

SSDT-YOLO: An Underwater Object Detection Framework for Efficient Small-Target Recognition in Complex Environments

Chaoyang Chen

ouzk@163.com

Hunan University of Science and Technology

Gangzhi Hu

Hunan University of Science and Technology

Zunhui Yi

Hunan University of Science and Technology

Lei He

Hunan University of Science and Technology

Dan Yang

Hunan University of Science and Technology

Research Article

Keywords: Benthic marine products, Complex underwater environment, Object detection, Deep learning

Posted Date: October 15th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-8271380/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

SSDT-YOLO: An Underwater Object Detection Framework for Efficient Small-Target Recognition in Complex Environments

Chaoyang Chen^{1,2*}, Gangzhi Hu^{2,3†}, Zunhui Yi^{1,2}, Lei He^{1,2},
Dan Yang^{1,2}

^{1*}School of Information and Electrical Engineering, Hunan University of Science and Technology, Xiangtan, 411201, Hunan, China.

²Sanya Institute of Hunan University of Science and Technology, Hunan University of Science and Technology, Sanya, 572024, Hainan, China.

*Corresponding author(s). E-mail(s): ouzk@163.com;

Contributing authors: hu.gangzhi@mail.hnust.edu.cn;
yizunhui@hnust.edu.cn; heleixb@hnust.edu.cn; dan.yang@hnust.edu.cn;

[†]These authors contributed equally to this work.

Abstract

Underwater object recognition represents a fundamental requirement for marine environmental monitoring and benthic ecosystem investigation. The complex underwater environment poses significant challenges to existing marine detection models. Current models commonly suffer from inaccurate and imprecise recognition. In low-light underwater conditions, their accuracy deteriorates drastically, and dense underwater scenes further exacerbate this issue. Therefore, addressing these problems has become an urgent necessity. To tackle these issues, this research introduces SSDT-YOLO, a high performance underwater object detection model built upon an optimized YOLOv8 framework, specifically optimized for benthic organism identification. The presented approach incorporates four major enhancements into the foundational YOLOv8 architecture: First, a TEfficientNet module is integrated into the backbone network to address chromatic interference caused by non-uniform underwater illumination, enhances feature extraction capability. Second, Spatial-Preserving Depth Convolution (SPDConv) replaces standard convolution layers to enhance fine-grained feature retention for small targets. Third, a C2f-Dual module is developed in the neck network to reduce computational complexity while improving feature fusion efficiency. Fourth, a Spatial and Channel Synergistic Attention (SCSA) mechanism is incorporated to strengthen discriminative capability for densely distributed microscopic targets. Experimental validation on the DUO dataset (7,782 images) and URPC2019 benchmark (4,707 images) demonstrates superior performance, achieving mAP scores of 84.9% and 83.0%, respectively, surpassing the baseline YOLOv8 by 2.0% and 2.5%. The proposed model achieves a 35.5% parameter reduction (2.0M vs. 3.1M) while maintaining real-time processing capability. Comparative analyses confirm the detector's advantages regarding precision, model compactness, and computational efficiency compared to cutting-edge approaches.

Keywords: Benthic marine products, Complex underwater environment, Object detection, Deep learning

1 Introduction

Underwater object detection technology plays a vital role in identifying and tracking targets of interest within challenging aquatic environments, where visibility is restricted and imaging conditions are complex. Its applications encompass, but are not limited to, underwater robotics, environmental monitoring, three-dimensional (3D) reconstruction, and map generation[1]. This method has been widely utilized in studies of aquatic environments, navigation systems for subsea applications, and biomass quantification in marine species [2, 3]. It also serves critical functions in offshore fishery management and the identification of submerged explosive devices, contributing to enhanced resource exploration and sustainability efforts in marine domains[4]. Presently, marine activities predominantly depend on manual intervention, with labor-intensive processes such as aquatic harvesting and seabed organism collection posing significant occupational hazards to personnel. These challenges not only escalate operational expenses but also introduce ethical dilemmas regarding human safety. Thus, the development and deployment of autonomous sensing systems are imperative to address these limitations[5, 6]. Automated systems and advanced sensing methods offer substantial opportunities to streamline marine workflows, accelerate scientific investigations, and reduce human operational dependencies, fostering innovation in marine research and resource utilization.

Computer vision represents another milestone in the evolution of computing technology[7]. By advancing computer vision techniques, we can use camera sensors to capture light information from the real world, convert it into matrices stored in computers, and then enable machines to perceive objects in a manner similar to human vision through a series of image processing, computational programming, and matrix transformation operations. This capability allows computers to detect and classify objects, thereby assisting humans in making more informed decisions [8, 9]. Computer vision has been widely adopted across various domains: it is used in industrial inspection to detect cracks and defects, in psychology to recognize human emotions, in medicine to efficiently identify abnormal tissues, in agriculture to detect pests, in autonomous vehicles to recognize surrounding obstacles, and in remote sensing to investigate ground-based facilities. All these examples collectively demonstrate that computer vision is profoundly transforming modern human life[10–15].

Deep learning (DL), on the other hand, refers to the use of mathematical functional relationships to simulate the neural network structure of the human brain. It enables artificial neural networks to automatically learn patterns and features from massive datasets, thereby eliminating the reliance on manually engineered features[16]. Traditional computer vision methods rely heavily on purely mathematical computations for underwater image preprocessing, object recognition, and tracking. Although these conventional algorithms are computationally efficient, they often face difficulties in underwater environments, which limits their robustness and generalization capability. By training models on large-scale datasets encompassing diverse lighting conditions, noise, and underwater distortions, deep learning approaches achieve superior robustness in underwater computer vision tasks. These models can adapt to the complexities of underwater environments—such as dynamic illumination variations and diverse noise patterns—exhibiting remarkable performance in object recognition and tracking. This adaptability makes deep learning the preferred solution for modern underwater computer vision applications[17].

Contemporary object detection architectures utilizing deep learning are systematically divided into two primary methodologies: single-phase and dual-phase frameworks. Single-stage systems execute recognition through unified end-to-end processing, whereas two-stage frameworks implement hierarchical refinement strategies. Representative two-stage detectors include RCNN[18], SPPNet[19], Fast RCNN[20], Faster RCNN[21], EfficientDet[22], and Feature Pyramid Networks (FPN)[23]. In contrast, single-stage detectors encompass the You Only Look Once (YOLO) series[24], Single Shot MultiBox Detector (SSD)[25], RetinaNet[26], CornerNet[27], CenterNet[28], DETR[29], and RT-DETR[30]. Compared to two-stage detectors, single-stage detectors require lower computational overhead and exhibit faster inference speeds, making them well-suited for real-time applications. However, two-stage detectors, which involve additional candidate box refinement steps, typically incur higher computational costs and slower processing speeds.

Deep learning techniques, especially Convolutional Neural Networks (CNNs) such as those in the YOLO architecture, have demonstrated notable potential for underwater target recognition, owing to their robust adaptability and automated feature learning mechanisms. This methodology overcomes constraints inherent in conventional machine vision approaches, facilitating efficient identification of seabed-dwelling species in challenging subaquatic conditions.

The latter methodology is grounded in deep learning, specifically CNN. These networks constitute a robust category of differentiated architectures that learn autonomously from raw data inputs, eliminating dependence on manual feature engineering[31]. Consequently, CNNs have advanced traditional Artificial Neural Network (ANN) frameworks, including regularized Multi-Layer Perceptron (MLP) systems. Individual CNN layers are engineered to refine salient feature representations while pruning model intricacy via meticulously calibrated parameters. Furthermore, CNNs integrate dropout methodologies[32] to counteract overfitting—a prevalent challenge in classical network designs. Architected for interpreting multidimensional structures, CNNs demonstrate broad utility across domains such as visual recognition, biomedical imaging, semantic segmentation, and linguistic processing[31, 33]. Their capacity for self-directed derivation of critical representations from data confers superior functionality compared to conventional networks.

To ensure sufficient real-time detection performance, we chose the YOLOv8 framework, a CNN-based object detection network. Although YOLOv8 demonstrates outstanding performance in general scenarios, it still faces challenges in specialized and complex underwater applications, such as dense object detection and small target recognition. In order to address the aforementioned challenges, this study undertakes a systematic investigation and architectural refinement of the YOLOv8 framework to improve its suitability for the unique demands of subaquatic target recognition. While YOLOv8 improves multi-scale feature fusion and context-aware feature aggregation via its C2f module, the architecture exhibits limitations in reliably identifying small-scale and clustered targets, especially within visually complex marine settings. Additionally, while the backbone network supports hierarchical feature integration, it incurs high computational complexity and suboptimal inference speeds due to parametric redundancies. To address these limitations, we propose SSDT-YOLO, an enhanced framework tailored for subaquatic detection tasks. The primary innovations and contributions of this work are summarized as follows:

(1) This paper presents a novel underwater object detector, referred to as SSDT-YOLO, based on the YOLOv8 architecture, designed for the detection and classification of marine benthic organisms. The proposed method has been extensively evaluated on standardized benchmark datasets, demonstrating superior performance. Leveraging the NVIDIA GeForce RTX 3090, SSDT-YOLO achieves an inference speed of 214.8 FPS and a compact model size of 2.0 MB, demonstrating a promising improvement for underwater object detection applications.

