frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 12 August 2025
DOI 10.3389/fmars.2025.1586510

OPEN ACCESS

EDITED BY
Yimian Dai,
Nanjing University of Science and Technology, China

REVIEWED BY
Ying Liang,
Guilin University of Electronic Technology, China
Chunlei Xia,
Chinese Academy of Sciences (CAS), China

*CORRESPONDENCE
Gang Chen
rwb@cug.edu.cn

RECEIVED 03 March 2025
ACCEPTED 30 June 2025
PUBLISHED 12 August 2025

CITATION
Rao W, Chen G, Zhang Y, Cang J, Chen S and Wang C (2025) Benthos-DETR: a high-precision efficient network for benthic organisms detection.
Front. Mar. Sci. 12:1586510.
doi: 10.3389/fmars.2025.1586510

Benthos-DETR: a high-precision efficient network for benthic organisms detection

Weibo Rao[1], Gang Chen[1*], Yifei Zhang[2], Jue Cang[3], Shusen Chen[1] and Chenyang Wang[1]

[1]College of Marine Science and Technology, China University of Geosciences, Wuhan, China, [2]Institute of Surveying and Mapping, Hubei Institute of Water Resources Survey and Design CO., LTD., Wuhan, China, [3]Lhasa Water Resources Survey Hydrology Branch, Tibet Autonomous Region Bureau of Hydrology, Lhasa, China

The intelligent, automated, and high-precision detection of underwater targets represents a challenging yet pivotal issue in marine science. Enhancing the localization accuracy of marine organisms holds significant importance for marine scientific research fields such as ecological conservation and fisheries management, especially in complex seabed environments where accurately identifying benthic organisms characterized by small size, large quantities, and diverse species offers considerable economic benefits and practical value. This study proposes Benthos-DETR, a benthic organisms detection network based on the RT-DETR network. In the backbone of Benthos-DETR network, the Efficient Block with the C2f module reinforces the shallow feature extraction operation in Benthos-DETR, enhancing the algorithm's multi-scale perception. To reduce the computational load and make the algorithm lightweight, a cascaded group attention module has been added to the Benthos-DETR network, it enhances the feature interaction within the same scale. In the neck, the original concatenation module is replaced with the Fusion Focus Module, effectively aggregating feature layer information from different stages of the backbone to achieve cross-scale feature fusion. The proposed Benthos-DETR ensures high target detection accuracy while minimizing hardware requirements for network deployment. The outcomes of the ablation experiment revealed that the various modules introduced in this research optimize the baseline network, and their integration markedly elevates the performance of Benthos-DETR. In tests on an open-source dataset, Benthos-DETR achieved a detection accuracy of 92.1% and mAP_{50} of 91.8% for sea cucumbers, 91.6% accuracy and 92.2% mAP_{50} for sea urchins, and 92.4% accuracy and 93.7% mAP_{50} for scallops. Through a series of experimental analyses, it was evident that the performance of the Benthos-DETR network surpasses existing target detection algorithms, achieving an optimal equilibrium between high recognition precision and a trim network scale.

KEYWORDS

benthic organisms, RT-DETR, attention mechanism, deep learning, underwater target detection

1 Introduction

The economic cost of Marine investigation is high, and the traditional methods employed by scientists to track marine organisms pose certain risks and have a great impact on biological populations (Li et al., 2022). Enhancing the positioning precision of marine organisms holds substantial significance within marine scientific research arenas like ecological conservation and fisheries administration. The intelligent, automated, and high-precision detection of underwater targets is a challenging and critical issue in marine science (Yan et al., 2022; Yu et al., 2022). Therefore, the realization of high-precision underwater biological detection provides scientific support for marine biodiversity conservation and resource management, helping researchers collect long-term and systematic data, analyze the health status of ecosystems, and lay the data foundation for sustainable environmental management decisions (Tamou et al., 2021).

In recent years, research by scientists on underwater biological detection has primarily focused on the following two aspects: In light of the intricate nature of the marine environment and the vast array of marine organism species, certain scholars have tackled the issue by gathering underwater images and handling the data, building various datasets for underwater target detection. These datasets have laid a foundation for underwater biological target detection tasks. For example, Martin et al. established the squid dataset (Martin-Abadal et al., 2020), Wageeh et al. created a dataset of 2000 goldfish images (Wageeh et al., 2021), and Gray et al. developed a marine biological dataset of 326 whale images and 1059 sea turtle images (Gray et al., 2019). Pedersen et al. put together a public dataset of marine organisms, which encompasses 14,518 pictures and includes such marine life as big fish, crabs, squid, shrimp, small fish and starfish, along with 25,613 annotated entries (Pedersen et al., 2019). Ditria et al. carried out a research on target detection by relying on the Mask R-CNN model within the self-constructed Luderick dataset, and the accuracy of intelligent detection surpassed that of both marine fish experts and ordinary citizens during manual detection (Ditria et al., 2020).

Conversely, in response to the diverse requirements of different application scenarios and research objectives, many scholars have conducted a series of optimization and improvement on the target detection algorithm. Alfonso et al. improved the model's generalization capacity with a fish detection approach based on R-CNN. The algorithm used attention to extract key features (Labao and Naval, 2019). Raza and Song improved the YOLO model through incorporating candidate anchor boxes, applying transfer learning and modifying the loss function, which elevated the detection accuracy (Raza and Hong, 2020). Han et al. enhanced underwater images and used a CNN for underwater recognition, achieving notable results (Han et al., 2020). Zhang et al. enhanced the YOLO model's precision by integrating the Swin-Transformer. However, this approach has challenges, including slower detection speeds and a complex model structure (Zhang et al., 2023b).

Presently, under the influence of the marine economic effect, the density of aquatic organisms, fish, sediments and other suspended matter in offshore fisheries has been gradually on the rise. As a result, it becomes challenging for traditional target detection and biometric identification approaches to fulfill the requirements of marine fisheries and ecological management (Ruan et al., 2024; Wang et al., 2024c). Meanwhile, the underwater target recognition environment with large range, multiple types, small targets and complex environment poses a major challenge for computer vision-based underwater target detection practices. Achieving high-precision localization of underwater targets and accurate classification and identification of multiple categories of underwater targets has become a difficult problem (Li et al., 2023a; Xu et al., 2023). In order to tackle the problems of low localization precision and the tendency to have category confusion in target recognition with multi-view underwater images, we put forward an enhanced target detection algorithm named Benthos-DETR, which is particularly devised for complex underwater environments. The primary contributions made by this study can be listed as follows:

1. Inspired by the RT-DETR network, the original backbone of Benthos-DETR network is redesigned, small target feature layer is introduced, and we approve a new network structure (Efficient Feature Extractor), which ensures the fast efficiency of computing and improves the positioning accuracy and recognition accuracy of the underwater target detection task;
2. We redesign the neck part the network. Firstly, on the basis of CGAM module, we modify the original AIFI module of RT-DETR network, which greatly reduces the amount of redundant network calculation; Secondly, a cross-feature fusion module based on attention mechanism (Focus Fusion Module) is proposed, which enhances the feature information flow and greatly improves the recognition effect of Benthos DETR network in benthic organisms detection.
3. On the public dataset EUDD, the coupling efficacy of multiple modules within the Benthos-DETR network was analyzed meticulously via the ablation experiment, and the comparison experiment was carried out by combining multi-class target detection algorithms. The results demonstrate that the Benthos-DETR network proposed by us attains a favorable equilibrium regarding recognition accuracy and network size. Although the network computing cost increases, it yields more accurate results for the detection tasks involving small-sized, large-quantity and multiple-types of marine biological targets.

2 Methodology

Figure 1 depicts the framework of the Benthos-DETR proposed by us. Our approach is based on RT-DETR, which is one of the state-of-the-art end-to-end target detectors (Carion et al., 2020). The RT-DETR network is renowned for balancing speed and accuracy across a variety of tasks (Dai et al., 2024; Lin et al., 2024; Zhao et al., 2024b), includes the backbone, hybrid encoders,

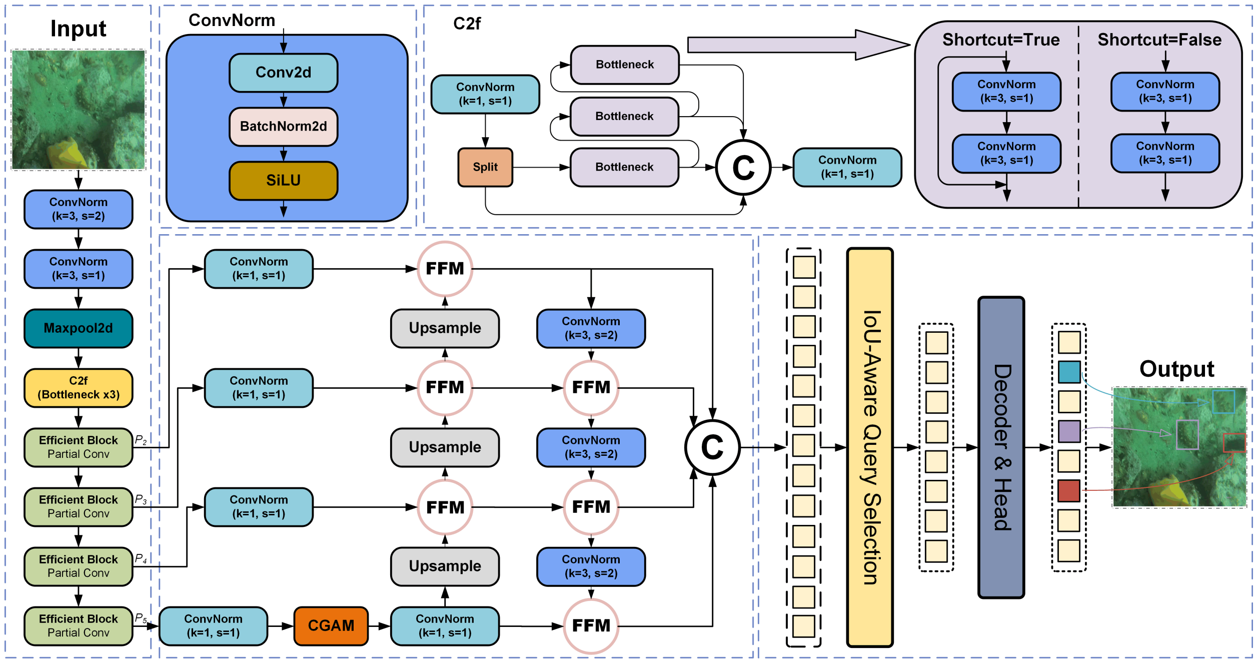

FIGURE 1
Architecture of the proposed Benthos-DETR.

decoders, and predicted the first four major network architectures (Zong et al., 2023; Zhao et al., 2024a). The core innovation of our proposed Benthos-DETR algorithm mainly concentrates on optimizing the backbone and hybrid encoder sections of the architecture, making the network lighter while preserving contextual integrity, and enhancing the accuracy and efficiency of the algorithm for detecting benthic organisms.

Firstly, the backbone network Efficient Feature Extractor ((detailed description in Section 3.1) captures essential information from the input seabed AUV sensor images and generates multi-scale feature maps from the last four stages {P_2, P_3, P_4, P_5}. Among them, the P_2 stage of the network involves shallow features in the image and contains tiny target information for subsea target detection, which is enhanced to ensure a lightweight design while facilitating richer gradient flows.

Secondly, these four-stage feature maps {P_2, P_3, P_4, P_5} are fused through a hybrid encoder that introduces a cascade group attention module, improving feature interaction capabilities at the same scale and reducing the network computational load (detailed description in Section 3.2). In the neck portion of the encoder network, the Fusion Focus Module effectively aggregates feature information from different stages of the backbone to achieve cross-scale feature fusion (detailed description in Section 3.3).

