The Visual Computer (2024) 40:1415–1438
https://doi.org/10.1007/s00371-023-02858-2

ORIGINAL ARTICLE

Multiple information perception-based attention in YOLO for underwater object detection

Xin Shen[1] · Huibing Wang[1] · Tianxiang Cui[1] · Zhicheng Guo[1] · Xianping Fu[1,2]

Accepted: 31 March 2023 / Published online: 25 May 2023

Abstract
Underwater object detection is a prerequisite for underwater robots to achieve autonomous operation and ocean exploration. However, poor imaging quality, harsh underwater environments, and concealed underwater targets greatly aggravate the difficulty of underwater object detection. In order to reduce underwater background interference and improve underwater object perception, we propose a multiple information perception-based attention module (MIPAM), which is mainly composed of five processes. In information preprocessing, spatial downsampling and channel splitting control parameters and computations of attention module by reducing dimension sizes. In information collection, channel-level information collection and spatial-level information collection enhance the semantic information expression by perceiving multi-dimensional dependency information, multi-dimensional structure information and multi-dimensional global information. In information interaction, channel-driven information interaction and spatial-driven information interaction stimulate the intrinsic interaction potential by further perceiving multi-dimensional diversity information. Adaptive feature fusion further improves the information interaction quality by allocating learnable parameters. In attention activation, the multi-branch structure enhances the attention calibration efficiency by generating multiple attention. In information postprocessing, channel concatenation and spatial upsampling realize the plug-and-play of attention module by restoring original feature states. In order to meet the high-precision and real-time requirements for underwater object detection, we integrate MIPAM into YOLO detectors. The experimental results indicate that our work brings significant performance gains for underwater detection tasks. Our work also provides some performance improvements for other detection tasks, which shows the ideal generalization ability.

Keywords Underwater object detection · Information perception · Attention mechanism · YOLO detector

1 Introduction

With the development of computer vision, object detection has achieved exciting results in many environments. However, facing with underwater environments, detection performance suffers from severe degradation. There are multiple irresistible factors that make underwater object detection become an extremely challenging task. First, underwater imaging quality is poor. During underwater propagation, light is often affected by suspended particles in the water. The absorption and scattering of light cause low contrast and colour cast in underwater images. The underwater robot is easily affected by ocean current during its movement. The irregular dithering causes texture distortion and detail blurring in underwater images. Second, underwater environments have strong randomness. A large number of sands, reefs, waterweeds and other interferences seriously block the underwater targets. The moving and grasping operations of underwater robot lead to the underwater dynamic turbidity. Third, underwater targets have high concealment. The underwater targets tend to have protective color and small size after long-term evolution. These underwater creatures

✉ Xianping Fu
fxp@dlmu.edu.cn

Xin Shen
shenxin@dlmu.edu.cn

Huibing Wang
huibing.wang@dlmu.edu.cn

Tianxiang Cui
a1065242944@dlmu.edu.cn

Zhicheng Guo
gzc15735162249@dlmu.edu.cn

[1] The School of Information Science and Technology, Dalian Maritime University, Dalian 116026, China

[2] The Peng Cheng Laboratory, Shenzhen 518000, China

always blend in with their surroundings to avoid attack. The poor imaging quality, harsh underwater environments, and concealed underwater targets lead to strong underwater background interference and weak underwater object perception, which greatly aggravates the difficulty of underwater detection tasks. It is worth noting that attention mechanism has been widely used in computer vision [1–5], which can extract important information from massive information by recalibrating features. In order to reduce the underwater background interference and improve the underwater object perception, we focus on the selective attention in this paper.

For attention modules, information collection and information interaction are two crucial components. Information collection is responsible for capturing intrinsic information from input features. Information interaction is responsible for stimulating the potential of intrinsic information. In information collection, the channel-wise global average pooling [6–13], channel-wise L2-norm [14] and channel-wise discrete cosine transform [15] process features from (C, H, W) to $(C, 1, 1)$, which capture spatial global information and channel structure information. The spatial-wise global average pooling [8, 16] processes features from (C, H, W) to $(1, H, W)$, which captures channel global information and spatial structure information. The spatial-wise 1×1 convolution [7, 17–19] processes features from (C, H, W) to $\left(C', H, W\right)$, which also captures channel global information and spatial structure information. The cross-channel global covariance pooling [18] and cross-spatial global covariance pooling [18] process features from (C, H, W) to $(C, C, 1)$ and $(1, HW, HW)$, which capture channel dependency information and spatial dependency information, respectively. In information interaction, almost all attention modules follow the traditional convolution idea. By assigning different parameters in channel dimensions and sharing same parameters in spatial dimensions, the feature information realizes active channel interaction and passive spatial interaction.

Although various attention modules have made great contributions, there are still two problems. First, the deficiency of information collection leads to the weakening of feature expression ability. Second, the passive interaction of spatial features reduces the quality of intrinsic information interaction. The negative effects brought by these problems are exacerbated in harsh underwater environments. In order to design an attention module more suitable for underwater detection tasks, we enhance the semantic information expression through richer information perception and stimulate the intrinsic interaction potential through more comprehensive active interaction.

In this paper, we propose a multiple information perception-based attention module (MIPAM). For information collection, channel-level information collection and spatial-level information collection are designed to perceive multi-dimensional dependency information, multi-dimensional structure information and multi-dimensional global information. In channel-level information collection, the cross-channel global covariance pooling perceives channel dependency information. The channel-wise global average pooling perceives channel structure information and spatial global information. In spatial-level information collection, the spatial-wise global average pooling perceives spatial structure information and channel global information. The cross-spatial global covariance pooling perceives spatial dependency information. For information interaction, channel-driven information interaction and spatial-driven information interaction are designed to further perceive multi-dimensional diversity information. In channel-driven information interaction, channel diversity information was perceived by allocating different parameters in channel dimension and sharing same parameters in spatial dimension. In spatial-driven information interaction, spatial diversity information was perceived by allocating different parameters in spatial dimension and sharing same parameters in channel dimension. Our MIPAM is integrated into the YOLO detector to achieve efficient object detection in harsh underwater environments. The main contributions of our work are summarized as follows:

- We propose a multiple information perception-based attention module(MIPAM), which reduces underwater background interference and improves underwater object perception.
- We design channel-level information collection and spatial-level information collection to perceive multi-dimensional dependency information, multi-dimensional structure information and multi-dimensional global information. This richer information perception enhances the semantic information expression.
- We design channel-driven information interaction and spatial-driven information interaction to further perceive multi-dimensional diversity information. This more comprehensive active interaction stimulates the intrinsic interaction potential.
- We integrate MIPAM into YOLO detector, which meets the high-precision and real-time requirements for underwater object detection.

The remainder of this paper is organized as follows. In Sect. 2, we review the related works on underwater object detection, attention mechanism and YOLO detection algorithm. In Sect. 3, we introduce the proposed method in detail. Experiments and results are provided in Sect. 4. The conclusion about our work is summarized in Sect. 5.

2 Related works

In this section, we analyze underwater object detection, attention mechanism and YOLO detection algorithm from three different perspectives, and discuss the differences and connections between our work and other works.

2.1 Underwater object detection

According to the different underwater imaging systems, underwater object detection algorithms can be divided into acoustic image-based underwater object detection algorithm [20, 21] and optical image-based underwater object detection algorithm [22, 23]. Acoustic underwater detection has great advantages in underwater remote detection tasks, and has a good detection effect on large underwater objects. However, in the marine ranching application, we need to complete accurate underwater detection tasks in a close range, so as to facilitate autonomous capture and dynamic statistics of small marine treasures by underwater robots. Optical underwater images have close-range imaging properties. Therefore, our research focuses on the optical underwater detection.

According to different underwater application technologies, underwater object detection algorithms can be further divided into traditional features-based underwater object detection algorithm [24, 25] and deep learning-based underwater object detection algorithm [26–29]. Traditional underwater detection uses manual feature design to extract low-level feature descriptors from underwater images. This method cannot describe complex target information effectively, and it cannot adapt to the strong randomness of underwater environments. Traditional underwater detection has problems such as weak feature extraction ability, poor robustness and low generalization, which cannot put into the actual underwater application.

It is worth noting that deep learning has driven the rapid development of the computer vision field. However, the development of underwater object detection has been relatively slow [30–32]. Although popular object detection algorithms using deep learning have achieved encouraging results, it is not ideal to apply these algorithms directly to the underwater environment. Obviously, common methods to improve the performance of neural networks, such as directly increasing the depth, width, and cardinality in the network, cannot effectively solve the severe problems faced by underwater object detection, which mainly refers to the poor imaging quality, harsh underwater environments, and concealed underwater targets. At present, underwater detection algorithms tend to improve the underwater detection performance from two different perspectives: 1. Data enhancement techniques [33], such as splicing and overlapping, are adopted to improve the dataset quality. 2. Network construction techniques [34, 35], such as residual connection and feature pyramid, are used to improve the network performance. This simple performance gain is mainly due to the improvement of dataset quality and network performance. The core problems of strong underwater background interference and weak underwater object perception have not been solved effectively. In practical underwater applications, underwater detection algorithms still have some problems, such as poor robustness and weak generalization.

Based on the above considerations, our work focuses on exploring the application potential of attention mechanisms in complex underwater environments and exploring the optimal attention design suitable for underwater detection tasks. With the core goal of reducing underwater background interference and improving underwater object perception, this paper is committed to addressing the underwater detection challenges from the essence of the problem, which plays a positive role in the research and development of underwater object detection.

2.2 Attention mechanism

According to different design needs, researchers have proposed various attention modules in computer vision. Channel attention focuses on adjusting the importance of channel dimensions. Spatial attention focuses on regulating the importance of spatial dimensions. Hybrid attention is responsible for simultaneously calibrating the importance of channel and spatial dimensions.

Channel attention The squeeze-and-excitation module (SEM) [6] learned the importance of each channel, and used bottleneck structure to reduce parameters and computations. The style-based recalibration module (SRM) [9] used global average pooling and global standard deviation pooling to collect channel-wise style information, and used channel-wise fully connected layer to achieve style integration. The efficient channel attention module (ECAM) [11] adaptively selected the kernel size of 1D convolution to better determine the coverage of local cross-channel interaction. The gated channel transformation module (GCTM) [14] used L2-norm with learnable parameters to replace GAP and FC in traditional attention modules, which captured the competition and cooperation between channel features. The frequency channel attention module (FCAM) [15] grouped the input features and used two-dimensional discrete cosine transform priors to capture the feature information of these groupings.

Spatial attention The double attention module (A2M) [17] used softmax to adaptively adjust the attention weight, and used bilinear pooling to collect the entire spatial information. The information was adaptively distributed to each spatial location. A2M generated two different attentions simultaneously. The spatial group-wise enhance module (SGEM) [10] grouped the channel dimensions and used global aver-

age pooling to gather spatial information for sub-features. The information was passed to all spatial locations for feature enhancement. SGEM learned rich information by generating spatial attention maps in each group, which was lightweight.

Hybrid attention The bottleneck attention module (BAM) [7] combined channel and spatial attentions in parallel, and used multiple dilated convolutions to expand the spatial receptive field. The convolutional block attention module (CBAM) [8] combined channel and spatial attentions in series, and used max pooling and average pooling to enrich receptive fields in different dimensions. The global second-order pooling module (GSoPM) [18] captured the second-order statistics by calculating the covariance matrices on channel and spatial dimensions. GSoPM considered long-range correlations through high-order modeling. The relation-aware global attention module (RGAM) [19] used two embedding functions to generate bi-directional correlations between feature points. For each feature position, the correlations between each feature and all features were stacked, and the features themselves were concatenated to activate attention at the current location.

Although the above attention modules have achieved exciting results in different applications, they still perform suboptimally in underwater environments. In order to design attention more suitable for underwater applications, here we focus on analyzing various attention modules from the perspective of information collection and information interaction. Table 1 reports the differences of these attention modules in detail, where checkmarked and unmarked positions indicate the factors considered and ignored in the module design, respectively

2.3 YOLO detection algorithm

In this paper, we focus on choosing the YOLO (You Only Look Once) series as the baseline methods. The main reason is that the one-stage YOLO detector can better balance detection accuracy and detection speed. Only detection algorithms with both high-accuracy and real-time performance can adapt to the complex and variable underwater detection tasks. In addition, the YOLO detectors can flexibly adjust the network size, where the parameters, computations and memory consumption can be controlled within the desired range. This will facilitate us to directly carry the algorithm to the underwater robot for practical underwater applications. Although the two-stage object detectors [36, 37] or transformer series [38, 39] can achieve high detection accuracy, its memory consumption and detection speed are not friendly for underwater detection tasks.