(2) A novel feature extraction backbone, TEfficientNet, is proposed to address the unique challenges of underwater environments. TEfficientNet adopts a top-down hierarchical architecture in which MBConv and Fused-MBConv layers are employed sequentially to extract features. During training, the input image resolution is progressively increased, accompanied by adaptive adjustment of the regularization parameters. By replacing the backbone network with this structure for feature extraction, the model demonstrates faster convergence and improved representation capabilities, which may contribute to enhanced training efficiency and detection performance in complex underwater scenes.

(3) The Spatial-Preserving Depth Convolution (SPDConv) module is adopted as a replacement for traditional convolutional layers in neck, strengthening the model's ability to extract small-object features while improving computational efficiency. The spatial and channel collaborative attention mechanism (SCSA) is introduced at the end of the backbone network to enhance the model's attention capability and support multi-scale feature extraction. Furthermore, by leveraging sparse feature separation in the head channel to aggregate features, the efficiency and effectiveness of the self-attention mechanism are potentially improved.

(4) To mitigate the high memory and computational resource demands posed by the deployment of CNN on edge devices in underwater environments, an improved C2f-Dual module is introduced at the neck for feature fusion. The C2f-Dual module utilizes both 1×1 and 3×3 convolutional kernels in a progressive manner to extract hierarchical features from the input feature maps. This design helps reduce

the model’s computational cost and parameter count, while contributing to faster processing efficiency.

2 RELATED WORK

Underwater object detection serves as a fundamental technology for marine environment monitoring and ecological conservation. However, it continues to encounter persistent challenges such as light absorption and scattering, motion blur, densely distributed small objects, and limited annotated data[34]. Conventional techniques, which often depend on sonar imaging or manual visual interpretation, tend to be inefficient and lack general adaptability[35]. With the rapid advancement of deep learning, end-to-end feature learning frameworks have markedly enhanced the detection accuracy and real-time capability of underwater vision systems. Among them, one-stage detectors—particularly the YOLO family—have emerged as mainstream solutions due to their favorable trade-off between detection precision and computational efficiency. Nevertheless, the inherent complexity of underwater scenes still necessitates further optimization of these models, especially in aspects such as lightweight architecture design, feature representation enhancement, and adaptive perception mechanisms[36].

Multiple research efforts have focused on evaluating and optimizing these computational frameworks for marine applications. Lyu et al.[37] proposed an underwater object detection algorithm based on YOLOX to address the challenges posed by dense and blurred objects. Their approach demonstrated unique effectiveness in marine organism identification, significantly improving the classification and localization capabilities of the model. However, experimental results indicate that the model exhibits insufficient detection capability for medium-sized targets. Xuan et al.[38] focused on small-object recognition in complex environments, using sea cucumbers as the research target. They constructed a dedicated marine sea cucumber dataset to overcome the lack of available data, achieving an impressive accuracy of 95.47% on their dataset. While their method holds potential for small-object detection in challenging environments, its applicability remains constrained by the limited diversity of target categories, necessitating further performance improvements. To tackle marine pollution issues, Tian et al.[39] developed a modified YOLOv4-based marine debris detection algorithm, which was tested on an autonomous underwater garbage collection robot. The algorithm achieved a processing speed of 66.67 FPS, offering heuristic insights for researchers in related fields. Recognizing the critical need for accurate detection and quantification of fish behavior in aquaculture, Li et al. [40] leveraged YOLOv9, replacing conventional convolutional modules with a multi-scale feature learning detection head and refining the loss function to enhance the identification and quantification of anomalous fish behaviors. However, due to the high computational demands of YOLOv9, its practical application in fishery settings remains debatable. Zhuo et al.[41] addressed the complexity of underwater treasure detection by incorporating an improved attention mechanism and additional convolutional modules into a YOLOv8-based model. Their approach achieved a striking accuracy of 98.9% on their dataset. However, the broad and loosely defined dataset annotations hindered the model’s generalization capabilities. Liu et al. [42], considering the diverse applications of underwater object detection, proposed a YOLOv8-based framework that replaces standard convolutions with deformable convolutions and introduces a novel attention mechanism without increasing network parameters. They also substituted the original loss function and evaluated generalization performance on both the Pascal VOC dataset and a custom-built dataset. While their approach demonstrated superior generalization compared to YOLOv8n, the increased computational demands posed challenges for deployment in resource-constrained underwater environments. Several underwater object detection networks based on recent object detection architectures have been proposed in recent years. To address the increasing pollution caused by marine debris, a lightweight object detection model based on RT-DETR was designed by Lin et al.[43], achieving promising results on the Trash-ICRA19 dataset without relying on pre-trained weights. Ji et al.[44] proposed a hybrid CNN–Transformer framework, referred to as FBDPN, whose effectiveness was verified through comprehensive ablation experiments. Wang et al.[45] developed FEFM-YOLOv11, and experimental evaluations demonstrated improvements of 2.0% and 1.9% in mAP@50 and mAP@95, respectively, compared with the baseline model. Nguyen et al.[46] enhanced underwater object detection based on the YOLOv12 architecture by incorporating physics-informed data augmentation, a Residual ELAN module, and a mechanism for maintaining a large

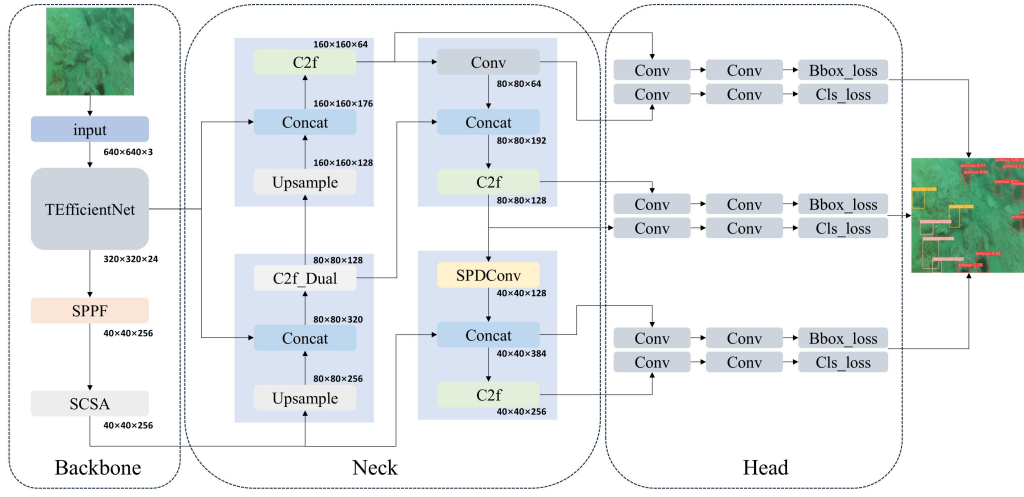


Fig. 1 SSDT-YOLO network structure is used to detect benthic organisms in complex underwater environments.

receptive field. Feng et al.[47] proposed IFEM-YOLOv13, which primarily focuses on detecting plastics, biological waste, and synthetic debris. Experimental results on their dataset indicated that the model achieved an impressive mAP@50 of 98.6%.

In summary, despite the remarkable progress achieved in recent years, current underwater object detection methods still suffer from several limitations: (1) color distortion and visual degradation caused by low illumination in underwater environments; (2) difficulty in identifying fine-grained features of small objects, leading to misdetection of microscopic targets in densely populated scenes; (3) challenges in effectively integrating lightweight computation with high-efficiency feature extraction while maintaining computational symmetry for fast inference. To address these issues, this study enhances the YOLOv8 framework by improving its backbone for stronger feature extraction, increasing feature fusion efficiency under reduced computational complexity, and incorporating a hybrid attention mechanism that jointly captures spatial and channel-level dependencies. These improvements aim to better meet the unique detection requirements of marine organisms in complex underwater environments.

To directly address the aforementioned challenges in underwater object detection, we propose a novel framework named SSDT-YOLO. The proposed architecture is designed as an integrated solution that simultaneously enhances detection accuracy and computational efficiency. Specifically, a top-down feature enhancement module is introduced, which adaptively adjusts regularization parameters through learnable mechanisms to automatically mitigate low-light distortions, thereby strengthening the backbone’s feature extraction capability and resolving the first limitation. In addition, the incorporation of SPDConv and the SCSA attention mechanism effectively tackles the second issue by improving the recognition of small and densely distributed targets. The overall framework maintains only 2.1 million parameters and achieves 214 FPS, directly overcoming the third challenge by achieving high precision without compromising efficiency—striking an elegant balance between performance and computational complexity.

3 METHODS

In underwater object recognition tasks, the original YOLOv8 model still faces challenges in feature extraction and feature fusion. These difficulties mainly arise from the high proportion of small objects in complex backgrounds, image distortions that blur fine details and edges, and the influence of underwater illumination and environmental factors. To meet the demanding requirements of high performance and real-time processing in underwater applications, this study enhances the original YOLOv8 framework and proposes a network architecture specifically optimized for small object detection in complex underwater scenes, referred to as SSDT-YOLO. The structural framework of the proposed model is illustrated as follows Fig. 1.