Finally, the comprehensive prediction outcomes generated by the Fusion Focus Module are conveyed to the decoder for prediction. A fixed quantity of image features are selected as the initial queries for the aforementioned decoder through an IoU-aware (Intersection over Union) query selection mechanism (Zhu et al., 2021a; Lv et al., 2024). By utilizing auxiliary headers, the decoder progressively refines the aforementioned queries, thereby generating bounding boxes and associated confidence scores (Zhang et al., 2023a; Wang et al., 2024b).

2.1 Efficient feature extractor

The backbone network of Benthos-DETR is analogous to ResNet (He et al., 2016) and is designated as the Efficient Feature Extractor (abbreviated as EFF). It consists of four stages for data feature processing (as shown in Figure 2). To reduce the influence of downsampling on feature extraction, the initial embedding layer consists of a ConvNorm module with a convolution kernel of 3×3 and a stride of 1, a ConvNorm module with a convolution kernel of 3×3 and a stride of 2, as well as a max pooling layer. The ConvNorm module processes feature maps through a convolution layer, a batch normalization layer and a SiLU activation function (Elfwing et al., 2018; Wang et al., 2021). In Stage 1, the C2f module and Efficient Block reinforce the shallow feature extraction process of underwater image data (Li et al., 2023b), the model's multi-scale sensing ability and outputting characteristic information from the P_2 detection layer (Yu and Zhou, 2023). In subsequent Stages 2, 3, and 4, the Efficient Block module downsamples the input feature maps, enabling the model to capture global information while retaining crucial features (feature information for the P_3, P_4, and P_5 detection layers).

As shown in the dashed box of Figure 2, unlike the residual network design of ResNet, the Efficient Block in the Efficient Feature Extractor consists of a special downsampling residual block and a residual block based on Partical Convolution (Chen et al., 2023). The special downsampling residuals of Efficient Block combine

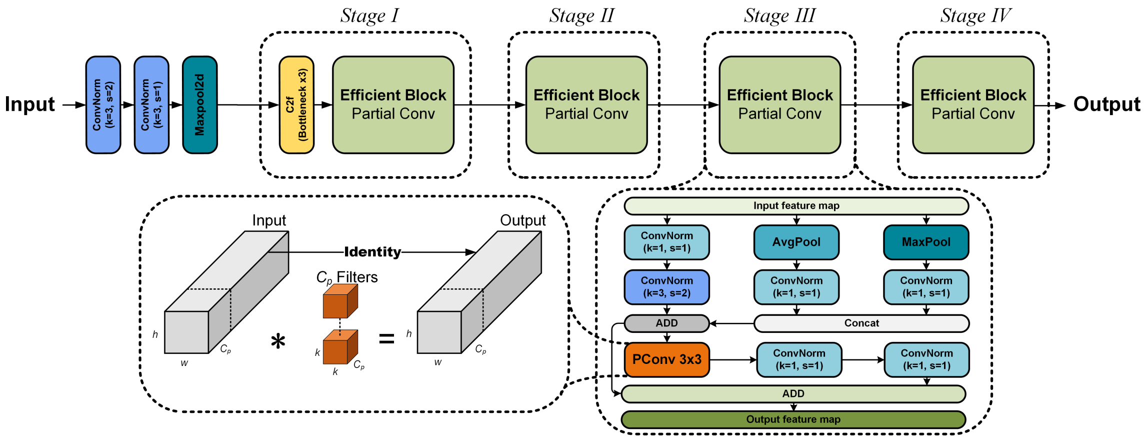

FIGURE 2
Structure of the efficient feature extractor (EFF).

maximum pooling layer and average pooling layer to construct a shortcut connection for spatial downsampling and channel expansion. Additionally, to optimize the traditional convolutional feature extraction process, a convolutional layer with a convolution kernel of 1×1 is employed to decrease the number of channels prior to the downsampling operation. The residual block based on Partical Convolution consists of one PConv layer and two convolution layers with a convolution kernel of 1×1 to construct a residual structure, which replaces the original residual module in ResNet. The PConv layer only conducts convolution operations on a part of the feature map, rather than applying it comprehensively, significantly reducing redundant computation and memory access (Fu et al., 2024; Lu et al., 2024).

The design of the Efficient Block aims to increase computational efficiency while maintaining or even improving model performance, particularly when handling large-scale and complex datasets. The Efficient Feature Extractor, based on Efficient Block, contributes to the construction of more lightweight and efficient deep learning models by reducing superfluous computations and parameters. Even when the depth of the network is increased, it remarkably enhances the feature extraction performance with only a slight increment in the number of parameters in the deep learning model. The model captures complex features of target organisms in underwater images. Stacking convolutional layers expands the range of the receptive field in the backbone network. Overlapping receptive fields compress image information, aiding the acquisition of more comprehensive details (Dumoulin and Visin, 2018). However, during downsampling, spatial information is compressed, which may result in the loss of small object details (Zhou et al., 2015; Gao et al., 2023). To tackle this problem, we have integrated an additional feature information layer, P2, in contrast to the original RTDETR, as illustrated in Figure 3 below.

The P_2 detection layer employs the C2f module to facilitate feature fusion by dividing the input data into two branches (Wang et al., 2023a). One transmits features directly, the other passes through bottleneck modules. This branching design improves the nonlinearity and representation of the network while extracting abstract features from the data (Yang et al., 2024b, 2024a). The two branches are concatenated along the channel dimension to create a feature map with integrated features of different scales. Feature fusion obtains contextual information and high-resolution data (Su et al., 2024; Wang et al., 2024d). This is important for object detection tasks, as it enables the model to identify objects accurately, low-contrast targets, and detailed information. Therefore, adding the C2f module before the output of the P_2 detection layer helps models identify low-contrast targets and detailed information, improving detection of objects and benthic organisms.

2.2 Cascaded grouped attention module

The multi-stage feature layers {P_2, P_3, P_4 and P_5} from the backbone will be fed into the improved encoder. The AIFI in the original RT-DETR is an attention-based multi-head model that increases complexity and parameters (Vaswani et al., 2017), which may affect performance (Zhao et al., 2024b). We have replaced the AIFI module with the Cascade Grouped Attention Module (CGAM), applied to feature layer P_5. CGAM is a key to the framework, integrating grouped attention and cascading to gradually extract key data features. This enhances the model's capacity to understand and process the data, while filtering out irrelevant noise (Liu et al., 2023, 2024). CGAM is especially useful in underwater AUV images, where marine organisms are frequently clustered in complex environments. Figure 4 shows how CGAM works.

CGAM is a flexible and efficient approach that adjusts feature map weights based on input image relevance. This improves the model's understanding of images and detection performance (Liu et al., 2023). In CGAM, the input image is divided into groups of pixels with different meanings. This grouping strategy improves the model's efficiency and allows it to focus on distinctive features. The input sequence is mapped to generate queries, keys and values. CGAM uses grouped attention, with Q, K, and V to calculate attention weights within each set, generating the attention output. This stage adapts the weights of feature maps to focus on important features while suppressing background noise, and improving feature extraction.

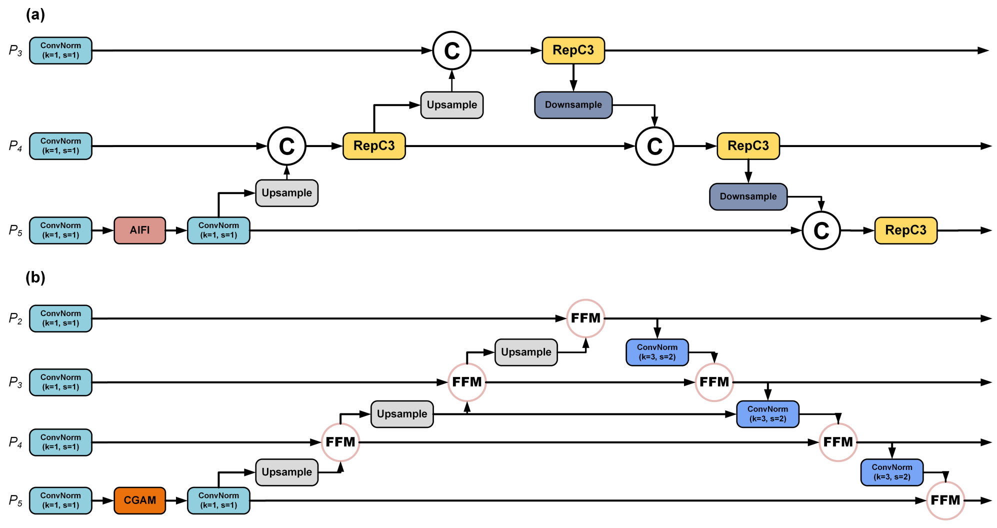

FIGURE 3
(a) Original structure in the RT-DETR; **(b)** Structure with extra P_2 feature layer in the proposed Benthos-DETR.

2.3 Focus fusion module

In this paper, besides the utilization of the CGAM module mentioned in the previous section, the most crucial improvement of the encoder in neck networks is the fusion module between multi-scale feature maps, which is termed the Focus Fusion Module (FFM). The overall structural diagram is shown in Figure 5 below.

The FFM uses spatial and channel attention to extract features from the upper and lower channels. The upper channel of the FFM uses deformable convolutions [DCNv2 (Zhu et al., 2019)] for local context aggregation and spatial feature extraction. To maintain the lightness of the algorithm, add local context to the global context within the attention module. The lower channel of FFM uses convolutions to extract features from adjacent sections of the feature map. Pooling layers achieve channel attention across multiple scales. The fused weighted points are multiplied back into the corresponding feature maps, providing the input for the decoder. The detailed implementation process of Fusion Focus Module is as follows:<I> Processing in the upper channel. The spatial attention formula (S_{att}) of global features at the upper part of FFM is shown in Equation 1. The CBR module extracts features through 1 × 1 convolution, and the DBR module represents the extraction of spatial features of different input path information through deformable convolution:

$$\begin{cases} CBR(X \oplus Y) = \delta(B(Conv_{1*1}^{k=1}(X \oplus Y))) \\ DBR(X \oplus Y) = \delta(B(DCNv2(X \oplus Y))) \\ S_{att} = DBR(DBR(CBR(X \oplus Y))) \end{cases} \tag{1}$$

where, X and Y are feature maps from different path. The symbol $\oplus$ denotes the channel dimension concatenation superposition. The symbol B denotes the BatchNorm layer, while the variable δ represents the ReLU activation function. The convolution layer with a kernel of 1 × 1 is represented by the symbol $Conv_{1*1}^{k=1}(\cdot)$, and $DCNv2(\cdot)$ means deformable convolution layers.

