Improved YOLOv7 Underwater Object Detection Based on Attention Mechanism

Junshang Fu, Ying Tian

Abstract—**The task of detecting marine target organisms has always been a challenging issue, despite the numerous machine learning detection methods proposed to improve precision. The underwater image blurriness caused by irregular light absorption and water quality remains a major obstacle to achieving accurate detection. This results in high misalignment rates and poor underwater scene recognition capabilities for detecting underwater targets. To address this, we put forward a YOLOv7-RNCA underwater target detection technology based on improvements to YOLOv7. This model adds residual modules and coordinate attention mechanisms (CA) at the end of the backbone network, as well as incorporating partial convolution (PConv) modules. The combination of these three components makes the model more precise during the detection process while reducing unnecessary computation and memory access. This allows for better optimization during deep network training and preserves more feature information. Additionally, we reconstructed the SPPCSPC structure and incorporated a global attention mechanism (GAM) to form the SPPCSPC-GAM module in the neck network, which improves the performance of the convolutional neural network (CNN) and ensures good data capabilities and robustness during training, thereby enhancing the target detection ability. We also improved the neck ELAN module by introducing PConv convolution modules, which continuously enhance network learning abilities without disrupting the original gradient path. The introduction of the PConv module reduces redundant computation and memory access, making the ELAN-PConv module more effective at extracting spatial features. Our outcomes of experimentation indicate YOLOv7-RNCA network an average precision of 86.6% on the URPC dataset, outperforming existing methods in accuracy detection and demonstrating great potential as a promising solution for marine target monitoring tasks.**

Index Terms—**Underwater Target Detection, Marine Resources, YOLOv7, Attention Mechanism**

I. INTRODUCTION

With the continuous exploitation of resources by mankind, natural resources on land are gradually depleting, and resource regeneration cannot keep pace with the rate of extraction. This situation necessitates finding new ways to obtain resources, leading us to turn our attention to the ocean. The ocean is the largest resource bank in the world, containing numerous precious treasures and holding the potential to provide food, medicine, and other essentials for human life[1-2]. However, to harvest marine resources, it is necessary to upgrade underwater equipment and effectively survey the ocean.

In recent years, advancements in marine robot technology have significantly enhanced our ability to explore the oceanic environment. This progress underscores the importance of preserving marine ecosystems and responsibly utilizing marine resources. Effective target detection is crucial in this context, and leveraging advanced machine vision technology is essential for oceanic exploration [3]. Through the use of ocean robots such as remote operating vehicles (ROVs) [4] and autonomous underwater vehicles (AUVs) [5], valuable underwater data can be acquired. For example, the URPC dataset used in this study contains underwater data gathered using ROVs and AUVs. These datasets provide crucial insights by addressing fundamental questions, such as the precise locations of objects. Underwater target detection predicts the location and category label for each detected target of interest using bounding boxes, enhancing comprehension of the intricate nuances of the oceanic realm.

Through the application of marine robots, we are able to acquire underwater datasets in the ocean and perform underwater target detection. Both deep sea and shallow sea can be targets for detecting hydrops within a certain range, and they can contribute significantly as well to seafood aquaculture in shallow waters [6]. For instance, by collecting specific marine organism datasets, such as those of holothurian, echinus, abalones, starfish, alongside other aquatic creatures, we can detect the presence of these organisms in certain areas, thereby facilitating seafood fishing. This can be of great assistance to seafood farmers in shallow seas. Through target detection, they can determine the density of various seafood in aquaculture areas, estimate seafood yields, and prevent underwater abnormalities by analyzing the target pictures detected. This greatly saves manpower, materials, and finances, and facilitates reuse of assets.

As deep learning target detection algorithms continue to evolve, we have observed significant improvements in feature extraction capabilities and robustness, laying a solid foundation for underwater target detection and driving its onward progress. Nevertheless, the application of these algorithms to underwater target detection presents a myriad of issues and challenges. For example, the intricate underwater environment gives rise to wavelength-related absorption and scattering, resulting in a notable decline in image quality when capturing underwater scenes using robotic devices. This degradation manifests in the form of reduced contrast, image blurring, and various other issues

Manuscript received November 23, 2023; revised May 20, 2024.

Junshang Fu is a postgraduate student majoring in software engineering at School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Liaoning 114051, China. (e-mail: fjs2284781073@163.com).

Ying Tian is a professor of School of Computer Science and Software Engineering, University of Science and Technology Liaoning, Liaoning 114051, China. (corresponding author to provide phone: +8613898015263; e-mail: astianying@126.com).

that hinder the collection of effective datasets. Moreover, a range of variables, including water quality, clarity, and temperature, across diverse aquatic environments—from shallow seas to the depths of the ocean and intricate coral reefs—can significantly impact the ability of underwater robots to detect targets accurately [7]. Additionally, obstacles such as small underwater targets, occlusion, overlap, and blurring pose considerable difficulties in achieving efficient target detection tasks within the oceanic environment.

Therefore, in response to these problems, we propose a method based on improving target detection of YOLOv7 on the basis of leveraging the ongoing advancements in deep learning technology. This is done as follows.

The fusion of the ResNet_CA_PConv unit within the core network structure signifies a substantial improvement. With the aid of the PConv module [8], both computations and memory access are efficiently reduced. Moreover, by flawlessly embedding the CA focus system [9-10], the module elevates network accuracy without imposing any additional computational burden. This synergistic approach is then tactically applied to the residual structure, resulting in a refined integration process that safeguards against feature loss, amplifies detection accuracy, and streamlines computational complexity.

