A Lightweight Sea Urchin Detection Algorithm Based on Improved Yolov8

Shengli Chen, Xiaohu Shi *

College of Computer Science and Technology, Jilin University, Changchun, China, 130012

* Corresponding Author Email: shixh@jlu.edu.cn

Abstract. In complex underwater environments with rocky reefs or gravelly substrates, an efficient sea urchin and starfish detection method is critical to replace traditional artificial inspection methods. This paper presents a sea urchin and starfish detection method based on an enhanced YOLOv8 (You Only Look Once) model. The backbone and head networks of the model utilize C3Ghost to reduce model size and computation, and an EMA attention mechanism is employed to improve recognition accuracy. The results show that the proposed Yolov8n-GNA model achieves an accuracy of 86.7%, an 11.3% improvement over the original Yolov8n model while maintaining similar recall and mean Average Precision (mAP) values. Furthermore, the computational effort is reduced to 5.4 GFLOPS, representing a 33% reduction compared to the original model. The Yolov8n-GNA model can rapidly and accurately detect sea urchins and starfish in challenging underwater environments, making it suitable for sea urchin aquaculture and fishing operations in complex marine habitats.

Keywords: Network Optimization, Deep Learning, Underwater Target Detection.

1. Introduction

The ocean is teeming with biological resources, and seafood fishing is a primary method for harnessing marine resources. However, in recent years, the conventional seafood fishing method has encountered several challenges. These include low efficiency, high labor intensity, potential damage to the marine ecological environment, and high labor costs. Sea urchin, an important seafood in China, is rich in both nutritional and medicinal value.

As of 2022, sea urchin culture in China covered an area of 10,710 hectares, yielding 5,154.819 tonnes. The market demand for sea urchins is on the rise. The traditional method of sea urchin fishing relies heavily on human visual perception to spot and capture the sea urchins in the water. However, this method presents several challenges. For instance, it can be hard for the human eye to differentiate the sea urchins from the marine background, especially in rocky reef or gravelly substrate environments. Moreover, the manual approach is time-consuming and requires significant manpower. Since the introduction of target detection algorithms in 1998, computer vision technology has offered alternative solutions for underwater target detection. Yanbang Zhang et al. [1] improved the salient target detection algorithm by fusing texture and colour features, and the experimental results show that the algorithm outperforms several other algorithms in terms of comprehensive performance on the MSRA and ECSSD datasets, displaying strong applicability and robustness. SUSANTOT et al. [2] used color features to detect underwater targets, and detection mainly uses RGB color detection and HSV color detection, but the results show that the accuracy rate is low in the case of insufficient or excessive lighting conditions. XIE Tao et al. [3] proposed a daytime sea fog recognition algorithm based on multiple grey scale covariance matrix eigenvalues for sea fog monitoring in the Yellow Sea and Bohai Sea waters, which effectively improves the recognition accuracy of sea fog by analysing texture features in satellite images. Jiakang Li and Shengmao Zhang et al. [4] proposed an approach based on image processing techniques for accurately identifying and analyzing calibrated color cards in aquaculture water bodies to monitor and assess water quality changes effectively.

Traditional computer vision algorithms emphasize feature description vector extraction more, which is simple to implement and has high computational efficiency. However, the robustness is poor, so the image processing ability is poor in complex environments. The rapid development of deep learning and improved device capabilities have given computer vision a new method. Deep learning-

based target detection algorithms are mainly divided into One-stage algorithms and Two-stage algorithms [5], of which One-stage algorithms primarily include: Overfat, YOLO series algorithms, SSD and Retina-Net, etc., which can directly extract features from the input image through convolutional neural networks, compared with the Two-stage algorithms. stage algorithm, the detection speed is greatly improved and can be used for real-time target detection. Li et al. [6] used Faster R-CNN to detect fish in the image, and the greedy algorithm and Hungarian algorithm will correlate the results of the detection of each frame to achieve the swimming fish tracking. Fang Shu et al. [7] studied the spotted catfish (Ictalurus punctatus) based on the HourglassNet network and obtained phenotypic data such as the body length and height of a single fish. Zhao et al. [8] used a composite backbone network based on Cascade R-CNN to optimize the fish identification and localization algorithm with better results. At present, YOLO (You Only Look Once) series of algorithms also have a wide range of applications in the field of underwater target detection [9,10], such as Xu Ruifeng et al. [11] used the optical flow method and GLCM to extract the motion characteristics of shrimp video clips and improve the YoloV8 algorithm to detect shrimp morbidity. xianyi Zhai et al. [12] added a detection layer and added CBAM attention mechanism to improve YoloV8 algorithm. and add a CBAM attention mechanism to improve YoloV5 algorithm for sea cucumber recognition. Deep learning-based neural networks generalize better and detect better in complex environments, and the models have good migration capabilities.

Song Xiaolu et al. [13] preprocessed the dataset with a dark-channel Apriori algorithm and adopted the improved YOLOv4 algorithm with the Swish function as the activation function for sea cucumber detection. The algorithm demonstrates that the YOLO series supports real-time monitoring of sea cucumbers while having the ability to migrate to other echinoderms such as sea urchins and starfishes. By reviewing the relevant domestic and international literature, no studies on real-time detection of sea urchins have been reported. This study aims to investigate an improved YoloV8-based sea urchin, starfish detection method that can strike a balance between computational effort and detection effectiveness. By adding GhostNet and EMA attention mechanism, the Yolov8n-GNA model is obtained. The Yolov8n-GNA model targets complex seabed environments such as rocky reefs and gravelly substrates, improving the detection algorithm's robustness and accuracy under such conditions. In addition, the improved detection method can help to promote the sustainable use of marine biological resources, and at the same time has a positive significance for marine ecological environment protection, which is in line with the current international trend and domestic policy requirements for marine ecological environment protection.