Redmon et al. proposed YOLOV1 [40], YOLOV2 [41] and YOLOV3 [42]. YOLOV1 used GoogleLeNet as the backbone, which had ideal inference speed and generalization

Table 1 Detailed analyses of attention modules from information collection and information interaction

Attention modules	Information collection						Information interaction			
	Channel global information	Channel global information	Channel structure information	Channel dependency information	Spatial information	Spatial global information	Spatial structure information	Spatial dependency information	Channel parameter drive	Spatial parameter drive
SEM [6]			✓		✓				✓	
BAM [7]	✓		✓		✓		✓		✓	
CBAM [8]	✓		✓		✓		✓		✓	
A2M [17]	✓						✓		✓	
SRM [9]			✓		✓				✓	
SGEM [10]			✓		✓				✓	
GSoPM [18]	✓			✓			✓	✓	✓	
ECAM [11]			✓		✓				✓	
GCTM [14]			✓		✓				✓	
FCAM [15]			✓		✓				✓	
MIPAM(Ours)	✓		✓	✓	✓		✓	✓	✓	✓

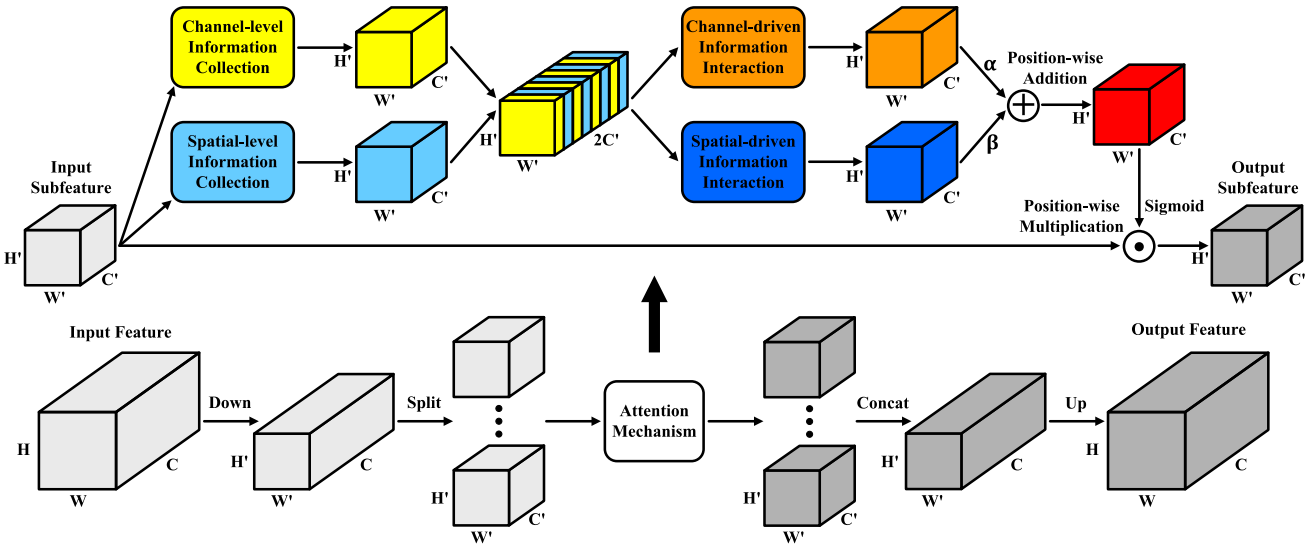


Fig. 1 The design architecture of multiple information perception-based attention module (MIPAM)

ability. YOLOV2 used DarkNet19 as the backbone and introduced the idea of anchor boxes. The multi-scale training method improved the robustness of YOLOV2 on images with different sizes. The backbone used by YOLOV3 was DarkNet53. YOLOV3 applied the residual structure to better extract features, and applied feature pyramid networks (FPN) for feature fusion. The multi-scale prediction strategy was used to better detect objects with different scales. Compared with YOLOV1 and YOLOV2, YOLOV3 can achieve a better balance of speed and accuracy.

Bochkovskiy et al. [43] proposed YOLOV4, which combined various tricks in deep learning. YOLOV4 introduced mosaic data augmentation and cross mini-batch normalization at the input. CSPDarkNet53, Mish activation function and DropBlock regularization were used in the backbone. The spatial pyramid pooling (SPP) module and path aggregation network (PAN) structure were borrowed in the neck. In the head, the loss computation and non-maximum suppression were performed based on complete-intersection over union (CIOU) and distance-intersection over union (DIOU), respectively. Compared with the previous versions, YOLOV4 has stronger performance. YOLOV5 was proposed in [44], which had the similar network structure to YOLOV4. In the backbone, YOLOV5 added Focus and SPP structures, and tweaked the implementation details, which can be called modified CSPDarkNet. The cross stage partial (CSP) structure is further used in the neck to strengthen the feature fusion ability of the network. Adaptive anchor box calculating and adaptive image scaling were applied at the input. YOLOV5 has stronger flexibility, which can achieve rapid deployment.

YOLOV6 [45] designed the EfficientRep backbone and the Rep-PAN neck based on RepVGG style. The decoupled head is further optimized by reducing overhead. YOLOV6 adopted the anchor-free training strategy and the SimOTA label assignment strategy to further improve the detection accuracy. For YOLOV7 [46], the extended efficient long-range attention network (Extended-ELAN) improved model learning ability without destroying the original gradient path. The concatenation-based model scaling method maintained the optimal structure of the model design. The planned re-parameterized convolution effectively increased model inference speed. The dynamic label assignment strategy with coarse-to-fine guidance provided better dynamic targets for different branches. Ge et al. [47] proposed YOLOX based on YOLOV3. YOLOX used an anchor-free strategy to reduce the complexity of the detection head, and used the decoupled head to improve the model convergence speed. The SimOTA strategy was applied to the loss computation, which is able to dynamically match positive samples for objects with different sizes. In general, YOLOX has more superior performance in terms of speed and accuracy.

3 Proposed method

In this section, we first introduce the design architecture of multiple information perception-based attention module (MIPAM). Then, we elaborate the information collection in MIPAM, which includes channel-level information collection and spatial-level information collection. Subsequently, we elaborate the information interaction in MIPAM, which includes channel-driven information interaction and spatial-driven information interaction. Finally, we provide the application of MIPAM in the YOLO detector.

3.1 Multiple information perception-based attention module (MIPAM)

Figure 1 highlights the design architecture of multiple information perception-based attention module (MIPAM). MIPAM is mainly composed of five processes: information preprocessing, information collection, information interaction, attention activation and information postprocessing.

Information preprocessing. Input feature $\mathbf{X} \in \mathbb{R}^{C\times H\times W}$ is first downsampled to feature $\mathbf{x} \in \mathbb{R}^{C\times H'\times W'}$ by using group convolution, batch normalization and ReLU function, where group is set to C. $\mathbf{x} \in \mathbb{R}^{C\times H'\times W'}$ is further split into input subfeature $\mathbf{x}_i \in \mathbb{R}^{C'\times H'\times W'}$ along the channel dimension, where non-overlapping split is set to g and $i \in [1, ..., g]$. The information preprocessing of MIPAM is formulated as:

$$\mathbf{x}_i = \text{Split}\left(\text{Down}\left(\mathbf{X}\right)\right) \tag{1}$$

where Down and Split represent downsampling and split operations, respectively. These two operations can reduce spatial and channel dimensions respectively, which are beneficial to control the subsequent parameter amount and computational cost.

Information collection Input subfeature $\mathbf{x}_i \in \mathbb{R}^{C'\times H'\times W'}$ is first processed into feature $\mathbf{x}_i^c \in \mathbb{R}^{C'\times H'\times W'}$ and feature $\mathbf{x}_i^s \in \mathbb{R}^{C'\times H'\times W'}$ by using channel-level information collection and spatial-level information collection, respectively. $\mathbf{x}_i^c \in \mathbb{R}^{C'\times H'\times W'}$ and $\mathbf{x}_i^s \in \mathbb{R}^{C'\times H'\times W'}$ are further cross-concatenated into feature $\mathbf{x}_i^{cs} \in \mathbb{R}^{2C'\times H'\times W'}$ in the channel dimension. The information collection of MIPAM is formulated as:

$$\mathbf{x}_i^{cs} = CConcat\left(f_{\text{clic}}\left(\mathbf{x}_i\right), f_{\text{slic}}\left(\mathbf{x}_i\right)\right) \tag{2}$$

where f_{clic}, f_{slic} and $CConcat$ represent channel-level information collection, spatial-level information collection and cross concatenation, respectively. Channel-level information collection can perceive channel dependency information, channel structure information and spatial global information by using cross-channel global covariance pooling and channel-wise global average pooling. Spatial-level information collection can perceive spatial dependency information, spatial structure information and channel global information by using cross-spatial global covariance pooling and spatial-wise global average pooling. Cross concatenation can organize the perceived multiple information to facilitate subsequent information interaction.

Information interaction Feature $\mathbf{x}_i^{cs} \in \mathbb{R}^{2C'\times H'\times W'}$ is first processed into feature $\mathbf{x}_i^{c's} \in \mathbb{R}^{C'\times H'\times W'}$ and feature $\mathbf{x}_i^{cs'} \in \mathbb{R}^{C'\times H'\times W'}$ by using channel-driven information interaction and spatial-driven information interaction, respectively. $\mathbf{x}_i^{c's} \in \mathbb{R}^{C'\times H'\times W'}$ and $\mathbf{x}_i^{cs'} \in \mathbb{R}^{C'\times H'\times W'}$ are further adaptively fused into feature $\mathbf{x}_i^{c's'} \in \mathbb{R}^{C'\times H'\times W'}$ by assigning learnable parameters $\alpha_i \in \mathbb{R}^{C'\times 1\times 1}$ and $\beta_i \in \mathbb{R}^{C'\times 1\times 1}$. The information interaction of MIPAM is formulated as:

$$\mathbf{x}_i^{c's'} = \alpha_i f_{\text{cdii}}\left(\mathbf{x}_i^{cs}\right) + \beta_i f_{\text{sdii}}\left(\mathbf{x}_i^{cs}\right) \tag{3}$$

where f_{cdii} and f_{sdii} represent channel-driven information interaction and spatial-driven information interaction, respectively. Channel-driven information interaction can perceive channel diversity information by assigning different parameters in the channel dimension and sharing same parameters in the spatial dimension. Spatial-driven information interaction can perceive spatial diversity information by assigning different parameters in the spatial dimension and sharing same parameters in the channel dimension.

Attention activation Feature $\mathbf{x}_i^{c's'} \in \mathbb{R}^{C'\times H'\times W'}$ is first activated into the attention map by using sigmoid function. The attention map is further applied to input subfeature $\mathbf{x}_i \in \mathbb{R}^{C'\times H'\times W'}$ to obtain output subfeature $\mathbf{x'}_i \in \mathbb{R}^{C'\times H'\times W'}$. The attention activation of MIPAM is formulated as:

$$\mathbf{x'}_i = \mathbf{x}_i Sigmoid\left(\mathbf{x}_i^{c's'}\right) \tag{4}$$

where Sigmoid represents the sigmoid function. The sigmoid function can achieve importance distinction by activating feature values between 0 and 1. It is worth noting that input subfeatures are processed as output subfeatures on all branches. This multi-branch structure is beneficial to activate diverse attention, which can perceive valuable feature information on different branches in a targeted manner.

Information postprocessing Output subfeature $\mathbf{x'}_i \in \mathbb{R}^{C'\times H'\times W'}$ on each branch is first concatenated into feature $\mathbf{y} \in \mathbb{R}^{C\times H'\times W'}$ along the channel dimension. $\mathbf{y} \in \mathbb{R}^{C\times H'\times W'}$ is further upsampled to output feature $\mathbf{Y} \in \mathbb{R}^{C\times H\times W}$ by using bilinear interpolation along the spatial dimension. The information postprocessing of MIPAM is formulated as:

$$\mathbf{Y} = Up\left(Concat\left(\mathbf{x'}_i\right)\right) \tag{5}$$

where Concat and Up represent concatenation and upsampling operations, respectively. These two operations can restore channel and spatial dimensions to the original state respectively, which are beneficial to realize the plug-and-play of attention module.

3.2 Information collection in MIPAM

In this subsection, we highlight more details about the information collection of MIPAM. For MIPAM, the information collection is mainly composed of two crucial processes:

channel-level information collection and spatial-level information collection.

