frontiers | Frontiers in Marine Science

TYPE Original Research
PUBLISHED 01 April 2025
DOI 10.3389/fmars.2025.1557965

OPEN ACCESS

EDITED BY
Zhaoqiang Xia,
Northwestern Polytechnical University, China

REVIEWED BY
Ling Yang,
Kunming University of Science and Technology, China
Yupeng Ma,
Ningxia Medical University, China

*CORRESPONDENCE
Lijie Cao
caolijie@dlou.edu.cn

RECEIVED 10 January 2025
ACCEPTED 07 March 2025
PUBLISHED 01 April 2025

CITATION
He Z, Cao L, Xu X and Xu J (2025) Underwater instance segmentation: a method based on channel spatial cross-cooperative attention mechanism and feature prior fusion. *Front. Mar. Sci.* 12:1557965. doi: 10.3389/fmars.2025.1557965

Underwater instance segmentation: a method based on channel spatial cross-cooperative attention mechanism and feature prior fusion

Zhiqian He[1], Lijie Cao[1,2]*, Xiaoqing Xu[1] and Jianhao Xu[1]

[1]College of Information Engineering, Dalian Ocean University, Dalian, China, [2]Key Laboratory of Marine Information Technology of Liaoning Province (Dalian Ocean University), Dalian, China

In aquaculture, underwater instance segmentation methods offer precise individual identification and counting capabilities. However, due to the inherent unique optical characteristics and high noise in underwater imagery, existing underwater instance segmentation models struggle to accurately capture the global and local feature information of objects, leading to generally lower detection accuracy in underwater instance segmentation models. To address this issue, this study proposes a novel Channel Space Coordinates Attention (CSCA) attention module and a Channel A Prior Attention Fusion (CAPAF) feature fusion module, aiming to improve the accuracy of underwater instance segmentation. The CSCA module effectively captures local and global information by combining channel and spatial attention weight, while the CAPAF module optimizes feature fusion by removing redundant information through learnable parameters. Experimental results demonstrate significant improvements when these two modules are applied to the YOLOv8 model, with the mAP@0.5 metric increasing by 3.2% and 2% on the UIIS underwater instance segmentation dataset. Furthermore, the instance segmentation accuracy is significantly improved on the UIIS and USIS10K datasets after these two modules are applied to other networks.

KEYWORDS

underwater instance segmentation, YOLO, attention mechanism, feature fusion, instance segmentation

1 Introduction

With the widespread application of underwater instance segmentation methods in aquaculture, it has significantly improved the precision and efficiency of management. For example, this method enables real-time monitoring (Yang et al., 2024) and growth tracking of cultured organisms (Fan et al., 2021), thereby optimizing feeding strategies and disease

prevention measures. Simultaneously, it has reduced manual intervention and promoted the development of aquaculture automation, further advancing the economic development and technological innovation of the aquaculture industry. Therefore, researching methods to enhance the accuracy of underwater image instance segmentation is of crucial importance for the continuous progress of aquaculture technology.

In recent years, with the rapid development of deep learning, many scholars have applied deep learning image instance segmentation methods to aquaculture. For example (Kannan, 2020), use analysis optimization techniques such as genetic algorithms, particle swarm optimization, and differential evolution to initialize the parameter set and finally detect the objects in the underwater images using evolution-based Gaussian mixture models and shape matching (Zhang L. et al., 2024). proposed an improved BoTS-YOLOv5s-seg model based on YOLOv5, which reached 90.9% mAP@0.5 in the individual counts of farmed fish (Zheng et al., 2024). proposed a video object segmentation-based method for fish individual recognition in underwater complex environments, and the key metric Rank1 value of the method achieved >96% accuracy on the public datasets DlouFish, WideFish, and Fish-seg datasets (Siri et al., 2024). proposed an improved two-stage R-FCN model achieving 99.94% accuracy, 99.58% accuracy and recall, and 99.27% F-measure on the Fish4knowledge dataset (Li et al., 2017). developed an adaptive thresholding underwater image segmentation method using image segmentation to detect fish feed consumption (Chen et al., 2024). proposed a new MPG-Net semantic segmentation model for aquaculture ponds to effectively improve the segmentation accuracy of individual aquaculture ponds using residual links and Global Context module and Polarized Self-Attention (Ma et al., 2018). proposed NWPU underwater image database to improve the quality of underwater images through enhanced image technique. However, the above scholars are applying the image instance segmentation method to the aquaculture field, not analyzing the underwater image characteristics to improve the performance of the underwater instance segmentation method.

Numerous scholars have recognized that efficient underwater instance segmentation methods are crucial for applications in the aquaculture sector and can significantly improve the efficiency of aquaculture management, thereby promoting the industry's sustainable development. Therefore, some scholars have proposed numerous underwater instance segmentation methods based on the characteristics of underwater images. For example, improving the performance of underwater instance segmentation models by means of image enhancement methods (Wang et al., 2024), such as adjusting brightness, contrast, color balance, and applying filters, can effectively improve the models' training effect and generalization ability (Farhadi Tolie et al., 2024). These enhancement operations can simulate different underwater environmental conditions, helping the model learn better and recognize complex underwater scenes, thus achieving higher accuracy and robustness in the segmentation task (Peng et al., 2023). In addition, some scholars improve the model performance by designing or selecting a network architecture that is more suitable for the characteristics of the underwater environment; for example, ZhiQian (He et al., 2024) propose an underwater image semantic segmentation network (UISS-Net) to improve the boundary accuracy of underwater image object segmentation (Zhang Z. et al., 2024). proposed a lightweight semantic segmentation model for underwater fish images with an improved U-Net to address the low progress of current underwater image segmentation (Zhang W. et al., 2024). proposed a real-time semantic segmentation network called WaterBiSeg-Net to address the problems of slow inference and large computation in existing underwater detection algorithms (Shen et al., 2024). proposed a multi-information-aware attention module (MIPAM) based on spatial downsampling and channel segmentation to solve the underwater image noise interference (Han et al., 2023). proposed an iterative attention feature fusion mechanism to fully perceive the features and contextual information at different scales and proposed an underwater fish segmentation method based on improved PSPNet network (IST-PSPNet) to greatly improve the segmentation accuracy of underwater fish. However, the above researchers and scholars have mainly relied on increasing the network depth of the model and enhancing the underwater image data to improve the segmentation accuracy of underwater images. These methods do not effectively use the channel and spatial information in the model, and the full use of contextual information in the feature fusion process has not been fully considered. Therefore, exploring how to effectively use spatial and channel information in the model and efficiently integrate contextual information in the feature fusion stage has become a key challenge in improving the performance of underwater image segmentation.

