	28.4
AWF-YOLO(Ours)	53.4	16,566,421	55.5

Additionally, at the identical performance metrics, our method demonstrates superior performance compared to YOLOv6 [30] by 2.4%, 4.3%, and 4.5%, and YOLOv7 [31] by 1.8%, 3.2%, and 3.3%. These comparative results further validate the effectiveness of our AWF-YOLO in underwater environments.

We have conducted a comprehensive comparative assessment, contrasting our proposed approach with YOLOv5, YOLOv6, and YOLOv8. The summarized comparative outcomes are presented in Table 5. The table distinctly indicates that compared to the benchmark techniques, our method achieves significant accuracy improvements while incurring some trade-offs in terms of speed and space. Given the notable accuracy enhancement and the absence of substantial differences in speed and space, we consider this trade-off acceptable. Nonetheless, we fully acknowledge the potential for further improvement in these dimensions, and addressing this challenge remains a significant objective for our future efforts.

3.6. Perceptual comparisons

Our proposed AWF-YOLO effectively solves the problem of low detection accuracy caused by dense small objects and underwater scenes with low contrast. To visually demonstrate the effectiveness of our method, we prepare image sets representing three different scenes: dense small objects, low contrast, and dense small objects at low contrast. For each scene, we choose three representative images. We apply the YOLOv8 method and our AWF-YOLO method to detect the objects in each scene. The results are shown in Figure 7, 8, and 9.

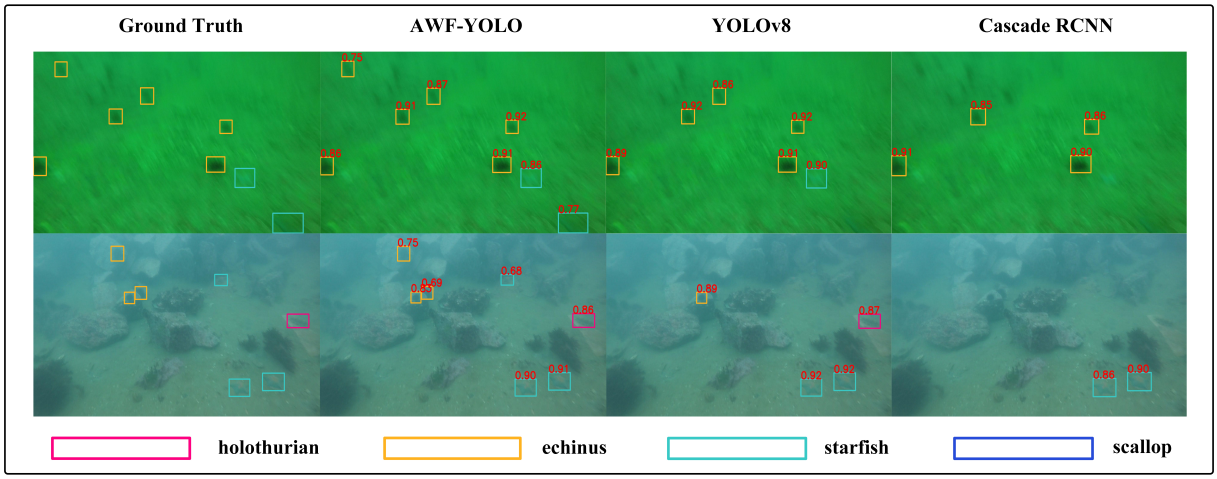

Figure 7. Qualitative comparison results on the low contrast scenario.