Channel-level information collection Input subfeature $\mathbf{x}_i \in \mathbb{R}^{C' \times H' \times W'}$ is processed into feature $\mathbf{x}_i^1 \in \mathbb{R}^{C' \times C' \times 1}$ by using cross-channel global covariance pooling, which computes the covariance statistic among all channel dimensions. More specifically, we perform the covariance computation on all channel features $\mathbf{x}_{i\dot{c}} \in \mathbb{R}^{1 \times H' \times W'}$ to capture channel dependency information, where $\dot{c} = \left[1, ..., C'\right]$. In cross-channel global covariance pooling, the covariance calculation is defined as:

$$Cov\left(\mathbf{x}_{i\dot{c}}, \mathbf{x}_{i\dot{c}}\right) = \frac{\sum_{a=1}^{H'W'} \left(\mathbf{x}_{i\dot{c}}^a - \bar{\mathbf{x}}_{i\dot{c}}\right)\left(\mathbf{x}_{i\dot{c}}^a - \bar{\mathbf{x}}_{i\dot{c}}\right)}{H'W' - 1} \tag{6}$$

where $\bar{\mathbf{x}}_{i\dot{c}}$ is the mean of $\mathbf{x}_{i\dot{c}}$. Here, feature $\mathbf{x}_i^1$ is represented as:

$$\mathbf{x}_i^1 = \begin{bmatrix} Cov\left(\mathbf{x}_{i1}, \mathbf{x}_{i1}\right) & \cdots & Cov\left(\mathbf{x}_{i1}, \mathbf{x}_{iC'}\right) \\ \vdots & \ddots & \vdots \\ Cov\left(\mathbf{x}_{iC'}, \mathbf{x}_{i1}\right) & \cdots & Cov\left(\mathbf{x}_{iC'}, \mathbf{x}_{iC'}\right) \end{bmatrix} \tag{7}$$

Input subfeature $\mathbf{x}_i \in \mathbb{R}^{C' \times H' \times W'}$ is processed into feature $\mathbf{x}_i^2 \in \mathbb{R}^{C' \times 1 \times 1}$ by using channel-wise global average pooling, which computes the average statistic for each channel dimension. More specifically, we perform the average computation on each channel feature $\mathbf{x}_{i\dot{c}} \in \mathbb{R}^{1 \times H' \times W'}$ to capture spatial global information and preserve channel structure information. In channel-wise global average pooling, the average calculation is defined as:

$$Ave\left(\mathbf{x}_{i\dot{c}}\right) = \frac{\sum_{a=1}^{H'W'} \mathbf{x}_{i\dot{c}}^a}{H'W'} \tag{8}$$

where $a = \left[1, ..., H'W'\right]$. Here, feature $\mathbf{x}_i^2$ is represented as:

$$\mathbf{x}_i^2 = \left[Ave\left(\mathbf{x}_{i1}\right), ..., Ave\left(\mathbf{x}_{iC'}\right)\right] \tag{9}$$

These two pooling operations are executed in parallel. We then fuse $\mathbf{x}_i^1 \in \mathbb{R}^{C' \times C' \times 1}$ and $\mathbf{x}_i^2 \in \mathbb{R}^{C' \times 1 \times 1}$ into feature $\mathbf{x}_i^c \in \mathbb{R}^{C' \times H' \times W'}$ using matrix multiplication and upsampling operations. The channel-level information collection is formulated as follows:

$$f_{clic}\left(\mathbf{x}_i\right) = Up\left(MM\left(CcGCP\left(\mathbf{x}_i\right), CwGAP\left(\mathbf{x}_i\right)\right)\right) \tag{10}$$

where $CcGCP$, $CwGAP$, MM and Up represent cross-channel global covariance pooling, channel-wise global average pooling, matrix multiplication and upsampling operations, respectively. Cross-channel global covariance pooling is responsible for perceiving channel dependency information. Channel-wise global average pooling is responsible for perceiving channel structure information and spatial global information. Matrix multiplication and upsampling operations are responsible for fusing the perceived information and adjusting the feature shape.

Spatial-level information collection Input subfeature $\mathbf{x}_i \in \mathbb{R}^{C' \times H' \times W'}$ is processed into feature $\mathbf{x}_i^3 \in \mathbb{R}^{1 \times H' \times W'}$ by using spatial-wise global average pooling, which computes the average statistic for each spatial dimension. More specifically, we perform the average computation on each spatial feature $\mathbf{x}_{i\dot{s}} \in \mathbb{R}^{C' \times 1 \times 1}$ to capture channel global information and preserve spatial structure information, where $\dot{s} = \left[1, ..., H'W'\right]$. In spatial-wise global average pooling, the average calculation is defined as:

$$Ave\left(\mathbf{x}_{i\dot{s}}\right) = \frac{\sum_{b=1}^{C'} \mathbf{x}_{i\dot{s}}^b}{C'} \tag{11}$$

where $b = \left[1, ..., C'\right]$. Here, feature $\mathbf{x}_i^3$ is represented as:

$$\mathbf{x}_i^3 = \left[Ave\left(\mathbf{x}_{i(1)}\right), ..., Ave\left(\mathbf{x}_{i(H'W')}\right)\right] \tag{12}$$

Input subfeature $\mathbf{x}_i \in \mathbb{R}^{C' \times H' \times W'}$ is processed into feature $\mathbf{x}_i^4 \in \mathbb{R}^{1 \times H'W' \times H'W'}$ by using cross-spatial global covariance pooling, which computes the covariance statistic among all spatial dimensions. More specifically, we perform the covariance computation on all spatial features $\mathbf{x}_{i\dot{s}} \in \mathbb{R}^{C' \times 1 \times 1}$ to capture spatial dependency information. In cross-spatial global covariance pooling, the covariance calculation is defined as:

$$Cov\left(\mathbf{x}_{i\dot{s}}, \mathbf{x}_{i\dot{s}}\right) = \frac{\sum_{b=1}^{C'} \left(\mathbf{x}_{i\dot{s}}^b - \bar{\mathbf{x}}_{i\dot{s}}\right)\left(\mathbf{x}_{i\dot{s}}^b - \bar{\mathbf{x}}_{i\dot{s}}\right)}{C' - 1} \tag{13}$$

where $\bar{\mathbf{x}}_{i\dot{s}}$ is the mean of $\mathbf{x}_{i\dot{s}}$. Here, feature $\mathbf{x}_i^4$ is represented as:

$$\mathbf{x}_i^4 = \begin{bmatrix} Cov\left(\mathbf{x}_{i(1)}, \mathbf{x}_{i(1)}\right) & \cdots & Cov\left(\mathbf{x}_{i(1)}, \mathbf{x}_{i(H'W')}\right) \\ \vdots & \ddots & \vdots \\ Cov\left(\mathbf{x}_{i(H'W')}, \mathbf{x}_{i(1)}\right) & \cdots & Cov\left(\mathbf{x}_{i(H'W')}, \mathbf{x}_{i(H'W')}\right) \end{bmatrix} \tag{14}$$

These two pooling operations are also performed in parallel. We then fuse $\mathbf{x}_i^3 \in \mathbb{R}^{1 \times H' \times W'}$ and $\mathbf{x}_i^4 \in \mathbb{R}^{1 \times H'W' \times H'W'}$ into feature $\mathbf{x}_i^s \in \mathbb{R}^{C' \times H' \times W'}$ using matrix multiplication, reshape and upsampling operations. The spatial-level information collection is formulated as follows:

$$f_{\text{slic}}\left(\mathbf{x}_i\right) = Up\left(MM\left(SwGAP(\mathbf{x}_i)^{\Delta}, CsGCP\left(\mathbf{x}_i\right)\right)^{\Delta}\right) \tag{15}$$

where $SwGAP$, $CsGCP$ and Δ represent spatial-wise global average pooling, cross-spatial global covariance pool-

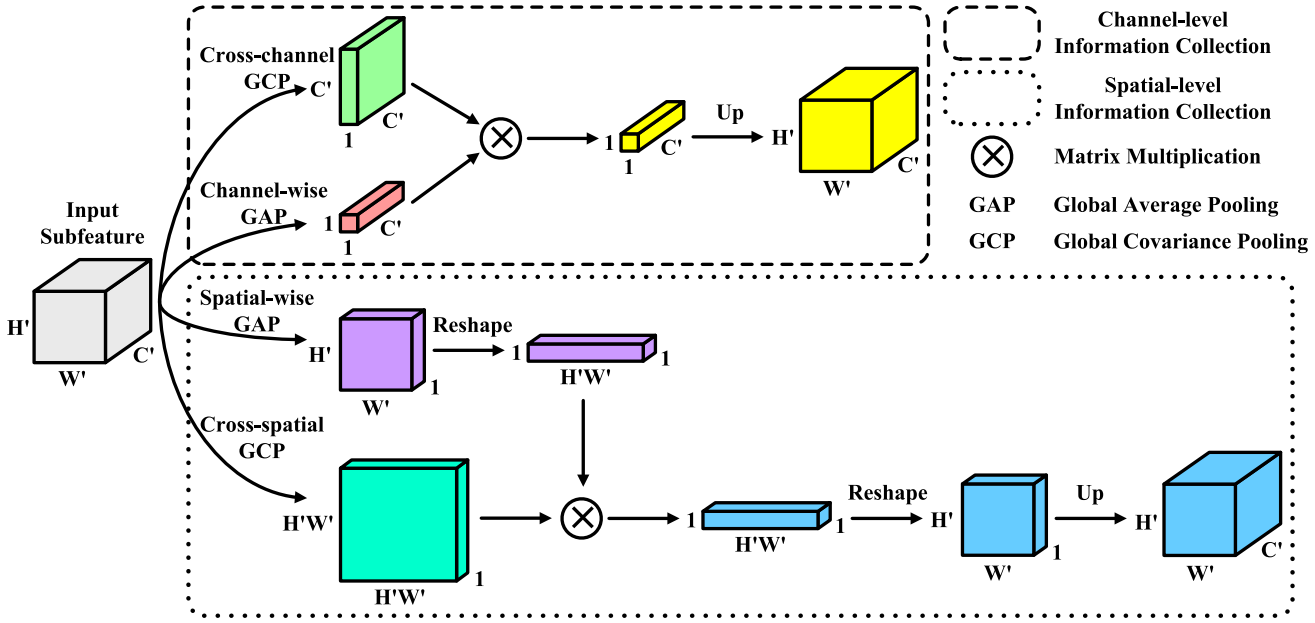


Fig. 2 The channel-level information collection and spatial-level information collection in MIPAM

ing and reshape operations, respectively. Spatial-wise global average pooling is responsible for perceiving spatial structure information and channel global information. Cross-spatial global covariance pooling is responsible for perceiving spatial dependency information. The reshape operation is also responsible for adjusting the feature to the desired shape for subsequent processing.

Figure 2 shows the channel-level information collection and spatial-level information collection in MIPAM. Our attention module simultaneously realizes the perception of multi-dimensional dependency information, multi-dimensional global information, and multi-dimensional structure information in information collection. We enhance the feature expression abilities with richer information collection.

3.3 Information interaction in MIPAM

In this subsection, we highlight more details about the information interaction of MIPAM. For MIPAM, the information interaction is mainly composed of two crucial processes: channel-driven information interaction and spatial-driven information interaction.

Channel-driven information interaction At this stage, feature $\mathbf{x}_i^{cs} \in \mathbb{R}^{2C' \times H' \times W'}$ is directly processed into feature $\mathbf{x}_i^{c's} \in \mathbb{R}^{C' \times H' \times W'}$ by using group convolution, batch normalization and ReLU function. It is worth noting that $\mathbf{x}_i^{cs} \in \mathbb{R}^{2C' \times H' \times W'}$ is formed by cross-concatenating $\mathbf{x}_i^{c} \in \mathbb{R}^{C' \times H' \times W'}$ and $\mathbf{x}_i^{s} \in \mathbb{R}^{C' \times H' \times W'}$ along the channel dimension, where $\mathbf{x}_i^{c} \in \mathbb{R}^{C' \times H' \times W'}$ and $\mathbf{x}_i^{s} \in \mathbb{R}^{C' \times H' \times W'}$ are the features generated after channel-level information collection and spatial-level information collection, respectively. The channel-driven information interaction is formulated as follows:

$$f_{\text{cdii}}\left(\mathbf{x}_i^{cs}\right) = GConv_{++}\left(\mathbf{x}_i^{cs}\right) \tag{16}$$

where $GConv_{++}$ represents the combination of group convolution, batch normalization and ReLU function. Here, the input channels, output channels, kernel size, stride, padding and grouping in 2D group convolution are set as $2C'$, C', 3, 1, 1 and C' respectively. There is no interference in the information interaction of each group. By allocating different parameters in the channel dimension and sharing same parameters in the spatial dimension, the channel-driven information interaction not only realizes the interactive fusion of multiple information from the channel-level information collection and spatial-level information collection, but also further perceives the channel diversity information.

