Journal of
Marine Science and Engineering

MDPI

Article

Seeing Through Turbid Waters: A Lightweight and Frequency-Sensitive Detector with an Attention Mechanism for Underwater Objects

Shibo Song and Bing Sun *

Department of Electrical Engineering, Shanghai Maritime University, 1550 Haigang Avenue, Shanghai 201306, China; songshibo0055@stu.shmtu.edu.cn
* Correspondence: bingsun@shmtu.edu.cn

Abstract

Precise underwater object detectors can provide Autonomous Underwater Vehicles (AUVs) with good situational awareness in underwater environments, supporting a wide range of unmanned exploration missions. However, the quality of optical imaging is often insufficient to support high detector accuracy due to poor lighting and the complexity of underwater environments. Therefore, this paper develops an efficient and precise object detector that maintains high recognition accuracy on degraded underwater images. We design a Cross Spatial Global Perceptual Attention (CSGPA) mechanism to achieve accurate recognition of target and background information. We then construct an Efficient Multi-Scale Weighting Feature Pyramid Network (EMWFPN) to eliminate computational redundancy and increase the model's feature-representation ability. The proposed Occlusion-Robust Wavelet Network (ORWNet) enables the model to handle fine-grained frequency-domain information, enhancing robustness to occluded objects. Finally, EMASlideloss is introduced to alleviate sample-distribution imbalance in underwater datasets. Our architecture achieves 81.8% and 83.8% mAP on the DUO and UW6C datasets, respectively, with only 7.2 GFLOPs, outperforming baseline models and balancing detection precision with computational efficiency.

Keywords: object detection; deep learning; attention mechanism; multiscale feature fusion; autonomous underwater vehicles

check for updates

Academic Editor: Rafael Morales

Received: 8 July 2025
Revised: 5 August 2025
Accepted: 7 August 2025
Published: 9 August 2025

Citation: Song, S.; Sun, B. Seeing Through Turbid Waters: A Lightweight and Frequency-Sensitive Detector with an Attention Mechanism for Underwater Objects. *J. Mar. Sci. Eng.* **2025**, *13*, 1528. https://doi.org/10.3390/jmse13081528

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

As humankind continues to exploit the Earth's resources, especially as our understanding of the land has slowly deepened, there has been a gradual shift in focus towards the oceans, which cover most of the Earth's surface The oceans contain abundant resources, but their exploration faces significant challenges. Early exploration efforts were limited to manned vehicles, but the harsh and complex underwater environment posed a considerable threat to the safety of the vehicle operator [1]. In recent years, the development of Autonomous Underwater Vehicles (AUVs) had revolutionized approaches to this problem. AUVs use an object detector to receive information from the surrounding environment. However, underwater imaging is more challenging than imaging on land due to poor lighting, refraction, and other issues. The complex environment can also make it difficult for detectors to recognize targets [2]. Furthermore, underwater robots usually lack the computing power to support high-complexity detectors. Therefore, it is crucial to develop a lightweight detector that balances accuracy and computing burden.

The main types of object detectors in use today are two-stage detectors and one-stage detectors. Two-stage detectors are represented by Faster-RCNN [3], which separated the recognition task from the classification task and obtained better detection results. In contrast, one-stage detectors, as exemplified by SSD [4] and RetineNet [5], complete the object-detection task in a single step. In particular, the models within the YOLO [6] series have undergone numerous iterations and are recognized within the industry for their high recognition accuracy and lightweight network framework, particularly in computationally demanding situations such as those encountered by embedded devices, while still ensuring stable operation. In addition, some transformer-based object-detection methods have attracted the attention of scholars. Gao et al. [7] proposed a path-enhanced Transformer framework that employs flexible point detection and a hierarchical loss-function design, thereby improving accuracy and efficiency and enhancing the capacity for semantic representation of small underwater targets. In the context of RGB-IR object detection, Yuan et al. proposed CAGTDet [8], which integrates a geometric alignment module and a complementary fusion transformer to alleviate cross-modal alignment bias and fusion inaccuracy. Zhao et al. [9] designed the OrientedFormer model by introducing Gaussian position encoding and the Wasserstein attention mechanism. This model can effectively model angles, positions, and geometric relationships, thereby enabling the detection of oriented targets in remote sensing images. Furthermore, Mamba is a relatively novel deep learning method. Tang et al. [10] proposed an improved YOLO-based framework that integrates SPPF-Mamba and multiscale adaptive spatial attention modules to enhance the fusion of global and local features, thereby improving the detection of small targets in remote sensing images. SLB-Mamba [11] combines reward-weighted attention with visual state-space feature pyramids to achieve accurate and efficient action detection in both closed and open classroom scenarios. Xing et al. proposed LightEMNet [12], a scribble-supervised lightweight ORSI-SOD network that integrates edge-aware refinement and cross-scale semantic interactions to balance detection performance and deployment efficiency. However, there is a lack of research on target-detection tasks in complex underwater scenes.

For underwater fields, a number of effective underwater object detectors have been proposed. ROIAttn [13] used the multihead self-attention mechanism to process the region of interest in the feature map. This enabled modeling of the dependencies between features in the region of interest, thereby improving detection performance. Song et al. [14] proposed a two-stage underwater object detector based on RCNN improvement that could combine the candidate region score in the first stage with the classification and regression scores in the second stage for evaluation. ERL-Net [15] addressed the problem of recognition difficulties caused by the low contrast between target and background by focusing on edge information for objects with more distinctive features. Some recent studies have explored the instability of underwater features through theoretical modeling; for example, Zhou et al. [16] design a backbone network based on partial differential equations and combining it with spatial residual constraints to alleviate feature drift issues. Furthermore, to improve generalization capabilities in the underwater domain, SPIR-DFM [17] extracts representations that are both invariant and task-relevant through instance normalization and domain-specific feature decomposition. However, on the one hand, these methods struggle to process multiscale feature maps efficiently, resulting in weak detection of small objects and of objects of varying sizes, and their performance in underwater degraded image needs to be improved. On the other hand, these methods tend to be more computationally intensive, which inevitably adversely affects computational efficiency.

Therefore, we propose a lightweight real-time underwater object detector. It is able to achieve higher accuracy in detecting blurred, occluded, and multiscale objects and can function in complex underwater scenes with low illumination and low contrast while

ensuring computational efficiency and detection speed. The main contributions of this paper are as follows.

- To address difficulties associated with the low-contrast and low-light underwater environment, as well as with fuzzy objects and objects with complex backgrounds, we propose Cross Spatial Global Perceptual Attention (CSGPA), which can effectively model the global dependencies of underwater feature maps and avoid the loss of potential relationships of local feature information in order to improve the model's attention to the region of interest.
- To improve the ability of the model to detect multiscale objects and small objects and to reduce the computational burden of the model, we construct the Efficient Multi-Scale Weighting Feature Pyramid Network (EMWFPN), which is able to adaptively adjust the fusion weights of the feature maps of different scales and adopt a simple network for feature extraction to reduce the computational redundancy.
- To address the problem of low recognition rates caused by object occlusion and missing features, we construct a detection head based on our Occlusion-Robust Wavelet Network (ORWNet), which is able to achieve a large receptive field with small computational complexity and model the feature relationships between the feature channels.
- Due to the serious imbalance of sample distributions in underwater datasets, we introduce EMASlideloss as the loss function. This loss function can dynamically adjust the criteria for identifying the hard samples and strengthen the adjustment of the weights of the hard samples to increase the robustness of the model.

The remainder of this paper is organized as follows. Section 2 reviews the literature related to the methods presented in this paper. Section 3 provides a detailed introduction to the proposed detection framework, including CSGPA attention, EMWFPN feature fusion, ORWNet detection head, and EMA sliding loss. Section 4 presents the experimental setup and results. Section 5 provides an in-depth discussion of the contributions of each module and their impact on performance. Finally, Section 6 summarizes this paper and provides an overview.

2. Related Work

Attention Mechanism. In order to facilitate the model's focus on valid information in complex scenarios, attention mechanisms were frequently employed to assign suitable weights to important information in the feature map. Ouyang et al. [18] devised a method to avoid the potential information loss caused by channel compression. Frequency Channel Attention (FCA) Net [19] reexamined the global average pooling operation from the perspective of the frequency domain. The authors elucidated the correlation between global average pooling and frequency-domain feature decomposition. They demonstrated that, in order to optimize network performance, frequency components should be utilized in place of pooling operations in SE. The swin transformer [20], proposed by Liu et al., performed self-attention computation in a local window using the same structure used by the traditional CNN to perform this operation on feature maps of different scales. This adaptability to high-resolution task aliasing and reduced computational load is effective. Pan et al. [21] found a potential relationship between convolution and self-attention that enhances the performance of the attention mechanism. In addition, in a departure from traditional attention mechanisms, Shen et al. proposed UCOEA [22], which employs clustering strategies in both the channel and spatial dimensions to reduce redundancy and improve detection efficiency in underwater scenes.

Multi-Scale Feature Fusion Networks. Combining information containing a high level of complexity and detail with abstract semantic information can improve the feature representation of the network. Feature Pyramid Network (FPN) [23] utilized a layer-by-

layer propagation of high-level features into the underlying network, thus achieving a synergistic integration of semantic and detailed information. Bi-directional Feature Pyramid Network(BiFPN) [24] was characterized by its incorporation of Cross-Scale Connections. This serves to enhance information transfer efficiency and mitigate the appearance of feature loss. In addition, Generalized Feature Pyramid Network(GFPN) [25] integrates the benefits of the aforementioned networks with internal block connectivity and enhanced cross-scale connectivity to optimize the expressiveness of the feature of the network. Meanwhile, STAMF [26] demonstrates that adding input dimensions can significantly improve the recognizability of underwater objects by fusing RGB and polarization information through a novel cross-modal framework. Furthermore, Li et al. proposed the ARAN algorithm [27], which integrates multiscale attention mechanisms with dynamic feature-aggregation techniques to improve spatial resolution and the consistency of feature fusion. Nevertheless, this approach results in a substantial computational burden.

Large Receptive Field Convolution. Larger convolution kernels allow for a larger receptive field than standard convolution, allowing for more shape features to be learned. However, simply increasing the size of the convolutional kernel increases the computational complexity exponentially. The introduction of dilates in the convolution process has been proposed as a means of expanding the effective receptive field of the convolution kernel [28]. Nevertheless, this approach has been observed to result in the appearance of artifacts in the output feature map. Ding et al. [29] successfully extended the convolution kernel to 31×31 by using multiple 3×3 convolution kernels to simulate a larger convolution kernel. This avoids the huge computational effort associated with increasing the number of receptive fields. However, each of these methods can use larger convolutional sums for feature extraction on the feature map only at a single scale.

3. Method

3.1. Overall Architecture

The structure of our model is shown in Figure 1. The network is divided into three primary components: the backbone, the neck, and the detection head. The backbone is responsible for extracting the features of the input image and generating feature maps at different scales. We devise and add the Cross Spatial Global Perceptual Attention (CSGPA) mechanism after the backbone network to improve the perceptual ability of the model, enhance the focus on the target region, and achieve higher feature representation. The neck is responsible for feature fusion, and we design an Efficient Multi-Scale Weighting Feature Pyramid (EMWFPN) to enhance the diversity of feature maps fused at each stage, using feature-weighted fusion to achieve a more efficient fusion strategy. A concise feature-extraction module is also used to reduce the redundancy of convolutional computation. Subsequent to this, we construct a detection head based on the Occlusion-Robust Wavelet Network (ORWNet) proposed by us, which detects objects in the output of the neck network and returns the gradient required for model training. The detection head is able to process features at varying frequency-domain scales independently while concurrently reducing computation and achieving a higher receptive field. This head also exhibits superior adaptability in the detection of underwater occluded objects. Finally, to address the imbalance between hard and easy samples in the underwater object-detection task, we introduce EMASlideloss to dynamically assign different learning weights to hard and easy samples during the training process in order to improve the training results of the model. The subsequent sections of this chapter will provide comprehensive explanations of the different parts of the our model.

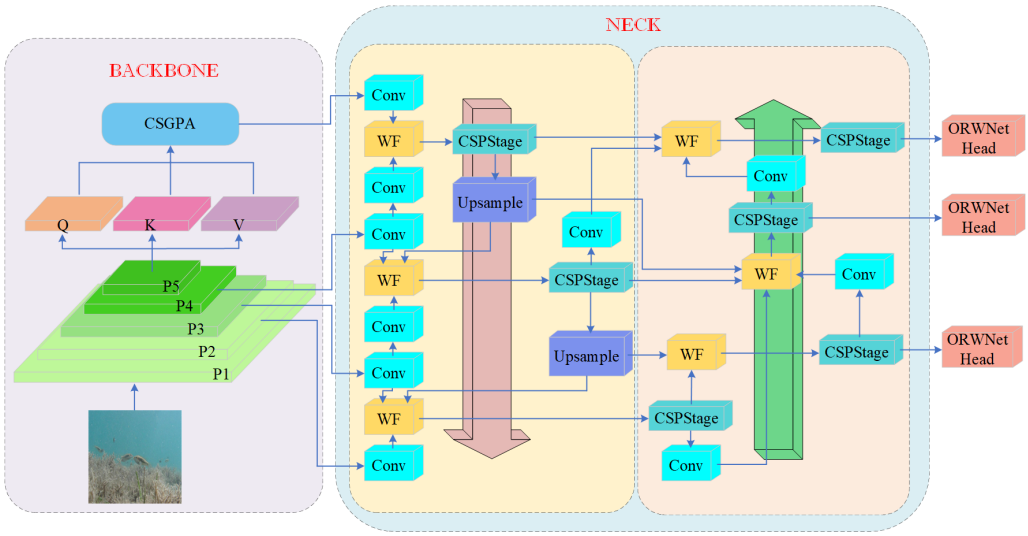


Figure 1. Overall architecture.