From Equation 1 and Figure 6, deformable convolution layers from DCNv2 at the FFM enhance feature representation and positioning (Wang et al., 2023b). Conventional networks struggle with geometric transformations due to inflexible convolution and pooling layers. This hinders their ability to adaptively detect objects of varying sizes in seabed environments. A deformable convolution layer has been added to enhance the adaptability of feature extraction (Dai et al., 2017). The deformable convolutional kernel allows for an offset at each sampling point, enhancing the model's ability to fit the input data.

$$\begin{cases} CBR(X \oplus Y) = \delta(B(Conv_{1*1}^{k=1}(X \oplus Y))) \\ G(X \oplus Y) = Gap(CBR(X \oplus Y)) \\ C_i(X \oplus Y) = Conv_{1D}^{k=i}(G(X \oplus Y)) \\ C_{att} = CBR(C_{i=3}(X \oplus Y) \oplus C_{i=5}(X \oplus Y) \oplus C_{i=7}(X \oplus Y)) \end{cases} \tag{2}$$

The definition of the symbol in Equation 2 is consistent with that previously provided. The number of channels is reduced to half through 1 × 1 convolution. The *Gap* is the global average pooling layer (Lin et al., 2014a), which inputs the globally averaged feature maps into 1D convolution with kernels of sizes 3, 5, and 7. The superimposition is performed based on the channel dimensions and the channels are restored to their original count through a 1 × 1 convolution.<III>Adding of the upper and lower channels. The broadcast mechanism employed for the purpose of aggregating the spatial and channel attention feature maps. The resulting formula, obtained through the application of a sigmoid activation function, is as follows:

$$w = Sigmoid(S_{att} + C_{att}) \tag{3}$$

In Equation 3, $S_{att} + C_{att}$ represents that the spatial adjustment through the broadcast mechanism is compatible with the channel attention feature map. The two feature maps are added element-wise (Y Adarbah and Ahmad, 2019). This operation integrates

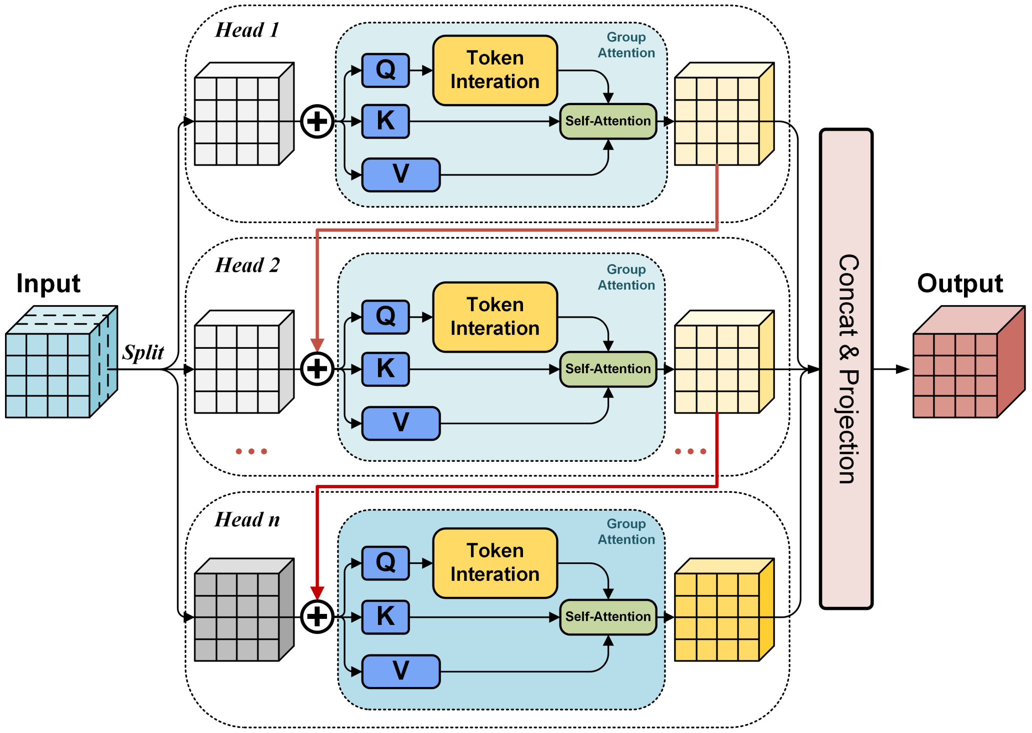

FIGURE 4
Diagram of the cascaded grouped attention module (CGAM).

spatial and channel attention to create a feature map that incorporates both (Ren et al., 2023). The fused map contains both spatial and channel information, allowing for a more comprehensive description of image features.<IV> Weighted output. Applying a sigmoid activation function constrains the output to the 0 to 1 range. The overall FFM computation is shown in Equation 4:

$$Output\ Z = (X \otimes w) \oplus (Y \otimes (1 - w)) \tag{4}$$

In Equation 4, The symbol $\otimes$ denotes element-wise multiplication. The fusion weights w consists of real numbers between 0 and 1, so are the $(1 - w)$, enabling the network to conduct a soft selection or weighted averaging between the feature maps of X and Y (Chen and Kassen, 2020). The attention weights are allocated to the feature maps in a dynamic manner and the resulting outputs are combined along the channel dimension.

```
  input: Multistage features $\{P_2, P_3, P_4, P_5\}$ obtained by
Efficient Feature Extractor
    output: Output of Hybrid Encoder in Benthos-
DETR Network
```

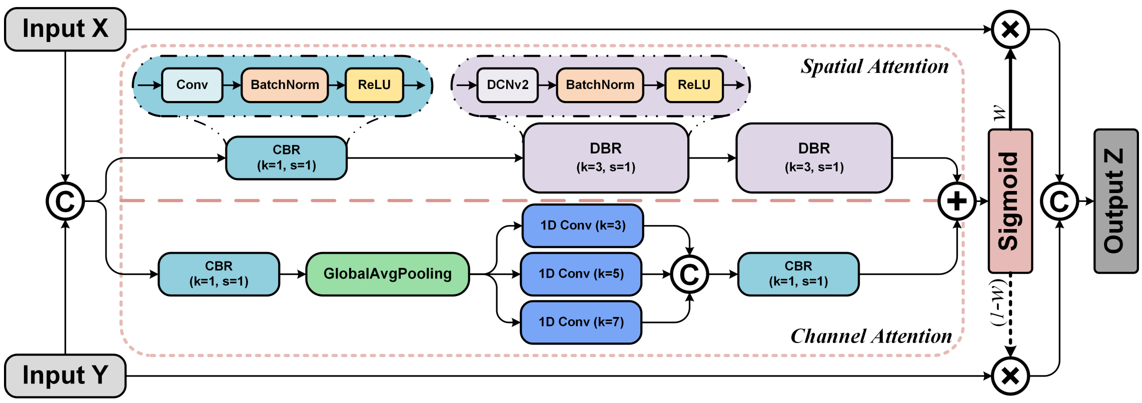

FIGURE 5
Diagram of the fusion focus module (FFM).

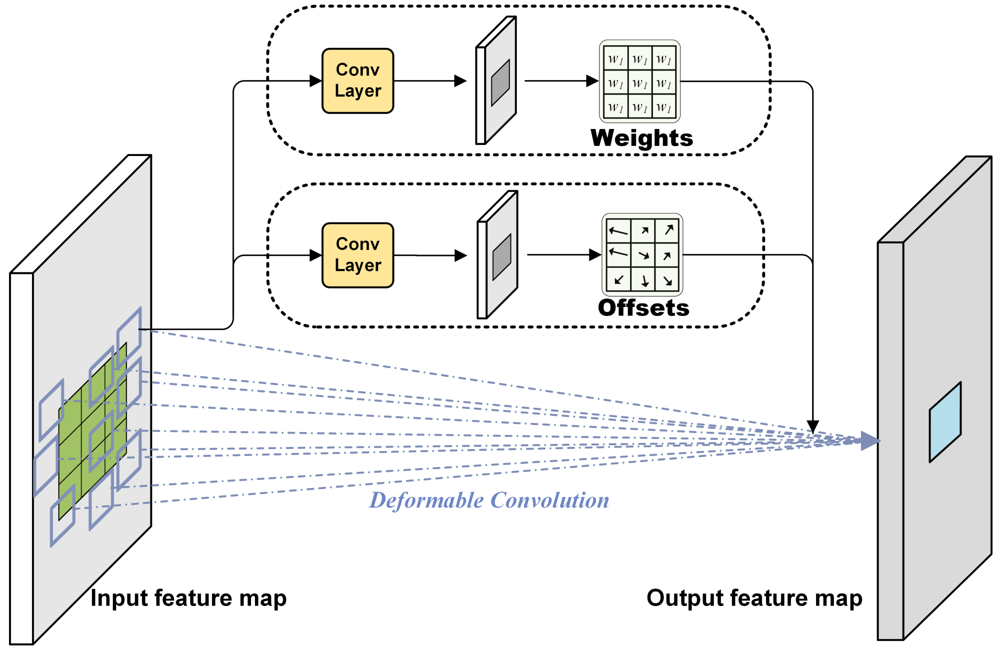

FIGURE 6
Structure of the deformable convolution (DCNv2).<II> Processing in the lower channel. In the following Equation 2, the channel attention formula (C_{att}) for global features is demonstrated, wherein correlation is examined between features at disparate scales via a multitude of one-dimensional convolutions:.

```
Let CGAM denote the Cascaded Grouped Attention Module, FFM denote the Fusion Focus Module, Up denote the Upsample module and Down mean the ConvNorm module;
for each $i \in \{5, 4, 3, 2\}$ do
    $X_i \leftarrow P_i$
    // For the stage 5 feature ($P_5$) of the backbone
    if $i$=5 then
        $X_5 \leftarrow$ CGAM ($P_5$)
    end
    for each $j$ = 1 to 3 do
        $Y_j \leftarrow$ Up($X_{j+1}$)
        $Z_j \leftarrow$ FFM ($X_i$, $Y_j$)
    end
end
$Out_1$ = $Z_3$
$Out_2$ = FFM (Down ($Z_3$), $Z_2$)
$Out_3$= FFM (Down ($Out_2$), $Z_1$)
//Fusion of the Cascaded Grouped Attention Module
$Out_4$ = FFM (Down ($Out_3$), $X_5$)
//The output of the whole Encoder
$Output$ = $Concat$ ($Out_1$, $Out_2$, $Out_3$, $Out_4$)
```

Algorithm 1. Implementation steps of hybrid encoder.

The complete flow of the encoder in the Benthos-DETR is shown in Algorithm 1. As stated in previous papers, detecting very small objects stands out as the key performance bottleneck of state-of-the-art networks (Singh et al., 2018). For example, the difficulty of COCO is largely due to the fact that most object instances are smaller than 1% of the image area (Lin et al., 2014b; Singh and Davis, 2018). Therefore, inspired by SENet (Hu et al., 2018), CBAM (Woo et al., 2018), CA (Hou et al., 2021), and SimAM (Yang et al., 2021) attention modules, we proposed a focus fusion module (FFM), which adds local channel contexts to the global channel-wise statistics. In the encoder of Benthos-DETR network proposed in this paper, FFM replaces the conventional concatenation module and effectively aggregates the feature information from different stage layers of the backbone to achieve cross-scale feature fusion. While ensuring lightweight, focusing on objects with less background clutter, and the recognition ability of small objects has been further improved.

3 Data and parameters

3.1 Underwater object datasets

The submarine small target detection network plays a pivotal role in the underwater picking system deployed in the AUV. It is instrumental in facilitating a range of underwater operations, including rapid positioning, automated monitoring of marine biological growth, and intelligent fishing. To enhance its performance, it is essential to train the network with images captured in actual picking environments. To improve the submarine object recognition task and simulate the real selection environment of Underwater object recognition, the Enhanced Underwater Detection Dataset (EUDD) from the UDD (Liu et al., 2022) based on the real farm image of the open sea was selected in this paper.

The EUDD is obtained from video recordings at two underwater locations approximately 500 meters from Zhangzi Island. The video recording is done by robots and divers working together to follow specific loop routes. The video samples and cuts multiple categories of images according to the uniform number of frames, depending on the sharpness (720P, 1080P, and 4K video), the shooting Angle (head-up, top-down), and the terrain scene (for example, flat, slope, and stone). The finalized underwater open sea farm object detection dataset comprises 2227 original images, categorized into three groups: sea cucumber, sea urchin, and scallop. The original images of the three types of marine organisms are presented below. In Figure 7a, the sea urchins are shown in blue frame lines, while sea cucumbers are shown in red frame lines in Figure 7b, and the scallops are shown in the green frame lines in Figure 7c.

