applied system innovation

MDPI

Article

FPH-DEIM: A Lightweight Underwater Biological Object Detection Algorithm Based on Improved DEIM

Qiang Li and Wenguang Song *

School of Computer Science and Engineering, Guangdong Ocean University, Yangjiang 529500, China; li.qiang@stu.gdou.edu.cn

* Correspondence: songwg@gdou.edu.cn

Abstract

Underwater biological object detection plays a critical role in intelligent ocean monitoring and underwater robotic perception systems. However, challenges such as image blurring, complex lighting conditions, and significant variations in object scale severely limit the performance of mainstream detection algorithms like the YOLO series and Transformer-based models. Although these methods offer real-time inference, they often suffer from unstable accuracy, slow convergence, and insufficient small object detection in underwater environments. To address these challenges, we propose FPH-DEIM, a lightweight underwater object detection algorithm based on an improved DEIM framework. It integrates three tailored modules for perception enhancement and efficiency optimization: a Fine-grained Channel Attention (FCA) mechanism that dynamically balances global and local channel responses to suppress background noise and enhance target features; a Partial Convolution (PConv) operator that reduces redundant computation while maintaining semantic fidelity; and a Haar Wavelet Downsampling (HWDown) module that preserves high-frequency spatial information critical for detecting small underwater organisms. Extensive experiments on the URPC 2021 dataset show that FPH-DEIM achieves a mAP@0.5 of 89.4%, outperforming DEIM (86.2%), YOLOv5-n (86.1%), YOLOv8-n (86.2%), and YOLOv10-n (84.6%) by 3.2–4.8 percentage points. Furthermore, FPH-DEIM significantly reduces the number of model parameters to 7.2 M and the computational complexity to 7.1 GFLOPs, offering reductions of over 13% in parameters and 5% in FLOPs compared to DEIM, and outperforming YOLO models by margins exceeding 2 M parameters and 14.5 GFLOPs in some cases. These results demonstrate that FPH-DEIM achieves an excellent balance between detection accuracy and lightweight deployment, making it well-suited for practical use in real-world underwater environments.

Keywords: underwater biology; object detection; deep learning; attention mechanism; lightweight architecture

check for updates

Received: 25 July 2025
Revised: 21 August 2025
Accepted: 25 August 2025
Published: 26 August 2025

Citation: Li, Q.; Song, W. FPH-DEIM: A Lightweight Underwater Biological Object Detection Algorithm Based on Improved DEIM. *Appl. Syst. Innov.* **2025**, *8*, 123. https://doi.org/10.3390/asi8050123

Copyright: © 2025 by the authors. Published by MDPI on behalf of the International Institute of Knowledge Innovation and Invention. 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

With the continuous advancement of ocean exploration and exploitation technologies, underwater object detection has played an increasingly vital role in intelligent ocean observation, autonomous underwater vehicle (AUV) navigation, and marine ecological monitoring. In particular, underwater image analysis is becoming a key foundation for intelligent decision-making in ecological protection and biological resource surveys [1]. Compared with terrestrial environments, underwater image acquisition is often subject to severe environmental interference, such as uneven illumination, light refraction, suspended

particulate occlusion, color distortion, and blurring. These factors result in underwater images characterized by low contrast, poor clarity, and high noise, posing significant challenges to the feature extraction, object perception, and recognition capabilities of deep learning models [2,3].

In recent years, deep learning approaches based on convolutional neural networks (CNNs) and vision Transformers have achieved remarkable progress in object detection tasks. Beyond architecture design, self-supervised learning [4] and multimodal fusion [5] have emerged as complementary paradigms to address data scarcity and environmental degradation in underwater scenarios. Among them, the YOLO series has been widely adopted in real-world applications due to its end-to-end architecture, high speed, and compact structure [6]. However, YOLO models typically rely on a one-to-many (O2M) strategy, where each object corresponds to multiple anchors or candidate boxes. Although this approach increases the number of positive samples during training and enhances supervision density, it also tends to cause redundant proposals, overlapping bounding boxes, and unstable Non-Maximum Suppression (NMS), particularly under scenarios involving densely distributed small underwater objects [7,8].

To overcome the limitations of traditional anchor-based methods, Carion et al. proposed the DETR (Detection Transformer) model, which introduced the Transformer mechanism into object detection and adopted a one-to-one (O2O) Hungarian matching strategy. This approach enables end-to-end detection while eliminating the need for NMS and anchor design [9]. However, standard DETR suffers from slow convergence, sparse positive samples, and poor utilization of low-quality matches—issues that significantly impair its effectiveness in detecting small or weak-edged objects [10].

To address these problems, the DEIM (DETR with Improved Matching) framework was proposed. The framework incorporates a dense one-to-one matching scheme together with a Matchability-Aware Loss (MAL), which jointly increase the availability of positive samples and refine the matching process, leading to accelerated convergence and improved detection accuracy [10]. Specifically, Dense O2O leverages data augmentation techniques such as Mosaic and MixUp to create multiple target regions in one image, this process augments the quantity of effective training samples. Meanwhile, MAL assigns differentiated weights to low-quality matches to encourage the network to learn from a wider range of supervision signals.

Although DEIM outperforms YOLOv10 and RT-DETR series in general scenarios in terms of both accuracy and efficiency, its architecture remains relatively complex and less suitable for resource-constrained platforms such as underwater robots, which demand compact models and low-latency inference [11]. Furthermore, DEIM still has room for improvement in feature representation, feature compression, and small object retention. To meet the specific requirements of underwater biological object detection under constraints such as low light, high noise, small object density, and limited computing resources, in this study, we design a lightweight detection framework enhanced from DEIM: FPH-DEIM.

The main contributions of this study are as follows:

1. Introduction of the Fine-grained Channel Attention (FCA) mechanism
   This module models the interaction between global average features and local channel context to dynamically generate channel-wise weighting vectors. Compared with conventional SE modules, FCA better adapts to underwater images where local noise interference and low contrast are prominent. It effectively suppresses redundant background information, enhances salient biological features, and improves robustness in complex environments [12].
2. Adoption of the efficient Partial Convolution (PConv) operator to replace standard convolution modules

PConv selectively activates parts of the input feature map for convolution, significantly reducing memory access and computational redundancy without sacrificing recognition accuracy. This makes it well-suited for deployment on embedded underwater platforms and mobile AUVs with limited computational capacity [13].

3. Integration of the Haar Wavelet Downsampling (HWDown) module to enhance small object preservation
   Traditional downsampling methods, such as max-pooling or strided convolution, often discard critical information while reducing spatial resolution, especially when object sizes approach the lower bound of the receptive field. The HWDown module employs Haar wavelet transforms for information compression, retaining the main energy distribution in the frequency domain. This enhances the model's semantic representation and discriminative power for small targets such as jellyfish, fish larvae, and micro-crustaceans [14].

To validate the effectiveness of the proposed methods, extensive experiments were conducted on the public URPC 2021 underwater image dataset. Results demonstrate that FPH-DEIM achieves a 3.2% improvement in mAP@0.5 while reducing the number of parameters and FLOPs by approximately 10% and 5%, respectively, compared to the original DEIM. Deployment tests on real underwater robotic platforms further confirm its superior real-time performance and robustness, significantly outperforming baseline models such as DEIM and YOLOv10.