To address the prevalent noise issues and detail blurring in underwater images, which lead to insufficient feature extraction in instance segmentation models and the failure to exploit contextual information fully, this work proposes a novel Channel Space Coordinates Attention (CSCA) mechanism and a Channel A Prior Attention Fusion (CAPAF) feature fusion module. The CSCA attention is achieved by extracting spatial and channel attention weights from the feature maps and, drawing on the method of the Transformer architecture, multiplying these weights to obtain a matrix of spatial and channel correlation attention weights. By multiplying this matrix with the input feature maps, we can obtain attention weights with spatial-channel coordination. This allows the CSCA mechanism to possess both a global receptive field for capturing long-range dependencies and a local receptive field for channel and spatial feature maps, thereby effectively enhancing the depth of feature extraction. The CAPAF module, on the other hand, extracts key information by filtering contextual information and adjusting the proportion of feature fusion using learnable parameters, solving the insufficient integration of contextual information in the model and ensuring the full utilization of contextual information. Finally, we embed the CSCA attention mechanism and the CAPAF module into several mainstream instance segmentation models, and the experimental results consistently show that the performance of these models has been significantly verified, corroborating their effectiveness.

The structure of the remainder of this work is as follows: *Section 2* delves into the related literature; *Section 3* outlines the CSCA

attention mechanism and the CAPAF feature fusion module proposed herein; *Section 4* presents the experimental findings, encompassing a detailed account of the experimental setup and outcomes; finally, *Section 5* draws conclusions from the study.

2 Related work

2.1 YOLO

In recent years, instance segmentation has become an important research area in computer vision and has attracted widespread attention. The YOLO network proposed by (Redmon et al., 2016) is significantly innovative compared to traditional two-stage target detection models. The network achieves direct prediction of target coordinates through the design of a fully convolutional neural network, which simplifies the detection process and improves the detection efficiency. The objective detection and instance segmentation task networks were subsequently provided in the YOLOv5 release by Ultralytics LLC, respectively. Then, to solve the problem of anchor-based needs to set many hyperparameters, the YOLOv8 version based on anchor-free was subsequently introduced. YOLOv8 does not need to preset the anchor and only needs to regress the target centroid and width-height of feature maps at different scales, which reduces the time-consuming and arithmetic power. Due to this, the YOLO series network has fast detection speed and high accuracy and is easy to train and deploy for many scholars who are applied to the aquaculture field. Therefore, the CSCA attention and CAPAF modules proposed in this work validate their effectiveness through application in the YOLO series networks.

2.2 Attention mechanism

The human visual system can spontaneously focus attention on information-rich key areas in complex visual scenes. As a computer vision technique, the attention mechanism aims to simulate this property of human vision. The core of the mechanism lies in shifting the focus of visual processing from the global image to key feature points in the image. In this way, the attention mechanism effectively motivates the model to focus on the more discriminative elements of the input features, thus significantly improving the model's performance (Woo et al., 2018). proposed the Convolutional Block Attention Module (CBAM) by adjusting spatial and channel attention weights (Hou et al., 2021). proposed coordinate attention by embedding the location information into the channel attention, thus allowing the mobile network to acquire information about a larger area without introducing large overheads (Huang et al., 2024). proposed a new Channel Prior Convolutional Attention (CPCA) method, which adopted multi-scale depth-separable convolutional modules to constitute spatial attention and could dynamically assign attentional weights in both channel and spatial dimensions (Yang et al., 2021). proposed a conceptually simple and very effective attention SimAM module that derives 3D attention weights for feature maps without additional parameters. The above attention mechanism focuses too much on the channel and spatial information. However, it ignores the coordination between the channel and spatial information. Therefore, this work proposes the CSCA attention mechanism, which obtains the characteristic channel-space coordination attention weight matrix by establishing the channel and spatial correlation matrix.

2.3 Feature fusion

In order to maximize the utility of features extracted from the backbone network (Lin et al., 2017), innovatively proposed the Feature Pyramid Network (FPN). This network skillfully integrates high-resolution shallow features with deep features rich in semantic information by introducing a top-down structure and lateral connections, significantly enhancing feature utilization efficiency. Consequently, FPN has become crucial in various visual tasks such as object detection and instance segmentation. Building upon this, numerous researchers have introduced improved feature fusion methods based on FPN, aiming to enhance model performance further (Liu et al., 2018). proposed the Path Aggregation Network (PAN), which, particularly in image segmentation tasks, effectively addresses the issue of low-level detail information loss during feature transmission in FPN and insufficient detail refinement of high-level features (Tan et al., 2020). proposed Bidirectional Feature Pyramid Network (BiFPN) to solve the problems of unidirectional information flow and unequal feature weight allocation at the feature level in the feature fusion process of traditional FPN. Nevertheless, these feature fusion networks effectively combine high- and low-level semantic information by integrating feature maps of different scales (Zhou et al., 2024). finally proposed MW-YOLO network for small target detection through the proposed multi-scale feature fusion module for neck network to enhance the fusion effect of different scale features. However, feature fusion solely in the dimensionality does not fully exploit contextual semantic information. In response, this study introduces a novel CAPAF module, which dynamically adjusts the distribution of attention weights through learnable parameters during the feature fusion process, thereby more effectively utilizing the contextual semantic information within the features.