2. Materials and methods

2.1. Datasets

2.1.1 Real-world-Underwater-Image dataset

The Realworld-Underwater-Image dataset [14] was gathered by Risheng Liu's team in the waters near Zhangzi Island in the Yellow Sea, situated at 39.186°N latitude and 44.625°E longitude. This area's water is significantly influenced by tides, exhibiting fluctuations from 5 to 9 meters. The dynamic changes in water depth create abundant natural experimental conditions for examining the performance of underwater light and color. The dataset's imagery captures the natural ecology of the Yellow Sea waters, showcasing diverse marine organisms such as fish, sea urchins, sea cucumbers, and scallops. These images are accompanied by detailed labels, supporting underwater target detection and classification tasks.

2.1.2 Dataset processing

In this study, for the scientific goal of real-time monitoring of sea urchins and starfish, the Make Sense online labeling tool was used in this research to label 311 underwater images in the Realworld-Underwater-Image dataset, and an example of the labeling is shown in Figure 1. The labeling process involves the precise delineation of the bounding boxes of starfish and sea urchins in the images and

the categorical labeling of their types. This operation provides the supervisory signals required by the model to ensure that the model is able to identify and classify the target objects accurately. After completing the labeling, following the principle of random samples, these images were divided into a training set, a cross-validation set, and a test set. Among them, 239 images were included in the training set for the initial learning of the model; 40 images constituted the cross-validation set for the evaluation of the model's performance and the adjustment of hyper-parameters; and the remaining 32 images were used as the test set for evaluating the model's detection accuracy and robustness in real-world application scenarios.

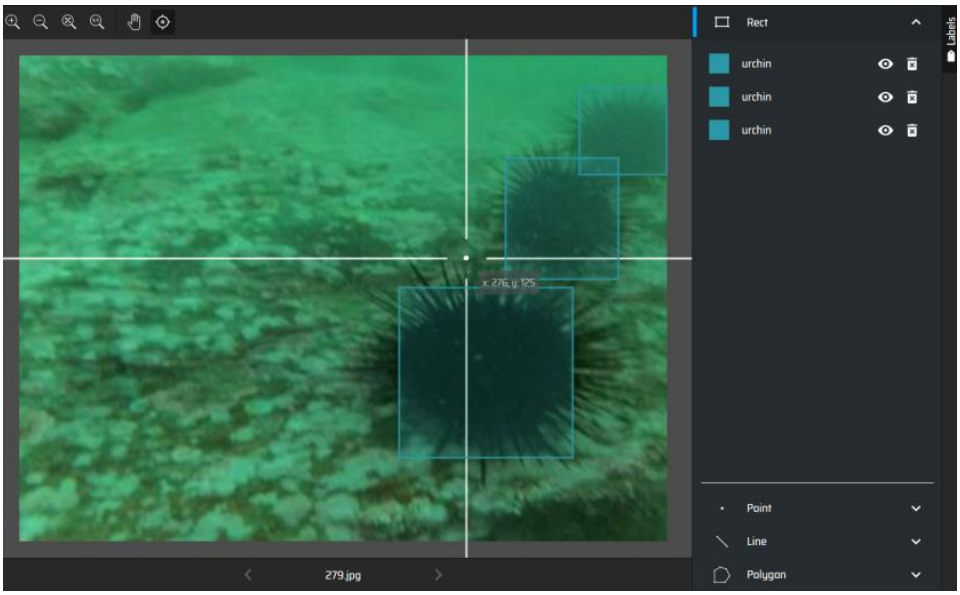

Fig. 1 Online annotation using Make-Sense

2.2. Overview of Yolov8n-GNA model

YOLOv8 [15] is a target detection network based on the PyTorch framework released in 2023, which can be used for object detection, image classification, and instance segmentation tasks, and in this paper, it is used for the object detection task. According to the network depth and width, YOLOv8 is divided into five model structures. As the depth and width of the model increase, the number of parameters and the computational amount of each model also increases. The YOLOv8 network structure consists of three parts: the Backbone network, the Neck structure, and the Head structure. The backbone consists of CBS, C2f, and SPPF modules to achieve multi-scale feature information integration and enhance the feature extraction capability; Neck achieves the simultaneous extraction of low-dimensional visual features and high-dimensional semantic features through Con‑cat and cross-layer fusion of down sampling after up-sampling; Head adopts the mainstream Decoupled-Head structure to integrate the classification and semantic features; and Head adopts the mainstream Decoupled-Head structure to separate the classification and detection heads, and calculating the Classification Loss and bounding Box Loss through two CBS modules and Conv2d respectively. Based on the demand for embedding small underwater electronic devices under actual industrial production, the improved network in this paper should meet the requirements of lower parameter count and computation to improve the detection speed and deployment performance of the network on low-computing power platforms. Therefore, YOLOv8n, which has the smallest model size, is adopted as the base model framework.

YOLOv8 considers the detection accuracy, speed, and multi-scale feature fusion of the detection model, but faces the problems of a complex underwater environment, much noise as well as low pixels, so Yolov8 cannot fully meet the requirements of underwater small target detection. Aiming at the above problems, this paper takes Yolov8n as the base model and optimizes the model in terms of lightweight and attention mechanism. The algorithm in this paper can effectively separate sea urchins and starfish from the background environment of complex rocky reefs or gravel substrate, and at the

same time detect and classify the specific location, to achieve the purpose of real-time monitoring, and the improved Yolov8n network structure in this paper is shown in Figure 2. The red module is the part of the model modified in this paper, and the inside of the surrounding box is the expansion description of the corresponding module. The improved Yolov8n network framework mainly contains four parts: the input side, the backbone network, the Head side, and the output side.

To reduce the size of the dataset, the input side sets the adaptive settings for the image data: no matter what the size of the image in the dataset is, the input side will automatically scale the image to the size of 640×640 with the standard of the short side, which ensures that the input data remains consistent and speeds up the model fitting speed. After the standardization of the data, the image data will be uniformly input into the backbone network. At the Head side, to ensure that the model algorithm focuses on the detection of small targets, the EMA attention mechanism is introduced. At the same time, GhostConv and C3Ghost in GhostNet are used to replace the Conv and C2f modules at the Backbone side, respectively, where Ghostconv is a lightweight convolution module that can replace the ordinary convolution in GhostNet, and C3Ghost effectively reduces the number of parameters and computational cost of the network, based on maintaining the size of the original convolutional feature maps and the size of the channels. C3Ghost is a lightweight convolutional module that can replace the normal convolutional method. After the feature learning operation of the C3Ghost module at the head side, the feature data will be transmitted to the output side for small target detection and recognition.

