Journal of *Marine Science and Engineering*

MDPI

Article

Lightweight Underwater Object Detection Algorithm for Embedded Deployment Using Higher-Order Information and Image Enhancement

Changhong Liu [1,*,†], Jiawen Wen [2,†], Jinshan Huang [2], Weiren Lin [1], Bochun Wu [1], Ning Xie [2] and Tao Zou [1,3,*]

[1] School of Mechanical and Electrical Engineering, Guangzhou University, Guangzhou 510006, China; 2007300052@e.gzhu.edu.cn (W.L.); 32107300017@e.gzhu.edu.cn (B.W.)
[2] School of Electronics and Communication Engineering, Guangzhou University, Guangzhou 510006, China; 32107400043@e.gzhu.edu.cn (J.W.); 32119500102@e.gzhu.edu.cn (J.H.); 2112007068@e.gzhu.edu.cn (N.X.)
[3] Guangdong-Hong Kong-Macao Key Laboratory of Multi-Scale Information Fusion and Collaborative Optimization Control of Complex Manufacturing Process, Guangzhou 510006, China
* Correspondence: lch@gzhu.edu.cn (C.L.); tzou@gzhu.edu.cn (T.Z.)
† These authors contributed equally to this work.

Abstract: Underwater object detection is crucial in marine exploration, presenting a challenging problem in computer vision due to factors like light attenuation, scattering, and background interference. Existing underwater object detection models face challenges such as low robustness, extensive computation of model parameters, and a high false detection rate. To address these challenges, this paper proposes a lightweight underwater object detection method integrating deep learning and image enhancement. Firstly, FUnIE-GAN is employed to perform data enhancement to restore the authentic colors of underwater images, and subsequently, the restored images are fed into an enhanced object detection network named YOLOv7-GN proposed in this paper. Secondly, a lightweight higher-order attention layer aggregation network (ACC3-ELAN) is designed to improve the fusion perception of higher-order features in the backbone network. Moreover, the head network is enhanced by leveraging the interaction of multi-scale higher-order information, additionally fusing higher-order semantic information from features at different scales. To further streamline the entire network, we also introduce the AC-ELAN-t module, which is derived from pruning based on ACC3-ELAN. Finally, the algorithm undergoes practical testing on a biomimetic sea flatworm underwater robot. The experimental results on the DUO dataset show that our proposed method improves the performance of object detection in underwater environments. It provides a valuable reference for realizing object detection in underwater embedded devices with great practical potential.

Keywords: lightweight network; YOLO; underwater object detection; attention mechanism; embedded deployment; image enhancement

check for updates

Citation: Liu, C.; Wen, J.; Huang, J.; Lin, W.; Wu, B.; Xie, N.; Zou, T. Lightweight Underwater Object Detection Algorithm for Embedded Deployment Using Higher-Order Information and Image Enhancement. *J. Mar. Sci. Eng.* **2024**, *12*, 506. https://doi.org/10.3390/jmse12030506

Academic Editors: Adriano Mancini, Anna Nora Tassetti and Pierluigi Penna

Received: 23 February 2024
Revised: 12 March 2024
Accepted: 14 March 2024
Published: 19 March 2024

Copyright: © 2024 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

The oceans constitute one of the largest and most biodiverse ecosystems on earth [1]. The 2010 Global Census of Marine Life suggests that there may be over 2 million species of marine organisms, with only approximately 200,000 species of marine macro-organisms known to us. The ocean concurrently serves as a significant repository for oil, gas, minerals, chemicals, and various other aquatic resources. An increasing number of professionals from diverse fields are involved in the exploration of marine resources [2]. In the past decade, underwater robotics and detection technologies, including Autonomous Underwater Vehicles (AUVs) and Remotely Operated Vehicles (ROVs) [3,4] have experienced rapid development. They play a crucial role in the exploitation and conservation of marine resources and have attracted the attention of many scholars. In this context, the technology of underwater object detection plays a crucial role. Detection of underwater objects can be classified into

two main categories: acoustic system detection and optical system detection [5]. It relies on acquired image information for image analysis, encompassing classification, identification, and detection. In comparison to acoustic images, optical images offer higher resolution and greater information content [6,7]. Presently, there is a growing focus on underwater object detection using optical systems. Nevertheless, the intricate underwater environment and lighting conditions, along with the inevitable noise during visual information acquisition, pose significant challenges to the implementation of vision-based underwater object detection.

The objective of general object detection is to ascertain the location of a target instance within a natural image, relying on a vast array of predefined categories. This involves such essential tasks as object categorization and orientation. Presently, object detection methods can be broadly classified into two groups: traditional object detection methods and convolutional neural network (CNN)-based object detection methods [8].

Traditional object detection begins with a sliding window strategy [9] for scanning the entire image, considering the target's position and aspect ratio variations. Features are extracted using methods like SIFT [10] and HOG [11], and machine learning classifiers, including SVM, classify these features to detect objects within the window. However, traditional methods face challenges such as a lack of universality, high time complexity, and feature robustness issues, impacting their effectiveness. The emergence of convolutional neural networks (CNN) in recent years has significantly improved accuracy and speed, particularly in handling complex scenarios and targets, including underwater object detection.

Numerous studies have demonstrated the superiority of convolutional neural network (CNN)-based object detection over traditional algorithms. Currently, the presence or absence of anchors in the input data is the primary classification used by object detection algorithms in deep learning. Anchorless object detection models, such as CenterNet [12], CornerNet [13], and Transformer [14]-based end-to-end object detection (DETR) [15], eliminate the need for target anchors. This enhances model flexibility to adapt to varying sizes and shapes of targets, ultimately improving object detection accuracy and robustness. However, their suitability for real-time applications is limited due to the more complex network structure and higher hardware cost requirements. Anchor-based object detection models [16] mainly include two-stage object detection methods based on region suggestions such as RCNN [17], Fast-RCNN [18], Faster-RCNN [19], Mask-RCNN [20], Cascade-RCNN [21], etc., and regression-based object detection algorithms such as the SSD [22] and YOLO [23] (You Only Look Once) series of algorithms [24–29], and so on. One-stage object detection algorithms skip the generation of candidate regions and directly execute operations like feature extraction, classification, and regression on the entire image, resulting in quicker detection times. The YOLO series of algorithms, as a classical one-stage object detection approach, is widely utilized for its exceptional detection performance. Through continuous improvements and innovations, its performance in object detection tasks has become increasingly outstanding.

Amidst the advancements in underwater robotics, researchers seek an efficient and easily deployable underwater object detection model to augment the commercial value of underwater robots. In 2015, Li et al. [30] pioneered the use of deep CNN for underwater object detection, concurrently creating the ImageCLEF dataset. In 2017, Zhou et al. [19] enhanced VGG16 by integrating image enhancement techniques and the Faster R-CNN network for underwater target detection on the URPC dataset. In 2019, Weihong Lin et al. [31] introduced a generalization model designed to address challenges such as target overlapping and blurring in underwater object detection tasks. In 2021, to tackle the heterogeneity of underwater passive targets and classification challenges, Weibiao Qiao et al. [32] presented a design for timely and accurate underwater target classifiers using Local Wavelet Acoustic Patterns (LWAP) and Multi-Layer Perceptron (MLP). With the focus on accuracy improvement, Minghua Zhang et al. [33] shifted attention to the lightweight aspect of underwater object detection models. To improve real-time and lightweight

performance, they introduced a lightweight underwater object detection method that incorporates MobileNet v2, the YOLOv4 algorithm, and an attention mechanism. In 2023, to address the challenges posed by the intricate underwater scenes and the limited ability to extract object features, Zhengwei Bao et al. [34] proposed a parallel high-resolution network for underwater object detection. Kaiyue Liu et al. [35] enhanced the model performance by incorporating a residual module and integrating a global attentional mechanism into the object detection network. Dulhare UN et al. [36] employed Faster R-CNN and data augmentation algorithms to tackle the issue of low accuracy in detecting humans in underwater environments. In 2024, Rakesh Joshi et al. [37] addressed the issue of underwater scattering caused by suspended particles in water, severely degrading signal detection performance. They proposed a degradation condition-based three-dimensional (3D) integral imaging (InIm) integrated deep learning bifunctional approach for underwater object detection and classification.

Additionally, deploying these object networks on embedded hardware presents a new challenge. The prevalent approach involves employing embedded development boards like Openmv [38], k210 [39], and Jetson series [40] development boards, along with image acquisition modules, to form a vision system. Xin Feng et al. [41] conducted a comprehensive analysis of recent advances in computer vision algorithms and their corresponding hardware implementations. In 2020, Yu-Chen Chiu et al. [42] introduced a lightweight object detection model based on Mobilenet-v2, demonstrating a balance between speed and accuracy through real-world tests on the Nvidia Jetson AGX Xavier platform. Subsequently, in 2023, Sichao Zhuo et al. [43] proposed a lightweight meter reading recognition network using deformable features and aggregation, deploying it on a Jetson TX1 smart vehicle for meter reading applications, and realized the performance of SOTA. Unlike high-performance graphics cards on the host side, limitations such as storage speed and maximum arithmetic support of embedded development boards constrain the practical application of various underwater object detection networks [44].

In conclusion, owing to the complexity of the underwater environment, current underwater object detection algorithms encounter numerous challenges in practical applications. These challenges include images affected by seawater refraction, scattering, and other factors, resulting in blurred images; multi-scale transformation of underwater targets due to different angles and distances of collected data; target occlusion caused by the aggregation of underwater organisms; the substantial computational overhead of traditional object detection models, making deployment on embedded devices difficult, and more. In the face of these challenges, the field of underwater object detection still has a long way to go.

Among the YOLO series algorithms, YOLOv7 stands out with superior accuracy, faster speed, heightened stability, and increased suitability for industrial applications. Consequently, in this paper, we introduce the YOLOv7-GN model, built upon the YOLOv7 framework, and validate the efficacy of our approach through experiments conducted on underwater images. Our contributions are outlined as follows:

(1) In this study, FUnIE-GAN is employed for underwater image enhancement, and data enhancement methods, including Mixup, are applied during dataset preprocessing to capture the feature information of the object effectively.

(2) In this study, an enhanced head network is designed for multiscale higher-order information interaction, aiming to enhance the spatial interaction capability of YOLOv7-GN. This improvement allows the extension of self-attention's second-order interactions to arbitrary orders by integrating recursive gated convolution (Conv) with a high degree of flexibility and customizability, all without introducing a significant computational burden.

(3) In order to capture more global features in the image, ACmix is incorporated as the fundamental model for self-attention and convolution in YOLOv7-GN. Building on the characteristics of ACmix and Conv, we introduce a novel lightweight higher-order attentional layer aggregation network, denoted as ACC3-ELAN. ACC3-ELAN optimizes the gradient length of the entire network compared to ELAN, reduces the number of parameters, and improves the network's fusion sensing ability for higher-order features.

(4) To assess the feasibility of deploying our model on embedded devices and its effectiveness in practical applications, we designed a biomimetic underwater robot. Subsequently, we deployed our model into this underwater biomimetic robot to conduct real world testing scenarios.

2. Related Works

In this section, we provide an overview of the underwater image data enhancement method and introduce Convolutional and Attentional Fusion and Spatial Information Interaction, along with the YOLOv7 model, to better elucidate our solution.

2.1. Underwater Image Data Enhancement

Feature engineering plays a pivotal role in deep learning. Currently, underwater datasets face two primary challenges: (1) limited size and (2) typical issues in underwater data, including color shift, blurriness, low contrast, color distortion, significant noise, and unclear details. Addressing the issue of dataset size presents a wide array of data enhancement methods. In the initial stages, researchers employed data expansion techniques involving geometric transformations such as random scaling, cropping, and panning. Subsequently, there has been a proposal for data enhancement methods grounded in color transformations. For instance, enlarging a dataset through adjustments in brightness, hue, saturation, and other factors. Another technique for data enhancement is Mixup [45]. It involves training a neural network with a convex combination of sample pairs and their respective labels. Randomly selecting two images from each stack and mixing them in a given ratio to create a new image is the central concept of Mixup. The formula is expressed in Equations (1) and (2).

$$\widetilde{x} = \lambda x_i + (1 - \lambda) x_j \tag{1}$$

$$\widetilde{y} = \lambda y_i + (1 - \lambda) y_j \tag{2}$$

where x_i, x_j represent raw input vectors, y_i, y_j denote one-hot label encodings, and λ is a number randomly sampled from the beta distribution $\lambda \in [0, 1]$. Thus, Mixup extends the training distribution by integrating linear prior knowledge, achieving expansion with minimal computational cost. The visual representation of Mixup is depicted in Figure 1. Additionally, other data enhancement methods include Cut-out [46] and CutMix [47].

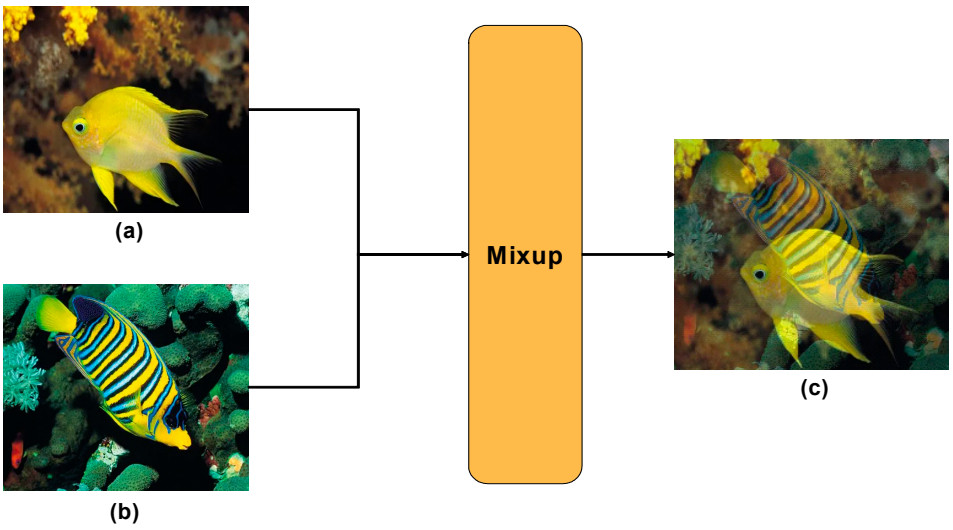

Figure 1. Mixup visualization process. (**a**) Deep-sea fish sample A. (**b**) Deep-sea fish sample B. (**c**) Sample plots generated by Mixup data enhancement.

In addressing underwater image degradation, traditional methods primarily involve the histogram technique, which directly manipulates underwater image pixels to enhance visual effects. Huang et al. [48] introduced the Relative Global Histogram Stretching (RGHS) method, rooted in the RGB and CIE-Lab color models. In addition, there are methods based on image restoration. These methods use an image model to estimate the relationship between clear, blurred, and transparent images. Subsequently, the restored image is generated through the dark channel prior (DCP) algorithm [49]. Amidst the rapid evolution of deep learning in computer vision, underwater image enhancement techniques have made significant progress. Examples include UIE-Net, based on convolutional neural networks [50], UIEC^2-Net [51], GAN-based multi-scale dense generative adversarial underwater image enhancement network [52], FUnIE-GAN [53], and more. The effect of various underwater image enhancements is illustrated in Figure 2.

