frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 11 April 2025
DOI 10.3389/fmars.2025.1544839

OPEN ACCESS

EDITED BY
Xuebo Zhang,
Northwest Normal University, China

REVIEWED BY
Keyu Chen,
Xiamen University, China
Zhiping Xu,
Jimei University, China
Victor Sineglazov,
National Aviation University, Ukraine

*CORRESPONDENCE
Zhibin Yu
yuzhibin@ouc.edu.cn
Mengxing Huang
mxhuang1129@163.com

RECEIVED 13 December 2024
ACCEPTED 17 March 2025
PUBLISHED 11 April 2025

CITATION
Wang W, Yu Z and Huang M (2025) Refining features for underwater object detection at the frequency level.
Front. Mar. Sci. 12:1544839.
doi: 10.3389/fmars.2025.1544839

Refining features for underwater object detection at the frequency level

Wenling Wang[1], Zhibin Yu[2,3]* and Mengxing Huang[1]*

[1]College of Information and Communication Engineering, Hainan University, Haikou, China, [2]Sanya Oceanographic Institution, Ocean University of China, Sanya, China, [3]College of Electronic Engineering, Ocean University of China, Qingdao, China

In recent years, underwater object detection (UOD) has become a prominent research area in the computer vision community. However, existing UOD approaches are still vulnerable to underwater environments, which mainly include light scattering and color shifting. The blurring problem caused by water scattering on underwater images makes the high-frequency texture edge less obvious, affecting the detection effect of objects in the image. To address this issue, we design a multi-scale high-frequency information enhancement module to enhance the high frequency features extracted by the backbone network and improve the detection effect of the network on underwater objects. Another common issue caused by scattering and color shifting is that it can easily change the low-frequency information in the background of underwater images, leading to performance degradation of the same target in different underwater scenes. Therefore, we have also designed a multi-scale gated channel information optimization module to reduce the scattering and color shifting effects on the channel information of underwater images and adaptively compensate the features for different underwater scenes. We tested the detection performance of our designed method on three typical underwater object detection datasets, RUOD, UDD and UODD. The experimental results proved that our method performed better than existing detection methods on underwater object detection datasets.

KEYWORDS

underwater object detection, deep learning, frequency utilization, feature extraction, DINO

1 Introduction

Recently, the availability of underwater imagery has increased exponentially due to the widespread use of digital cameras deployed in autonomous underwater vehicles (AUVs), remotely operated vehicles (ROVs), and unmanned underwater vehicles (UUVs). These devices provide a sound platform to collect various underwater images and opportunities for learning-based underwater object detection (UOD) automatic image analysis techniques (Xu et al., 2023a; Han et al., 2023). Accurate detection of organisms, geology,

and marine debris in deep-sea environments is vital for human society and marine environmental protection (Zhang et al., 2022). However, the complex underwater environment results in high noise, low visibility, blurred edges, low contrast, and color deviation in underwater images, posing significant challenges to underwater object detection tasks (Xu et al., 2023b).

Advancements in deep learning have led to many high-performance object detection algorithms (Girshick, 2015; Ren et al., 2017; Redmon, 2016; Liu et al., 2016). Recently, the transformer-based end-to-end object detector DETR (DEtection TRansformer) (Carion et al., 2020) proposed by Carion et al. has received widespread attention. DETR (Carion et al., 2020) eliminates the need for manually set pre-processing and post-processing operations, greatly simplifying the object detection pipeline. Networks such as DABDeTR (Liu et al., 2022), Deformable-DeTR (Zhu et al., 2021), DN-DeTR (Li et al., 2022b), and DINO (Zhang et al., 2023) have further optimized the DETR (Carion et al., 2020)model, achieving excellent detection results.

The common solution for UOD is to retrain existing detectors (e.g., CNNs) (Fu et al., 2023). However, the complex underwater environment results in high noise, low visibility, blurred edges, low contrast, and color deviation in underwater images, posing significant challenges to underwater object detection tasks (Xu et al., 2023b). While many existing UOD works focus on efficient feature enhancement (Fan et al., 2020), augmentation (Lin et al., 2020) or small object detection (Jiang et al., 2021a), few UOD works noticed the frequency information utilization. As shown in Figure 1, the influence of water scattering causes blurring in underwater images, which can make the original high-frequency texture edges and other distinguishing features of the image less obvious and degrade the detection effect. Therefore, we propose a multi-scale high-frequency information strengthening module (MHFIEM) to strengthen high-frequency information for multi-scale features extracted by the backbone network, optimizing the detection effect of the network on blurred underwater images.

Furthermore, different underwater environments usually result in unfixed information variation in different channels, leading to low-frequency feature shifting (e.g., color biases or illumination variation) in underwater images. To address this, we propose a multi-scale gated channel information refinement module (MGCIRM), which combines a gating mechanism to take advantage of low-frequency features to optimize multi-scale underwater image channel features and reduce the impact of scattering and color shifting on channel information in underwater images.

Additionally, the significant scale differences among underwater objects (e.g., sea cucumbers, sea turtles, and divers) and the scale imbalance pose a huge challenge for underwater object detection (Fu et al., 2023). Inspired by the multi-scale strategy of DINO, we design both the MHFIEM and MGCIRM modules to optimize underwater image features at multiple scales to better adapt to the scale changes of underwater detection targets.

In summary, this paper designs an underwater object detection network, named underwater DINO (UDINO), based on the current excellent detection model DINO (Zhang et al., 2023). UDINO introduces a multi-scale high-frequency information strengthening Module (MHFIEM) and a multi-scale gated channel information refinement module (MGCIRM) on top of the DINO network to enhance the network's detection performance for underwater objects. Experimental validation