2. Related Work

2.1. Overview of Underwater Object Detection Methods

With the rapid integration of artificial intelligence into the marine domain, underwater object detection—serving as a core component of intelligent perception systems—has become a prominent research focus in algorithm design and performance optimization. Traditional underwater image analysis methods mainly rely on handcrafted features based on color, edges, or texture. Representative methods include Canny edge detection and HOG feature matching [15–17]. However, in underwater environments with uneven lighting and complex backgrounds, these approaches are often highly sensitive to noise and fail to meet accuracy and robustness requirements.

The advent of deep neural networks has significantly boosted the capabilities of detecting objects in underwater environments. In particular, convolutional neural networks (CNNs), empowered by large-scale data, demonstrate strong feature extraction capabilities. Architectures such as Faster R-CNN, YOLO, and SSD have been widely adapted to underwater tasks [17,18]. Among t hem, the YOLO series has gained popularity in industry for its end-to-end architecture, high inference speed, and competitive accuracy. Nevertheless, YOLO faces two major challenges in underwater environments:

- The standard YOLO uses a one-to-many (O2M) matching strategy, leading to high prediction box density and difficulty in learning stable representations for low-contrast, occluded, or small targets;
- Its feature extraction backbone is not specifically designed for the unique characteristics of underwater imagery, resulting in limited ability to perceive fine-grained semantic information.

To improve YOLO's adaptability in underwater scenes, various modifications have been proposed. For instance, Yu et al. introduced U-YOLO, which integrates multi-scale fusion modules to enhance small object detection [15]; Zhang et al. embedded channel attention mechanisms into YOLOv3, significantly improving the handling of low-quality images [19]. Although these methods partially alleviate accuracy degradation

in underwater scenarios, their increasingly complex structures hinder deployment on embedded platforms and do not fundamentally overcome the performance bottlenecks caused by the O2M matching strategy.

2.2. Transformer-Based Object Detection

In recent years, Transformer architectures have emerged as a new paradigm in object detection. Since Carion et al. introduced DETR [7], which leverages self-attention mechanisms to model global contextual relationships, the field has seen a shift away from anchor boxes and heuristic matching strategies, toward a new end-to-end detection framework. DETR employs Hungarian matching to implement one-to-one correspondence between queries and objects, eliminating redundant bounding boxes and post-processing via NMS, thereby simplifying the detection pipeline.

However, the original DETR has notable drawbacks: it suffers from slow convergence (typically requiring over 500 training epochs), low sensitivity to small objects, and poor matching quality in early training stages. To tackle these problems, researchers have introduced several enhancements focusing on convergence speed and detection accuracy. In Deformable DETR [8], multi-scale deformable attention is employed. Sparse and localized perception is thereby improved. to improve sparse and localized perception; Conditional DETR [20] optimizes the decoder by introducing conditional queries to enhance target representation; DN-DETR and DINO further boost matching efficiency and accuracy through noise-perturbed anchor learning and multiple positive sample augmentation strategies [21,22].

Despite their advantages in global modeling, Transformer-based methods are typically associated with high computational complexity and resource consumption, making them less suitable for resource-constrained platforms such as underwater robots. Consequently, optimizing DETR architectures for efficiency and lightweight deployment without compromising accuracy has become a key research trend.

2.3. The DEIM Framework

DEIM (DETR with Improved Matching) is a recently proposed and efficient enhancement of the DETR framework. While retaining the one-to-one matching strategy, DEIM significantly improves training efficiency and performance, especially in small object detection scenarios [10]. Its key contributions include the following:

- Dense One-to-One Matching: DEIM increases the density of detectable objects within a single image via data augmentation techniques such as Mosaic and MixUp, thereby improving positive sample utilization without altering the core DETR structure.
- Matchability-Aware Loss (MAL): This novel loss function uses confidence scores to guide IoU-based supervision, allowing the model to optimize for low-quality matched samples and improve overall detection robustness.

On datasets such as COCO and CrowdHuman, DEIM has shown superior performance over detectors like RT-DETR and YOLOv7, particularly excelling in small object detection (AP_S). Moreover, DEIM maintains a relatively simple architecture, making it adaptable to further modification.

However, DEIM still adopts a relatively heavy backbone and lacks specialized design for underwater imagery. Its downsampling pathway may also cause spatial information loss, limiting its performance in detecting fine-scale biological targets underwater. These limitations highlight the need for task-specific structural improvements focusing on lightweight design and enhanced perception.

2.4. Model Lightweighting and Perception Enhancement Techniques

Current strategies for lightweighting deep detection networks primarily focus on three aspects: efficient convolutional operator design, integration of attention mechanisms, and information-preserving downsampling techniques.

On one hand, to improve inference efficiency, researchers have proposed various lightweight convolutional structures, such as Depthwise Separable Convolutions, Ghost Modules, and Partial Convolution (PConv). As a core module in the FasterNet architecture, PConv restricts convolution operations to a subset of input channels, significantly reducing memory access and redundant computation without compromising accuracy [13]. Benchmark tests show that PConv achieves lower inference latency than MobileViT across CPU/GPU/ARM platforms, demonstrating its suitability for real-world deployment.

On the other hand, attention mechanisms have been widely employed to boost model sensitivity to object features. SE [23] and CBAM [24], for example, apply attention along channel and spatial dimensions, respectively, achieving notable performance gains. The Fine-Grained Channel Attention (FCA) mechanism further enhances this by modeling both global and local channel interactions. FCA employs a stripe-and-diagonal matrix fusion strategy to adaptively assign channel weights, effectively suppressing redundant noise common in underwater imagery [12]. This mechanism has shown strong generalization in dehazing and underwater enhancement tasks and is well-suited for object detection in challenging environments.

Moreover, maintaining expressive features while compressing spatial resolution remains a challenge in small object detection. Traditional max pooling and strided convolutions reduce computational costs at the expense of spatial details. To tackle this, Haar Wavelet Downsampling (HWDown) has been proposed, introducing wavelet basis transformation into the downsampling process. HWDown retains edge and texture information in the frequency domain while reducing dimensionality [14], showing low entropy loss and high feature fidelity in tasks like semantic segmentation and structural recognition, making it a promising approach for underwater detection.

In summary, the proposed FPH-DEIM integrates all three categories of techniques—lightweight convolution (PConv), channel-wise attention (FCA), and frequency-preserving downsampling (HWDown)—to optimize DEIM along two dimensions: lightweight design and perceptual enhancement. This allows the model to more effectively handle the complexities of underwater settings and meet the deployment requirements of embedded platforms.

3. Materials and Methods

3.1. FPH-DEIM: An Improved DEIM-Based Lightweight Approach for Underwater Object Detection

The proposed FPH-DEIM algorithm is built upon the DEIM (DETR with Improved Matching) architecture and systematically tailored for underwater object detection tasks. It forms a lightweight deep neural network with enhanced perceptual capability and computational efficiency. Figure 1 shows the complete network architecture, which comprises five main components:

- Image input and preprocessing.
- Lightweight backbone for feature extraction, integrating FCA and PConv.
- Multi-scale feature downsampling and fusion, including HWDown.
- Encoder–decoder with one-to-one target matching, keeping DEIM's MAL.
- Output module for bounding box regression and object classification.