3.2. Cross-Spatial Global Perceptual Attention

Convolutional neural networks with inductive biases are better able to capture geometric and topological information. However, this also results in a limited receptive field for convolution, which can only focus on local information and lack the ability to model long-range dependencies [30]. In summary, the aim is to enable the model to focus its attention on the region of interest when the objects cannot be distinctly distinguished, as in underwater images with low contrast and complex backgrounds, etc. We proposed the Cross Spatial Global Perceptual Attention(CSGPA), the structure of which is shown in Figure 2; the detailed structures of two parts are shown in Figure 3.

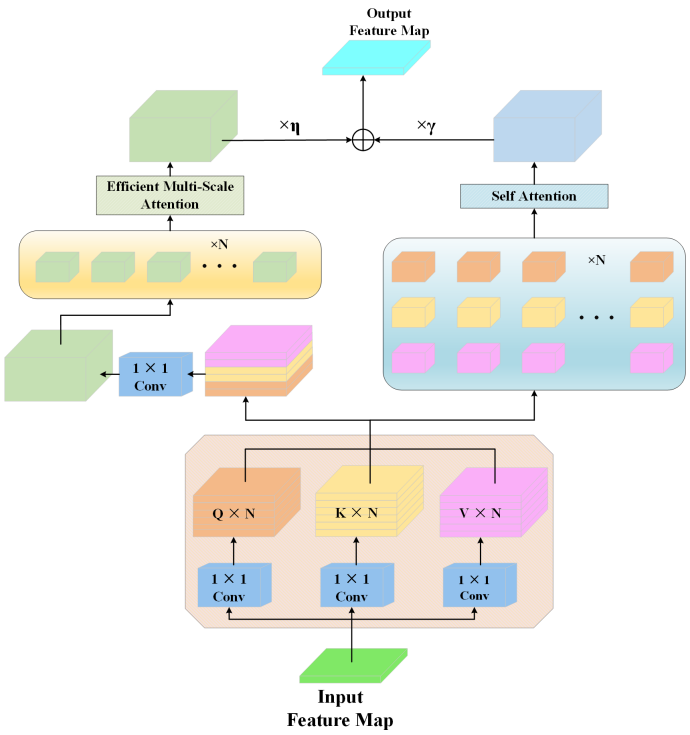


Figure 2. An illustration of the proposed CSGPA module. The input feature map is first mapped by a convolution to N groups Q, K, V, N being the number of heads of the multihead Self-Attention. The right side corresponds to the self-attention mechanism, while the left side corresponds to the efficient multi-scale attention mechanism. Finally, the outputs from both sides are concatenated by weighting them with γ and η, respectively, γ and η being two learnable parameters.

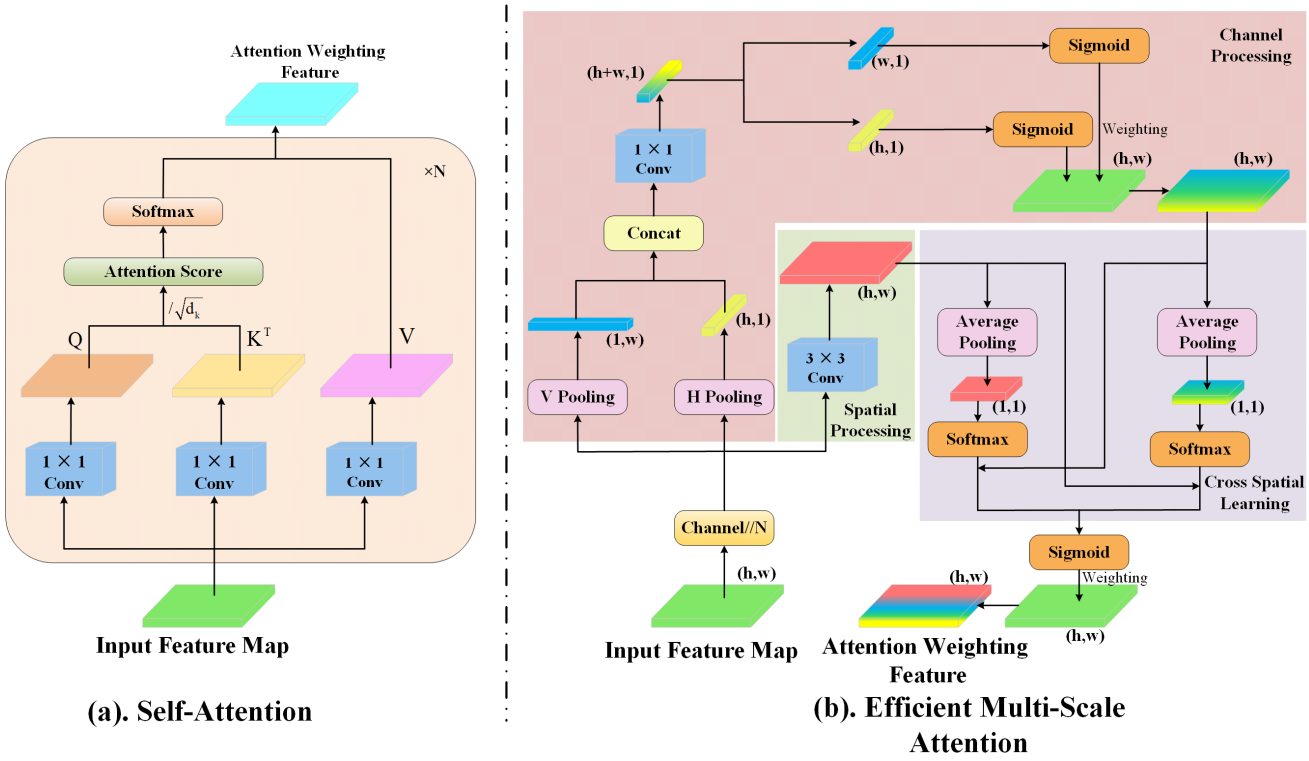


Figure 3. Detailed structural diagrams of Self-Attention and Efficient Multi-Scale Attention.

Firstly, by reconstruction of the structure of self-attention, mapping of the input feature map into N sets Q, K, and V. The formula for Self-Attention proposed by the Transformer architecture is as follows:

$$Q = XW_Q, \quad K = XW_K, \quad V = XW_V \tag{1}$$

$$Output = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{2}$$

where Q, K, and V are the query, key, and value of the input X mapped through the parameter matrix, $Q,\ K,\ V \in \mathbb{R}^{B\times h\times w\times C}$. B is the batch size, and C represents the channel number. The width and height of the feature map are denoted by w and h. For multihead attention, with the aim of making each head responsible for learning independent attentional representations, the shapes of Q, K, and V need to be reshaped according to (3) as $Q,\ K,\ V \in \mathbb{R}^{B\times N\times N_c\times hw}$, N_c is the number of channels each head is responsible for.

$$N_c = \frac{C}{N} \tag{3}$$

Similarly, let I be the input feature map where (m, n) represents a pixel on the feature map such that $m \in \{0, 1, \ldots, w\}, n \in \{0, 1, \ldots, h\}$. We equate the mapping of the input feature map through the parameter matrices to Q, K, and V as a mapping after 1×1 convolutions and reconstruct the self-attention formula as follows:

$$\begin{cases} Q_{mn} = Conv_Q^{1\times1}(I_{mn}), \quad K_{mn} = Conv_K^{1\times1}(I_{mn}), \quad V_{mn} = Conv_V^{1\times1}(I_{mn}) \\ Conv_Q^{1\times1} = W_Q, \quad Conv_K^{1\times1} = W_K, \quad Conv_V^{1\times1} = W_V \end{cases} \tag{4}$$

where W_Q, W_K, and W_V are parameter matrices. The feature map will then go down two separate paths, one path into the self-attention mechanism and the other path into the

Efficient Multi-Scale Attention mechanism. For self-attentive paths, combining (1) and (2), the expression of self-attention on a feature map can be written as follows:

$$O_{mn} = (softmax(\frac{(I_{mn}W_Q)(I_{wh}W_K)^T}{\sqrt{d_k}}))I_{wh}W_V \tag{5}$$

where O is the output feature map of Self-Attention. Then, combine (4) and (5); in the case of a multihead self-attention system with N heads, the output of the self-attention part O_{SA}^{mn} can be formulated as follows:

$$O_S^{mn} = \sum_{i=1}^{N}(softmax(\frac{Q_{mn}^i(K_{mn}^i)^T}{\sqrt{d_k}})V_{mn}^i) \tag{6}$$

(6) corresponds to the stream of Figure 3a, and reconstruction of the self-attention mechanism enables the implementation of the structure shown in Figure 3b.

The other path allows the model to have a richer feature representation while avoiding the loss of semantic information due to the reduction of channel dimensions when processing deep visual features. It is necessary to concatenate Q, K, and V along the N dimension and linearly transform them by a 1×1 convolution to generate the feature map $I_E \in \mathbb{R}^{B\times(k_c\times N_c)\times h\times w}$, which is input to the EMA module; k_c indicates the number of weights after convolution kernel expansion.

As shown in Figure 3b, the EMA module is mainly composed of four parts: Feature Grouping, Channel Processing, Spatial Processing, and Cross Spatial Learning. First, in the Feature Grouping stage, we group the feature maps which are fed into the EMA according to the number of heads N to enable collaboration between EMA and self-attention and maintain the consistency of computation. That is, the number of channels each head is responsible for processing is $\frac{k_c\times C}{N^2}$.

In the channel-processing stage, two one-dimensional global average poolings are used to encode channel information in the horizontal and vertical directions, respectively. As a result, channel relationships can be captured in one direction and precise position information can be retained in the other. Afterwards, after the two outputs have been spliced along the h-dimension and linearly transformed using a 1×1 convolution, the formula for this process is as follows:

$$O_C = Conv^{1\times1}(Concat(\frac{1}{w}\sum_{0\le m\le w} I_E(m,h),\ \frac{1}{h}\sum_{0\le n\le h} I_E(w,n))) \tag{7}$$

Next, reconstruct it into two one-dimensional vectors representing the channel information in the horizontal and numerical directions. The channel attention weights in different spatial directions are acquired using a sigmoid function and weighted to the input feature map. The output of the channel process is as follows:

$$O_{CP} = Sigmoid(O_C) \cdot I_E \tag{8}$$

After that, the 3×3 convolution is used to capture multiscale spatial feature representations; it can be formulated as follows:

$$O_{SP} = Conv^{3\times3}(I_E) \tag{9}$$

In the cross-space learning part, the outputs of the channel-processing stage and the spatial-processing stage are each global-average pooled and then passed into the softmax function to obtain the attention matrix; the formula for two-dimensional global average pooling is expressed as follows:

$$GAP = \frac{1}{w \times h} \sum_{m=1}^{w} \sum_{n=1}^{h} I(m,n) \tag{10}$$

Afterwards, they are separately weighted to each other's output feature maps to enable local features to guide global semantic information and enhance local feature representation, while global channel features guide local feature learning to mitigate global information loss. The resulting operation is described by the following equation:

$$O_{CS} = Concat \begin{cases} Softmax(\frac{1}{w \times h} \sum_{m=1}^{w} \sum_{n=1}^{h} O_{SP}(m,n)) \cdot O_{CP} \\ Softmax(\frac{1}{w \times h} \sum_{m=1}^{w} \sum_{n=1}^{h} O_{CP}(m,n)) \cdot O_{SP} \end{cases} \tag{11}$$

$$O_E = Sigmoid(O_{CS}) \cdot I_E \tag{12}$$

Finally, the output of the CSGPA is obtained as follows:

$$O = \gamma \cdot O_s + \eta \cdot O_E \tag{13}$$

where γ and η are two learnable parameters used to dynamically adjust the weighting of the fusion of the two paths. As demonstrated in the CSGPA module, EMA and self-attention can work in parallel and reinforce each other through effective integration and streamlined processing. CSGPA not only builds long-range feature-dependent models to capture cross-regional information to better discriminate between objects and backgrounds; it also aggregates cross-spatial information of different spatial dimensions to enrich feature representations and enhance feature learning in semantic regions to improve the model's focus on target regions in underwater degraded images.