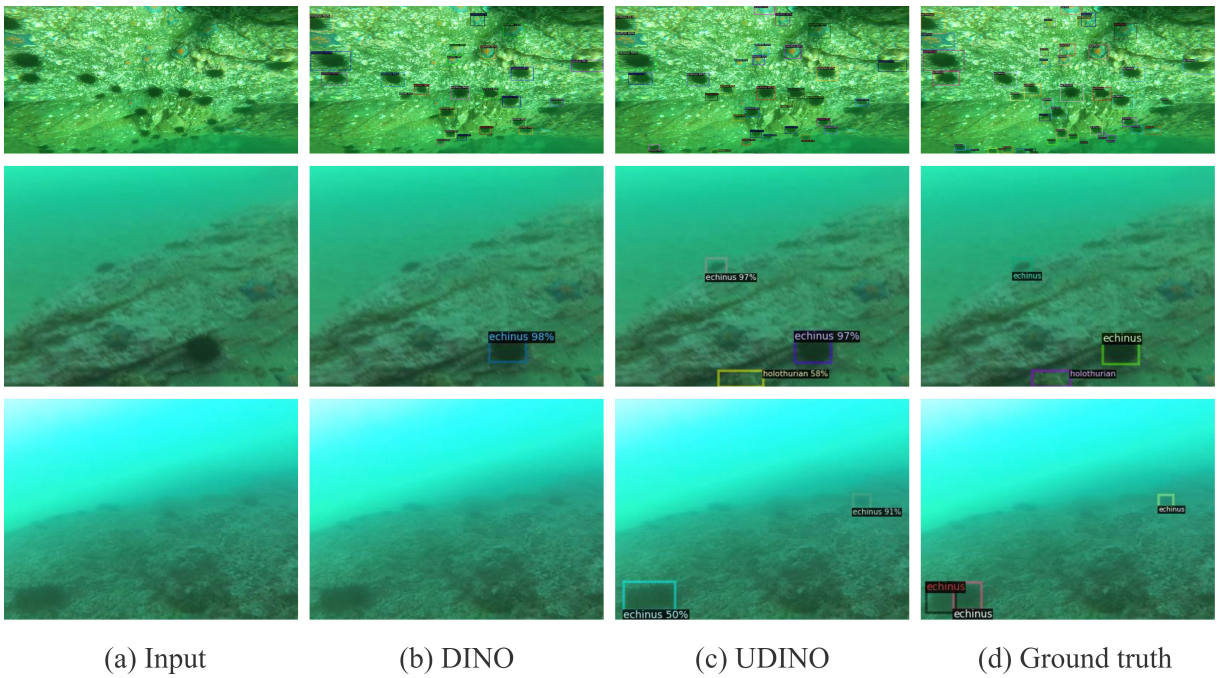

FIGURE 1
The improvement of underwater image object detection performance by UDNIO compared to DINO.

demonstrates that UDINO achieves better detection results than existing methods on current underwater object detection datasets. The contributions of this paper are as follows:

- This paper proposes a new underwater object detection network, UDINO, which achieves the best detection results on the existing underwater object detection datasets, UDD, RUOD and UODD.
- This paper introduces the MHFIEM module to optimize high-frequency information in underwater images at multiple scales, mitigating the adverse effects of underwater image blurring on detection.
- This paper proposes the MGCIRM module to optimize channel features in underwater images based on low-frequency information, reducing the impact of background interference caused by scattering in underwater images.
- Through experiments, we demonstrate the effectiveness of modules MHFIEM and MGCIRM, and find that these two modules also contribute to the underwater object detection performance of other DETR-based object detection models.

2 Research background

2.1 Underwater object detection

Object detection is a fundamental problem in computer vision and has gained significant attention in recent years. The success of deep-learning spawns various deep learning based object detection models. There are two typical groups of deep learning based object detectors: "two-stage" and "one-stage" (Zhang et al., 2020b). The two-stage detectors follows a "coarse-to-fine" process, while the latter completes the object detection task "in one step."

The core idea of single-stage object detection algorithms is to directly predict the category and location of objects in a single forward pass, eliminating the step of generating candidate regions in traditional two-stage algorithms, thereby achieving faster detection speeds. SSD (Liu et al., 2016) predicts the location and category of objects on feature maps at multiple scales, enabling effective detection of objects of different sizes. The RetinaNet (Ross and Dollár, 2017) algorithm addresses the issue of imbalance between positive and negative samples in single-stage object detection by introducing Focal Loss, significantly improving detection accuracy. The YOLO series object detection algorithms are a typical branch of fast object detection approaches (Redmon, 2016; Redmon and Farhadi, 2017; Wang et al., 2023).

The two-stage object detection method first determines the area where the target is located and then determines the target category during object detection. For example, the Faster R-CNN (Ren et al., 2017) introduces the Region Proposal Network (RPN), which enables the process of generating candidate regions to be optimized through learning. Cascade RCNN (Cai and Vasconcelos, 2018) improves detection accuracy based on Faster RCNN by cascading multiple detectors to gradually correct the localization and recognition results of the target.

Similar as the object detection tasks, the purpose of underwater object detection (UOD) is to identify the type and location of an object in an underwater image (Fu et al., 2023). However, the UOD tasks always suffer from high noise, low visibility, blurred edges, low contrast, and color deviation in underwater images. To address these issues, some researchers redesigned architectures based on the existing object detection frameworks with efficient feature refinement (Fan et al., 2020), augmentation (Lin et al., 2020) or small object detection (Jiang et al., 2021a). Unlike these works, this paper focus on the exploration of frequency information utilization.

2.2 DETR-based object detection

Compared to classical detection algorithms, DETR (Carion et al., 2020) is a novel detection algorithm based on transformers. DETR (Carion et al., 2020) models object detection as a set prediction task and assign labels through bipartite matching. While DETR demonstrates good performance, it has a slow training convergence rate. Conditional DETR (Meng et al., 2021) proposed the conditional cross attention mechanism. It explicitly searches for the extreme region of an object through conditional spatial queries, thereby narrowing down the search range and accelerating the convergence speed of the DETR model training. UP-DETR (Dai et al., 2021) learns target localization capability through random query patch detection, which also significantly improves the performance and convergence speed of the DETR model. Deformable DETR (Zhu et al., 2021) addresses this issue by designing a Deformable attention module that focuses only on certain sampling points around reference points to improve the training convergence speed of the DETR algorithm. DAB-DETR (Liu et al., 2022) proposes defining the DETR queries as dynamic anchor boxes (DAB), bridging the gap between traditional anchor-based detectors and DETR-like detectors. DN-DETR (Li et al., 2022b) further addresses the instability of bipartite matching by introducing denoising (DN) techniques. DINO (Zhang et al., 2023), on the other hand, proposes a contrastive denoising training approach that rejects useless anchor boxes to assist in model training convergence. Based on the DINO (Zhang et al., 2023) model, we designed the UDINO model for underwater object detection, which addresses the issues of large target scale span, blurring of detection images due to water scattering, and loss of channel information. This model achieves better performance in underwater object detection.