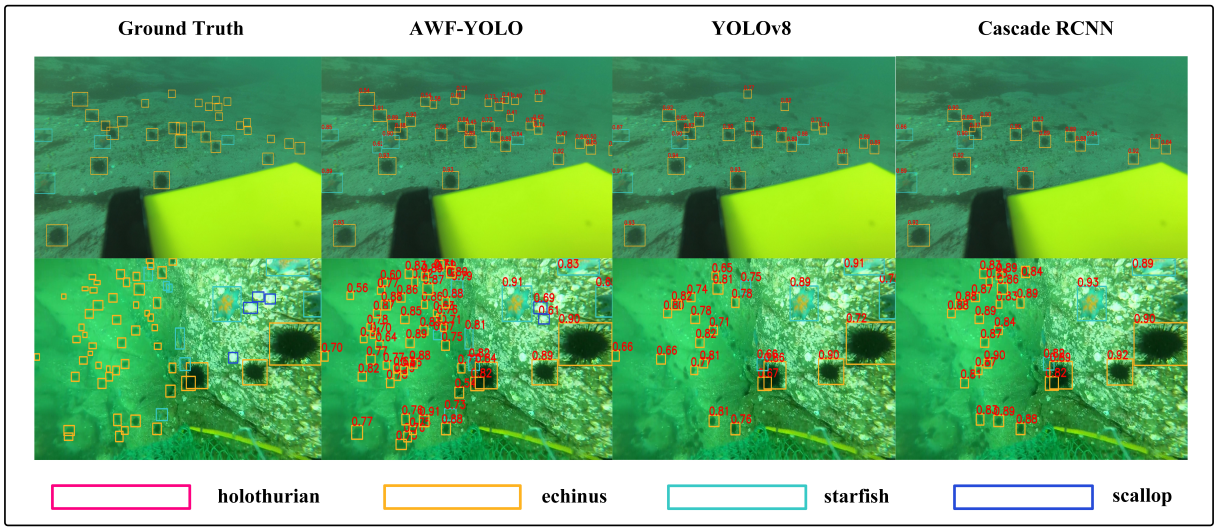

Figure 8. Qualitative comparison results on the dense small objects scenario.

In Figure 7, we showcase the detection outcomes of ground truth bounding boxes, YOLOv8, Cascade RCNN, and AWF-YOLO. It is evident from the illustration that YOLOv8 and Cascade RCNN fail to detect certain sea cucumbers, starfish, and sea urchins. Particularly in underwater blurry environments, their performance exhibits poor noise resilience. In comparison, AWF-YOLO demonstrates superior recall and accuracy in detecting low-contrast underwater objects, thereby exhibiting heightened noise resilience.

In Figure 8, we present the detection results of ground truth bounding boxes, YOLOv8, Cascade RCNN, and AWF-YOLO. It is evident that both YOLOv8 and Cascade RCNN fail to detect numerous sea urchins, particularly in densely populated areas. The accuracy and recall rates of YOLOv8 and Cascade RCNN perform poorly. In contrast, AWF-YOLO demonstrates higher recall and precision in detecting densely packed small

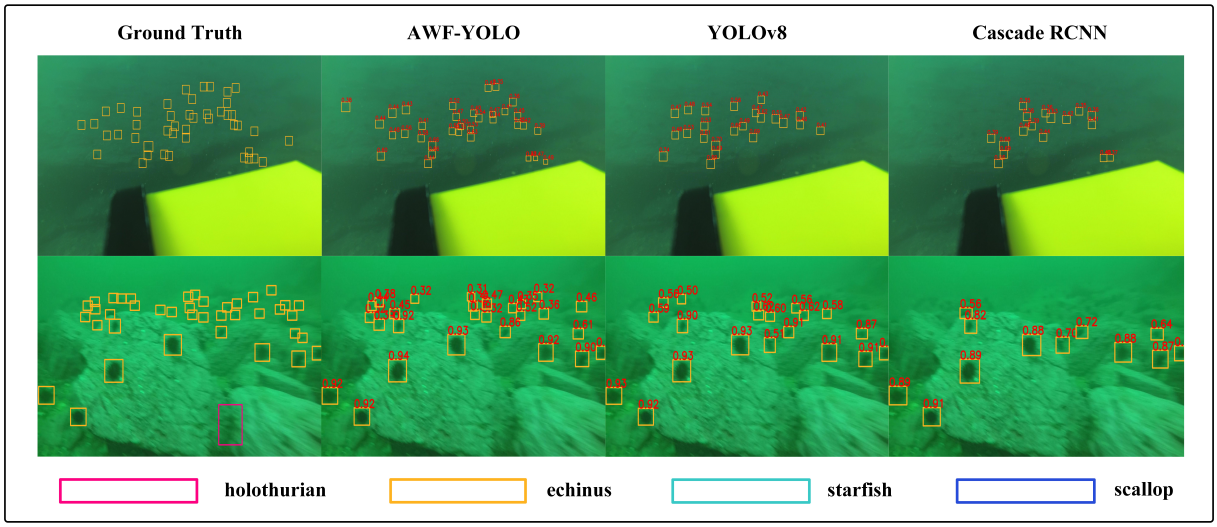

Figure 9. Qualitative comparison results on the dense small objects under low contrast scenario.

underwater objects when compared to YOLOv8 and Cascade RCNN.