3.3. Efficient Multi-Scale Weighting Feature Pyramid Network

In the CNN-based object-detection task, shallow feature maps contain extensive fine-grained information such as texture and edges due to their superior image resolution. The deep feature carries more abstract semantic information, such as the overall contour of the target. For this reason, it is necessary to effectively fuse different scales of feature information to improve the performance of the object-detection network [31]. Feature Pyramid Network (FPN) fuses deep feature maps with shallow feature maps through simple up-sampling and concatenating operations to generate multiscale feature maps. The PANet in Figure 4b adds top-down feature channels to the bottom-up channels to integrate richer features. More cross-connections are introduced in the structure of Figure 4c, the approach resembles a residual network in concept, but this architecture may impose some constraints on the network. Although all of the above methods fuse feature information from different layers to some extent, one common disadvantage is that none of them incorporate feature information from the P2 layer into the network.

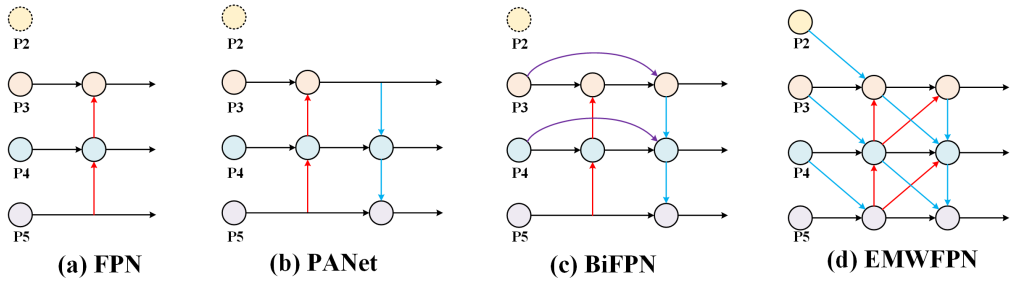


Figure 4. A sketch of different Feature Pyramid Network structures.

However, this layer is the backbone network with the feature map with the highest spatial resolution, making it essential for the detection of small and clustered objects. Thus, we propose the EMWFPN shown in Figure 4d. In the initial stage of this network, we propagate the shallow features into the deeper layers of the network to help the model learn the features; in the second stage of feature fusion, we adopt as many connections as possible so that the feature maps of the bottom-up channel and the top-down channel can have more feature interactions, which enhances the network's capacity for feature representation. EMWFPN enriches the fusion capability of multiscale feature maps, enables the network to autonomously enhance attention to more important scale features, improves the capacity of the model for feature expression, and reduces. computational redundancy.

Weighted Fusion. The conventional concatenation operation merely links two feature maps without distinguishing the relative importance of different features, which can lead to unbalanced information. To enhance the effectiveness of feature fusion, we propose to replace concatenation with weighted fusion. To illustrate this approach, consider the up-sampling stage and down-sampling stage of the P4 layer in Figure 4d. The fusion process is outlined as follows:

$$P_4^M = \frac{\alpha^M \cdot P_4^I + \beta^M \cdot DS(P_3^I) + \lambda^M \cdot Up(P_5^M)}{\alpha^M + \beta^M + \lambda^M + \theta} \tag{14}$$

$$P_4^O = \frac{\alpha^O \cdot P_4^M + \beta^O \cdot DS(P_3^M) + \lambda^O \cdot Up(P_5^M) + \mu^O \cdot DS(P_3^O)}{\alpha^O + \beta^O + \lambda^O + \mu^O + \theta} \tag{15}$$

where Up and DS are upsampling and downsampling operations, respectively. The four learnable parameters α, β, λ, and μ enable the dynamic adjustment of the proportion of weights when fusing feature maps at different scales. Parameter θ is used to avoid the denominator being equal to 0; furthermore, $0 < \theta \ll 1$.

Lightweight Feature-Extraction Module. As the convolution neural network becomes deeper, a massive inference computation problem arises because there is too much repetitive gradient information in the network, which produces severe computational redundancy and leads to a loss of efficiency [32]. Therefore, in EMWFPN, we adopt CSPStage [33] as the feature-extraction module. It can effectively reduce gradient redundancy, improve computational efficiency, and lower model complexity. The structure of CSPStage is shown in Figure 5. In order to minimize repetitive gradient computation, the input is first divided into two parts along the channel dimension and feature mapped separately using 1×1 convolutions. Subsequently, one part undergoes feature extraction by repeated applications of RepConv and 3×3 convolutions. RepConv incorporates a parallel structure of 3×3 convolution, 1×1 convolution, and cross-connection, enabling multiscale feature extraction. The other part bypasses this manipulation to avoid gradient redundancy. Eventually, the output is derived by splicing the two parts and passing them through a single 1×1 convolution for channel compression.

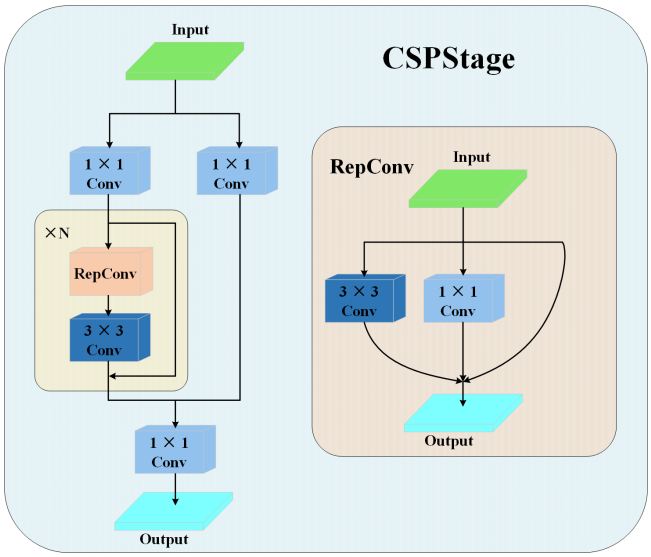


Figure 5. The structure of CSPSage.

3.4. Occlusion-Robust Wavelet Network-Based Head

It is notable that underwater organisms frequently exhibit collective behavior and that the intricate nature of the underwater environment can result in object occlusion within underwater images. This phenomenon can lead to local aliasing and problems of missing features, with a subsequent impact on detection performance [34]. To address these problems, processing of the feature map in the subfrequency domain is an effective approach [35]. Therefore, we construct a frequency-sensitive detection head for occluded objects. This head divides the feature information into separate frequency domains for processing. This significantly reduces computational load while maintaining a large receptive field. Meanwhile, it achieve better results for the detection of occluded targets. The network structure is illustrated in Figure 6.

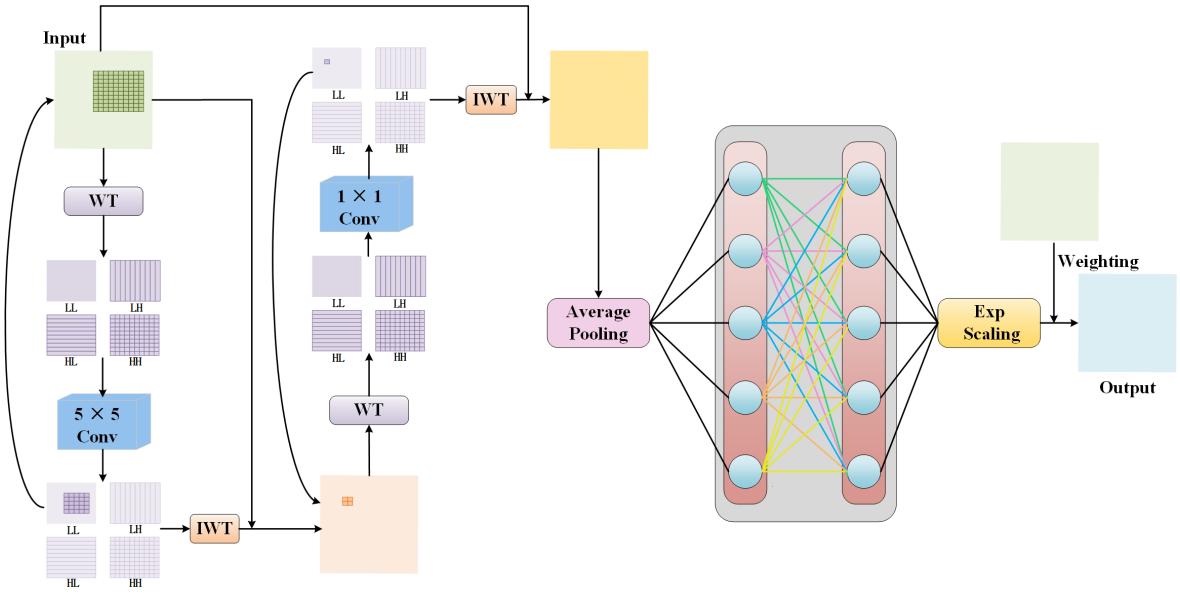


Figure 6. The network structure of ORWNet.

First, due to the characteristics of wavelet transform, the use of $k \times k$ convolution kernels for the wavelet-transformed feature maps is equivalent to the use of $2k \times 2k$ convolution kernels for the original feature maps, which greatly improves the receptive field of

the network. The convolution process is shown in Figure 6, with the equivalent formula as follows:

$$\sum_{p,q} W_{p,q} I([m/2] + p - [k/2], [n/2] + q - [k/2]) = \sum_{u,v} W_{u,v} I(m + u - [k], n + v - [k]) \quad (16)$$

where W is the kernel weight, I is the input feature, k represents the kernel size of the convolution and $p, q \in \{0, 1, \ldots, k-1\}$, $u, v \in \{0, 1, \ldots, 2k-1\}$. The wavelet convolution process commences with the wavelet transform of the input feature map, which results in the generation of four sub-bands: the low frequency component (LL), the horizontal high-frequency component (LH), the vertical high-frequency component (HL), and the diagonal high-frequency component (HH). These components can be represented as follows:

$$\xi^{\theta}_{j,k_m,k_n} = \sum_{m} \sum_{n} I(m, n) \phi^{\theta}_{j,k_m,k_n}(m, n) \quad (17)$$

where ξ is defined as the wavelet coefficients, θ is the direction index, j represents the frequency scale, k_X and k_Y are the position indexes, and ϕ represents the wavelet basis function. The following formula is employed for the wavelet basis functions:

$$\phi_{j,k_m,k_n}(m, n) = 2^j \phi(2^j m - k_m, 2^j n - k_n) \quad (18)$$

Subsequently, to ensure the stability of the feature-extraction flow and to improve its hierarchical nature, the 5×5 convolution kernel and the 1×1 convolution kernel are used for feature extraction at different scales. This approach is used to enhance the expressiveness of the model. In addition, the inverse wavelet transform is implemented to restore the image, as demonstrated in the subsequent equation:

$$O_{WC}(m, n) = \sum_{j,k_m,k_n}^{\theta} Conv(\tilde{\xi}^{\theta}_{j,k_m,k_n}) \tilde{\phi}^{\theta}_{j,k_m,k_n}(m, n) \quad (19)$$

where $\tilde{\phi}$ is the inverse wavelet basis function and O_{WC} represents the output of the wavelet convolution. Finally, following the implementation of a global average pooling layer and the transmission of data through two fully connected layers to obtain the attention weights, the model is adjusted to focus on the important features. The attention weights are then exponentially scaled to further enhance the salience of the important features, and the output is obtained after weighting to the input feature map. The formula is as follows:

$$O = Exp(FC_2 FC_1(GAP(O_{WC})))I \quad (20)$$

where FC denotes fully connected layers and Exp denotes an exponential scaling operation. It is evident that as a result, ORWNet is capable of processing features across a variety of frequency-domain scales. This, in turn, leads to a substantial reduction in computational redundancy and a concomitant expansion of the receptive field. The upshot of this is that the model is better able to comprehend the contextual relationship that exists between the occluded objects and the surrounding features. In addition, it enhances the model's capacity to adapt to an occluded target.

3.5. Loss Function

There is a critical imbalance in the distribution of categories in the underwater dataset. Thus, if equal attention is given to hard and easy samples in the training stage, the model will focus more on the features of the easy samples and less on those of the hard samples. This will have a detrimental effect on the final performance of the model. In this

paper, we introduce the improved slideloss function based on the Exponential Moving Average(EMA) [36] approach to dynamically modify the focus of the model on hard versus easy samples to optimize the model performance.

Slideloss is a loss function that assigns greater weights to hard samples by setting a weighting function that allows the model to pay more attention to learning hard sample features. It can be written as follows:

$$weight(x, \sigma) = \begin{cases} 1, & x \leq \sigma - 0.2 \\ e^{1-\sigma}, & \sigma - 0.2 < x < \sigma \\ e^{1-x}, & x \geq \sigma \end{cases} \tag{21}$$

where x represents the Intersection over Union (IoU) value of the current sample predicted and σ refers to the set IoU threshold. By comparing the values of x and σ, the sample can be classified as hard or easy. In Slideloss, σ is plotted as the average of the IoUs of all samples and holds a fixed value.