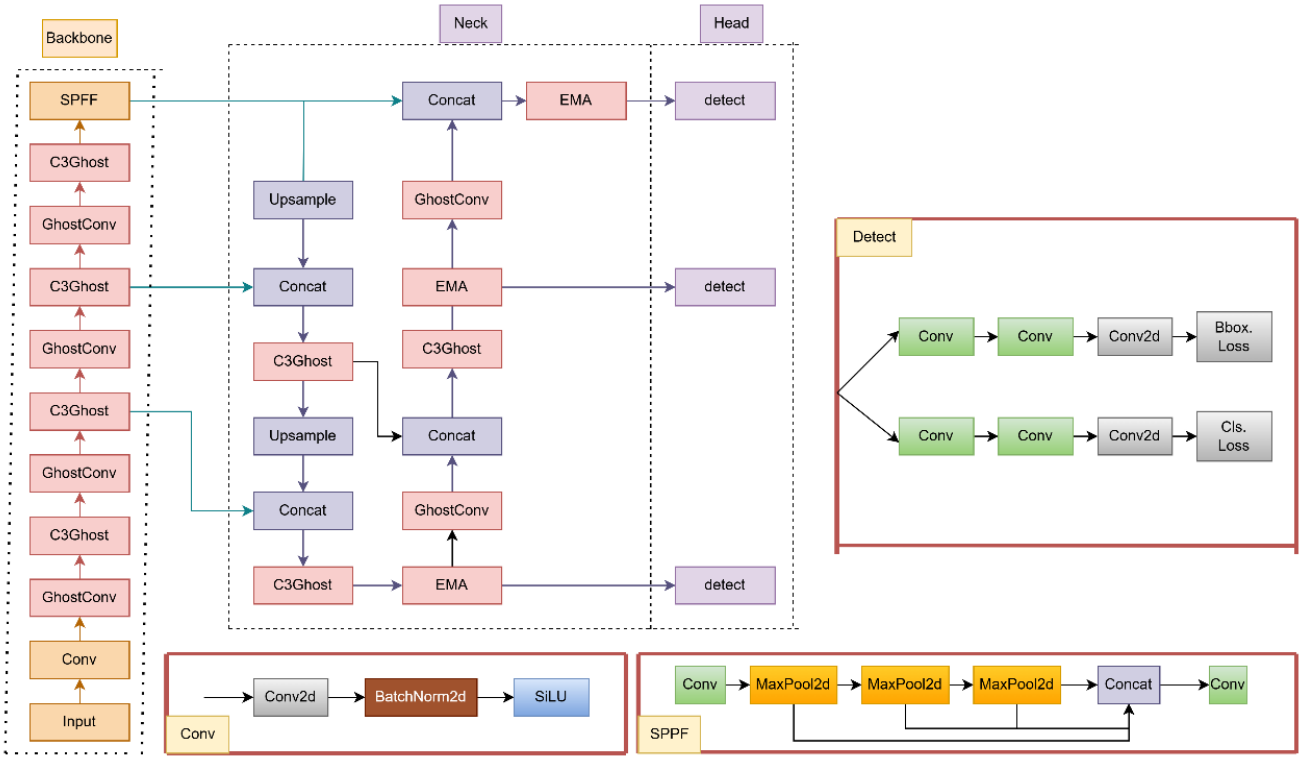

Fig. 2 Structure of the Yolov8n-GNA model

2.3. EMA Attention Mechanism

The EMA (Efficient Mutli-scale Attention) mechanism [16] is an efficient multiscale attention module to abate the side effects of constructing cross-channel relationships in extracting deep vision, thus reducing the amount of information and computation on each channel and allowing the spatial semantics to present a uniform distribution across feature groups. Adding the EMA mechanism to the YOLOv8n backbone network can solve the problem that the backbone network reduces the dimensional interaction of channel and spatial aspects of feature information due to the lightweight improvement so that the improved model further aggregates the output features of the parallel

branches through cross-dimensional interactions to achieve the purpose of capturing pixel-level pairwise relationships, which improves detection accuracy. The overall structure of the EMA is shown in Fig. 3, “Groups” denotes the number of groups into which the input channels are divided. “X Avg Pool” and “Y Avg Pool” respectively represent one-dimensional horizontal and vertical global pooling operations. The inputs are first grouped and then processed in different branches: one branch performs 1D global pooling and the other performs feature extraction by convolution. The output features of both branches are later modulated by a sigmoid function and a normalization operation and finally merged by a cross-dimensional interaction module. After the final sigmoid modulation, the output features are mapped to augment or attenuate the original input features to achieve cross-branch interaction to expand the weight of the target feature information.