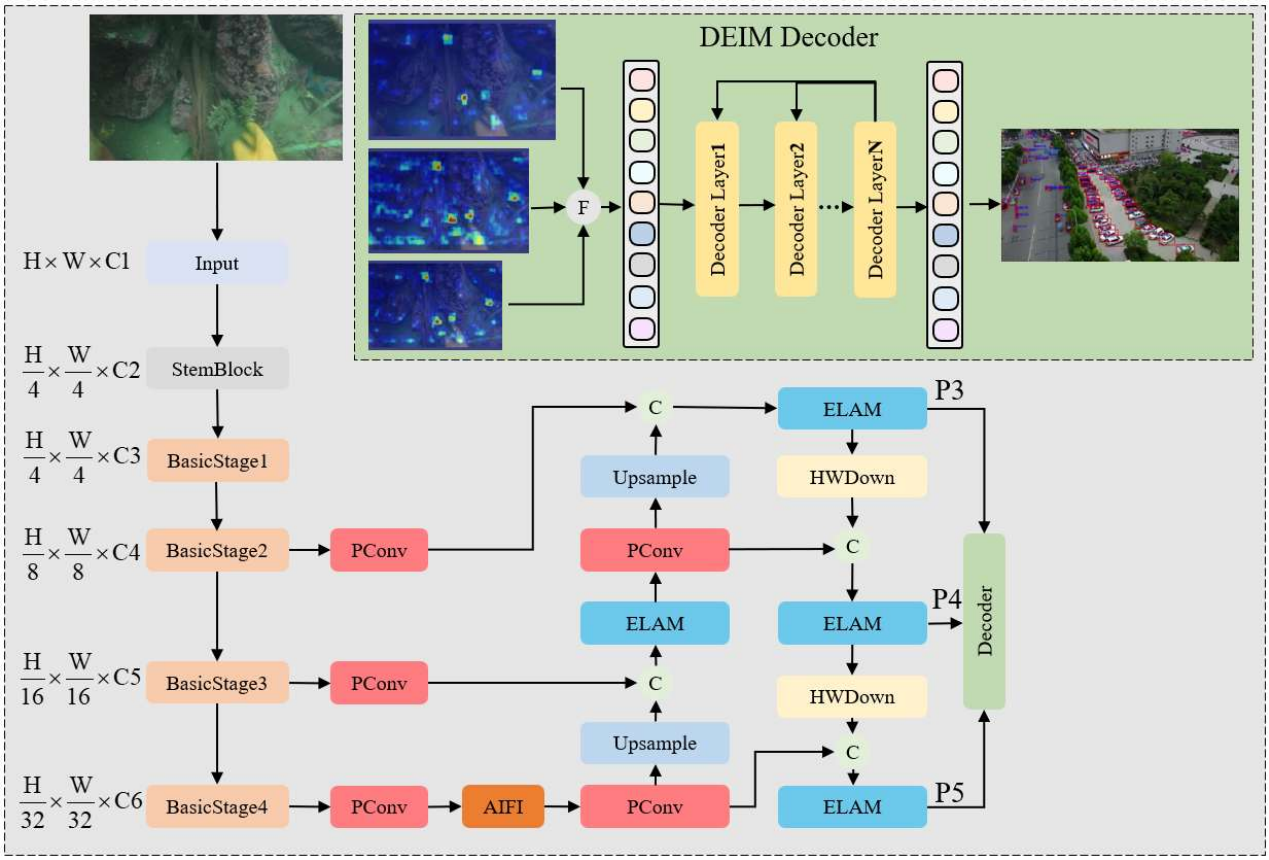


Figure 1. Architecture of the FPH-DEIM Network.

Building on DEIM's one-to-one matching and Matchability-Aware Loss, this network incorporates key improvements to enhance perception and efficiency. The next subsections describe the design and integration of three main modules.

Compared with the original DEIM and other multi-scale fusion methods, FPH-DEIM introduces three key architectural innovations. First, the Feature Channel Attention (FCA) module is designed to preserve fine-grained inter-channel variations, enabling the network to capture subtle feature differences that are often ignored in conventional DEIM. Second, the Partial Convolution (PConv) selectively updates feature maps by focusing on informative regions while skipping irrelevant areas, which differs fundamentally from simply reducing the number of convolutional filters. This allows the model to maintain strong local feature extraction ability with lower computational cost. Third, the Haar Wavelet Downsampling (HWDown) module replaces stride convolution or pooling to retain structural details of small objects, thereby improving detection robustness in underwater environments. Together, these innovations distinguish FPH-DEIM from prior architectures by enhancing both efficiency and accuracy.

3.2. Fine-Grained Channel Attention Module

Traditional CNN feature extractors depend on local convolutions, limiting their ability to capture long-range semantics, especially in underwater images with blurred edges and variable colors. The FCA (Fine-grained Channel Attention) mechanism improves feature discrimination by adaptively weighting channels using global–local information interactions [12], making it well-suited for noisy underwater environments. The overall structure of FCA is illustrated in Figure 2.

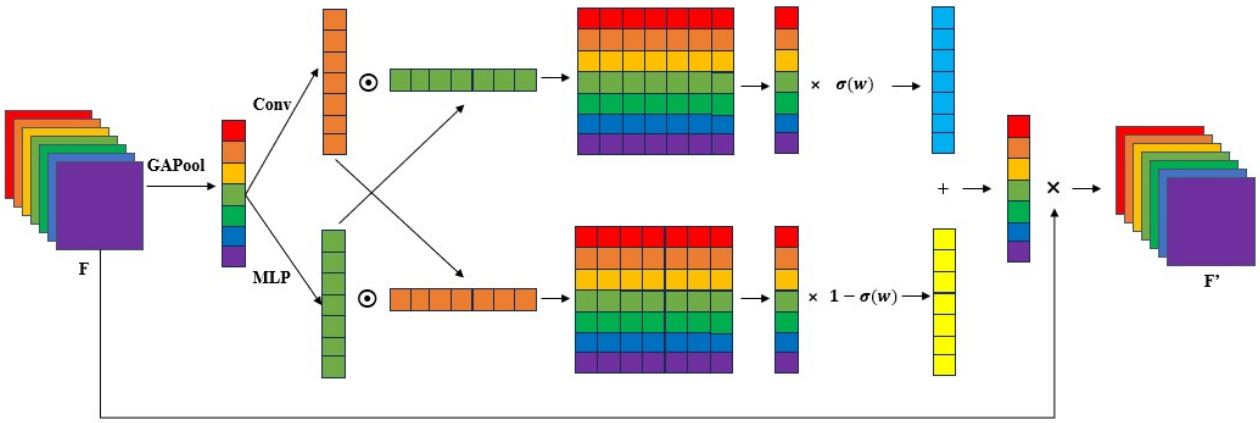


Figure 2. Architecture of the FCA Module.