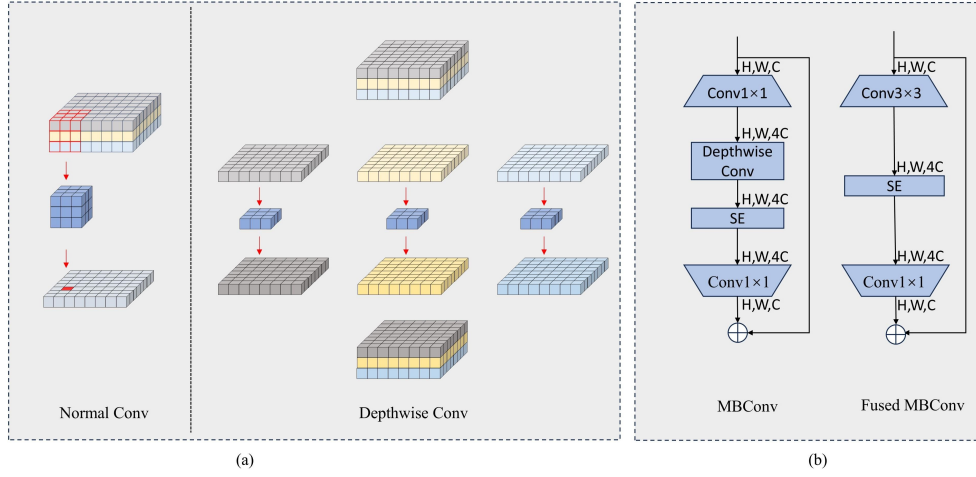


Fig. 2 Structure of MBConv and Fused-MBConv. (a) presents a schematic illustration of convolution operations, encompassing both standard convolution and depthwise convolution, using an $8 \times 8 \times 3$ input tensor and a $3 \times 3 \times 3$ kernel as an example. (b) depicts the detailed architectural design of MBConv and Fused-MBConv modules.

3.1 TEfficientNet

The complexity and variability of underwater environments lead to noticeable variations in the difficulty of detecting underwater objects. Efficient adaptation to such variations remains difficult. EfficientNet is a family of models designed to optimize both FLOPs and parameter efficiency. While deep convolutions are slow in the initial layers, they prove effective in the later stages. A major training bottleneck in efficient networks arises from the use of wide deep convolutions. Although deep convolutions have fewer parameters and less computational overhead than standard convolutions, they typically fail to fully leverage the capabilities of modern accelerators. Recently, Gupta and Tan[48] introduced the Fused-MBConv, which replaces the deep 3×3 convolutions and dilated 1×1 convolutions in MBConv with a single regular 3×3 convolution, as illustrated in Fig. 2. This approach facilitates a systematic comparison of these two building blocks.

EfficientNet V2 enhances feature extraction by refining the balanced scaling approach introduced in its predecessor, EfficientNet V1. While V1 improves CNN performance by uniformly scaling network depth, width, and resolution, V2 introduces an optimized progressive learning strategy. This method incrementally increases image resolution during training while adaptively adjusting regularization, thereby accelerating training speed without compromising accuracy.

Compared to its V1 counterpart, EfficientNet V2 primarily improves efficiency and computational speed. As a next-generation CNN, it is characterized by faster training and higher parameter efficiency. By integrating training-aware neural architecture search and optimized scaling strategies, the model achieves significant gains in both training speed and parameter efficiency. The progressive learning strategy ensures a smoother adaptation to larger input resolutions, enhancing both model generalization and convergence stability.

To accelerate network training, enhance key channel information extraction, and retain essential features for handling complex underwater environments, we propose an improved version of the EfficientV2 network—TEfficientNet, a backbone feature extraction network tailored for underwater object detection. Our approach optimizes the intricate feature extraction structure of EfficientV2 by reducing redundant information and minimizing the excessive and redundant use of Fused-MBConv and MBConv layers during training. A detailed comparison between the two architectures is illustrated in Fig. 3. This refinement enables more efficient key feature extraction while strategically reducing feature extraction hierarchy where appropriate. Compared to the original backbone network, TEfficientNet significantly lowers computational complexity, making it better suited for underwater applications while maintaining competitive performance.

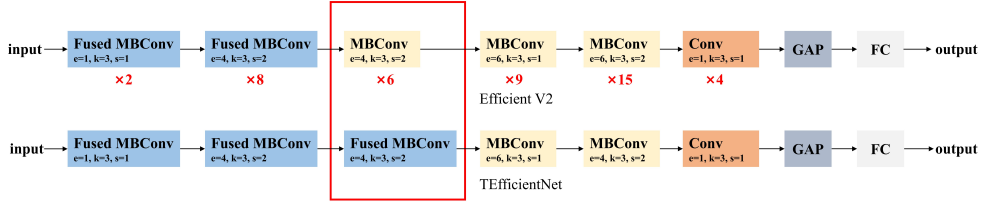


Fig. 3 Efficient V2 network structure and TEfficientNet network structure proposed in this paper. Compared to the original EfficientNetV2 architecture, the proposed TEfficient network eliminates redundant convolutional operations and redefines the information transmission pathways to enhance computational efficiency.

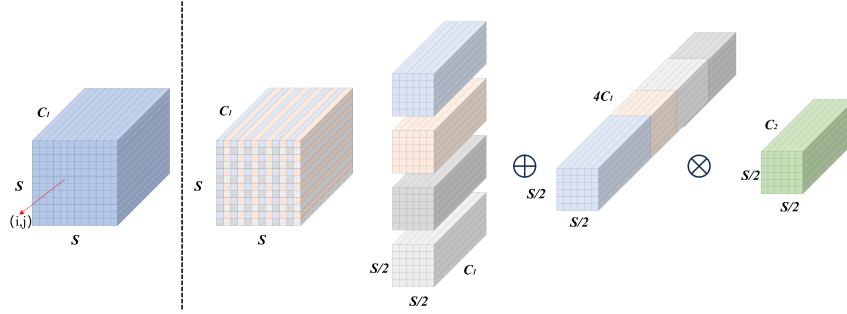


Fig. 4 An example demonstrating the use of SPDConv for enhancing the detection of small-scale objects. Illustrates the case where the scaling factor is set to 2.

3.2 Space-to-Depth Convolution

Since the advent of AlexNet, CNNs have demonstrated outstanding performance in numerous computer vision tasks, particularly in the detection and recognition of real-world objects. Models such as the R-CNN series, YOLO series, SSD, and EfficientDet have been widely used. However, all these CNN models require "high-quality" inputs for training and inference, such as finely detailed images and well-shaped medium to large objects for optimal detection. Unfortunately, these theoretically high-quality inputs do not hold true in underwater environments during image training. In fact, underwater small-object detection is a particularly challenging task. Objects in underwater environments are often poorly lit and unclear due to the effects of light, and lens distortions during image capture further complicate recognition. Additionally, the limited contextual information available for model learning exacerbates the difficulty in detecting underwater objects. Consequently, existing models struggle to recognize underwater images.

To address these challenges and improve object recognition capabilities in resource-constrained underwater edge computing systems, we employ SPDConv to enhance feature extraction and reduce computational complexity. Let's briefly introduce this

approach.

$$\begin{aligned}
f_{0,0} &= X[0 : S : scale, 0 : S : scale], f_{1,0} = X[1 : S : scale, 0 : S : scale], \dots, \\
f_{scale-1,0} &= X[scale - 1 : S : scale, 0 : S : scale]; \\
f_{0,1} &= X[0 : S : scale, 1 : S : scale], f_{1,1}, \dots, \\
f_{scale-1,1} &= X[scale - 1 : S : scale, 1 : S : scale]; \\
&\vdots \\
f_{0,scale-1} &= X[0 : S : scale, scale - 1 : S : scale], f_{1,scale-1}, \dots, \\
f_{scale-1,scale-1} &= X[scale - 1 : S : scale, scale - 1 : S : scale].
\end{aligned} \tag{1}$$

Given an arbitrary input feature map X , its sub-feature maps $f_{x,y}$ are formed by selecting entries $X(i, j)$ at intervals defined by the scale factor along both spatial dimensions. Consequently, each sub-feature map represents a downsampled version of X by a factor of scale.

For instance, when $scale = 2$, the original feature map is decomposed into four sub-feature maps: $f_{0,0}, f_{0,1}, f_{1,0}, f_{1,1}$, each with a spatial resolution of $(S/2, S/2, C_1)$. These sub-feature maps are then concatenated along the channel dimension, resulting in a transformed feature map X' . This transformation reduces the spatial resolution while increasing the number of channels by a factor of 2. After the SPD feature conversion layer, a stride 1 convolution layer is added using a filter and further converted $X'(\frac{S}{scale}, \frac{S}{scale}, scale^2 C_1) \rightarrow X''(\frac{S}{scale}, \frac{S}{scale}, scale^2 C_2)$.

SPDConv facilitates the transformation of spatial features into depth features, enabling the capture of higher-level semantic information and thereby enhancing the model's capability to recognize underwater targets. This mechanism effectively expands the receptive field and improves detection accuracy. The structural diagram of SPDConv is presented below. In our proposed approach, we primarily replace the Conv modules within the backbone network with SPDConv, thereby reinforcing the feature extraction capacity of the improved architecture.

3.3 Spatial and Channel Synergistic Attention

The attention mechanism is designed to refine and localize the network's focus, ultimately forming attention hotspots, which are critical for object detection. By leveraging attention mechanisms, perceptual information is distilled, and contextual relationships are enhanced. These computational strategies are broadly classified into three paradigms: channel-wise attention, spatial attention, and combined channel-spatial attention. Each paradigm adaptively quantifies the relevance of input features to optimize detection accuracy through noise reduction and targeted information aggregation.