SPPCSPC module is reconstructed in neck network, and GAM (Global Attention Mechanism) [11] is added to replace CBS module in SPPCSPC module, to attain excellent capability in feature detection and resilience during the training procedure.

Finally, by enhancing the ELAN-W module within the neck network, the PConv module is incorporated into the ELAN module, replacing one Conv module within the Conv_BN_SiLu structure. This improvement enhances the effectiveness of the ELAN-W module in feature extraction.

II. RELATED ALGORITHMS

Advanced neural networks have achieved noteworthy advancements in feature extraction over the past few years, with object recognition systems utilizing this technology, finding widespread applications in terrestrial, aerial, and military domains. Despite their extensive use on land, the research on target detection in the ocean remains limited. As target detection methods continue to advance, a growing number of algorithms tailored for oceanic target detection are expected to emerge continuously in the future.

Deep neural network object identification methods can generally be grouped into two main groups. The group involves two-phase procedures like R-CNN [12], FastR-CNN [13], and FasterR-CNN [14]. These algorithms identify candidate regions in an image and verify their content. They achieve high accuracy but suffer from slow detection speeds due to their computational complexity. The second group consists of one-stage detection algorithms, exemplified by OverFeat [15], YOLO [16], and SSD [17]. These algorithms directly locate objects and calculate losses using a single regression classification approach.

With the continuous development of single-stage detection techniques, increasingly, researchers have utilized the YOLO algorithm to detect objects beneath the water's surface. For instance, Lei et al. [18] enhanced the images by applying the CLAHE algorithm and histogram equalization to the raw data, and then applied it to YOLOv2 and YOLOv3 to meet the requirements of better underwater recognition. Xu et al. [19] fused MobileNet V2 with depth-wise separable convolution, enhanced the AFFM module, balancing speed and accuracy for underwater object detection. Jian et al. [20] proposed a dual-domain data augmentation method to enhance data diversity. They used self-attention and convolutional operations to increase the detection efficiency of YOLOv7, introduced the SIoU loss function for faster convergence, and enhanced model performance. Yi et al. [21] integrated the SENet attention framework into YOLOv7 to focus on capturing more crucial information of small targets in the network, enhancing network topology to reduce model complexity, and implemented the EIoU loss function to boost underwater model detection precision. Shen et al. [22] launched the MIPAM module and integrated it into the YOLO detector model to strengthen detection task effectiveness by collecting more feature information through MIPAM. In summary, their research on single-stage detection techniques not only demonstrated the effectiveness of single-stage object detection but also provided a solid platform for studying underwater object detection using YOLO.

YOLOv7 [23] stands as a cutting-edge single-stage real-time target detector. Its architecture primarily comprises two components: the Backbone, alongside modules such as MP, ELAN, CSB, and SPPCSPC, synergistically employed for extracting crucial image information. The Head segment seamlessly integrates upper and lower sampling, detection layers, and additional modules, ensuring further refinement and judicious application of the extracted features. Subsequently, through a meticulous analysis of these features, YOLOv7 excels in both speed and precision, outperforming its predecessors. Remarkably, YOLOv7 reduces parameters by 40% and computations by 50% compared to the foundational YOLO model, making it the preferred choice for enhancing target detection research in underwater marine organisms. However, despite its promise, there remains a paucity of studies exploring the YOLOv7 model's efficacy in marine organism target detection, highlighting the imperative for refining its accuracy and paving the way for future advancements.

III. IMPROVEMENTS

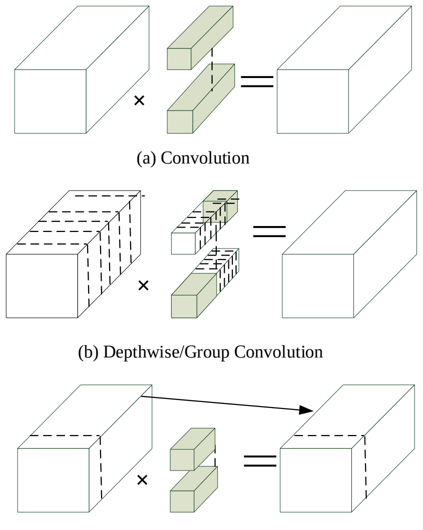
A. Introduction to the PConv Module

To underscore the proficiency of the ELAN module, this paper avails itself of the highly effective feature extraction technique denominated as PConv, which was introduced by Chen et al. This methodology adeptly extracts spatial features through the application of traditional convolution to a select group of input channels, thereby preserving the total count of channels. The calculation of memory access pertaining to interconnections is determined by considering either the initial or the terminal channel as an emblematic representation of the entire feature map. The corresponding FLOPs are then tallied in the manner outlined below:

$$h\times w\times 2C_P+k^2\times C_P^2 \tag{1}$$

The variables *h* signifies height, *w* signifies width, and *d* signifies depth. of the input channels, severally. In this scenario, the calculated quantity is merely 1/16 of the standard calculated amount. Additionally, the memory access of the PConv module is minimal, approximately $h\times w\times 2C_P$, constituting 1/4 of that required by standard convolution.

A schematic of the PConv module work is shown in Fig. 1.