Let the input feature map be $F \in R^{C\times H\times W}$. The first step is global average pooling per channel to obtain the channel descriptor:

$$U_n = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} F_n(i,j),\ U \in R^C \tag{1}$$

Then, a band matrix $B \in R^{C\times k}$ is constructed to extract local context vectors:

$$U_{lc} = \sum_{i=1}^{k} U \cdot b_i \tag{2}$$

Simultaneously, a diagonal matrix $D \in R^{C\times C}$ is used to derive global context vectors:

$$U_{gc} = \sum_{i=1}^{C} U \cdot d_i \tag{3}$$

An interaction matrix is computed as

$$M = U_{gc} \cdot U_{lc}^T \tag{4}$$

By aggregating row and column statistics of M, a channel-wise attention vector $W \in R^C$ is generated using a sigmoid activation and a learnable scaling factor, which is then applied to the input feature map:

$$F* = W \otimes F \tag{5}$$

In our implementation, the FCA module is inserted after every two convolutional layers, enhancing the model's ability to emphasize low-contrast objects and improve feature discriminability.

3.3. Partial Convolution Module

While traditional standard and depthwise separable convolutions reduce parameter count and FLOPs, recent studies show that FLOPs reduction does not always lead to faster inference due to memory access overhead and inefficient inter-channel operations [13].

PConv (Partial Convolution) addresses this by selectively applying convolution to only a subset of input channels, effectively reducing redundant computation and improving runtime efficiency. Specifically, during each convolution operation, only the first portion of input channels is processed via standard convolution, while the remaining channels

are either directly passed or linearly fused. The overall structure of PConv is illustrated in Figure 3. Formally, for an input feature map $X \in R^{C\times H\times W}$ and channel utilization rate $r \in (0,1)$, the output is defined as

$$Y = [\text{Conv}(X_{1:rC}), X_{rC:C}] \tag{6}$$

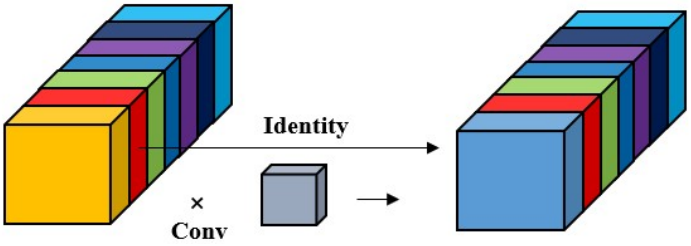


Figure 3. Architecture of the PConv Module.

This design retains discriminative power while significantly accelerating inference, making it suitable for real-time underwater applications. In our network, all standard 3×3 convolutions in the DEIM backbone are replaced by PConv layers, with $r = 0.5$ to balance accuracy and speed.

The main benefit of PConv is that it achieves an effective trade-off between efficiency and representation capacity. Instead of uniformly reducing the number of filters, which may degrade the ability to capture fine-grained details, PConv focuses computation on informative channels while keeping the model compact. As a result, FPH-DEIM achieves lower latency and reduced computational cost without sacrificing detection accuracy, which is particularly important for real-time underwater applications where both efficiency and robustness are required.

3.4. HWDown Downsampling Module

Underwater organisms are often small and subtle in visual appearance. Conventional downsampling methods such as max pooling or strided convolution tend to lose fine-grained details, leading to missed detections of small objects. To tackle this problem, we introduce the HWDown module, a downsampling method based on discrete wavelet transform.

Haar Wavelet Transform is a classic method for decomposing spatial information. For an input feature map $F \in R^{C\times H\times W}$, HWDown applies a 2D Haar transform to each channel, reducing spatial resolution to $\frac{H}{2} \times \frac{W}{2}$ while preserving high-frequency edge features:

$$F_{LL}, F_{LH}, F_{HL}, F_{HH} = \text{Haar}(F) \tag{7}$$

Here, *LL* captures the low-frequency components, and *LH*, *HL*, *HH* capture directional high-frequency edge information. These components are fused or concatenated to form the output feature map, which is passed to the decoder for further processing.

HWDown modules can be flexibly embedded in any downsampling layer. In our implementation, HWDown replaces conventional pooling in the feature pyramid network (FPN), thereby retaining semantic structure while enhancing the sensitivity to small-scale features—crucial for accurate underwater detection.

4. Experiments

4.1. Experimental Setup

All experiments in this work were performed on an Ubuntu 22.04 LTS system using the PyCharm-2025.1.3.1 integrated development environment. The experiments were run on a

server equipped with 15 vCPUs (Intel(R) Xeon(R) Platinum 8474C) and an NVIDIA GeForce RTX 4090D GPU with 24 GB of VRAM. The programming environment used Python 3.10 and PyTorch 2.0.1, with all experiments performed under identical configurations. The hyperparameter settings are shown in Table 1.

Table 1. Experimental Hyperparameter Settings.

Method	Configuration
Learning rate	0.0002
Weight_decay	0.0001
Batch size	8
Optimizer	Adamw
Image size	640 × 640
Epochs	300

4.2. Experimental Dataset

The URPC dataset [25] was used in this study. It originates from the National Underwater Robot Competition and comprises images captured by underwater robots across various depths and conditions, including diverse backgrounds, contrasts, brightness levels, blur degrees, and chromatic aberrations. The dataset includes four classes of underwater biological targets: echinus, holothurian, scallop, and starfish. A total of 8200 images were randomly split into training, test and validation sets at an 8:1:1 ratio, yielding 6560 training images, 820 test images, and 820 validation images.

4.3. Performance Metrics

To assess the performance of the proposed model, we used the standard COCO metrics, including Precision (P), Recall (R), mAP@0.5 (*mean Average Precision*), GFLOPs (Giga Floating-point Operations), and the number of model parameters. The formulas for key metrics are as follows:

$$P = \frac{TP}{TP + FP} \tag{8}$$

$$R = \frac{TP}{TP + FN} \tag{9}$$

$$mAP = \frac{\sum_{i=1}^{N} \int_0^1 P(R)\mathrm{d}R}{N} \tag{10}$$

where TP denotes true positives, FP denotes false positives, and FN denotes false negatives.

4.4. Experimental Results

4.4.1. Comparison of Attention Modules

To validate the effectiveness of the FCA module, we compared DEIM with several popular attention mechanisms (CBAM, ECA, SSA) under identical training settings using the URPC dataset. As shown in Table 2, the addition of FCA significantly improved Precision, Recall, mAP@0.5, and mAP@0.95 by 1.57%, 1.81%, 1.82%, and 1.04%, respectively.

Table 2. Comparison of various Attention Mechanisms.

Method	P (%)	R (%)	mAP@0.5 (%)	Parameters (M)	GFLOPs (B)
DEIM	86.3	85.6	86.2	8.1	7.5
DEIM + CBAM	81.7	79.8	82.2	9.1	8.6
DEIM + ECA	83.7	81.4	86.3	10.7	9.7
DEIM + SSA	83.9	80.5	85.2	8.1	7.6
DEIM + FCA	88.9	87.6	88.6	10.4	9.5

Figure 4 compares the heatmaps of DEIM models enhanced with different attention modules, illustrating their effectiveness in focusing on target regions within underwater biological images. Due to interference from the complex underwater environment, the original model lacks sufficient attention to the target areas. Although the ECA and SSA modules improve attention to some extent, they fail to clearly distinguish targets from the background. The CBAM module achieves a better separation between targets and background, but it still assigns insufficient attention to the target regions compared to the FCA module. In contrast, the FCA mechanism establishes an interaction between global and local information and dynamically adjusts feature weights at the channel level. This enhances target information while suppressing redundant noise, resulting in more accurate focus on underwater biological targets and reduced background interference, thereby improving detection accuracy.