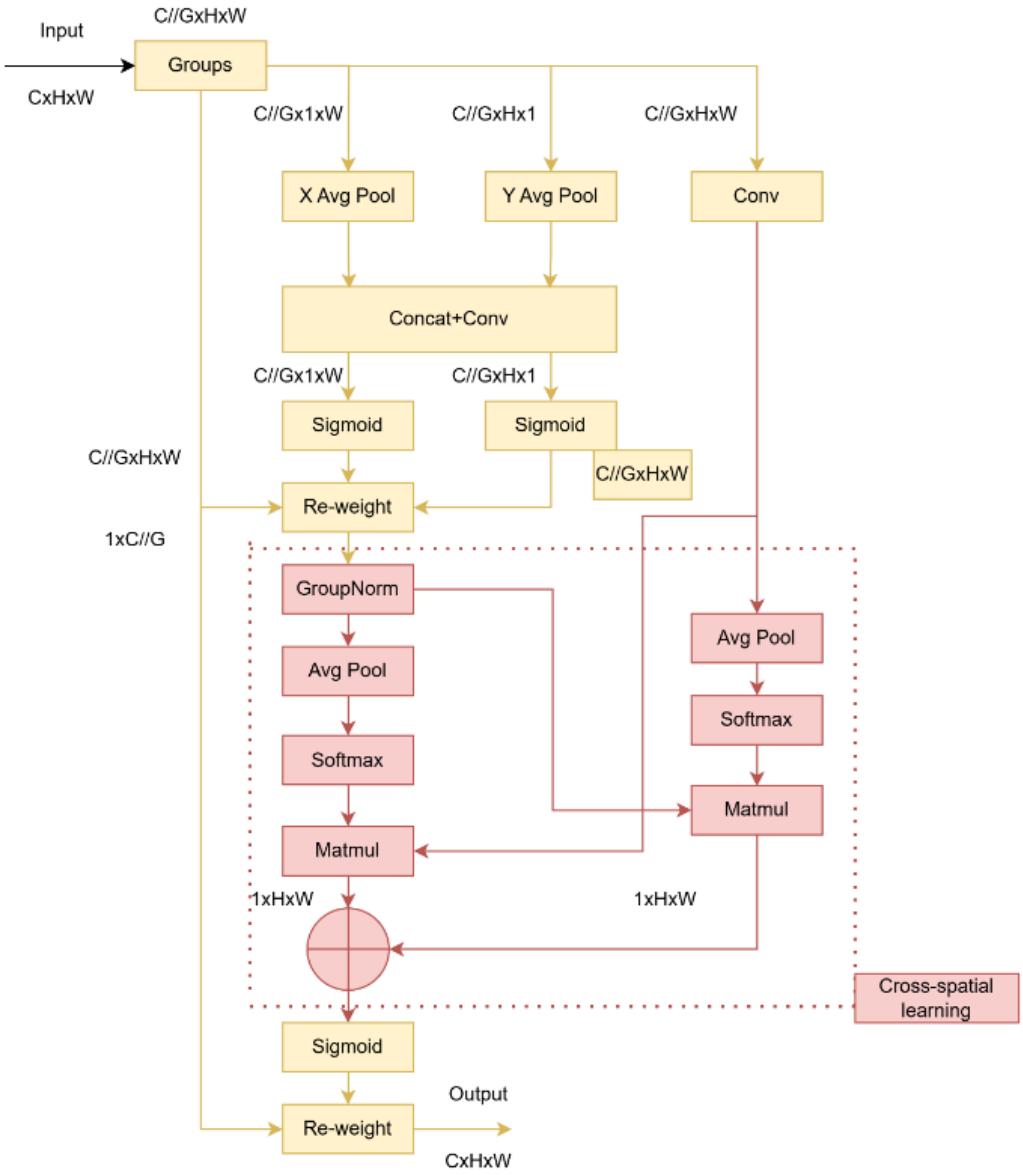

Fig. 3 Schematic diagram of the Efficient Mutli‑scale Attention mechanism

2.4. GhostNet Additions

GhostNet [17] is a lightweight network developed by Huawei Noah's Ark Lab in 2020. In GhostNet, the Ghostconv module serves as an alternative to conventional convolution and has advantages in model lightweight tasks. Modifying the conventional convolution to a GhostNet convolution can significantly reduce model complexity. The network principle involves dividing a

convolutional computation into two steps: ordinary convolution and simplified linear computation, where the number of computations and parameters is significantly reduced due to the convolutional operation, resulting in a model lightening effect. The comparison between conventional convolution and GhostNet convolution is shown in Fig. 4: first, the input image is compressed in terms of the number of channels using ordinary convolution, then depth-separable convolution (layer-by-layer convolution) is performed to obtain more feature maps, and then the different feature maps are spliced together and combined to form a new outturn. Based on the above operations, when the same size of convolution kernel is used, the computation and parameter count of GhostNet will be greatly reduced, so the GhostNet network is combined with the C3 module and the new C3Ghost structure is shown in Fig. 5