Due to the different economic benefits of seafood in Marine fisheries and the different number of varieties (Wang et al., 2024c), the original UDD has the problem of class imbalance (Chawla et al., 2002; Liu et al., 2022). Poisson GAN is used to balance categories, addressing class imbalance in data sets (Zhu et al., 2017; Deng et al., 2018; Huang et al., 2018). EUDD is introduced as follows: Three categories of underwater organisms are extracted from the UDD and synthesized via Poisson GAN. Each image undergo a specified number of paste operations with probabilities of 0.1, 0.35, 0.30 and 0.25. In each paste operation, Poisson mixing is performed with a probability of 0. The results are included as supplementary material to the EUDD, which contain 18,661 images. The images include 15,615 sea cucumbers, 47,893 sea urchins, and 8,798 scallops, the pie chart of categories is shown in Figure 7d.

Furthermore, the capacity to detect small objects must be significantly enhanced in accordance with the evaluation criteria established by MS COCO (Wu et al., 2020). In MS COCO (Lin et al., 2014b) and PASCAL VOC (Everingham et al., 2010), the number of instances per image is 7.7 and 3, with about 50% of the objects occupying no more than 10% of the image itself, and the other evenly occupying 10% to 100%. Compared to the UDD, EUDD contains an increased proportion of instances of small objects, with a percentage of 3.08% and an average of 12.3 for EUDD in terms of instances per image, as shown in Figure 7e. The resulting EUDD better reflects reality by having more categories and instances, which makes the submarine target detection network evaluation more comprehensive. The detailed comparison is shown in the following Table 1:

3.2 Implementation details

In this study, JPEG images from EUDD ranged from 720 × 405 to 3840 × 2160 pixels. These images are acquired by marine students to ensure data authenticity and usability. The data is divided into three sets: 70% for training, 20% for validation, and 10% for testing. During training, the hyperparameters are set as follows: input image size 640 × 640, batch size 8, epoch 200. The optimizer is AdamW with an initial learning rate of 0.0001 and weight decay of 0.0001. Table 2 shows the specific hyperparameter configurations.

The experimental system environment is shown in Table 3.

3.3 Evaluation metrics

In order to evaluate the effectiveness of the Benthos-DETR in improving the situation, a number of indicators have been introduced (Fisher, 1936; Zheng et al., 2015). The efficacy of the model can be gauged by considering the number of model parameters (Params) and the number of giga floating-point operations per second (GFLOPs). A reduction in parameters and GFLOPs results in the creation of a more straightforward model. The precision (P), recall (R), and mean average precision (mAP) are used to assess detectors. Precision is the proportion of correctly identified positive samples, while recall is the ratio of actual to predicted positive samples. The following definitions are provided for clarity:

$$\begin{cases} Precision = \frac{TP}{TP+FP} \\ Recall = \frac{TP}{TP+FN} \\ mAP = \frac{1}{n}\sum_{i=1}^{n} AP_i \end{cases} \tag{5}$$

In Equations 5, "true positive" (*TP*) denotes samples correctly identified as positive, and "true negative" (*TN*) samples correctly identified as negative. The "false positive" (*FP*) is a sample incorrectly classified as positive, while "false negative" (*FN*) is a sample incorrectly classified as negative. Figure 8 illustrates a visual representation of those relationships.

Specifically, mAP_{50} and $\mathrm{mAP}_{50:95}$ are used to evaluate the precision of target detection, with higher values denoting greater accuracy. mAP_{50} is formed by precision and recall. Area under P-R curve (Precision-Recall curve) for $\mathrm{mAP}_{50:95}$ is calculated by dividing it into 10 IoU thresholds (0.5 to 0.05 to 0.95) and averaging the results. FPS shows the number of images detected per second, indicating detection speed:

$$FPS = S/T \tag{6}$$

In Equations 6, *S* is the count of samples, and *T* is the required processing time.

4 Experiment and results

4.1 Ablation experiment

In this paper, we evaluated the efficacy of each module in the Benthos-DETR using the EUDD dataset. The baseline model was RT-DETR-r18. To achieve high-precision recognition of underwater objects, we made a series of improvements to the original network: (1) The backbone network had been enhanced to become an efficient feature extractor, replacing the basic blocks with efficient blocks and producing an additional feature layer of P_2 while maintaining network computing efficiency; (2) In the neck network that processes features extracted from the backbone, the

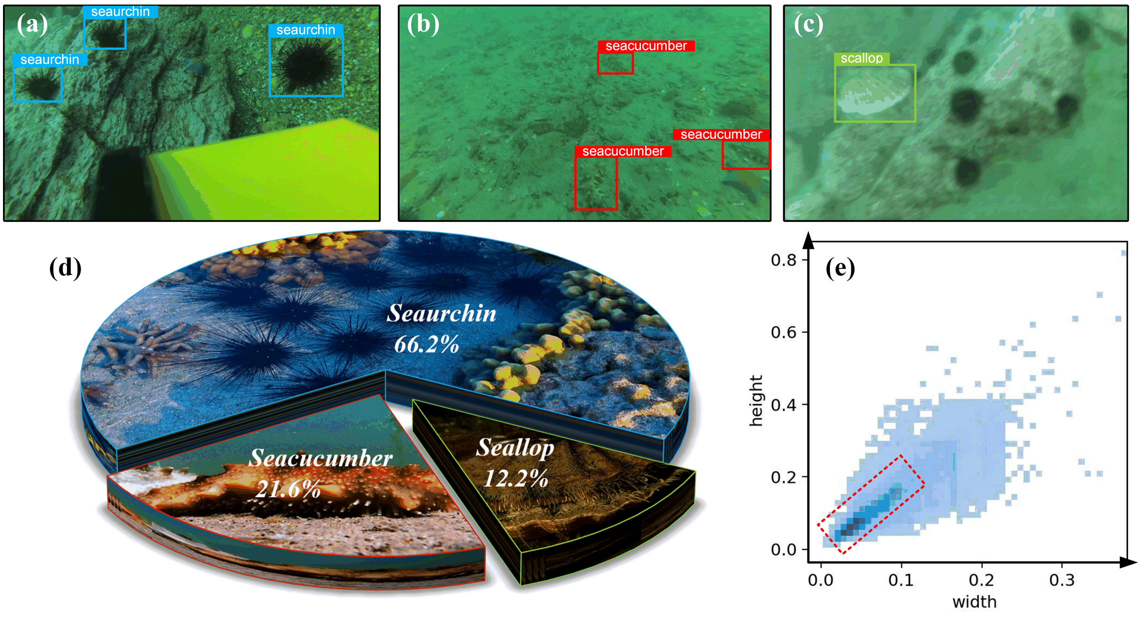

FIGURE 7
(a-e) Overview of the enhanced underwater detection dataset (EUDD).

Cascaded Grouped Attention Module had been introduced to replace the AIFI module in the original RT-DETR, providing a lightweight improvement to the feature layer of P_5; (3) In the neck feature hybrid network of the Benthos-DETR network, the concatenation module was further optimized by cross-feature attention mechanism, strengthening the feature perception effect of Benthos DETR network on multi-scale, complex scenes and tiny targets during underwater object recognition.

Table 4 presented the results of the ablation experiments conducted on the three main improved modules of Benthos-DETR. EFF referred to the Efficient Feature Extractor, which forms the backbone network. CGAM was the Cascaded Grouped Attention Module, which was applied to the feature layer of P_5. FFM standed for Focus Fusion Module, which was used in conjunction with the neck feature hybrid network. mAP was a metric used in object detection. It assessed the accuracy of detection across multiple categories. Parameters indicated the number of network parameters, and GFLOPs measured network complexity.

By comparing Group 1 (Baseline) with Group 2 (Baseline + EFF), and Group 3 (Baseline + FFM), we could observe the significant roles played by the proposed modules in enhancing network performance and reducing complexity. When EFF replaced the original RT-DETR backbone, the network recognition accuracy improved, with the mAP value rising to 91.5%. However, due to the additional computation for the feature layer of P_2, the network parameters increased from the 19.9M to 22.5M, and the GFLOPs also increased from 57.3 to 65.2. CGAM has a more pronounced impact on lightweight networks. By replacing the original AIFI module in RT-DETR, the network

TABLE 1 Comparisons of different object detection datasets.

Dataset	Primary Resolution	Ins./Image	Ins./Percentage	Year
PASCAL VOC (Everingham et al., 2010)	*500 * 375*	3	12.35%	2012
MS COCO (Lin et al., 2014b)	*640 * 480*	7.7	5.67%	2014
URPC 2017 (Ruan et al., 2024)	*720 * 405*	9.3	0.73%	2017
URPC 2018 (Ruan et al., 2024)	*1920 * 1080*	8.2	0.58%	2018
UDD (Liu et al., 2022)	*3840 * 2160*	10	0.47%	2020
EUDD (Our dataset)	***Variety*** ***from720 * 405 to 3840 * 2160***	**12.3**	**3.08%**	**2024**

The bold values indicate the results from our models.

TABLE 2 Hyperparameter settings of network training.

Epochs	Batch Size	Image Size	Optimizer	Learning rate	Weight decay
200	*8*	*640 * 640*	*Adam with Weight Decay Correction*	*0.0001*	*0.0001*

parameters decreased from 19.9M to 14.6M, and GFLOPS also dropped from 57.3 to 43.5. However, this change also affects the network's recognition accuracy, with the mAP value decreasing from 88.5% to 83.9%. Compared to the first two modules, the introduction of FFM in the neck network achieved a more balanced result. With only a slight increase in network parameters, the mAP value increased from 88.5% to 89.7%, indicating that FFM could effectively combined network performance improvement with model lightweight.

It was worth noting that, as shown in Group 5 to Group 8 in Table 4, combining modules yielded better results than the original baseline. To visually present the ablation experimental results, we had plotted a comparative statistical graph of ablation experiments, as shown in Figure 9. Two types of indicators were selected as representatives: the left y-axis represented the mAP value, which measured model accuracy, denoted by a rose-red line; y-axis represented the GFLOPs value, which indicated model complexity, denoted by gray rectangles.

It could be observed that the combination of multiple modules produced a more pronounced effect. The addition of both EFF and CGAM to the baseline model resulted in an increase in mAP from 88.5% to 91.1%, accompanied by a reduction in network GLOPs from 57.3 to 54.2. At this juncture, the network demonstrated enhanced precision in object detection while retaining its lightweight configuration. Ultimately, the network Benthos-DETR, which combined all three modules, achieved the highest object detection result (highlighted in red on the right side of Figure 9). Compared to Group 7 (Baseline + EFF + FFM), the GFLOPs decreased from 67.2 to 62.3. Although the complexity of the Benthos-DETR network, compared with the baseline model (highlighted in blue on the left side of Figure 9), increased from 57.3 to 62.3 in terms of GFLOPs, the network performance increased by 4.7%, meeting the requirement of high-precision detection in the task of benthic organisms detection.

TABLE 3 Experimental system environment.