3 Methods

To address the challenges of coarse contour edges, high noise, and insufficient utilization of contextual information in underwater image instance segmentation, we propose the CSCA attention mechanism and the CAPAF feature fusion module, designed to capture the contour features of various organisms in aquaculture accurately. Figure 1 illustrates the structure of CSCA, which

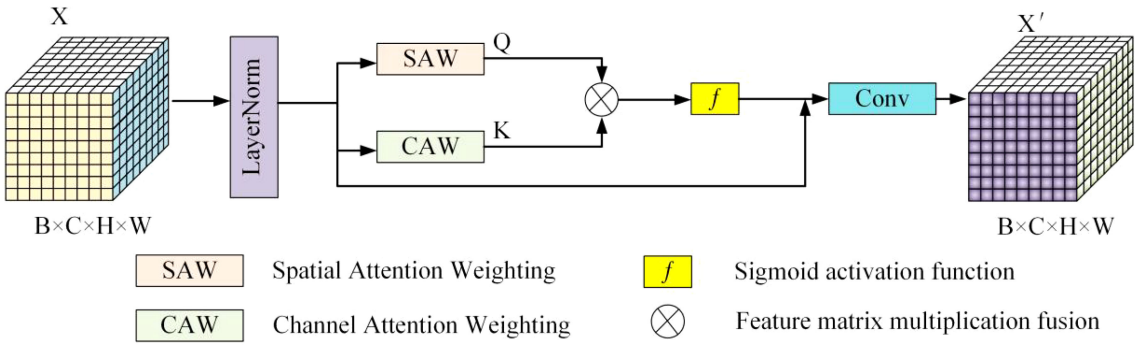

FIGURE 1
Structure of the CSCA attention module.

primarily processes the spatial and channel information of feature maps to obtain respective spatial and channel attention weights. By employing a Transformer architecture, long-range dependencies are established between the spatial and channel attention weights, endowing the CSCA attention module with a global receptive field while maintaining local perception. Figure 2 details the structure of the CAPAF module, which significantly increases the importance of key information in the feature fusion phase by introducing learnable parameters to optimize the attention weights of the input features. This mechanism ensures that the model uses contextual information more effectively when integrating features, enhancing the understanding and representation of scene details.

3.1 CSCA attention

In order to solve the problem of low efficiency of instance segmentation model in underwater image feature extraction, this study proposes a CSCA attention mechanism, whose structure is shown in Figure 1. The CSCA attention mechanism first performs normalization on the input features X to enhance the robustness of the model. Subsequently, the Spatial Attention Weighting (SAW) and Channel Attention Weighting (CAW) modules are utilized to extract the local receptive fields of the features, respectively. Following this, a self-attention mechanism is applied to multiply the obtained Q (query) and K (key) matrices point-wise, thereby

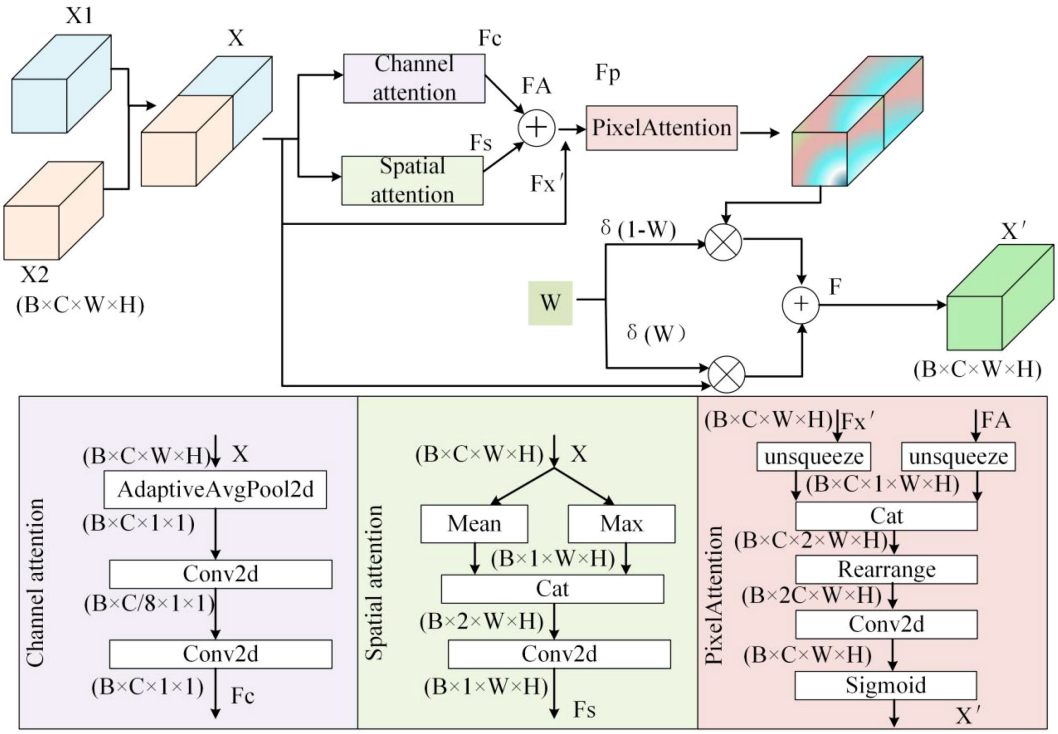

FIGURE 2
Structure of the attention fusion module for CAPAF dynamic focusing.

constructing the global receptive field. Finally, after processing with the Sigmoid activation function, the activation results are multiplied by the normalized X and then subjected to convolution operations, resulting in an enhanced feature map X' that integrates both the global receptive field with long-distance dependencies and the local receptive fields within the channel space.

In the SAW module, the average values of the height and width of the normalized feature map$X$ are first calculated to extract key information from the feature map. Next, dilated convolution is used to expand the receptive field, performing further feature extraction on the height and width of the feature map. Subsequently, these features are concatenated to integrate the common characteristics of height and width. Finally, the attention weights are generated through grouped convolution and activated using the Sigmoid function, which are then multiplied by the input X to obtain the attention weights Q, realizing the spatial attention operation. The structure of this module is shown in Figure 3.

In the CSCA attention mechanism, Q and K feature maps are mainly obtained by feature adjustment of the input feature map X, respectively.

First, by adjusting the width (W) and height (H) of the input feature map X, F_W and F_H are obtained as shown in Equation 1:

$$\begin{aligned} F_H &= Mean(X, \dim = 2) \\ F_W &= Mean(X, \dim = 3) \end{aligned} \tag{1}$$

The obtained F_W and F_H feature maps are split by channel $^{C}/_{4}$. Convolution operations are performed on the four feature maps obtained by inputting convolution kernels of (Li et al., 2017; Ma et al., 2018; Kannan, 2020; Zheng et al., 2024) size, respectively. The feature weight matrices under different perceptual fields on different channels can be obtained, and then, the weight matrices are fused to obtain F_{WM} and F_{HM} by features according to the channel C. After group normalization of the F_{WM} and F_{HM} feature maps, respectively, the feature maps are subjected to Sigmoid activation to obtain the attention matrix, and then, the attention matrix is multiplied by X and then convolved with a 3×3 convolution to obtain the feature map K as shown in Equation 2.

$$\begin{aligned} F_{HM} &= Sigmoid(GroupNorm(F_H)) \\ F_{WM} &= Sigmoid(GroupNorm(F_W)) \\ Q &= Conv(X \times (F_{HM} + F_{WM}) \end{aligned} \tag{2}$$

Additionally, to obtain the local attention weights for the channels, the input X is adjusted along the channel dimension to derive the channel attention weights K. The schematic diagram of this structure is shown in Figure 4.