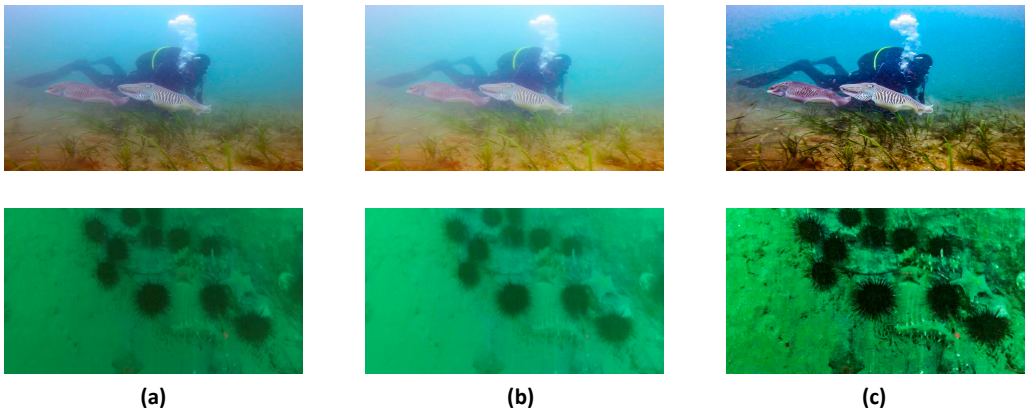

Figure 2. Effect of underwater image enhancement. (**a**) sample initial image, (**b**) color histogram-based method, (**c**) GAN-based deep learning image enhancement method.

As depicted in Figure 2, deep learning-based image enhancement methods exhibit superior generalization ability and more accurate color reproduction compared to traditional enhancement methods. The application of visual enhancement techniques effectively resolves inherent issues in underwater video images, including color distortion, low contrast, and fogging. Enhanced underwater images not only improve visual perception but also offer robust support for the subsequent execution of diverse visual tasks.

2.2. Convolutional and Attentional Fusion

The attentional mechanism's most notable feature is its ability to capture global features effectively. Vision Transformer [54] pioneered the application of Transformer to Backbone, achieving state-of-the-art (SOTA) performance. Although this approach enhances model accuracy, it comes at a higher computational cost. Convolutional networks can surpass Transformer-based networks in terms of speed for equivalent computational resources. Thus, a paradigm that combines convolution and self-attention can integrate both high-frequency and low-frequency information while maintaining a balance between speed and accuracy [55]. X.R. Pan et al. [56] sought a closer connection between self-attention [57] and convolution. Upon decomposing the operations of these two modules, it becomes evident that they heavily rely on the same 1×1 convolution operation. Based on this observation, they introduced a hybrid model named ACmix, skillfully integrating self-attention and convolution with minimal computational overhead. The model is structured into the following two phases.

In the first stage, the input features are projected by three 1×1 convolutions and then reshaped into N pieces, thus obtaining a rich set of intermediate features containing $3 \times N$ feature mappings.

In the second stage, convolution and self-attention follow different paradigms. In the self-attention path, intermediate features are assembled into N groups, each containing

three features, each from a 1×1 convolution. To follow the traditional multi-head self-attention mechanism [58], the three feature maps corresponding to the self-attention path are used as query, key, and value. For the convolution path, a lightweight fully connected layer is used to generate the feature maps by convolving the input features. By shifting and aggregating the generated features, it is possible to collect information from the local receptive field like traditional convolution. The overall model is shown in Figure 3.

Finally, the outputs of the two paths are added to obtain the final output, whose strength is controlled by two learnable scalars:

$$F_{out} = \alpha F_{att} + \beta F_{conv} \tag{3}$$

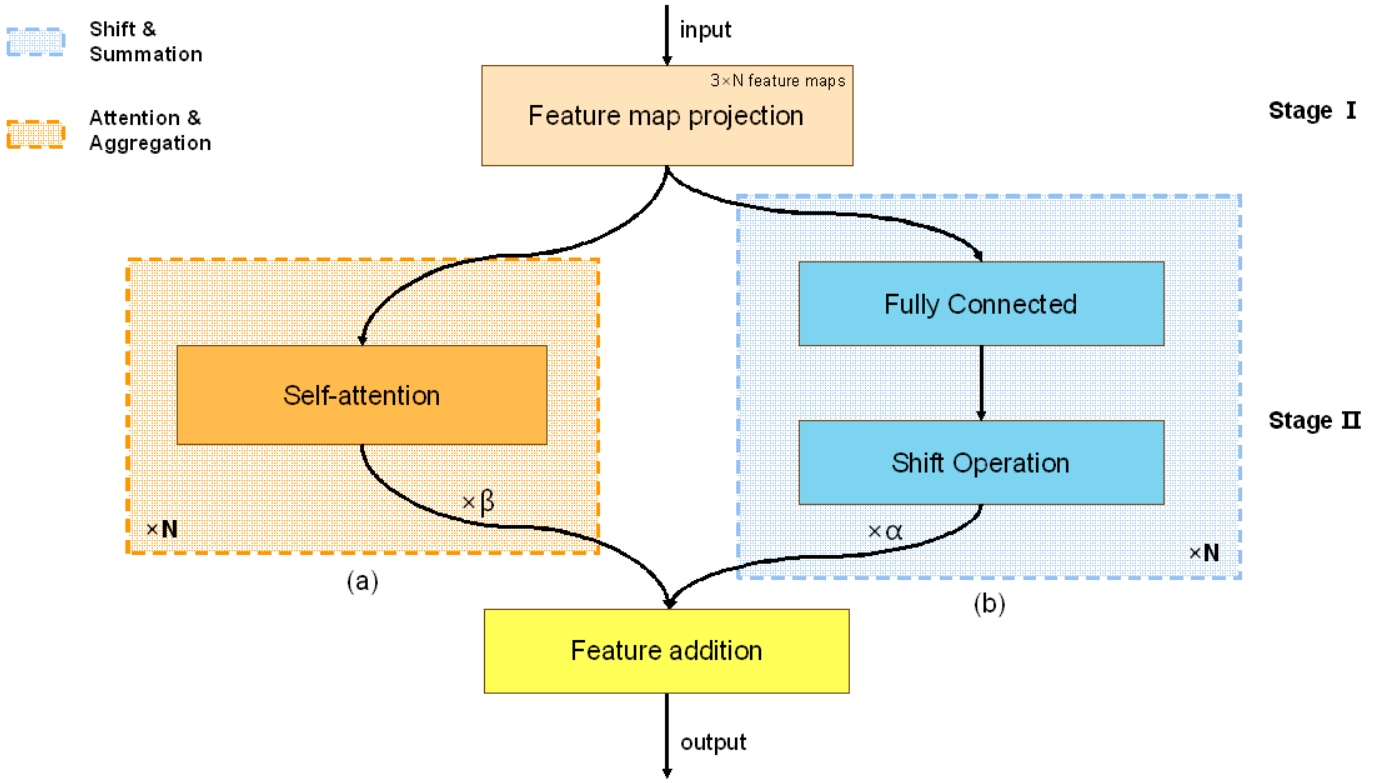

Figure 3. Schematic diagram of ACmix model. (**a**) Self-attention part. (**b**) Convolutional part. The intermediate features in the convolutional paths follow the shift and sum operations in the conventional convolutional module.

2.3. Spatial Information Interaction

Spatial interactions are fundamental in deep learning models as they facilitate understanding the relationships between different spatial locations within images or feature maps. Traditional convolutional neural networks (CNNs) [59] have limited ability to capture high-order spatial interactions due to their local receptive fields and shared weights across spatial dimensions. To overcome this limitation, dynamic convolutional operations, as employed in Dynamic Convolutional Neural Networks (DCNNs) [60], dynamically generate weights for convolutional filters based on input data, enabling adaptive attention to different spatial regions. Moreover, self-attention mechanisms [57], popularized by Transformer models, have been adapted for computer vision tasks to capture long-range dependencies and spatial interactions within images [54]. By integrating dynamic convolution or self-attention mechanisms into CNN architectures, models can capture higher-order spatial dependencies and effectively exploit spatial context for improved feature representation and prediction accuracy. Visualizations and comparisons of feature maps produced by various operations can illustrate how explicit modeling of spatial interactions leads to

more informative and context-aware representations, ultimately enhancing the modeling capacity of vision models.

The authors of Hornet [61] also highlight that the explicit higher-order spatial interactions introduced by architectural design contribute to enhancing the modeling capabilities of visual models. Their convolution-based framework fulfills the role of input adaptation, long-range interaction, and higher-order information interaction similar to the vision transformer.

2.4. YOLOv7 Model

YOLO employs convolutional neural networks in two distinct phases: feature extraction and object prediction. The network analyzes the input image using convolutional layers to extract relevant feature information in the feature extraction phase. Subsequently, in the prediction phase, YOLO partitions the image into grids and predicts multiple boxes for each grid. These predictions are refined by the head network to generate the final result.

The YOLOv7 model [24], an enhanced iteration of the YOLO series, surpasses its predecessors across multiple performance metrics. The YOLOv7 network consists of three distinct modules: the input module, the backbone network, and the head network. It introduces a novel gradient path design network called ELAN [62], which addresses the degradation issue encountered in depth model convergence during model scaling. This design strategy leverages existing computational units, resulting in reduced parameters, computation, and hardware resource usage, thereby enhancing inference speed and learning efficiency.

YOLOv7 is divided into several versions according to the complexity of the network, such as YOLOv7-tiny, YOLOv7, YOLOv7-X, YOLOv7-W6, YOLOv7-E6, etc., among which YOLOv7-tiny has 83.19% smaller number of parameters and 86.81% fewer GFLOPs compared to the original network of YOLOv7, but this brings about a performance loss in recognition accuracy. The work in this paper is based on the YOLOv7-tiny network, which aims to further lighten the model while improving the accuracy of object detection and providing a better object detection solution for underwater embedded devices.

3. Methods

In this section, we provide a detailed description of the implementation process and specifics of our approach. As illustrated in Figure 4, the complete architecture of our proposed solution involves the image enhancement module and YOLOv7-GN model, comprising the enhanced backbone network and head network.

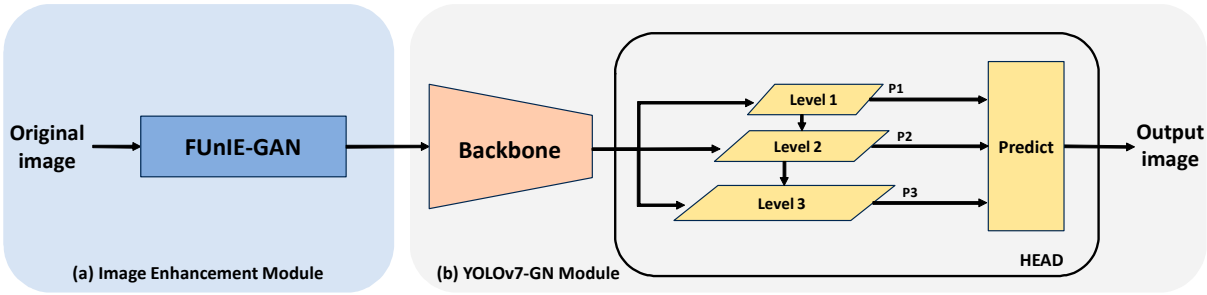

Figure 4. Complete architecture of the proposed solution. (**a**) Image enhancement module. (**b**) YOLOv7-GN module.

3.1. Underwater Image Enhancement Based on FUnIE-GAN

A Generative Adversarial Network(GAN) [63] consists of a Generator and a Discriminator. The Discriminator acts as a two-dimensional classifier to distinguish between real and simulated samples, while the Generator is used to model the real sample distribution and generate simulated samples based on that distribution. The Generator and Discriminator undergo iterative training to enable the Discriminator to accurately distinguish the

source of training samples. Simultaneously, this process aims to enhance the similarity between the simulated samples generated by the Generator and the standard samples, ultimately achieving the goal of underwater image enhancement. GAN has demonstrated significant generative capabilities in the domain of image generation [64] and is extensively utilized for dataset enhancement.

The core idea of FUnIE-GAN [53] is to find the nonlinear mapping relationship between the distorted image and the enhanced image through a condition-based GAN model. The generator in FUnIE-GAN employs a fully convolutional network with a Unet structure. Both the encoder and decoder consist of five convolutional layers, and the decoder utilizes anti-convolution. Regarding the discriminator, FUnIE-GAN adopts a Markovian PatchGAN [65] to identify local texture and content similarity. The complete network structure is illustrated in Figure 5.

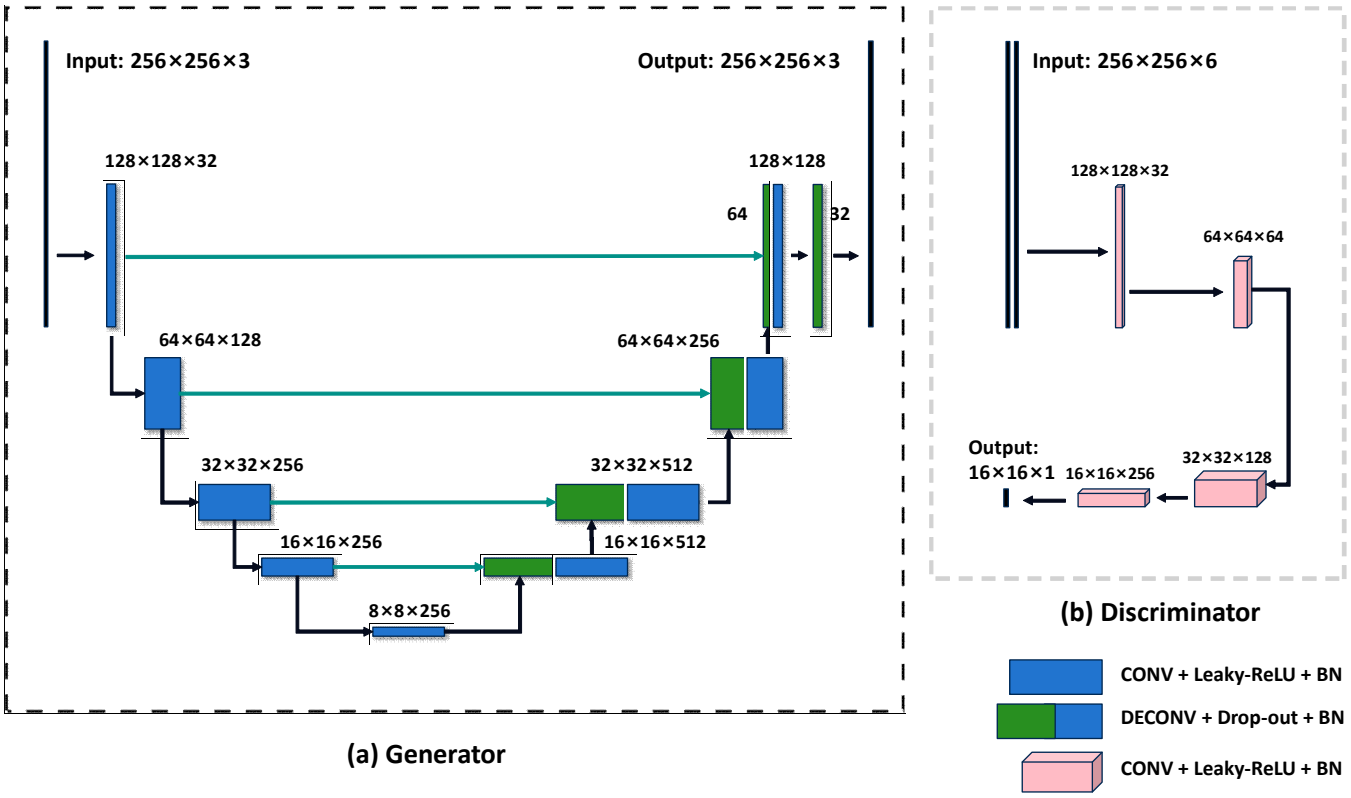

Figure 5. The overall network structure of FUnIE-GAN is divided into two parts: Generator and Discriminator. (**a**) Generator: a fully convolutional network using the U-net structure, in which the encoder and decoder have 5 convolutional layers each, and the decoder uses inverse convolution. (**b**) Discriminator: a Markovian PatchGAN is used to seek similarity between local texture and content.