Configuration	Parameters
Operating System	*Ubuntu 16.04*
Programming Language	*Python 3.9*
CPU	*12th Gen Intel(R) Core(TM) i7-12700K 3.60 GHz*
GPU	*GeForce RTX 3090*
GPU Memory	*24 G*
CUDA	*12.0*
cuDNN	*11.7*
RAM	*64 G*
Algorithm Framework	*Pytorch-2.0.1+cu117*
IDM	*Spyder 3.3.0*

4.2 Analysis of detection

The ablation experiments demonstrated that the Benthos-DETR network exhibited a notable improvement in underwater target detection performance compared to the RT-DETR network. Despite a slight increase in network complexity due to the addition of P_2 feature layer in the Efficient Feature Extractor and the introduction of Focus Fusion Module in the neck part of network, the enhanced feature extraction capability and stronger feature information flow laid a solid foundation for potential future improvements. In this section, we would showcase the effectiveness of the proposed Benthos-DETR in the actual seabed benthic organisms detection, and conduct a detailed analysis of the network optimization effects through comparison experiments with the original RT-DETR network.

Following the application of predefined hyperparameters to the training process, the recognition results of the Benthos-DETR network on the validation dataset were presented in Figure 10. In Figure 10a, the red bounding boxes represent sea urchins, and the numbers on the boxes indicate the confidence scores of the detections. The blue bounding boxes in Figure 10b represent sea

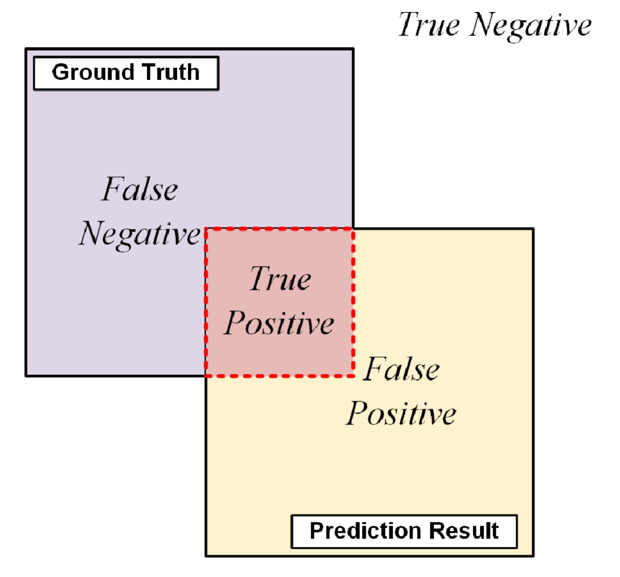

FIGURE 8
Sample relationship chart.

TABLE 4 Ablation experiments.

Groups	EFF	CGAM	FFM	mAP/%	Parameters/M	GFLOPs
1				88.5	19.9	57.3
2	✓			91.5	22.5	65.2
3		✓		83.9	14.6	43.5
4			✓	89.7	20.7	59.1
5	✓	✓		91.1	18.6	54.2
6		✓	✓	84.9	17.3	49.7
7	✓		✓	92.2	23.4	67.2
8	✓	✓	✓	92.7	21.8	62.3

The symbols "✓" indicates that the components of the current ablation experiment include the checked modules.

cucumbers, and the green bounding boxes represent scallops, as shown in Figure 10c. Due to the complex biological situation on the seabed, there are large clusters of organisms, as shown in Figure 10d. In cases where recognition results were located at the edges of the image or are densely overlapping, the bounding box colors served as the primary means of distinction, and only the recognition confidence was displayed on the boxes, with the specific label names being omitted for clarity. As can be seen from Figure 10, the proposed Benthos-DETR network could obtain relatively accurate results for seabed benthic organisms detection tasks with complex conditions, multiple categories and tiny targets. However, a comprehensive analysis of recognition accuracy should also consider the results of training and evaluations of test set.

Figure 11 below showed the network training outcomes. Figure 11a showed the confusion matrix of the Benthos-DETR network's detection results. The matrix showed that the network often failed to detect sea urchins and scallops due to their light colors and background mimicry. As shown in Figure 10, sea urchins, which had a spherical body shape and were mostly black in color, had the highest recall rate of 89% when detected by the Benthos-DETR network. However, due to the cluster distribution of black sea urchins and their similarity to complex backgrounds such as underwater holes or gaps, the probability of the background being misclassified as sea urchins was 58% during testing, which was higher than the probability of the background being incorrectly identified as one of the other two categories. The accuracy of the Benthos-DETR network in identifying three types of seabed benthic organisms was shown in Figure 11b through the P-R curve. The zoomed-in area was highlighted with an orange box line. During the object detection process for the EUDD, the Benthos-DETR network

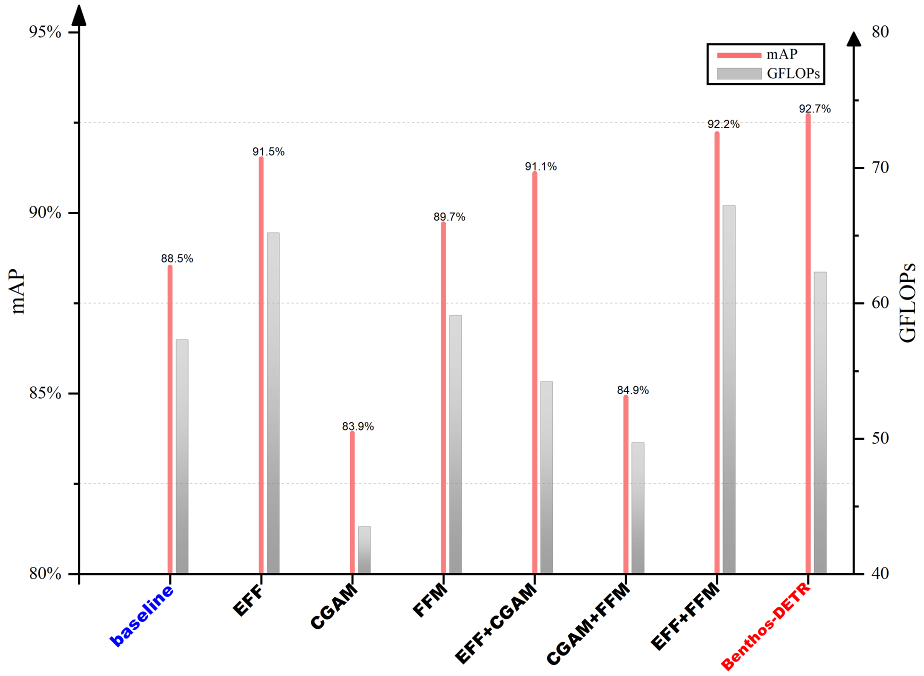

FIGURE 9
Comparative chart of ablation experiments.

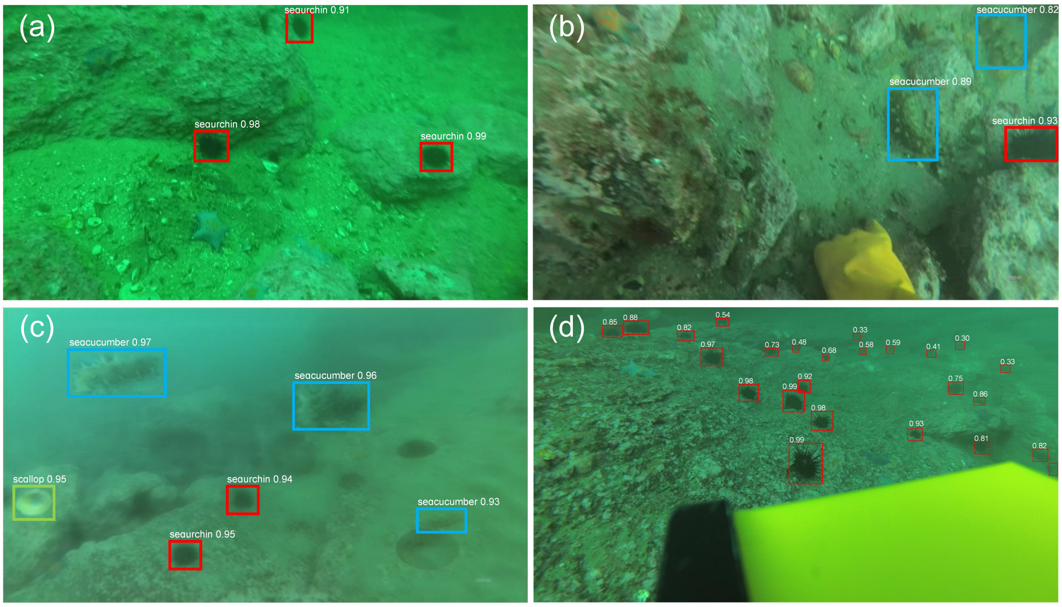

FIGURE 10
(a-d) The recognition results of Benthos-DETR on EUDD dataset.

achieved the highest mAP$_{50}$ value of 93.7% for scallops, 92.2% for sea cucumbers, and 91.8% for sea urchins. From Figure 11c, it can be found that the Benthos-DETR network not only performs well in detecting sea cucumbers and scallops under complex background interference but also excels in detecting sea urchins in large numbers and clusters. Although the detection accuracy for seabed benthic organisms could not reach the level of scallops, the overall mAP$_{50}$ value for all categories combined still reached an impressive 92.7%. The comprehensive statistical analysis of the network recognition outcomes and accuracy is presented in Table 5 below.

Table 5 presents the test results of the Benthos-DETR and RT-DETR networks on an underwater object detection task, with input data from the test set partitioned by EUDD. The detected number of sea urchins was greater than the sum of sea cucumbers and scallops, which aligned well with the actual species distribution. Compared with the RT-DETR network, Benthos DETR achieved higher accuracy in the identification of three types of seabed benthic organisms. However, in the detection of sea cucumbers, RT-DETR identified more instances and images than Benthos-DETR, with a higher recall rate. Nevertheless, the recognition accuracy of RT-DETR significantly lagged behind Benthos-DETR. In the detection of sea urchins, which were numerous in number and small in size, Benthos-DETR demonstrated its superior accuracy by identifying more sea urchin instances from fewer images, with both precision and recall rates surpassing those of RT-DETR. For the detection accuracy of the three underwater organism categories, both networks exhibited the highest mAP$_{50}$ for scallops, which was related to the biological attributes of the underwater shell characteristics.

The Benthos-DETR network, as demonstrated in this paper, not only achieved network lightweighting but also improved the accuracy of target recognition compared to the RT-DETR network before optimization. This was based on the recognition performance of the benthic organisms in the EUDD dataset. Specifically, the Precision had increased from 89.2% to 91.4%, the Recall had risen from 85.2% to 87.1%, and the mAP$_{50}$ had improved from 88.5% to 92.7%. The network recognition accuracy had been enhanced by 4.7%. The detailed comparison statistics were shown in Table 6 below.

4.3 Comparison experiment

In this chapter, a comparative analysis of the proposed Benthos-DETR network is conducted alongside other target detection algorithms, including both qualitative assessments and quantitative metrics. The algorithms involved in the comparison encompassed classic two-stage algorithms such as Faster R-CNN (Ren et al., 2017), Cascade R-CNN (Cai and Vasconcelos, 2018), TOOD (Akyon et al., 2022), and Retina Net (Lin et al., 2020). Additionally, multiple versions of the single-stage target detection YOLO algorithm were included, such as YOLOv5 (Jocher, 2020), YOLOv8 (Jocher et al., 2023), and YOLOv10 (Wang et al., 2024a). The following example in Figure 11 demonstrated the target detection capabilities of different algorithms on the underwater target detection dataset EUDD.