First, global pooling is performed on the normalized feature map X to obtain the $C1$ and $C2$ feature channels. The $C1$ and $C2$ matrices are then transposed to obtain the $T1$ and $T2$ matrices, respectively. The obtained $C1$ and $C2$ matrices and their transposed matrices are cross-multiplied to obtain the $CT1$ and $CT2$ matrix weights for channel information fusion. As shown in Equation 3,

$$\begin{aligned} C1 &= \text{Adaptive Pooling(X)} \\ C2 &= \text{Adaptive Pooling(X)} \\ CT1 &= \text{C1} \otimes \text{C2}^{\text{T}} \\ CT2 &= C2 \otimes \text{C1}^{\text{T}} \end{aligned} \tag{3}$$

The weights $CT1$ and $CT2$ rectangles after channel fusion are obtained through Equation 3 and then activated by the Sigmoid function. At this point, the matrix with blended attention weights is obtained. However, there is channel redundancy information in the reweight matrix. Therefore, the two weight matrices $W1$ and $W2$ are fused after adjusting the valid and invalid features by the learnable hyperparameter θ. Finally, the Sigmoid function is activated again to obtain the channel attention weights of the output X_C feature map, which is multiplied with the input feature X and then mixed with the channel information after 3×3 convolution to the K matrix. The formula is shown in Equation 4.

$$\begin{aligned} X_C &= (W1 \otimes \sigma(\theta)) \oplus (W2 \otimes (1 - \sigma(\theta))) \\ K &= \text{Conv}(Sigmoid(X_C)) \end{aligned} \tag{4}$$

Finally, the Q and K weight matrices obtained through spatial and channel processing are multiplied with the output X, which is the Q and K matrix in the self-attention mechanism that can be used

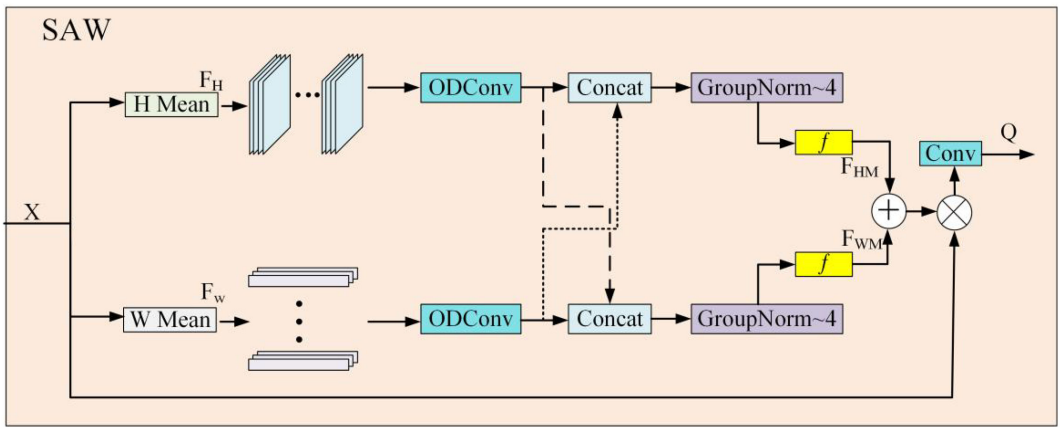

FIGURE 3
Structure diagram of the SAW module.

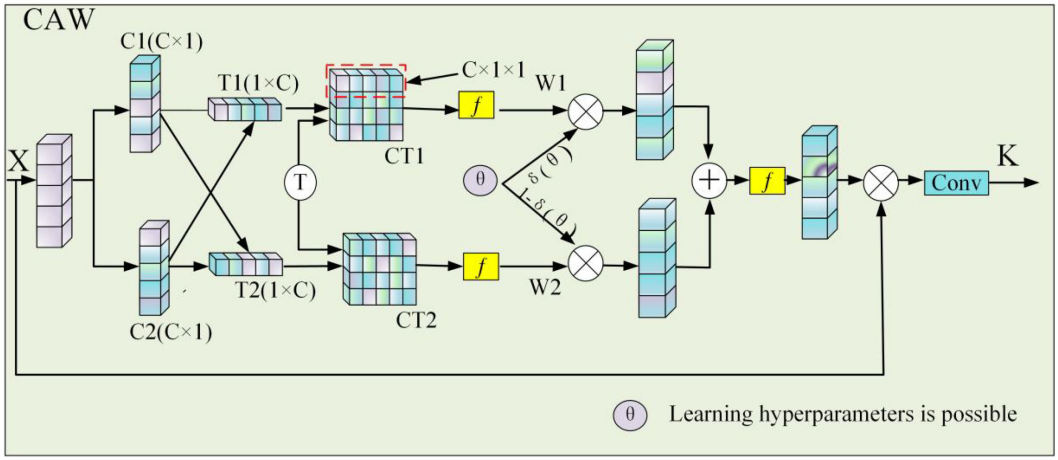

FIGURE 4
Structure diagram of the CAW module.

to establish long textual dependencies. The matrix is then multiplied with the input X matrix by the Sigmoid function activation. The spatial channel coordinated attention weights X' after spatial and channel processing are obtained. as shown in Equation 5.

$$X' = X \otimes (Sigmoid(Q \otimes K)) \tag{5}$$

3.2 CAPAF feature fusion module

Concatenation (Concat) operation is used for feature fusionin deep learning, but it increases the computational burden, raises the risk of overfitting, and may introduce redundant information. In addition, another Element-wise Addition (ADD) feature fusion operation maintains feature dimensionality. However, it cannot deal with features of different dimensions, may ignore the importance of features, and cannot capture non-linear relationships, which affects its effectiveness in complex feature interaction processing. Therefore, this work proposes the attention-fusion module CAPAF with dynamic focusing, the structure of which is shown in Figure 2.