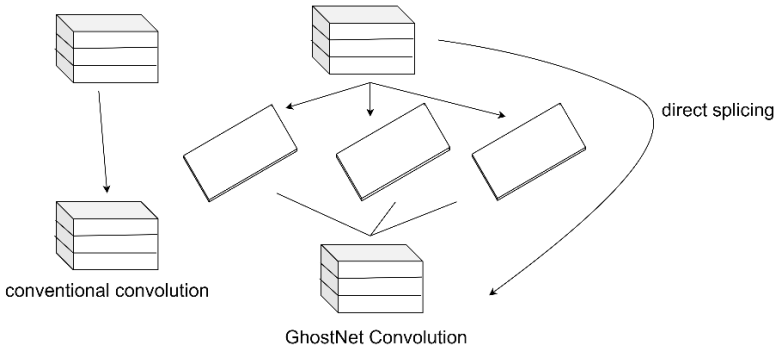

Fig. 4 Difference between Ghost convolution and conventional convolution

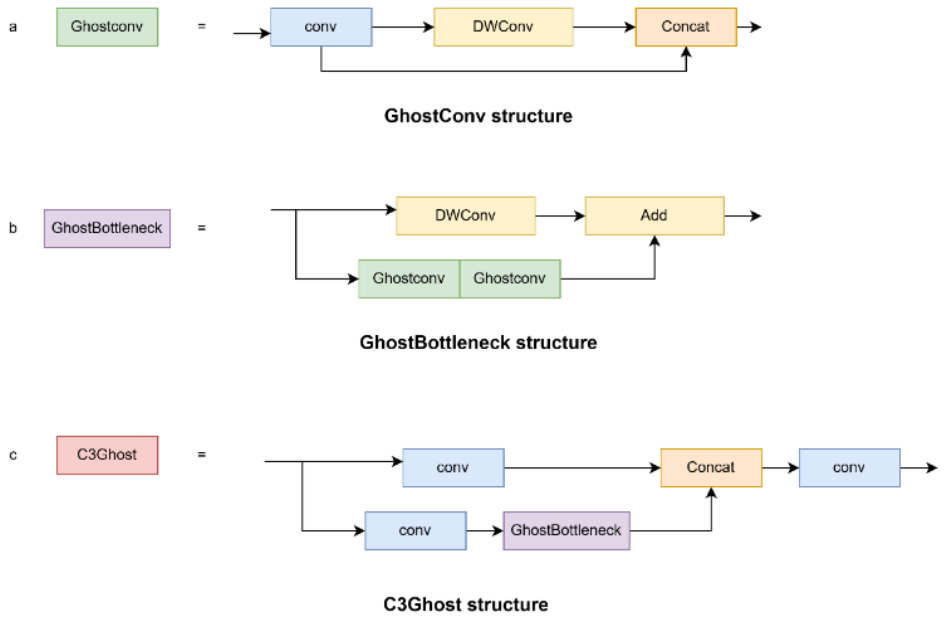

Fig. 5 The C3Ghost structure including its composition

3. Experimental results and analysis

3.1. Experimental Setup

The experiments were divided into comparison and ablation experiments. Faster-RCNN and SSD algorithms were used as comparisons. The environment for the ablation experiments was NVIDIA GeForce RTX 3060 Laptop GPU uses 6144 MiB of host memory. The programming language was Python 3.9.19, accelerated using CUDA v0. The training was performed based on the deep learning framework PyTorch 2.0.0. The parameters in the training were set to a maximum iteration number of 300, batch size of 8, a learning rate of 0.01, threshold of intersection and merger ratio (IOU) of 0.7,

bounding box loss weight (Box) of 7.5, loss weight of categories (cls) of 0.5, and deep feature loss weight (DFL) of 1.5. In this setting, the EMA mechanism and GhostNet are added under the basic YoloV8n model respectively, as a way to analyze the accuracy and lightweight improvement effect of the YoloV8n-GNA model.

3.2. Evaluation indicators

In this study, we mainly use Precision, mean average precision (mAP), and evaluation metrics to measure the execution time of the network model, using giga floating-point operations per second (GFLOPS) as the unit, and smaller computation volume means fewer resources can be saved. GFLOPS (giga floating-point operations per second) is used as the unit of measurement, where smaller computation means smaller resources can be saved, and Mb is used as the unit of measurement, while model size is also used as the metric. In addition, the Confusion Matrix Normalized can show the accuracy and confusion rate and can show the confusion situation. The Precision-Recall Curve (PRC) is a graphical representation used to summarize the performance of a classification model, particularly when dealing with imbalanced datasets. Precision and mean average precision (mAP) are calculated as follows:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

where TP denotes the true predicted positive example number, and FP is the wrong predicted positive example number.

$$\text{m}AP = \sum \frac{AP}{N} \tag{2}$$

where N is the total number of categories, AP is the area enclosed by a curve consisting of recall as the horizontal axis and precision as the vertical axis, and mAP@0.5 means the Mean average precision at an IOU threshold of 0.5.

3.3. Analysis of Performance Results

In the above environment, the Yolov8n-GNA model was used to train the dataset. The best model obtained after 300 rounds of training had a mAP50 value of 82.8% for sea urchins and 74.8% for starfish, which was able to achieve the goal of real-time monitoring. The overall accuracy was 86.7%, which was a good result. The computational volume and model size were 5.4 GFLOPS and 3.7Mb, which meets the lightweight index, so the improved Yolov8n-GNA model meets the practical application needs. The normalized confusion matrix of the training results of the model is shown in Fig. 6, in which the values on the diagonal line are the correct examples, the data on the off-diagonal line are the confusing cases, and the change of the precision-recall curve during the training process is shown in Fig. 7, and the loss functions, as well as the parameters as a whole change, is shown in Fig. 8.