3.2. YOLOv7-GN

3.2.1. Overview of YOLOv7-GN

There are considerable differences between the underwater and land environments, including factors like light attenuation, scattering, absorption, etc. These factors significantly degrade the quality of underwater images. Thus, a robust backbone is essential for feature extraction. To enhance model accuracy while minimizing parameter count and computational load, a lightweight higher-order attention layer aggregation network (ACC3-ELAN) is introduced into the Backbone. This enhanced Backbone reduces parameter count while capturing more intricate higher-order feature information across different scales, thereby facilitating robust feature extraction. In the head network, we leverage the interaction of multi-scale higher-order information to enhance it and fuse higher-order semantic informa-

tion from features at different scales. Furthermore, we introduce the AC-ELAN-t module, derived from pruning based on ACC3-ELAN, to further streamline the entire network while improving the model performance. Figure 6 illustrates the comprehensive network structure of YOLOv7-GN.

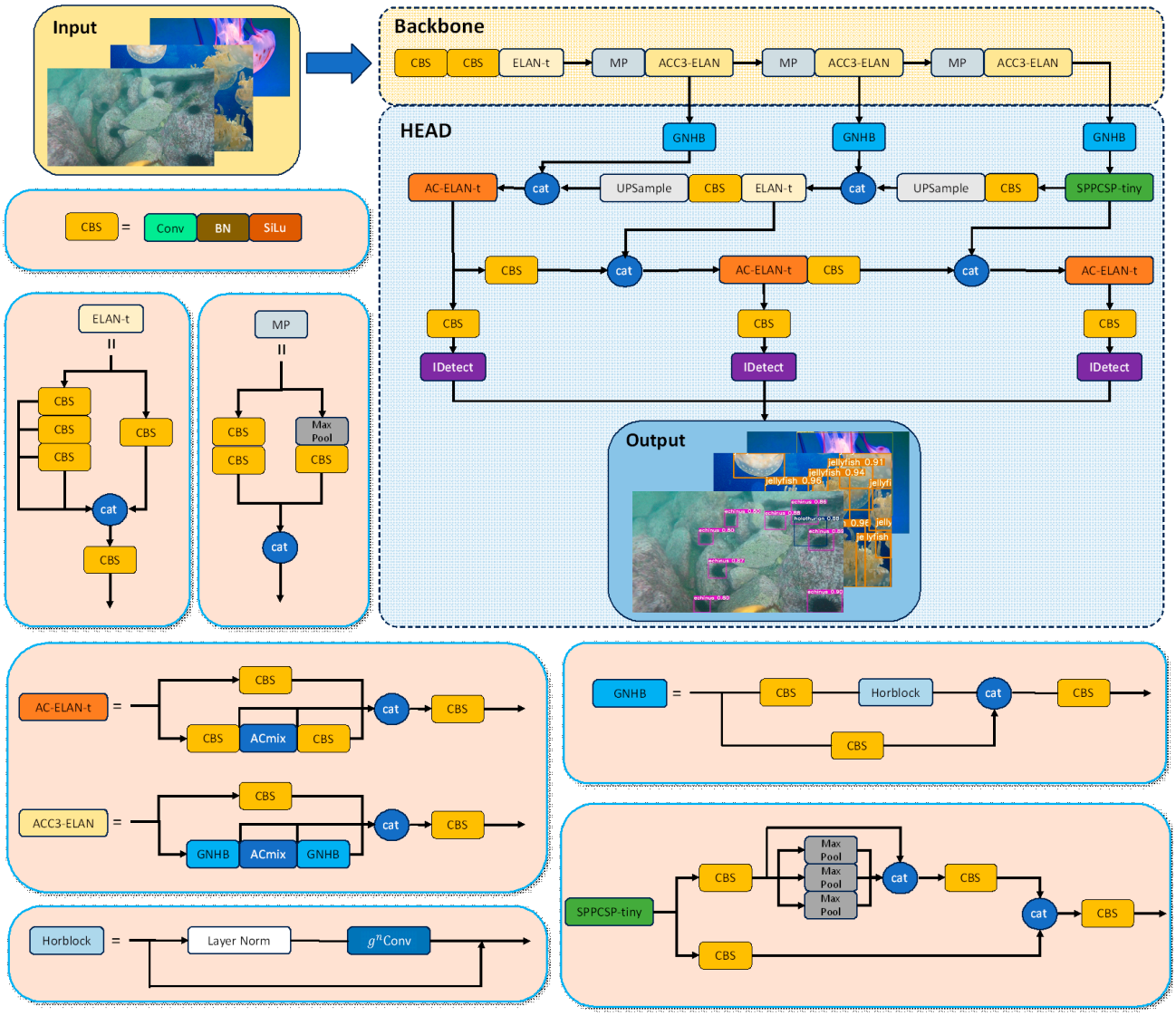

Figure 6. The overall network structure of YOLOv7-GN.

3.2.2. Backbone Network Based on ACC3-ELAN

ACmix [56] allows our proposed network to gain the ability to globally capture low-frequency information with a self-attention mechanism at minimal overhead. However, self-attention alone cannot achieve the same performance as Transformers. The three key factors for the success of Transformers can be attributed to the input adaptive weight generation strategy, the long-range spatial modeling capability, and the higher-order spatial interactions, the first two of which are implemented in the self-attention mechanism. However, there is no relevant research in the field of higher-order spatial interactions.

The introduction of g^n Conv recursive gated convolution [61] serves the purpose of extending the higher-order spatial interactions. This approach implements spatial interactions through gated convolution and recursive design, providing flexibility and customizability. Importantly, it can be compatible with a wide range of convolutional variants and extends the second-order interaction in self-attention to arbitrary orders

without introducing significant extra computation. Taking the first-order spatial interaction as an example, for the input feature $x \in R^{HW \times C}$, its specific operation is as follows:

In the first step, x is upscaled to twice its original size by a linear projection layer:

$$[m_0 \in R^{HW \times C}, n_0 \in R^{HW \times C}] = \phi_{in}(x) \in R^{HW \times 2C} \tag{4}$$

where H is the feature height, W is the feature width, C is the number of channels of the feature, and ϕ_{in} is the input fully connected layer.

In the second step, the features are matrix partitioned, and the DWConv (Depth-wise convolution) operation is performed on n_0:

$$n_0{}' = f(n_0) \in R^{HW \times C} \tag{5}$$

where f denotes a depth-separable convolutional layer.

The third step is to multiply $n_0{}'$ and m_0 to realize the spatial interaction:

$$m_1 = n_0{}' \cdot m_0 \in R^{HW \times C} \tag{6}$$

In the fourth step, the output is mapped to the desired dimension by a linear projection layer of channel mixing.

$$y_{out} = \phi_{out}(m_1) \in R^{HW \times C} \tag{7}$$

The detailed operation flow is shown in Figure 7.

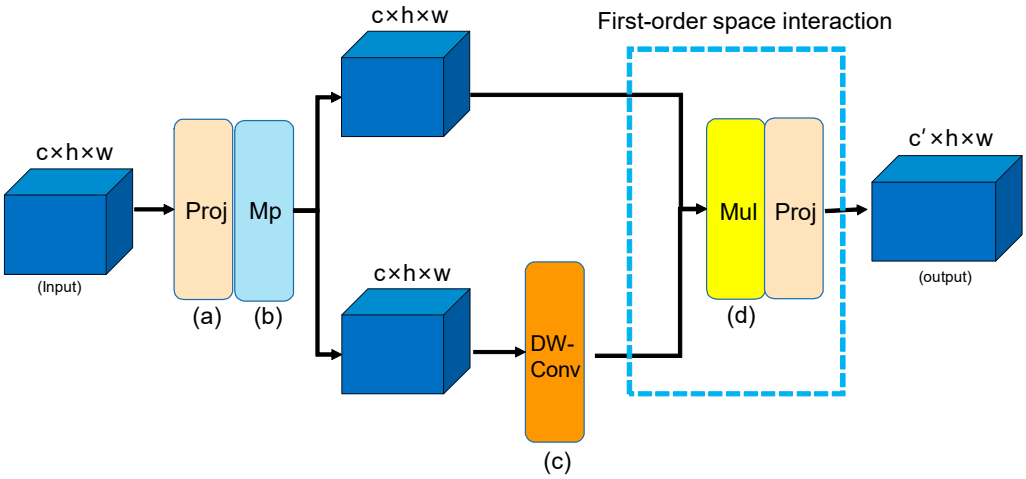

Figure 7. Principles of first-order spatial interaction. (**a**) Linear projection layer, which can map the input c-channels to any dimension. (**b**) Matrix partitioning splits the input into two outputs of equal size. (**c**) Depth-wise convolution. (**d**) Matrix Multiplication, which is used to calculate the product between two matrix.

The fundamentals of higher-order spatial interactions remain consistent with first-order interactions. The key distinction involves altering the dimension of the output derived from the linear projection layer. By doing so, spatial interactions can be realized, as depicted in Figure 8 for the third-order spatial information interactions of g^3 Conv.

Underwater images are commonly affected by lighting conditions, scattering, and absorption, leading to diminished image quality. In this study, we employ the GNHB residual connection module based on g^n Conv recursive gated convolution. The residual structure introduces additional information paths through jump connections and feature passing, facilitating improved learning and capturing of underwater object features.

ACmix combines self-attention with convolution. g^n Conv recursive gated convolution extends the self-attention mechanism with higher-order spatial interactions of feature information, and neither introduces additional computational overhead. We combine

the two and propose a lightweight higher-order attentional layer aggregation network (ACC3-ELAN) for use in Backbone. Compared to the original ELAN structure, ACC3-ELAN acquires more extensive global higher-order attentional features while concurrently decreasing the number of parameters.

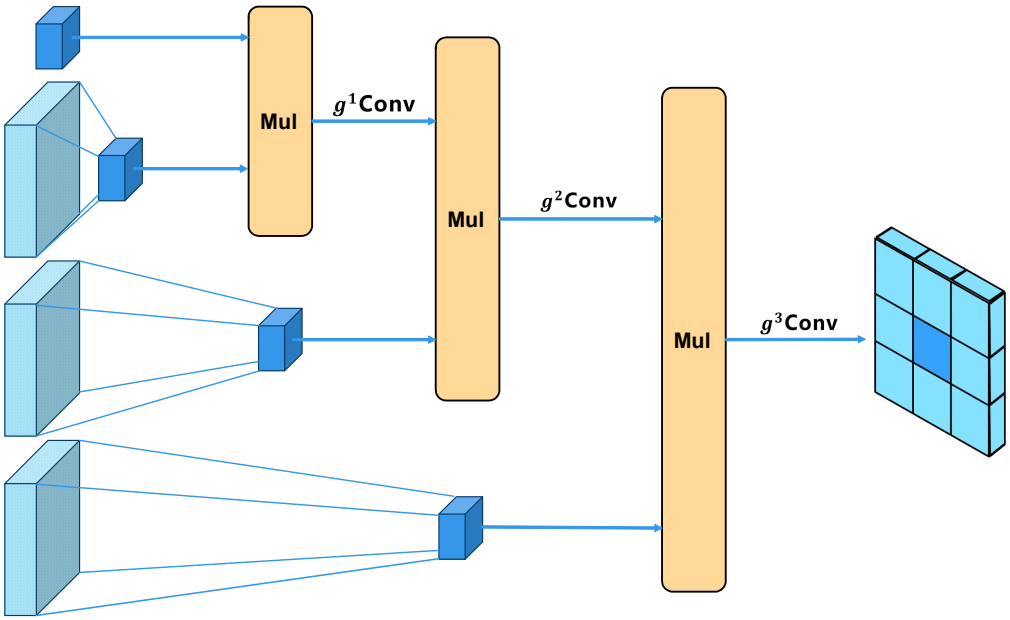

Figure 8. Conv. where Mul is Multiplication, which realizes arbitrary order spatial information interaction by using recursive gated convolution.

As illustrated in Figure 9, ACC3-ELAN builds upon the state-of-the-art ELAN architecture [62]. ELAN introduces a layer aggregation architecture with efficient gradient propagation paths, utilizing the shortest and longest gradient paths of each layer in the network. It optimizes the gradient lengths of the entire network through the stack structure in the computational blocks. ELAN enables the network to learn more abstract and advanced feature representations, capturing richer information in the input data and enhancing the model's expressive power. However, stacking multiple convolutional blocks leads to a significant increase in parameters and computational load, which we aim to mitigate.

To address this concern, we initiate model pruning on the original ELAN module. We evaluate the importance of each convolutional layer in ELAN, filter out redundant branch convolutional modules, and remove them. The pruned module is more lightweight, albeit with some performance loss. Subsequently, we replace the convolutional layers on the non-residual links in the pruned model with GNHB and ACmix. A combination of GNHB and ACmix is utilized on the main link. The input features initially pass through a GNHB to acquire preliminary higher-order semantic information. Subsequently, an ACmix is applied to enable the self-attention mechanism to globally capture low-frequency information with minimal computational expenditure. Finally, another GNHB is employed to further enhance the higher-order semantic information. In comparison to the original pure convolutional layers, this hybrid paradigm combines the two approaches to obtain richer global higher-order attentional features with reduced computational effort.

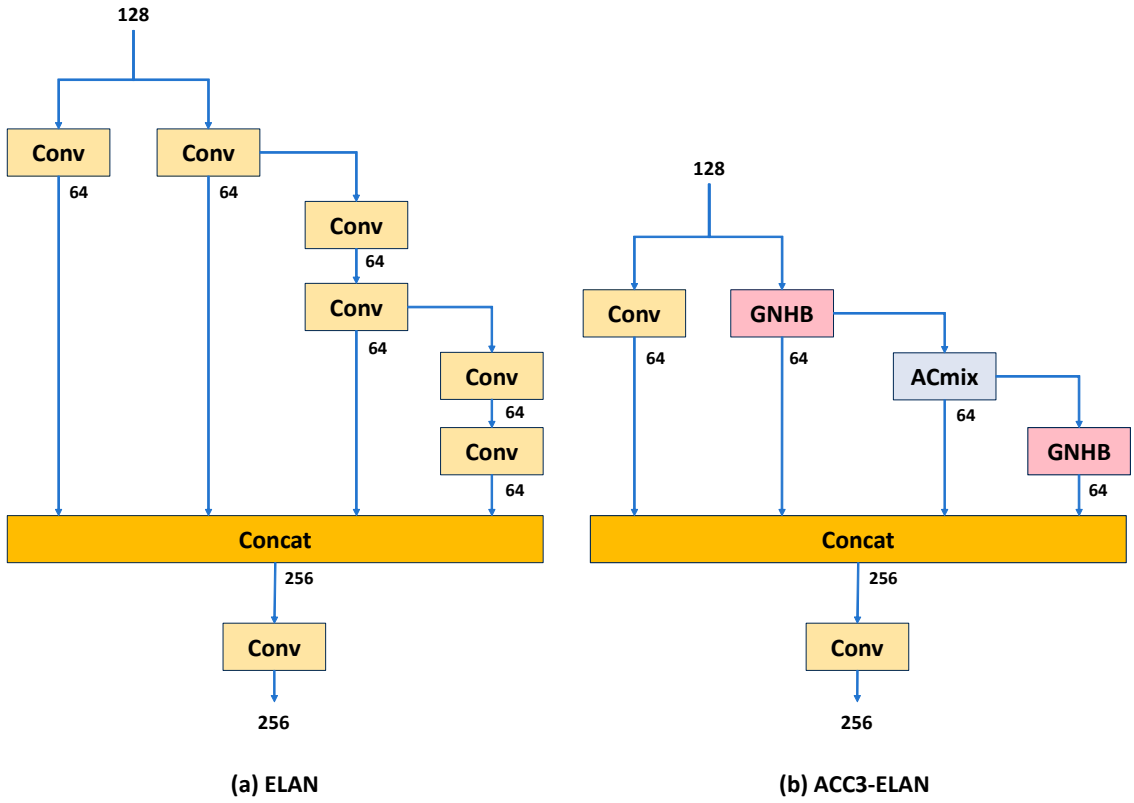

Figure 9. Structures of ELAN and ACC3-ELAN. (**a**) ELAN. (**b**) ACC3-ELAN.