In underwater environments, objects are often small and densely clustered, making them difficult to detect using conventional attention mechanisms. To resolve this limitation, the SCSA attention framework is integrated within the final layer of the backbone architecture to amplify the network's perceptual scope and strengthen feature discriminability. The SCSA's architectural composition comprises two modular units: Shareable Multi-Semantic Spatial Attention (SMSA) and Progressive Channel-wise Self-Attention (PCSA). Comprehensive implementation schematics for these modules are delineated in Fig. 5. Following the disaggregation of distinct feature components and the encoding of spatially distributed attributes with varying contextual meanings, the synthesis of a region-selective weighting matrix becomes imperative. An element-wise sigmoid operator is applied to compute spatial attention coefficients, dynamically modulating the prominence of targeted sub-regions within the input feature space. The operational mechanics of the proposed SMSA module are mathematically formalized as follows:

$$Attn_H = \sigma(GN_H^K(Concat(\tilde{X}_H^1, \tilde{X}_H^2, \dots, \tilde{X}_H^K))) \tag{2}$$

$$Attn_W = \sigma(GN_W^K(Concat(\tilde{X}_W^1, \tilde{X}_W^2, \dots, \tilde{X}_W^K))) \tag{3}$$

$$SMSA(X) = X_s = Attn_H \times Attn_W \times X \tag{4}$$

σ represents the normalization of Sigmoid, While GN_H^K and GN_W^K represent the GN of K along H and W dimensions, respectively.

To retain and exploit the multi-scale contextual attributes extracted via the SMSA framework, we integrate these features into subsequent processing stages to enhance discriminative performance in downstream tasks, the framework employs a gradual compression strategy utilizing average pooling as its core mechanism. Unlike

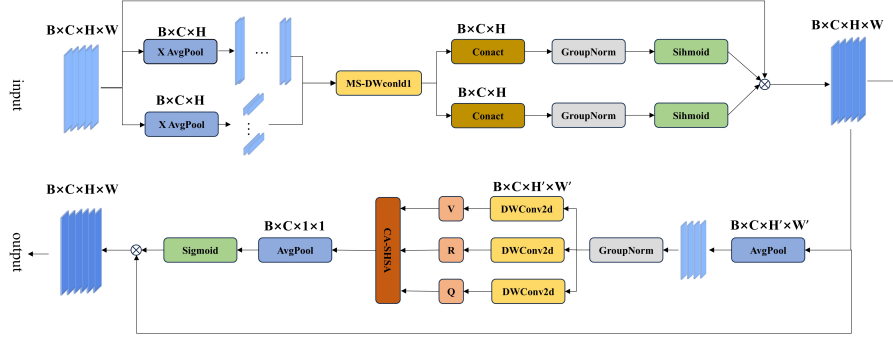


Fig. 5 A brief description of SCSA. Designed to enhance the learning of channel-wise self-attention, thereby improving the model's capability to discriminate among benthic organisms. The upper section illustrates the architecture of the SMSA module, while the lower section presents the structure of the PCSA module. Here, b denotes the batch size, c the number of channels, and H and W represent the height and width of the feature map, respectively. The variable n indicates the number of sub-function groups, and 1P corresponds to a single pixel.

traditional convolutional techniques for modeling channel relationships, PCSA demonstrates improved adaptability to input fluctuations and efficiently incorporates spatial priors from SMSA to strengthen feature extraction. The specific operational steps of PCSA are outlined below:

$$X_p = POOL_{(7,7)}^{(H,W) \rightarrow (H',W')}(X_S) \quad (5)$$

$$F_{proj} = DWConv1d_{(1,1)}^{C \rightarrow C'} \quad (6)$$

$$Q = F_{proj}^Q(X_p) = F_{proj}^V(p) = F_{proj}^V(X_p) \quad (7)$$

$$X_{attn} = Attn(Q, K, V) = Soft \max\left(\frac{QK^T}{\sqrt{C}}\right)V \quad (8)$$

$$PCSA = X_C = X_S \times \sigma(POOL_{(H',W')}^{(H',W') \rightarrow (1,1)}(X_{attn})) \quad (9)$$

$POOL_{(k,k)}^{(H,W) \rightarrow (H',W')}$ indicates that the image resolution is reduced from (H, W) to (H', W') through a $k \times k$ kernel pooling operation. F_{proj} represents linear projection of generated queries, keys and values.

The synergistic spatial and guided attention mechanism is designed to complement each other. A sequential structure is adopted to integrate the SMSA and PCSA modules, forming the SCSA mechanism. The former extracts multi-semantic spatial information from each feature map, providing precise spatial priors for the latter. The latter refines the semantic understanding of local subfields X_i by leveraging the holistic feature map X , thereby mitigating the semantic discrepancies introduced by the multi-scale convolutions in the former. The constructed SCSA is as follows:

$$SCSA(X) = PCSA(SMSA(X)) \quad (10)$$

The SCSA mechanism produces a combined effect between spatial and channel attention by merging the SMSA and PCSA modules. Within SMSA, the integration of multi-semantic spatial cues coupled with a progressive compression approach effectively establishes spatial priors for channel attention. Concurrently, PCSA employs this spatial context to further refine channel features via a self-attention process, thereby reducing inconsistencies among different semantic levels. Given that the backbone network is pivotal for feature extraction within the YOLOv8 framework, integrating the SCSA module amplifies pertinent information while diminishing redundant data, ultimately enhancing both feature extraction performance and detection precision.

3.4 C2f-Dual

Within underwater image detection, reducing the parameter count is crucial for enhancing the real-time performance of object detection models. The C2f module in YOLOv8 contains a high parameter count due to its intricate convolutional operations and the repeated design of the Bottleneck module. Although this architecture enhances the model's ability to process features, it also introduces considerable computational overhead. This becomes particularly problematic in resource-limited environments, as

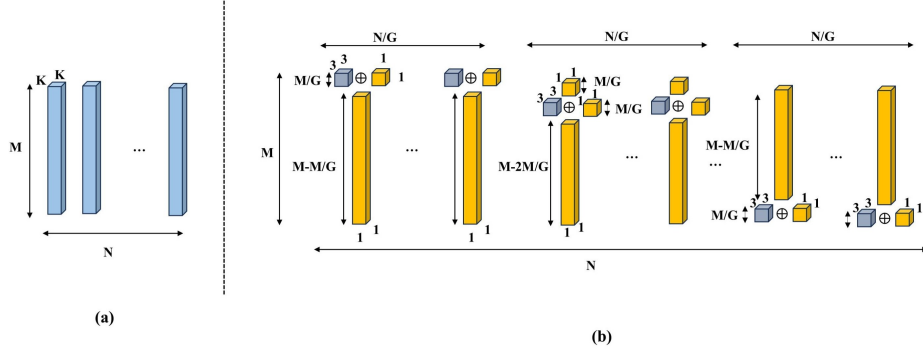


Fig. 6 The figure illustrates the feature extraction process employing a dual convolutional structure to enhance representation learning. M denotes the depth of the input feature map, n denotes the depth of the output feature map, and G represents the number of groups used in dual convolution operations.

it leads to a reduction in inference speed. Efficient feature extraction is the core challenge in real-time object detection tasks. To reduce the parameter count, we propose the C2f-Dual module as a replacement for C2f. Specifically, DualConv is a lightweight convolutional network structure designed to construct compact deep neural networks. By integrating 3×3 and 1×1 convolutional kernels to process the same input feature map channels, it optimizes information processing and feature extraction. The dual convolution is shown in Fig. 6. Assuming the output feature map has dimensions $D_o \times D_o \times n$, where D_o represents the width and height of the output feature map, the standard convolution operation, as illustrated in Figure 7(a), applies convolutional filters to the input feature map. Each filter in the convolutional layer has dimensions $K \times K \times M$ where $K \times K$ denotes the kernel size. Consequently, the total number of FLOPs performed by the standard convolution layer, denoted as $FLOPs_{sc}$, is given by:

$$FLOPs_{sc} = D_o^2 \times K^2 \times M \times N \quad (11)$$

In DualConv architectures, the cardinality parameter G governs the proportional distribution of $K \times K$ convolutional kernels within filter groups. For a specified group count G , the composite convolution operations comprising simultaneous $K \times K$ and 1×1 kernels constitute $1/G$ of the total channel allocation, while the residual 1×1 convolutional operations account for $(1 - 1/G)$ proportion. Consequently, the parametric composition of a dual convolutional layer structured with G filter groups is mathematically expressed as:

$$FLOPs_{CC} = (D_o^2 \times K^2 \times M \times N + D_o^2 \times M \times N)/G \quad (12)$$

The number of parameters for the remaining 1×1 is:

$$FLOPs_{PC} = (D_o^2 \times K^2 \times M) \times (1 - 1/G) \quad (13)$$

The computational cost reduction ratio $R_{DC/SC}$ when comparing the dual convolution layer to the standard convolution layer can be expressed as:

$$R_{DC/SC} = \frac{FL_{DC}}{FL_{SC}} = \frac{1}{G} + \frac{1}{K^2} \quad (14)$$

The DualConv architecture employs group convolution techniques to optimize filter arrangement, significantly reducing computational costs and parameter quantities. This structural design integrates 3×3 and 1×1 convolutional kernels to leverage their complementary advantages: 3×3 kernels capture richer spatial information during feature extraction, while 1×1 kernels facilitate inter-channel feature interaction and integration without introducing excessive parameters or computational complexity. In DualConv, 3×3 kernels extract spatial features from feature maps, whereas 1×1 kernels consolidate these features while minimizing model parameters. Each filter group processes partitioned input channels independently, followed by output concatenation

to enable efficient cross-channel information flow and fusion. This design preserves network depth and representational capacity while reducing computational complexity and model size, rendering it suitable for resource-constrained environments.