In Figure 2, the structure of CAPAF feature fusion module is shown. First, the input $X1$ and $X2$ are fused to obtain the fused baseline feature X. Second, the feature matrix X is subjected to channel attention and spatial attention operations, respectively, to obtain the feature map of the attentional weights of feature X in channel and space as shown in Equation 6, and at the same time, the learnable parameter W is activated and the value domain is adjusted to be between [0,1] by using the Sigmoid function.

$$\begin{aligned} X &= Concat(X1, X2) \\ Fc &= Channela\ Attention(X) \\ Fs &= Spatial\ Attention(X) \\ FA &= Fc + Fs \end{aligned} \tag{6}$$

Subsequently, in order to further highlight the importance of different informative features in the feature map, this work employs the Unsqueezee function to insert an additional dimension to the channel dimensions of Fx' and FA, respectively. This is done to learn the key features at the pixel locations and to perform the fusion of the features after this step is completed. Next, through the activation of the Sigmoid function, this work obtains the feature weights reflecting the pixel's attention Fp. On this basis, this work uses the learnable parameter W to adjust the weights of X and Fp in feature fusion to generate the final weighted feature map F. By performing a convolution operation on the feature map F, this work successfully constructs an attention fusion feature map with dynamic focusing capability X'. The detailed process and computation are described in Equation 7.

$$\begin{aligned} Fp &= Pixel\ Atention(Fx', FA) \\ F &= ((W \otimes X) + (1 - W) \otimes Fp) \\ X' &= Conv2D(F) \end{aligned} \tag{7}$$

4 Experiments and discussion

4.1 Dataset

This work aims to verify the effectiveness of the proposed attention and feature fusion modules in underwater scenes. For this purpose, we used the publicly available Underwater Image Instance Segmentation (UIIS) dataset (Lian et al., 2023), which contains 4,628 images, each with pixel-level annotations for seven underwater instance segmentation task categories. The dataset is divided into 3,937 training images and 691 validation images. The quality of the dataset images is shown in Figure 5A. Additionally, we used the large dataset USIS10K (Lian et al., 2024) proposed by Lian Shijie et al. for underwater prominent instance segmentation tasks. This dataset contains 10,632 images, with 7,442 images for training,

1,594 images for validation, and 1,596 images for testing. The dataset provides pixel-level annotations for seven categories, with specific examples shown in Figure 5B.

4.2 Experimental parameters and performance assessment indicators

The GPU version required for the experiments is NVIDIA GTX 4080. The software environment was Pytorch 1.8.0 on Python 3.7, Anaconda 3, CUDA 10.0, and CUDNN 7.3.0. A total of 200 training epochs were employed to ensure the model training's convergence. Batch size was set to 12 based on the graphics card's performance. using the SGD optimizer. The learning rate was initially 0.01. The point cloud measurement model was deployed on a homemade aquatic animals' measurement platform.

In this work, the average prediction accuracy mAP@0.5 and the number of model parameters, the computational complexity of the model, are used as evaluation metrics. This Mean Average Precision (mAP) can be calculated based on the precision P and recall R, as shown in Equations 8, 9.

$$P = \frac{TP}{TP + FP} \times 100\,\% \tag{8}$$

$$R = \frac{TP}{TP + FN} \times 100\,\% \tag{9}$$

where TP and FP are the predicted positive and negative samples, respectively, and FN is the incorrectly predicted sample. The mAP is a combination of recall and precision that effectively evaluates model detection performance, as shown in Equation 10.

$$mAP = \frac{1}{n}\sum_{i=1}^{n}\int_{0}^{1} P(R)dR \tag{10}$$

where n represents the number of detected categories.

4.3 Experimental results and analysis

To validation the effectiveness of the proposed CSCA attention mechanism and CAPAF module in underwater scenarios, in this work, we conduct ablation experiments in the UIIS underwater scene dataset based on the YOLOv8 network. First, YOLOv8 network is used as the baseline model. Second, CSCA attention and CAPAF feature fusion module are added, respectively. The results are shown in Table 1.

From the ablation experiments in Table 1, it can be seen that our proposed CSCA attention and CAPAF feature fusion modules have an improved effect on the segmentation performance of

(a)Publicly available images of the UIIS dataset.

(b)Publicly available images of the USIS10K dataset.

FIGURE 5
Display of the image section of the dataset. **(A)** Publicly available images of the UIIS dataset. **(B)** Publicly available images of the USIS10K dataset.

TABLE 1 CSCA attention and CAPAF module in YOLOv8 model results.

NO	CSCA	CAPAF	Parameters/10^6	GFLOPs	mAP@0.5/%
1			3.5	12.0	37.5
2	✓		4.5	14.8	40.7
3		✓	5.3	18.7	39.5
4	✓	✓	7.7	24.9	41.2

underwater instances. Specifically, the mAP@0.5 reaches 40.7% and 39.5% after using the CSCA attention and CAPAF feature fusion model in the baseline model, respectively. Compared with the baseline model, mAP@0.5 is improved by 3.2% and 2%, respectively. When the CSCA attention mechanism and the CAPAF feature fusion module are simultaneously applied to the baseline model, the number of parameters increases by 4.2M, and the computational cost rises by 12.9 GFlops. However, the mAP@0.5 improves to 41.2%, which is 3.7% higher than the baseline model, demonstrating the significant advantage of these modules in enhancing the accuracy of underwater instance segmentation.

In order to verify the superiority of the CSCA attention proposed in this work, it is compared with the current mainstream attention models. The results are shown in Table 2.

From the experimental results in Table 2, it is concluded that the embedding of the CSCA attention module significantly improves the performance of the YOLOv8 model. Compared to the original model, the CSCA attention module not only significantly increased precision (from 41.5% to 48.7%) but also appropriately increased recall (from 41.5% to 43.2%), which is the best performance among all compared attention modules. Despite the increase in the number of parameters and GFLOPs, the performance improvement brought by the CSCA attention module proved its value in the target detection task, resulting in a mAP@0.5 of 40.7%, outperforming the other attention modules.

In order to understand the region of attention of the network after adding the attention mechanism, in this work, the GradCAM method is used to map the last layer of gradient weights of the backbone network to get the model heat map. Figure 2 shows the results of the heat map of several attention mechanisms compared, and from the demonstration in Figure 6, it can be understood that the CSCA attention mechanism proposed in this work pays more attention to the target region and is able to extract effective feature information.

Finally, in order to examine the performance of the CSCA attention mechanism and the CAPAF function fusion module proposed in this study in the current mainstream network architectures, we embedded the CSCA attention mechanism into the backbone networks of YOLOv5, YOLOv8, YOLOv9, and YOLOv11, respectively, and applied the CAPAF module to the Neck network. The corresponding experimental results are displayed in Tables 3, 4, respectively.