3.2.3. Improved Head Network Based on Multi-Scale Higher-Level Information Interaction

The primary role of the Head part in YOLOv7 is to conduct multi-scale feature fusion of feature maps and regression prediction. Given the varied sizes and scales of targets in underwater environments, we replace the original convolutional layer with three GNHBs at the connection between the Backbone and Head. This replacement enhances the fusion of higher-order semantic information across different scales, enabling the extraction of more robust feature representations. The output channels of the three GNHB modules at the connection are 64, 128, and 128, respectively, arranged in a gradient based on the scale of the features. The receptive field can be dynamically adjusted to meet the detection and localization requirements of objects with different scales.

Instead of using ACC3-ELAN in the Head, we adopt the lighter weight AC-ELAN-t. In particular, the input and output channels remain unchanged, and the GNHB modules in ACC3-ELAN are replaced by CBS convolutional modules. This substitution further reduces the complexity of the overall network, facilitating easy deployment to edge devices.

4. Experiments and Results

In this section, we begin by presenting the experimental configurations and materials. Next, specific ablation experiments are conducted on YOLOv7-GN, followed by a detailed evaluation of the performance of the FUnIE-GAN underwater image enhancement. Finally, the model is compared with mainstream object detection models, and its deployment on an embedded device is discussed.

4.1. Experimental Environment

In our algorithm design and implementation, we employed the PyTorch deep learning framework. For model training, we utilized stochastic gradient descent to optimize the model parameters, setting the momentum to 0.937 and the weight decay parameter to 0.0005, and incorporating the Mixup image mashup data enhancement technique with a probability of 0.15. Additionally, for effective model training, we employed the dynamic

learning rate method, initializing the learning rate to 0.01 and gradually decreasing it using the cosine learning rate strategy [66]. The batch size was set to 16, and the number of workers was set to 30 to enable multi-threaded data loading. The complete training process of the model spanned 300 epochs, with each epoch representing the entire training dataset being fed into the network once.

In this experiment, all parameters are consistent with YOLOv7 [24] except for batch size and epochs, ensuring experimental fairness. Setting the epochs to 300 aims to prevent model overfitting/underfitting and reduce training hardware costs. The choice of batch size was primarily informed by studies [67,68] from other scholars and tasks of similar size. Experimental equipment and environmental setup details are provided in Table 1.

Table 1. Experimental setup and specific parameters.

Parameter	Configuration
CPU model	Intel Xeon Gold 6248R
GPU model	NVIDIA RTX A6000 (48 GB)
CUDA version	CUDA 11.4
Python version	Python 3.8.10
Deep learning framework	Pytorch 1.12.1
Operating system	Ubuntu 20.04.3 LTS

4.2. Evaluation Metrics and Dataset

4.2.1. Evaluation Metrics

To accurately evaluate the object detection model's performance in this study, we utilized precision, recall, Mean Average Precision (mAP), and $F1$ *score* as the performance evaluation metrics. Firstly, *IoU* (Intersection over Union) is a commonly used evaluation metric in the field of object detection to measure the degree of overlap between the model detection results and the actual target location. It quantifies the accuracy of detection by calculating the ratio of the intersection area to the union area of the detected box. If A is used to denote a real object box in the dataset and B is used to denote the predicted box after model detection, then the expression of IoU is as follows:

$$IoU = \frac{A \cap B}{A \cup B} \tag{8}$$

Precision measures how many of the samples that the model recognizes as positive categories are true positive categories. The expression is shown in Equation (9), where TP is the number of true cases (samples correctly predicted by the model to be positive), and FP is the number of false positive cases (samples incorrectly predicted by the model to be positive). Recall is a measure of how many of the true positive categories are successfully recognized by the model, and the formula is shown in Equation (10), where TP is the true cases, and FN is the false negative cases (samples incorrectly predicted by the model as negative categories).

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

Mean Average Precision (mAP) is a widely used, comprehensive metric used to evaluate object detection. mAP accounts for the differences between various object categories, and its calculation involves the area under the Precision–Recall curve. In object detection, AP is typically calculated at different confidence thresholds to generate the Precision–Recall (P–R) curve. However, the model may only be effective at a specific threshold (e.g., 50%).

Therefore, we specifically use $mAP50$ to evaluate performance. Generally, a higher mAP value indicates better performance of the model in object detection.

$$AP_c = \sum \int_0^1 p(R_c) dR_c \tag{11}$$

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{12}$$

The *F*1 *Score* is the reconciled average of Precision and Recall, which combines the precision and recall of the model.

$$F1\ score = \frac{2 \times P \times R}{P + R} \tag{13}$$

In this study, we denote accuracy, recall, and detection class as P, R, and c, respectively. Additionally, TP, FN, and FP represent true-positive, false-negative, and false-positive instances, respectively. Here, N denotes the total number of categorized classes.

For underwater image quality evaluation, we utilize two non-reference evaluation metrics, Naturalness Image Quality Evaluator (*NIQE*) [69] and Underwater Color Image Quality Evaluation (*UCIQE*) [60]. The computation of *NIQE* relies on the fitting process of both the natural and distorted images. First, *NIQE* establishes the fitting relationship between natural and distorted images by obtaining their means (k_1, k_2) and variance matrices (λ_1, λ_2). Then, it evaluates the quality of the images, calculating the distance between the fitting parameters of the warped images and the fitting parameters of the natural images. The *NIQE* value is a quantification of this difference, with lower values indicative of higher image quality.

$$NIQE = \sqrt{(k_1 - k_2)^T \left(\frac{\lambda_1 + \lambda_2}{2}\right)^{-1} (k_1 - k_2)} \tag{14}$$

UCIQE involves key parameters for image quality assessment. First, δ_c represents the standard deviation of color intensity, con_l denotes the contrast of luminance, and μ_s is the mean of saturation. These parameters are weighted by specific constant value $(\partial_1, \partial_2, \partial_3)$, with values of 0.4680, 0.2745, and 0.2576, respectively. These constant values align with the data in the literature [70].

$$UCIQE = \partial_1 \times \delta_c + \partial_2 \times con_l + \partial_3 \times \mu_s \tag{15}$$

The *UCIQE* formula quantifies the overall image quality in a weighted manner by combining the averages of color intensity, brightness contrast, and saturation. A higher *UCIQE* value corresponds to better image quality.

4.2.2. Dataset

DUO [71] is an open-source underwater dataset created in 2021. The dataset comprises a total of 74,515 objects including sea cucumbers, sea urchins, scallops, and starfish, with quantities of 7887, 50,156, 1924, and 14,548, respectively. As depicted in Figure 10, sea urchins are the most numerous, constituting 67% of the overall count. Due to variations in the economic values of different seafood products, leading to differences in species numbers, the overall data distribution exhibits a clear long-tail pattern. Figure 10 also illustrates the distribution of center coordinates and target sizes of the boxes in the dataset. The overall dataset sample distribution is observed to be unbalanced, with small targets (w < 0.3, h < 0.4) constituting the majority of the data, posing a more significant challenge for each object detection model.

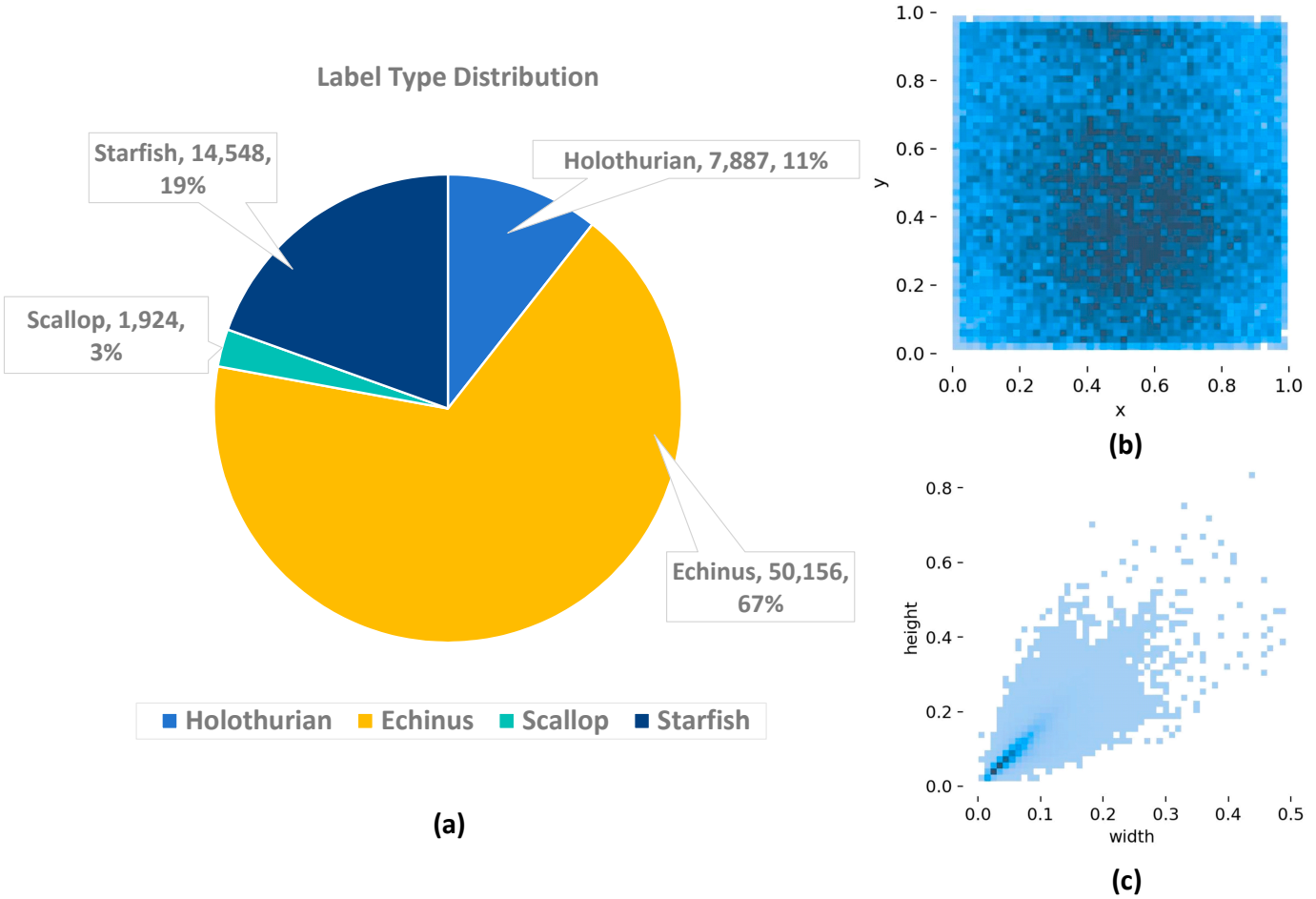

Figure 10. Distribution of targets in the DUO dataset. (**a**) Distribution of categories in the dataset. (**b**) Object center. (x,y) coordinate distribution; darker color means more concentrated distribution. (**c**) Distribution of target widths and heights; darker colors represent more concentrated distributions.

Simultaneously, we employ two non-reference underwater image evaluation metrics, NIQE and UCIQE, to assess the underwater image quality of the dataset. This facilitates the subsequent quantitative comparison of underwater image enhancement effects. Table 2 presents all the key information of the DUO dataset.

Table 2. Overall information about the DUO dataset.

Dataset	Environment	Image Number	Holothurian	Echinus	Scallop	Starfish	NIQE	UCIQE
DUO	Sea Floor	7782	7887	50,156	1924	14,548	11.893	0.4209

4.3. Ablation Experiments

To comprehensively assess the effectiveness of each scheme for YOLOv7-GN, we conducted ablation experiments on the DUO dataset, outlining the improvement strategies in Table 3. The performance evaluation focuses on precision (P), recall (R), F1 score, Mean Average Precision (*mAP*50 and *mAP*50:95), Parameters, Giga Floating-point Operations Per Second (GFLOPs), and the size of the weights file generated after completing model training (Model Size). The results of each scheme are presented in Table 4. Additionally, the results of the ablation experiments are depicted in Figure 11 to facilitate the observation of data variations.

Table 3. Different improvement programs based on YOLOv7.

Scheme	GNHB	AC-ELAN-t	ACC3-ELAN
YOLOv7-tiny (baseline)			
ACEL-YOLOv7t		√	
ACC3-YOLOv7t			√
C3-YOLOv7t	√		
C3-ACEL-YOLOv7t	√	√	
YOLOv7-GN (ours)	√	√	√

Table 4. Ablation experiment results. Parameters in bold are optimal.

Model	P (%)	R (%)	F1 (%)	mAP50 (%)	mAP50:95 (%)	Parameters (M)	GFLOPs	Model Size (MB)
YOLOv7 (baseline)	84.45	77.55	80.85	84.79	62.7	6.23	13.9	11.72
ACEL-YOLOv7t	85.35	77.8	81.4	85.34	62.49	5.91	13.1	11.52
C3-YOLOv7t	84.62	78.52	81.46	86.02	63.73	6.04	13.6	11.87
C3-ACEL-YOLOv7t	87.2	78.28	82.5	86.21	64.22	5.93	13.4	11.68
ACC3-YOLOv7t	85.88	**79.81**	**82.73**	86.25	64.19	5.54	**12.4**	11.11
YOLOv7-GN	**87.48**	78.46	82.72	**86.78**	**65.11**	**5.45**	12.6	**11.05**

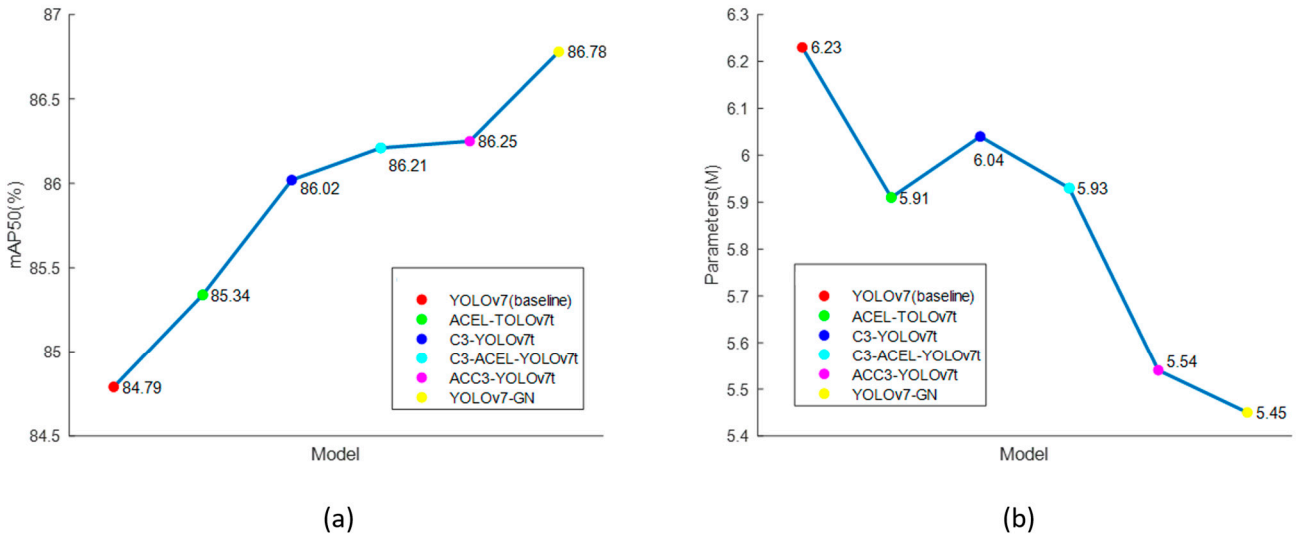

Figure 11. Comparison chart of ablation experiment results. (**a**) Comparison of mAP changes in different models. (**b**) Comparison of parameter changes in different models.