(a) Convolution

(b) Depthwise/Group Convolution

(c) Partial Convolution (ours)

Fig. 1. PConv module basic working principle

B. ELAN-PC Module

The YOLOv7 ELAN module is engineered to regulate the shortest and longest path gradients, facilitating effective learning and convergence in deeper networks. This maintenance of the original gradient path continuously amplifies the network's learning capacity and fortifies its robustness, thereby expanding its ability to acquire a broader range of features. In contrast to the standard Conv convolution module, the PConv module swiftly captures input branch features and accesses memory expediently. This not only enhances the model's capability to detect small targets but also streamlines computations and memory access, ultimately improving the extraction of spatial characteristics when integrated with ELAN. By supplanting the common Conv module with PConv, the channel features remain intact while upholding a high computational speed. The modified ELAN-PC module is illustrated in Fig. 2 below.

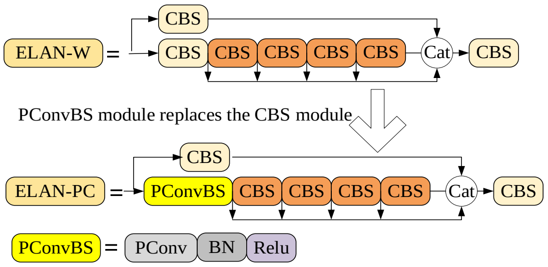


Fig. 2. ELAN compare with ELAN-PC

C. ResNet_CA_PConv Module

The YOLOv7 backbone network has been augmented with the integration of the ResNet_CA_PConv module, which effectively maintains the consistency of the captured feature characteristics. This module relies on ResNet [24] compressed design.

Ongoing investigations into the design of mobile networks reveal that channel attention holds paramount significance in elevating model performance. Nevertheless, a noteworthy constraint of this methodology lies in its disregard for location information. To bridge this gap, we present a groundbreaking and computationally streamlined CA mechanism. This innovative approach seamlessly integrates location nuances with channel data, empowering mobile networks to effectively concentrate on diverse regions without imposing substantial computational burdens.

To overcome the potential positional details lost through 2D holistic pooling, we refine the channel attention mechanism by decomposing the global average pooling. This decomposition involves the utilization of two 1D global pooling operations. These operations encourage attention blocks to identify spatial distant dependencies while maintaining exact location context. Concisely, for input X, we codify each channel individually along horizontal and vertical axes, exploiting spatial ranges of (H, 1) and (1, H), respectively. Consequently, the output for a given channel *C* at a specific height h and width w is determined as follows:

$$z_c^w(h)=\frac{1}{W}\sum_{0\leq i<W} x_c(h,i) \tag{2}$$

$$z_c^w(w)=\frac{1}{H}\sum_{0\leq j<H} x_c(j,w) \tag{3}$$

The aforementioned two transformations to create a set of two feature maps for directional sensing, specifically $C\times H\times 1$ and $C\times 1\times W$ trait diagrams. This allows our attention to detect remote distance connections in a dimensional alignment and maintain accurate positional data, thus enhancing the network's precision in localization. Subsequently, the resulting feature map of $C\times 1\times W$ undergoes a 1×1 dimensional convolution function F1 to obtain:

$$f=\delta(F1([z^h,z^w])) \tag{4}$$

[,] marks the progression of operations along the spatial plane, δ is a nonlinear activation mechanism, and $f\in R^{\frac{C}{r}\times(H+W)}$ is an intermediary feature representation that encodes spatial data in both the lateral and vertical planes. *r* serves as the factor to regulate the scaling proportion of the block size within SE blocks, Subsequently *f* is separated into two independent tensors $f^h\in R^{\frac{C}{r}\times H}$ and $f^w\in R^{\frac{C}{r}\times W}$ across the spatial axis, Then, a 1×1 convolution is applied to perform the dimension lifting operation on each tensor, respectively, followed by the final attention vectors are derived by applying the sigmoid activation function, as follows:

$$g^h=\sigma(F_h(f^h)) \tag{5}$$

$$g^w=\sigma(F_w(f^w)) \tag{6}$$

By utilizing the sigmoid function σ, we commonly reduce the number of channels of *f* by an suitable scaling ratio *r* in order to streamline the model. Subsequently, we examine the impact of various decline ratios on execution and generate g^h and g^w extensions, which subsequently affect the attention weights. Ultimately, this culminates in the Coordinate Attention output formula.

$$g_c(i,j)=x_c(i,j)\times g_c^h(i)\times g_c^w(j) \qquad (7)$$

The CA module not only considers the relationship between space and channel, but also tackles the long-range dependency problem. By doing so, it avoids excessive information loss, leading to improved accuracy. Furthermore, the module boasts fewer parameters and reduced computational requirements. Labeled as 'CA Module Algorithm Flowchart', Fig. 3 illustrates the step-by-step process of the CA module's algorithm.