Spatial-driven information interaction At this stage, feature $\mathbf{x}_i^{cs} \in \mathbb{R}^{2C' \times H' \times W'}$ is processed into feature $\mathbf{x}_i^{cs'} \in \mathbb{R}^{C' \times H' \times W'}$ through three steps. First, $\mathbf{x}_i^{cs} \in \mathbb{R}^{2C' \times H' \times W'}$ is reshaped and split into feature $\mathbf{x}_{ij}^{cs} \in \mathbb{R}^{1 \times H'W' \times 2}$, where $j \in [1, ..., C']$. Next, we process $\mathbf{x}_{ij}^{cs} \in \mathbb{R}^{1 \times H'W' \times 2}$ into $\mathbf{x}_{ij}^{cs} \in \mathbb{R}^{1 \times H' \times W'}$ using group convolution, batch normalization, ReLU function, row-wise sum and reshape operations in sequence. Finally, $\mathbf{x}_{ij}^{cs} \in \mathbb{R}^{1 \times H' \times W'}$ is concatenated into $\mathbf{x}_i^{cs'} \in \mathbb{R}^{C' \times H' \times W'}$ along the channel dimension. The spatial-driven information interaction is formulated as follows:

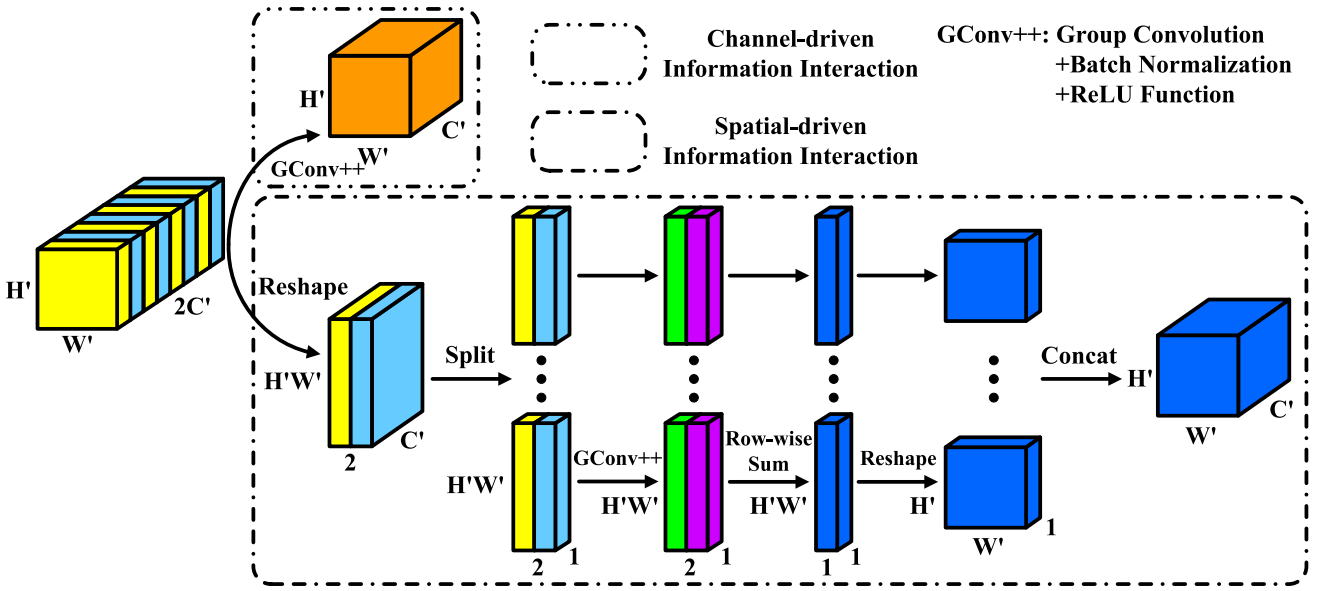


Fig. 3 The channel-driven information interaction and spatial-driven information interaction in MIPAM

$$f_{\text{sdii}}\left(\mathbf{x}_i^{cs}\right) = \text{Concat}\left(\text{Sum}\left(GConv_{++}\left(\text{Split}\left(\mathbf{x}_i^{cs\Delta}\right)\right)\right)^{\Delta}\right) \tag{17}$$

where *Sum* represents the row-wise summation. Reshaping and splitting operations are used to adjust the feature shape for subsequent specific information interactions. Here, the input channels, output channels, kernel size, stride, padding and grouping in 1D group convolution are set as $H'W'$, $H'W'$, 1, 1, 0 and $H'W'$ respectively. By allocating different parameters in the spatial dimension and sharing same parameters in the channel dimension, the spatial-driven information interaction not only realizes the interactive fusion of multiple information from the channel-level information collection and spatial-level information collection, but also further perceives the spatial diversity information.

Figure 3 shows the channel-driven information interaction and spatial-driven information interaction in MIPAM. Our attention module simultaneously realizes the perception of multi-dimensional diversity information in information interaction. We stimulate the intrinsic information potentials through more comprehensive active interaction.

3.4 Attention application in YOLO

In order to sort out the proposed attention module more comprehensively, here we further show the implementation details of MIPAM in Fig. 4, including parameter configuration, feature change, and specific process. Our MIPAM perceives multi-dimensional dependency information, multi-dimensional structure information and multi-dimensional global information in channel-level information collection and spatial-level information collection, and perceives multi-dimensional diversity information in channel-driven information interaction and spatial-driven information interaction. In this paper, the proposed MIPAM is integrated into YOLO algorithms to achieve a better trade-off between detection speed and detection accuracy in complex underwater environments.

The main reason we focus on the YOLO series [40–44, 47] is that YOLO detectors are one-stage detectors. Compared with two-stage detectors, they have great advantages in inference speed. This is crucial for the real-time requirement of underwater detection tasks. It is worth noting that YOLO detectors [42–44, 47] have a similar design architecture, which mainly consists of three modular processes. The backbone is responsible for extracting image features, which can obtain high-level semantic information. The neck is responsible for fusing features at different scales, which can further enhance semantic information. The head is responsible for classifying and regressing the enhanced features at different scales, which can obtain the object category and bounding box position. We add plug-and-play attention modules to ten important positions in YOLO detector, as shown in Fig. 5. The six attention modules located at the front and back of YOLO neck are responsible for recalibrating the features at three different scales, which can further enhance the perception of underwater objects with different sizes. The four attention modules located at the inside of YOLO neck are responsible for recalibrating the features between two adjacent scales, which can achieve more effective multi-scale feature fusion and further reduce the underwater background interference.

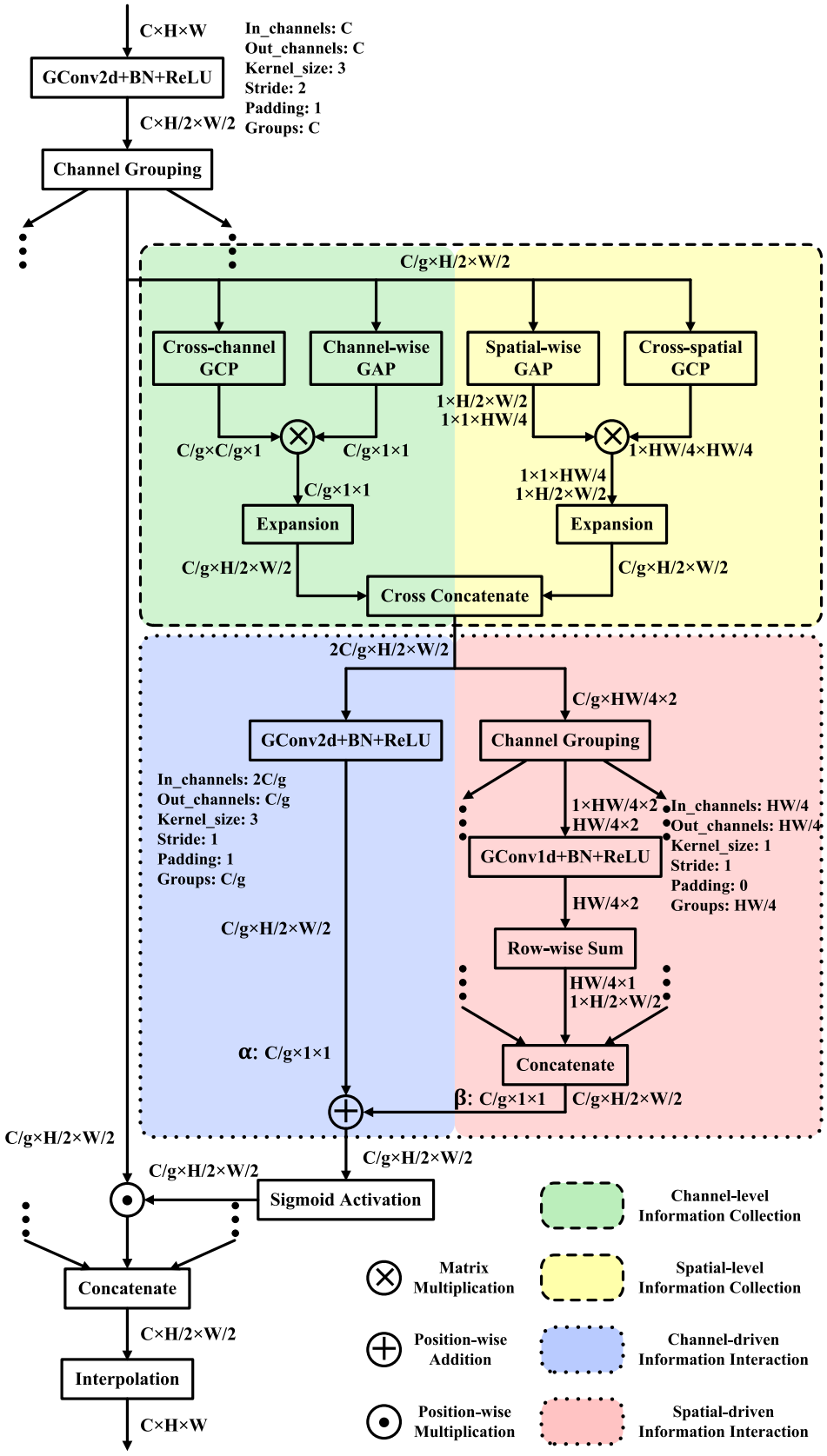


Fig. 4 The implementation details about multiple information perception-based attention module (MIPAM), including parameter configuration, feature change, and specific process

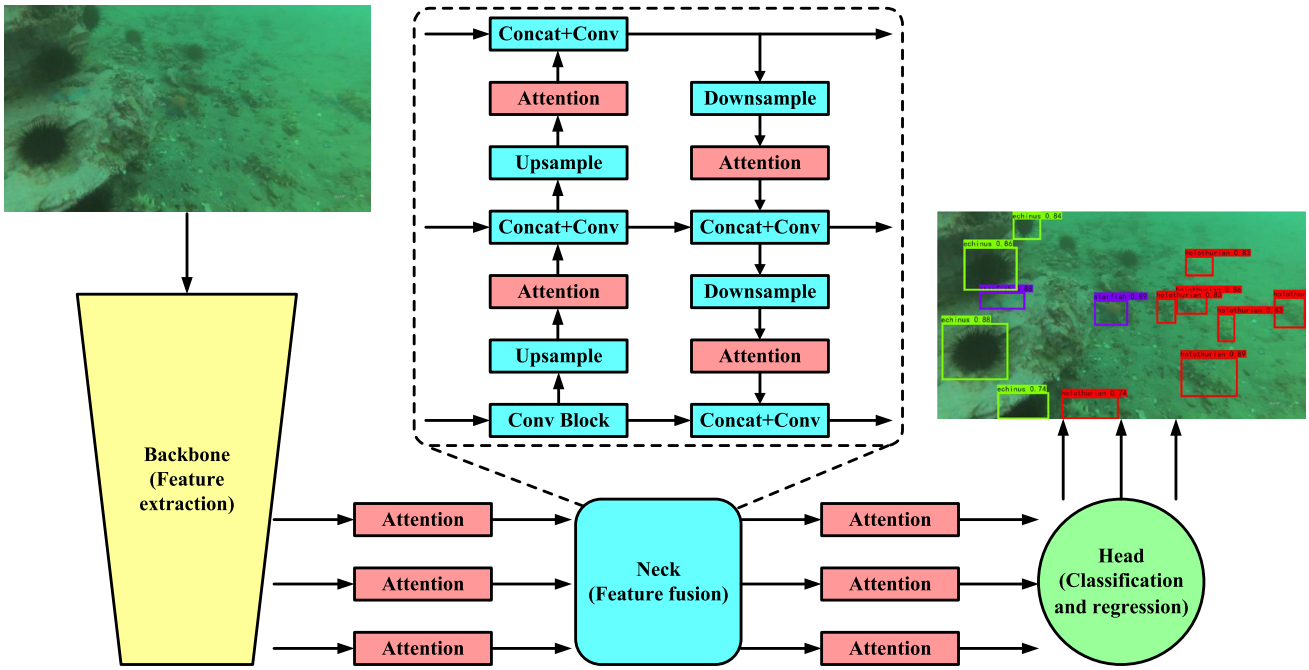


Fig. 5 Combining attention with YOLO for underwater object detection

4 Experiments and results

In order to verify the effectiveness of our work, we conduct extensive detection experiments on the underwater image dataset [48] and the PASCAL VOC dataset [49, 50], and analyze the experimental results in detail. In this section, we first provide training details about the network model. We then conduct ablation experiments on the proposed attention from three design perspectives, and decide the most suitable attention design for the underwater detection task. We further perform comparative experiments on state-of-the-art attention modules and provide attention visualization results on the underwater image dataset. Finally, some experiments are implemented on the PASCAL VOC dataset to demonstrate the generalization ability of our attention module on other detection tasks.