Specifically, in the ACEL-YOLOv7t scheme, we integrate the AC-ELAN-t obtained from pruning ACC3-ELAN into the head network. This replaces the original ELAN module connected to the detection head, and we arrange the AC-ELAN-t modules in a gradient according to the size of the feature pyramid. The other ELAN modules in the network remain unchanged. To investigate the impact of ACC3-ELAN on the extraction ability of high and low-frequency features and the overall network size, we primarily replace and improve the ELAN in the Backbone of the ACC3-YOLOv7t scheme. After each MP module in the Backbone, we replace the ELAN module with ACC3-ELAN, using it as a connection to the HEAD feature pyramid (FPN) at a key location. In the C3-YOLOv7t scheme, we primarily explore the gain of higher-order information interaction brought by GNHB to the overall network. We replace the two 1×1 convolutional layers connecting the Backbone in the original HEAD feature pyramid (FPN) with GNHB modules. Subsequently, we insert another GNHB module at the end of the Backbone to further enhance the network's

ability to extract higher-order information. The C3-ACEL-YOLOv7t scheme is our proposed improved head network based on multi-scale higher-order information interaction. Combining the above three makes up our improved algorithm called YOLOv7-GN.

The results of the ablation experiments are presented in Table 4. From Table 4, we can see that the YOLOv7-tiny algorithm has lower F1 scores, mAP, and other object detection metrics. Simultaneously, there is still potential for further compression of model parameter counts, GFLOPs, and the size of the trained model. ACEL-YOLOv7t introduces our lightweight ELAN module for the first time compared to the original algorithm, resulting in decreased model parameter counts, GFLOPs, and Model Size. Crucially, while the model is lightweighted, the integration of self-attention and convolution through ACmix cleverly improves global dependency modeling. ACEL-YOLOv7t improves mAP by 0.55% and reduces the number of model parameters by 0.32.

In the C3-YOLOv7t scheme, we employ recursive gated convolution at the key connection location of the Backbone and feature pyramid (FPN), providing higher-order feature information of different sizes. This simple design brings a 1.23% performance improvement to the mAP of the network and increases the F1 score by 0.61%. However, the size of the network is increased. For the C3-ACEL-YOLOv7t scheme, compared to the previous scheme, the improved head network based on multi-scale higher-level information interaction achieves a more significant performance improvement. It increases mAP by 1.42%, P by 2.25%, and F1 by 1.65%. The model parameters are reduced by 0.3 million while taking advantage of the higher-level spatial interaction with gated convolution and recursive design.

The ACC3-YOLOv7t scheme demonstrates the performance of the lightweight higher-order attention layer aggregation network. ACC3-ELAN enables Backbone to possess three simultaneous capabilities: an input adaptive weight generation strategy, long-distance spatial modeling capability, and higher-order spatial interaction. This scheme increases R, F1, and mAP by 2.26%, 1.88%, and 1.46%, respectively, while reducing the number of model parameters by 0.69 million and GFLOPs by 1.5. Notably, ACC3-YOLOv7t achieves the best recall (R) and GFLOPs among all scenarios, illustrating that the lightweight higher-order attentional layer aggregation network (ACC3-ELAN) efficiently reduces the calculated load and captures most actual positive samples, minimizing the chances of missing the target.

YOLOv7-GN is a fusion of the previously mentioned schemes, and in comparison with the baseline, YOLOv7-GN achieves significant improvements. Specifically, YOLOv7-GN enhances precision (P) by 3.03%, recall (R) by 0.91%, F1 score by 1.87%, and mean Average Precision (mAP) by 1.99%. Moreover, the number of model parameters and GFLOPs are reduced by 0.78 million and 1.3, respectively. The model size is also decreased by 0.67 MB. Compared to the baseline, YOLOv7-GN showed better performance in terms of accuracy, recognition speed, and model size.

To further characterize the enhancement effect of our method on the original model, we compare YOLOv7-GN with the baseline model's Precision–Recall curves, as illustrated in Figure 12. This curve illustrates the trade-off between precision (Precision) and recall (Recall) at varying confidence thresholds, with the area between the curve and the axes representing the Average Precision (AP). The results show that the AP of YOLO-GN is higher, and the overall performance is better, than that of the benchmark model.

We additionally conducted a comparison of the F1–Confidence curves for both models, as depicted in Figure 13. Combining the F1 Score and Confidence in the object detection task, the F1–Confidence curve serves as a metric for evaluating the model's performance. The curve illustrates the correlation between the F1 Score and Confidence across various confidence thresholds, facilitating the analysis of the model's performance under different confidence levels. Experimental results indicate that YOLOv7-GN exhibits superior recall compared to the benchmark model across most confidence levels. This suggests that YOLOv7-GN strikes a balance between Precision and Recall, enhancing robustness, especially in scenarios with an imbalanced distribution of samples in the target category.

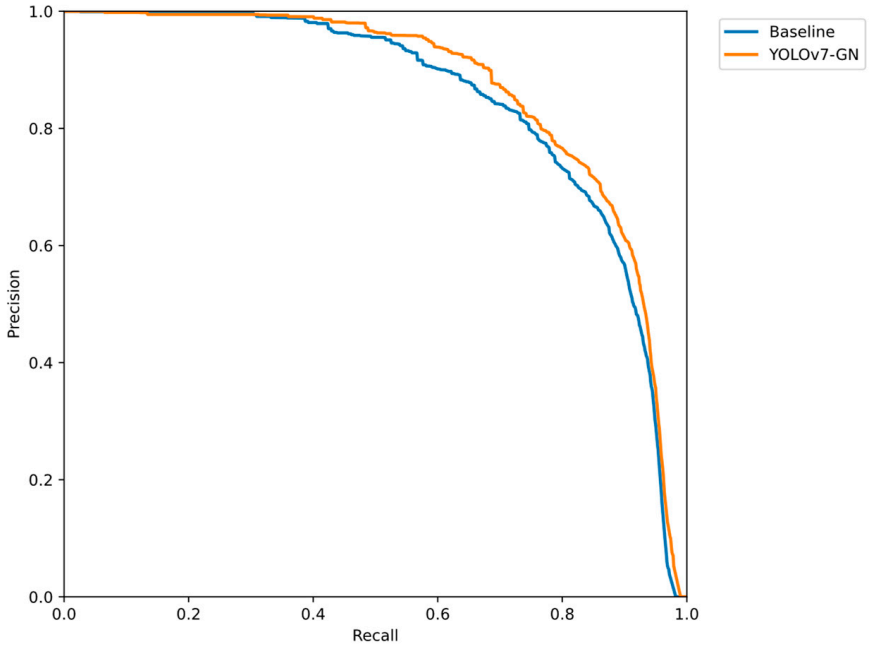

Figure 12. P–R curves of YOLOv7-GN compared to baseline.

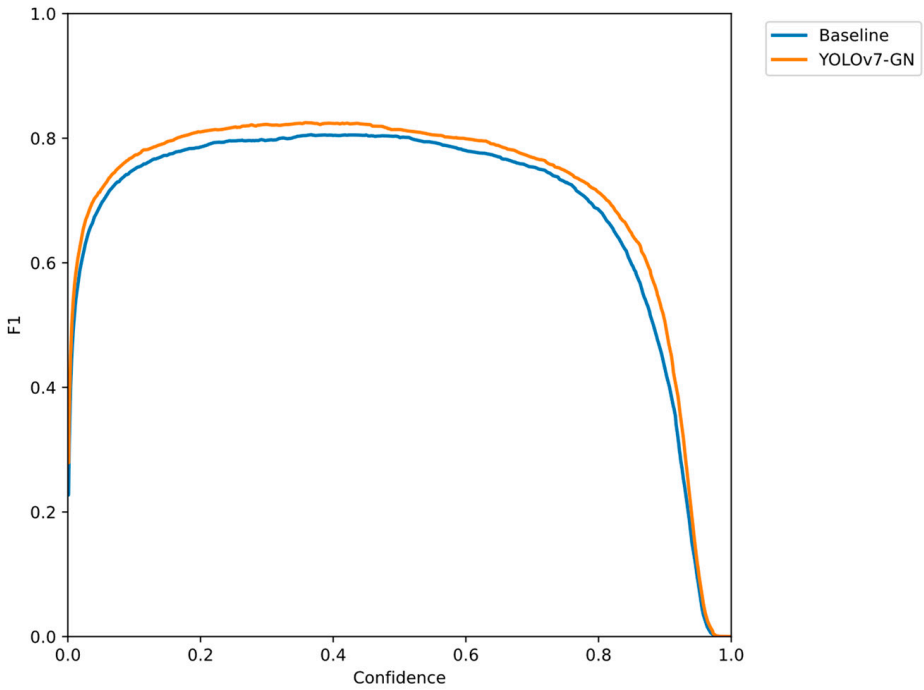

Figure 13. F1–Confidence curves of YOLOv7-GN compared to the baseline model.

4.4. Underwater Image Enhancement

In this section, we examine the impact of diverse image enhancement methods on underwater image enhancement and compare traditional image enhancement methods with the FUnIE-GAN method. Initially, we illustrate the effects of various image enhancements through subjective evaluation. Subsequently, we employ two non-reference evaluation metrics to assess the enhancement effect objectively. Finally, we implement the optimal underwater image enhancement scheme in YOLOv7-GN and assess the performance of the enhanced version.

4.4.1. Evaluation of Underwater Image Enhancement

We employed various common underwater image enhancement methods for comparative analysis, namely Gamma Correction (GC), CLAHE [72], HE [73], ICM [74], Rayleigh Distribution (RD) [75], RGHS [48], UCM [76], and FUnIE-GAN. Table 5 presents a comparison of the definitions, advantages, and disadvantages of common image enhancement methods. Illustrated in Figure 14, we selected four scenarios with distinct challenges from the DUO dataset as test samples, encompassing green, turbid, and low-visibility scenarios, respectively.

Table 5. Comparison table of common image improvement methods.

Method	Definitions	Advantages	Drawbacks
GC	Apply a power transformation to every gray value of the image	Easy to use and performs effectively in low-light conditions	Susceptible to distortion in high-contrast images
HE [73]	Uniform distribution of the grayscale histogram	Effectively increases image contrast without relying on parameter settings	Following the homogenization process, dark areas may exhibit noise
CLAHE [72]	Optimization of image block-to-block transition problem in HE method using linear interpolation approach	Enhanced naturalness of images and noise suppression	For images with low global contrast, the effect may not be as good as global histogram homogenization
ICM [74]	Stretching the contrast of the RGB algorithm and the saturation and intensity stretching of the HIS	Equalizes the color contrast in the images and also address the problem of lighting.	Parameters need to be adjusted for different underwater environments
RGHS [48]	A shallow water image enhancement method based on adaptive parameter acquisition	Adaptive to different image conditions	Fine-tuning parameters is necessary for complex environments
UCM [76]	An unsupervised color correction method	Efficiently removes the bluish color cast and improve the low red color	The enhancements exhibit instability and incur high computational costs
RD [75]	The histograms of the RGB and HSV models of the image are adjusted and fused separately	Effectively improves image contrast and reduces the effect of blue-green color difference	Oversaturation may occur in some scenarios

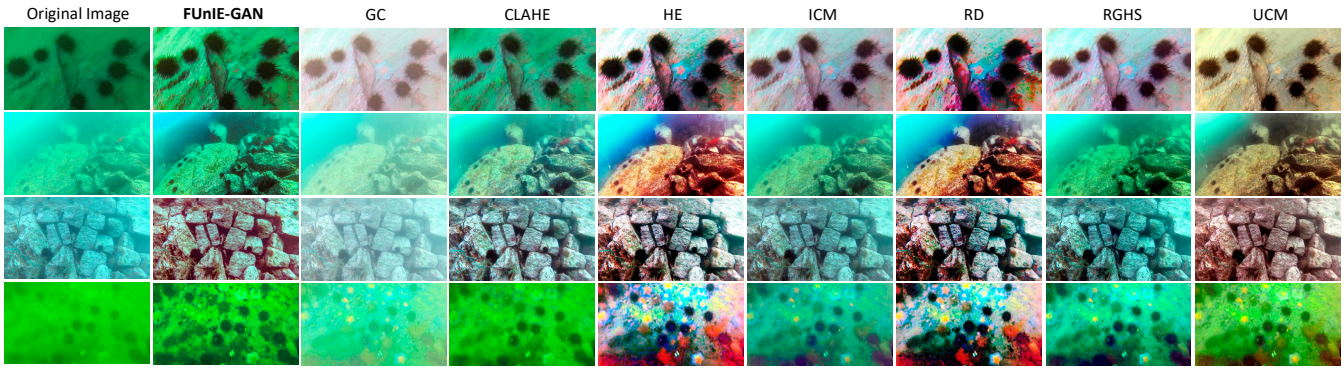

Figure 14. Comparison of common underwater image enhancements.

From the results depicted in Figure 14, it is evident that the FUnIE-GAN method attains a more realistic color restoration of underwater images while maintaining a certain level of clarity during dehazing. The GC method enhances the overall brightness of the photo and is effective in high-contrast scenes between foreground and background, but its overall enhancement effect is relatively weak. The CLAHE, ICM, and RGHS methods all achieve varying levels of dehazing and color restoration on the original image, with RGHS

performing the best in the fourth group of photos by naturally eliminating greenish hues. The HE and RD methods achieve strong color restoration on underwater images, resulting in images with richer colors than the real-world images and serious purple fringing around the objects. The UCM method performs comparably to FUnIE-GAN, but FUnIE-GAN outperforms in terms of clarity and color restoration. In the second group of photos, FUnIE-GAN successfully restores the blue ocean in the background, while UCM exhibits a brownish-green hue. Overall, the FUnIE-GAN method achieves the most balanced performance, with clear distinctions between targets and backgrounds in the enhanced images and restoration of more realistic textures and color features.

Then, we employed two non-reference evaluation metrics, UCIQE and NIQE, to objectively assess these various image enhancement methods. A random image from the dataset was selected as the original image, and the score of this image served as the baseline. The final results of the objective evaluation are presented in Table 6.

Table 6. The UCIQE and NIQE scores of different image enhancement methods. The numbers in parentheses represent the amount of change compared to the baseline. The bolded parameters are optimal.

Method	UCIQE	NIQE
Original	0.4766	12.877
GC	0.4487 (−0.0279)	12.666 (−0.211)
CLAHE	0.5567 (+0.0801)	13.143 (+0.266)
HE	0.6139 (+0.1373)	12.924 (+0.047)
ICM	0.5255 (+0.0489)	13.178 (+0.301)
RGHS	0.5779 (+0.1013)	12.833 (−0.044)
UCM	0.5912 (+0.1146)	13.317 (+0.44)
RD	**0.6251 (+0.1485)**	11.294 (−1.583)
FUnIE-GAN	0.6035 (+0.1269)	**15.691 (+2.814)**

From the results in Table 6, it is evident that, except for GC and RGHS, all other methods exhibit improvement in image quality compared to the original images. FUnIE-GAN achieves the best performance in terms of the EIQE metric, with a score increase of 2.814 points compared to the original image. UCM obtains a score of 13.317, while HE and RD achieve scores of 12.924 and 11.294, respectively. Regarding the UCIQE evaluation metric, all methods, excluding GC, demonstrate improvement compared to the baseline. RD and HE receive scores of 0.6251 and 0.6139, respectively, while FUnIE-GAN achieves a score of 0.6035. It is noteworthy that RD and HE attain slightly higher scores in the UCIQE metric. We speculate that this might be associated with the evaluation criteria of UCIQE, which quantitatively assesses non-uniform color shifts, blurriness, and low contrast using a linear combination of chromatic standard deviation, saturation, and contrast. The higher saturation and contrast in RD and HE could contribute to their better performance in UCIQE. Furthermore, Figure 15 illustrates the detailed performance of FUnIE-GAN and HE methods on the fourth group of photos.

From Figure 15, it is evident that the images enhanced by the HE method display a broad spectrum of colors but also experience significant local color distortions. In the magnified detail image, it is apparent that the sea urchin has lost its original color and is even subjected to hue shifts, particularly noticeable in the bottom right corner. This clearly impacts on the object detection task.

In conclusion, we assess the image enhancement effects from both subjective and objective perspectives. The evaluation results unequivocally illustrate that the FUnIE-GAN underwater image enhancement method attains a more comprehensive enhancement effect.