It is noticeable that not all hard samples have higher weights. This is because we consider that samples with IoU values significantly lower than σ probably represent noise or background information or are just negative samples that are relatively easy to classify. Thus, paying more attention to these samples may not result in a greater optimization of the model. In contrast, samples with IoU values that are around σ are possibly samples that are prone to misclassification by the model and can be critical in affecting the performance of the model.

However, as we mentioned, the thresholds for hard and easy samples in Slideloss are set as the average of the IoUs of all the samples, whereas the mean IoUs of the samples in different training phases are not the same as the mean IoUs of all the samples. This approach may thus lead to biased classification of the hard and easy samples. Therefore, we adopt EMA to dynamically adjust the IoU threshold to achieve more accurate classification of hard and easy samples. The formulation for automatically modifying the threshold using EMA is as follows:

$$d = d_0 \cdot (1 - e^{-i/\tau}) \tag{22}$$

$$\sigma_i = d \cdot \sigma_{i-1} + (1 - d) \cdot \sigma_i \tag{23}$$

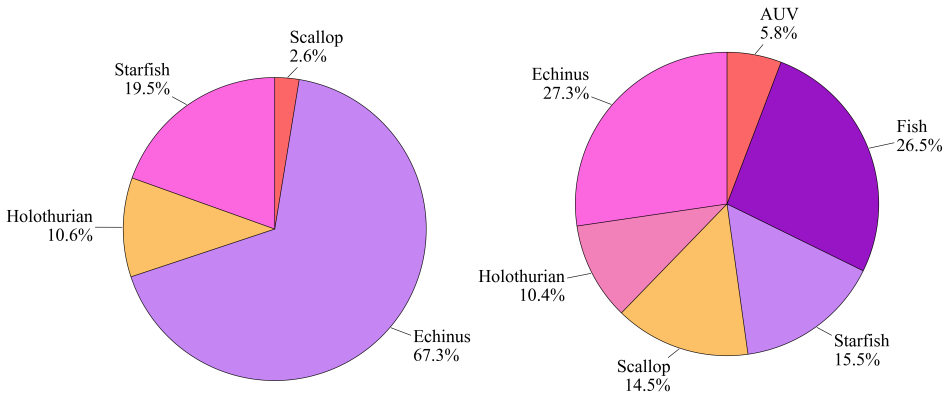
where d is a decay factor dependent on the number of loss updates, d_0 is a decay parameter, i is the number of current loss updates, and τ is a time constant controlling the exponential movement that is used to ensure the smoothness of the updated IoU. Then in (23), by using the IoU mean σ_i of the i-th training round and the IoU mean σ_{i-1} of the previous round in the current training round, incorporation of the decay factor d can enable the computation of an exponential moving average of IoU. With the aforementioned method, EMASlideloss effectively mitigates the sample-imbalance issue in underwater object detection.

4. Experiments

4.1. Datasets

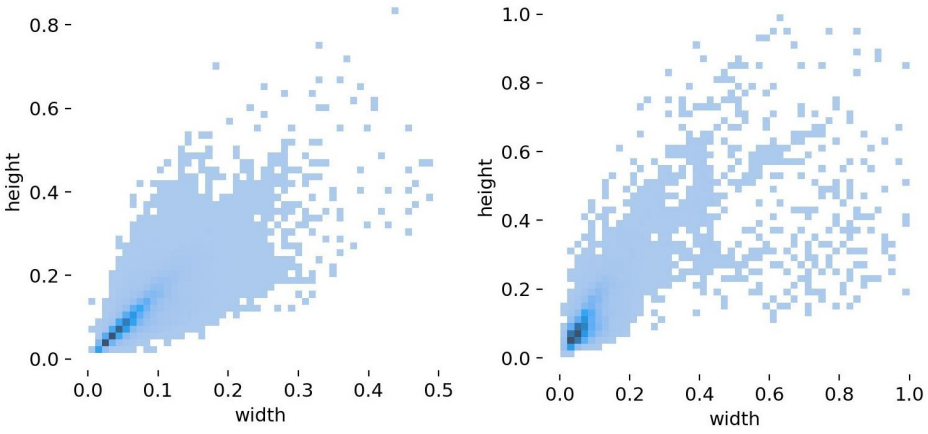
The model is trained using the DUO [37] and UW6C datasets. The DUO dataset is a public underwater object-detection dataset containing 7782 images with careful labels. The four categories of objects detected in this dataset are holothurians, echinus, scallops, and starfish. However, the dataset exhibits significant imbalance in the various categories, as illustrated in Figure 7a. The echinus category accounts for the highest proportion of samples (67.3%), while the scallop category accounts for the lowest (2.6%). This represents an order-of-magnitude difference, which poses challenges in terms of training. The UW6C dataset is a dataset was compiled by us because the DUO dataset has only four categories

and does not include fish, even though fish represent the greatest number of underwater objects. We thus added fish as a new category, using data from DeepFish datasets [38]. We also included Autonomous Underwater Vehicle (AUV) as a new category. Then, we attempted to ensure that the numbers of samples from different categories were as balanced as possible. Accordingly, the UW6C dataset comprises a total of 5064 images categorized into six distinct groups: holothurians, echinus, scallops, starfish, fish, and AUV. These categories account for 10.4%, 27.3%, 14.5%, 15.5%, 26.5%, and 5.8% of the total number of instances in the dataset, respectively. Figures 7 and 8 illustrate the distribution of the different categories of samples and the pixel values of the samples, respectively.



(**a**) DUO (**b**) UW6C

Figure 7. Dataset class distributions.



(**a**) DUO (**b**) UW6C

Figure 8. Dataset sample sizes.

4.2. Experimental Environment and Evaluation Metrics

The experiments in this paper were conducted on the Linux operating system, version Ubuntu 24.04.1 LTS. The CPU model employed is Inter(R) Xeon(R) Platinum 8368 × 152, and the GPU model is NVIDIA A30. The experimental environment was conducted using Python 3.9.20, the version of CUDA was 12.6, and PyTorch version 1.12.1 was used as the deep learning framework. The SGD optimizer was utilized with a batch size of 64, and the training process was executed for a total of 200 epochs. In addition, the learning rate was set to 0.01, the momentum to 0.937, and the weight-decay to 0.0005, all of which are consistent with the baseline default parameters. The random seed was set to 42.

In order to evaluate the performance of the model, the mean average precision (mAP) metric is employed as the evaluation criterion. This metric calculates the mean of the

average precision of all categories contained within the dataset. The calculation of this metric is as outlined below:

$$mAP = \frac{1}{N} \cdot \sum_{i=1}^{N} \int_0^1 P_i(R_i)\, dR_i \tag{24}$$

where N signifies the total number of categories, i denotes the current category, P represents Precision, and R denotes Recall. The following definitions are applicable to P and R:

$$\begin{cases} P = \frac{TP}{TP+FP} \\ R = \frac{TP}{TP+FN} \end{cases} \tag{25}$$

where TP is used to denote correctly detected targets, FP is used to denote incorrectly detected targets, and FN is used to denote missed objects. It can thus be concluded that P and R are partially mutually exclusive, reflecting a trade-off relationship between these two variables. To evaluate P and R in a balanced way, the F1-score is also used to evaluate them, and the formula for the F1-score is as shown in (26). In addition, the parameters and GFLOPs were selected as evaluation criteria for the computational cost of the models.

$$F1 = \frac{2 \times P \times R}{P + R} \tag{26}$$

4.3. Comparison Experiments

In this section, a comparison is drawn between our model and various excellent object-detection models to validate the superiority of the former. Furthermore, a comparison is made between the proposed attention mechanism and the feature-fusion network and other methods to validate their effectiveness.

Comparison with Different Attention Mechanism. In this study, we presented the CSGPA attention mechanism, which we had designed for the purpose of enhancing the model's ability to discriminate between object and background in underwater degraded images and to improve the model's attention to the target region. We then proceeded to compare our designed attentional mechanism with other excellent attentional mechanisms, and the results of these comparisons are shown in Tables 1 and 2. In addition, we present a visual comparison of the evaluation metrics of different attentional mechanisms on different datasets.

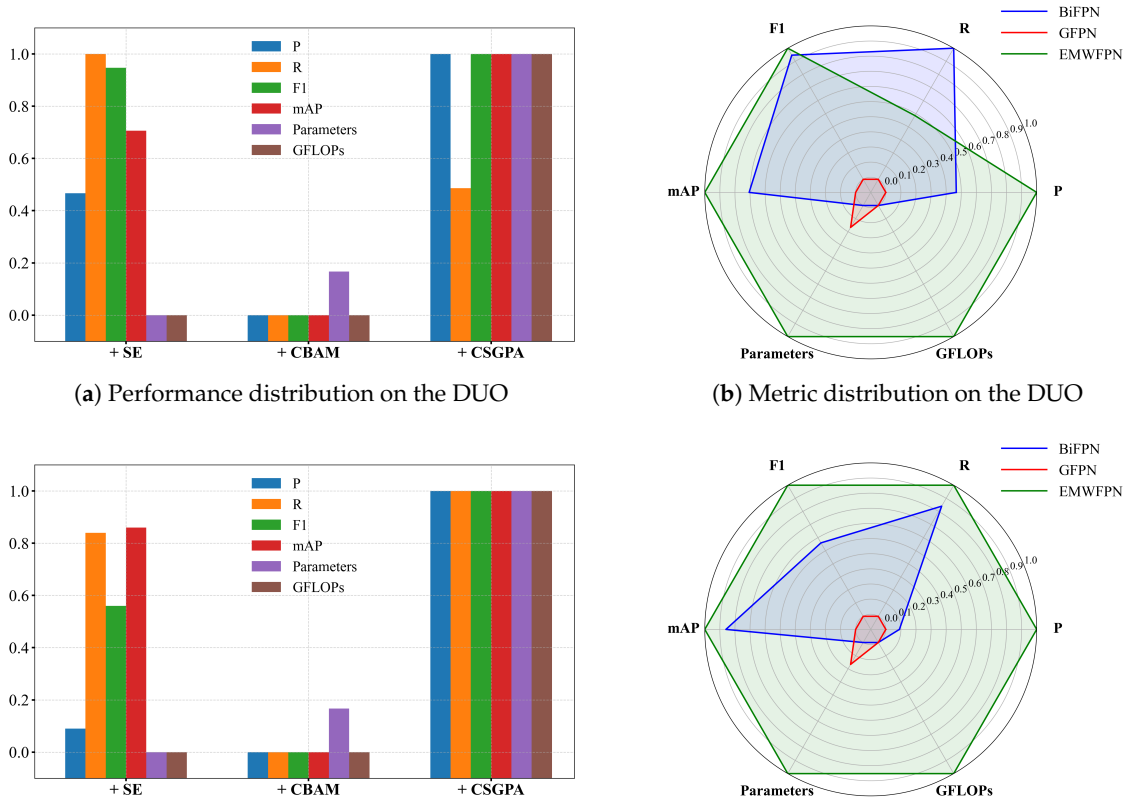
Table 1. Comparison with different attention mechanisms on the DUO datasets.

Model	P	R	F1	mAP@0.5	Parameters	GFLOPs
+SE	83.8	75.1	79.2	83.1	3.0	8.3
+CBAM	83.1	72.5	77.4	81.9	3.1	8.3
+CSGPA	84.6	74.7	79.3	83.6	3.6	9.2

Table 2. Comparison with different attention mechanisms on the UW6C datasets.

Model	P	R	F1	mAP@0.5	Parameters	GFLOPs
+SE	84.0	74.9	79.1	80.9	3.0	8.3
+CBAM	83.9	73.3	78.2	80.3	3.1	8.3
+CSGPA	85.0	75.2	79.8	81.0	3.6	9.2

In both datasets, CBAM yields the lowest score, indicating that this attention mechanism performs suboptimally on underwater images. Conversely, CSGPA performs similarly to SE and has a slightly higher Recall score than CSGPA on the DUO dataset for SE. However, P, F1 and mAP are better for CSGPA, indicating that the proposed attention mechanism leads to a better overall performance of the model and higher prediction precision. This conclusion is further substantiated by the visual representations in the histogram and the radar plot of Figure 9.



(**a**) Performance distribution on the DUO

(**b**) Metric distribution on the DUO

(**c**) Performance distribution on the UW6C

(**d**) Metric distribution on the UW6C

Figure 9. Visual comparison of different attention mechanisms on the DUO and UW6C Datasets.