3 Methods

3.1 Overall structure of UDINO

The overall structure diagram of Underwater DINO (UDINO) proposed by us is shown in Figure 2. We design our model based on a popular detection backbone DINO. Since DINO includes a pyramid structure to obtain multi-scale features using the ResNet50 (He et al., 2016) backbone network, we design the multi-scale high-frequency

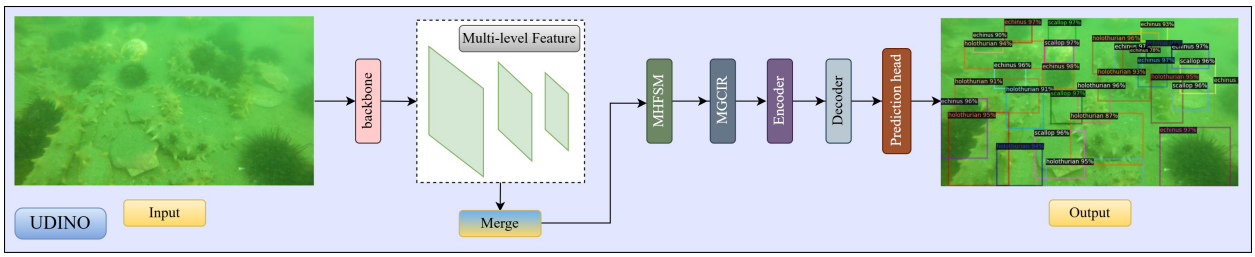

FIGURE 2
Overview of the UDINO.

information strengthen module (MHFIEM) and multi-scale gated channel information refine module (MGCIRM) to utilize the multi-scale features, separately. Finally, the optimized features are passed through the DINO's transformer encoder, decoder, and prediction head to obtain the prediction results.

3.2 Multi-scale high-frequency information enhancement module

Refining features is a typical way to improve the detection performance (Fan et al., 2020). Specially, the key features of the underwater objects (e.g., sea urchin and holothurian) always includes high frequency information (e.g., edges, textures and contours). However, the degradation of underwater image quality caused by water scattering brings difficulties in underwater object detection. Although many underwater image enhancement methods can generally improve the image contrast (Wang et al., 2024b), pointed out that enhancing the degraded underwater images before detection often can hardly improve the underwater object detection effect. On the one hand, underwater image enhancement may increase noise interference, edge blur, and texture corruption problems. These problems can further damage the high-frequency texture edge information and lead to performance degradation of an underwater objection detection task. On the other hand, the independent enhancement and the detection process result in the lack of connection between the adjustment of enhanced images and the performance optimization of the detection network. Considering these factors, we design an embedded multi-scale high-frequency information enhancement module (MHFIEM), which can adaptively enhance the high-frequency features according to the detection requirements of the detection network.

FcaNet (Qin et al., 2021) pointed out that the low-frequency characteristics of the input information are linearly correlated with its average pooling result. This inspires us to obtain high-frequency information through the pooling operation with subtraction. As shown in left part of Figure 3, suppose MHFIEM receives multi-scale features X extracted from the backbone network and uses the dictionary structure to store and process different scale features separately. We use the average pooling operation to process the different scale features X of the input image individually, obtaining the corresponding low-frequency features F_l in different scales as shown in Equation 1:

$$F_l = Avgpool(X) \tag{1}$$

Next, we subtract the original features X_m under the m^{th} scale from $F_{l,m}$ to obtain the high-frequency feature $F_{h,m}$ of the input feature X_m as shown in Equation 2.

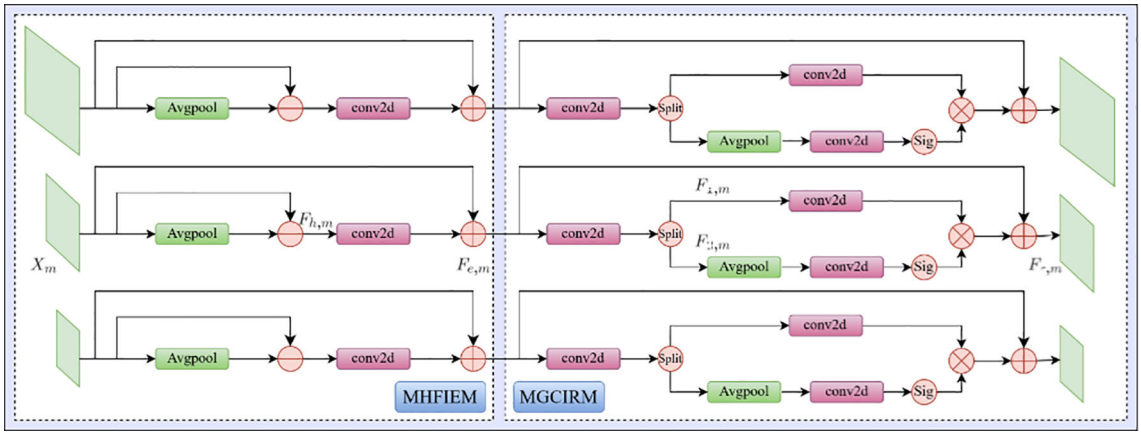

FIGURE 3
The network structure of MHFIEM and MGCIRM.

$$F_{h,m} = X_m - F_{l,m} = X_m - Avgpool(X_m) \quad (2)$$

where F_{h_m} is adaptively weighted by a 1×1 convolutional layer. The enhanced feature F_e can be simply represented by the following formulas (Equation 3):

$$F_{e,m} = Conv(F_{h,m}) + X_m \quad (3)$$

Since the convolutional operation *Conv* includes trainable parameters, the whole feature enhancement process is also an adaptive enhancement progress.

3.3 Multi-scale gated channel information refinement module

Water bodies' absorption and scattering can easily cause channel information severely changing in underwater images, resulting in varying degrees of color cast in the underwater image background. The color cast difference in the background can interfere with the stable detection of the same target. To alleviate this issue, we design a multi-scale gate-controlled channel information refinement module (MGCIRM) to adaptively optimize the channel information of underwater images. Unlike the features of underwater objects, the underwater backgrounds always contain low-frequency features with different color and brightness. Thus, the purpose of MGCIRM is to capture the background (low-frequency) information and adaptively compensate the features for different underwater scenes and reduce the impact of scattering on channel information changing of underwater images.