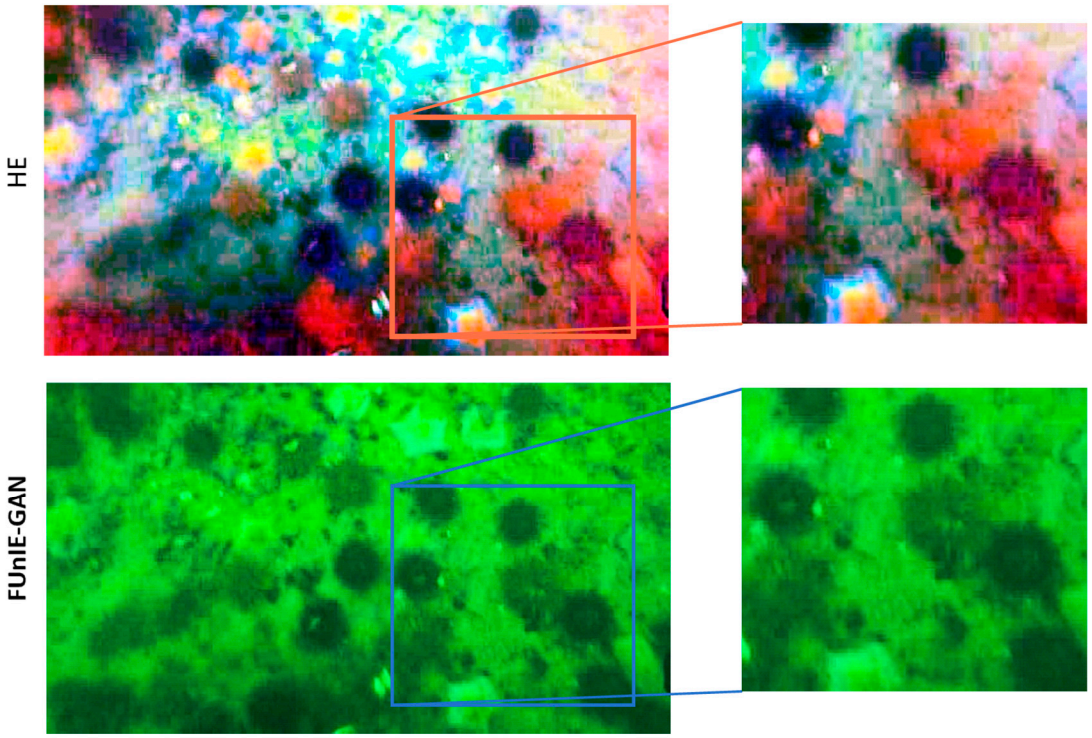

Figure 15. FUnIE-GAN and HE in photo details.

4.4.2. Evaluation of Object Detection Models after Underwater Image Enhancement

In order to investigate the potential impact of underwater image enhancement on object detection performance, we selected two relatively balanced methods, CLAHE and FUnIE-GAN, for testing. The enhanced datasets were incorporated into the YOLO-GN network during training, and the test results are presented in Table 7. CLAHE-YOLOv7-GN denotes the use of CLAHE as the underwater enhancement method, while FUnIE-YOLOv7-GN designates the utilization of FUnIE-GN as the underwater enhancement method. From the experimental outcomes, we found that the CLAHE method achieved the highest improvement in object detection accuracy, but there was a decrease in recall rate and other metrics. This could be attributed to the fact that, in certain scenarios, CLAHE may render it challenging to differentiate between objects and the background, consequently resulting in an elevated rate of missed detections. The FUnIE-GAN method showed improvements in recall rate, F1 score, and mAP, with a 2.64% increase in mAP50:95 and a 0.16% increase in mAP50. In comparing the results of CLAHE and FUnIE-GAN, it is apparent that image enhancement algorithms aim to improve the perceptual quality of images to the human eye. However, improvements in image evaluation metrics do not necessarily correlate directly with improvements in object detection performance. Additionally, diverse underwater image enhancement methods may necessitate fine-tuning for better adaptation to various object detection tasks in different scenarios.

Table 7. Effect of CLAHE and FUnIE-GAN image enhancement methods on object detection performance. The bolded parameters are optimal.

Model	mAP50 (%)	mAP50:95 (%)	R (%)	P (%)	F1 (%)
Original Image-YOLOv7-GN	86.76	65.18	77.98	88.13	82.74
CLAHE-YOLOv7-GN	85.84	64.53	75.31	**88.35**	81.31
FUnIE-YOLOv7-GN	**86.92**	**67.82**	**79.87**	87.56	**83.54**

Figure 16 employs the Class Activation Mapping (CAM) feature visualization technique to produce weighted heatmaps, enabling a thorough comprehension and comparison of the detection performance and decision-making procedures among different enhancement methodologies. From the figure, we can see that predicting on the original unprocessed image led to many false positives due to the network showing a high interest in regions where there were no objects. The use of FUnIE-GAN reduced this issue without losing any targets. Although the RD method performed better than FUnIE-GAN in terms of the UCIQE evaluation metric, its actual object detection performance was consistent with our analysis and prediction. The excessive color enhancement interfered with the network's judgment, making it difficult for the network to detect targets in the lower regions of test images. The YOLOv7-GN combined with the FUnIE-GAN paradigm was the final method proposed.

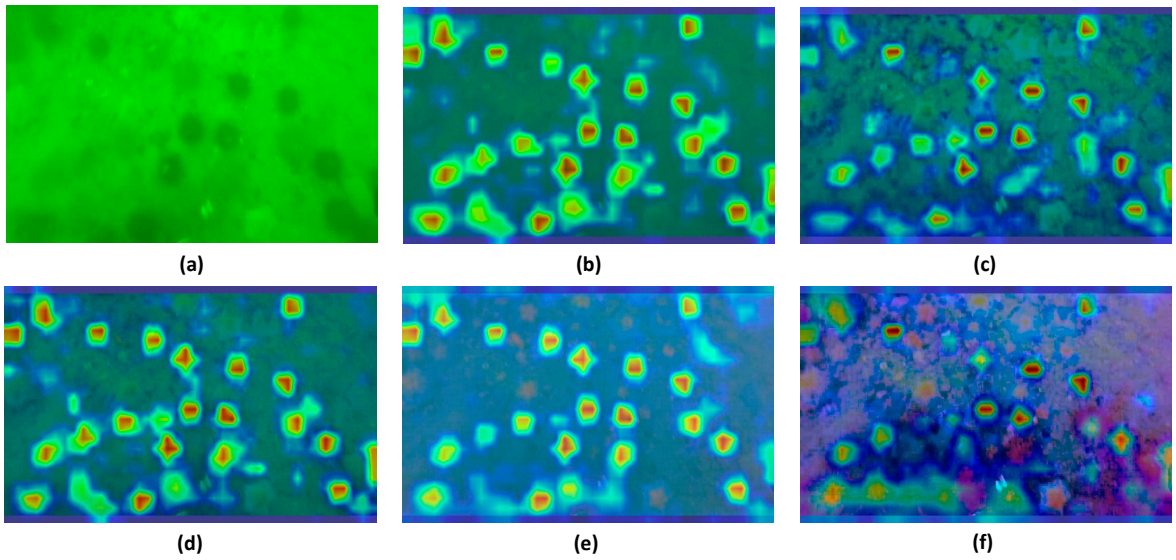

Figure 16. Comparison of attention maps trained by different underwater image enhancement methods. Visualization of the attentional map of the detector neck onto the enhanced input image using the CAM method. (**a**) Original input image. (**b**) Attention map of the original image. (**c**) FUnIE-GAN. (**d**) CLAHE. (**e**) GC. (**f**) RD. The darker color in the attention map represents the higher probability that the network thinks there is a target here, and the more interested the network is in this region.

4.5. Comparison with Other Object Detection Methods

To objectively assess the effectiveness of YOLOv7-GN, we compare our method with several state-of-the-art (SOTA) object detection methods in this section, including Faster R-CNN [18], the YOLOv5 series, and the latest YOLOv8 series. All experiments were conducted under identical environmental conditions and datasets, and the results are presented in Table 8. The visual analysis of the results is depicted in Figure 17.

The F1 score reflects the network's balanced performance in terms of precision and recall. YOLOv7-GN achieved an F1 score of 82.72, which is second only to the best result. Although YOLOv3 showed superior results in the F1 score, it experienced a 2.87% decrease in mAP50% compared to YOLOv7-GN, and the generated model size increased by 10.67 times. YOLOv7-GN demonstrated the best performance in terms of the mAP50% metric, with a 12.98% improvement over Faster R-CNN and a 2.88% improvement over YOLOv5s. It is noteworthy that YOLOv7-GN achieved these performance improvements with only 5.45 million parameters, fewer even than top-performing models such as YOLOv8s and YOLOv5s. In terms of network size comparison, YOLOv5s exhibited similar performance to YOLOv7-GN but significantly compromised accuracy.

Table 8. Comparison experiment. Bolded parameters are optimal.

Model	Backbone	F1 (%)	mAP50 (%)	Parameters (M)	GFLOPs	Model Size (MB) *
TOOD [77]	ResNet18	-	76.5	18.85	33.52	-
Faster R-CNN [78]	ResNet50	79.78	73.8	41.14	63.26	-
RetinaNet [79]	ResNet50	67.81	70.4	36.17	52.62	-
YOLOv3	DarkNet53	**83.41**	83.91	61.53	155.3	117.82
YOLOv5m	CSPDarkNet53	82.63	84.27	20.88	48.3	40.25
YOLOv5s	CSPDarkNet53	82.21	83.9	7.03	16.0	13.75
YOLOv8s	CSPDarkNet53(c2f)	82.55	86.12	11.17	28.8	21.49
YOLOv7-GN	ACC3-DarkNet53	82.72	**86.78**	**5.45**	**12.6**	**11.05**

* Model Size: Due to the different coding formats used to generate weights, only the Model Size of the YOLO series will be compared here.

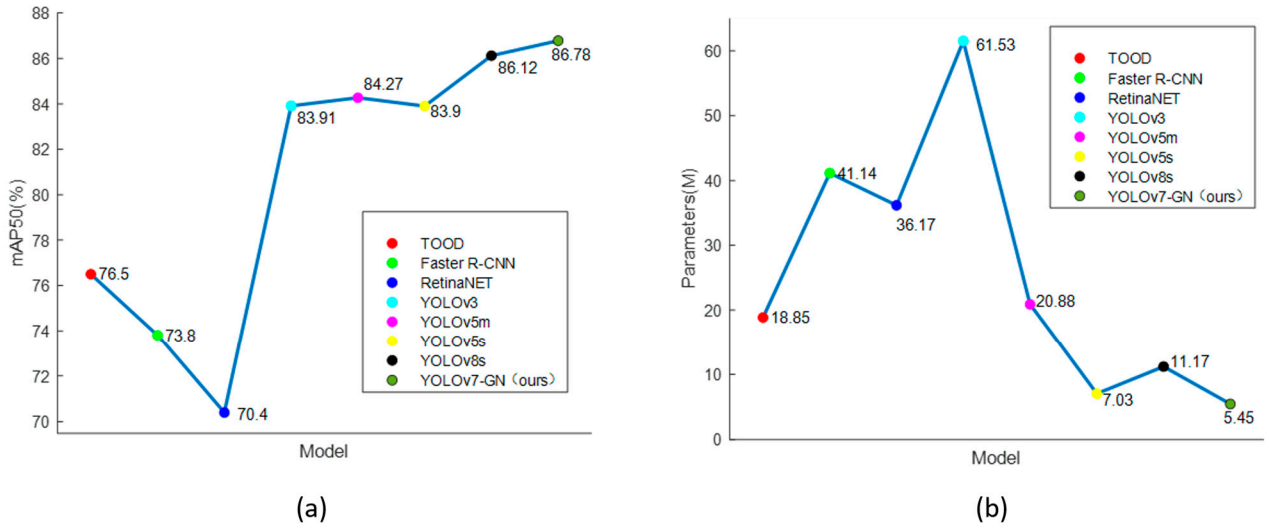

Figure 17. Comparison chart of comparative experiment results. (**a**) Comparison of mAP changes in different models. (**b**) Comparison of parameter changes in different models.

4.6. Visual Result Analysis

In this section, we visualize the detection results and further analyze the differences in our method's detection performance. Underwater object detection is affected not only by factors such as refraction and scattering of seawater, which can cause image blurring, but also by the clustering of underwater organisms. This can lead to multiple overlapping targets, and the varying distances between the captured images and the targets can result in multi-scale transformations, which pose a great challenge to underwater object detection models. We randomly selected six test images for visual comparison and further analysis of the detection results, as shown in Figure 18.

From the detection results, both models exhibit good performance when the image background is simple and the target stands out. However, in cases where target overlap occurs, such as in the first group of images, the baseline model experiences missed detections. When the target background is complex, YOLOv7-GN detects objects with higher confidence and demonstrates fewer instances of missed detections compared to the baseline model. We have achieved superior performance while compressing the model size, further validating the effectiveness of the improvements proposed in YOLOv7-GN.

To visually illustrate the advantages of YOLOv7-GN, we present the feature attention effects in the form of heatmaps, as depicted in Figure 19.

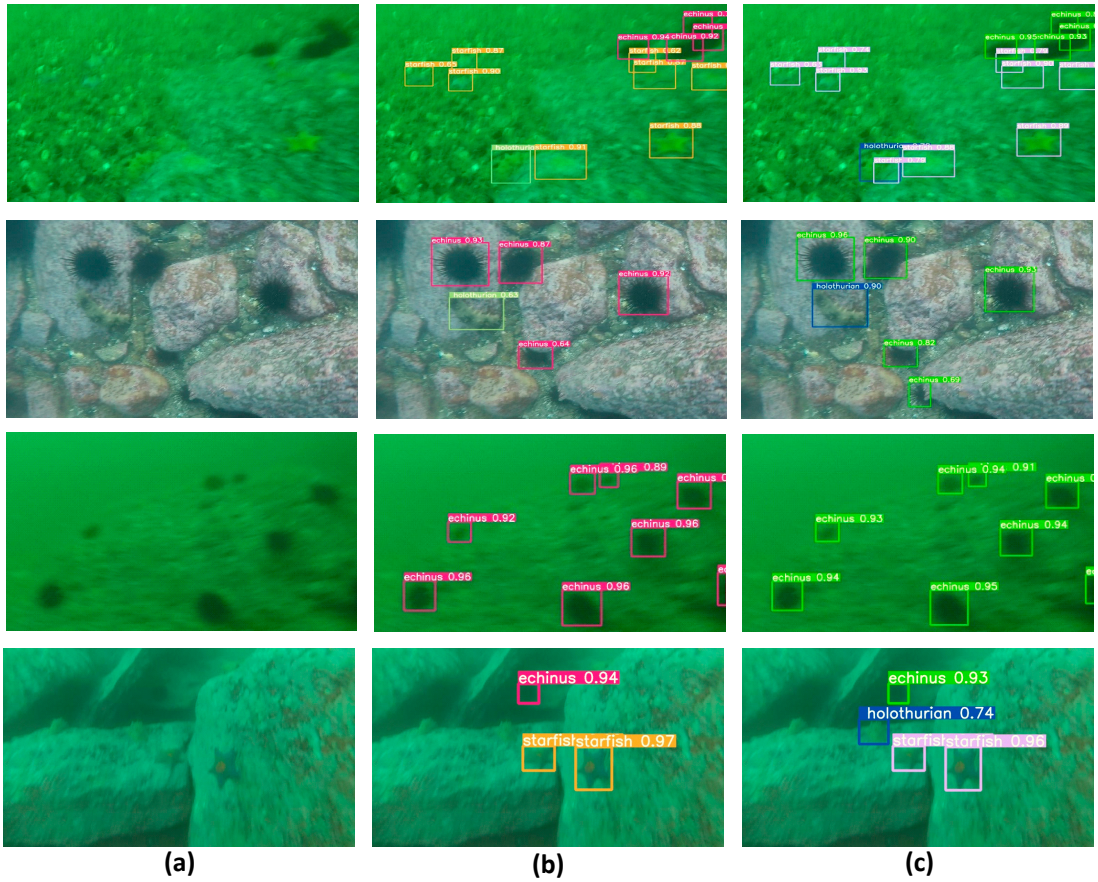

Figure 18. Detection results for the DUO dataset. The figure serves for qualitative analysis, with distinct colors of the object boxes denoting various categories. (**a**) The original input image. (**b**) The detection result of YOLOv7 (baseline). (**c**) The detection result of YOLOv7-GN (ours).