In this paper, the mean average precision (mAP) under specified intersection over union (IoU) is used to measure detection accuracy. mAP0.5 refers to mAP at IoU=0.5, which is the general metric. mAP0.75 refers to mAP at IoU=0.75, which is the strict metric. mAP0.5:0.95 refers to mAP at IoU=0.5:0.05:0.95, which is the primary challenge metric. The parameters (Params) and floating point operations (FLOPs) are used to measure network size and model computational complexity.

4.1 Training details

The underwater image dataset (URPC 2017–2020) consists of URPC 2017(17655), URPC 2018(2901), URPC 2019(4757) and URPC 2020(6575), which has a total of 25747 images and 4 categories after removing duplicate images. The underwater image dataset (URPC 2021) has a total of 8200 images and 4 categories. The PASCAL VOC dataset consists of VOC 2007 test(4952), VOC 2007 trainval(5011), and VOC 2012 trainval(11540), which has a total of 21503 images and 20 categories. In this paper, we first divide the dataset into test set and trainval set in a 5:5 ratio. The trainval set is further divided into training set and validation set in a 5:5 ratio. For URPC 2017-2020, the test set, training set and validation set have 12875, 6436 and 6436 images respectively. For URPC 2021, the test set, training set and validation set have 4100, 2050 and 2050 images respectively. For PASCAL VOC dataset, the test set, training set and validation set have 10753, 5375 and 5375 images respectively.

During training, the input image is set to 640 × 640 size and further processed using mosaic data enhancement. We use the stochastic gradient descent (SGD) optimizer with weight decay of 5e-4 and momentum of 0.937. The network model is trained for a total of 500 epochs based on pretrained weights, where mosaic data enhancement is turned off at the last 30 percent. We first perform frozen training with a batch size of 32 for 50 epochs. We then perform unfrozen training with a batch size of 16 for 450 epochs. The cosine annealing algorithm is used to control the learning rate decay, where the initial learning rate is set to 0.01 and the minimum learning rate is set to 0.0001. All experiments are run on a personal

Fig. 6 Compare our underwater image dataset with the PASCAL VOC dataset. The first line represents PASCAL VOC images in some traditional environments. The last three lines represent our underwater images in real marine environments

computer with NVIDIA GeForce RTX 3090/PCIe/SSE2 and Intel(R) Core(TM) i9-10980XE CPU @ 3.00GHz×36.

4.2 Experiments on underwater image dataset

The harsh underwater environments bring great difficulties to the collection and annotation of underwater datasets. At present, the underwater robot picking contest (URPC) [48] is a public underwater detection dataset, where underwater images are captured by underwater robots and divers in the near-shallow sea. URPC mainly includes four detection categories: Holothurian, Echinus, Scallop, and Starfish. Many underwater work studies are based on this underwater dataset. Underwater images in real marine environments and VOC images in other traditional environments are shown in Fig. 6. Compared with the images in other environments, the images in underwater environments obviously show low contrast, color cast, texture distortion and so on. It is worth noting that underwater objects have strong concealment capabilities and evolve natural protective colors. The above phenomena make the underwater detection task face severe problems about strong underwater background interference and weak underwater object perception. In this paper, our work is dedicated to reducing underwater background interference and improving underwater object perception for efficient underwater object detection.

4.3 Ablation experiments

In order to explore the optimal attention design for underwater object detection, we focus on designing ablation experiments from three different perspectives, including information collection and information interaction, grouping and fusion, and attention location. The detectors used in ablation experiments are uniformly specified as the medium (M) model.

Information collection and information interaction It can be seen from Sect. 3.2 that the information collection of our MIPAM is mainly composed of channel-level information collection and spatial-level information collection. The cross-channel GCP and channel-wise GAP are two important components in channel-level information collection. The spatial-wise GAP and cross-spatial GCP are two important components in spatial-level information collection. It can be seen from Sect. 3.3 that the information interaction of our MIPAM is mainly composed of channel-driven information interaction and spatial-driven information interaction.

Tables 2 and 3 report the ablation experiments on information collection and information interaction, where various attention modules are integrated onto state-of-the-art YOLO detectors for underwater object detection. MIPAM(c) considers spatial global information, channel structure information and channel diversity information by

Table 2 Ablation experiments on information collection and information interaction (YOLOV5)

Attention modules	Channel-level information collection		Spatial-level information collection		Channel-driven information interaction	Spatial-driven information interaction	Params	FLOPs	Memory	mAP0.5 (%)
	Cross-channel GCP	Channel-wise GAP	Spatial-wise GAP	Cross-spatial GCP						
MIPAM(c)		✓			✓		21.11 M	25.26 G	574.07 M	87.9
MIPAM(s)			✓			✓	21.13 M	25.24 G	573.36 M	88.1
MIPAM(cs)		✓	✓		✓	✓	21.14 M	25.28 G	575.43 M	88.4
MIPAM(C)	✓	✓			✓		21.11 M	25.26 G	574.10 M	88.6
MIPAM(S)			✓	✓		✓	21.13 M	25.24 G	605.21 M	88.9
MIPAM(CS)	✓	✓	✓	✓	✓	✓	21.14 M	25.28 G	607.31 M	**89.1**

The results with optimal detection accuracy are marked in bold

Table 3 Ablation experiments on information collection and information interaction (YOLOX)

Attention modules	Channel-level information collection		Spatial-level information collection		Channel-driven information interaction	Spatial-driven information interaction	Params	FLOPs	Memory	mAP0.5 (%)
	Cross-channel GCP	Channel-wise GAP	Spatial-wise GAP	Cross-spatial GCP						
MIPAM(c)		✓			✓		25.33 M	36.79 G	665.78 M	89.0
MIPAM(s)			✓			✓	25.34 M	36.77 G	665.07 M	89.2
MIPAM(cs)		✓	✓		✓	✓	25.36 M	36.81 G	667.14 M	89.4
MIPAM(C)	✓	✓			✓		25.33 M	36.79 G	665.81 M	89.5
MIPAM(S)			✓	✓		✓	25.34 M	36.77 G	696.92 M	89.7
MIPAM(CS)	✓	✓	✓	✓	✓	✓	25.36 M	36.81 G	699.02 M	**89.9**

The results with optimal detection accuracy are marked in bold

selecting channel-wise GAP and channel-driven information interaction. MIPAM(s) considers channel global information, spatial structure information and spatial diversity information by selecting spatial-wise GAP and spatial-driven information interaction. MIPAM(cs) considers multi-dimensional global information, multi-dimensional structure information and multi-dimensional diversity information by combining MIPAM(c) and MIPAM(s). MIPAM(C) considers channel dependency information, channel structure information and spatial global information and channel diversity information by selecting cross-channel GCP, channel-wise GAP and channel-driven information interaction. MIPAM(S) considers spatial dependency information, spatial structure information, channel global information and spatial diversity information by selecting spatial-wise GAP, cross-spatial GCP and spatial-driven information interaction. MIPAM(CS) considers multi-dimensional dependency information, multi-dimensional structure information, multi-dimensional global information and multi-dimensional diversity information by combining MIPAM(C) and MIPAM(S).

As can be seen from both Tables 2 and 3, this design strategy of MIPAM(CS) achieves the best detection results. This indicates that multiple information perception-based attention is more suitable for underwater object detection. Through further analysis, we draw three conclusions about MIPAM in underwater detection tasks. First, the spatial branch is stronger than the channel branch in terms of detection accuracy, and the dimensional branch can achieve better performance improvements by perceiving richer information. Second, joint design strategies outperform single design strategies in harsh underwater environments. Third, MIPAM(CS) not only brings significant performance gains, but also the parameters, computations and memory are controlled within a reasonable range.

Grouping and fusion It can be seen from Sect. 3.1 that grouping and fusion operations are designed in information preprocessing and information interaction, respectively. The grouping operation is responsible for splitting the input feature into input subfeatures without overlapping along the channel dimension. This multi-branch structure not only controls the parameters and computations by reducing channels, but also generates multiple targeted attentions by dividing information. The fusion operation is responsible for integrating features derived from channel-driven information interaction and spatial-driven information interaction by assigning learnable parameters. This adaptive fusion strategy effectively integrates different features and selectively delivers more valuable information to subsequent processes.

The ablation experiments on grouping and fusion are reported in Tables 4 and 5, where attention modules under different configurations are integrated on YOLOV5 detector and YOLOX detector. For information preprocessing, we here set the number of groups to 2, 4, 8, 16 and 32, which can generate 2, 4, 8, 16 and 32 different subfeatures, respectively. This multi-branch structure of our attention module can correspondingly activate 2, 4, 8, 16 and 32 diverse attentions. For information interaction, we further configure the learnable parameters on each branch. When choosing not to assign learnable parameters, we directly fuse the features by location-wise addition. When choosing to assign learnable parameters, we first perform importance calibration on the features, and then perform information fusion.

As can be seen from both Tables 4 and 5, attention performance can be effectively improved by setting a moderate number of groups and assigning learnable parameters. The attention module with 16 groups and learnable parameters is more beneficial to the underwater detection task. Compared with other design methods, this design method not only achieves optimal detection accuracy, but also reduces the amount of parameters and memory consumption.

Attention location It can be seen from Sect. 3.4 that our attention is embedded in ten locations of YOLO detector to enhance the underwater detection performance. Six attention modules located at the front and back of YOLO neck are responsible for recalibrating the features at three different scales, which improve the perception of underwater objects with different sizes. Four attention modules located at the inside of YOLO neck are responsible for recalibrating the features between two adjacent scales, which achieve efficient multi-scale fusion and reduce underwater background interference.

Tables 6 and 7 report the ablation experiments on attention location. We first add attention modules to the front of neck, the middle of neck, and the back of neck to test the effect of this individual embedding strategy on detection performance, where the number of attentions is 3, 4, and 3, respectively. We then add attention modules to the front-middle of neck, the front-back of neck, and the middle-back of neck to test the effect of this combined embedding strategy on detection performance, where the number of attentions is 7, 6, and 7, respectively. We finally add attention modules to the front-middle-back of neck to test the effect of this full embedding strategy on detection performance, where the number of attentions is 10.

As can be seen from both Tables 6 and 7, embedding attention modules on the front-middle-back of neck significantly improves the underwater detection performance. This shows that adding our attention module to ten important locations of YOLO detectors can effectively reduce underwater background interference and significantly enhance underwater object perception. After further analysis, we find that the number of attention modules at key locations is proportional to the improvement of detection performance. When embed-

Table 4 Ablation experiments on grouping and fusion (YOLOV5)

Number of groups	$\alpha\&\beta$	Params	FLOPs	Memory	mAP0.5 (%)
2		21.17 M	25.28 G	625.65 M	88.5
	✓	21.18 M	25.28 G	625.65 M	88.5
4		21.15 M	25.28 G	614.83 M	88.6
	✓	21.16 M	25.28 G	614.83 M	88.6
8		21.14 M	25.28 G	609.76 M	88.7
	✓	21.15 M	25.28 G	609.76 M	88.8
16		21.14 M	25.28 G	607.31 M	89.0
	✓	21.14 M	25.28 G	607.31 M	**89.1**
32		21.13 M	25.28 G	606.11 M	88.7
	✓	21.14 M	25.28 G	606.11 M	88.9

The results with optimal detection accuracy are marked in bold

Table 5 Ablation experiments on grouping and fusion (YOLOX)

Number of groups	$\alpha\&\beta$	Params	FLOPs	Memory	mAP0.5 (%)
2		25.38 M	36.81 G	717.36 M	89.5
	✓	25.39 M	36.81 G	717.36 M	89.5
4		25.36 M	36.81 G	706.54 M	89.5
	✓	25.37 M	36.81 G	706.54 M	89.6
8		25.35 M	36.81 G	701.47 M	89.7
	✓	25.36 M	36.81 G	701.47 M	89.7
16		25.35 M	36.81 G	699.02 M	89.8
	✓	25.36 M	36.81 G	699.02 M	**89.9**
32		25.35 M	36.81 G	697.82 M	89.6
	✓	25.35 M	36.81 G	697.82 M	89.6

The results with optimal detection accuracy are marked in bold

Table 6 Ablation experiments on attention location (YOLOV5)