In Tables 3, 4, attention is given to the significant enhancement of the YOLO model performance by the introduction of both CSCA and CAPAF modules. By integrating the CSCA attention and CAPAF functional fusion modules, the detection accuracy of the YOLOv8 and YOLOv11 models in several underwater scenarios is significantly enhanced. For example, the addition of these two modules improves the YOLOv8 model's mAP@0.5 from 55.8% to 57.5% on the USIS10 dataset, and in particular, it performs more accurately in detecting human divers and robots. Similarly, the

TABLE 2 Results of CSCA attention compared to other attention mechanisms.

Model	Precision	Recall	Parameters/10^6	GFLOPs	mAP@0.5/%
YOLOv8	41.5	41.5	3.5	12.0	37.5
+Coord Attention	45.3	41.8	3.2	12.0	39.8
+CPCA Attention	44.9	38.3	3.3	12.4	38.7
+SimAM Attention	40.0	45.2	3.2	12.0	40.1
+CBAM Attention	45.0	41.2	3.3	12.6	39.7
+SE Attention	45.0	38.9	3.2	12.0	39.2
+AFGCAttention (Han et al., 2025)	46.0	39.9	3.2	12.0	40.2
+FCAttention (Sun et al., 2024)	42.2	46.5	3.2	12.0	40.0
+CSCA Attention	48.7	43.2	4.5	14.8	40.7

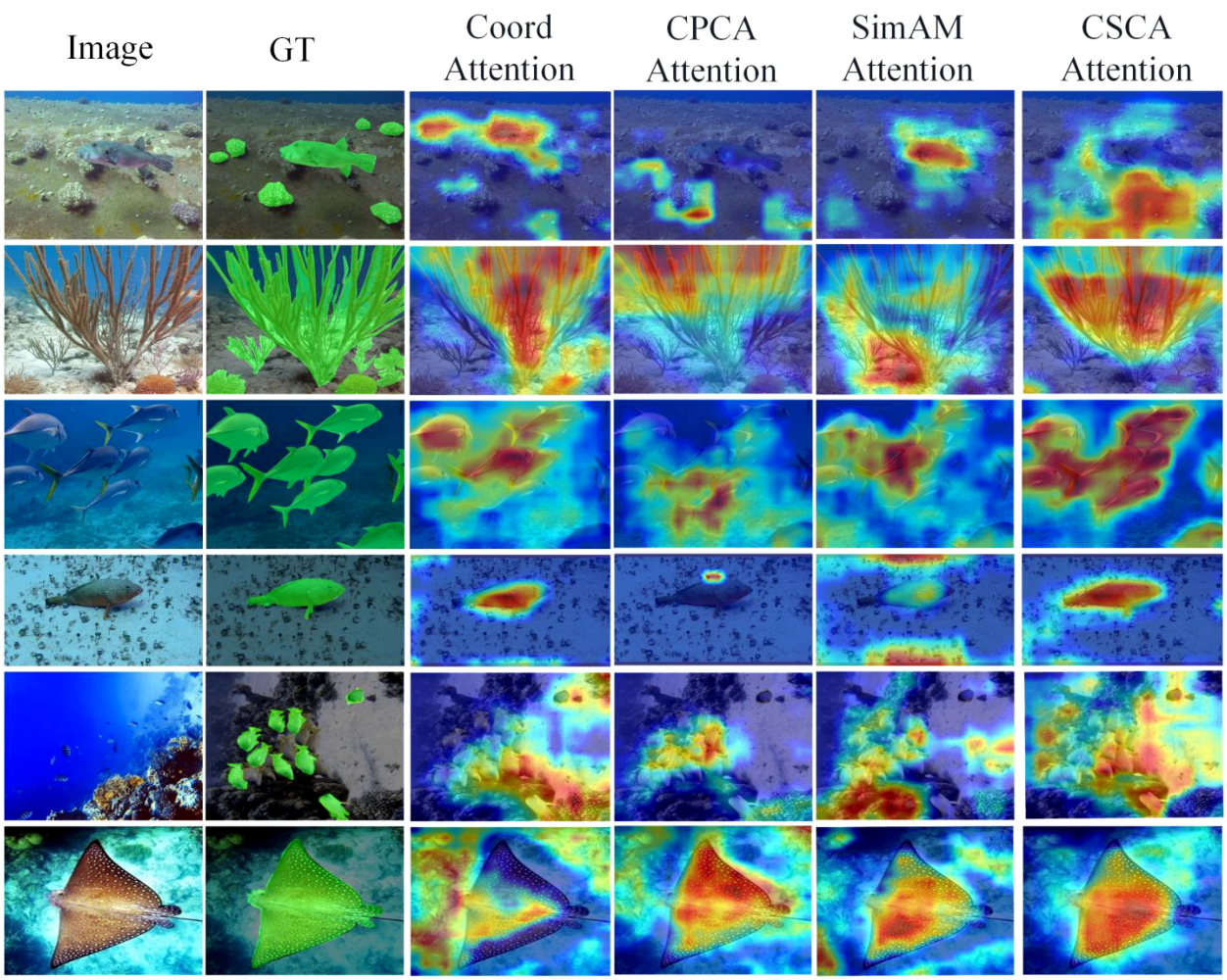

FIGURE 6
Comparison of GradCAM heat map results for several attentions.

TABLE 3 Experimental results in the UIIS dataset.

Model		AP@0.5/%							Parameters/ 10^6	GFLOPs	mAP@0.5/%
		Fish	Reefs	Aquatic plants	Wrecks/ ruins	Human divers	Robots	Sea-floor			
YOLOv11		63.7	44.1	12.6	25.8	79.1	1.7	17.0	2.8	10.2	34.9
YOLOv9		68.3	47.2	19.1	29.9	86.5	33.6	21.2	13.7	54.9	43.7
YOLOv8		66.6	43.9	15.0	22.9	83.2	19.4	18.3	3.2	12.0	37.5
YOLOv5		63.5	34.3	14.4	19.1	79.5	3.3	12.9	1.8	6.8	32.5
YOLOv11	+CSCA	63.6	43.2	16.2	29.1	83.8	6.8	14.1	4.2	13.4	36.7
YOLOv9	+CSCA	70.0	45.6	18.1	35.1	87.0	31.6	22.9	13.9	62.4	44.3
YOLOv8	+CSCA	66.6	43.8	17.0	27.0	88.0	24.0	16.6	4.5	14.8	40.7
YOLOv5	+CSCA	63.3	39.8	16.1	19.1	83.0	3.5	17.0	4.5	10.2	34.6
YOLOv11	+CAPAF	64.4	43.0	16.1	29.9	82.9	2.8	15.1	6.6	24.7	36.3
YOLOv9	+CAPAF	70.8	48.7	20.5	34.4	89.3	23.4	20.8	14.7	70.0	44.0