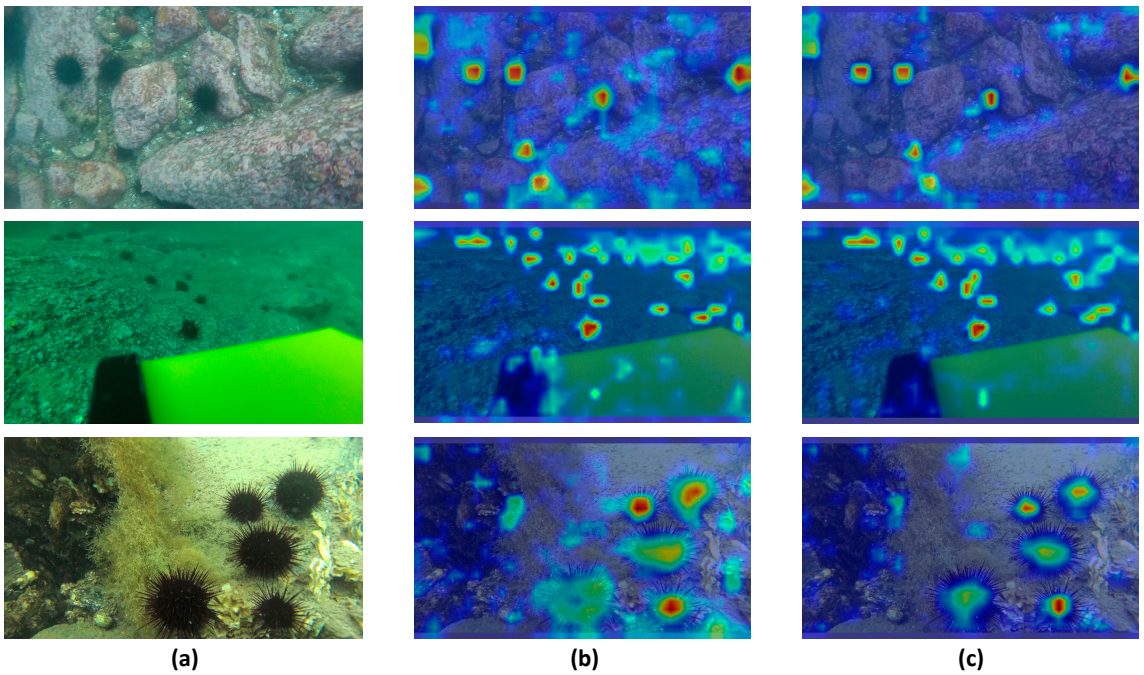

Figure 19. Visual heat maps of the output properties of YOLOv7-GN and YOLOv7 in different scenarios. (**a**) Input image, (**b**) YOLOv7, (**c**) YOLOv7-GN. Darker areas represent the more attention the network pays to them during recognition.

From Figure 19, it is evident that compared to YOLOv7, YOLOv7-GN significantly enhances the focus on the target region by incorporating a lightweight higher-order attention aggregation network. The network concentrates attention more on the core features of the target, effectively suppressing other noise and unrelated features in the image. This observation suggests that YOLOv7-GN exhibits enhanced robustness.

4.7. Embedded Deployed Results

To assess the feasibility of deploying our proposed lightweight model on embedded devices, we designed a biomimetic flatworm underwater robot to validate the algorithm's practical applications. Illustrated in Figure 20, this robot utilizes a reciprocating crank rocker mechanism as a power transmission component, employing a drivable flexible skeleton to transmit power. It also features drive fins that oscillate reciprocally, generating a stable waveform to propel the robot forward. We integrated a Jetson Nano [80] as the visual processing module at the tail of the biomimetic robot, connecting it to the camera module at the robot's head via a data cable, thereby establishing the visual system of the biomimetic robot.

After assembling the biomimetic robot based on the model, we evaluated the feasibility of deploying our method on embedded devices in an actual underwater environment. Initially, we implemented YOLOv7-GN on the Jetson Nano embedded host computer. Jetson Nano, a small and cost-effective artificial intelligence (AI) computer developed by NVIDIA, incorporates a high-performance NVIDIA Maxwell architecture GPU and a quad-core ARM Cortex-A57 processor known for its exceptional energy efficiency. This device supports various AI frameworks, including TensorFlow, PyTorch, and Caffe, enabling the development and deployment of deep learning models on edge devices. In comparison to the mainstream embedded deep learning development board Jetson TX2, Jetson Nano has 2.75 times less computing power, imposing more stringent requirements on model size deployment. Nevertheless, our method operates smoothly on Jetson Nano, providing a comprehensive visual system. As illustrated in Figure 21, we fabricated the biomimetic robot and conducted relevant underwater tests.

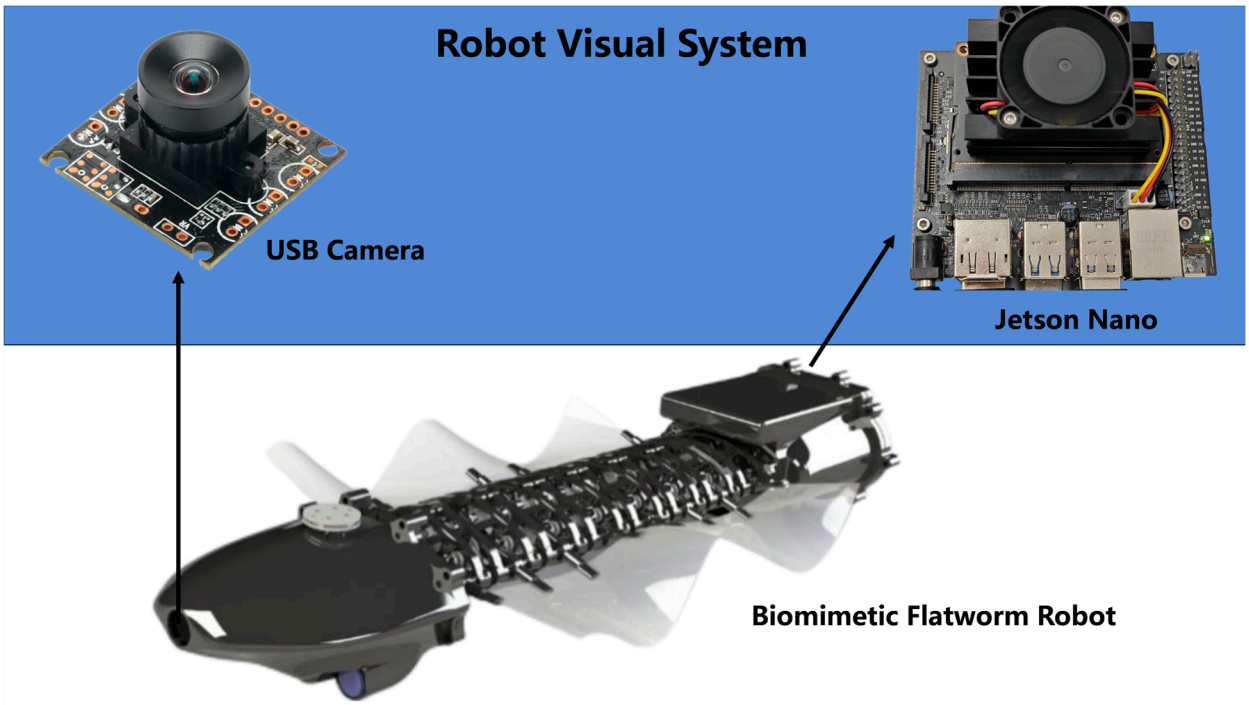

Figure 20. Biomimetic flatworm underwater robot vision system, consisting of a USB camera in the head and a Jetson Nano master control in the tail.

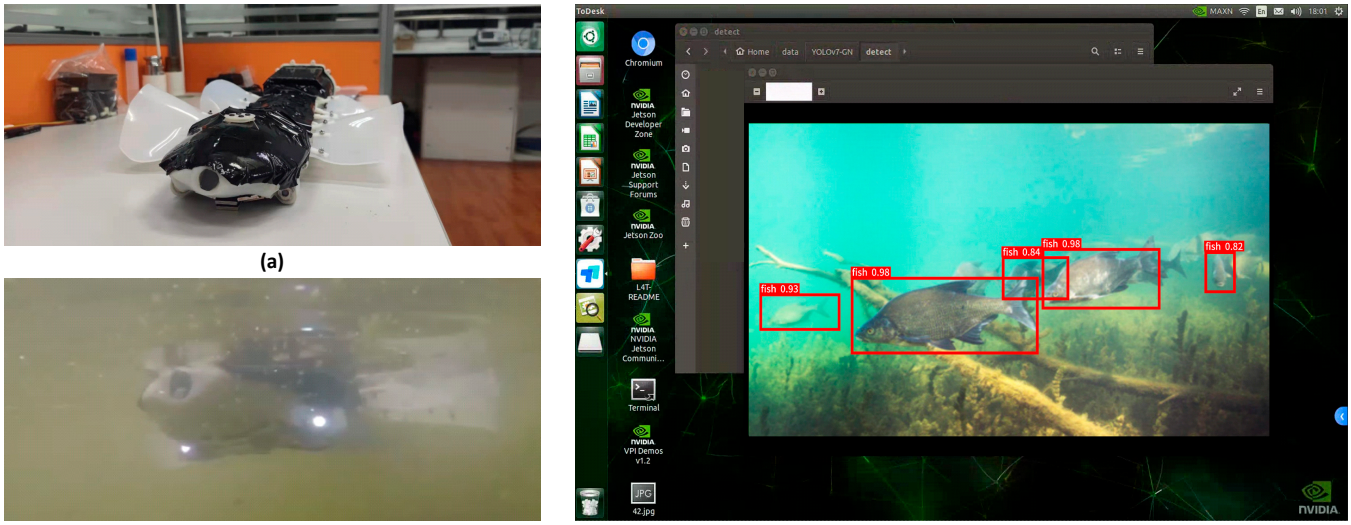

Figure 21. Real-world testing of the biomimetic flatworm robot. (**a**) Photograph of the robot. (**b**) The underwater working state of the robot. (**c**) Ubuntu recognition result interface.

In the tests, we compare mainstream embedded-oriented object detection models with YOLOv7-GN, including YOLOv7-tiny and YOLOv5s, on this robot. Their corresponding scheme EM-YOLOv7-tiny, EM-YOLOv5s, and EM-YOLOv7-GN. The results are presented in Table 9. Besides the key metrics, the table displays the accuracy for each category. According to the experimental results, EM-YOLOv7-GN achieves the highest mAP and boasts the smallest parameters and GFLOPs. EM-YOLOv5s exhibits the highest AP value for object recognition in both Echinus and Starfish categories, while EM-YOLOv7-tiny demonstrates a more balanced performance. These results demonstrate the superior balance between accuracy and network size that EM-YOLOv7-GN provides on embedded platforms.

Table 9. Comparative experiments on embedded platforms. The bolded parameters are optimal.

Model	mAP (%)	AP (%)				Parameters (M)	GFLOPs
		Holothurian	Echinus	Scallop	Starfish		
EM-YOLOv7-tiny	78.5	80.1	88.1	57.5	88.4	6.23	13.9
EM-YOLOv5s	77.7	80.8	**91.3**	48.3	**90.3**	7.03	16.0
EM-YOLOv7-GN	**81.0**	**83.1**	88.7	**63.1**	89.1	**5.45**	**12.6**

In conclusion, we have conducted both theoretical verification and practical deployment testing of the algorithm. Through the final practical validation, we have seen the possibility of actual application implementation of YOLOv7-GN, rather than just staying at data comparison and theoretical analysis.

5. Discussion

In the domain of underwater object detection, relying solely on 2D convolutions is insufficient for capturing deeper feature information due to substantial environmental variations and diverse target scales and morphologies. The utilization of global attention to capture features would not only escalate the training model's cost but also substantially increase the overall network size, contradicting our initial goal of developing lightweight networks deployable on embedded devices [55,56]. We assert that the key to achieving superior performance with YOLOv7-GN lies in the paradigm of combining convolution and self-attention. This approach not only integrates high and low-frequency information but also strikes a balance between speed and accuracy. Moreover, through network structure

design, we have empowered the network to enhance its fusion perception capability of multi-scale higher-order features without an increase in the number of parameters.

In this study, we completed the deployment of the proposed YOLOv7-GN on an underwater robot for practical testing. During the testing phase, we observed limitations in our method. Firstly, the initialization of the model takes longer due to hardware performance constraints. Secondly, the camera transmits jittery images during actual operation due to the robot's motion mode, posing challenges for the object detection model. The robot scheme discussed in the paper represents our preliminary design, and we plan to enhance both the hardware and software aspects of our program for further improvements.

In our future work, we plan to explore the utilization of our method as a pre-task for object tracking, a prominent research area in computer vision. Object tracking involves continuously monitoring the position of an object in a video sequence and updating its location in response to movement, deformation, or occlusion. Object tracking is widely used in various fields such as remote sensing video target tracking, species conservation, etc. Many object tracking systems utilize object detection to initialize the tracker, providing the initial position and bounding box of the target. We hypothesize that employing our method as a pre-task for object tracking will not only enhance the performance of object tracking but also contribute to making the entire system more lightweight. In the upcoming research, we aim to validate this hypothesis.

Furthermore, we will consider the migration capability of our method in different application scenarios. In our practical research, we have identified several application scenarios in the object detection that pose challenges, such as color shifts in the original image, transformations of the target scale, and similarities between the target and the background. These scenarios include satellite remote sensing images and target detection in UAV aerial images. In UAV aerial image object detection, the color shift of the original image occurs due to the weather or to the noise source generated by the photosensitive element in the equipment. Most notably, UAVs are embedded devices like underwater robots, and the vision system of UAVs also requires a lightweight model to enable more sophisticated control and versatile functionality. Thus, we will extend the application of our method to the task of object detection in remote sensing images from UAVs and explore the possibility of its application in additional fields.

6. Conclusions

Complex underwater environments present significant challenges to object detection models, particularly for deployment in underwater UAVs, where lightweight and efficient models are essential. Existing object detection models like YOLOv7 and Faster R-CNN suffer from limitations such as poor robustness and high computational cost. This paper proposes an innovative approach to underwater object detection by leveraging deep learning techniques and image enhancement methods to address the aforementioned challenges. Underwater images are enhanced using FUnIE-GAN and fed into YOLOv7-GN. To reduce parameters and enrich global higher-order attention features, a lightweight ACC3-ELAN network based on ACmix and g^n Conv recursive gated convolution is introduced. Additionally, a multi-scale higher-order information interaction head network is proposed for robust feature extraction by fusing semantic information from different scales. For further efficiency, we also integrate a lighter AC-ELAN-t module into the head network. To assess the feasibility of deploying our proposed lightweight model on embedded devices, we conduct underwater tests using a biomimetic flatworm underwater robot equipped with Jetson Nano as the visual processing module. We compare the FUnIE-GAN image enhancement with various other methods to evaluate its effectiveness in enhancing the object detection model's performance. Through detailed experiments, we confirm the efficacy of FUnIE-GAN for image enhancement. For YOLOv7-GN, we conduct ablation experiments and compare it with other mainstream object detection models. Our method outperforms these common object detection models in terms of Mean Average Precision (mAP) and model size, achieving well-balanced performance between underwater object

detection accuracy and model size. In the future, the proposed method will be utilized as an initial task for target tracking, and we will explore its transferability to fields such as UAV aerial photography, thereby expanding the deployment and application of the algorithm presented in this paper.