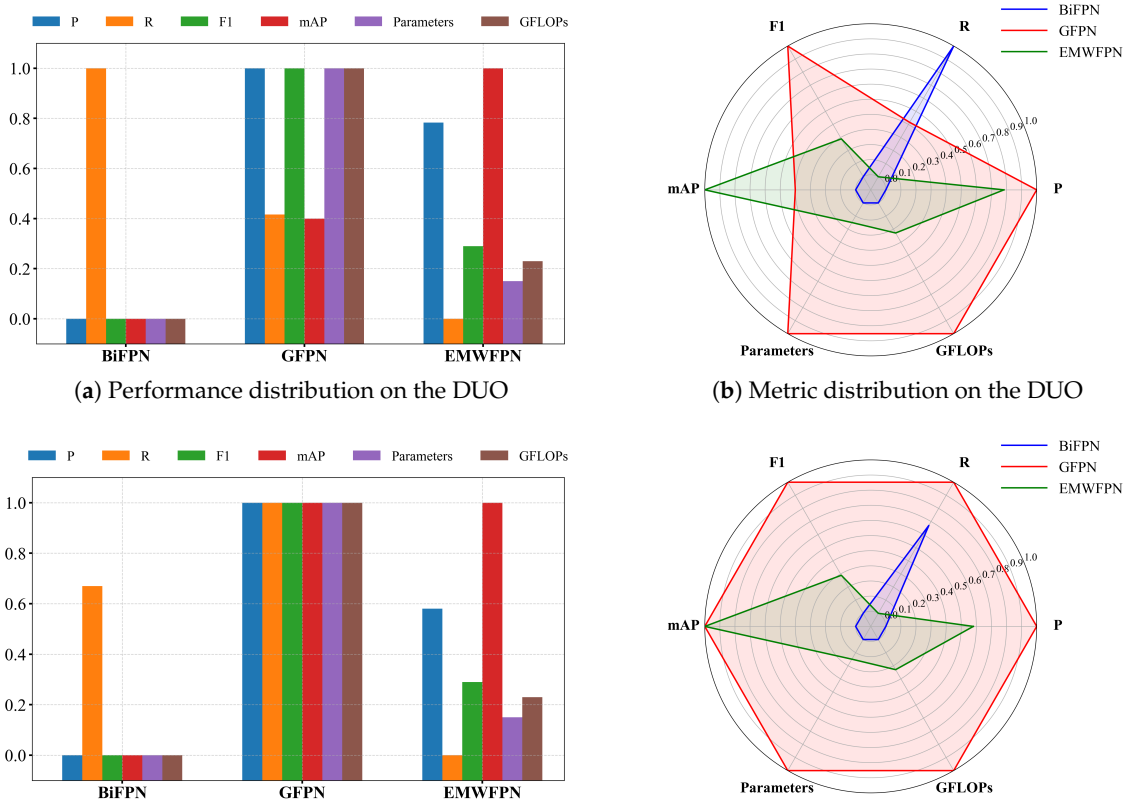
Comparison with different feature-fusion networks. The proposal of EMWFPN is driven by the objective of enhancing the feature representation of the model and reducing the computational redundancy of convolutional operations. To this end, a comparative analysis of different classical feature pyramid networks is shown in Tables 3 and 4, with the aim of substantiating the efficacy of the proposed approach. A visual representation of this comparison is shown in Figure 10.

Table 3. Comparison with different feature-fusion networks on the DUO datasets.

Model	P	R	F1	mAP@0.5	Parameters	GFLOPs
BiFPN [24]	82.4	75.5	78.7	82.6	2.0	7.2
GFPN [25]	84.7	74.8	79.4	82.8	3.3	8.5
EMWFPN	84.2	74.3	78.9	83.1	2.2	7.5

Table 4. Comparison with different feature-fusion networks on the UW6C datasets.

Model	P	R	F1	mAP@0.5	Parameters	GFLOPs
BiFPN	83.0	74.6	78.6	80.4	2.0	7.2
GFPN	86.5	74.9	80.3	80.7	3.3	8.5
EMWFPN	85.0	74.0	79.1	80.7	2.2	7.5



(**a**) Performance distribution on the DUO

(**b**) Metric distribution on the DUO

(**c**) Performance distribution on the UW6C

(**d**) Metric distribution on the UW6C

Figure 10. Visual comparison of different feature-fusion approaches on the DUO and UW6C datasets.

The proposed EMWFPN achieves optimal mAP values of 83.1% and 80.7% on both datasets. It should be noted that the values of P, R, and F1-score of GFPN are higher than those of the proposed network, indicating that the feature-fusion approach employed by our model enhances its ability for target localization and extraction of spatial information. It is conceivable that the classification ability of GFPN may be superior to that of our network. However, the number of parameters of GFPN is 50% higher than ours, suggesting that our approach better balances model efficacy and computational consumption.

Comparison with other object-detection models. In order to evaluate the effectiveness and generalization performance of the methods presented in this manuscript, a comparison was made of different advanced object-detection methods on the DUO and UW6C datasets. The following methods were included in the comparison: Faster-RCNN, SSD, RetinaNet, RT-DETR [39], YOLOv5 [40], YOLOv6 [41], YOLOv8 [42], YOLOv10 [43], YOLOv11 [44]. The experimental results are presented below.

Tables 5 and 6 illustrate the performance of different detection methods on the DUO and UW6C datasets, respectively, while Figures 11 and 12 present their corresponding visual images. These images clearly demonstrate the advantages and disadvantages of the different detection methods in each evaluation index. According to the data presented in the tables, the mAP of our method on the two datasets reaches 83.8% and 81.7%, constituting a substantial improvement compared to alternative methods. Furthermore, the model

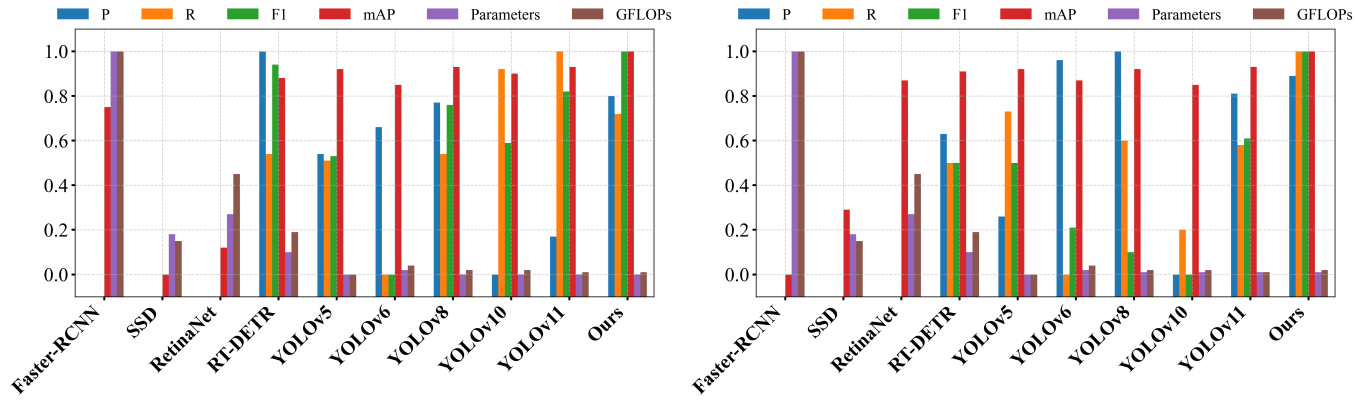
demonstrates remarkable efficiency, with a reduction in parameters to 2.6 and GFLOPs to 7.2, placing it second only to YOLOv5 in terms of lightness. The superior generalization capability and commendable performance of the proposed model on challenging underwater degraded images are evidenced through a range of comparative experiments.

Table 5. Comparison experiment on the DUO datasets.

Model	P	R	F1	mAP@0.5	Parameters(M)	GFLOPs
Faster-RCNN	/	/	/	80.4	137.1	185.1
SSD	/	/	/	70.0	26.3	31.4
RetinaNet	/	/	/	71.7	38.0	85.0
RT-DETR	85.7	74.2	79.5	82.2	15.5	37.9
YOLOv5	84.1	74.1	78.8	82.7	1.8	4.2
YOLOv6	84.5	72.1	77.9	81.7	4.2	11.9
YOLOv8	84.9	74.2	79.2	82.8	3.0	8.3
YOLOv10	82.2	75.7	78.9	82.3	2.7	8.4
YOLOv11	82.8	76.0	79.3	82.9	2.6	6.4
Ours	85.0	74.9	79.6	83.8	2.6	7.2

Table 6. Comparison experiment on the UW6C datasets.

Model	P	R	F1	mAP@0.5	Parameters	GFLOPs
Faster-RCNN	/	/	/	63.0	137.1	185.1
SSD	/	/	/	68.4	26.3	31.4
RetinaNet	/	/	/	79.2	38.0	85.0
RT-DETR	85.0	73.8	79.0	80.1	15.5	37.9
YOLOv5	84.0	74.7	79.0	80.2	1.8	4.2
YOLOv6	85.9	71.8	78.2	79.3	4.2	11.9
YOLOv8	86.0	74.2	79.6	80.2	3.0	8.3
YOLOv10	83.3	72.6	77.6	78.9	2.7	8.4
YOLOv11	85.5	74.1	79.3	80.4	2.6	6.4
Ours	85.7	75.8	80.4	81.8	2.6	7.2



(**a**) Performance distribution on the DUO

(**b**) Performance distribution on the UW6C

Figure 11. Performance distributions on the DUO and UW6C datasets.