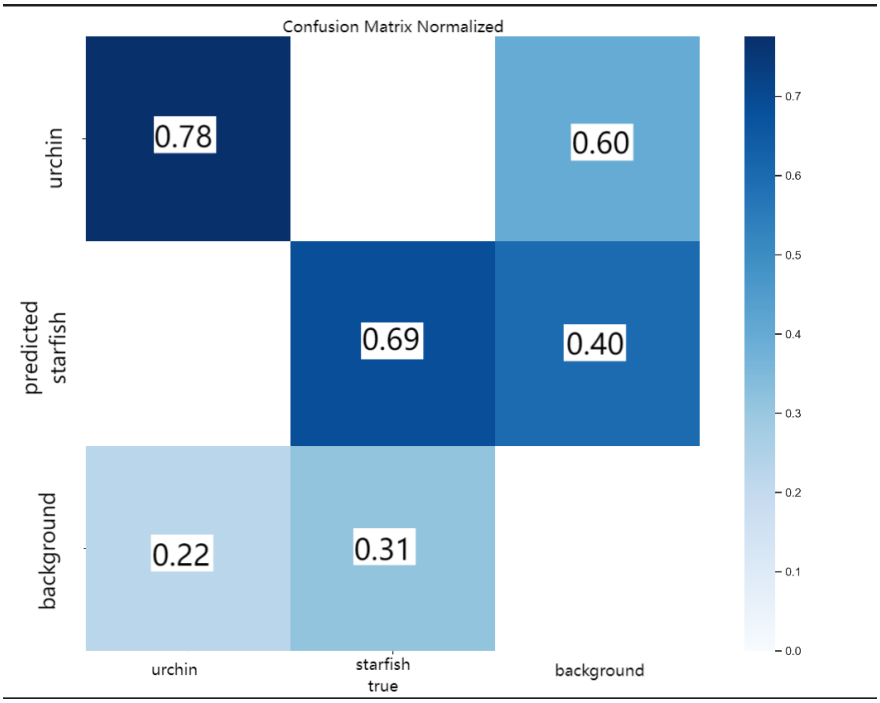

Fig. 6 Normalized confusion matrix

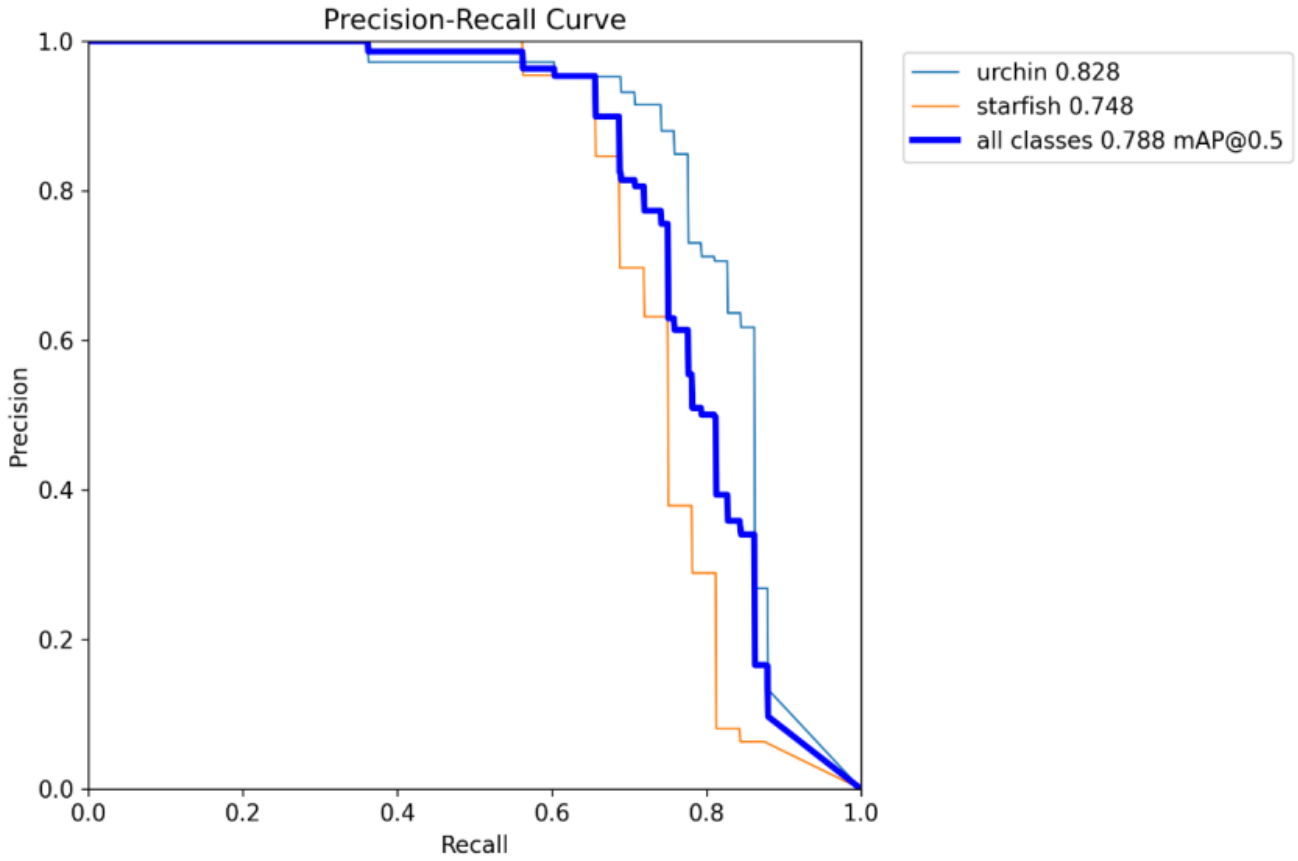

Fig. 7 Precision-recall curve

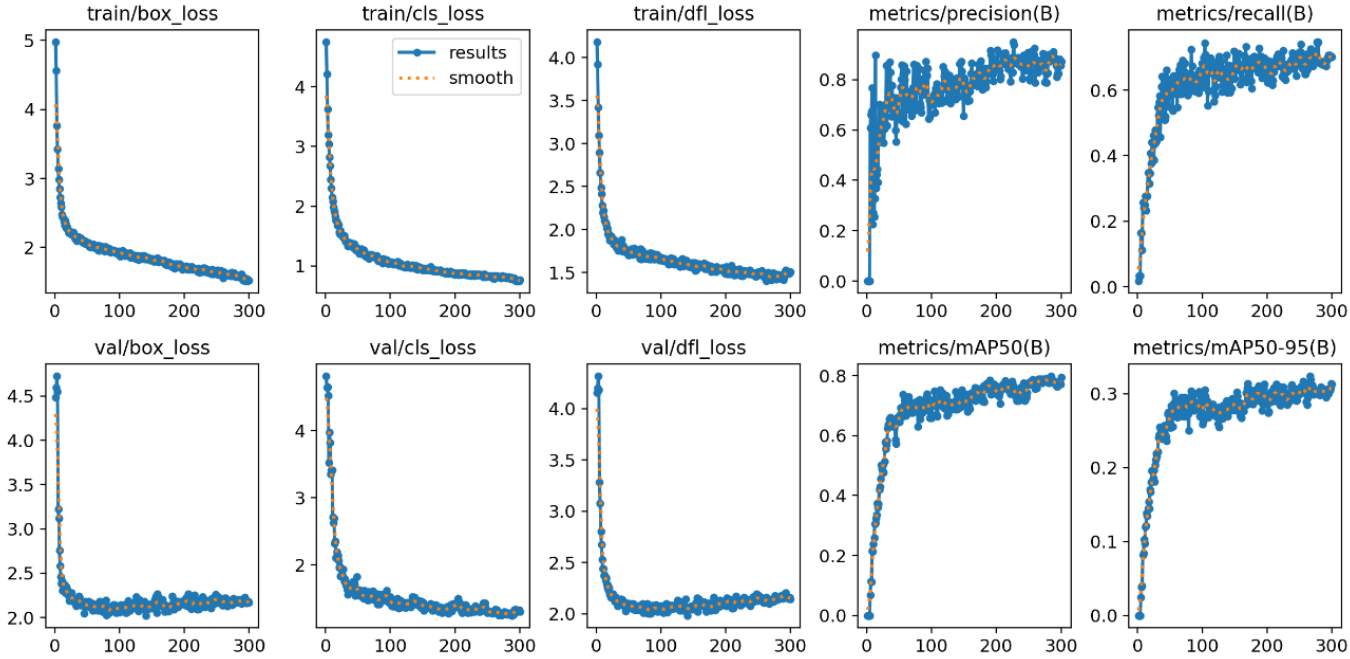

Fig. 8 Change curve of machine learning training metrics

3.4. Comparison and ablation experiments

Through comparison experiments, it can be seen that the Yolo algorithm is better than the traditional convolutional neural network algorithm for target detection in complex environments, and the model size is reduced substantially.