As shown in the right part of Figure 3, the multi-scale feature F_e strengthened by the MHFIEM module is convolved by 1 × 1 convolution to stretch the channel to twice its previous size. Next, the information flow is split into feature F_1, and feature F_2 through the split operation. The number of feature channels for F_1 and F_2 is consistent with the input feature F_e as shown in Equation 4.

$$F_{1,m}, F_{2,m} = Split(Conv(F_{e,m})) \quad (4)$$

where F_1 undergoes a layer of 1 × 1 convolution to further tune the features and obtain feature X_e. F_2 is processed through an average pooling, a 1 × 1 convolution layer to obtain a low-frequency feature $Avgpool(F_1)$. Since a sigmoid function can provide outputs between 0 and 1, we design a sigmoid function after $Avgpool(F_1)$ to obtain the channel adaptive adjustment factor f_c. Next, we can obtain the channel-wise attention result of by multiplying F_c and X_e. Finally, we can get the output of MGCIRM by adding the original feature F_e and the channel-wise attention, achieving channel information optimization for the multi-scale feature $F_e$and get the refined feature F_r. The entire forward propagation process can be represented by the following formulas (Equations 5, 6):

$$F_{c,m} = Sigmoid(Conv(Avgpool(F_{2,m}))) \quad (5)$$

$$F_{r,m} = X_m + Conv(F_{1,m}) \times F_{c,m} \quad (6)$$

4 Experiment

In this section, we first describe the implementation details and introduce the experimental settings. Then, we compare our method with representative object detection methods on underwater object detection datasets RUOD (Fu et al., 2023), UDD (Liu et al., 2021) and UODD (Jiang et al., 2021a). Then, we show the underwater object detection visual effect of UDINO on (Fu et al., 2023) and UODD (Jiang et al., 2021a). In the ablation experiment, we verify the efficiency of MHFIEM and MGCIRM. Finally, we show the MHFIEM and MGCIRM are also applicable to other DETR (Carion et al., 2020) models.

4.1 Implementation details

UDINO includes 69.541M parameters, and the FLOPS for a an image with 1920×1080 resolution is 209.6 GFLOPs. We implement our method with PyTorch and train it on 4 NVIDIA Tesla A40 GPUs. The batch size is 8. The number of iterations of ablation and verification experiments is 100K. We adopt average precision (AP) and AP50 as the primary metrics for model accuracy evaluation (Zhu, 2004), with precision (P) and recall (R) as supplementary indicators. We use APs, APm, and APl to evaluate the detection effect of detectors on objects with different scales. To showcase the generalizability of our network, we train and test our method on the RUOD dataset, the UODD dataset and UDD dataset. The RUOD dataset contains various underwater scenes and consists of 10 categories. It includes 9800 training images and 4200 test images. The UODD dataset consists of 3 categories, which include 2688 training images and 506 test images. The UDD dataset includes 2227 underwater images, where 1,827 ones are for training and 400 for testing.

4.2 Quantitative comparisons

We compared the detection performance of our method with single-stage, two-stage, and DETR-based object detection methods on the RUOD dataset in Table 1. We also evaluated the performance of some YOLO series on RUOD dataset in Table 2. As shown in Tables 1, 3, compared to other methods, our method achieved the best detection performance on all indicators. Compared to the suboptimal DINO model (Zhang et al., 2023), our UDINO improved the overall AP metric by 1.1 (from 60.9 to 62.0). Our APl and APm metrics improved by 1.0 (from 66.8 to 67.8) and 0.2 (from 46.4 to 46.6) respectively in large and medium-sized object detection on the dataset. Compared to the suboptimal FCOS method for small object detection, our method improves the APs metric by 0.1.

We compare the detection performance of our method with other methods on the UODD dataset in Table 4. We also evaluated the performance of some YOLO series on UODD dataset in Table 3. As shown in Tables 3, 4, compared to other methods, our method

TABLE 1 Quantitative comparisons of underwater object detection effects on RUOD dataset.

Methods	AP↑	AP50↑	AP75↑	APs↑	APm↑	APl↑
Faster RCNN (Ren et al., 2017)	52.8	81.8	57.5	17.2	40.9	58.2
Cascade RCNN (Cai and Vasconcelos, 2018)	54.8	81.1	59.7	16.8	42.2	60.6
Dynamic RCNN (Zhang et al., 2020a)	54.4	81.3	60.3	17.1	42.8	60.0
Libra RCNN (Pang et al., 2019)	54.8	82.8	60.5	16.5	43.1	60.6
RetinaNet (Ross and Dollár, 2017)	50.7	79.3	54.5	14.3	39.2	56.1
FCOS (Tian et al., 2022)	50.7	79.5	54.0	18.0	40.0	56.2
ATSS (Zhang et al., 2020b)	52.9	80.3	56.9	16.4	41.1	58.6
Deformable-DETR (Zhu et al., 2021)	57.4	85.6	63.2	17.4	43.0	63.1
Dab-DETR (Tian et al., 2022)	55.7	85.0	61.1	13.8	41.7	61.3
DINO (Zhang et al., 2023)	60.9	85.7	66.3	17.6	46.4	66.9
UDINO (ours)	62.0	86.1	67.8	18.1	46.6	67.8

achieved the best detection performance on all indicators. Compared to the suboptimal DINO model (Zhang et al., 2023), our UDINO improved the overall AP metric by 1.1 (from 60.9 to 62.0). Our APl and APm metrics improved by 1.0 (from 66.8 to 67.8) and 0.2 (from 46.4 to 46.6) respectively in large and medium-sized object detection on the dataset. Compared to the suboptimal FCOS method for small object detection, our method improves the APs metric by 0.1.

We further conducted experiments on UDD datasets (Liu et al., 2021) and compared them with some other methods in Table 5. Comparing with the other two datasets, UDD is a challenging dataset for underwater object detection with less samples and small sizes. Compared to the suboptimal FCOS method for small object detection, our method improves the AP metric by 0.5 (from 25.8 to 26.3). Since most of the targets of UDD are objects with small sizes, our APs and APm metrics improved by 1.9 (from 14.2 to 16.1) and 0.5 (from 25.1 to 25.6) respectively in small and medium-sized object detection tasks on the UDD dataset.

TABLE 2 Quantitative comparisons with YOLO series on RUOD dataset (Fu et al., 2023).