This study proposes an innovative integration of DualConv with conventional C2f modules, resulting in the C2f-Dual architectural module that substantially reduces parameter count while simultaneously enhancing feature fusion efficiency. The proposed configuration augments network performance through accelerated feature extraction rates and improved representational capacity, thereby achieving heightened computational efficiency in neural network operations.

4 EXPERIMENT

4.1 Experiment Introduction

The experimental configuration employed a software stack comprising Ubuntu 20.04.2 LTS, PyTorch 2.5.1, CUDA 12.1, Python 3.10.16, and Ultralytics YOLOv8.3.75 framework. Hardware acceleration was implemented through an Intel® Xeon® Gold 6326 CPU @ 2.90 GHz paired with an NVIDIA GeForce RTX 3090 GPU featuring 24GB GDDR6X VRAM. Table 1 details the standardized hyperparameter configuration adopted for network training across all architectural variants.

Table 1 Values or Methods for Training and Testing Hyperparameters.

Parameter	Configuration
Training Epochs	100
Batch Size	16
Learning Rate	0.01
Momentum	0.937
Optimizer	SGD
Image size	640×640
IoU	0.7

4.2 Dataset Introduction

The Deep-sea Underwater Object Detection (DUO)[49] dataset comprises a total of 7,782 high-quality underwater images, primarily extracted from video frames recorded by underwater robots in natural marine environments. This dataset includes images annotated with bounding boxes, facilitating precise object detection. The evaluation in our experiments focuses on four categories of marine organisms: sea cucumbers, sea urchins, scallops, and starfish.

To construct the experimental dataset, the images were randomly divided into training, validation, and test sets in a 7:1:2 ratio. This resulted in 5,547 images for training, 778 images for validation, and 1,557 images for testing. Additionally, we incorporated the URPC2019 underwater object detection dataset as a complementary data source to enhance the DUO dataset. The Underwater Robotics Professional Competition 2019(URPC2019)[50] dataset contains 4,707 high-quality underwater images, covering the same four categories of marine objects as DUO. Specifically, 3,767 images were allocated for model training, 695 images for validation, and 245 images for testing during the training process.

To validate the effectiveness of constituent modules in the proposed SSDT-YOLO framework, systematic ablation studies were conducted to quantify the individual and collective impacts of architectural enhancements on detection performance. The evaluation protocol employed established metrics including Precision(P), Recall(R), Average Precision(AP), mean Average Precision(mAP), and inference speed quantified through per-image processing latency. Within the dataset schema, defect-free samples were designated as positive instances, while defective specimens were categorized as negative cases. The trained detection model generates probabilistic predictions regarding sample polarity, producing binary classifications through threshold operations. Correctly identified positives constitute True Positives (TP), whereas accurately recognized negatives form True Negatives (TN), unrecognized errors constitute are

respectively captured as False Positives (FP) and False Negatives (FN). The detection metrics are formally defined as:

$$precision = \frac{TP}{TP + FP} \quad (15)$$

$$recall = \frac{TP}{TP + FN} \quad (16)$$

The calculation formulas for AP and mAP are as follows:

$$AP = \int_0^1 P(r)dr \quad (17)$$

$$mAP = \frac{\sum AP}{k} \quad (18)$$

The improved SSDT-YOLO is built upon the Ultralytics-YOLO framework, which offers a user-friendly interface, outstanding performance, and high scalability. This framework supports various object detection tasks, including image classification, object localization, and instance segmentation. To achieve better convergence, the network undergoes training for 100 epochs.

4.3 Comparative experiment

4.3.1 Comparison of Backbone

To substantiate the superiority of the proposed backbone in enhancing feature extraction capability, we conducted a comparative study using YOLOv8 as the baseline. We replaced the baseline backbone with various high-performance alternatives and evaluated their performance. As shown in table 2, our improved backbone module outperforms YOLOv8 models incorporating Swin Transformer, FasterNet, and EMO backbones in detecting Holothurian, Scallop, and Starfish. Furthermore, it also surpasses the original backbone in both accuracy and efficiency. Specifically, our proposed backbone achieves the highest AP improvement of 5.2% for Holothurian, followed by 0.2% for Echinus, 0.6% for Scallop, and 0.8% for Starfish compared to the original network. Additionally, our model exhibits a 1.7% increase in mAP@50 while reducing parameter size by 0.8M compared to the baseline and other backbones.

Table 2 Backbone comparison results.

Model	AP(%)				mAP@50(%)	Para(M)	FLOPs(G)
	Holothurian	Echinus	Scallop	Starfish			
baseline	78.0	93.4	68.6	91.5	82.9	3.1	8.1
Swin	75.6	92.4	62.1	90.3	80.1	29.9	79.1
FasterNet	79.6	93.2	66.8	91.2	82.7	15.2	37.0
EMO	81.4	93.7	66.6	91.9	83.4	6.3	9.2
TEfficientNet	83.2	93.6	69.2	92.3	84.6	2.3	24.8

4.3.2 Comparison of Convolution

We conducted a comparative analysis of SPDCConv against the original convolution module, RFACConv, Dynamic Conv, and Dynamic Snake Conv to validate the efficiency improvement achieved by replacing the convolution module in the baseline network. As shown in Table 3, although SPDCConv does not surpass the aforementioned convolution modules in terms of accuracy, it exhibits significantly lower parameter count and computational complexity. While alternative convolution modules enhance accuracy, they also introduce substantial computational overhead, which is detrimental in resource-constrained underwater environments. In contrast, replacing the traditional convolution with SPDCConv contributes significantly to model light weighting without compromising accuracy.

In order to verify the effectiveness of SPDCConv for underwater small target detection, we compared the basic model with the model with SPDCConv added when there are many underwater small targets. The comparison chart is shown in Fig7. It can be

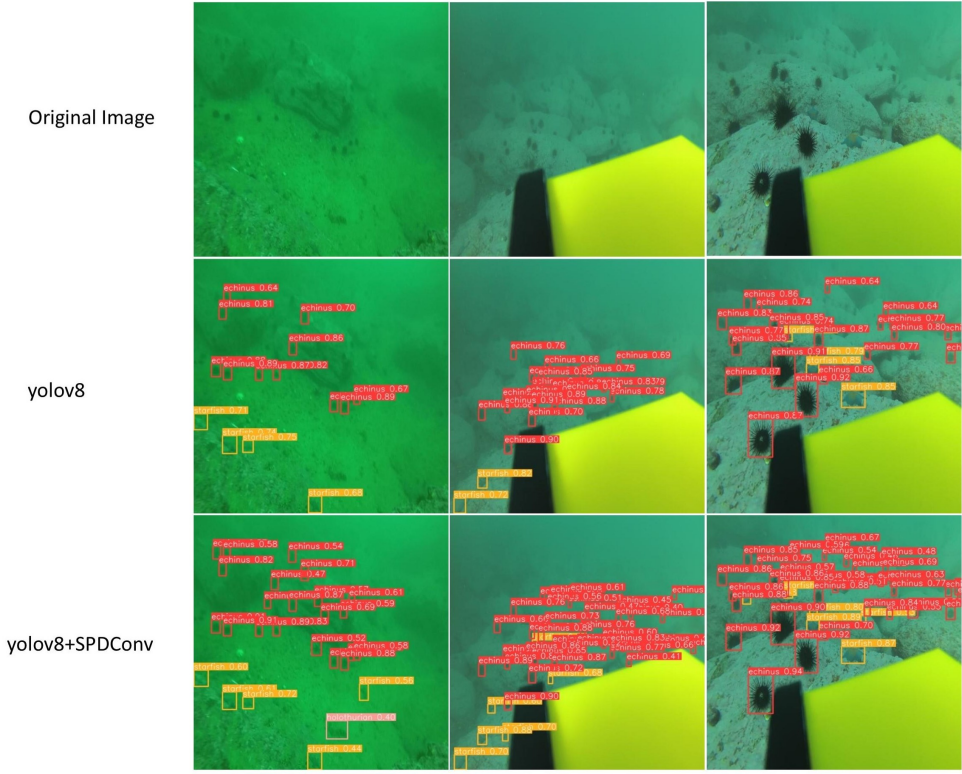


Fig. 7 Comparison of underwater small target detection. The first column illustrates the original input images, followed by detection results from the baseline YOLOv8 in the second column, and those from the YOLOv8+SPDConv model in the third column. The integration of SPDConv leads to noticeably denser object detections, particularly in cluttered regions.

seen that compared with the basic model, the model with SPDConv can better detect underwater small targets under the same circumstances.

Table 3 Convolution comparison results.

Model	$AP(\%)$				$mAP@50(\%)$	$Para(M)$	$FLOPs(G)$
	Holothurian	Echinus	Scallop	Starfish			
baseline	78.0	93.4	68.6	91.5	82.9	3.1	8.1
RFACnv	79.9	93.5	68.4	91.7	83.4	3.1	8.3
Dynamic	78.9	93.5	68.2	91.3	83.0	4.7	6.9
DynamicSnake	79.0	93.4	66.3	91.1	82.4	3.8	9.5
SPDConv	76.8	93.2	68.4	90.9	82.3	2.6	7.4

4.3.3 Comparison of C2f

We conducted a comparative analysis of our novel C2f-Dual module against the baseline C2f module in YOLOv8, as well as C2f-CSPPC, C2f-RCS, and C2f-OREPA, to demonstrate the superior generalization capability and parameter efficiency of our proposed design. Compared to the baseline model, our module reduces the parameter count by 0.8M and decreases computational complexity by 1.2 GFLOPs. Similar to the rationale behind adopting SPDConv, while the C2f-Dual module exhibits slightly lower accuracy than the aforementioned variants, it significantly outperforms them in terms of model compactness and efficiency. Thus, we have strong reason to believe that our proposed module is an effective choice. The detailed results of various C2f modules on the DUO dataset are presented in table 4.

