Enhanced YOLOv5 for Efficient Marine Debris Detection

Shicheng Li，Xiaoxia Zhang，Ruiqing Shan

Abstract—**To address the issue of large model parameter size and computational complexity in existing garbage detection models deployed on underwater embedded devices or underwater mobile devices, we propose an improved YOLOv5 network based on lightweight mechanism. Firstly, lightweight C3-Faster and Ghost conv with smaller parameters and computational complexity are adopted to replace the original C3 module and some Conv modules in the YOLOv5s network. Secondly, a squeeze-and-excitation (SE) attention mechanism is embedded in the network to enhance feature extraction capabilities. Finally, the bounding box regression loss function is modified to *EIoU* loss function to achieve faster training convergence. Experimental evaluations were conducted on the Trash-ICRA19 dataset. The results indicate that the mean Average Precision of the optimized algorithm reached 98.3%. Compared to the original YOLOv5s, the optimized algorithm reduced the parameter size by 35% and achieved a processing speed of over 40 frame/s, meeting the real-time detection requirements. This research indicates that the proposed improvement method can develop more efficient underwater garbage detection models for embedded devices or mobile devices, providing better support for real-time marine debris detection.**

Index Terms—**Underwater garbage, Yolov5s, C3-Faster, GhostNet, Attentional mechanism, *EIoU* loss function**

I. Introduction

THE entire world is currently grappling with a growing and serious problem: pollution caused by marine debris. The presence of garbage in underwater environments not only poses a significant threat to ecosystems but also has indirect negative impacts on human society. Consequently, this issue demands immediate attention and resolution. Marine debris is pervasive across various marine habitats, including beaches, sea surfaces, seabeds, and marine life. Currently, mainstream approaches to studying marine debris cleanup have focused on treating beaches and floating debris; the reality is that nearly seventy percent of the waste sinks to the seabed [1]. Even low-density polymers, when combined with the weight of biofouling, can lose buoyancy and sink to the seabed [2]. However, due to the unique challenges posed by underwater operations, cleanup efforts often rely on manual methods, which inevitably result in higher costs and certain risks. Consequently, the cleaning work of underwater trash still faces significant challenges.

Since the construction of the CNN network AlexNet by Hinton in 2012 [3], deep learning has made an entrance into an era of rapid expansion. A highly influential event occurred in 2016 when AlphaGo, utilizing deep learning techniques, defeated the world champion in the game of Go [4]. This brought artificial intelligence into the public spotlight and sparked global attention and extensive discussions on deep learning. Due to its advantages in adaptability, data-driven nature, scalability, and high precision [5], deep learning has been proven to be a highly effective method for replacing traditional manual labor in numerous fields. With the increasing maturity of autonomous underwater vehicles in the hardware domain, using autonomous intelligent machines to replace manual labor for marine debris detection and cleaning has become an efficient method for underwater garbage removal.

An excellent detection algorithm can provide real-time and reliable target information to machines, assisting them in completing garbage recognition and detection tasks. This paper proposes optimization and improvement research based on YOLOv5s target detection algorithm, aiming to solve the limitations of hardware computing power and real-time detection in actual mobile device applications. Through lightweight and high-precision optimization, it provides technical support for accurate and fast marine debris cleanup. This article's subsequent sections are organized as follows: Section Two outlines the current research on underwater target detection. Section Three briefly introduces the YOLOv5s algorithm, focusing on its lightweight networks, FasterNet and Ghost Convolution. It also discusses improvements like SE attention and the modified *EIoU* loss function. Section Four details the experimental results and provides an investigation of the enhanced algorithm. Lastly, Section Five draws conclusions from the experiments and proposes future enhancements.

II. Related work

Traditional object detection algorithms typically utilize sliding windows or region proposal methods for object detection. These methods involve dense region searching and feature extraction on the image, resulting in high computational complexity and slow processing speeds. They are also sensitive to changes in object scale and orientation, making it challenging to perform detection in complex

Manuscript received Sep 9, 2023; revised Jun 8, 2024.