Methods	AP↑	AP50↑
YOLOv3 (Farhadi and Redmon, 2018)	49.1	80.3
YOLOv5 (Bochkovskiy et al., 2020)	53.8	81.4
YOLOv6 (Li et al., 2022a)	60.1	84.9
YOLOv7 (Wang et al., 2023)	57.9	84.3
YOLOv8n (Sohan et al., 2024)	58.2	84.2
YOLOv9t (Wang et al., 2024a)	59.2	83.3
YOLOv10s (Wang et al., 2025)	59.8	84.6
YOLOv11n (Khanam and Hussain, 2024)	56.5	81.9
YOLOv11s (Khanam and Hussain, 2024)	61.7	85.8
UDINO (Ours)	62.0	86.1

4.3 Qualitative comparisons

We compare our method with recent excellent object detection methods in Figures 4, 5. The proposed method has outperformed the other detection methods. As shown in Figure 4, our method performs better than other methods, especially in detecting unclear, small targets at the bottom of the input image in the first line. For the input image in the first line, our method performs better in detecting blurry small targets on the right side of the image, detecting more targets overall and fewer missed targets. For the second line of Figure 4, our method has higher detection accuracy for jellyfish targets in deep blue waters.

As shown in Figure 5, there are many underwater objects located in different underwater scenes with significant color shifting. Our method has better localization ability on the holothurian target at the bottom of the input image in the first line compared to other methods. Our method detects box boundaries more accurately. In the second line of Figure 5, our method can detect more correct scallops that are close to the background color. In the third and fourth lines of Figure 5, our method can accurately locate the sea urchin target sandwiched among starfishes and sea urchins.

TABLE 3 Quantitative comparisons of underwater object detection effects on the UODD dataset (Jiang et al., 2021a).

Methods	AP↑	AP50↑
YOLOv3 (Farhadi and Redmon, 2018)	48.4	88.9
YOLOX (Ge et al., 2021)	48.8	86.3
YOLOv8s (Sohan et al., 2024)	50.7	89.8
YOLOv9t (Wang et al., 2024a)	48.7	86.8
YOLOv10s (Wang et al., 2025)	49.2	88.5
YOLOv11s (Khanam and Hussain, 2024)	50.8	89.9
UDINO (Ours)	51.1	90.2

TABLE 4 Quantitative comparisons of underwater object detection effects on the UODD dataset.

Methods	AP↑	AP50↑	AP75↑	APs↑	APm↑	APl↑
Faster RCNN (Ren et al., 2017)	47.1	86.5	44.7	30.7	46.9	56.6
Cascade RCNN (Cai and Vasconcelos, 2018)	48.5	86.4	49.1	33.9	48.2	58.2
Dynamic RCNN (Zhang et al., 2020a)	46.9	84.8	47.2	30.9	47.2	55.1
Libra RCNN (Pang et al., 2019)	47.5	87.0	45.3	32.0	47.7	58.0
RetinaNet (Ross and Dollár, 2017)	45.2	84.2	42.2	33.0	45.5	53.5
FCOS (Tian et al., 2022)	44.8	86.3	39.2	32.1	44.9	52.4
ATSS (Zhang et al., 2020b)	48.3	88.1	45.5	34.5	47.7	57.9
Deformable-DETR (Zhu et al., 2021)	48.4	84.2	51.5	32.6	48.0	62.6
Dab-DETR (Tian et al., 2022)	49.8	89.4	49.8	31.0	50.1	60.8
DINO (Zhang et al., 2023),	49.8	87.2	51.7	33.4	50.5	64.1
UDINO (ours)	51.1	90.2	52.6	34.3	51.1	62.7

4.4 Verify the effectiveness of module design

4.4.1 Ablation of MHFIEM

To verify the impact of the proposed module on the network performance, we conducted a series of ablation experiments. Table 6 shows that after adding the MHFIEM module, the network's AP detection performance on the RUOD dataset improved by 0.78. At the same time, we noticed that the model with the MHFIEM module performed better when detecting large and medium-scale detection targets, with the model's APm index improving by 0.17 and the APl index significantly increasing by 0.92.

To validate the effectiveness of our design, we also conducted an experiment of low-frequency information enhancement to observe the effect of low-frequency information enhancement. We named the low-frequency information enhancement module MLFIEM, and its effect with the previous high-frequency information enhancement module MHFIEM. From the experimental results in Table 6, the effect of MHFIEM module is better than MLFIEM. The high-frequency and low-frequency information with different scale features is also different, and the effect of strengthening the high-frequency information from multiple scales will be better. To verify this point, we constructed a single scale high frequency information enhancement module HFIEM, and compared its effect with our MHFIEM module in Table 6. The experimental results show that enhancing high-frequency information from multi-scale performs better.

4.4.2 Ablation of MGCIRM

From Table 6, we can see that after further adding the MGCIRM module, the network's Ap detection performance on the RUOD dataset has significantly improved by 0.33. At the same time, we noticed that the model with the MHFIEM module can greatly improve the detection performance of the model for small-scale targets, with a significant increase in the APs metric of 2.23. The model's detection performance for large targets has slightly improved, with an APl index increase of 0.07. Due to the large proportion of large-scale targets in the dataset, the slight improvement in detection performance for large targets also has a significant positive impact on overall detection performance.

4.4.3 MHFIEM and MGCIRM modules are also effective in other DETR models

To further investigate the potential of our design, we added our MHFIEM module and MGCIRM module to other DETR (Carion et al., 2020) structures. Table 7 shows the changes in the detection performance of our DETR models Dab-DETR (Liu et al., 2022) and Deformable-DETR (Zhu et al., 2021) after adding MHFIEM and

TABLE 5 Quantitative comparisons of underwater object detection effects on the UDD dataset (Liu et al., 2021).