Table 4 C2f convolution comparison results.

Model	AP(%)				mAP@50(%)	Para(M)	FLOPs(G)
	Holothurian	Echinus	Scallop	Starfish			
baseline	78.0	93.4	68.6	91.5	82.9	3.1	8.1
C2f-CSPPC	77.5	92.9	65.2	90.2	81.5	2.5	6.8
C2f-RCS	80.2	93.6	70.2	91.6	83.9	4.1	10.5
C2f-OREPA	78.6	93.4	67.0	91.2	82.6	3.6	9.6
C2f-Dual	78.2	93.1	68.5	91.1	82.7	2.3	6.9

4.3.4 Comparison of Attention Mechanisms

To validate the effectiveness of the attention mechanism incorporated in our model and demonstrate that the SCSA attention module enhances performance in small-scale image tasks, we conducted a comparative analysis against the baseline YOLOv8 and classical attention mechanisms, including CA, ECA, and CBAM. As shown in table 5, the SCSA attention mechanism outperforms the others in recognizing Echinus, Scallop, and Starfish. Moreover, it achieves a 0.8% higher mAP@50 than the baseline while reducing the parameter count by 0.1M.

Table 5 Attention Mechanisms Comparison results.

Model	AP(%)				mAP@50(%)	Para(M)	FLOPs(G)
	Holothurian	Echinus	Scallop	Starfish			
baseline	78.0	93.4	68.6	91.5	82.9	3.1	8.1
CA	78.6	93.1	68.6	91.3	82.9	3.0	8.1
ECA	79.5	93.6	68.5	91.7	83.3	3.0	8.1
CBAM	79.4	93.6	68.0	91.6	83.2	3.2	9.6
SCSA	79.3	93.6	70.1	91.8	83.7	3.0	8.2

4.3.5 Model Comparison

We first conducted a literature review and experimental evaluation of various models on the DUO dataset. The comparative results with different models are presented in table 6. Furthermore, to validate the generalization capability of our model, we performed additional experiments on the URPC2019 dataset under identical hyperparameter settings. This study verifies the effectiveness of the improved SSDT-YOLOv8 model through comparative experiments with state-of-the-art object detection algorithms. The proposed model consistently outperforms the baseline across both the

Table 6 Comparison between SSDT-YOLO and other algorithms.

Model	Metrics					
	Backbone	Size	mAP@50(%)		Para(M)	FLOPs(G)
			DUO	URPC2019		
Faster RCNN	ResNet-50	640×640	45.6	79.1	41.4	207.5/178.1
Retinanet	ResNet-50	640×640	47.6	76.3	37.9	149.1/175.5
YOLO-v3	Darknet-53	640×640	80.4	80.5	12.1	18.9
YOLO-v5	CSPDarknet	640×640	81.5	79.5	2.5	7.1
YOLO-v6	EfficientRep	640×640	80.7	78.2	4.2	11.8
YOLO-v8	CSPDarknet	640×640	82.9	80.5	3.1	8.1
YOLO-v10	C3K	640×640	81.0	78.3	2.7	6.7
YOLO-v11	CSPNet	640×640	82.5	80.6	2.6	6.3
YOLO-v12	R-ELAN	640×640	81.9	78.5	2.5	6.3
DETR	ResNet-50	640×640	59.0	79.0	41.6	60.5/81.6
ours	TEfficientNet	640×640	84.9	83.0	2.1	24.2

DUO and URPC2019 datasets. Specifically, on the DUO dataset, the mAP@50 of our algorithm surpasses Faster R-CNN, RetinaNet, YOLOv3, YOLOv5, YOLOv6, YOLOv8, YOLOv10, YOLOv11, YOLOv12 and DETR by 39.3%, 37.3%, 4.5%, 3.4%, 4.2%, 2.0%, 3.9%, 2.4%, 3.0% and 25.9%, respectively. Similarly, on the URPC2019 dataset, our approach achieves mAP@50 improvements of 3.9%, 6.3%, 2.5%, 3.5%, 4.8%, 2.5%, 2.0%, 2.4%, 4.5% and 4.0% over the aforementioned models. Furthermore, the proposed model features a smaller footprint and reduced computational complexity, making it well-suited for deployment on embedded devices. A comprehensive evaluation of various performance metrics indicates that the proposed algorithm effectively meets the requirements for rapid and precise detection of benthic organisms in complex underwater environments.

In terms of FLOPs, SSdT-YOLO achieves a significant improvement over two-stage detection models such as Faster R-CNN and RetinaNet, reducing FLOPs by nearly an order of magnitude. Compared to one-stage models in the YOLO series, our model shows a slight increase in FLOPs but achieves superior accuracy and a lower parameter count. Our proposed model achieves the highest accuracy on both the DUO and URPC datasets while maintaining the smallest number of parameters. Compared to the DETR family, our model surpasses DETR in terms of accuracy, parameter efficiency, and computational complexity. These results confirm the effectiveness of our proposed improvements.

4.4 Visual analysis

Heatmaps are a key visualization tool in deep learning-based object detection. The location information of detected objects can be reflected through heatmaps. By using heatmaps, specific regions of an image can be selected, assisting the neural network in making final classification decisions. High-response areas in heatmaps provide essential information to support the final differentiation of the object detection network. To ensure experimental fairness, we kept the parameter settings consistent for heatmap comparisons, with the input image size fixed at 640×640 and GradCAM used as the heatmap generation method. We selected a few images from the DUO dataset and conducted a comparative experiment on heatmap visualization. In the figures, we demonstrate the detection capabilities of YOLOv8 and SSdT-YOLO for four representative underwater marine organisms: echinus, starfish, holothurian, and scallop, along with the detection capabilities of various classical attention mechanisms Fig. 8. is an experimental comparison of the heat map.

From the highlighted hotspots, it is evident that, due to the large receptive field of feature information, the traditional YOLOv8 struggles to focus on the regions of interest. This leads to imprecise detection and recognition of echinus, starfish, and scallop. In contrast, in SSdT-YOLO, the attention mechanism is employed to enhance feature extraction capabilities. During echinus detection, it is clear that the target to be detected is concentrated in the central part of the image, which our algorithm effectively focuses on, while the original YOLOv8 algorithm emphasizes the peripheral areas, neglecting the center region that requires closer monitoring. For starfish detection, where the targets are scattered and relatively small, our algorithm captures these subtle differences well, while YOLOv8 appears somewhat confused. In holothurian detection, we observe that the target is concentrated in the bottom half of the image, which our algorithm accurately focuses on, while YOLOv8 directs its attention to the upper-left section. For scallop detection, the targets are relatively large and widely distributed, with only the upper-left area being empty; our algorithm detects this effectively, whereas the performance of YOLOv8 is not as satisfactory. However, when detecting some challenging targets, both YOLOv8 and the SSdT-YOLO proposed in this paper encounter some issues. This is not only due to limited parameters and computational resources but also because many of the required targets are obstructed by underwater lighting or sediments. Therefore, further research is needed to address these challenges.

To highlight the exceptional capabilities of the SSdT-YOLO model, several images selected from the test set were tested, as shown in the Fig. 9. The visual evidence demonstrates that the SSdT-YOLO model excels in detecting smaller objects and achieves higher accuracy in complex scenarios. This performance underscores the enhanced effectiveness of the model in addressing the challenges of object detection with precision. Overall, the SSdT-YOLO model effectively reduces both

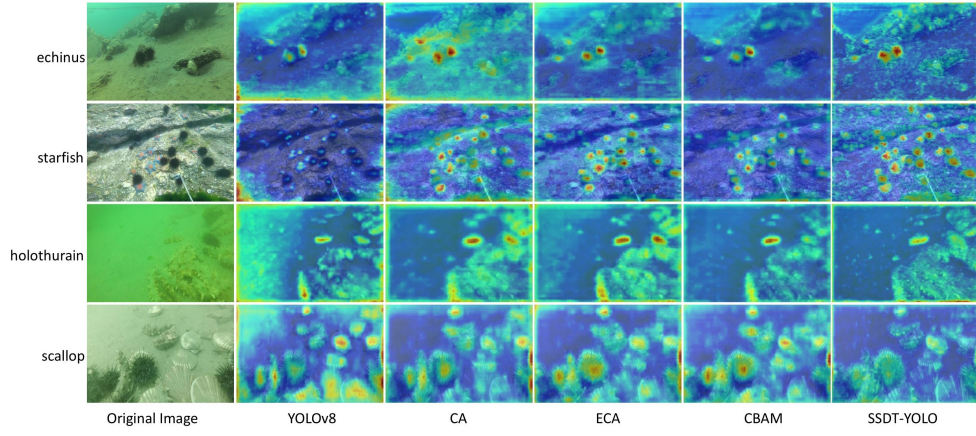


Fig. 8 A comparative analysis is conducted on the proposed model against alternative attention mechanisms, with results visualized through heatmaps. The visualization is structured into four rows, each depicting the attention responses of different benthic organisms under various attention mechanisms.