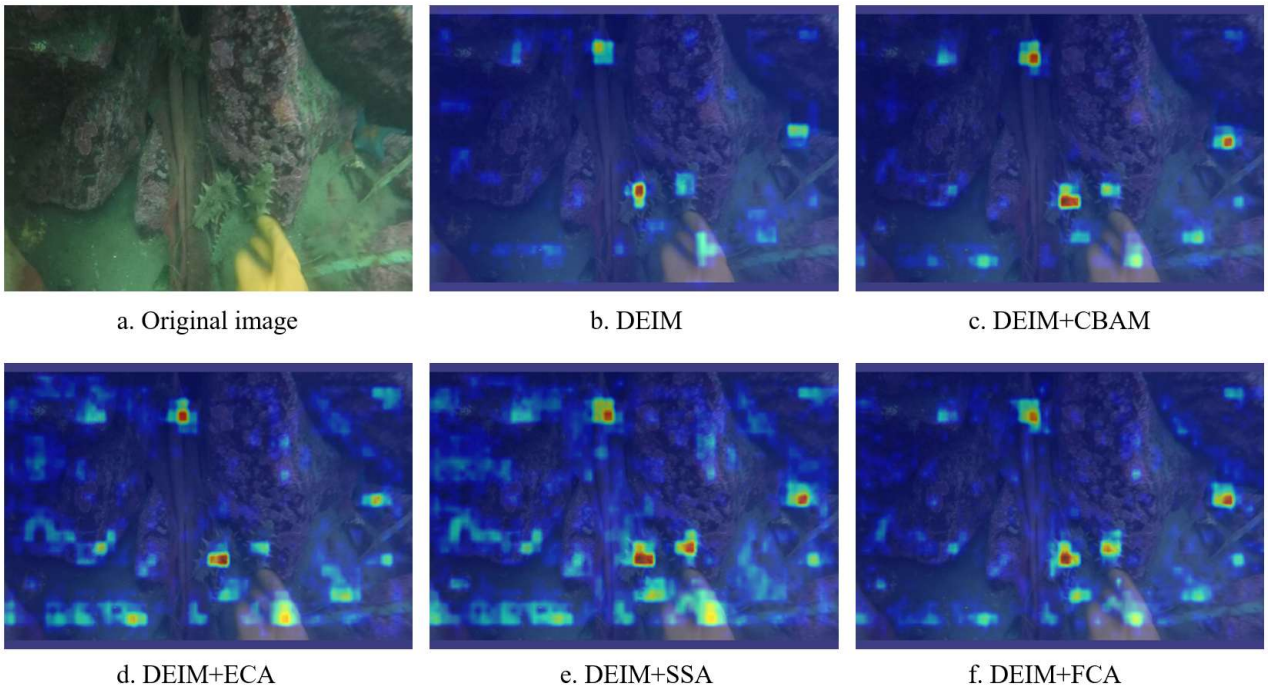


Figure 4. Comparison of heatmaps after adding different attention modules to the DEIM model.

4.4.2. Ablation Study

To assess the contribution of each module in the FPH-DEIM framework, ablation experiments were carried out under the same experimental settings. The results are summarized in Table 3.

Table 3. Comparative results of ablation experiment.

ID	FCA	PConv	HWDown	P (%)	R (%)	mAP (%)	Parameters (M)	GFLOPs (B)
1	×	×	×	86.3	85.6	86.2	8.1	7.5
2	√	×	×	87.9	87.8	88.5	9.4	10.2
3	×	√	×	85.8	84.7	85.4	4.6	4.3
4	×	×	√	87.8	87.7	87.4	8.6	9.2
5	√	√	×	88.6	89.7	88.9	8.2	10.5
6	√	×	√	86.5	88.8	88.9	10.4	11.2
7	×	√	√	89.2	88.1	88.2	5.2	4.9
8	√	√	√	89.8	87.7	89.4	7.2	7.1

Experiment 1 serves as the baseline model without any additional modules. With 8.1 M parameters and 7.5 GFLOPs, it achieves a precision (*P*), recall (*R*), and *mAP* of 86.3%, 85.6%, and 86.2%, respectively, indicating mediocre performance in complex underwater scenarios. In Experiment 2, the FCA (Feature Channel Attention) module was added to the backbone, markedly improving feature discrimination. As a result, P, R, and mAP increase to 87.9%, 87.8%, and 88.5%, respectively, with noticeable improvement in focusing on target regions.

In Experiment 3, the lightweight PConv module replaces part of the standard convolution operations. Although the parameter count drops to only 4.6 M, the detection performance slightly degrades, suggesting that PConv alone may be insufficient to enhance high-level semantic representations. Experiment 4 adopts the HWDown structure for downsampling, better preserving spatial structure, leading to an improved mAP of 87.4%.

Further combinations were tested in Experiment 5, where FCA and PConv modules are jointly used. The model achieves 88.6% precision, 89.7% recall, and 88.9% mAP, demonstrating the complementary effects of attention enhancement and lightweight convolution. Experiment 6 combines FCA and HWDown, reaching a comparable mAP of 88.9% but with increased computational complexity. Experiment 7 explores the combination of PConv and HWDown, achieving a balanced trade-off: despite only 5.2 M parameters, the model achieves 88.2% mAP, confirming the complementary advantages of the two modules.

Finally, Experiment 8 integrates all three modules—FCA, PConv, and HWDown—yielding the best overall performance: 89.8% precision, 87.7% recall, and 89.4% mAP, while keeping the model compact (7.2 M parameters and 7.1 GFLOPs). This confirms that the synergistic use of attention mechanisms, efficient convolutions, and structural downsampling significantly improves both accuracy and efficiency in underwater object detection tasks.

The ablation study results presented in Section 4 clearly demonstrate that each of the three core modules—FCA, PConv, and HWDown—positively contributes to performance improvement when integrated into the FPH-DEIM framework. Notably, the combination of FCA and HWDown yields a significant boost in detection accuracy without introducing substantial computational overhead, suggesting strong complementarity between feature enhancement and small object retention.

When all three modules are integrated, as in Experiment 8, the model achieves the highest mAP@0.5 of 89.4% with only 7.2 M parameters—surpassing all tested YOLO variants. This validates the effectiveness of the modular integration strategy in achieving both compactness and strong representation capability. Compared to using FCA or HWDown alone, the combined configuration significantly improves object boundary recognition and suppresses background interference, especially in dense-target and visually degraded underwater environments.

4.4.3. Performance Comparison Among Detection Models

To validate the performance advantages of the proposed FPH-DEIM algorithm in underwater biological object detection, a series of comparative experiments were conducted under identical settings. We compared FPH-DEIM with several mainstream detection algorithms, including Faster-RCNN, SSD, YOLOv5-n, YOLOv7-n, YOLOv8-n, YOLOv10-n, and the D-FINE series on the URPC dataset. Key evaluation metrics such as detection precision, model parameters, and computational complexity (GFLOPs) are summarized in Table 4.

Table 4. Performance Comparison among Detection Models.