S. C. Li is a Postgraduate Student of School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, AnShan, 114051, China (e-mail: 2741594782@qq.com).

X. X. Zhang is a Professor of School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, Anshan, 114051, China (corresponding author, phone:86-0412-5929812; e-mail: aszhangxx@163.com).

Ruiqing Shan is a Postgraduate Student of School of Computer Science and Software Engineering, University of Science and Technology LiaoNing, AnShan, 114051, China (e-mail: 1310846716@qq.com).

environments. Still, deep learning has become a hot topic in object detection due to its excellent robustness and effective representation of image detection features. DL algorithms exhibit strong capabilities in handling complex underwater scenes and have the potential to overcome the limitations of traditional methods [6].

In order to identify underwater debris, Valdenegro-Toro [7] used a convolutional neural network that had been developed on proactive sonar imagery, yielding an accuracy of approximately 80%. However, this approach relied on a simulated dataset created by introducing typical objects encountered in marine debris into a water tank and capturing forward-looking sonar images. While this study showcased the effectiveness of CNN and other deep models in identifying small-scale marine debris, their suitability in natural marine environments remains uncertain.

In 2019, Lin et al. proposed the ROIMIX image augmentation technique to address the issue of overlapping or occluded underwater biological targets [8]. This method utilizes the fusion of candidate bounding boxes and the ROI module to improve the model's detection capabilities. There has been a slight improvement in detection accuracy for different data. However, when it comes to underwater debris targets, they are typically distributed in a more scattered manner. Even if some debris targets overlap, it does not significantly affect the detection of underwater debris. Moreover, there are already existing data augmentation methods that can simulate target overlap. Therefore, in the context of detection algorithms, enhancing the detection capabilities of garbage targets is more crucial than addressing the issue of overlap.

In 2021, Shi et al. [9] introduced a method that employed ResNet as the backbone for feature extraction in Faster R-CNN. By implementing a bidirectional feature pyramid network, they achieved substantial advancements in both feature extraction and multi-scale feature fusin, leading to a notable improvement in underwater object detection accuracy to 88.94%. However, the frame per second achieved was only 4.3, which falls short of meeting real-time detection requirements. Therefore, there is a need for further exploration of lightweight networks and model compression techniques to achieve faster and more efficient underwater object detection.

In 2022, Wei et al. [10] introduced an enhanced architecture based on U-Net for underwater image semantic segmentation. They employed deeper contraction and expansion pathways to achieve end-to-end image semantic segmentation, resulting in improved detection accuracy and speed. However, it should be noted that the dataset created for underwater debris semantic segmentation in their experiments was relatively small. It consisted of only 410 training images generated from 205 images after undergoing image enhancement, along with over 50 images used for testing. The model reached saturation after 30 training epochs during the training process. Therefore, building and training the model across a larger dataset is advised in order to better imitate genuine marine settings.

In summary, the detection accuracy of underwater object identification techniques based on deep learning has improved. However, these algorithms typically produce a substantial number of parameters and fail to satisfy the real-time demands of useful applications. Additionally, the adoption of big models is hampered by the limited processing and storage power of mobile devices. Lightweight networks were established to resolve these issues. In this study, the first step is to switch out the YOLOv5 backbone network's C3 module with FasterNet, a lighter network. Taking advantage of its lower parameter and computational requirements to achieve preliminary lightweightization. Second, in the feature fusion section, we use the Ghost module from GhostNet in place of the Conv module. Similar feature maps may be produced with the Ghost module using less expensive linear methods, further achieving lightweightization of YOLOv5s. We embed the SE attention module with fewer parameters and computational requirements in the improved network to enhance detection accuracy. Ultimately, the *EIoU* loss function has replaced the bounding box regression loss, resulting in faster convergence and higher regression accuracy.