Methods	AP↑	AP50↑	AP75↑	APs↑	APm↑	APl↑
Faster RCNN (Ren et al., 2017)	25.7	60.3	14.3	13.6	25.3	34.0
Dynamic RCNN (Zhang et al., 2020a)	25.0	58.6	15.1	13.6	24.6	36.5
Libra RCNN (Pang et al., 2019)	25.5	60.1	14.7	14.1	25.4	29.0
RetinaNet (Ross and Dollár, 2017)	19.4	45.6	11.4	10.3	17.4	26.2
FCOS (Tian et al., 2022)	25.8	61.0	15.5	14.2	25.1	31.2
UDINO (Ours)	26.3	66.6	13.5	16.1	25.6	25.4

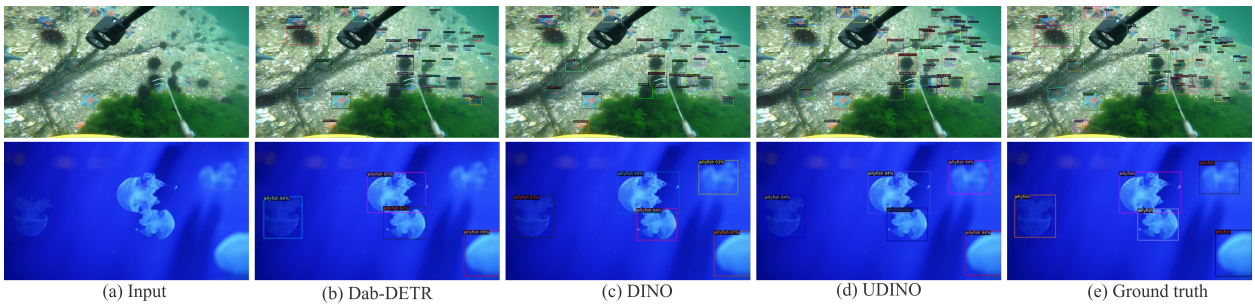

FIGURE 4
Qualitative comparison on the RUOD dataset.

MGCIRM modules. After introducing our module, the detection performance indicator AP of the Dab-DETR (Liu et al., 2022) model has improved by 0.50, and the detection performance of the Deformable-DETR model has improved from 57.44 to 59.81. Thus, we found that MHFIEM and MGCIRM modules can enhance other DETR models' underwater object detection performance as well.

To further demonstrate the efficiency of MHFIEM and MGCIRM modules, we added Figure 6 to demonstrate the impact of our design in the feature dimension. From the figure, we can find that the attention maps with MHFIEM and MGCIRM are brighter than the baseline, which indicate that the proposed UDNIO can effectively capture the underwater features.

5 Conclusion

In this paper, we propose a new underwater object detection network UDINO which can refine underwater features at the frequency level. The blurring and color cast of underwater images caused by water scattering pose challenges to underwater object detection. We design MHFIEM and MGCIRM modules specifically to adaptively enhance the high-frequency and channel features of underwater images, respectively, to improve the detection performance of the detection network for underwater targets. We tested the detection performance of our designed method on representative underwater object detection datasets RUOD and

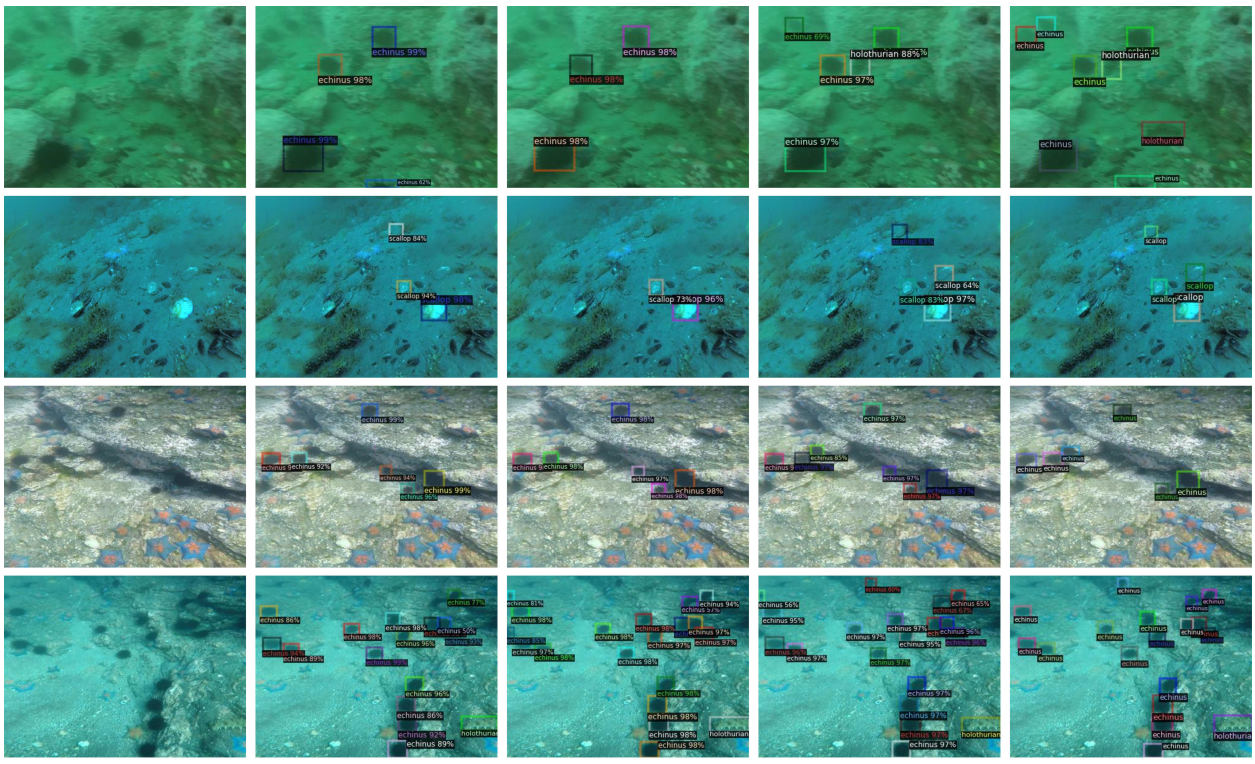

FIGURE 5
Qualitative comparison on the UODD dataset.

TABLE 6 Ablation study.

Methods	AP↑	AP50↑	AP75↑	APs↑	APm↑	APl↑
Base	60.87	85.66	66.28	17.60	46.43	66.85
Base + MHFIEM	61.65	85.86	67.62	15.84	46.60	67.77
Base + MLFIEM + MGCIRM	61.20	85.81	66.47	19.44	46.43	66.08
Base + HFIEM + MGCIRM	61.57	85.52	67.20	19.28	46.31	67.90
Base + MHFIEM + MGCIRM (Ours)	61.98	86.14	67.76	18.07	46.56	67.84

TABLE 7 Effectiveness of the proposed modules with different DETR backbones.