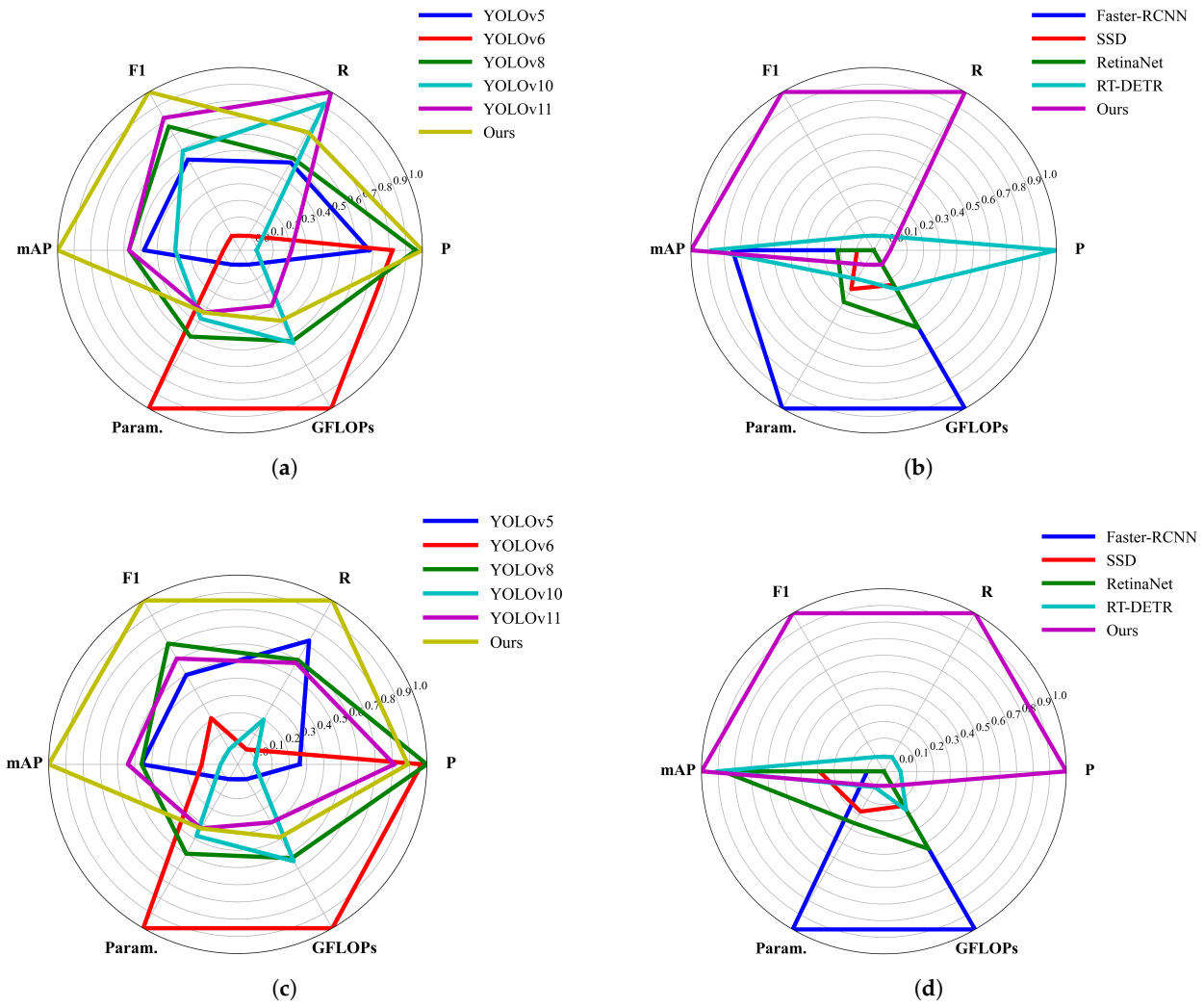


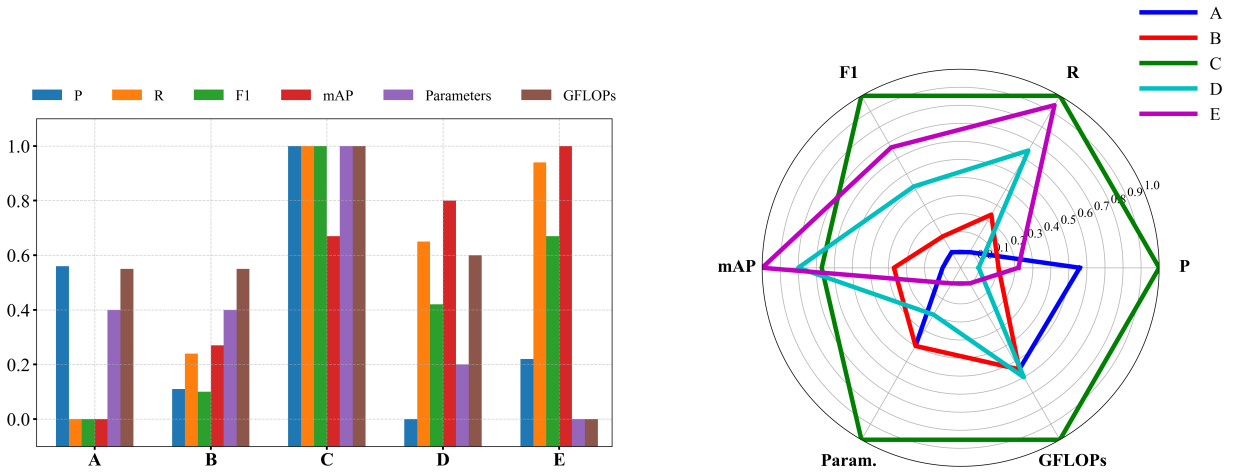
Figure 12. Metric distributions on the DUO and UW6C datasets. (**a**) Metric distribution on the DUO of YOLO Series Method. (**b**) Metric distribution on the DUO of Other Method. (**c**) Metric distribution on the UW6C of YOLO Series Method. (**d**) Metric distribution on the UW6C of Other Method.

4.4. Ablation Study

In this manuscript, the CSGPA attention mechanism is proposed as a solution to the challenges posed by water fog, target blurring, and the difficulty in distinguishing background information from targets in low-quality underwater imaging. Secondly, we constructed the EMWFPN feature-fusion network to address the problem of shallow detail information loss in deep convolutional networks, especially for small targets, due to the presence of objects at multiple scales in underwater images. In addition, we designed an ORWNet-based detection head to address the problem of object occlusion in underwater images. Furthermore, EMA Slideloss was introduced to address the problem of poor model training caused by the imbalance between hard and easy samples in the underwater object detection dataset. To verify the contribution of each component in the model and the computational consumption, they were sequentially added to the baseline for ablation experiments. The experimental results are presented in Table 7, where A denotes the baseline, B signifies the model augmented with the loss function, C represents the CSGPA integrated with it, D illustrates the implementation of EMWFPN, and F indicates the substitution of the baseline detection head with our proposed detection head. Figure 13 provides a visual depiction of the contribution of each component.

Table 7. Ablation study results.

Model	Loss	CSGPA	EMWFPN	ORWNet	P	R	F1	mAP@0.5	Parameters	GFLPOs
A	–	–	–	–	86.0	74.2	79.6	80.2	3.0	8.3
B	✓	–	–	–	85.6	74.6	79.7	80.6	3.0	8.3
C	✓	✓	–	–	86.4	75.9	80.8	81.2	3.6	9.2
D	✓	✓	✓	–	85.5	75.3	80.1	81.4	2.8	8.4
E	✓	✓	✓	✓	85.7	75.8	80.4	81.8	2.6	7.2



(**a**) Performance distribution for ablation study

(**b**) Metric distribution for ablation study

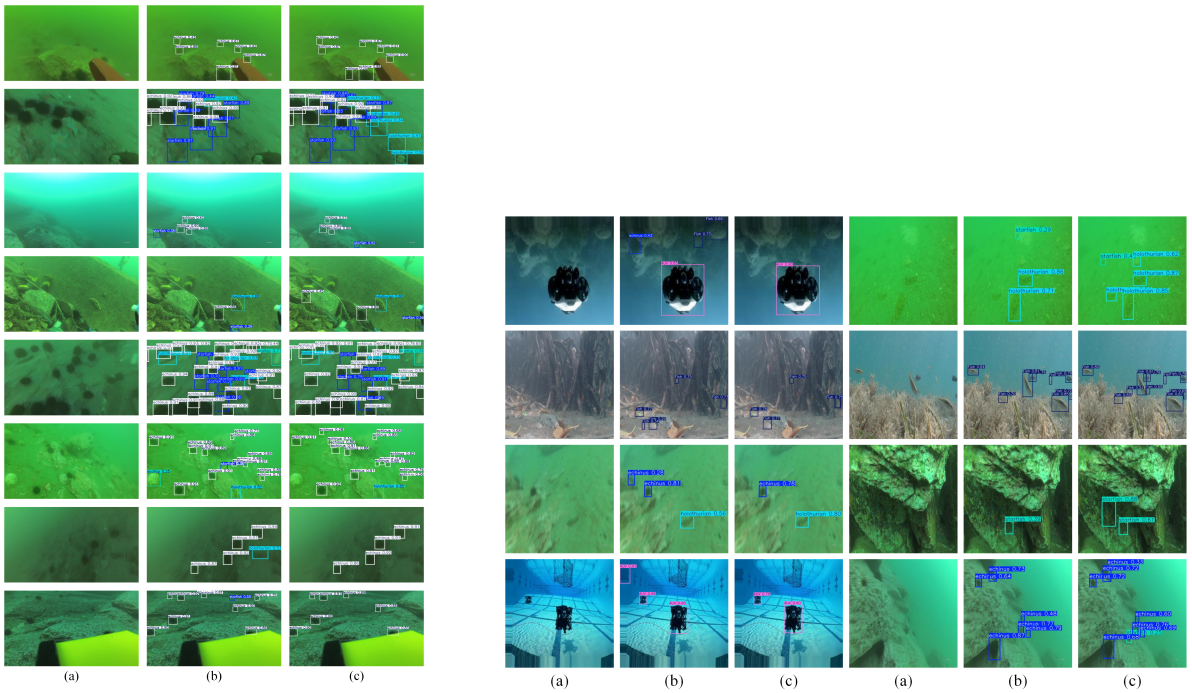
Figure 13. Visual comparison of ablation study results.

Incorporation of EMA Slideloss into the model enhances its mAP to 80.6%, indicating the efficacy of this loss function in mitigating the issue of sample imbalance. Subsequently, the CSGPA mechanism increases precision to 81.2%, underscoring the mechanism's contribution to the model. However, it was observed that the implementation of CSGPA resulted in a considerable increase in computational load. The application of EMWFPN resulted in a reduction in the computational load imposed by the model while improving the mAP, suggesting that the method effectively enhances the representation of the features of the model and reduces computational redundancy. Finally, the ORWNet-based detection head elevates the model's mAP to 81.8%, concurrently reducing the model's computational demands. Figure 13 provides a visual representation of the contribution of each component. Following the incorporation of CSGPA, each accuracy index attains a high value; however, the computational burden remains substantial. Conversely, the complete model is associated with the lowest computational demands and the second-highest accuracy index, thereby attaining an equilibrium between accuracy and computational burden.

4.5. Visualization Results

In this section, we present a range of visualization results to further validate the efficacy of the proposed model. Figure 14 illustrates the visualization results of the proposed model and the baseline model, respectively, for prediction of a variety of scenes in the DUO dataset and the UW6C dataset. These include underwater degraded images with features such as blurring, color bias, and low contrast, as well as complex images, such as multiscale objects and occluded objects. As illustrated in the figures, (a) represents the original image; (b) is the image predicted using the baseline model; and (c) is the result of the prediction using the proposed model. It is evident that the baseline model is more missed and false detections of objects in complex underwater scenes. The proposed model also exhibits

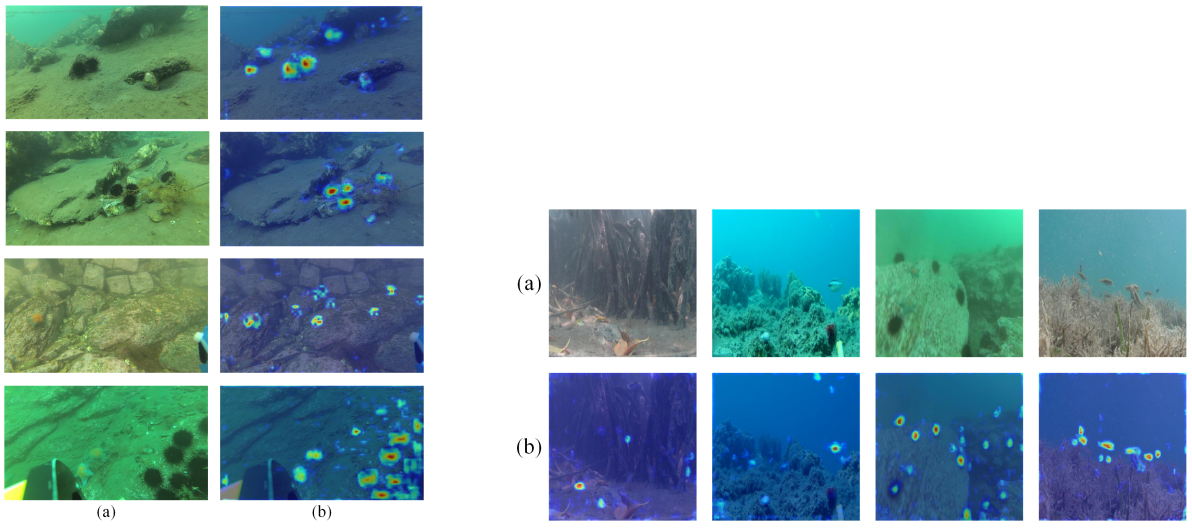
a small number of these errors, but the rates are significantly reduced compared to the baseline model.



(**A**) DUO Dateset (**B**) UW6C Dataset

Figure 14. Predicted results: (a) represents the original image; (b) is the Baseline; (c) is the result of our model.

Furthermore, Figure 15 is the heatmap, which is derived from the feature maps during the model-prediction process. Utilizing Grad-CAM [45], the heatmap visualizes the amount of the model's attention allocated to the target region. In (a), the original image is presented, while (b) depicts the heatmap. It is apparent that the model accurately captures the object of interest and focuses more intently on the region of interest. The combination of the prediction results of the image and the visualization results of the heatmap demonstrates the superior detection accuracy as well as the generalization performance of the proposed model in complex underwater environments.



(**A**) DUO Dateset (**B**) UW6C Dataset

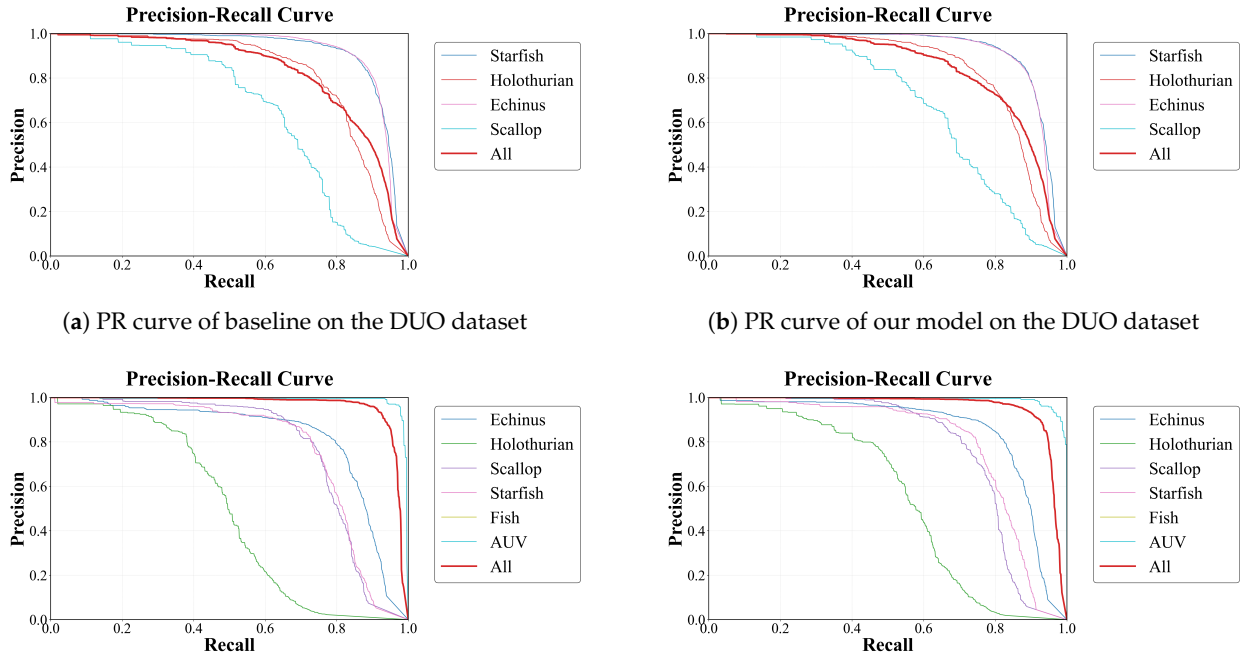
Figure 15. Heatmap results: (a) is the original image; (b) is the heatmap.