III. Network Architecture

Glenn Jocher introduced the YOLOv5 object detection algorithm in 2020, presenting significant improvements in both speed and accuracy when compared to its predecessors within the YOLO series. The main idea behind the algorithm is to immediately perform item detection and classification on an image by feeding it into a neural network using deep learning techniques, without the need for pre-generated candidate bounding boxes as in traditional methods. Specifically, the input image is divided into S x S grids. B-bounding boxes are predicted by each grid, along with their class probabilities and confidence ratings. Subsequently, overlapping bounding boxes are removed using non-maximum suppression (NMS), yielding the final discovered items [11].

The article chooses YOLOv5s as the basic model for garbage detection and classification. Four components make up the YOLOv5s network: input, neck, backbone, and prediction [12]. The network architecture is illustrated in Fig. 1.

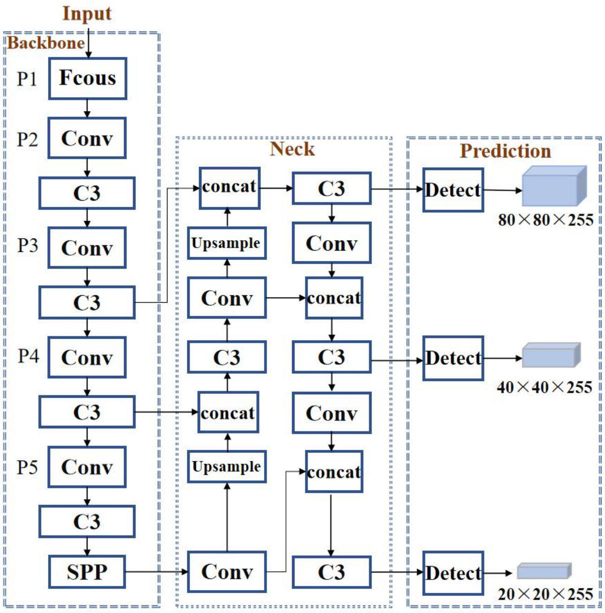


Fig. 1. YOLOv5s network structure

In the first part, input techniques such as mosaic data augmentation, adaptive image scaling, and automatic calculation of optimal anchor box values are employed. In the second part, Backbone, the main layers consist of Focus, CBS (Conv+Batch Normalization+SiLU), C3, SPP, and other modules, which are in charge of extracting features from the input images. A Feature Pyramid Network (FPN) and a Path Aggregation Network (PAN) form the neck of YOLOv5s, generating a feature pyramid that enhances the fusing of neck network features. The bounding box regression loss function in the prediction section is the *CIoU* loss. This section will introduce our optimization work on YOLOv5s in detail, focusing on four parts. (a) Although the YOLOv5s' original backbone network is capable of strong feature extraction, the great similarity of several convolutions leads to a lot of duplication in feature maps. To reduce the model parameters, the YOLOv5s backbone replaces the original C3 module with a lightweight C3-Faster module, achieving an initial lightweight model. (b) Introduce the Ghost module in GhostNet to replace the Conv module in the feature fusion part. The feature of the Ghost module is that it can generate similar feature maps with low-cost linear operations. With this architecture, the model's computational load is further decreased, and the network is further made lighter. (c) describes adding parameters and calculations to the network. Fewer SE attention mechanisms are used to enhance the extraction effect of lightweight backbones. (d) For faster training convergence and better positioning accuracy, change the *CIoU* loss to the *EIoU* loss. After the above adjustments, the model shows higher robustness and accuracy when handling complex scenes.

A. FasterNet Network

The C3 module is an important component in the YOLOv5s backbone network. However, due to its shortcomings, such as the large number of parameters, slow detection speed, and limited application, we propose a new C3-Faster module based on the FasterNet network [13]. The C3-Faster module combines the idea of a lightweight FasterNet network and uses the FasterNet Block structure to replace the Bottlenecks structure in the original C3 module. This modification seeks to improve the model's efficiency by maintaining the network's depth while reducing the module's computational complexity and parameter count.