(Continued)

TABLE 3 Continued

Model		AP@0.5/%							Parameters/ 10^6	GFLOPs	mAP@0.5/%
		Fish	Reefs	Aquatic plants	Wrecks/ ruins	Human divers	Robots	Sea-floor			
YOLOv8		66.4	45.1	17.1	28.5	85.5	15.4	18.7	5.3	18.7	39.5
YOLOv5		63.2	37.9	14.0	20.2	80.8	1.7	13.9	4.5	10.2	33.1
YOLOv11	+CSCA +CAPAF	65.4	44.0	15.4	28.8	83.6	3.5	19.8	8.0	27.7	37.2
YOLOv9		70.1	47.0	19.2	32.0	88.2	33.7	22.7	18.8	78.5	44.7
YOLOv8		67.5	44.8	15.5	24.7	86.8	29.8	19.0	7.7	24.9	41.2
YOLOv5		62.1	38.5	15.5	20.6	82.4	6.7	19.1	4.5	10.2	35.0

TABLE 4 Experimental results in the USIS10K dataset.

Model		AP@0.5/%							Parameters/ 10^6	GFLOPs	mAP@0.5/%
		Fish	Reefs	Aquatic plants	Wrecks/ ruins	Human divers	Robots	Sea-floor			
YOLOv11		68.0	85.9	49.3	11.5	81.1	51.6	21.5	2.8	10.2	52.7
YOLOv9		75.7	89.9	53.6	13.4	86.0	62.5	27.1	13.7	54.9	58.3
YOLOv8		69.9	87.0	51.3	13.0	82.5	60.0	26.9	3.2	12.0	55.8
YOLOv5		62.5	84.3	44.3	10.1	78.0	58.2	24.4	1.8	6.8	51.7
YOLOv11	+CSCA +CAPAF	66.2	86.1	49.8	10.6	78.2	57.2	25.7	8.0	27.7	53.5
YOLOv9		76.1	91.4	56.1	14.2	87.9	60.7	31.2	18.8	78.5	59.7
YOLOv8		73.1	88.1	51.7	14.5	83.5	59.3	32.5	7.7	24.9	57.5
YOLOv5		63.5	86.0	48.4	11.0	82.0	64.6	29.7	4.5	10.2	55.0

mAP@0.5 of the YOLOv11 model improved from 52.7% to 53.5% after the introduction of the CSCA and CAPAF modules, which indicates that these two modules effectively enhance the model's ability to perceive complex underwater environments, thus verifying the important role of the CSCA and CAPAF modules in enhancing the performance of existing YOLO models.

5 Conclusion

In this study, we innovatively propose a CSCA attention mechanism and a CAPAF feature fusion module for underwater instance segmentation, aiming at solving the problem of insufficient segmentation accuracy of complex underwater scene images. Through the visual analysis of feature mapping, we find that the proposed CSCA attention module can focus more on the target region. During the experimental process, we integrated the CSCA attention mechanism with the CAPAF feature fusion module into the YOLOv8 model, which resulted in a 3.2% and 2% improvement in the mAP@0.5 metrics of the model, respectively. In addition, we apply this method to multiple networks for validation, and the experimental comparison results fully demonstrate the significant effectiveness of the method proposed in this paper in terms of segmentation accuracy of underwater image instances.

The current limitation of the CSCA attention and CAPAF feature fusion modules lies in the need for further lightweight optimization of the parameters and computational complexity. Additionally, these modules have only been validated in YOLO-based networks and not in other networks. However, experimental results have demonstrated that our proposed CSCA attention and CAPAF feature fusion modules can enhance model detection accuracy and prevent issues with missing segmentation edges. In the future, our work will continue to improve the CSCA attention and CAPAF modules and validate them in other networks.

Data availability statement

Publicly available datasets were analyzed in this study. This data can be found here: doi: 10.1109/ICCV51070.2023.00126.

Author contributions

ZH: Conceptualization, Data curation, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. LC: Formal Analysis, Funding acquisition, Writing – review & editing. XX: Validation, Writing – original draft. JX: Validation, Conceptualization, Writing – original draft.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. This work was supported by a grant from the Liaoning Provincial Education Department Scientific Research Funding Project (LJKZ0731).

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.

Generative AI statement

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

Publisher's note

References

Chen, Y., Zhang, L., Chen, B., Zuo, J., and Hu, Y. (2024). MPG-net: A semantic segmentation model for extracting aquaculture ponds in coastal areas from sentinel-2 MSI and planet superDove images. *Remote Sens.* 16 (20), 3760. doi: 10.3390/rs16203760

Fan, Z., Xia, W., Liu, X., and Li, H. (2021). Detection and segmentation of underwater objects from forward-looking sonar based on a modified Mask RCNN. *SIViP.* 15, 1135–1143. doi: 10.1007/s11760-020-01841-x

Farhadi Tolie, H., Ren, J., and Elyan, E. (2024). DICAM: deep inception and channel-wise attention modules for underwater image enhancement. *Neurocomputing.* 584, 127585. doi: 10.1016/j.neucom.2024.127585

Han, D., Ye, T., Han, Y., Xia, Z., Pan, S., Wan, P., et al. (2025). "Agent attention: on the integration of softmax and linear attention," in *Computer vision – ECCV 2024. Lect. Notes comput. Sci.* Eds. A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler and G. Varol (Cham: Springer), 15108. doi: 10.1007/978-3-031-72973-7_8

Han, Y., Zheng, B., Kong, X., Huang, J., Wang, X., Ding, T., et al. (2023). Underwater fish segmentation algorithm based on improved PSPNet network. *Sensors.* 23 (19), 8072. doi: 10.3390/s23198072

He, Z., Cao, L., Luo, J., Xu, X., Tang, J., Xu, J., et al. (2024). UISS-Net: Underwater Image Semantic Segmentation Network for improving boundary segmentation accuracy of underwater images. *Aquacult Int.* 32, 5625–5638. doi: 10.1007/s10499-024-01439-x