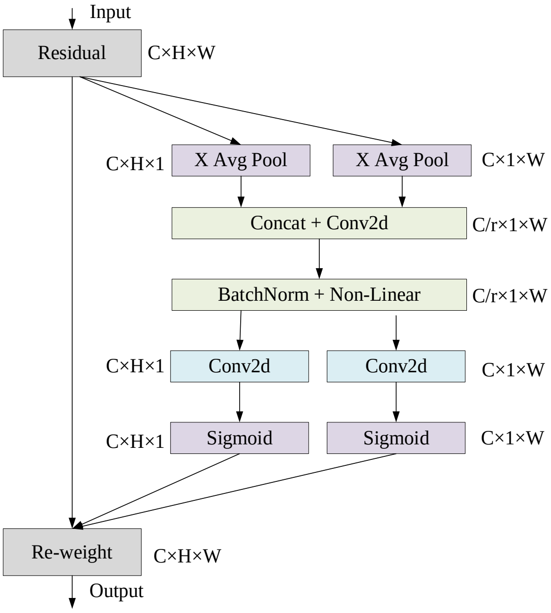


Fig. 3. CA attention mechanism network structure

Based on the residual structure of ResNet, the CA module and PConv module are integrated within the ResNet module. The 1×1 convolution is replaced with PConv, facilitating rapid extraction of input branch characteristics, reducing redundant calculations, and enhancing memory access speed. Subsequently, the 3×3 convolution is replaced with the CA module. This allows for the aggregation of input characteristics along the two spatial dimensions, enabling the capture of long-distance dependencies and precise location information. This not only helps prevent information loss but also reduces parameters and calculations. The improved ResNet_CA_PConv module is illustrated in Fig. 4.

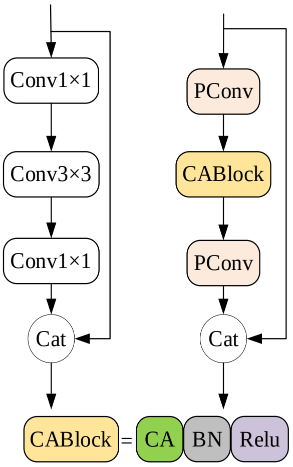


Fig. 4. ResNet_CA_PConv module structure diagram (left: ResNet; Right: ResNet_CA_PConv)

D. *The SPPCSPC Module is Reconstructed*

The SPPCSPC module in YOLOv7 enhances the model's receptive field and feature expression capability by pooling the feature maps across the input's multi-scale spatial pyramid. Attention mechanism is a technique used to enhance complex feature identification environments by distributing varying weights to different parts of the neural network input. Notably, the GAM attention mechanism demonstrates strong performance. GAM consists of a channel-focused attention module and a spatial attention module. During the channel attention phase, GAM initially transforms the dimensionality of the input feature map. It then feeds it into two layers of Multi-Layer Perceptron (MLP) followed by sigmoid processing. Meanwhile, the spatial attention mechanism integrates utilizing spatial info two convolutional layers and sigmoid processing, allowing the network to prioritize context-sensitive areas in the image. The schematic of GAM is depicted in Fig 5. Replacing the Conventional Block Structure (CBS) with GAM in the SPPCSPC module leads to significant accuracy improvements. The updated SPPCSPC module is illustrated in Fig 6.