Figure 12 was a representative visual example, showing the visual detection results of our Benthos DETR compared to other advanced target detection networks. The labels in the bottom right corner of each subfigure indicated the names of the respective target detection networks. The "origin" image displayed the ground truth target detection labels annotated by professional marine science researchers sea urchin. By comparing the detection results of various models with the actual distribution of seabed benthic

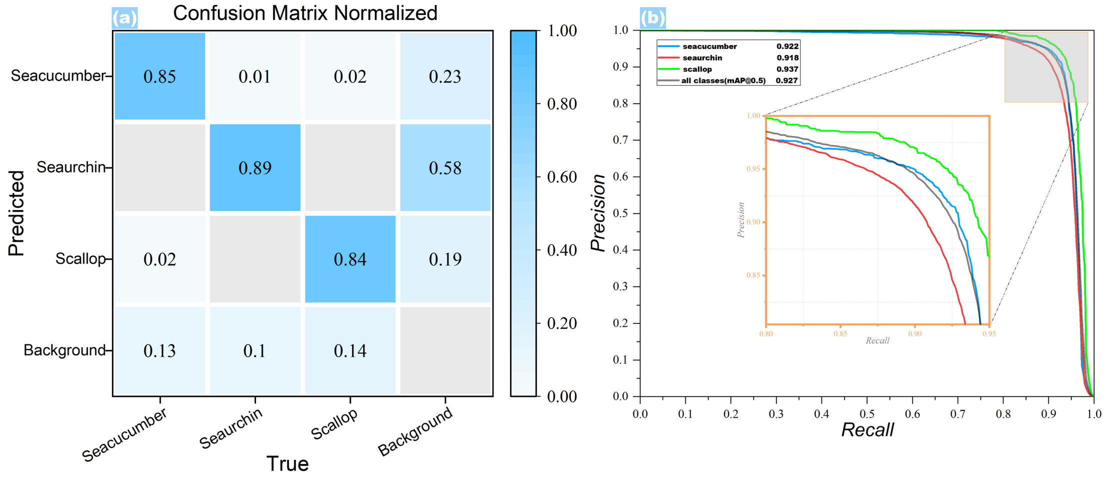

FIGURE 11
The results of the Benthos-DETR network training. **(a)** Normalized confusion matrix; **(b)** Precision-Recall curve.

organisms, we could qualitatively assess the practicability and effectiveness of the target detection algorithm. The comparative images in Figure 12 highlighted the detection accuracy of the Benthos-DETR network in challenging underwater scenarios with multiple types, tiny targets, and a large number of objects. The proposed Benthos-DETR network was able to accurately identify the types of targets and precisely locate their positions, avoiding interference from complex environments. Rigorous quantitative analysis requires the participation of more network evaluation metrics, and a detailed summary of accuracy metrics from various network comparison experiments was provided in Table 7 below.

According to Table 7, the Benthos-DETR network outperformed the two-stage algorithms in terms of computational cost and detection speed, achieving an mAP$_{50}$ of 92.7%. Although it did not match the real-time detection speed of single-stage YOLO algorithms, its accuracy had seen a notable improvement. In particular, the recognition accuracy of the proposed Benthos-DETR network had increased from 79.7% (YOLOv5), 83.5% (YOLOv8) and 86.3% (YOLOv10) to 91.4%. When compared to the DETR and RT-DETR algorithms, RT-DETR showed a 15.5% improvement in accuracy over DETR, while Benthos-DETR demonstrated an 18.4% enhancement. Furthermore, Benthos-DETR achieved the highest mAP$_{50:95}$ among the comparative experiments, reaching 75.2%. Despite the proposed Benthos-DETR network implementing a series of enhancements to the backbone and neck components of the RT-DETR network, with the objective of enhancing the accuracy of target detection, this inevitably resulted in an increase in the amount of network computation. However, these changes in network complexity were deemed to be worthwhile for underwater target detection tasks. Our GFLOPs reached 60.5, higher than some lightweight models but much lower than computationally intensive ones, such as Cascade R-CNN (184.3GFLOPs) and TOOD (232.8GFLOPs). The moderate performance, computational cost and model size (20.8M parameters) of the Benthos-DETR network represented an optimal balance between performance and efficiency, facilitating effective training and deployment of the algorithm. In summary, the benthos-DETR network proposed in this paper was capable of effectively identifying and accurately locating a multitude of categories, in considerable quantities, and of a diminutive size, of seabed benthic organisms in complex underwater environments. The network contributed to advancing underwater target detection

TABLE 5 The detection results of Benthos-DETR and RT-DETR on EUDD dataset.

Class	Groups	Images	Instances	Precision/%	Recall/%	mAP$_{50}$/%	mAP$_{50:95}$/%
sea cucumber	RT-DETR	953	2360	88.1	86.2	89.3	75.2
	Benthos-DETR	937	2293	92.1	85.6	91.8	74.3
sea urchinsea urchin	RT-DETR	1782	4456	88.9	84.3	87.1	71.2
	Benthos-DETR	1773	4631	91.6	89.0	92.2	75.9
scallop	RT-DETR	618	718	91.3	85.3	90.4	77.5
	Benthos-DETR	620	731	92.4	84.6	93.7	78.1

TABLE 6 The detection comparison between Benthos-DETR and RT-DETR.

Groups	Precision/%	Recall/%	mAP_{50}/%	$mAP_{50:95}$/%	Parameters/M	GFLOPs
RT-DETR	89.2	85.2	88.5	74.1	19.9	57.3
Benthos-DETR	91.4	87.1	92.7	75.2	20.8	60.5
Performance	↑ 2.4%	↑ 2.2%	↑ 4.7%	↑ 1.4%	↑ 4.5%	↑ 5.6%

The symbol "↑ " means the improvement of the experimental results.

tasks and provided a reliable solution for target detection in actual complex marine scenes.

5 Discussion

This paper used heatmaps to demonstrate the effectiveness of feature utilization in the Benthos-DETR network, as shown in Figure 13 below. The first column of images in Figure 12 shown the original images fed into the network, showcasing diverse benthic organisms across environments. The second column displayed the feature heatmaps of the RT-DETR target detection network. The original RT-DETR network focused on the background of recognition images because the P_2 feature layer was ignored. The feature information from the feature layer of P_5 had a significant impact on the network's recognition heatmap, which inadvertently diminished the focus on small targets such as sea urchins, sea cucumbers, cave entrances, and underwater crevices. Consequently, the network's recognition accuracy for these underwater small targets was somewhat lacking. The third column of images in Figure 12 showed the feature heatmaps of the Benthos-DETR target detection network. The Benthos-DETR network's capacity to discern the characteristics of seabed benthic organisms has been enhanced by the incorporation of a multi-path attention mechanism and data from the P_2 feature layer. The recognition features captured by the network were more detailed. The heat map was also clearer. Therefore, the optimized Benthos-DETR network could capture more detailed features and was more discernible in complex underwater environments, achieving superior results in target detection.

6 Conclusion

This study proposed the Benthos-DETR network as an extension of the RT-DETR network, with the objective of detecting seabed benthic organisms. Firstly, in the backbone of Benthos-DETR network, the C2f module and Efficient Block were used to enhance the shallow feature extraction process of data, improving the model's multi-scale perception capabilities. Secondly, to reduce the computational load of the network and achieve algorithmic lightweight, a cascaded group attention module was introduced into the encoder of the Benthos-DETR network, enhancing feature interaction at the same scale. Finally, in the neck part of the network encoder, the original concatenation module was replaced with the Fusion Focus Module, effectively aggregating feature layer information from different stages of the backbone to achieve cross-scale feature fusion. Those improvements of the proposed Benthos-DETR network ensure high performance in target detection accuracy while minimizing the hardware requirements for network deployment.

Through a series of experimental analyzed in this paper, the Benthos-DETR network demonstrated superior performance

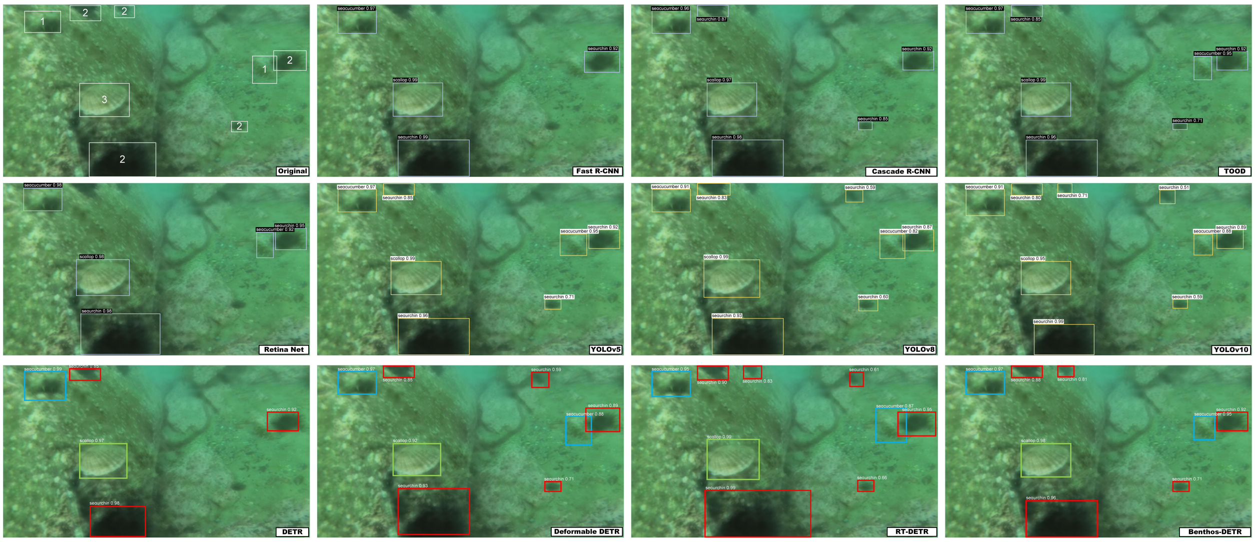

FIGURE 12
A representative visual example of several target detection networks.

TABLE 7 Comparison with target detection networks on the EUDD dataset.

Methods	Reference	Precision /%	Recall /%	mAP$_{50}$ /%	mAP$_{50:95}$ /%	Params /M	GFLOPS	FPS
Faster R-CNN	(Ren et al., 2017)	64.8	59.5	63.6	43.3	41.3	251.4	38
Cascade R-CNN	(Cai and Vasconcelos, 2018)	62.3	60.7	66.2	47.5	37.6	184.3	13
TOOD	(Akyon et al., 2022)	70.3	68.7	73.4	56.6	38.8	232.8	34
Retina Net	(Lin et al., 2020)	59.2	54.5	58.9	39.2	34.2	152.3	36
YOLOv5	(Jocher, 2020)	79.7	76.8	80.1	60.3	21.2	49.0	235
YOLOv8	(Jocher et al., 2023)	83.5	79.2	85.7	63.4	25.9	78.9	223
YOLOv10	(Wang et al., 2024a)	86.3	88.3	89.0	68.0	15.4	59.1	220
DETR	(Carion et al., 2020)	77.2	73.1	78.4	59.2	14.3	44.3	59
Deformable DETR	(Zhu et al., 2021b)	82.1	79.2	84.7	69.2	13.2	41.5	41
RT-DETR	(Zhao et al., 2024b)	89.2	85.2	88.5	74.1	19.9	57.3	105
Benthos-DETR	**Our Research**	**91.4**	**87.1**	**92.7**	**75.2**	**20.8**	**60.5**	**84**

The bold values indicate the results from our models.

compared to several existing object detection algorithms. The results of the ablation experiment demonstrated that the multiple modules have a beneficial effect on the performance of the baseline network. Furthermore, the integration of these modules had led to a notable enhancement in the network performance of Benthos-DETR. In tests conducted on the EUDD dataset, the Benthos-DETR network achieves a detection accuracy of 92.1% and mAP$_{50}$ of 91.8% for sea cucumbers, 91.6% accuracy and 92.2% mAP$_{50}$ for