Methods	AP↑	AP50↑	AP75↑	APs↑	APm↑	APl↑
Deformable-DETR (Zhu et al., 2021)	57.44	85.57	63.18	17.41	43.02	63.11
Deformable-DETR + MHFIEM + MGCIRM	59.81	85.05	66.32	16.27	44.76	65.78
Dab-DETR (Tian et al., 2022)	55.74	84.98	61.11	13.79	41.65	61.26
Dab-DETR + MHFIEM + MGCIRM	56.24	85.51	61.25	15.03	42.03	61.63
DINO (Zhang et al., 2023)	60.9	85.7	66.3	17.6	46.4	66.9
DINO + MHFIEM + MGCIRM (Ours)	61.98	86.14	67.76	18.07	46.56	67.84

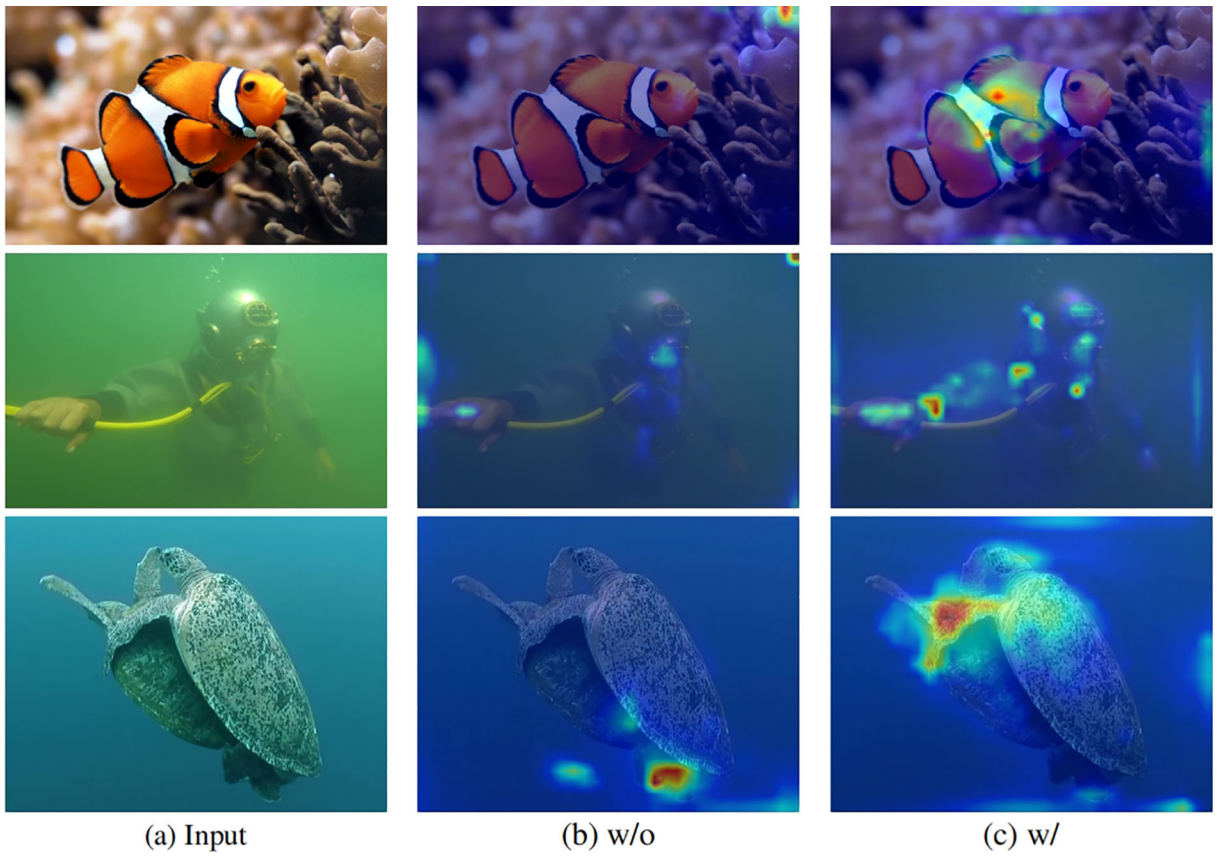

FIGURE 6
Demonstrate the impact of our design in the feature dimension. w/o and w/o means without and with our network design(MHFIEM and MGCIRM).

UODD, and relevant experiments proved that our method has better detection accuracy than existing detection methods on underwater object detection datasets.

Data availability statement

The original contributions presented in the study are included in the article/supplementary material. Further inquiries can be directed to the corresponding authors.

Ethics statement

The manuscript presents research on animals that do not require ethical approval for their study.

Author contributions

WW: Writing – original draft. ZY: Writing – review & editing. MH: Funding acquisition, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by National Natural Science Foundation of China (Grant: 82260362), the Hainan Province Science and Technology Special Fund, China (ZDYF2022SHFZ318) and the National Natural Science Foundation of China (Grant No. 62171419).

Conflict of interest

The 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.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

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

Publisher's note

References

Bochkovskiy, A., Wang, C.-Y., Liao, H.-Y. M., and Liu, Y.-J. (2020). YOLOv5.

Cai, Z., and Vasconcelos, N. (2018). "Cascade r-cnn: Delving into high quality object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (New York, USA: IEEE). 6154–6162.

Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., and Zagoruyko, S. (2020). "End-to-end object detection with transformers," in *European conference on computer vision*. (Berlin, Germany: Springer). 213–229.

Dai, Z., Cai, B., Lin, Y., and Chen, J. (2021). "UP-DETR: Unsupervised pre-training for object detection with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (New York, USA: IEEE). 1601–1610.

Fan, B., Chen, W., Cong, Y., and Tian, J. (2020). "Dual refinement underwater object detection network," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XX*, (Berlin, Germany: Springer). 16. 275–291.

Farhadi, A., and Redmon, J. (2018). "YOLOv3: An incremental improvement," in *Computer vision and pattern recognition*, vol. 1804. (Springer Berlin, Heidelberg, Germany), 1–6.