Author Contributions: Conceptualization, C.L. and J.W.; data curation, J.W.; funding acquisition, C.L.; investigation, J.W. and J.H.; methodology, C.L., J.W., N.X. and T.Z.; project administration, C.L. and T.Z.; writing—original draft preparation, J.W.; resources, C.L., J.H. and T.Z.; software, J.W., N.X. and W.L.; visualization, J.W. and B.W.; supervision, C.L. and T.Z.; validation, B.W., N.X. and W.L.; writing—review and editing, J.W., W.L. and N.X.; formal analysis, J.W. and B.W. All authors have read and agreed to the published version of the manuscript.

Funding: The authors acknowledge the funding of the following science foundations: the College Students' Science and Technology Innovation Cultivation Project of Guangdong Province, China (pdjh2023a0404), the College Students' Innovation and Entrepreneurship Training Project of China (202211078115), the Science and Technology Planning Project of Guangzhou, China (202102010392), the Science and Technology Planning Project of Guangdong Province, China (2020A1414050067), Guangzhou City School Joint Laboratory Project (No. 2023A03J0120).

Data Availability Statement: The DUO Dataset used in this article can be downloaded at https://github.com/chongweiliu/DUO (accessed on 21 February 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Costanza, R. The ecological, economic, and social importance of the oceans. *Ecol. Econ.* **1999**, *31*, 199–213. [CrossRef]
2. Zhou, X.; Ding, W.; Jin, W. Chapter 17—Microwave-assisted extraction of lipids, carotenoids, and other compounds from marine resources. In *Innovative and Emerging Technologies in the Bio-Marine Food Sector*; Garcia-Vaquero, M., Rajauria, G., Eds.; Academic Press: New York, NY, USA, 2022; pp. 375–394.
3. Bryson, M.; Johnson-Roberson, M.; Pizarro, O.; Williams, S.B. True Color Correction of Autonomous Underwater Vehicle Imagery. *J. Field Robot.* **2016**, *33*, 853–874. [CrossRef]
4. Kim, H.G.; Seo, J.; Kim, S.M. Underwater Optical-Sonar Image Fusion Systems. *Sensors* **2022**, *22*, 8445. [CrossRef] [PubMed]
5. Ghafoor, H.; Noh, Y. An Overview of Next-Generation Underwater Target Detection and Tracking: An Integrated Underwater Architecture. *IEEE Access* **2019**, *7*, 98841–98853. [CrossRef]
6. Liu, K.; Liang, Y. Enhancement of underwater optical images based on background light estimation and improved adaptive transmission fusion. *Opt. Express* **2021**, *29*, 28307–28328. [CrossRef] [PubMed]
7. Jing, S.; Xuan, Z.; Chao, Z.; Yin Xu, B.; Hua, S. Research on key technologies of underwater target detection. In Proceedings of the Seventh Symposium on Novel Photoelectronic Detection Technology and Applications, Kunming, China, 5–7 November 2020; SPIE: Bellingham, WA, USA, 2021; p. 1176345.
8. Zhang, W.; Sun, W. Research on small moving target detection algorithm based on complex scene. *J. Phys. Conf. Ser.* **2021**, *1738*, 012093. [CrossRef]
9. Forsyth, D. Object Detection with Discriminatively Trained Part-Based Models. *Computer* **2014**, *47*, 6–7. [CrossRef]
10. Lowe, D.G. Distinctive Image Features from Scale-Invariant Keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
11. Bakheet, S.; Al-Hamadi, A. A Framework for Instantaneous Driver Drowsiness Detection Based on Improved HOG Features and Naïve Bayesian Classification. *Brain Sci.* **2021**, *11*, 240. [CrossRef] [PubMed]
12. Zhou, X.; Wang, D.; Krähenbühl, P. Objects as points. *arXiv* **2019**, arXiv:1904.07850.
13. Law, H.; Deng, J. Cornernet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750.
14. Vaswani, A.; Shazeer, N.M.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is All you Need. In Proceedings of the NIPS, Long Beach, CA, USA, 4–9 December 2017.
15. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
16. Shen, Z.; Liu, Z.; Li, J.; Jiang, Y.G.; Chen, Y.; Xue, X. DSOD: Learning Deeply Supervised Object Detectors from Scratch. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; pp. 1937–1945.
17. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
18. Girshick, R. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.

19. Zhou, H.; Huang, H.; Yang, X.; Zhang, L.; Qi, L. Faster R-CNN for Marine Organism Detection and Recognition Using Data Augmentation. In Proceedings of the International Conference on Video and Image Processing, Singapore, 27–29 December 2017; pp. 56–62.
20. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. *IEEE Trans. Pattern Anal. Mach. Intell.* **2020**, *42*, 386–397. [CrossRef]
21. Cai, Z.; Vasconcelos, N. Cascade R-CNN: Delving Into High Quality Object Detection. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6154–6162.
22. 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, Amsterdam, The Netherlands, 11–14 October 2016; pp. 21–37.
23. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
24. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**, arXiv:2207.02696.
25. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A Single-Stage Object Detection Framework for Industrial Applications. *arXiv* **2022**, arXiv:2209.02976.
26. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
27. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
28. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
29. Lou, H.; Duan, X.; Guo, J.; Liu, H.; Gu, J.; Bi, L.; Chen, H. DC-YOLOv8: Small-Size Object Detection Algorithm Based on Camera Sensor. *Electronics* **2023**, *12*, 2323. [CrossRef]
30. Xiu, L.; Min, S.; Qin, H.; Liansheng, C. Fast accurate fish detection and recognition of underwater images with Fast R-CNN. In Proceedings of the OCEANS 2015—MTS/IEEE, Washington, DC, USA, 19–22 October 2015; pp. 1–5.
31. Lin, W.-H.; Zhong, J.-X.; Liu, S.; Li, T.; Li, G. *RoIMix: Proposal-Fusion among Multiple Images for Underwater Object Detection*; IEEE: New York, NY, USA, 2019.
32. Qiao, W.; Khishe, M.; Ravakhah, S. Underwater targets classification using local wavelet acoustic pattern and Multi-Layer Perceptron neural network optimized by modified Whale Optimization Algorithm. *Ocean Eng.* **2021**, *219*, 108415. [CrossRef]
33. Zhang, M.; Xu, S.; Song, W.; He, Q.; Wei, Q. Lightweight underwater object detection based on yolo v4 and multi-scale attentional feature fusion. *Remote Sens.* **2021**, *13*, 4706. [CrossRef]
34. Bao, Z.; Guo, Y.; Wang, J.; Zhu, L.; Huang, J.; Yan, S. Underwater Target Detection Based on Parallel High-Resolution Networks. *Sensors* **2023**, *23*, 7337. [CrossRef]
35. Liu, K.; Sun, Q.; Sun, D.; Peng, L.; Yang, M.; Wang, N. Underwater Target Detection Based on Improved YOLOv7. *J. Mar. Sci. Eng.* **2023**, *11*, 677. [CrossRef]
36. Dulhare, U.N.; Hussam Ali, M. Underwater human detection using faster R-CNN with data augmentation. *Mater. Today Proc.* **2023**, *80*, 1940–1945. [CrossRef]
37. Joshi, R.; Usmani, K.; Krishnan, G.; Blackmon, F.; Javidi, B. Underwater object detection and temporal signal detection in turbid water using 3D-integral imaging and deep learning. *Opt. Express* **2024**, *32*, 1789–1801. [CrossRef]
38. Abdelkader, I.; El-Sonbaty, Y.; El-Habrouk, M. Openmv: A python powered, extensible machine vision camera. *arXiv* **2017**, arXiv:1711.10464.
39. Pan, J.; Zhu, Z.; Liu, X.; Yan, X. Design of Fire Alarm System Based on K210 and Deep Learning. In Proceedings of the 2021 3rd International Conference on Artificial Intelligence and Advanced Manufacture, Manchester, UK, 23–25 October 2022; pp. 768–772.
40. Mittal, S. A Survey on optimized implementation of deep learning models on the NVIDIA Jetson platform. *J. Syst. Archit.* **2019**, *97*, 428–442. [CrossRef]
41. Feng, X.; Jiang, Y.; Yang, X.; Du, M.; Li, X. Computer vision algorithms and hardware implementations: A survey. *Integration* **2019**, *69*, 309–320. [CrossRef]
42. Chiu, Y.C.; Tsai, C.Y.; Ruan, M.D.; Shen, G.Y.; Lee, T.T. Mobilenet-SSDv2: An Improved Object Detection Model for Embedded Systems. In Proceedings of the 2020 International Conference on System Science and Engineering (ICSSE), Kagawa, Japan, 31 August–3 September 2020; pp. 1–5.
43. Zhuo, S.; Zhang, X.; Chen, Z.; Wei, W.; Wang, F.; Li, Q.; Guan, Y. DAMP-YOLO: A Lightweight Network Based on Deformable Features and Aggregation for Meter Reading Recognition. *Appl. Sci.* **2023**, *13*, 11493. [CrossRef]
44. Chen, Y.; Zheng, B.; Zhang, Z.; Wang, Q.; Shen, C.; Zhang, Q. Deep Learning on Mobile and Embedded Devices: State-of-the-art, Challenges, and Future Directions. *ACM Comput. Surv.* **2020**, *53*, 84. [CrossRef]
45. Zhang, H.; Cissé, M.; Dauphin, Y.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. *arXiv* **2017**, arXiv:1710.09412.
46. Devries, T.; Taylor, G.W. Improved Regularization of Convolutional Neural Networks with Cutout. *arXiv* **2017**, arXiv:1708.04552.
47. Yun, S.; Han, D.; Oh, S.J.; Chun, S.; Choe, J.; Yoo, Y.J. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Seoul, Republic of Korea, 27–28 October 2019; pp. 6022–6031.

48. Huang, D.; Wang, Y.; Song, W.; Sequeira, J.; Mavromatis, S. Shallow-Water Image Enhancement Using Relative Global Histogram Stretching Based on Adaptive Parameter Acquisition. In Proceedings of the MultiMedia Modeling, Bangkok, Thailand, 5–7 February 2018; pp. 453–465.
49. He, K.; Sun, J.; Tang, X. Single Image Haze Removal Using Dark Channel Prior. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 2341–2353. [CrossRef] [PubMed]
50. Wang, Y.; Zhang, J.; Cao, Y.; Wang, Z. A deep CNN method for underwater image enhancement. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 1382–1386.
51. Wang, Y.; Guo, J.; Gao, H.; Yue, H. UIEC^2-Net: CNN-based underwater image enhancement using two color space. *Signal Process. Image Commun.* **2021**, *96*, 116250. [CrossRef]
52. Guo, Y.; Li, H.; Zhuang, P. Underwater Image Enhancement Using a Multiscale Dense Generative Adversarial Network. *IEEE J. Ocean Eng.* **2020**, *45*, 862–870. [CrossRef]
53. Islam, M.J.; Xia, Y.; Sattar, J. Fast Underwater Image Enhancement for Improved Visual Perception. *IEEE Robot. Autom. Lett.* **2020**, *5*, 3227–3234. [CrossRef]
54. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale. *arXiv* **2020**, arXiv:2010.11929.
55. Yu, A.W.; Dohan, D.; Luong, M.-T.; Zhao, R.; Chen, K.; Norouzi, M.; Le, Q.V. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv* **2018**, arXiv:1804.09541.
56. Pan, X.R.; Ge, C.J.; Lu, R.; Song, S.J.; Chen, G.F.; Huang, Z.Y.; Huang, G.; IEEE Computer Society. 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, 18–24 June 2022; pp. 805–815.
57. Shaw, P.; Uszkoreit, J.; Vaswani, A. Self-Attention with Relative Position Representations. In Proceedings of the North American Chapter of the Association for Computational Linguistics, Seattle, WA, USA, 10–15 July 2018.
58. Voita, E.; Talbot, D.; Moiseev, F.; Sennrich, R.; Titov, I. Analyzing Multi-Head Self-Attention: Specialized Heads Do the Heavy Lifting, the Rest Can Be Pruned. *arXiv* **2019**, arXiv:1905.09418.
59. Alzubaidi, L.; Zhang, J.; Humaidi, A.J.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.A.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef]
60. Chen, Y.; Dai, X.; Liu, M.; Chen, D.; Yuan, L.; Liu, Z. Dynamic convolution: Attention over convolution kernels. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 14–19 June 2020; pp. 11030–11039.
61. Rao, Y.; Zhao, W.; Tang, Y.; Zhou, J.; Lim, S.N.; Lu, J. HorNet: Efficient High-Order Spatial Interactions with Recursive Gated Convolutions. *arXiv* **2022**, arXiv:2207.14284.
62. Wang, C.-Y.; Liao, H.; Yeh, I.-H. Designing Network Design Strategies Through Gradient Path Analysis. *arXiv* **2022**, arXiv:2211.04800.
63. Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; Bengio, Y. Generative adversarial networks. *Commun. ACM* **2020**, *63*, 139–144. [CrossRef]
64. Gecer, B.; Ploumpis, S.; Kotsia, I.; Zafeiriou, S. GANFIT: Generative Adversarial Network Fitting for High Fidelity 3D Face Reconstruction. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 1155–1164.
65. Isola, P.; Zhu, J.-Y.; Zhou, T.; Efros, A.A. Image-to-Image Translation with Conditional Adversarial Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5967–5976.
66. Loshchilov, I.; Hutter, F. SGDR: Stochastic Gradient Descent with Warm Restarts. *arXiv* **2016**, arXiv:1608.03983.
67. Smith, S.L.; Kindermans, P.-J.; Ying, C.; Le, Q.V. Don't decay the learning rate, increase the batch size. *arXiv* **2017**, arXiv:1711.00489.
68. He, F.; Liu, T.; Tao, D. Control batch size and learning rate to generalize well: Theoretical and empirical evidence. In Proceedings of the 33rd International Conference on Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Curran Associates Inc.: New York, NY, USA, 2019; pp. 1143–1152.
69. Mittal, A.; Soundararajan, R.; Bovik, A.C. Making a "completely blind" image quality analyzer. *IEEE Signal Process. Lett.* **2012**, *20*, 209–212. [CrossRef]
70. Yang, M.; Sowmya, A. An underwater color image quality evaluation metric. *IEEE Trans. Image Process.* **2015**, *24*, 6062–6071. [CrossRef]
71. 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.
72. Reza, A.M. Realization of the Contrast Limited Adaptive Histogram Equalization (CLAHE) for Real-Time Image Enhancement. *J. VLSI Signal Process. Syst. Signal Image Video Technol.* **2004**, *38*, 35–44. [CrossRef]
73. Hummel, R. Image enhancement by histogram transformation. *Comput. Graph. Image Process* **1977**, *6*, 184–195. [CrossRef]
74. Kashif, I.; Salam, R.A.; Azam, O.; Talib, A.Z. Underwater Image Enhancement Using an Integrated Colour Model. *IAENG Int. J. Comput. Sci.* **2007**, *34*, 239–244.
75. Abdul Ghani, A.; Mat Isa, N. Underwater image quality enhancement through composition of dual-intensity images and Rayleigh-stretching. *SpringerPlus* **2014**, *3*, 757. [CrossRef] [PubMed]

76. Iqbal, K.; Odetayo, M.O.; James, A.E.; Salam, R.A.; Talib, A.Z. Enhancing the low quality images using Unsupervised Colour Correction Method. In Proceedings of the 2010 IEEE International Conference on Systems, Man and Cybernetics, Istanbul, Turkey, 10–13 October 2010; IEEE: New York, NY, USA, 2010.
77. Feng, C.; Zhong, Y.; Gao, Y.; Scott, M.R.; Huang, W. Tood: Task-aligned one-stage object detection. In Proceedings of the 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal, BC, Canada, 11–17 October 2021; pp. 3490–3499.
78. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *arXiv* **2015**, arXiv:1506.01497. [CrossRef]
79. Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988.
80. Kurniawan, A.; Kurniawan, A. Introduction to nvidia jetson nano. In *IoT Projects with NVIDIA Jetson Nano: AI-Enabled Internet of Things Projects for Beginners*; Apress: Berkeley, CA, USA, 2021; pp. 1–6.

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.