Based on the Yolov8n backbone network, different algorithmic structures are combined with the Yolov8n model to show the effect of different combinations of models on accuracy and computation. The experimental results show that the use of the EMA attention mechanism can effectively improve the accuracy of the model, and the use of C3Ghost convolution can effectively reduce the model size and computation volume, at the same time, the addition of the EMA attention mechanism and the C3Ghost convolution has a lighter weight and improved accuracy compared with the original model.

As can be seen from Table 1, using the C3Ghost convolution alone reduces the model size from the original 6.3Mb to 3.8Mb and the computation from 8.1GFLOPS to 5.0GFLOPS, which are 60% and 61% of the original model, respectively, and there is a 5% improvement in the Box value; using the EMA Attention Mechanism alone makes the Box value increasing from 76.7% of the original model to 86.0%, which is a 12% improvement compared with the original. When the combination of C3Ghost and EMA mechanism is added, the model size, computation, and Box value are 3.7MB, 5.4GFFLOPS, and 86.7%, respectively, which is a substantial optimization and improvement relative to the original model. The rest of the relevant reference quantities, such as recall, and map50 are the same. The comprehensive experimental results show that the model of the combination of C3Ghost and EMA can not only effectively reduce the amount of model computation, but also ensure the accuracy of model computation.

Table. 1 Comparison of models

	Ghost	EMA	Size	Volume	P	R	map50
Faster-RCNN			108Mb	370.2GFLOPS	0.438	0.961	0.748
SSD			91.1Mb	62.7GFLOPS	0.424	0.835	0.718
YoloV8n			6.3Mb	8.1GFLOPS	0.767	0.758	0.788
YoloV8n	√		3.8Mb	5.0GFLOPS	0.812	0.731	0.786
YoloV8n		√	6.3Mb	8.3GFLOPS	0.860	0.717	0.793
YoloV8n	√	√	3.7Mb	5.4GFLOPS	0.867	0.697	0.788

3.5. Model testing results

Thirty-two images from the dataset were selected for testing, including sea urchins, starfish, and other organisms. The test results are shown below in Table 2. Statistics on the number of sea urchins and starfish contained in the test images and the labels obtained from the detection model trained by the Yolov8n-GNA model, the correct, faulty, and missed detection data are recorded and the accuracy rate is calculated respectively, it can be concluded that the accuracy rate of sea urchins and starfish in the actual detection is 91.1% and 92.4%, which is a better performance in the practical application, the recognition images and the statistical results are shown in the following diagrams The recognition images and statistical results are shown in the following table.

Table. 2 Model detection results for each tag

label	Total tags	correct	miss	wrong	correctness rate
urchins	34	31	3	3	91.1%
starfish	53	49	4	0	92.4%

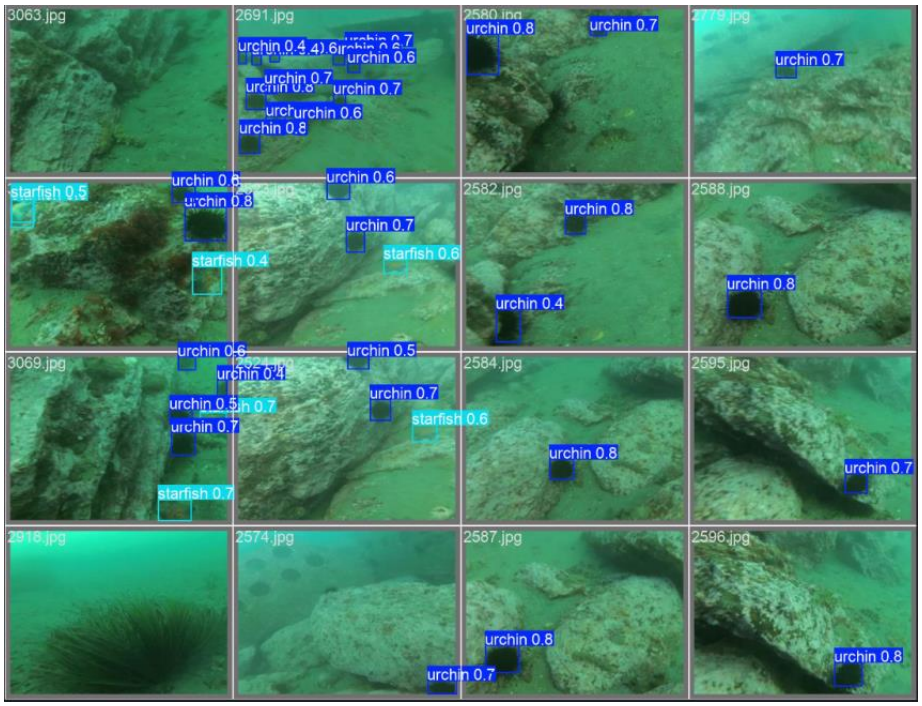

Fig. 9 Visualization of detection results

3.6. Presentation of model results in different environments

The lighting conditions in the underwater environment are significantly different from those in the land environment, and the absorption and scattering effects of water have a significant impact on image quality. In addition, changes in water depth led to changes in light intensity and spectral distribution, which puts higher requirements on the robustness of underwater image processing algorithms. Therefore, three each of the blue-green, blue, and green marine environments are randomly selected from the dataset and detected using the trained Yolov8n-GNA model, and the detection results are shown in Fig. 10. The experimental results show that the model can cope with the complex marine background environment with high robustness.