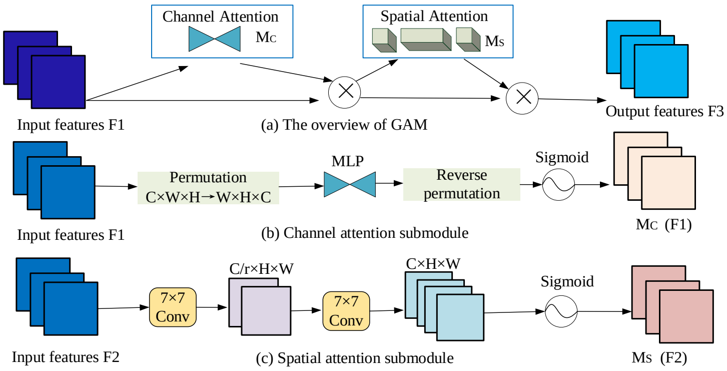


Fig. 5. GAM model

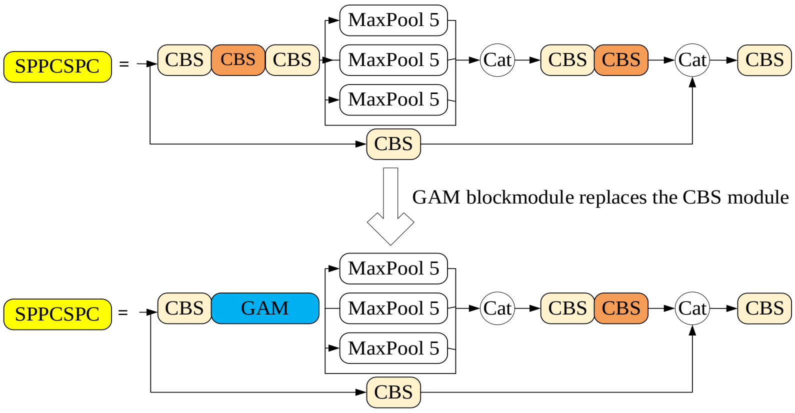


Fig. 6. The structure diagram of SPPCSPC-GAM module

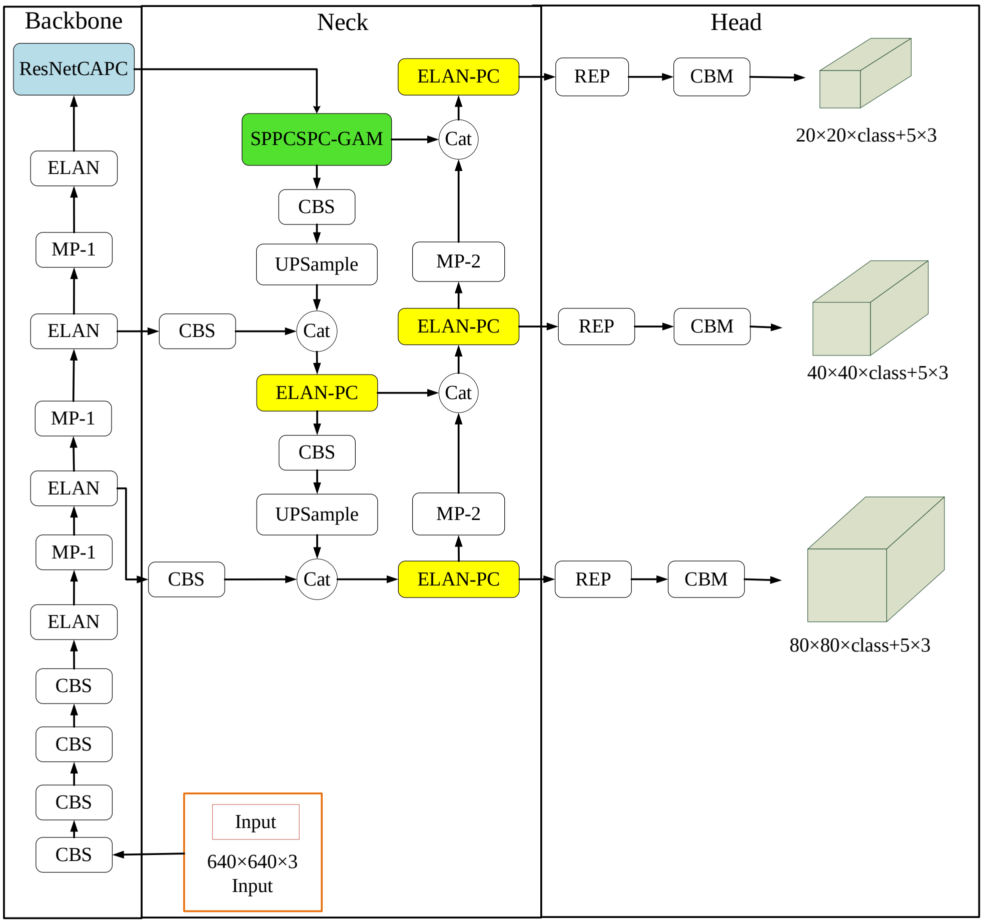


Fig. 7. Schematic of YOLOv7-RNCA structure

In summary, the proposed YOLOv7-RNCA model effectively preserves the Backbone features captured by introducing the ResNet_CA_PConv structure at the bottom layer of the backbone network. This model is able to extract feature information from small targets and complex backgrounds, thereby enhancing the model's ability to focus on valuable content and specific locations within input image samples. By integrating PConv into the ELAN module, rather than using the original 1×1 convolution module of the neck network, computation redundancy is reduced, inference time is shortened, while maintaining detection accuracy. Simultaneously, the introduction of the GAM into the SPPCSPC module enhances the network's feature extraction capabilities and boosts the original network's deep feature distillation skill. The modified YOLOv7-RNCA model is shown in Fig. 7.

IV. EXPERIMENTS

A. *Experiment Environments*

Hardware specifications include an NVIDIA GeForce RTX 4070Ti graphics card with 12GB of video memory. The software environment comprises Windows 10, CUDA 11.3, Pytorch 1.11.0, and Python 3.8.0. For the training parameters, the inputted pictures are resized to 640×640 pixels, the total set iteration limit to 300, and a batch size of 4 is used. The chosen optimizer is Stochastic Gradient Descent (SGD) momentum value 0.937. Set the learning rate to 0.01 and apply a weight decay of 0.0005. Regulate the learning rate as training progresses, a cosine annealing learning algorithm is employed.

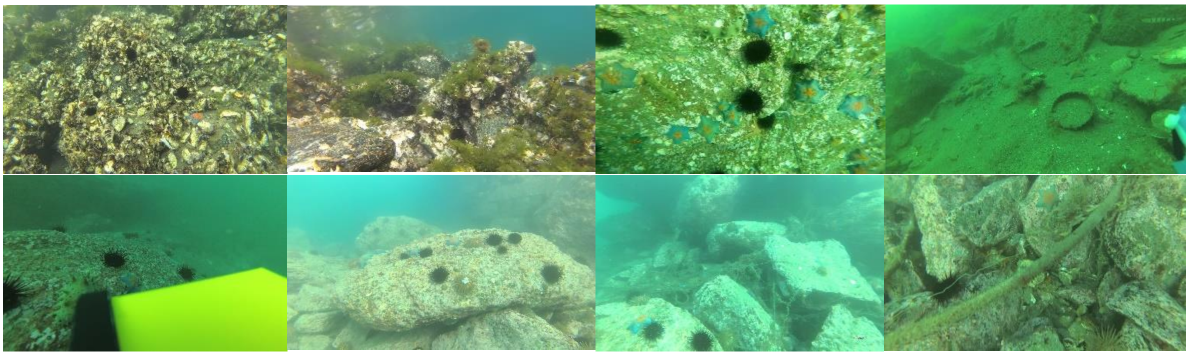


Fig. 8. The URPC marine organism datasets

B. Datasets and Settings

In this research, we chose an underwater dataset originating from the Underwater Robot Professional Competition (URPC). This dataset comprises images captured by autonomous underwater vehicles (AUVs) in the aquatic meadows of Zhangzi Island, Dalian, offering an authentic depiction of the marine ecosystem. As the official competition test set is not publicly accessible, we curated a collection of 6671 images, highlighting four marine species: sea urchins, sea cucumbers, sea stars, and scallops. These are labeled as echinus, holothurians, starfish, and scallops in the dataset. Fig 8 showcases some representative images from this dataset. For experimental purposes, we allocated the images into training, validation, and test sets in a 7:1:2 ratio, resulting in 4669 pictures for training, 667 for validation, and 1335 for testing, all randomly distributed.