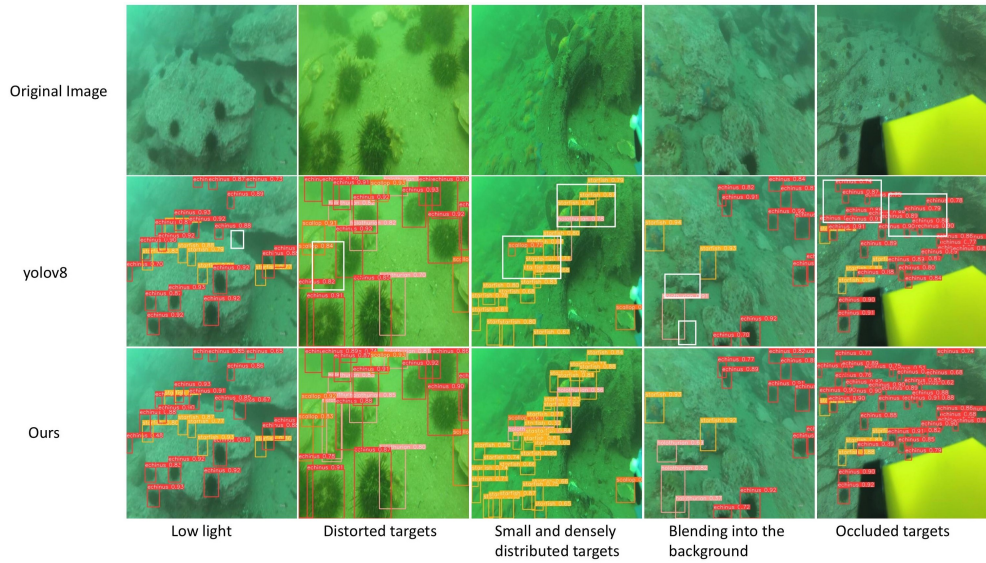


Fig. 9 Detection capabilities are evaluated across various complex underwater environments. Each column corresponds to the same base image, with five columns in total representing different challenging conditions: low-light environments, distorted targets, small and densely distributed targets, targets that are difficult to recognize, and occluded targets. Objects enclosed by white bounding boxes are correctly identified by the proposed SSDT-YOLO model, in contrast to the traditional baseline model, which fails to detect them.

missed and incorrect detections, while enhancing optimization in complex underwater environments. Particularly in scenarios involving numerous small targets, intricate terrains, and distorted images, it ensures exceptional accuracy and detection speed. In summary, our work proves to be effective.

4.5 Ablation experiment

In this section, ablation experiments are conducted to validate the effectiveness of different modules. A higher $mAP@50$ value indicates better model performance in terms of detection and recognition. The $mAP@50-95$ metric is particularly valuable as it demonstrates the model's detection and recognition capabilities in more complex scenarios. Additionally, the model's parameter size is another critical factor: fewer parameters result in faster processing and easier deployment in embedded systems. Frame rate is also an important metric; for underwater object detection, a higher frame rate means the robot utilizing our model can identify targets more quickly and make more efficient decisions for subsequent actions. During the measurement of the model's frame rate, we pre-warm the model for 200 iterations to better enable it to showcase its performance. After this, we perform a loop of 1000 iterations to calculate the average frame rate, which helps minimize error. Different modules were sequentially used to

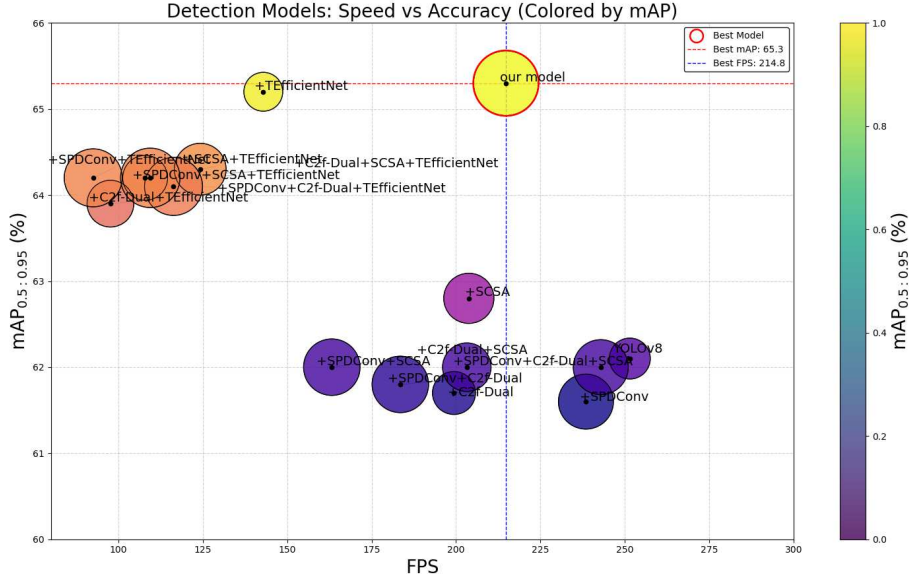


Fig. 10 A scatter plot is used to compare the detection speed and accuracy of the models evaluated in the ablation experiments.

improve the YOLOv8 algorithm. The results of this ablation experiment are shown in the table below. In the first row of Table 7, the detection results of the original YOLOv8 network are shown as the baseline.

Compared to the baseline model, when only the SPDCConv module is added, both mAP@50 and mAP@50-95 show a decrease to varying extents. Specifically, mAP@50 decreases by 0.6%, while mAP@50-95 drops by 0.5%, and the frame rate decreases by 13%. However, the model parameters are reduced by 319,328, and the model size decreases by 0.7MB. When only our innovative C2f-Dual module is added, the overall model sees a decrease of 0.2% in mAP@50 and a 0.4% drop in mAP@50-95, with the model size reducing by 0.7MB, parameters decreasing by 299,648, and the frame rate experiencing the most significant decline, dropping by 52.1%. When only the SCSA module is incorporated, mAP@50 increases by 0.2%, mAP@50-95 rises by 0.7%, but the number of parameters increases by 35,854, the model size reduces by 0.2MB, and the frame rate drops by 47.7%. When the innovative Efficient V2 module is added, mAP@50 increases by 1.7%, mAP@50-95 rises by 3.1%, parameters decrease by 714,360, model size reduces by 1.3MB, and the frame rate increases by 37.4%. When all of the above modules are combined with the baseline model, mAP@50 increases by 2%, mAP@50-95 rises by 3.2%, the parameters reduce by 815,352, corresponding to a 27.1% reduction in model parameters, the model size decreases by 2.5MB, equivalent to a 39.6% reduction in model size, and the frame rate only drops by 36.7%. From this analysis, we can conclude that our designed model achieves the expected results and meets the real-time and accuracy requirements for underwater robots.

5 Conclusion

This paper proposes SSDT-YOLO, a specialized detection framework for benthic organisms in marine environments, enhancing YOLOv8 in feature extraction, feature fusion, and boundary localization. Addressing the challenges posed by real-world underwater scenarios—such as object blurriness, occlusions, and small object sizes due to complex environments, this study introduces four key innovations. First, to mitigate the issues of uneven underwater illumination and the high degree of similarity between target objects and their surroundings, the backbone network incorporates the proposed TEfficientNet module. This module accelerates network training while enhancing feature extraction capabilities, enabling the model to more effectively identify targets, extract critical channel features, and preserve essential information, thereby improving detection accuracy. Second, to address the loss of fine-grained information for small objects and the insufficient learning of effective feature representations during training, the standard convolutional module in the original model is replaced with SPDCConv for feature extraction, thereby improving model efficiency. Third, considering the limitations of underwater computational resources, a novel C2f-Dual module

Table 7 Performance comparison of removing or replacing improved modules in SSDT-YOLO.

Model	Method	SPDConv	C2f-Dual	SCSA	TEfficientNet	$mAP@50(\%)$	$mAP@50 - 95(\%)$	$Parameters$	$Modelsize(MB)$	$FPS(frames\ s^{-1})$
baseline	a					82.9	62.1	3006428	6.3	251.5
	b		✓			82.7	61.7	2706780	5.6	199.4
	c		✓	✓	✓	84.3	64.2	2272868	4.7	107.9
	d		✓		✓	83.8	63.9	2245476	4.7	97.6
	e		✓	✓		82.9	62.0	2984540	5.9	203.2
	f			✓		83.7	62.8	3042292	6.1	203.8
	g			✓	✓	84.6	64.3	2296164	4.8	124.2
	h				✓	84.6	65.2	2292068	5.0	142.9
	i	✓				82.3	61.6	2687100	5.6	238.5
	j	✓	✓			82.2	61.8	2604572	5.2	183.5
	k	✓	✓		✓	84.2	64.1	2225060	4.6	116.3
	l	✓	✓	✓		82.9	62.0	2946332	5.9	242.9
	m	✓		✓		82.7	62.0	2987036	5.9	163.2
	n	✓		✓	✓	84.2	64.2	2519356	4.6	109.4
	o	✓			✓	84.1	64.2	2271652	4.7	92.5
	ours	✓	✓	✓	✓	84.9(+ 2.0%)	65.3(+ 3.2%)	2191076	4.8	214.8

is introduced in the neck network to replace the traditional C2f module for feature fusion. This enhances the model’s generalization ability while, in conjunction with SPDConv, significantly reducing computational costs and parameter count. Finally, given that underwater targets are predominantly small, densely distributed, and difficult to discern, an SCSA attention mechanism is integrated into the backbone to enhance the model’s attentional capacity. This mechanism models the interactions among local regions at various scales, improving the network’s ability to handle complex small-object detection tasks, thereby ensuring reliable recognition of small underwater targets.