5. Discussion

In this section, an in-depth analysis of our model and its components is provided. The initial phase of the study involves a comparison of the overall performance with the baseline using PR curves, an IoU scattergraph, and confidence distributions. This is followed by a discussion of the contribution of each module via the presentation of visual evidence.

5.1. Discussion for Overall Model Performance

To evaluate the comprehensive performance of our model, we present the Precision–Recall (PR) curves in Figure 16, the IoU scattergraph in Figure 17, and the detection-confidence histogram in Figure 18.



(**a**) PR curve of baseline on the DUO dataset

(**b**) PR curve of our model on the DUO dataset

(**c**) PR curve of baseline on the UW6C Dataset

(**d**) PR Curve of our model on the UW6C Dataset

Figure 16. PR Curves of baseline and our model.

First, as shown in Figure 16, our model consistently outperforms the baseline on both datasets. On the DUO dataset, the PR curves for our model are larger than the region enclosed by the baseline curve, indicating better overall mAP. This improvement is particularly evident in classed such as starfish and scallops, which are often challenging due to low contrast or irregular textures in underwater environments. On the UW6C dataset, our model also has better results based on the PR curves, especially on small or dense objects such as fish and echinus, which perform better compared to the baseline model. These observations highlight the fact that our model maintains a balance of recall and precision across a wide range of underwater object classes.

Second, We present the IoU Scattergraphs in Figure 17. Compared to the baseline, our model's predictions are more densely clustered in the upper-right region, suggesting that high-confidence predictions also exhibit high IoU values—demonstrating both reliability and accuracy. Specifically, the baseline model shows a large range of low IoU predictions even at moderate confidence (0.6–0.8), while our model significantly reduces such noise. This suggests that our model is better able to combine classification confidence with true localization quality, which is crucial in practical applications.

Finally, An idea of how certain the models are about their predictions can be gleaned from the histogram of detection confidence in Figure 18. The predictions of the baseline model are biased towards low to medium confidence levels, with a flat or bimodal distribution (particularly evident in UW6C). In contrast, our model shows a sharp peak in the high-confidence region ([0.85, 1.0]), indicating more confident and decisive predictions. This is particularly important in underwater scenes where image quality is degraded due to scattering and turbidity. Our model demonstrates better calibration of prediction confidence, reflecting more reliable and robust decision boundaries in complex underwater scenes.

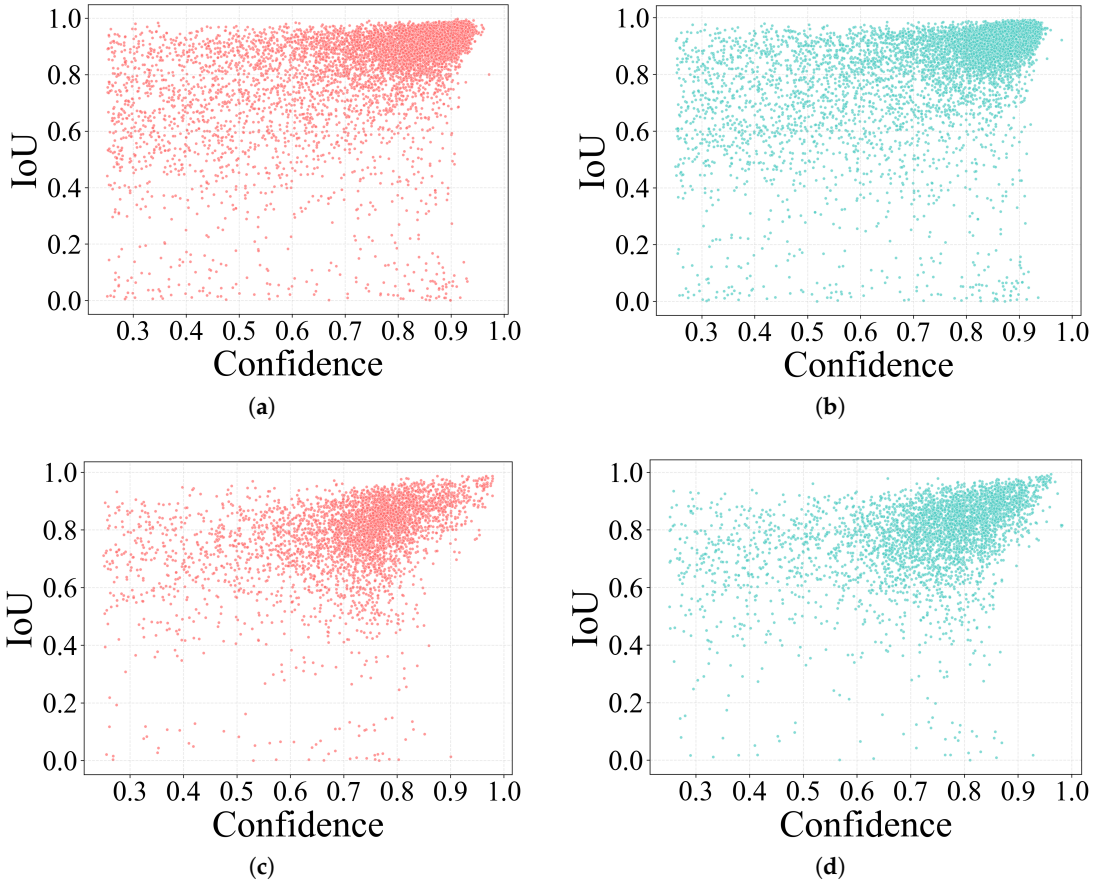


Figure 17. IoU Scattergraphs of baseline and our model. (**a**) IoU Scattergraphs of baseline on the DUO dataset. (**b**) IoU Scattergraphs of our model on the DUO dataset. (**c**) IoU Scattergraphs of baseline on the UW6C Dataset. (**d**) IoU Scattergraphs of our model on the UW6C Dataset.

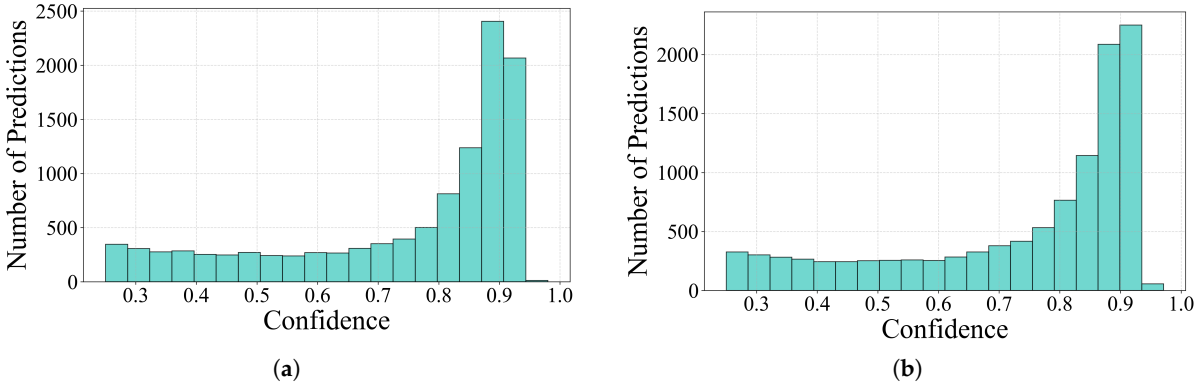


Figure 18. *Cont.*

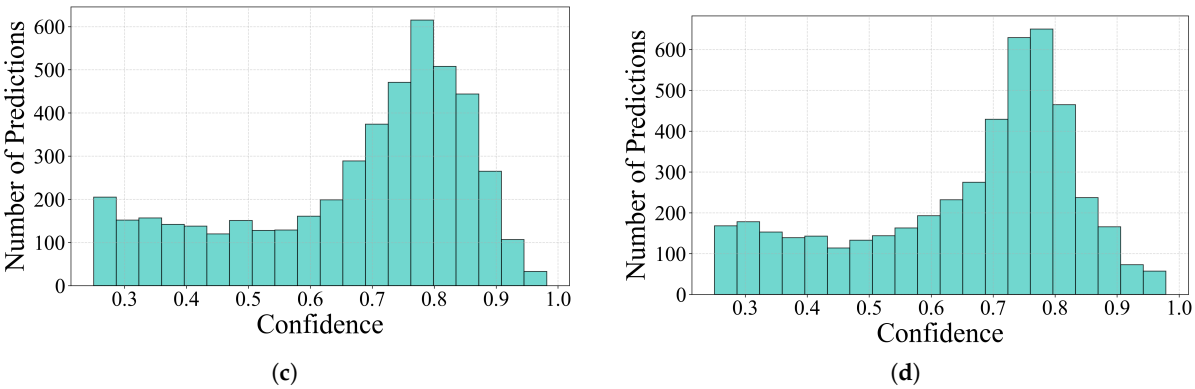


Figure 18. Confidence histogram of baseline and our model. (**a**) Confidence histogram of baseline on the DUO dataset. (**b**) Confidence histogram of our model on the DUO dataset. (**c**) Confidence histogram of baseline on the UW6C Dataset. (**d**) Confidence histogram of our model on the UW6C Dataset.

5.2. Discussion for CSGPA

The superiority of CSGPA compared to SE and CBAM can be attributed to its architectural design. This design also facilitates smoother gradient propagation, although the main performance gain stems from its ability to jointly model context and spatial structure. Whereas SE captures only channel dependence and CBAM processes channel and spatial attention sequentially, CSGPA employs a two-branch parallel architecture that combines multiheaded self-attention with efficient multiscale attention. This enables the simultaneous modeling of long-range dependencies and cross-space interactions, addressing the limitations of local receptive fields and spatial biases in traditional modules. Such joint modeling is essential in underwater environments with low contrast and target–background blurring. Experimental results in Tables 1 and 2 show that CSGPA consistently outperforms both SE and CBAM in terms of accuracy and mAP. In addition, an ablation study, the results of which are given in Table 7, confirms the impact of CSGPA. Although this approach leads to a higher computational cost, the improved detection capability justifies the computational overhead. This improvement stems from better contextual reasoning and spatial modeling.

5.3. Discussion for EMWFPN

EMWFPN is designed to improve the accuracy of localization of small and clustered objects while maintaining computational efficiency. In a departure from BiFPN and GFPN, EMWFPN integrates shallow high-resolution features and employs a dynamic weight-fusion strategy, which enables the model to adaptively adjust the importance of features at different scales. As shown in Tables 3 and 4, EMWFPN achieves the highest mAP on both datasets while using significantly fewer parameters than GFPN. However, its F1 score is slightly lower, which we attribute to a design trade-off prioritizing precise localization over classification. This trade-off is evident in Figure 19, where the feature maps processed by EMWFPN outperform the baseline model in terms of spatial sharpness, edge clarity, and activation intensity around object contours. In contrast, the baseline model exhibits blurrier and more redundant activations, indicating weaker spatial focus capabilities. By reducing feature redundancy and enhancing spatial-discrimination capabilities, EMWFPN improves localization accuracy at the cost of slightly sacrificing classification performance. This helps explain why mAP improves while F1 decreases slightly. We accept this trade-off because we believe that prioritizing localization accuracy is more important in noisy underwater environments.

5.4. Discussion for ORWNet Based Head

To address the occlusion problem in underwater scenes, we propose ORWNet, a frequency-aware detection head based on wavelet decomposition of feature maps. This method efficiently separates shape information and detail information, enabling the model to better capture contextual clues and reduce interference from overlapping targets. As shown in Table 7, use of the ORWNet-based detection head increases the mAP@0.5 from 81.4% to 81.8% while reducing the number of parameters and GFLOPs. This indicates that ORWNet improves accuracy while reducing computational cost. By operating in the frequency domain, ORWNet expands the receptive field, enhancing the model's ability to identify occluded objects. This improves detection stability in complex underwater environments.

5.5. Discussion for EMASlideloss

In underwater object detection, sample imbalance causes the model to overfit simple examples while underfitting challenging ones. EMA Slideloss addresses this issue by dynamically adjusting weights rather than using a fixed IoU threshold. Through the use of an exponential moving average, the definition of 'difficult' continuously adapts during training, ensuring the model focuses on cases that are misclassified but learnable at each stage. This targeted weighting mechanism enhances the model's ability to learn discriminative features, improves generalization performance, and avoids interference from noisy or misleading samples. As a result, the model achieves superior detection performance across complex underwater object categories.