C. Model Evaluation Metrics

Target identification key metrics: Precision, Recall, IoU, AP, mAP, the weighted harmonic average F1, and the number of parameters (Params).

We can assess the effectiveness of underwater marine organism detection by considering the computational complexity of the model, quantified by FLOP count. Additionally, we can use TP, FP, and FN to represent the number of correctly detected, falsely detected, and missed marine organisms, respectively. Another evaluation metric is the AP, which denotes precision-recall curve's area under a curve, and mAP represents the average AP across all categories. These metrics provide a comprehensive assessment of the model's performance in detecting marine organisms in underwater environments:

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

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

$$AP=\int_0^1 P(R)\mathrm{d}R \tag{10}$$

$$F1=\frac{2PR}{P+R} \tag{11}$$

$$\mathrm{mAP}=\frac{1}{n}\sum_{j=1}^{n} AP(j) \tag{12}$$

Among them, *TP* is true positive, *TN* is true negative, *FP* is false positive, and FN is false negative. N is the number of classes.

D. Results and Analysis of the P-R Curve on the URPC Dataset

P-R curves were compared to assess the detection capabilities of the introduced YOLOv7-RNCA model and the YOLOv7 model on the URPC dataset, as shown in Fig. 9 and Fig. 10. Relying on P-R curve, the area under the PR curve for the YOLOv7-RNCA model is larger than that for the original YOLOv7 model. This suggests that the newly introduced model, YOLOv7-RNCA, has good generalization ability.

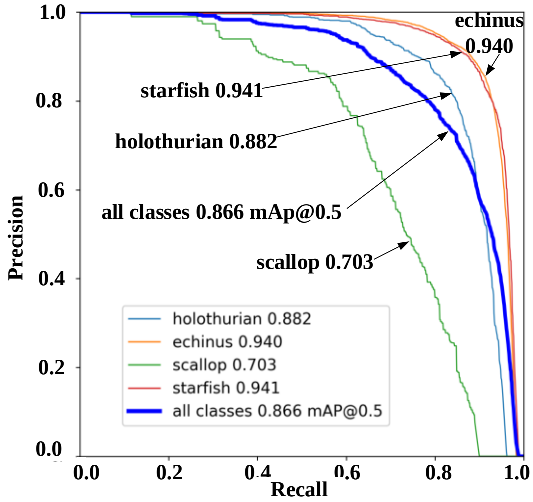


Fig. 9. P-R curve of YOLOv7-RNCA model

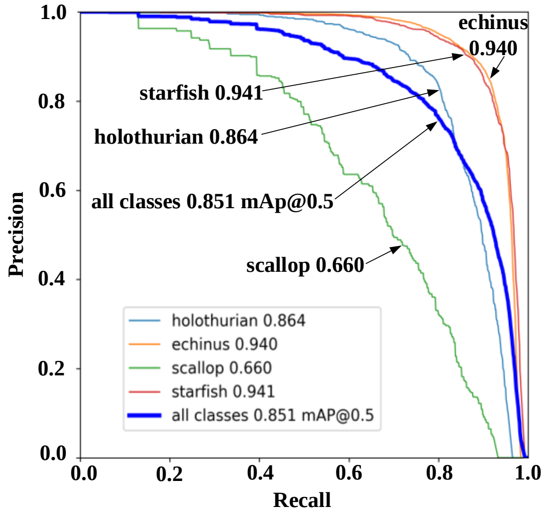


Fig. 10. P-R curve of YOLOv7 model

E. Comparative Experimental Results of Different Models

To demonstrate the superiority of the YOLOv7-RNCA model, we trained and tested it on the URPC dataset, using mAP as an evaluation metric, and compared it with currently popular object detection models such as YOLOv5s, YOLOv6n, YOLOv7, YOLOv8m, and YOLOv8l. The findings of the analogous comparison are outlined in Table I. As evident from the table, the YOLOv7-RNCA model surpasses the performance of alternative detection methods, achieving a 1.5% higher mAP than YOLOv7, and 3.9%, 3.7%, 4.1%, and 4.4% higher mAP than YOLOv8m, YOLOv8l, YOLOv6n, and YOLOv5s, respectively. These experimental results Exhibit the practical applications and advantages of this method of Recognition of underwater objects.