In Figure 9, we present visual detection results of real bounding boxes, YOLOv8, Cascade RCNN, and AWF-YOLO. It is evident that both Cascade RCNN and YOLOv8 failed to detect numerous densely packed small sea urchins. In comparison, AWF-YOLO exhibits higher recall and accuracy in detecting dense small objects in low-contrast underwater scenarios when compared to them.

4. CONCLUSIONS

In this study, we present AWF-YOLO, an advanced underwater object detection algorithm meticulously designed to address the intricate challenges posed by dynamic and complex underwater environments. Specifically, it effectively handles issues such as image blurring due to low contrast and reduced detection accuracy caused by the presence of closely-packed small objects. Our approach revolves around the novel AWFPN. This architectural innovation adeptly integrates multi-scale feature semantics by capturing both pertinent and distinctive information, thereby augmenting underwater object detection capabilities. To further elevate the performance of underwater object detection, we introduce an AWFE Module. This module is dedicated to extracting task-relevant and distinctive features, thereby elevating the overall detection accuracy. Furthermore, we seamlessly integrate a specialized small object detection head into the framework. This component is meticulously tailored for the precise detection of densely-clustered small objects, which often pose challenges in underwater scenarios. Rigorous qualitative and quantitative experiments conducted on the DUO dataset affirm the superiority of our proposed AWF-YOLO over existing methodologies, showcasing remarkable performance. AWF-YOLO stands as a powerful tool for advancing efficient underwater exploration. It significantly bolsters the efficiency of underwater intelligence gathering, enabling devices to navigate and comprehend underwater ecological resources more effectively. Our work makes valuable contributions by enhancing the efficacy of underwater object detection tasks, and its insights extend to detection tasks across diverse fields. For instance, our methodology has the potential to extend to target detection from the vantage point of unmanned aerial vehicles, which frequently encounter challenges such as detecting small-sized objects with unclear features. However, it is important to acknowledge the hardware limitations of certain underwater intelligence gathering devices, which impose stringent constraints on model parameter counts. Therefore, the development of lightweight underwater object detection algorithms has become imperative. Tradition-

ally, reducing model parameters might result in a discernible accuracy drop. The incorporation of multiple fusion structures in our model has led to the introduction of a notable number of parameters. Hence, a pivotal research trajectory in underwater object detection revolves around investigating approaches that achieve lightweight models through parameter reduction without compromising accuracy. This avenue holds substantial promise for the future of underwater object detection, allowing for the deployment of efficient algorithms that strike a delicate balance between model size and detection performance.

DECLARATIONS

Authors' contributions

Conception and design of the study: Jiang Y, Qin H
Methodology, experimentation, and writing- original draft: Guo Q, Qi H
Data validation and analysis: Wang Y, Zhang Y

Availability of data and materials

Not applicable.

Financial support and sponsorship

This work was supported by the National Natural Science Foundation of China under Grant 62072211, Grant 51939003, and Grant U20A20285.

Conflicts of interest

All authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

REFERENCES

1. Sun K, Cui W, Chen C. Review of underwater sensing technologies and applications. *Sensors* 2021;21:7849. DOI
2. Jiang Y, Zhao M, Zhao W, et al. Prediction of sea temperature using temporal convolutional network and lstm-gru network. *Complex Eng Syst* 2021;1:6. DOI
3. Maamoun KSA, Karimi HR. Reinforcement learning-based control for offshore crane load-landing operations. *Complex Eng Syst* 2022;2:13. DOI
4. Er MJ, Chen J, Zhang Y, Gao W. Research challenges, recent advances, and popular datasets in deep learning-based underwater marine object detection: a review. *Sensors* 2023;23:1990. DOI
5. Lecun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 1998;86:2278-324. DOI
6. Redmon J, Divvala S, Girshick R, Farhadi A. You only look once: unified, real-time object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016;779-88. DOI
7. Al Muksit A, Hasan F, Emon MFHB, Haque MR, Anwary AR, Shatabda S. Yolo-fish: a robust fish detection model to detect fish in realistic underwater environment. *Ecol Inform* 2022;72:101847. DOI
8. Redmon J, Farhadi A. Yolov3: an incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. DOI
9. Zhang M, Xu S, Song W, He Q, Wei Q. Lightweight underwater object detection based on YOLO v4 and multi-scale attentional feature fusion. *Remote Sens* 2021;13:4706. DOI
10. Bochkovskiy A, Wang CY, Liao HYM. YOLOv4: optimal speed and accuracy of object detection. *arXiv preprint arXiv:2004.10934* 2020. DOI

11. Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC. MobileNetV2: inverted residuals and linear bottlenecks. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018. p. 4510-20. DOI
12. Zhao S, Zheng J, Sun S, Zhang L. An improved YOLO algorithm for fast and accurate underwater object detection. *Symmetry* 2022;14:1669. DOI
13. Peng F, Miao Z, Li F, Li Z. S-FPN: a shortcut feature pyramid network for sea cucumber detection in underwater images. *Expert Syst Appl* 2021;182:115306. DOI
14. Shang Y. Resilient multiscale coordination control against adversarial nodes. *Energies* 2018;11:1844. DOI
15. Dai Y, Gieseke F, Oehmcke S, Wu Y, Barnard K. Attentional feature fusion. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021. p. 3560-9. DOI
16. Liu K, Peng L, Tang S. Underwater object detection using TC-YOLO with attention mechanisms. *Sensors* 2023;23:2567. DOI
17. Lei F, Tang F, Li S. Underwater target detection algorithm based on improved YOLOv5. *J Mar Sci Eng* 2022;10:310. DOI
18. Li B, Liu B, Li S, Liu H. Underwater target detection based on improved YOLOv4. In: 2022 41st Chinese Control Conference (CCC), 2022 Jul 25-27; Hefei, China. 2022. p. 7012-7. DOI
19. Ge H, Dai Y, Zhu Z, Zang X. Single-stage underwater target detection based on feature anchor frame double optimization network. *Sensors* 2022;22:7875. DOI
20. Qi J, Gong Z, Xue W, Liu X, Yao A, Zhong P. An unmixing-based network for underwater target detection from hyperspectral imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 2021;14:5470-87. DOI
21. Li C, Guo C, Ren W, et al. An underwater image enhancement benchmark dataset and beyond. *IEEE Trans Image Process* 2020;29:4376-89. DOI
22. Zhuang P, Wu J, Porikli F, Li C. Underwater image enhancement with hyper-laplacian reflectance priors. *IEEE Trans Image Process* 2022;31:5442-55. DOI
23. Tan M, Le QV. Efficientnet: rethinking model scaling for convolutional neural networks. In: International Conference on Machine Learning, PMLR, 2019, p. 6105-14. DOI
24. Lin TY, Dollar 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, 2017. p. 2117-25. DOI
25. 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, 2018. p. 8759-68. DOI
26. Liu C, Li H, Wang S, et al. A dataset and benchmark of underwater object detection for robot picking. In: 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shenzhen, China. 2021. p. 1-6. DOI
27. Lin TY, Maire M, Belongie S, et al. Microsoft COCO: common objects in context. In: Fleet D, Pajdla T, Schiele B, Tuytelaars T, editors, Computer Vision - ECCV 2014, Springer, Cham; 2014. p. 740-55. DOI
28. Ren S, He K, Girshick R, Sun J. Faster R-CNN: towards real-time object detection with region proposal networks. *Adv Neur Inf Process Sys* 2015;28. DOI
29. Cai Z, Vasconcelos N. Cascade R-CNN: delving into high quality object detection. In *CVPR*, 2018;6154-62. DOI
30. Li C, Li L, Jiang H, et al. YOLOv6: a single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*, 2022. DOI
31. Wang CY, Bochkovskiy A, Liao HYM. 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 (CVPR), 2023. p. 7464-75. DOI
32. Girshick R. Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV), 2015. p. 1440-8. DOI
33. Lu X, Li B, Yue Y, Li Q, Yan J. Grid R-CNN. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019. p. 7363-72. DOI
34. Liu W, Anguelov D, Erhan D, et al. SSD: single shot multibox detector. In: Leibe B, Matas J, Sebe N, Welling M, editors. Computer Vision - ECCV 2016. Springer, Cham; 2016. p. 21-37. DOI
35. Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 2017. p. 2980-8. DOI
36. Tian Z, Shen C, Chen H, He T. FCOS: fully convolutional one-stage object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019. p. 9627-36. DOI
37. Jocher G. YOLOv5 by Ultralytics, May 2020. https://github.com/ultralytics/yolov5[Last accessed on 10 Sep 2023].
38. Ge Z, Liu S, Wang F, Li Z, Sun J. YOLOX: exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. DOI
39. Lv W, Zhao Y, Xu S, et al. DETRs beat YOLOs on real-time object detection. *arXiv preprint arXiv:2304.08069*, 2023. DOI
40. Jocher G, Chaurasia A, Qiu J. YOLO by Ultralytics, January 2023. https://github.com/ultralytics/ultralytics [Last accessed on 10 Sep 2023].