Front of neck	Middle of neck	Back of neck	Number of attentions	Params	FLOPs	Memory	mAP0.5 (%)
✓			3	21.09 M	25.24 G	571.39 M	87.6
	✓		4	21.09 M	25.24 G	572.50 M	88.1
		✓	3	21.09 M	25.24 G	571.39 M	87.8
✓	✓		7	21.12 M	25.26 G	589.91 M	88.5
✓		✓	6	21.12 M	25.26 G	588.80 M	88.2
	✓	✓	7	21.12 M	25.26 G	589.91 M	88.8
✓	✓	✓	10	21.14 M	25.28 G	607.31 M	**89.1**

The results with optimal detection accuracy are marked in bold

Table 7 Ablation Experiments on Attention Location (YOLOX)

Front of neck	Middle of neck	Back of neck	Number of attentions	Params	FLOPs	Memory	mAP0.5 (%)
✓			3	25.31 M	36.77 G	663.10 M	88.3
	✓		4	25.31 M	36.77 G	664.21 M	88.9
		✓	3	25.31 M	36.77 G	663.10 M	88.7
✓	✓		7	25.33 M	36.79 G	681.61 M	89.2
✓		✓	6	25.33 M	36.79 G	680.51 M	89.0
	✓	✓	7	25.33 M	36.79 G	681.61 M	89.7
✓	✓	✓	10	25.36 M	36.81 G	699.02 M	**89.9**

The results with optimal detection accuracy are marked in bold

Table 8 Underwater detection results of different attention modules on YOLOV5.(URPC 2017–2020)

Settings	Params	FLOPs	mAP0.5 (%)	mAP0.75 (%)	mAP0.5:0.95
YOLOV5	21.07 M	25.22 G	87.5	62.7	56.7%
+SEM [6]	21.31 M	25.22 G	88.8	63.6	57.1% (+0.4)
+SRM [9]	21.08 M	25.22 G	88.7	64.0	57.4% (+0.7)
+SGEM [10]	21.07 M	25.22 G	88.8	64.2	57.4% (+0.7)
+ECAM [11]	21.07 M	25.22 G	88.8	64.0	57.3% (+0.6)
+GCTM [14]	21.08 M	25.22 G	88.2	64.1	57.3% (+0.6)
+CoAM [52]	21.26 M	25.23 G	88.4	63.1	56.7% (+0.0)
+ShAM [12]	21.07 M	25.22 G	88.6	63.7	57.2% (+0.5)
+PSAM [13]	21.84 M	26.53 G	88.0	63.3	56.9% (+0.2)
+FCAM [15]	21.31 M	25.23 G	88.2	63.8	57.1% (+0.4)
+MIPAM (Ours)	21.14 M	25.28 G	**89.1**	**64.6**	**57.7% (+1.0)**

The results with optimal detection accuracy are marked in bold

Table 9 Underwater detection results of hybrid attention modules and their variants on YOLOV5.(URPC 2017–2020)

Settings	Params	FLOPs	mAP0.5	mAP0.75	mAP0.5:0.95
YOLOV5	21.07 M	25.22 G	87.5%	62.7%	56.7%
+BAM [7]	21.69 M	25.63 G	88.9%	64.0%	57.5% (+0.8)
+BAM(C)	21.31 M	25.22 G	88.8%	64.0%	57.5% (+0.8)
+BAM(S)	21.45 M	25.63 G	88.7%	64.0%	57.4% (+0.7)
+CBAM [8]	21.55 M	25.22 G	88.5%	63.4%	56.9% (+0.2)
+CBAM(C)	21.55 M	25.22 G	88.7%	63.8%	57.2% (+0.5)
+CBAM(S)	21.07 M	25.22 G	88.5%	63.7%	57.1% (+0.4)
+GSoPM [18]	25.67 M	33.75 G	88.8%	64.4%	57.5% (+0.8)
+GSoPM(C)	21.09 M	25.24 G	88.7%	63.6%	57.2% (+0.5)
+GSoPM(S)	21.40 M	25.24 G	88.6%	64.1%	57.3% (+0.6)
+MIPAM (Ours)	21.14 M	25.28 G	**89.1%**	**64.6%**	**57.7% (+1.0)**
+MIPAM(C)	21.11 M	25.26 G	88.6%	64.3%	57.4% (+0.7)
+MIPAM(S)	21.13 M	25.24 G	88.9%	64.3%	57.5% (+0.8)

The results with optimal detection accuracy are marked in bold

Table 10 Underwater detection results of different attention modules on YOLOX.(URPC 2017–2020)

Settings	Params	FLOPs	mAP0.5 (%)	mAP0.75 (%)	mAP0.5:0.95
YOLOX	25.28 M	36.75 G	88.2	66.3	58.6%
+SEM [6]	25.52 M	36.76 G	89.3	66.6	58.9% (+0.3)
+SRM [9]	25.30 M	36.75 G	89.3	67.0	59.1% (+0.5)
+SGEM [10]	25.28 M	36.75 G	89.1	66.5	58.9% (+0.3)
+ECAM [11]	25.28 M	36.75 G	89.5	66.5	58.9% (+0.3)
+GCTM [14]	25.29 M	36.75 G	89.5	66.7	59.1% (+0.5)
+CoAM [52]	25.47 M	36.76 G	89.4	66.8	59.0% (+0.4)
+ShAM [12]	25.28 M	36.75 G	89.1	66.6	58.8% (+0.2)
+PSAM [13]	26.06 M	38.06 G	89.7	66.8	59.0% (+0.4)
+FCAM [15]	25.52 M	36.76 G	89.1	66.5	58.7% (+0.1)
+MIPAM (Ours)	25.36 M	36.81 G	**89.9**	**67.0**	**59.3% (+0.7)**

The results with optimal detection accuracy are marked in bold

Table 11 Underwater detection results of hybrid attention modules and their variants on YOLOX.(URPC 2017–2020)

Settings	Params	FLOPs	mAP0.5	mAP0.75	mAP0.5:0.95
YOLOX	25.28 M	36.75 G	88.2%	66.3%	58.6%
+BAM [7]	25.91 M	37.16 G	89.3%	66.3%	58.9% (+0.3)
+BAM(C)	25.53 M	36.76 G	89.2%	66.4%	58.7% (+0.1)
+BAM(S)	25.66 M	37.16 G	89.4%	66.6%	58.8% (+0.2)
+CBAM [8]	25.76 M	36.76 G	89.5%	66.7%	58.9% (+0.3)
+CBAM(C)	25.76 M	36.76 G	89.6%	66.6%	58.8% (+0.2)
+CBAM(S)	25.28 M	36.76 G	89.2%	66.8%	58.9% (+0.3)
+GSoPM [18]	29.88 M	45.28 G	89.0%	66.6%	58.8% (+0.2)
+GSoPM(C)	25.31 M	36.77 G	88.6%	66.4%	58.6% (+0.0)
+GSoPM(S)	25.62 M	36.78 G	88.8%	66.5%	58.7% (+0.1)
+MIPAM(Ours)	25.36 M	36.81 G	**89.9%**	**67.0%**	**59.3% (+0.7)**
+MIPAM(C)	25.33 M	36.79 G	89.5%	66.6%	59.0% (+0.4)
+MIPAM(S)	25.34 M	36.77 G	89.7%	66.9%	59.1% (+0.5)

The results with optimal detection accuracy are marked in bold

ding the same amount of attention, recalibrating high-level semantic information located in deeper layers can lead to more effective performance gains.

4.4 Comparative experiments

Here, we still focus on the YOLOV5 detector [44] and the YOLOX detector [47], and uniformly set the network size to the M model. These two detectors are state-of-the-art YOLO detectors, which show superior performance in both speed and accuracy. In order to further explore the optimal attention mechanism for underwater object detection, we select popular attention modules in computer vision and compare them with the proposed attention module MIPAM. These plug-and-play attention modules are combined into detectors in the same way, where the attention application in YOLO is provided in Sect. 3.4. There are three points worth noting for the specific configuration of our MIPAM. First, channel-level information collection, spatial-level information collection, channel-driven information interaction and spatial-driven information interaction are simultaneously configured in information collection and information interaction. Second, the number of groups in information preprocessing is set to 16, and the learnable parameters are assigned in fusion stage of information interaction. Third, attention modules are added to detectors using the full embedding strategy.

Tables 8 and 9 report the test results of various attention modules on YOLOV5. The different attention modules are compared with the proposed MIPAM in Table 8. The hybrid attention modules and their variants in channel and spatial dimensions are compared with our MIPAM, MIPAM(C) and MIPAM(S) in Table 9. Similarly, the comparison results of various attention modules on YOLOX are reported in Tables 10 and 11. Compared to other attention modules, our attention module obviously exhibits more excellent potential for underwater detection tasks. MIPAM brings significant performance gains on general, strict, and primary challenge metrics while maintaining network size and model complexity. This benefits from MIPAM's full perception of multi-dimensional global information, multi-dimensional dependency information, multi-dimensional structure information and multi-dimensional diversity information in information collection and information interaction.

In order to more intuitively demonstrate the detection advantages brought by the proposed attention module in complex underwater environments, we focus on selecting the top 3 attention modules that perform best in the underwater dataset to achieve attention visualization. Figures 7 and 8 show the attention visualization results of BAM, GSoPM, SRM, GCTM and MIPAM on YOLO detectors in different marine environments, where we use Grad-CAM [51] and choose YOLO head as the visualization layer. It is worth noting here that BAM, GSoPM and MIPAM are the top three attention modules that perform best on YOLOV5, and SRM, GCTM and MIPAM are the top three attention modules that perform best on YOLOX. BAM perceives multi-dimensional global information and multi-dimensional structure information using channel-wise global average pooling and spatial-wise 1×1 convolution. GSoPM perceives multi-dimensional dependency information, channel global information and spatial structure information using cross-channel global covariance pooling, cross-spatial global covariance pooling and spatial-wise 1×1 convolution. SRM perceives spatial global information and channel structure information using channel-wise standard deviation pooling and channel-wise global average pooling. GCTM perceives spatial global information and channel structure information using channel-wise L2-norm. These four atten-

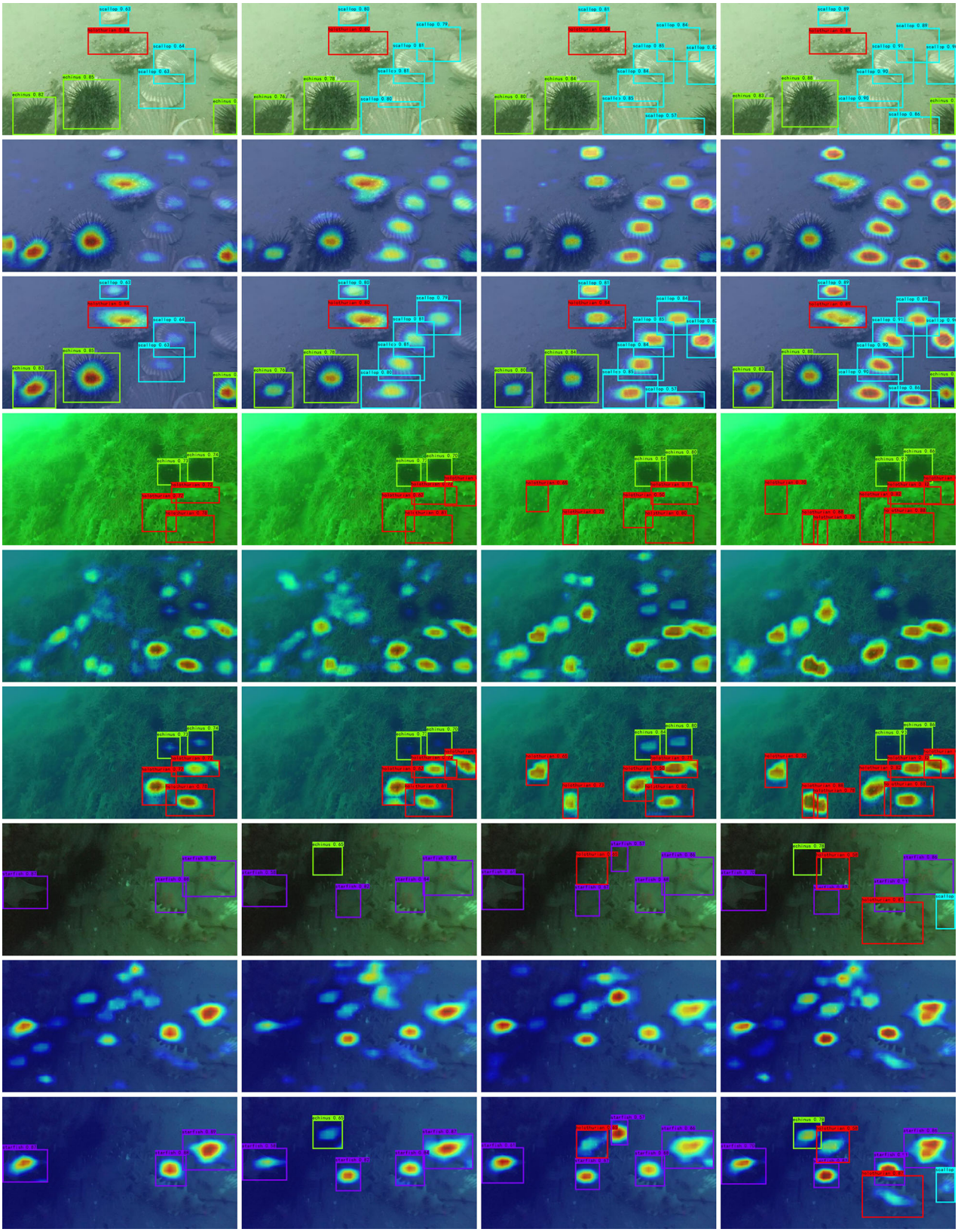


Fig. 7 Attention visualization results in different marine environments. The attention modules are integrated into the YOLOV5 detector, including no-attention module, BAM, GSoPM, and MIPAM from left to right. The experimental results in various marine environments are represented from top to bottom, including detection results, attention visualization results, and combined results