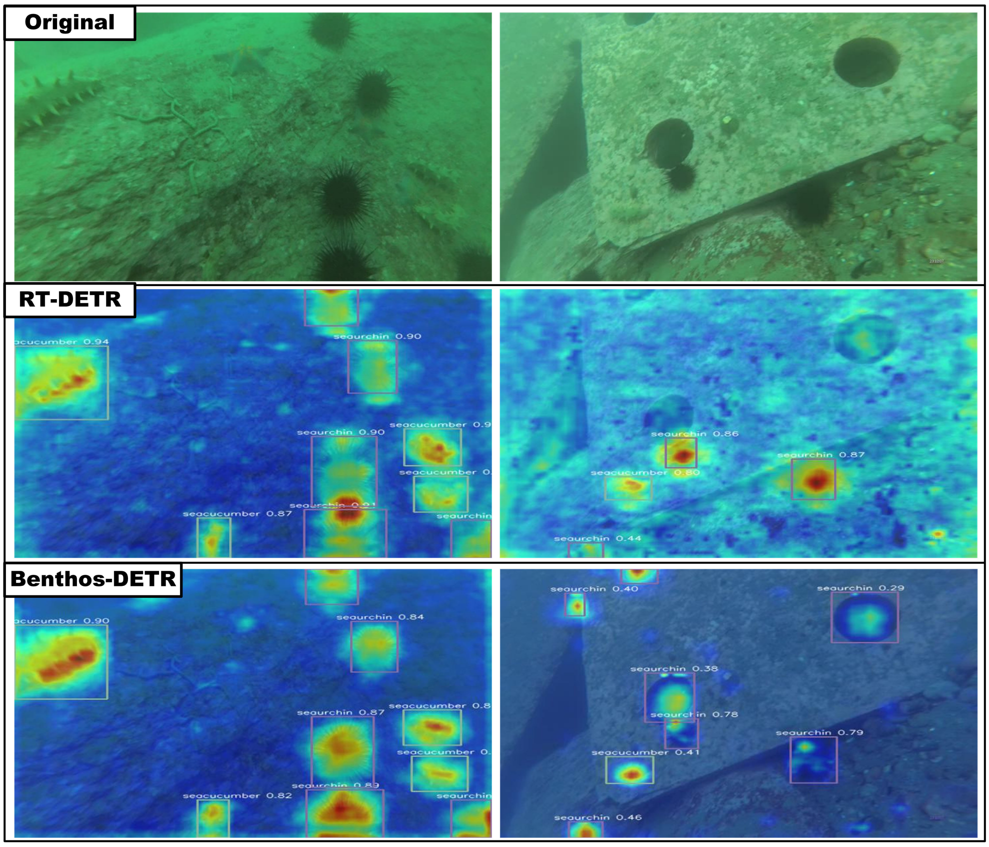

FIGURE 13
Comparative experiment for object detection analysis.

sea urchins, and 92.4% accuracy and 93.7% mAP_{50} for scallops. Combining the detection accuracy results for these three types of underwater biological targets, Benthos-DETR achieved an overall mAP_{50} of 92.7%, representing a 4.7% improvement in mAP_{50} compared to the RT-DETR network. A comprehensive comparison with alternative object recognition algorithms demonstrated that the proposed algorithm struck an optimal balance between recognition accuracy and network size. Despite the increased computational cost of the network, higher accuracy metrics were achieved in tasks involving the detection of small, numerous, and diverse underwater objects. In the future, the variety of underwater targets for detection will be expanded, with the incorporation of additional species that are both dynamically active and widely distributed, as part of the network training process. Concurrently, a series of lightweight algorithms will be developed to achieve high-precision and real-time underwater target detection, while maintaining high-precision target detection. These algorithms will provide technical support and algorithmic reference for research fields such as marine fisheries management, marine ecological protection, and marine biological surveys, among others.

Data availability statement

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

Author contributions

WR: Conceptualization, Writing – original draft. GC: Validation, Writing – review & editing. YZ: Funding acquisition, Investigation, Writing – review & editing. JC: Data curation, Writing – review & editing. SC: Supervision, Writing – review & editing. CW: Visualization, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The research is supported by the Scientific Research Program Project of Hubei Provincial Department of Natural Resources (grant no. ZRZY2024KJ03), the National Natural Science Foundation of China (grant no. 42104024), the Natural Science Foundation of Jiangxi Province (grant no. 20242BAB20126), the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (grant no. CUGL200805), and the National Natural Science Foundation of China under Grant 42101390.

Acknowledgments

We thank Rowan John from University of British Columbia for a university educational license for Visio Pro, provided to us, for network visualization and interpretation. We thank Michel J. from the University of Lausanne, Yuyang Ye from Wuhan University and Sui R. from Colorado School of Mines for their helpful and insightful embellishments on the research.

Conflict of interest

Author YZ was employed by the company Institute of Surveying and Mapping, Hubei Institute of Water Resources Survey and Design CO., LTD.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Akyon, F. C., Altinuc, S. O., and Temizel, A. (2022). "Slicing aided hyper inference and fine-tuning for small object detection," in 2022 IEEE International Conference on Image Processing, ICIP 2022, Bordeaux, France, 16–19 October *2022* (IEEE), 966–970. doi: 10.1109/ICIP46576.2022.9897990

Cai, Z., and Vasconcelos, N. (2018). "Cascade R-CNN: delving into high quality object detection," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22*, 2018* (Computer Vision Foundation/IEEE Computer Society), 6154–6162. doi: 10.1109/CVPR.2018.00644

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in Computer Vision – ECCV *2020*. Eds. A. Vedaldi, H. Bischof, T. Brox and J.-M. Frahm (Springer International Publishing, Cham), 213–229. doi: 10.1007/978-3-030-58452-8_13

Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *J. Artif. Intell. Res.* 16, 321–357. doi: 10.1613/JAIR.953

Chen, J., Kao, S., He, H., Zhuo, W., Wen, S., Lee, C.-H., et al. (2023). "Run, don't walk: chasing higher FLOPS for faster neural networks," in *2023 IEEE/CVF Conference* *on Computer Vision and Pattern Recognition* (CVPR), 12021–12031. doi: 10.1109/CVPR52729.2023.01157

Chen, P., and Kassen, R. (2020). The evolution and fate of diversity under hard and soft selection. *Proc. R. Soc. B: Biol. Sci.* 287, 20201111. doi: 10.1098/rspb.2020.1111

Dai, J., Qi, H., Xiong, Y., Li, Y., Zhang, G., Hu, H., et al. (2017). "Deformable convolutional networks," in *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017* (IEEE Computer Society), 764–773. doi: 10.1109/ICCV.2017.89

Dai, L., Wang, D., Song, F., and Yang, H. (2024). "Concrete bridge crack detection method based on an improved RT-DETR model," in *2024 3rd International Conference on Robotics, Artificial Intelligence and Intelligent Control (RAIIC)*, 172–175. doi: 10.1109/RAIIC61787.2024.10670904

Deng, W., Zheng, L., Ye, Q., Kang, G., Yang, Y., and Jiao, J. (2018). "Image-image domain adaptation with preserved self-similarity and domain-dissimilarity for person re-identification," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22*, 2018* (Computer Vision Foundation/IEEE Computer Society), 994–1003. doi: 10.1109/CVPR.2018.00110

Ditria, E. M., Lopez-Marcano, S., Sievers, M., Jinks, E. L., Brown, C. J., and Connolly, R. M. (2020). Automating the analysis of fish abundance using object detection: optimizing animal ecology with deep learning. *Front. Mar. Sci.* 7. doi: 10.3389/fmars.2020.00429

Dumoulin, V., and Visin, F. (2018). A guide to convolution arithmetic for deep learning. doi: 10.48550/arXiv.1603.07285

Elfwing, S., Uchibe, E., and Doya, K. (2018). Sigmoid-weighted linear units for neural network function approximation in reinforcement learning. *Neural Networks* 107, 3–11. doi: 10.1016/j.neunet.2017.12.012

Everingham, M., Gool, L. V., Williams, C. K. I., Winn, J. M., and Zisserman, A. (2010). The pascal visual object classes (VOC) challenge. *Int. J. Comput. Vis.* 88, 303–338. doi: 10.1007/S11263-009-0275-4

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Ann. Eugenics* 7, 179–188. doi: 10.1111/j.1469-1809.1936.tb02137.x

Fu, Q., Zheng, Q., and Yu, F. (2024). LMANet: A lighter and more accurate multiobject detection network for UAV remote sensing imagery. *IEEE Geosci. Remote Sens. Lett.* 21, 1–5. doi: 10.1109/LGRS.2024.3432329

Gao, S., Li, Z.-Y., Han, Q., Cheng, M.-M., and Wang, L. (2023). RF-next: efficient receptive field search for convolutional neural networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 45, 2984–3002. doi: 10.1109/TPAMI.2022.3183829

Gray, P. C., Fleishman, A. B., Klein, D. J., McKown, M. W., Bezy, V. S., Lohmann, K. J., et al. (2019). A convolutional neural network for detecting sea turtles in drone imagery. *Methods Ecol. Evol.* 10, 345–355. doi: 10.1111/2041-210X.13132

Han, F., Yao, J., Zhu, H., and Wang, C. (2020). Underwater image processing and object detection based on deep CNN method. *J. Sens.* 2020, 6707328. doi: 10.1155/2020/6707328

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in 2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30*, 2016* (IEEE Computer Society), 770–778. doi: 10.1109/CVPR.2016.90

Hou, Q., Zhou, D., and Feng, J. (2021). "Coordinate attention for efficient mobile network design," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 13708–13717. doi: 10.1109/CVPR46437.2021.01350

Hu, J., Shen, L., and Sun, G. (2018). "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7132–7141. doi: 10.1109/CVPR.2018.00745

Huang, S.-W., Lin, C.-T., Chen, S.-P., Wu, Y.-Y., Hsu, P.-H., and Lai, S.-H. (2018). "AugGAN: cross domain adaptation with GAN-based data augmentation," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part IX*. Eds. V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss (Springer), 731–744. doi: 10.1007/978-3-030-01240-3_44

Jocher, G. (2020). *YOLOv5 by Ultralytics*. doi: 10.5281/zenodo.3908559

Jocher, G., Qiu, J., and Chaurasia, A. (2023). *Ultralytics YOLO*. Available online at: https://github.com/ultralytics/ultralytics (Accessed April 1, 2025).

Labao, A. B., and Naval, P. C. (2019). Cascaded deep network systems with linked ensemble components for underwater fish detection in the wild. *Ecol. Inform.* 52, 103–121. doi: 10.1016/j.ecoinf.2019.05.004

Li, J., Xu, W., Deng, L., Xiao, Y., Han, Z., and Zheng, H. (2023a). Deep learning for visual recognition and detection of aquatic animals: A review. *Rev. Aquac.* 15, 409–433. doi: 10.1111/raq.12726

Li, X., Hao, Y., Zhang, P., Akhter, M., and Li, D. (2022). A novel automatic detection method for abnormal behavior of single fish using image fusion. *Comput. Electron. Agric.* 203, 107435. doi: 10.1016/J.COMPAG.2022.107435

Li, Y., Fan, Q., Huang, H., Han, Z., and Gu, Q. (2023b). A modified YOLOv8 detection network for UAV aerial image recognition. *Drones* 7, 304. doi: 10.3390/drones7050304

Lin, H., Liu, J., Li, X., Wei, L., Liu, Y., Han, B., et al. (2024). DCEA: DETR with concentrated deformable attention for end-to-end ship detection in SAR images. *IEEE J. Selected Topics Appl. Earth Observ. Remote Sens.* 17, 17292–17307. doi: 10.1109/JSTARS.2024.3461723

Lin, M., Chen, Q., and Yan, S. (2014a). "Network in network," in *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*. Eds. Y. Bengio and Y. LeCun Available online at: http://arxiv.org/abs/1312.4400.