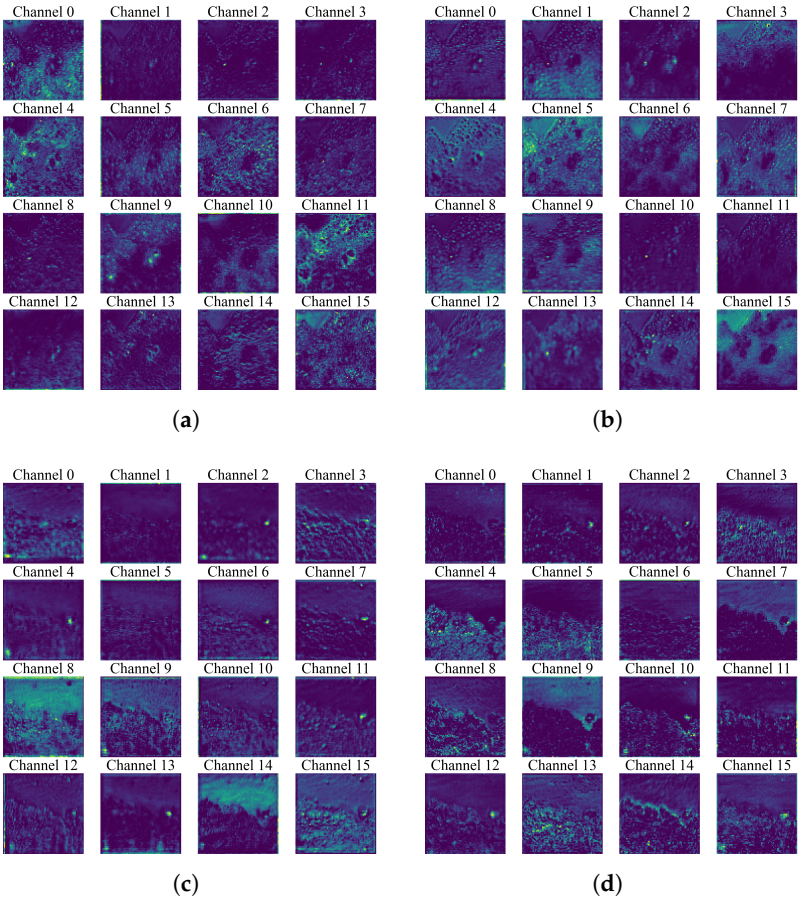


Figure 19. EMWFPN's feature map visualization of the baseline and our model. (**a**) Feature map visualization of the baseline on the DUO dataset. (**b**) Feature map visualization of our model on the DUO dataset. (**c**) Feature map visualization of the baseline on the UW6C Dataset. (**d**) Feature map visualization of our model on the UW6C Dataset.

Overall, the proposed improvements—including the CSGPA attention module, EMWFPN feature fusion, ORWNet detection head, and EMA sliding loss—collectively address key challenges in underwater target detection, such as occlusion, small targets, low contrast, and sample imbalance. Each component contributes to specific aspects of performance, and their integration forms a coherent detection pipeline. This synergistic effect enables more accurate, stable, and generalizable predictions in underwater scenarios.

6. Conclusions

This paper proposes a lightweight and efficient framework for underwater object detection that effectively addresses the object-recognition challenges caused by low-quality optical images when AUVs perform tasks underwater. Firstly, we propose a novel attention mechanism that enables the model to focus on both long-range dependencies and local features so that they are synergistic and fused under a unified framework. Second, an efficient multiscale feature-fusion network is constructed by incorporating more detailed information into the fusion process and adding more cross-scale connections. We also use weighted fusion to improve model efficiency and introduce lightweight feature extraction to reduce computational redundancy. Finally, we design a lightweight detection head with a large receptive field that processes features from the frequency domain perspective and effectively identifies occluded objects. Numerous experiments show that our model achieves better performance and excellent generalization ability on both the DUO and UW6C datasets compared with the baseline model and other excellent models, making it possible to achieve more efficient underwater exploration by AUVs.

Author Contributions: S.S., Conceptualization, Methodology, Writing—Original Draft, Software. B.S., Writing—Review & Editing, Supervision, Resources. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China under Grant 52271321 and Natural Science Foundation of Shanghai under Grant 22ZR1426700.

Data Availability Statement: The original contributions presented in the study are included in the article material. Further inquiries can be directed to the corresponding author.

Conflicts 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.

References

1. Vedachalam, N.; Ramesh, R.; Jyothi, V.B.N.; Doss Prakash, V.; Ramadass, G.A. Autonomous underwater vehicles-challenging developments and technological maturity towards strategic swarm robotics systems. *Mar. Georesour. Geotechnol.* **2019**, *37*, 525–538. [CrossRef]
2. Li, Q.; Shi, H. Yolo-ge: An attention fusion enhanced underwater object detection algorithm. *J. Mar. Sci. Eng.* **2024**, *12*, 1885. [CrossRef]
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *39*, 1137–1149. [CrossRef]
4. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In Proceedings of the Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
5. Lin, T.Y.; Goyal, R.; Girshick, R.; He, K.; Dollar, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 2980–2988.
6. 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), Las Vegas, Nevada, 26 June–1 July 2016; pp. 779–788.
7. Gao, J.; Zhang, Y.; Geng, X.; Tang, H.; Bhatti, U.A. PE-Transformer: Path enhanced transformer for improving underwater object detection. *Expert Syst. Appl.* **2024**, *246*, 123253. [CrossRef]
8. Yuan, M.; Shi, X.; Wang, N.; Wang, Y.; Wei, X. Improving RGB-infrared object detection with cascade alignment-guided transformer. *Inf. Fusion* **2024**, *105*, 102246. [CrossRef]

9. Zhao, J.; Ding, Z.; Zhou, Y.; Zhu, H.; Du, W.; Yao, R.; EI Saddik, A. OrientedFormer: An End-to-End Transformer-Based Oriented Object Detector in Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2024**, *62*, 5640816. [CrossRef]
10. Tang, Y.; Zhao, Z.; Huang, O. Mamba enhanced you only look once network with multiscale spatial attention for remote sensing object detection. *Eng. Appl. Artif. Intell.* **2025**, *159*, 111732. [CrossRef]
11. Wang, Z.; Li, L.; Zeng, C.; Dong, S.; Sun, J. SLB-Mamba: A vision Mamba for closed and open-set student learning behavior detection. *Appl. Soft Comput.* **2025**, *180*, 113369. [CrossRef]
12. Xing, G.; Wang, M.; Wang, F.; Sun, F.; Li, H. Lightweight Edge-Aware Mamba-Fusion Network for Weakly Supervised Salient Object Detection in Optical Remote Sensing Images. *IEEE Trans. Geosci. Remote Sens.* **2025**, *63*, 5631813. [CrossRef]
13. Liang, X.; Song, P. Excavating roi attention for underwater object detection. In Proceedings of the 2022 IEEE International Conference on Image Processing (ICIP), Bordeaux, France, 16–19 October 2022; pp. 2651–2655.
14. Song, P.; Li, P.; Dai, L.; Wang, T.; Chen, Z. Boosting R-CNN: Reweighting R-CNN samples by RPN's error for underwater object detection. *Neurocomputing* **2023**, *530*, 150–164. [CrossRef]
15. Dai, L.; Liu, H.; Song, P.; Tang, H.; Ding, R.; Li, S. Edge-guided representation learning for underwater object detection. *CAAI Trans. Intell. Technol.* **2024**, *9*, 1078–1091. [CrossRef]
16. Zhou, J.; He, Z.; Zhang, D.; Liu, S.; Fu, X.; Li, X. Spatial Residual for Underwater Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 4996–5013. [CrossRef]
17. Chen, H.; Wang, Z.; Qin, H.; Mu, X. Self-supervised domain feature mining for underwater domain generalization object detection. *Expert Syst. Appl.* **2025**, *265*, 126023. [CrossRef]
18. Ouyang, D.; He, S.; Zhang, G.; Luo, M.; Guo, H.; Zhan, J. Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In Proceedings of the ICASSP 2023–2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes, Greece, 4–10 June 2023; pp. 1–5.
19. 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), Virtual, 11–17 October 2021; pp. 783–792.
20. 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), Virtual, 11–17 October 2021; pp. 10012–10022.
21. Pan, X.; Ge, C.; Lu, R.; Song, S.; Chen, G.; Huang, Z.; Huang, G. On the Integration of Self-Attention and Convolution. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 21–24 June 2022; pp. 815–825.
22. Shen, X.; Yuan, G.; Wang, H.; Fu, X. Unsupervised clustering optimization-based efficient attention in YOLO for underwater object detection. *Artif. Intell. Rev.* **2025**, *58*, 219. [CrossRef]
23. Lin T.Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
24. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 10781–10790.
25. Jiang, Y.; Tan, Z.; Wang, J.; Sun, X.; Lin, M.; Li, H. GiraffeDet: A heavy-neck paradigm for object detection. *arXiv* **2022**, arXiv:2202.04256.
26. Ma, Q.; Li, X.; Li, B.; Zhu, Z.; Wu, J.; Huang, F.; Hu, H. STAMF: Synergistic transformer and mamba fusion network for RGB-Polarization based underwater salient object detection. *Inf. Fusion* **2025**, *122*, 103182. [CrossRef]
27. Li, S.; Wang, Z.; Dai, R.; Wang, Y.; Zhong, F.; Liu, Y. Efficient Underwater Object Detection with Enhanced Feature Extraction and Fusion. *IEEE Trans. Ind. Inform.* **2025**, *21*, 4904–4914. [CrossRef]
28. Salehi, A.; Balasubramanian, M.; Li, H. DDCNet: Deep dilated convolutional neural network for dense prediction. *Neurocomputing* **2023**, *523*, 116–129. [CrossRef] [PubMed]
29. Ding, X.; Zhang, X.; Han, J.; Ding, J. Scaling up Your Kernels to 31x31: Revisiting Large Kernel Design in CNNs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–23 June 2022; pp. 11963–11975.
30. Shen, J.; Chen, Y.; Liu, Y.; Zuo, X.; Fan, H.; Yang, W. ICAFusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognit.* **2024**, *145*, 109913. [CrossRef]
31. Jiang, L.; Yuan, B.; Wang, Y.; Ma, Y.; Du, J.; Wang, F. Guo, J. MA-YOLO: A Method for Detecting Surface Defects of Aluminum Profiles with Attention Guidance. *IEEE Access* **2023**, *11*, 71269–71286. [CrossRef]
32. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A New Backbone That Can Enhance Learning Capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Virtual, 14–19 June 2020; pp. 390–391.
33. Xu, X.; Jiang, Y.; Chen, W.; Huang, Y.; Zhang, Y.; Sun, X. Damo-yolo: A report on real-time object detection design. *arXiv* **2022**, arXiv:2211.15444.

34. Yu, Z.; Huang, H.; Chen, W.; Su, Y.; Liu, Y.; Wang, X. Yolo-facev2: A scale and occlusion aware face detector. *Pattern Recognit.* **2024**, *155*, 110714. [CrossRef]
35. Finder, S.E.; Amoyal, R.; Treister, E.; Freifeld, O. Wavelet Convolutions for Large Receptive Fields. In Proceedings of the European Conference on Computer Vision, Milan, Italy, 29 September–4 October 2024; pp. 363–380.
36. Cai, Z.; Ravichandran, A.; Maji, S.; Fowlkes, C.; Tu, Z.; Soatto, S. Exponential moving average normalization for self-supervised and semi-supervised learning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR), Virtual, 21–25 June 2021; pp. 194–203.
37. Liu, C.; Li, H.; Wang, S.; Zhu, M.; Wang, D.; Fan, X.; Wang, Z. A Dataset and Benchmark of Underwater Object Detection for Robot Picking. In Proceedings of the 2021 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), Shenzhen, China, 5–9 July 2021; pp. 1–6.
38. Saleh, A.; Laradji, I.H.; Konovalov, D.A.; Bradley, M.; Vazquez, D.; Sheaves, M. A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. *Sci. Rep.* **2020**, *10*, 14671. [CrossRef] [PubMed]
39. Zhao, Y.; Lv, W.; Xu, S.; Wei, J.; Wang, G.; Dang, Q.; Liu, Y.; Chen, J. DETRs Beat YOLOs on Real-time Object Detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle WA, USA, 17–21 June 2024; pp. 16965–16974.
40. Ultralytics YOLOv5. Available online: https://github.com/ultralytics/yolov5 (accessed on 19 September 2024).
41. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Wei, X. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976. [CrossRef]
42. Ultralytics YOLOv8. Available online: https://github.com/ultralytics/ultralytics (accessed on 8 July 2025).
43. Wang, A.; Chen, H.; Liu, L.; Chen, K.; Lin, Z.; Han, J.; Ding, G. YOLOv10: Real-Time End-to-End Object Detection. In Proceedings of the Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Vancouver, BC, Canada, 10–15 December 2024; pp. 107984–108011.
44. Ultralytics YOLOv11. Available online: https://github.com/yt7589/yolov11 (accessed on 28 May 2025).
45. 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), Venice, Italy, 22–29 October 2017; pp. 618–626.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.