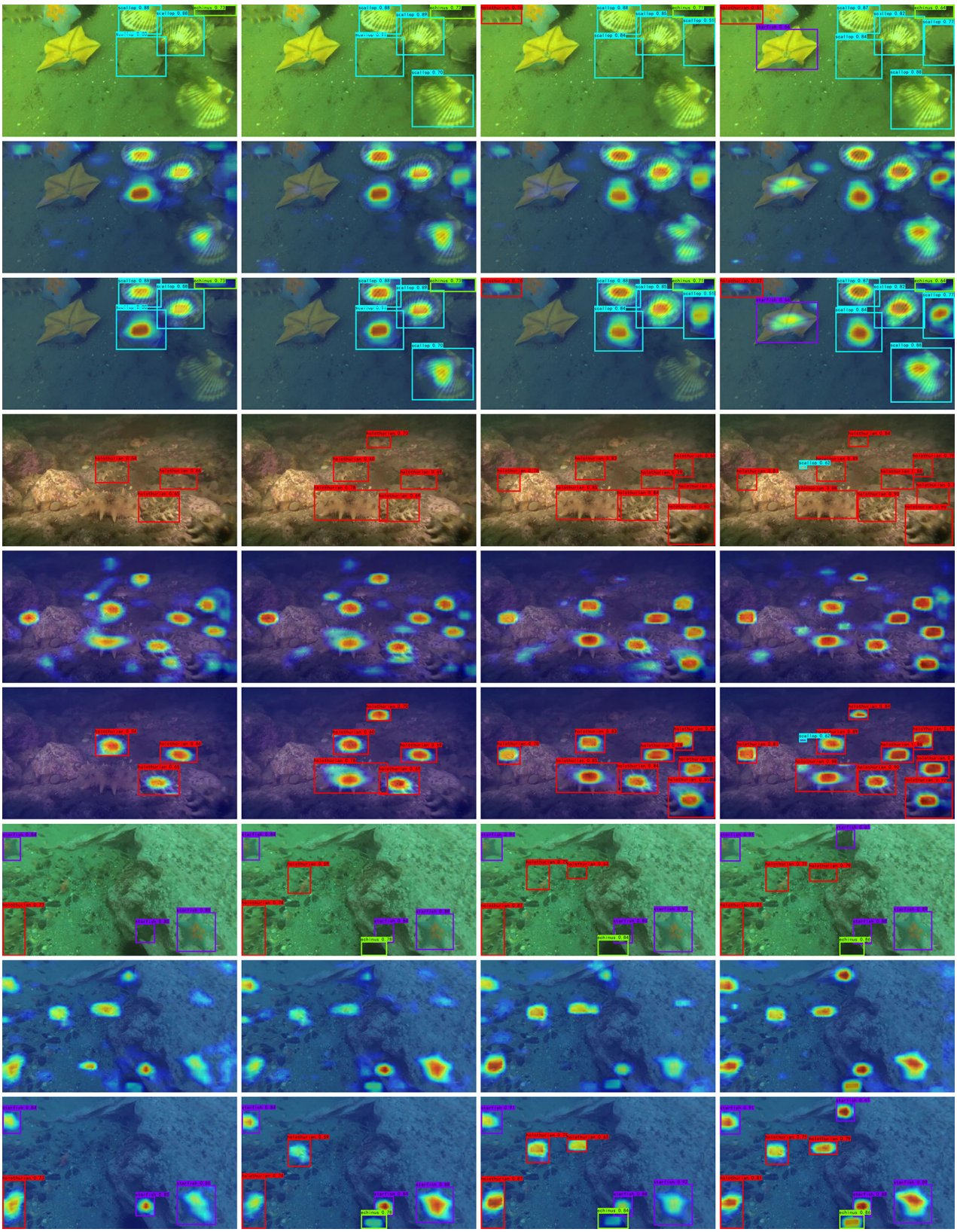


Fig. 8 Attention visualization results in different marine environments. The attention modules are integrated into the YOLOX detector, including no-attention module, SRM, GCTM, and MIPAM from left to right. The experimental results in various marine environments are represented from top to bottom, including detection results, attention visualization results, and combined results

Table 12 Underwater performance test of our work on different YOLO detectors. (URPC 2021)

Detector	Params	MAdd	Memory	FPS	mAP0.5	mAP0.5:0.95
YOLOV3 [42]	61.54 M	154.86 G	1042.66 M	29	79.4%	48.3%
+MIPAM	61.60 M	154.89 G	1091.86 M	24	81.3%	49.5%(+1.2)
YOLOV4 [43]	63.95 M	141.38 G	1434.07 M	25	82.3%	56.4%
+MIPAM	64.03 M	141.42 G	1485.91 M	21	84.0%	57.3%(+0.9)
YOLOV5 [44]	46.65 M	114.27 G	907.12 M	30	78.5%	57.6%
+MIPAM	46.74 M	114.32 G	967.54 M	23	80.0%	58.4%(+0.8)
YOLOV6 [45]	59.60 M	150.70 G	1046.87 M	26	82.0%	61.2%
+MIPAM	59.68 M	150.74 G	1100.69 M	20	83.2%	61.7%(+0.5)
YOLOV7 [46]	37.21 M	104.79 G	688.76 M	32	81.4%	60.8%
+MIPAM	37.27 M	104.82 G	735.97 M	25	82.4%	61.3%(+0.5)
YOLOX [47]	54.15 M	155.29 G	1029.59 M	28	78.9%	58.1%
+MIPAM	54.24 M	155.34 G	1090.01 M	23	80.3%	58.7%(+0.6)

tion modules use active channel interaction to perceive channel diversity information. Although the above attention modules show some potential in underwater detection tasks through rich information perception, they are still slightly insufficient in detection performance compared to our attention modules. MIPAM uses channel-level information collection and spatial-level information collection to perceive multi-dimensional dependency information, multi-dimensional structure information and multi-dimensional global information, which enhance the feature expression abilities. MIPAM further uses channel-driven information interaction and spatial-driven information interaction to further perceive multi-dimensional diversity information, which stimulate the intrinsic information potentials. As can be seen from the attention visualization, our attention module achieves more efficient underwater object detection compared with other attention modules. Our MIPAM effectively reduces underwater background interference and significantly improves underwater object perception through richer information perception and more comprehensive active interaction.

In order to further verify the effectiveness of our attention module on different baseline methods, we provide the underwater performance test in YOLOV3, YOLOV4, YOLOV5, YOLOV6, YOLOV7 and YOLOX, as shown in Table12. During the experiment, the input image size is uniformly set to 640×640 and the network model size is uniformly set to L. Our attention is integrated into the YOLO detectors according to the proposed method. From the experimental results, we can see that our work has good robustness and can achieve significant performance gains in various YOLO detectors.

In order to further demonstrate the effectiveness of our work in underwater detection tasks, we provide the results of comparison between the proposed underwater work and other underwater work, as shown in Table13. Xu et al. [34] proposed a scale-aware feature pyramid network(SAFPN) for marine object detection, which used a special backbone subnetwork to provide richer fine-grained features for small underwater targets, and used a multi-scale feature pyramids to enhance semantic features. Xu et al. [35] further proposed an attention-based spatial pyramid pooling network(ASPPN) for marine object detection, which expanded receptive fields to enrich the interesting information, and fused bidirectional features to improve the feature robustness. As can be seen from the experimental results, our work showed the excellent performance in terms of detection accuracy and detection speed, which can better meet the requirements of high-precision and real-time for underwater object detection. Compared with other works, our high-intensity collaborative attention calibration strategy specifically for underwater detection tasks has higher flexibility and extensibility in practical applications.

Table 13 The comparison of our work with other underwater detection works. (URPC 2021)

Methods	FPS	mAP0.5
SAFPN [34]	10	77.5%
ASPPN [35]	16	78.2%
YOLOV5+MIPAM	23	80.0%
YOLOX+MIPAM	23	80.3%

4.5 Experiments on PASCAL VOC dataset

In this subsection, we further conduct experiments on PASCAL VOC dataset. Table 14 reports the test results of CoAM, ShAM, PSAM, FCAM and MIPAM on YOLOV5 and YOLOX, where the network model is set to M size. For VOC detection tasks, these original YOLO detectors achieve 80.8% mAP and 82.2% mAP, respectively. We add MIPAM to YOLOV5 detector and YOLOX detector, which

Table 14 VOC detection results of different attention modules on YOLO detectors

Settings	Params	FLOPs	mAP0.5
YOLOV5	21.13 M	25.33 G	80.8%
+CoAM [52]	21.33 M	25.33 G	81.4% (+0.6)
+ShAM [12]	21.13 M	25.33 G	81.4% (+0.6)
+PSAM [13]	21.91 M	26.63 G	81.2% (+0.4)
+FCAM [15]	21.37 M	25.33 G	81.1% (+0.3)
+MIPAM(Ours)	21.21 M	25.38 G	**81.5% (+0.7)**
YOLOX	25.29 M	36.78 G	82.2%
+CoAM [52]	25.48 M	36.79 G	82.6% (+0.4)
+ShAM [12]	25.29 M	36.78 G	82.8% (+0.6)
+PSAM [13]	26.07 M	37.66 G	82.9% (+0.7)
+FCAM [15]	25.53 M	36.79 G	82.7% (+0.5)
+MIPAM(Ours)	25.37 M	36.82 G	**82.9% (+0.7)**

The results with optimal detection accuracy are marked in bold

improves the detection accuracy by 0.7% and 0.7%, respectively. Compared to other attention modules, MIPAM brings the greatest performance gain. Our attention module performs the best in terms of accuracy and is also competitive in terms of parameters and computations. It is worth noting that the main reason for detection performance improvement is not the simple capacity increase. This is due to the reasonable correction of feature information by our attention module, which activates high-quality attention by perceiving multiple information. The experimental results in Table14 demonstrate the generalization ability of MIPAM on different detection tasks. After further analyses of experimental results, we find that MIPAM shows more significant performance gains in underwater detection environments compared to VOC detection environments. This means that our attention module can make a greater contribution to solving the problems of strong underwater background interference and weak underwater feature discriminability.

5 Conclusion

In this paper, we proposed a multiple information perception-based attention module(MIPAM) in YOLO for underwater object detection. In information preprocessing, we used spatial downsampling and channel splitting to control parameters and computations of attention module. In information collection, we designed channel-level and spatial-level information collections to enhance feature expression capabilities. For channel-level information collection, the cross-channel GCP perceived channel dependency information. The channel-wise GAP perceived channel structure information and spatial global information. For spatial-level information collection, the spatial-wise GAP perceived spatial structure information and channel global information. The cross-spatial GCP perceived spatial dependency information. In information interaction, we proposed channel-driven and spatial-driven information interactions to further stimulate intrinsic information potentials. For channel-driven information interaction, channel diversity information was perceived by allocating different parameters in channel dimension and sharing same parameters in spatial dimension. For spatial-driven information interaction, spatial diversity information was perceived by allocating different parameters in spatial dimension and sharing same parameters in channel dimension. In attention activation, we introduced the multi-branch structure to generate multiple attention, which facilitated targeted calibration of feature information on different branches. In information postprocessing, we applied channel concatenation and spatial upsampling to realize the plug-and-play of attention module.

We embedded MIPAM into ten important positions of YOLO detector, which met the high-precision and real-time requirements for underwater object detection. Our work provided more significant performance gains for underwater detection tasks, which reduced underwater background interference and improved underwater object perception. Our work also brought some performance improvements for other detection tasks, which showed a certain generalization ability.