TABLE I
THE PERFORMANCE OF DIFFERENT MODELS ON THE URPC DATASET

Method	Precision	Recall	mAP@0.5	mAP@0.9
YOLOv5s	85.7%	75.5%	82.2%	64.8%
YOLOv6n	85.1%	76.1%	82.5%	63.7%
YOLOv7	84.5%	76.7%	85.1%	64.8%
YOLOv8m	84.2%	77.1%	82.7%	67.0%
YOLOv8l	88.2%	74.2%	82.9%	67.1%
YOLOv7-RNCA	85.7%	79.7%	86.6%	66.9%

F. Ablation Experiments of the URPC Dataset

To enhance the accuracy of YOLOv7-RNCA, various improvements were implemented and evaluated through ablation experiments. Specifically, the study initially substituted the ELAN module in network of the initial YOLOv7 model with an enhanced ELAN-PC module. This allowed us to assess the AP values across various category and the mAP values for a particular category within the dataset. Subsequently, building upon the original YOLOv7 model, ResNet_CA_PConv was integrated into the trunk network's terminus. This addition was evaluated by analyzing its mAP value. Furthermore, the ELAN-PC module of the neck network was integrated, and the combined effect of these two modifications was assessed by examining the AP and mAP values. Lastly, the original SPPCSPC structure was replaced with a newly designed SPPCSPC-GAM structure, and the AP and mAP values were observed after the integration of all three modifications. These findings are summarized in Table II.

G. Inference Result

Fig. 11 and Fig. 12 show the result inference diagrams of the YOLOv7 model and the YOLOv7-RNCA model. From these result diagrams, we can see that the detection correctness of sea urchins has markedly improved in most cases. Specifically, from Fig. 12, we can observe regarding the accuracy of sea stars is greatly enhanced in cases of occlusion, indicating that the YOLOv7-RNCA module presented efficiently achieved in this paper enhance the correctness of occlusion detection. Furthermore, from Fig. 11, we can also note that scallops were missed in the original YOLOv7 model, but they can be detected in this model with high accuracy, as further confirmed by Table II. Therefore, we indeed conclude that the improved network structure exhibits higher accuracy.

TABLE II
THE RESULTS OF THE ABLATION EXPERIMENTS FOR THE YOLOV7 MODEL ON THE URPC DATASET

Model	ELAN-PC	ResNet_CA_PConv	SPPCSPC-GAM	AP(echinus)	AP(holothurian)	AP(starfish)	AP(scallop)	mAP
YOLOv7	×	×	×	94.0%	86.4%	94.1%	66.0%	85.1%
	√	×	×	94.4%	86.9%	94.4%	67.4%	85.8%
	×	√	×	94.0%	87.3%	94.3%	68.5%	86.0%
	√	√	×	93.9%	87.3%	94.1%	70.3%	86.4%
	√	√	√	94.0%	88.2%	94.1%	70.3%	86.6%

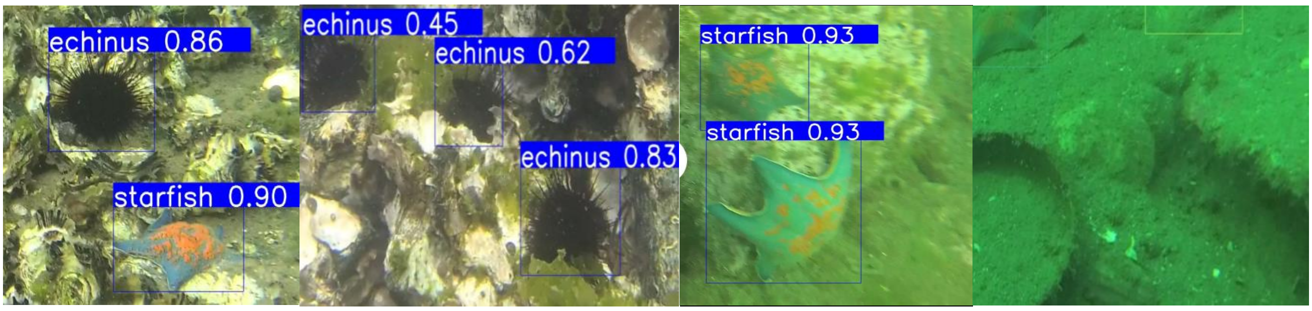


Fig. 11. The inference result on YOLOv7

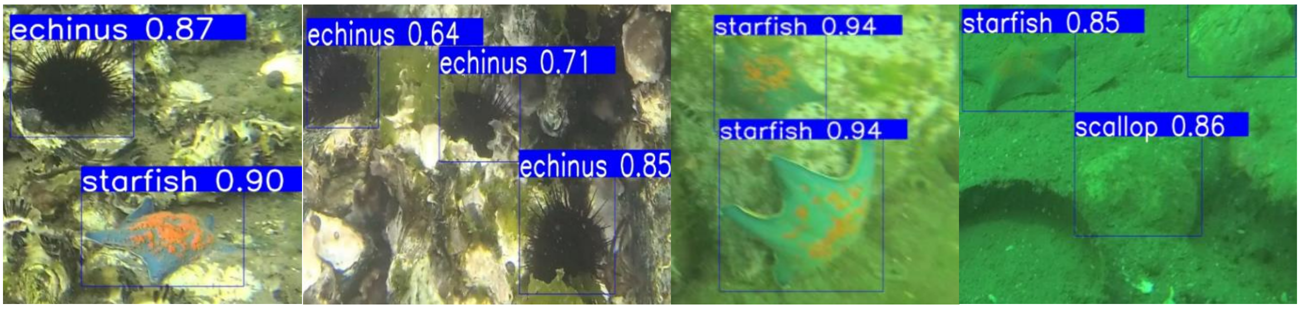


Fig. 12. The inference result on YOLOv7-RNCA

V. CONCLUSION

In the research herein, an advanced algorithm for underwater marine biological target detection, YOLOv7-RNCA, is proposed. This algorithm enhances the original YOLOv7 by integrating the PConv structure into ELAN, significantly improving both computational efficiency and memory access speed. Furthermore, the original SPPCSPC module has been enhanced to enhance its ability to extract features. Furthermore, the incorporation of the ResNet_CA_PConv module into the main network further enhances the overall accuracy of the detection model. Consequently, these improvements result in excellent detection accuracy for detecting underwater marine targets. However, submerged target detection still faces several challenges, such as scattered targets and unclear images that need to be addressed. We must overcome these dataset-related challenges and continue to make progress through continued efforts and improvements.