Lin, T.-Y., Goyal, P., Girshick, R. B., He, K., and Dollár, P. (2020). Focal loss for dense object detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 42, 318–327. doi: 10.1109/TPAMI.2018.2858826

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014b). "Microsoft COCO: common objects in context," in *Computer Vision – ECCV 2014*. Eds. D. Fleet, T. Pajdla, B. Schiele and T. Tuytelaars (Springer International Publishing, Cham), 740–755. doi: 10.1007/978-3-319-10602-1_48

Liu, X., Peng, H., Zheng, N., Yang, Y., Hu, H., and Yuan, Y. (2023). "EfficientViT: memory efficient vision transformer with cascaded group attention," in IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24*, 2023* (IEEE), 14420–14430. doi: 10.1109/CVPR52729.2023.01386

Liu, C., Wang, Z., Wang, S., Tang, T., Tao, Y., Yang, C., et al. (2022). A new dataset, poisson GAN and AquaNet for underwater object grabbing. *IEEE Trans. Circuits Syst. Video Technol.* 32, 2831–2844. doi: 10.1109/TCSVT.2021.3100059

Liu, S., Yue, W., Guo, Z., and Wang, L. (2024). Multi-branch CNN and grouping cascade attention for medical image classification. *Sci. Rep.* 14, 15013. doi: 10.1038/s41598-024-64982-w

Lu, W., Chen, S.-B., Shu, Q.-L., Tang, J., and Luo, B. (2024). DecoupleNet: A lightweight backbone network with efficient feature decoupling for remote sensing visual tasks. *IEEE Trans. Geosci. Remote Sens.* 62, 1–13. doi: 10.1109/TGRS.2024.3465496

Lv, W., Zhao, Y., Chang, Q., Huang, K., Wang, G., and Liu, Y. (2024). *RT-DETRv2: Improved Baseline with Bag-of-Freebies for Real-Time Detection Transformer* (CoRR abs/2407.17140). doi: 10.48550/ARXIV.2407.17140

Martin-Abadal, M., Ruiz-Frau, A., Hinz, H., and Cid, Y. G. (2020). Jellytoring: real-time jellyfish monitoring based on deep learning object detection. *Sensors* 20, 1708. doi: 10.3390/S20061708

Pedersen, M., Haurum, J. B., Gade, R., and Moeslund, T. B. (2019). "Detection of marine animals in a new underwater dataset with varying visibility," in IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20*, 2019* (Computer Vision Foundation/IEEE), 18–26. Available online at: http://openaccess.thecvf.com/content_CVPRW_2019/html/AAMVEM/Pedersen_Detection_of_Marine_Animals_in_a_New_Underwater_Dataset_with_CVPRW_2019_paper.html.

Raza, K., and Hong, S. (2020). Fast and accurate fish detection design with improved YOLO-v3 model and transfer learning. *Int. J. Adv. Comput. Sci. Appl.* 11, 7–16. doi: 10.14569/IJACSA.2020.0110202

Ren, S., He, K., Girshick, R. B., and Sun, J. (2017). Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* 39, 1137–1149. doi: 10.1109/TPAMI.2016.2577031

Ren, H., Zhang, Z., Peng, Z., Li, L., and Pan, C. (2023). Energy minimization in RIS-assisted UAV-enabled wireless power transfer systems. *IEEE Internet Things J.* 10, 5794–5809. doi: 10.1109/JIOT.2022.3150178

Ruan, Z., Wang, Z., and He, Y. (2024). DeformableFishNet: a high-precision lightweight target detector for underwater fish identification. *Front. Mar. Sci.* 11. doi: 10.3389/fmars.2024.1424619

Singh, B., and Davis, L. S. (2018). "An analysis of scale invariance in object detection \- SNIP," in 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22*, 2018* (Computer Vision Foundation/IEEE Computer Society), 3578–3587. doi: 10.1109/CVPR.2018.00377

Singh, B., Najibi, M., and Davis, L. S. (2018). "SNIPER: efficient multi-scale training," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. Eds. S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, 9333–9343. Available online at: https://proceedings.neurips.cc/paper/2018/hash/166cee72e93a992007a89b39eb29628b-Abstract.html.

Su, J., Qin, Y., Jia, Z., and Liang, B. (2024). MPE-YOLO: enhanced small target detection in aerial imaging. *Sci. Rep.* 14, 17799. doi: 10.1038/s41598-024-68934-2

Tamou, A. B., Benzinou, A., and Nasreddine, K. (2021). Multi-stream fish detection in unconstrained underwater videos by the fusion of two convolutional neural network detectors. *Appl. Intell.* 51, 5809–5821. doi: 10.1007/S10489-020-02155-8

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). "Attention is All you Need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*. Eds. I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus and S. V. N. Vishwanathan, 5998–6008. Available online at: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wageeh, Y., Mohamed, H. E.-D., Fadl, A., Anas, O., ElMasry, N., Nabil, A., et al. (2021). YOLO fish detection with Euclidean tracking in fish farms. *J. Ambient Intell. Humaniz. Comput.* 12, 5–12. doi: 10.1007/S12652-020-02847-6

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2023a). "YOLOv7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *2023 IEEE/CVF*

Conference on Computer Vision and Pattern Recognition (CVPR), 7464–7475. doi: 10.1109/CVPR52729.2023.00721

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al. (2024a). *YOLOv10: Real-Time End-to-End Object Detection* (CoRR abs/2405.14458). doi: 10.48550/ARXIV.2405.14458

Wang, W., Dai, J., Chen, Z., Huang, Z., Li, Z., Zhu, X., et al. (2023b). "InternImage: exploring large-scale vision foundation models with deformable convolutions," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023* (IEEE), 14408–14419. doi: 10.1109/CVPR52729.2023.01385

Wang, Z., Ruan, Z., and Chen, C. (2024c). DyFish-DETR: underwater fish image recognition based on detection transformer. *J. Mar. Sci. Eng.* 12, 864. doi: 10.3390/jmse12060864

Wang, J., Song, L., Li, Z., Sun, H., Sun, J., and Zheng, N. (2021). "End-to-end object detection with fully convolutional network," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25*, 2021* (Computer Vision Foundation/IEEE), 15849–15858. doi: 10.1109/CVPR46437.2021.01559

Wang, S., Xia, C., Lv, F., and Shi, Y. (2024b). *RT-DETRv3: Real-time End-to-End Object Detection with Hierarchical Dense Positive Supervision* (CoRR abs/2409.08475). doi: 10.48550/ARXIV.2409.08475

Wang, Z., Zhao, L., Li, H., Xue, X., and Liu, H. (2024d). Research on a metal surface defect detection algorithm based on DSL-YOLO. *Sensors* 24, 6268. doi: 10.3390/s24196268

Woo, S., Park, J., Lee, J.-Y., and Kweon, I. S. (2018). "CBAM: convolutional block attention module," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VII*. Eds. V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss (Springer), 3–19. doi: 10.1007/978-3-030-01234-2_1

Wu, X., Sahoo, D., and Hoi, S. C. H. (2020). Recent advances in deep learning for object detection. *Neurocomputing* 396, 39–64. doi: 10.1016/J.NEUCOM.2020.01.085

Xu, S., Zhang, M., Song, W., Mei, H., He, Q., and Liotta, A. (2023). A systematic review and analysis of deep learning-based underwater object detection. *Neurocomputing* 527, 204–232. doi: 10.1016/j.neucom.2023.01.056

Y Adarbah, H., and Ahmad, S. (2019). Channel-adaptive probabilistic broadcast in route discovery mechanism of MANETs. *JCOMSS* 15. doi: 10.24138/jcomss.v15i1.538

Yan, J., Zhou, Z., Zhou, D., Su, B., Zhe, X., Tang, J., et al. (2022). Underwater object detection algorithm based on attention mechanism and cross-stage partial fast spatial pyramidal pooling. *Front. Mar. Sci.* 9. doi: 10.3389/fmars.2022.1056300

Yang, R.-X., Lee, Y.-R., Lee, F.-S., Liang, Z., and Liu, Y. (2024b). An improved YOLOv5 algorithm for bamboo strip defect detection based on the ghost module. *Forests* 15, 1480. doi: 10.3390/f15091480

Yang, C., Xiang, J., Li, X., and Xie, Y. (2024a). FishDet-YOLO: enhanced underwater fish detection with richer gradient flow and long-range dependency capture through mamba-C2f. *Electronics* 13, 3780. doi: 10.3390/electronics13183780

Yang, L., Zhang, R.-Y., Li, L., and Xie, X. (2021). "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18–24 July 2021, Virtual Event*. Eds. M. Meila and T. Zhang (PMLR), 11863–11874. Available online at: http://proceedings.mlr.press/v139/yang21o.html.

Yu, K., Cheng, Y., Li, L., Zhang, K., Liu, Y., and Liu, Y. (2022). Underwater image restoration via DCP and yin-yang pair optimization. *J. Mar. Sci. Eng.* 10, 360. doi: 10.3390/jmse10030360

Yu, G., and Zhou, X. (2023). An improved YOLOv5 crack detection method combined with a bottleneck transformer. *Mathematics* 11, 2377. doi: 10.3390/math11102377

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., et al. (2023a). "DINO: DETR with improved deNoising anchor boxes for end-to-end object detection," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023* (OpenReview.net). Available online at: https://openreview.net/forum?id=3mRwyG5one.

Zhang, Q., Li, Y., Zhang, Z., Yin, S., and Ma, L. (2023b). Marine target detection for PPI images based on YOLO-SWFormer. *Alex. Eng. J.* 82, 396–403. doi: 10.1016/j.aej.2023.10.014

Zhao, Y., Lv, W., Xu, S., Wei, J., Wang, G., Dang, Q., et al. (2024b). "DETRs beat YOLOs on real-time object detection," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 16965–16974. doi: 10.1109/CVPR52733.2024.01605

Zhao, C., Sun, Y., Wang, W., Chen, Q., Ding, E., Yang, Y., et al. (2024a). "MS-DETR: efficient DETR training with mixed supervision," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 17027–17036. doi: 10.1109/CVPR52733.2024.01611

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., and Tian, Q. (2015). "Scalable person re-identification: A benchmark," in 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13*, 2015* (IEEE Computer Society), 1116–1124. doi: 10.1109/ICCV.2015.133

Zhou, B., Khosla, A., Lapedriza, À., Oliva, A., and Torralba, A. (2015). "Object detectors emerge in deep scene CNNs," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*. Eds. Y. Bengio and Y. LeCun Available online at: http://arxiv.org/abs/1412.6856.

Zhu, X., Hu, H., Lin, S., and Dai, J. (2019). "Deformable ConvNets V2: more deformable, better results," in IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20*, 2019* (Computer Vision Foundation/IEEE), 9308–9316. doi: 10.1109/CVPR.2019.00953

Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). "Unpaired image-to-image translation using cycle-consistent adversarial networks," in IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29*, 2017* (IEEE Computer Society), 2242–2251. doi: 10.1109/ICCV.2017.244

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021a). "Deformable DETR: deformable transformers for end-to-end object detection," in 9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7*, 2021* (OpenReview.net).

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021b). Deformable DETR: deformable transformers for end-to-end object detection. doi: 10.48550/arXiv.2010.04159

Zong, Z., Song, G., and Liu, Y. (2023). "DETRs with collaborative hybrid assignments training," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, 6725–6735. doi: 10.1109/ICCV51070.2023.00621