In future work, we will continue to take reducing underwater background interference and improving underwater object perception as the primary goal, and further explore the application potential of attention mechanism in underwater object detection. The attention mechanism mainly consists of three processes: information collection, information interaction and attention activation. In this paper, we studied the problems of information collection in detail and proposed the reasonable solutions. For underwater detection tasks, information interaction and attention activation also have improved directions. For information interaction, dimensionality reduction interaction strategy will lead to the destruction of direct information correspondence, and local interaction strategy will lead to the lack of global information interaction. For attention activation, single-dimensional attention will weaken the robustness of attention application, single-functional attention will reduce the flexibility of attention calibration, and single-level attention will lack the diversity of attention perception. In follow-up work, we will start from these two aspects to further improve the calibration intensity of underwater attention to detail features, and further explore the optimal attention design suitable for underwater detection tasks.

Acknowledgements The authors gratefully acknowledge the financial supports from the National Natural Science Foundation of China under Grant 61370142, Grant 61802043, Grant 61272368, Grant 62176037 and Grant 62002041, in part by the Fundamental Research

Funds for the Central Universities under Grant 3132016352 and Grant 3132021238, in part by the Dalian Science and Technology Innovation Fund under Grant 2018J12GX037, Grant 2019J11CY001 and Grant 2021JJ12GX028, in part by Liaoning Revitalization Talents Program under Grant XLYC1908007, in part by the Liaoning Doctoral Research Start-up Fund Project Grant 2021-BS-075, and in part by the China Postdoctoral Science Foundation under Grant 3620080307.

Data availability statement The datasets generated during and/or analysed during the current study are available from the corresponding author on reasonable request.

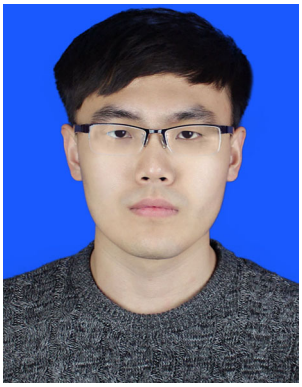
Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Jiang, M., Zhai, F.H., Kong, J.: Sparse attention module for optimizing semantic segmentation performance combined with a multi-task feature extraction network. Vis. Comput. **38**(7), 2473–2488 (2022)
2. Yang, Q.N., Shi, W.M., Chen, J., Tang, Y.: Localization of hard joints in human pose estimation based on residual down-sampling and attention mechanism. Vis. Comput. **38**(7), 2447–2459 (2022)
3. Cheng, Z.M., Qu, A.P., He, X.F.: Contour-aware semantic segmentation network with spatial attention mechanism for medical image. Vis. Comput. **38**(3), 749–762 (2022)
4. Li, Z.X., Lu, S.H., Dong, Y.S., Guo, J.Y.: Msffa: a multi-scale feature fusion and attention mechanism network for crowd counting. Vis. Comput. 1–12 (2022)
5. Li, X.L., Hua, Z., Li, J.J.: Attention-based adaptive feature selection for multi-stage image dehazing. Vis. Comput., 1–16 (2022)
6. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141 (2018)
7. Park, J., Woo, S., Lee, J.Y., Kweon, I.S.: Bam: bottleneck attention module. arXiv preprint arXiv:1807.06514 (2018)
8. Woo, S., Park, J., Lee, J.Y., Kweon, I.S.: Cbam: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
9. Lee, H., Kim, H.E., Nam, H.: Srm: A style-based recalibration module for convolutional neural networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 1854–1862 (2019)
10. Li, X., Hu, X.L., Yang, J.: Spatial group-wise enhance: Improving semantic feature learning in convolutional networks. arXiv preprint arXiv:1905.09646 (2019)
11. Wang, Q., Wu, B., Zhu, P., Li, P., Hu, Q.: Eca-net: Efficient channel attention for deep convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
12. Zhang, Q.L., Yang, Y.B.: Sa-net: Shuffle attention for deep convolutional neural networks. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2235–2239 (2021). IEEE
13. Zhang, H., Zu, K., Lu, J., Zou, Y., Meng, D.: Epsanet: an efficient pyramid split attention block on convolutional neural network. arXiv preprint arXiv:2105.14447 (2021)
14. Yang, Z.X., Zhu, L.C., Wu, Y., Yang, Y.: Gated channel transformation for visual recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 11794–11803 (2020)
15. Qin, Z., Zhang, P., Wu, F., Li, X.: Fcanet: Frequency channel attention networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 783–792 (2021)
16. Misra, D., Nalamada, T., Arasanipalai, A.U., Hou, Q.: Rotate to attend: convolutional triplet attention module. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 3139–3148 (2021)
17. Chen, Y.P., Kalantidis, Y., Li, J.S., Yan, S.C., Feng, J.S.: A ˆ 2-nets: double attention networks. Adv. Neural Inf. Process. Syst. **31** (2018)
18. Gao, Z.L., Xie, J.T., Wang, Q.L., Li, P.H.: Global second-order pooling convolutional networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3024–3033 (2019)
19. Zhang, Z.Z., Lan, C.L., Zeng, W.J., Jin, X., Chen, Z.B.: Relation-aware global attention for person re-identification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3186–3195 (2020)
20. Haining, H., Yu, L.: Underwater acoustic detection: current status and future trends. Bull. Chin. Acad. Sci. (Chin. Vers.) **34**(3), 264–271 (2019)
21. Cho, H., Gu, J., Joe, H., Asada, A., Yu, S.-C.: Acoustic beam profile-based rapid underwater object detection for an imaging sonar. J. Mar. Sci. Technol. **20**, 180–197 (2015)
22. Zhang, L.Y., Li, C.Y., Sun, H.F.: Object detection/tracking toward underwater photographs by remotely operated vehicles (ROVs). Futur. Gener. Comput. Syst. **126**, 163–168 (2022)
23. Moniruzzaman, M., Islam, S.M.S., Lavery, P., Bennamoun, M.: Faster r-cnn based deep learning for seagrass detection from underwater digital images. In: 2019 Digital Image Computing: Techniques and Applications (DICTA), pp. 1–7 (2019). IEEE
24. Tharwat, A., Hemedan, A.A., Hassanien, A.E., Gabel, T.: A biometric-based model for fish species classification. Fish. Res. **204**, 324–336 (2018)
25. Chuang, M.-C., Hwang, J.-N., Williams, K.: A feature learning and object recognition framework for underwater fish images. IEEE Trans. Image Process. **25**(4), 1862–1872 (2016)
26. Knausgård, K.M., Wiklund, A., Sørdalen, T.K., Halvorsen, K.T., Kleiven, A.R., Jiao, L., Goodwin, M.: Temperate fish detection and classification: a deep learning based approach. Appl. Intell., 1–14 (2022)
27. Pan, T.-S., Huang, H.-C., Lee, J.-C., Chen, C.-H.: Multi-scale ResNet for real-time underwater object detection. SIViP **15**, 941–949 (2021)
28. Ayob, A., Khairuddin, K., Mustafah, Y., Salisa, A., Kadir, K.: Analysis of pruned neural networks (mobilenetv2-yolo v2) for underwater object detection. In: Proceedings of the 11th National Technical Seminar on Unmanned System Technology 2019: NUSYS'19, pp. 87–98 (2021). Springer
29. Jalal, A., Salman, A., Mian, A., Shortis, M., Shafait, F.: Fish detection and species classification in underwater environments using deep learning with temporal information. Eco. Inf. **57**, 101088 (2020)
30. Jian, M.W., Liu, X.Y., Luo, H.J., Lu, X.W., Yu, H., Dong, J.Y.: Underwater image processing and analysis: a review. Signal Process. Image Commun. **91**, 116088 (2021)
31. Jian, M.W., Qi, Q., Dong, J.Y., Yin, Y.L., Lam, K.-M.: Integrating QDWD with pattern distinctness and local contrast for underwater saliency detection. J. Vis. Commun. Image Rep. **53**, 31–41 (2018)
32. Jian, M.W., Qi, Q., Yu, H., Dong, J.Y., Cui, C.R., Nie, X.S., Zhang, H.X., Yin, Y.L., Lam, K.-M.: The extended marine underwater environment database and baseline evaluations. Appl. Soft Comput. **80**, 425–437 (2019)

33. Lin, W.-H., Zhong, J.-X., Liu, S., Li, T., Li, G.: Roimix: proposal-fusion among multiple images for underwater object detection. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2588–2592 (2020). IEEE
34. Xu, F.Q., Wang, H.B., Peng, J.J., Fu, X.P.: Scale-aware feature pyramid architecture for marine object detection. Neural Comput. Appl. **33**, 3637–3653 (2021)
35. Xu, F.Q., Wang, H.B., Sun, X.D., Fu, X.P.: Refined marine object detector with attention-based spatial pyramid pooling networks and bidirectional feature fusion strategy. Neural Comput. Appl. **34**(17), 14881–14894 (2022)
36. Zhang, S., Wen, L., Bian, X., Lei, Z., Li, S.Z.: Mask r-cnn. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 2961–2969 (2017)
37. Cai, Z., Vasconcelos, N.: Cascade r-cnn: Delving into high quality object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6154–6162 (2018)
38. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10012–10022 (2021)
39. Yang, J., Li, C., Zhang, P., Dai, X., Xiao, B., Yuan, L., Gao, J.: Focal self-attention for local-global interactions in vision transformers. arXiv preprint arXiv:2107.00641 (2021)
40. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: Unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 779–788 (2016)
41. Redmon, J., Farhadi, A.: Yolo9000: better, faster, stronger. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7263–7271 (2017)
42. Redmon, J., Farhadi, A.: Yolov3: an incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
43. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
44. Jocher, G., et al: Yolov5. https://github.com/ultralytics/yolov5 (2021)
45. Yolov6: a single-stage object detection framework dedicated to industrial applications. https://github.com/meituan/YOLOv6 (2022)
46. Wang, C.Y., Bochkovskiy, A., Liao, H.Y.: Yolov7: trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. arXiv preprint arXiv:2207.02696 (2022)
47. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
48. Underwater robot picking contest. http://www.cnurpc.org/
49. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (VOC) challenge. Int. J. Comput. Vision **88**(2), 303–338 (2010)
50. Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vision **111**(1), 98–136 (2015)
51. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV), pp. 618–626 (2017)
52. Hou, Q., Zhou, D., Feng, J.: Coordinate attention for efficient mobile network design. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 13713–13722 (2021)



Xin Shen is currently pursuing the Ph.D. degree in computer science and technology from Dalian Maritime University, China. His current research interests focus on image processing, computer vision, and object detection.

Huibing Wang received the Ph.D. degree from the School of Computer Science and Technology, Dalian University of Technology, Dalian, Liaoning, China, in 2018. From 2016 to 2017, he was a Visiting Scholar at The University of Adelaide, Adelaide, SA, Australia. Now, he is an Associate Professor at Dalian Maritime University. He has authored or coauthored more than 50 papers in some famous journals and conferences, including IJCAI, ACM MM, IEEE Transactions On Neural Networks And Learning Systems, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS: SYSTEMS, IEEE MULTIMEDIA, ECCV, ICMR, and ICME. His research interests include computer vision and machine learning. Dr. Wang serves as a Reviewer for IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, IEEE TRANSACTIONS ON IMAGE PROCESSING, ACM TOIS, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, IEEE TRANSACTIONS ON COGNITIVE AND DEVELOPMENTAL SYSTEMS, ACM TOMM, Information Fusion, Neurocomputing, and Pattern Recognition Letters, and as a PC Member for AAAI 2022, AAAI 2021, IJCAI 2020, ICMR 2021, ICPR 2020, and ICME 2019.

Tianxiang Cui is currently working toward the M.S. degree with the Information Science and Technology College, Dalian Maritime University, Dalian, China. His research interests include image processing, computer vision, and object search.

Zhicheng Guo is currently working toward the M.S. degree with the Information Science and Technology College, Dalian Maritime University, Dalian, China. His research interests include image processing, computer vision, and object tracking.

Xianping Fu received the Ph.D. degree from Dalian Maritime University, Dalian, China, in 2005. He is currently a Full Professor at Dalian Maritime University (DMU). He previously worked as a Post-Doctoral Researcher at Tsinghua University, Beijing, China, in 2008, and a Senior Research Fellow at Harvard University, Cambridge, MA, USA, in 2009. Now, he is working as the Dean of the College of Information Science and Technology, DMU, and the Director of the Liaoning Underwater Robot Engineering Research Center. His major research interests are image processing for content recognition, multimedia technology, and underwater robot vision. He has authored over 100 journals and conference papers in these areas, which have been published in IJCAI, IEEE TRANSACTIONS ON MULTIMEDIA, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, ICME, ICMR, and OCEANs. Dr. Fu won the American RPB International Scholar Research Award in 2009. His group was included in the Liaoning Revitalization Talents Program.