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).* (Nashville, TN, USA: IEEE), 13708–13717. doi: 10.1109/CVPR46437.2021.01350

Huang, H., Chen, Z., Zou, Y., Lu, M., Chen, C., Song, Y., et al. (2024). Channel prior convolutional attention for medical image segmentation. *Comput. Biol. Med.* 178, 108784. doi: 10.1016/j.compbiomed.2024.108784

Kannan, S. (2020). Intelligent object recognition in underwater images using evolutionary-based Gaussian mixture model and shape matching. *SIViP* 14, 877–885. doi: 10.1007/s11760-019-01619-w

Li, D., Xu, L., and Liu, H. (2017). Detection of uneaten fish food pellets in underwater images for aquaculture. *Aquacultural Eng.* 78, 85–94. doi: 10.1016/j.aquaeng.2017.05.001

Lian, S., Li, H., Cong, R., Li, S., Zhang, W., Kwong, S., et al. (2023). "WaterMask: instance segmentation for underwater imagery," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV).* (Paris, France: IEEE), 1305–1315. doi: 10.1109/ICCV51070.2023.00126

Lian, S., Zhang, Z., and Li, H. (2024). "Diving into Underwater: Segment Anything Model Guided Underwater Salient Instance Segmentation and A Large-scale Dataset," in *Proceedings of the 41st International Conference on Machine Learning (ICML'24), Vol. 235. JMLR.org, Article 1190.* 29545–29559. Available online at: https://dl.acm.org/doi/10.5555/3692070.3693260.

Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., and Belongie, S. (2017). "Feature pyramid networks for object detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* (Honolulu, HI, USA: IEEE), 936–944. doi: 10.1109/CVPR.2017.106

Liu, S., Qi, L., Qin, H., Shi, J., and Jia, J. (2018). "Path aggregation network for instance segmentation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* (Salt Lake City, UT, USA: IEEE), 8759–8768. doi: 10.1109/CVPR.2018.00913

Ma, Y., Feng, X., and Chao, L. (2018). "A new database for evaluating underwater image processing methods," in *Proceedings of the Eighth International Conference on Image Processing Theory, Tools and Applications (IPTA).* (Xi'an, China: IEEE), 1–6. doi: 10.1109/IPTA.2018.8608131

Peng, L., Zhu, C., and Bian, L. (2023). U-shape transformer for underwater image enhancement. *IEEE Trans. Image Processing.* 32, 3066–3079. doi: 10.1109/TIP.2023.3276332

Redmon, J., Divvala, S., Girshick, R., and Farhadi, A. (2016). "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* (Las Vegas, NV, USA: IEEE), 779–788. doi: 10.1109/CVPR.2016.91

Shen, X., Wang, H., Cui, T., Gou, Z., and Fu, X. (2024). Multiple information perception-based attention in YOLO for underwater object detection. *Vis. Comput.* 40, 1415–1438. doi: 10.1007/s00371-023-02858-2

Siri, D., Vellaturi, G., Ibrahim, S. H. S., Molugu, S., Desanamukula, V. S., Kocherla, R., et al. (2024). Enhanced deep learning models for automatic fish species identification in underwater imagery. *Heliyon.* 10, e35217. doi: 10.1016/j.heliyon.2024.e35217

Sun, H., Wen, Y., Feng, H., Zheng, Y., Mei, Q., Ren, D., et al. (2024). Unsupervised Bidirectional Contrastive Reconstruction and Adaptive Fine-Grained Channel Attention Networks for image dehazing. *Neural Networks.* 176, 106314. doi: 10.1016/j.neunet.2024.106314

Tan, M., Pang, R., and Le, Q. V. (2020). "EfficientDet: scalable and efficient object detection," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).* (Seattle, WA, USA: IEEE), 10778–10787. doi: 10.1109/CVPR42600.2020.01079

Wang, C., Duan, W., and Luan, C. (2024). USNet: underwater image superpixel segmentation via multi-scale water-net. *Front. Mar. Science.* 11. doi: 10.3389/fmars.2024.1411717

Woo, S., Park, J., Lee, J. Y., and Kweon, I. (2018). "CBAM: convolutional block attention module," in *Computer vision – ECCV 2018. ECCV 2018. Lecture notes in computer science.* Eds. V. Ferrari, M. Hebert, C. Sminchisescu and Y. Weiss (Cham: Springer), 11211. doi: 10.1007/978-3-030-01234-2_1

Yang, Y., Li, D., and Zhao, S. (2024). A novel approach for underwater fish segmentation in complex scenes based on multi-levels triangular atrous convolution. *Aquacult Int.* 32, 5215–5240. doi: 10.1007/s10499-024-01424-4

Yang, L., Zhang, R.-Y., and Li, L. (2021). "SimAM: A simple, parameter-free attention module for convolutional neural networks," in *Proceedings of the 38th International Conference on Machine Learning MLR, Virtual*, 139, 11863–11874. Available online at: http://proceedings.mlr.press/v139/yang21o.html.

Zhang, L., Qiu, Y., Fan, J., Li, S., Hu, Q., Xing, B., et al. (2024). Underwater fish detection and counting using image segmentation. *Aquacult Int.* 32, 4799–4817. doi: 10.1007/s10499-024-01402-w

Zhang, W., Wei, B., Li, Y., Li, H., and Song, T. (2024). WaterBiSeg-Net: An underwater bilateral segmentation network for marine debris segmentation. *Mar. Pollut. Bulletin.* 205, 116644. doi: 10.1016/j.marpolbul.2024.116644

Zhang, Z., Li, W., and Seet, B. C. (2024). A lightweight underwater fish image semantic segmentation model based on U-Net. *IET Image Process.* 18, 3143–3155. doi: 10.1049/ipr2.13161

Zheng, T., Wu, J., Kong, H., Zhao, H., QU, B., Liu, L., et al. (2024). A video object segmentation-based fish individual recognition method for underwater complex environments. *Ecol. Informatics.* 82, 102689. doi: 10.1016/j.ecoinf.2024.102689

Zhou, W., Wang, J., Song, Y., Zhang, X., Liu, Z., Ma, Y., et al. (2024). "MW-YOLO: improved YOLOv8n for lightweight dense vehicle object detection algorithm," in *2024 3rd International Conference on Image Processing and Media Computing (ICIPMC).* (Hefei, China: IEEE), 28–35. doi: 10.1109/ICIPMC62364.2024.10586598