Method	P (%)	R (%)	mAP@0.5 (%)	Parameters (M)	GFLOPs (B)
Faster-RCNN	68.2	59.4	67.8	136.7	18.7
SSD	74.2	68.7	75.4	25.1	74.8
YOLOv5-n	83.1	76.0	86.1	8.2	16.5
YOLOv7-n	85.1	76.1	84	6.1	13.1
YOLOv8-n	83.8	76.7	86.2	11.2	28.6
YOLOv10-n	81.5	76.3	84.6	9.1	21.6
D-FINE-n	84.7	83.2	84.7	22.5	18.8
DEIM-D-FINE-n	86.3	85.6	86.2	8.1	7.5
FPH-DEIM-n	**89.8**	**87.7**	**89.4**	**7.2**	**7.1**

As observed in Table 4, FPH-DEIM-n achieves significant improvements in detection accuracy, with a mAP@0.5 of 89.4%, which is 4.8 percentage points higher than the baseline YOLOv10-n. Additionally, FPH-DEIM-n has fewer parameters (2 M) and lower computational complexity (7.1 GFLOPs) compared to YOLOv10-n (9.1 M, 21.6 GFLOPs), demonstrating both high detection performance and lightweight characteristics.

Compared with traditional two-stage detectors like Faster-RCNN (mAP@0.5 of 67.8%) and SSD (75.4%), FPH-DEIM-n achieves improvements of 21.6 and 14.0 percentage points, respectively, while reducing the parameter size by more than 90% and significantly lowering inference complexity.

In comparison with other lightweight one-stage detectors, FPH-DEIM-n also shows comprehensive superiority. Specifically, it achieves mAP@0.5 gains of 3.3%, 5.4%, and 3.2% over YOLOv5-n, YOLOv7-n, and YOLOv8-n, respectively. At the same time, its model size is substantially smaller—by more than 1 million parameters compared to YOLOv5-n, and 4 million parameters compared to YOLOv8-n—indicating a more optimal trade-off between detection accuracy and computational efficiency.

Notably, when compared with DEIM-D-FINE-n, although the two models have similar computational complexity (7.1 G vs. 7.5 G FLOPs), FPH-DEIM-n improves detection accuracy by 3.2% and reduces the number of parameters by approximately 13.6%, highlighting the effectiveness of our structural design.

As shown in Table 4, FPH-DEIM-n outperforms traditional detectors such as Faster-RCNN and SSD, as well as YOLOv5/7/8/10 series, in terms of detection accuracy, model complexity, and parameter size. Compared with YOLOv10-n, FPH-DEIM-n achieves a 4.8% improvement in mAP@0.5 while cutting nearly 2 M parameters and 14.5 GFLOPs—demonstrating a more optimal trade-off between lightweight design and detection performance.

Moreover, FPH-DEIM-n achieves a Recall of 87.7%, significantly higher than that of YOLOv5-n (76.0%) and YOLOv10-n (76.3%), indicating superior capability in detecting small and low-contrast underwater targets. This advantage stems from the HWDown module's ability to preserve fine structural details and the FCA module's enhancement of ambiguous features. The precision-recall curves in Figure 5 further support this observation, showing that FPH-DEIM-n maintains high accuracy even at high recall levels, reflecting robust and stable detection behavior.

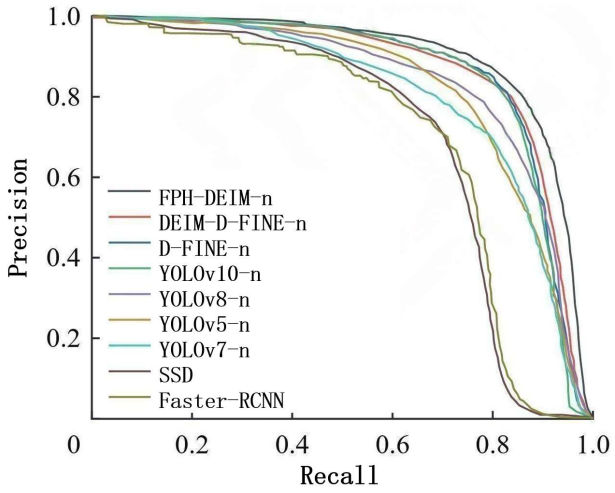


Figure 5. P-R Curves of Different Models.

4.5. Real-World Monitoring Experiment of the Improved Model

To further verify the deploy ability and real-time performance of the proposed FPH-DEIM algorithm in complex real-world underwater environments, we deployed the optimized model on an underwater intelligent inspection robot platform and conducted open-water tests. The configuration of the deployment platform is shown in Table 5, and the appearance of the underwater robot is presented in Figure 6.

Table 5. Hardware Configuration of the Underwater Robot.

Component	Description
Main Control Board	NVIDIA Jetson Xavier NX
Operating System	Ubuntu 20.04 + Jetpack 5.0
Camera	FLIR Blackfly S underwater industrial camera (1080 p, 30 fps)
Communication	Acoustic modem + Wi-Fi feedback
Power System	14.8 V lithium battery pack, rated at 10 Ah

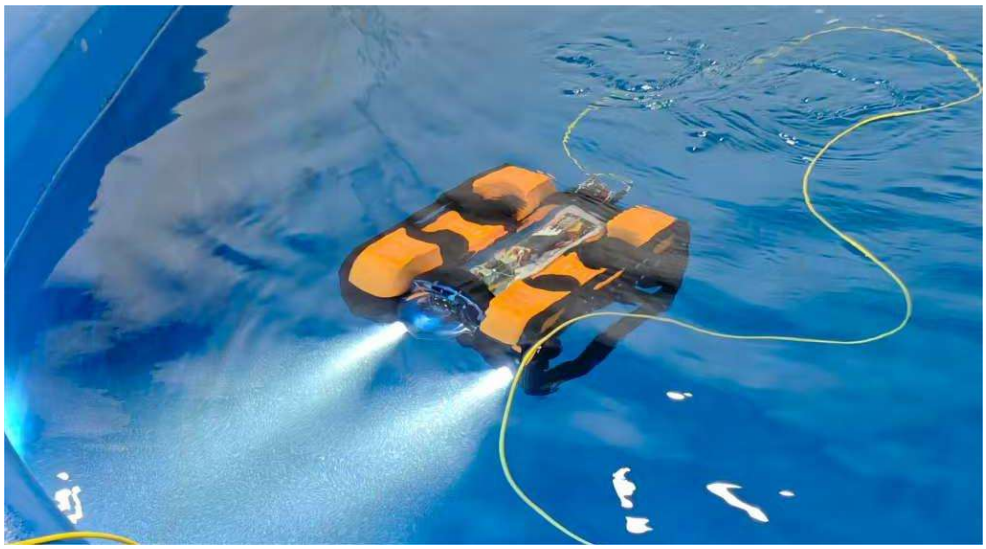


Figure 6. External View of the Underwater Robot.

To offer a clearer demonstration of the practical performance of the proposed FPH-DEIM model in underwater object detection tasks, we conducted a visual comparison with two representative models: the mainstream lightweight detector YOLOv8-n and the

original DEIM model. Figure 7 presents the comparison results, from which it is evident that under the same underwater image conditions, the YOLOv8-n model is susceptible to interference from uneven lighting and suspended particles, often resulting in missed detections or imprecise boundaries. Although the original DEIM model performs better in suppressing background interference, it still lacks the capability to distinguish fine-grained targets effectively.