Fu, C., Liu, R., Fan, X., Chen, P., Fu, H., Yuan, W., et al. (2023). Rethinking general underwater object detection: Datasets, challenges, and solutions. *Neurocomputing* 517, 243–256. doi: 10.1016/j.neucom.2022.10.039

Ge, Z., Liu, S., Wang, F., Li, Z., and Sun, J. (2021). YOLOX: exceeding YOLO series in 2021. *arXiv preprint arXiv:2107.08430*. 1–7.

Girshick, R. (2015). "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*. (New York, USA: IEEE). 1440–1448. doi: 10.1109/ICCV.2015.169

Han, L., Zhai, J., Yu, Z., and Zheng, B. (2023). See you somewhere in the ocean: few-shot domain adaptive underwater object detection. *Front. Mar. Sci.* 10, 1151112. doi: 10.3389/fmars.2023.1151112

He, K., Zhang, X., Ren, S., and Sun, J. (2016). "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (New York, USA: IEEE). 770–778.

Jiang, L., Wang, Y., Jia, Q., Xu, S., Liu, Y., Fan, X., et al. (2021a). "Underwater species detection using channel sharpening attention," in *Proceedings of the 29th ACM International Conference on Multimedia*. (New York, USA: ACM). 4259–4267.

Khanam, R., and Hussain, M. (2024). Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*. 1–17.

Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., et al. (2022a). YOLOv6: A single-stage object detection framework for industrial applications. *arXiv preprint arXiv:2209.02976*. 1–9.

Li, F., Zhang, H., Liu, S., Guo, J., Ni, L. M., and Zhang, L. (2022b). "DN-DETR: Accelerate DETR training by introducing query denoising," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (New York, USA: IEEE). 13619–13627.

Lin, W.-H., Zhong, J.-X., Liu, S., Li, T., and Li, G. (2020). "Roimix: proposal-fusion among multiple images for underwater object detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (IEEE)*. (New York, USA: IEEE). 2588–2592.

Liu, C., Wang, Z., Wang, S., Tang, T., Tao, Y., Yang, C., et al. (2021). 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., Li, F., Zhang, H., Yang, X., Qi, X., Su, H., et al. (2022). "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *International Conference on Learning Representations*. (Washington DC: International Conference on Learning Representations).

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., et al. (2016). "Ssd: Single shot multibox detector," in *Computer Vision–ECCV 2016: 14th European*

Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. (Berlin, Germany: Springer). 21–37.

Meng, D., Chen, X., Fan, Z., Zeng, G., Li, H., Yuan, Y., et al. (2021). "Conditional DETR for fast training convergence," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. (New York, USA: IEEE).

Pang, J., Chen, K., Shi, J., Feng, H., Ouyang, W., and Lin, D. (2019). "Libra r-cnn: Towards balanced learning for object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (New York, USA: IEEE). 821–830.

Qin, Z., Zhang, P., Wu, F., and Li, X. (2021). "Fcanet: Frequency channel attention networks," in *Proceedings of the IEEE/CVF international conference on computer vision*. (New York, USA: IEEE). 783–792.

Redmon, J. (2016). "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (New York, USA: IEEE).

Redmon, J., and Farhadi, A. (2017). "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE conference on computer vision and pattern recognition*. (New York, USA: IEEE). 7263–7271.

Ren, S., He, K., Girshick, R., 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

Ross, T.-Y., and Dollár, G. (2017). "Focal loss for dense object detection," in *proceedings of the IEEE conference on computer vision and pattern recognition*. (New York, USA: IEEE). 2980–2988.

Sohan, M., Sai Ram, T., Reddy, R., and Venkata, C. (2024). "A review on yolov8 and its advancements," in *International Conference on Data Intelligence and Cognitive Informatics*. (New York, USA: IEEE). 529–545.

Tian, Z., Chu, X., Wang, X., Wei, X., and Shen, C. (2022). Fully convolutional one-stage 3d object detection on lidar range images. *Adv. Neural Inf. Process. Syst.* 35, 34899–34911.

Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y. M. (2023). "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (New York, USA: IEEE). 7464–7475.

Wang, A., Chen, H., Liu, L., Chen, K., Lin, Z., Han, J., et al. (2025). Yolov10: Real-time end-to-end object detection. *Adv. Neural Inf. Process. Syst.* 37, 107984–108011.

Wang, Y., Guo, J., He, W., Gao, H., Yue, H., Zhang, Z., et al. (2024b). Is underwater image enhancement all object detectors need? *IEEE J. Oceanic Eng.* 49, 606–621. doi: 10.1109/JOE.2023.3302888

Wang, C.-Y., Yeh, I.-H., and Mark Liao, H.-Y. (2024a). "Yolov9: Learning what you want to learn using programmable gradient information," in *European conference on computer vision*. (New York, USA: IEEE). 1–21.

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

Xu, G., Zhou, D., Yuan, L., Guo, W., Huang, Z., and Zhang, Y. (2023a). Vision-based underwater target real-time detection for autonomous underwater vehicle subsea exploration. *Front. Mar. Sci.* 10, 1112310. doi: 10.3389/fmars.2023.1112310

Zhang, H., Chang, H., Ma, B., Wang, N., and Chen, X. (2020a). "Dynamic r-cnn: Towards high quality object detection via dynamic training," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV*, (Berlin, Germany: Springer). 16, 260–275.

Zhang, S., Chi, C., Yao, Y., Lei, Z., and Li, S. Z. (2020b). "Bridging the gap between anchor-based and anchor-free detection via adaptive training sample selection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (New York, USA: IEEE). 9759–9768.

Zhang, H., Li, F., Liu, S., Zhang, L., Su, H., Zhu, J., et al. (2023). "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *The Eleventh International Conference on Learning Representations*. (Washington DC: International Conference on Learning Representations).

Zhang, W., Zhuang, P., Sun, H.-H., Li, G., Kwong, S., and Li, C. (2022). Underwater image enhancement via minimal color loss and locally adaptive contrast enhancement. *IEEE Trans. Image Process.* 31, 3997–4010. doi: 10.1109/TIP.2022.3177129

Zhu, M. (2004). *Recall, precision and average precision* Vol. 2 (Waterloo: Department of Statistics and Actuarial Science, University of Waterloo), 6.

Zhu, X., Su, W., Lu, L., Li, B., Wang, X., and Dai, J. (2021). "Deformable DETR: Deformable transformers for end-to-end object detection," in *International Conference on Learning Representations*. (Washington DC: International Conference on Learning Representations).