Experimental evaluations on 7,782 images from the DUO dataset and 4,707 high-quality underwater images from URPC2019 demonstrate that SSDT-YOLO achieves mAP values of 84.9% and 83.0%, respectively, surpassing the baseline model’s 82.9% and 80.5%. Additionally, SSDT-YOLO reduces parameter count to 2.0M, significantly lower than YOLOv8’s 3.1M. When tested under a warm-up phase of 200 iterations followed by 1,000 cycles, the proposed model attains a frame rate of 214.8 FPS, outperforming various state-of-the-art underwater object detection algorithms in terms of accuracy, compactness, and processing speed. These results underscore the model’s effectiveness in handling complex underwater scenarios with limited computational resources, offering a practical and efficient solution while demonstrating its significant potential for real-world applications.

However, the model exhibits an increase in FLOPs. Future work will focus on reducing model complexity through knowledge distillation and network pruning to further optimize computational efficiency and enhance detection speed.

Funding

This research was supported by the National Natural Science Foundation of China Excellent Youth Fund (62222306). This research was supported by the National Natural Science Foundation of China (624030192).

References

- [1] Su, X., Ullah, I., Liu, X., Choi, D.: A review of underwater localization techniques, algorithms, and challenges. *Journal of Sensors* **2020**(1), 6403161 (2020)
- [2] 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* **527**, 204–232 (2023)
- [3] Wu, Y., Duan, Y., Wei, Y., An, D., Liu, J.: Application of intelligent and unmanned equipment in aquaculture: A review. *Computers and Electronics in Agriculture* **199**, 107201 (2022)
- [4] Kim, B., Yu, S.-C.: Imaging sonar based real-time underwater object detection utilizing adaboost method. In: 2017 IEEE Underwater Technology (UT), pp. 1–5 (2017). IEEE
- [5] Moniruzzaman, M., Islam, S.M.S., Bennamoun, M., Lavery, P.: Deep learning on underwater marine object detection: A survey. In: *Advanced Concepts for Intelligent Vision Systems: 18th International Conference, ACIVS 2017, Antwerp, Belgium, September 18–21, 2017, Proceedings 18*, pp. 150–160 (2017). Springer
- [6] Poongundran, M., Agnes, J.S., Das, S., Anusha, D., Amandykova, D., *et al.*: Role of underwater robots in ocean exploration research. In: 2022 International Conference on Applied Artificial Intelligence and Computing (ICAAIC), pp. 1789–1794 (2022). IEEE
- [7] Khan, A.A., Laghari, A.A., Awan, S.A.: Machine learning in computer vision: A review. *EAI Endorsed Transactions on Scalable Information Systems* **8**(32) (2021)
- [8] Chien, C.-F., Chen, Y.-J., Han, Y.-T., Hsieh, M.-K., Lee, C.-M., Shih, T., Wu, M.-Y., Yang, W.-W.: Ai and big data analytics for wafer fab energy saving and chiller

- optimization to empower intelligent manufacturing. In: 2018 e-Manufacturing & Design Collaboration Symposium (eMDC), pp. 1–4 (2018). IEEE
- [9] Jia, J.: A machine vision application for industrial assembly inspection. In: 2009 Second International Conference on Machine Vision, pp. 172–176 (2009). IEEE
 - [10] Hu, F., Gou, H.-y., Yang, H.-z., Yan, H., Ni, Y.-q., Wang, Y.-w.: Automatic paut crack detection and depth identification framework based on inspection robot and deep learning method. *Journal of Infrastructure Intelligence and Resilience* **4**(1), 100113 (2025)
 - [11] Srivastav, M., Mathur, P., Poongodi, T., Sagar, S., Yadav, S.A.: Human emotion detection using open cv. In: 2022 2nd International Conference on Innovative Practices in Technology and Management (ICIPTM), vol. 2, pp. 748–751 (2022). IEEE
 - [12] Hussain, D., Al-Masni, M.A., Aslam, M., Sadeghi-Niaraki, A., Hussain, J., Gu, Y.H., Naqvi, R.A.: Revolutionizing tumor detection and classification in multi-modality imaging based on deep learning approaches: Methods, applications and limitations. *Journal of X-Ray Science and Technology* **32**(4), 857–911 (2024)
 - [13] Zhang, Y., Lv, C.: Tinysegformer: A lightweight visual segmentation model for real-time agricultural pest detection. *Computers and Electronics in Agriculture* **218**, 108740 (2024)
 - [14] Kapadnis, J., Lahare, P., Bhadane, S., Patil, C., Joshi, A., Borase, J.: Implementation of autonomous vehicle using real-time image processing and computer vision algorithm. In: 2025 International Conference on Emerging Smart Computing and Informatics (ESCI), pp. 1–5 (2025). IEEE
 - [15] Liu, M., Dan, J., Lu, Z., Yu, Y., Li, Y., Li, X.: Cm-unet: Hybrid cnn-mamba unet for remote sensing image semantic segmentation. *arXiv preprint arXiv:2405.10530* (2024)
 - [16] González-Sabbagh, S.P., Robles-Kelly, A.: A survey on underwater computer vision. *ACM Computing Surveys* **55**(13s), 1–39 (2023)
 - [17] Er, M.J., Chen, J., Zhang, Y., Gao, W.: Research challenges, recent advances, and popular datasets in deep learning-based underwater marine object detection: A review. *Sensors* **23**(4), 1990 (2023)
 - [18] Uijlings, J.R., Van De Sande, K.E., Gevers, T., Smeulders, A.W.: Selective search for object recognition. *International journal of computer vision* **104**, 154–171 (2013)
 - [19] He, K., Zhang, X., Ren, S., Sun, J.: Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE transactions on pattern analysis and machine intelligence* **37**(9), 1904–1916 (2015)
 - [20] Girshick, R.: Fast r-cnn. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1440–1448 (2015)
 - [21] Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* **28** (2015)
 - [22] 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*, pp. 10781–10790 (2020)
 - [23] 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*, pp. 2117–2125 (2017)
 - [24] 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, pp. 779–788 (2016)
- [25] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., Berg, A.C.: Ssd: Single shot multibox detector. In: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14, pp. 21–37 (2016). Springer
 - [26] Lin, T.-Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2980–2988 (2017)
 - [27] Law, H., Deng, J.: Cornernet: Detecting objects as paired keypoints. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 734–750 (2018)
 - [28] Zhou, X., Wang, D., Krähenbühl, P.: Objects as points. arXiv preprint arXiv:1904.07850 (2019)
 - [29] Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: European Conference on Computer Vision, pp. 213–229 (2020). Springer
 - [30] Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., Liu, Y., Chen, J.: DETRs Beat YOLOs on Real-time Object Detection (2023)
 - [31] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
 - [32] Géron, A.: Concepts, tools, and techniques to build intelligent systems. Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow, 2nd Edition. London, United Kingdom: O’Reilly Media (2017)
 - [33] Sarker, I.: Deep cybersecurity: A comprehensive overview from neural network and deep learning perspective. SN COMPUT. SCI. 2, 154 (2021) (2021)
 - [34] 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, vol. 1, p. (2001). Ieee
 - [35] Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05), vol. 1, pp. 886–893 (2005). Ieee
 - [36] Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008). Ieee
 - [37] Lyu, L., Liu, Y., Xu, X., Yan, P., Zhang, J.: Efp-yolo: a quantitative detection algorithm for marine benthic organisms. Ocean & Coastal Management **243**, 106770 (2023)
 - [38] Xuan, K., Deng, L., Xiao, Y., Wang, P., Li, J.: So-yolov5: Small object recognition algorithm for sea cucumber in complex seabed environment. Fisheries Research **264**, 106710 (2023)
 - [39] Tian, M., Li, X., Kong, S., Wu, L., Yu, J.: A modified yolov4 detection method for a vision-based underwater garbage cleaning robot. Frontiers of Information Technology & Electronic Engineering **23**(8), 1217–1228 (2022)
 - [40] Li, Y., Hu, Z., Zhang, Y., Liu, J., Tu, W., Yu, H.: Ddeyolov9: network for detecting and counting abnormal fish behaviors in complex water environments. Fishes **9**(6), 242 (2024)
 - [41] Zhuo: Underwater treasure object detection algorithm based on improved yolov8.

- [42] 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* **12**(18), 3892 (2023)
- [43] Lin, C., Zhang, Y., Chen, G.: Lightweight underwater debris detection model based on improved rt-detr. *Marine Pollution Bulletin* **222**, 118537 (2026)
- [44] 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 Systems with Applications* **256**, 124978 (2024)
- [45] Wang, Q., Liu, Z.: Fefm-yolo11: Underwater object detection algorithm based on improved yolo11. *Neural Processing Letters* **57**(5), 82 (2025)
- [46] Nguyen, T.: Improve underwater object detection through yolov12 architecture and physics-informed augmentation. *arXiv preprint arXiv:2506.23505* (2025)
- [47] Feng, Z., Liu, F.: Balancing feature symmetry: Ifem-yolov13 for robust underwater object detection under degradation. *Symmetry* **17**(9), 1531 (2025)
- [48] Gupta, S.: EfficientNet-EdgeTPU: Creating Accelerator-Optimized Neural Networks with AutoML. 2020
- [49] 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: 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1–6 (2021). IEEE
- [50] Wang, L., Ye, X., Xing, H., Wang, Z., Li, P.: Yolo nano underwater: A fast and compact object detector for embedded device. In: Global Oceans 2020: Singapore–US Gulf Coast, pp. 1–4 (2020). IEEE