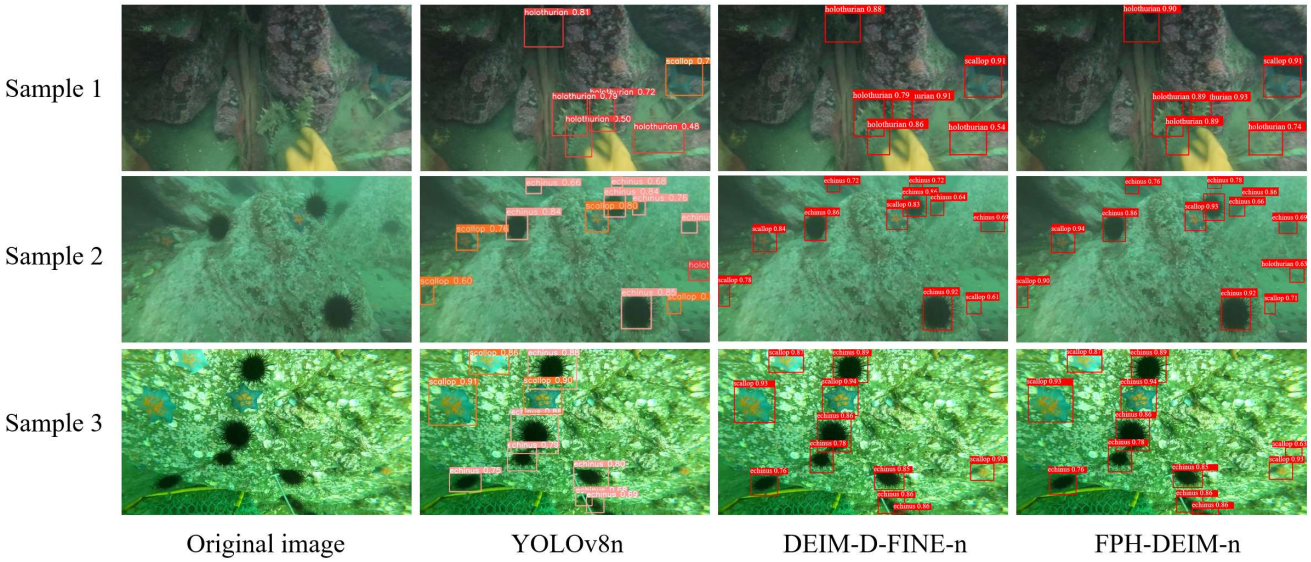


Figure 7. Comparative Detection Results of DEIM, FPH-DEIM, and YOLOv8.

In contrast, FPH-DEIM demonstrates superior target recognition performance. It not only accurately localizes multiple target instances but also exhibits improved boundary completeness and shape consistency. The advantages are particularly evident when dealing with images characterized by color distortion or significant pose variation.

These results further validate the effectiveness of the three core improvements proposed in this work: the FCA-enhanced perceptual mechanism strengthens the representation of blurred edge features; the PConv unit improves feature extraction efficiency while reducing redundancy; and the HWDown module effectively preserves spatial detail of small objects during downsampling. The synergy among these modules enables FPH-DEIM to achieve high detection accuracy and model compactness, while maintaining strong adaptability and robustness in complex underwater environments.

In real-world deployment tests, FPH-DEIM-n runs stably on the Jetson Xavier NX with an inference latency under 35 ms and a frame rate exceeding 28 FPS—demonstrating strong potential for real-time processing on autonomous underwater platforms. Compared to YOLOv8-n, it offers superior inference speed and lower memory consumption, enhancing its suitability for edge devices with strict power and bandwidth constraints.

As illustrated in Figure 7, FPH-DEIM performs robustly under challenging conditions such as heavy background interference, color distortion, and target pose variation. It delivers superior boundary completeness and stable localization, making it not only suitable for scientific underwater monitoring but also applicable to resource-constrained ecological inspection and task recognition scenarios. These qualities underscore the algorithm's practical value and deployment viability.

5. Discussion

The ablation experiments reveal that all three key modules—FCA, PConv, and HWDown—contribute positively to the overall detection performance when integrated

into different configurations of the FPH-DEIM framework. Specifically, FCA enhances the discriminative power of features by modeling both global and local channel interactions, while HWDown significantly helps retain fine-grained spatial details, especially for small objects. Although the inclusion of PConv alone results in a slight reduction in mAP, it substantially reduces computational load and parameter count, making it particularly well-suited for resource-constrained platforms. When all 3 modules are combined, the model attains the highest detection accuracy (mAP@0.5 of 89.4%) with only 7.2 million parameters and 7.1 GFLOPs, demonstrating an effective balance between performance and efficiency.

When compared with other mainstream detection models, FPH-DEIM-n consistently outperforms across several key metrics. Relative to YOLOv10-n, it achieves a 4.8% gain in mAP@0.5, reduces parameters by approximately 2 million and lowering the computational cost by 14.5 GFLOPs. Its Recall is 11.7% and 11.4% higher than YOLOv5-n and YOLOv8-n, respectively, indicating superior sensitivity to low-contrast and small underwater targets. As shown in the P-R curves in Figure 5, FPH-DEIM-n maintains high precision even at elevated recall levels, reflecting its robustness and stability in detection tasks.

In addition, real-world deployment tests on the Jetson Xavier NX platform further validate the model's practical value. FPH-DEIM-n achieves an inference latency of under 35 milliseconds and operates at over 29 FPS, while consuming only 2.8 GB of GPU memory—substantially lower than YOLOv8-n and DEIM-D-FINE. These results affirm its suitability for deployment in embedded environments such as underwater robots. As illustrated in Figure 7, the model delivers precise object localization and maintains contour integrity even in visually complex underwater scenes, demonstrating strong anti-interference capabilities and deployment readiness.

Despite the favorable results, several limitations remain. Under extreme conditions such as low lighting or high turbidity, the model may still produce false positives or miss detections. Furthermore, the current evaluation is restricted to the URPC dataset, which mainly covers coastal underwater scenes. This limitation may reduce the generalizability of the findings to other ecological conditions such as deep-sea or highly turbid waters. In addition, the present design relies solely on RGB imagery without leveraging complementary sensing modalities such as sonar, hyperspectral, or LiDAR, which limits its comprehensive perception capabilities in more complex underwater scenarios. Although real-world deployment validates the practicality of the proposed model, large-scale and long-term field testing across diverse environments is still required to fully confirm its reliability.

Future research may therefore explore several directions: integrating underwater image enhancement and low-light compensation mechanisms to improve robustness; adopting lightweight Transformer-based encoders to enhance multi-scale semantic representation; constructing geographically diverse datasets to improve generalization; and incorporating multi-modal sensory data to broaden the applicability of the algorithm in real-world marine perception tasks.