REFERENCES

[1] C. Costello, L. Cao, S. Gelcich, et al., "The future of food from the sea." *Nature*, 2020. 588, pp. 95-100.

[2] J. D. Santos, I. Vitorino, F. Reyes, F. Vicente, O. M. Lage, "From ocean to medicine: Pharmaceutical applications of metabolites from marine bacteria." *Antibiotics* 2020, 9, p. 455.

[3] A. Sahoo, S. K. Dwivedy, P. Robi, "Advancements in the field of autonomous underwater vehicle." *Ocean Eng*. 2019, 181, pp. 145-160.

[4] J. J. Leonard and A. Bahr, "Autonomous underwater vehicle navigation." in *Springer Handbook of Ocean Engineering*. 2016, pp. 341-358.

[5] J. C. Kinsey, M. R. Eustice, and L. L. Whitcomb, "A survey of underwater vehicle navigation: Recent advances and new challenges." in *Proc. IFAC Conf. Manoeuver Control Mar. Craft*, vol. 88, 2006, pp. 1-12.

[6] M. J. Er, J. Chen, Y. Zhang, W. Gao, "Research Challenges, Recent Advances, and Popular Datasets in Deep Learning-Based Underwater Marine Object Detection: A Review." *Sensors*. 2023; 23(4): p. 1990.

[7] J. Wang, P. Li, J. Deng, Y. Du, J. Zhuang, P. Liang and P. Liu, "CA GAN: Class condition attention GAN for underwater image enhancement." *IEEE Access*, vol. 8, pp. 130719-130728.

[8] J. Chen, S. Kao, H. He, et al., "Don't Walk: Chasing Higher FLOPS for Faster Neural Networks." *arXiv preprint arXiv*: 2303.03667, 2023.

[9] Q. Hou, D. Zhou and J. Feng, "Coordinate Attention for Efficient Mobile Network Design." *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, pp. 13708-13717.

[10] N. Yang, and J. Zhao, "Dangerous Driving Behavior Recognition Based on Improved YoloV5 and Open Pose." *IAENG International Journal of Computer Science*, vol. 49, no.4, pp. 1112-1122, 2022.

[11] Y. Liu, Z. Shao, "Hoffmann, N. Global attention mechanism: Retain information to enhance channel spatial interactions." *arXiv* 2021, *arXiv*: 2112.05561.

[12] R. Girshick, J. Donahue, T. Darrell, et al., "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 580-587.

[13] R. Girshick, "Fast R-CNN," *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440-1448.

[14] S. Ren, K. He, R. Girshick, et al., "Faster r-cnn: Towards real time object detection with region proposal networks." *Advances in Neural Information Processing Systems*, 2015, p. 28.

[15] P. Sermanet, D. Eigen, X. Zhang, M. Mathieu, R. Fergus, "LeCun, Y. Overfeat: Integrated recognition, localization and detection using convolutional networks." *arXiv* 2013, *arXiv*: 1312.6229.

[16] J. Redmon, S. Divvala, R. Girshick, et al., "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779-788.

[17] W. Liu, D. Anguelov, D. Erhan, et al., "Ssd: Single shot multi box detector." *Computer Vision-ECCV 14th European Conference*. 2016, pp. 21-37.

[18] T. Lei, X. Hong, W. Han, et al., "Research on Collaborative Object Detection and Recognition of Autonomous Underwater Vehicle Based on YOLO Algorithm," *2021 33rd Chinese Control and Decision Conference (CCDC)*. 2021, pp. 1664-1669.

[19] M. Zhang, S. Xu, W. Song, Q. He, Q. Wei, "Lightweight underwater object detection based on yolo v4 and multi scale attentional feature fusion." *Remote Sens*. 2021, 13, p. 4706.

[20] J. Zhang, et al., "Marine Organism Detection Based on Double Domains Augmentation and an Improved YOLOv7." in *IEEE Access*, vol. 11, pp. 68836-68852, 2023.

[21] W. Yi and B. Wang, "Research on Underwater Small Target Detection Algorithm Based on Improved YOLOv7." in *IEEE Access*, vol. 11, pp. 66818-66827, 2023.

[22] X. Shen, H. Wang, T. Cui, et al., "Multiple information perception-based attention in YOLO for underwater object detection." *Vis Comput* 40, pp. 1415–1438, 2024.

[23] C. Y. Wang, A. Bochkovskiy, H. Y. M. Liao, "YOLOv7: Trainable bag of freebies sets new state of the art for real time object detectors." *arXiv* 2022, *arXiv*: 2207.02696.

[24] A. Z. Zeyuan and Y. Z. Li, "What Can ResNet Learn Efficiently, Going Beyond Kernels?" *Adv. Neural Inf. Psrocess. Syst*. 2019, p. 32.