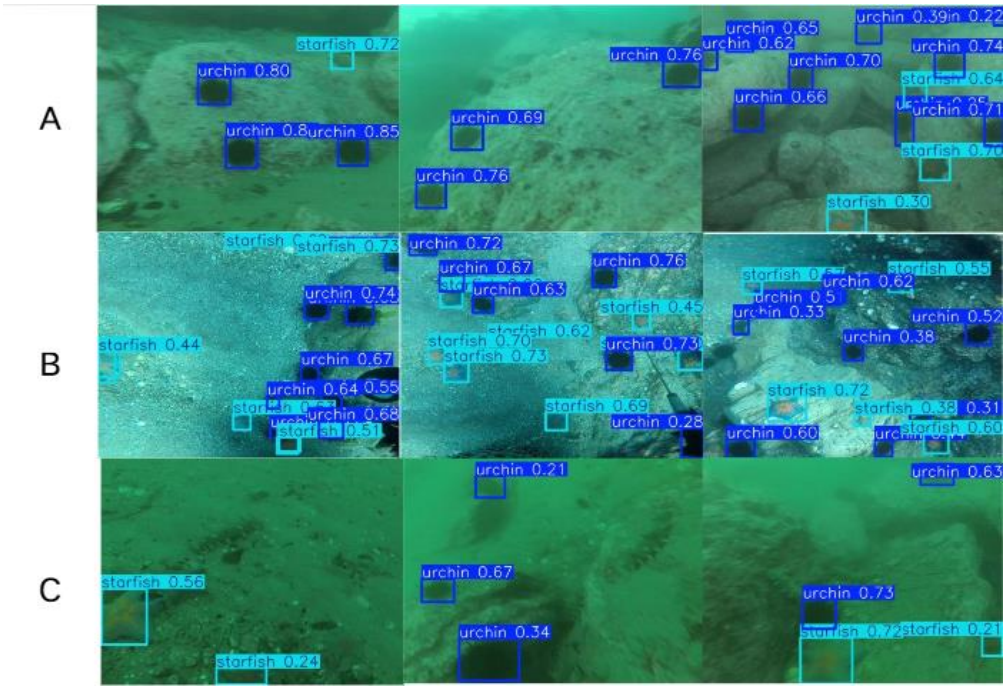

Fig. 10 Presentation of results in different marine environments (blue-green in group A, blue in group B and green in group C)

4. Conclusion

To address the lack of automation in sea urchin aquaculture management and harvesting, this study proposes a new detection method. An improved YoloV8n network with the addition of a GhostNet lightweight network and the EMA attention mechanism is used to detect sea urchins and starfish with a computer vision approach.

(1) The improved YoloV8n network uses GhostNet to replace the backbone network and introduces the EMA attention mechanism, and the model size and computation of the improved model are 3.7 Mb and 5.4 GFLOPS, which are 58% and 67% of the original model, respectively, and the accuracy is 86.7%, which is a 13% improvement over the original model. Therefore, this method is suitable for real-time monitoring on low-computing power platforms, which is convenient to put into practical application scenarios for sea urchin farming.

(2) This method is mainly oriented to the economy of agricultural production, so it reduces the model size as much as possible, strikes a balance between model size and recognition accuracy, and reduces the amount of computation under the condition of guaranteeing the accuracy, to reduce the cost. However, due to the limitation of the conditions, the recognition accuracy is relatively poor in the more extreme and complex sea conditions, and there is still room for improvement.

(3) The model provides a relatively high accuracy target detection model suitable for low-computing power platforms, which can be applied to more underwater target detection scenarios. The detection head and data enhancement algorithm can also be added to further improve its accuracy and robustness.

References

[1] Yanbang Zhang, et al. Salient Object Detection Based on Texture and Color Features[J]. Computer&Digital Engineering,2021,49(09).

[2] Susanto T, et al. Development of Underwater Object Detection Method Based on Color Feature [C]//2018 International Conference on Computer Engineering, Network and Intelligent Multimedia (CENIM). Indonesia: IEEE, 2018: 254-259.

[3] XIE Tao, et al. Daytime Sea fog recognition algorithm over the Yellow Sea and Bohai Sea based on multi-gray co-occurrence matrix eigenvalues[J]. Journal of the Meteorological Sciences, 2024, 44(1): 189-198.

[4] LI Jiakang, et al. Identification and analysis of color changes in aquaculture waters using color checker[J]. Fishery Information and Strategy, 2024, 39 (01): 49-60.

[5] Xiao Y, et al. A review of object detection based on deep learning[J]. Multimedia Tools and Applications, 2020, 79: 23729-23791.

[6] LI X H, et al. Fish trajectory extraction based on object detection[C]// 2020 39th Chinese Control Conference (CCC). USA: IEEE, 2020: 6584-6588.

[7] Fang S. Research on fish phenotypic data measurement method based on deep learning [D]. Hangzhou: Zhejiang University, 2021.

[8] Zhao Z, et al. Composited FishNet: Fish detection and species recognition from low‐quality underwater videos [J]. IEEE Signal Processing Society, 2021, 5(30): 4719-4734.

[9] Zeng T, et al. Lightweight tomato real-time detection method based on improved YOLO and mobile deployment [J]. Computers and Electronics in Agriculture, 2023(205): 107625.

[10] Wang Q, et al. Design, integration, and evaluation of a robotic peach packaging system based on deep learning [J]. Computers and Electronics in Agriculture, 2023, 211: 108013.

[11] XU Ruifeng, et al. Shrimp Diseases Detection Method Based on Improved YOLOv8 and Multiple Features [J]. Smart Agriculture, Mar. 2024 Vol. 6, No. 2

[12] Zhai Xianyi, et al. Underwater Sea cucumber recognition method based on improved YOLOv5[J]. Applied Science, 2022, 12: 9105.

[13] Song Xiaolu, et al. Recognition and counting algorithm of underwater sea cucumbers based on YOLOv4 network[J]. Journal of Chinese Agricultural Mechanization, 2024, 45(9): 258-264

[14] R. Liu, et al. Real-World Underwater Enhancement: Challenges, Benchmarks, and Solutions Under Natural Light[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2020, 30(1): 1-1.

[15] Varghese R, S M. YOLOv8: A Novel Object Detection Algorithm with Enhanced Performance and Robustness[C]//2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS). Chennai, India: IEEE, 2024: 1-6.

[16] D. Ouyang, et al. Efficient Multi-Scale Attention Module with Cross-Spatial Learning[C]. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 2023: 1-5.

[17] Han K, et al. GhostNet: More Features from Cheap Operations[C]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 2020: 1577-1586.