6. Conclusions

This paper proposes FPH-DEIM, a lightweight object detection algorithm that integrates perception enhancement and efficiency optimization to address the multiple challenges of underwater biological target detection. Based on the DEIM architecture and tailored to the characteristics of underwater environments, we introduce module-level innovations in feature extraction, convolutional efficiency, and spatial perception to construct a detection framework suitable for edge computing platforms. The main contributions are summarized as follows:

1. A feature-aware mechanism integrating FCA is introduced to effectively enhance the representational contrast between blurred targets and noisy backgrounds in underwater images, significantly improving detection accuracy.
2. A highly efficient PConv unit is designed to reduce computational redundancy, resulting in lower model latency and memory access load.
3. A novel HWDown module based on Haar wavelets is proposed, which preserves spatial detail during downsampling and enhances the model's capability to identify small objects.
4. Extensive experiments on public underwater datasets and real underwater robotic platforms demonstrate that FPH-DEIM outperforms mainstream lightweight detectors in terms of accuracy, model size, and inference speed, showing strong potential for practical deployment.

Future work may involve integrating underwater image enhancement preprocessing, exploring lightweight Transformer-based encoders, and extending the method to multimodal perception scenarios such as sonar imaging and LiDAR point clouds.

Author Contributions: Conceptualization, Q.L. and W.S.; Methodology, Q.L.; Formal analysis, Q.L.; Investigation, Q.L.; Resources, W.S.; Software, Q.L.; Writing—original draft preparation, Q.L.; Writing—review and editing, W.S.; Visualization, Q.L.; Supervision, W.S.; Project administration, W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by program for scientific research start-up funds of Guangdong Ocean University (YJR24010). China University Research Innovation Fund: Research on campus Internship and Training Teaching Platform for Artificial Intelligence Talent Training (2023ZY010). Laboratory Safety Research Project of Guangdong Ocean University in 2023: Research on Intelligent System for Laboratory Safety Monitoring, Early Warning and Emergency Management; Yangjiang Campus Artificial Intelligence Interdisciplinary Practice Teaching Platform (PX-1302024001). Guangdong Province College Students' Innovation and Entrepreneurship Training Program (Project No. S202510566003).

Data Availability Statement: There were no new data created, and there was no particular data used during the research.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Li, C.; Guo, J.; Cong, R.; Pang, Y.; Wang, B. Underwater Image Enhancement via Medium Transmission-Guided Multi-Color Space Embedding. *IEEE Trans. Image Process.* **2021**, *30*, 4985–5000. [CrossRef] [PubMed]
2. Islam, M.J.; Xia, Y.; Sattar, J. Fast Underwater Image Enhancement for Improved Visual Perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. [CrossRef]
3. 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, NV, USA, 27–30 June 2016; pp. 779–788. [CrossRef]
4. Li, J.; Yang, W.; Qiao, S.; Gu, Z.; Zheng, B.; Zheng, H. Self-Supervised Marine Organism Detection from Underwater Images. *IEEE J. Ocean. Eng.* **2025**, *50*, 120–135. [CrossRef]
5. Li, Y.; Zhang, L.; Wang, K.; Xu, L.; Gulliver, T.A. Underwater Acoustic Intelligent Spectrum Sensing with Multimodal Data Fusion: A Mul-YOLO Approach. *Future Gener. Comput. Syst.* **2025**, *173*, 107880. [CrossRef]
6. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
7. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. In Proceedings of the 14th European Conference on Computer Vision (ECCV 2020), Glasgow, UK, 23–28 August 2020; pp. 213–229. [CrossRef]

8. Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; Dai, J. Deformable DETR: Deformable Transformers for End-to-End Object Detection. In Proceedings of the International Conference on Learning Representations (ICLR), Virtual Conference, 3–7 May 2021. Available online: https://openreview.net/forum?id=gZ9hCDWe6ke (accessed on 21 August 2025).
9. Chen, Q.; Wang, Y.; Yang, T.; Zhang, X.; Cheng, J.; Sun, J. You Only Look One-Level Feature. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 19–25 June 2021; pp. 13039–13048. [CrossRef]
10. Huang, S.; Lu, Z.; Cun, X.; Yu, Y.; Zhou, X.; Shen, X. DEIM: DETR with Improved Matching for Fast Convergence. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 11–15 June 2025; pp. 15162–15171.
11. Musa, U.I.; Roy, A. Marine Robotics: An Improved Algorithm for Object Detection Underwater. *Int. J. Comput. Graph. Multimed.* **2022**, *2*, 1–8. [CrossRef]
12. Sun, H.; Wen, Y.; Feng, H.; Zheng, Y.; Mei, Q.; Ren, D.; Yu, M. Unsupervised Bidirectional Contrastive Reconstruction and Adaptive Fine-Grained Channel Attention Networks for Image Dehazing. *Neural Netw.* **2024**, *176*, 106314. [CrossRef] [PubMed]
13. Chen, J.; Kao, S.; He, H.; Zhuo, W.; Wen, S.; Lee, C.H.; Chan, S.H.G. Run, Don't Walk: Chasing Higher FLOPS for Faster Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 12021–12031.
14. Xu, G.; Liao, W.; Zhang, X.; Li, C.; He, X.; Wu, X. Haar Wavelet Downsampling: A Simple but Effective Downsampling Module for Semantic Segmentation. *Pattern Recognit.* **2023**, *143*, 109819. [CrossRef]
15. Yu, H.; Yin, Y.; Huang, T.; Liu, C. U-YOLO: An Improved YOLOv3 Model for Underwater Object Detection. *Appl. Sci.* **2022**, *12*, 3481. [CrossRef]
16. Canny, J. A Computational Approach to Edge Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **1986**, *8*, 679–698. [CrossRef] [PubMed]
17. Er, M.J.; Chen, J.; Zhang, Y.; Gao, W. Research Challenges, Recent Advances, and Popular Datasets in Deep Learning-Based Underwater Marine Object Detection: A Review. *Sensors* **2023**, *23*, 1990. [CrossRef] [PubMed]
18. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *Advances in Neural Information Processing Systems, Proceedings of the NIPS 2015*; MIT Press: Cambridge, MA, USA, 2015; Volume 1, pp. 91–99.
19. Zhang, Y.; Cui, Y.; Wu, X.; Wang, Y. Underwater Object Detection Based on Improved YOLOv3 Model. *J. Mar. Sci. Eng.* **2021**, *9*, 311. [CrossRef]
20. Meng, D.; Chen, X.; Fan, H.; Xu, G.; Xiang, S.; Pan, C.; Sun, J. Conditional DETR for Fast Training Convergence. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, QC, Canada, 11–17 October 2021; pp. 3651–3660.
21. Zhang, H.; Zhang, F.; Liu, S.; Zhu, J.; Ni, L.M.; Shum, H.Y. DINO: DETR with Improved DeNoising Anchor Boxes for End-to-End Object Detection. In Proceedings of the International Conference on Learning Representations (ICLR), Kigali, Rwanda, 1–5 May 2023. Available online: https://openreview.net/forum?id=3mRwyG5one (accessed on 21 August 2025).
22. Li, X.; Wang, Y.; Zhang, H.; Zhou, T.; Li, H.; Sun, J. DN-DETR: Accelerate DETR Training via Denoising Anchor Boxes. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 19–24 June 2022; pp. 13619–13628.
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
24. 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 2018), Munich, Germany, 8–14 September 2018; pp. 3–19.
25. 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. [CrossRef]

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.