FasterNet is a brand new neural network family proposed by Chen et al. in 2023. It surpasses rival networks' operating speeds on devices by a large margin while maintaining the accuracy of different visual tasks. The network re-examines the existing operators, especially the computational speed of DWConv-FLOPS. The results of the investigation showed that the operators' frequent memory access, especially during depthwise convolution, is the main reason for the poor FLOPS. In order to reduce unnecessary calculations and memory accesses while also extracting spatial characteristics more efficiently, a unique partial convolution (PConv) is presented. Based on PConv, FasterNet is further proposed. This innovation significantly improves the running speed of the network while maintaining its accuracy.

Fig. 2 shows the general design of FasterNet. The network shows a simple PConv module that simply applies a conventional convolution operation to a selected number of input channels, keeping the remaining channels unchanged, in order to extract spatial information. The first or final contiguous channels are regarded as indicative of the full feature map for calculation when using sequential or regular memory access. It is considered that the input and output feature maps' channel counts are the same without sacrificing generality. Memory accesses and computational redundancies are decreased with the addition of the PConv module. The FasterNet Block is suggested using PConv; each block consists of two Conv 1x1 layers after a PConv layer. Together, they are shown as inverted residual blocks, with a shortcut to reuse input features and an increased number of channels in the intermediate layer. Finally, a novel neural network family, FasterNet, is proposed via FasterNet Block, which consists of 4 layers. For spatial downsampling and channel expansion, an embedding layer (a conventional 4x4 convolution with a 4 stride) or a merging layer (a conventional 2x2 convolution with a 2 stride) comes before each level. FasterNet blocks are stacked in each level; moreover, because of their greater FLOPS consumption and reduced memory usage, more

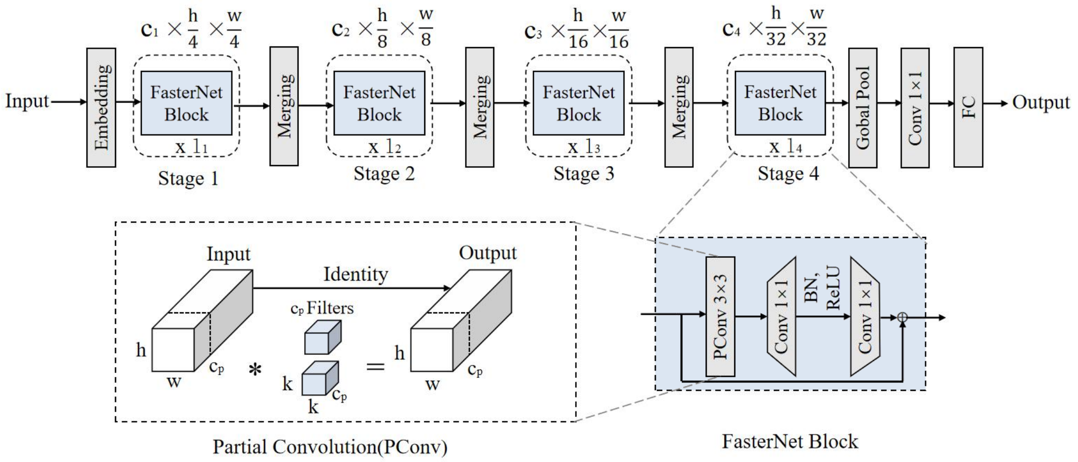


Fig. 2. FasterNet network structure.

FasterNet blocks are positioned in the final two stages. Accordingly, more computations are allocated to the last two stages.

With the help of the Fasternet block structure, a C3-faster module is proposed. In comparison to the previous C3 module, the new C3-faster requires less computing power and has fewer parameters. It can reduce the number of parameters in the algorithm and the computational load while maintaining good detection accuracy. This achieves the goal of creating a preliminary lightweight model and improves the performance of marine debris detection to some extent.

B. GhostNet

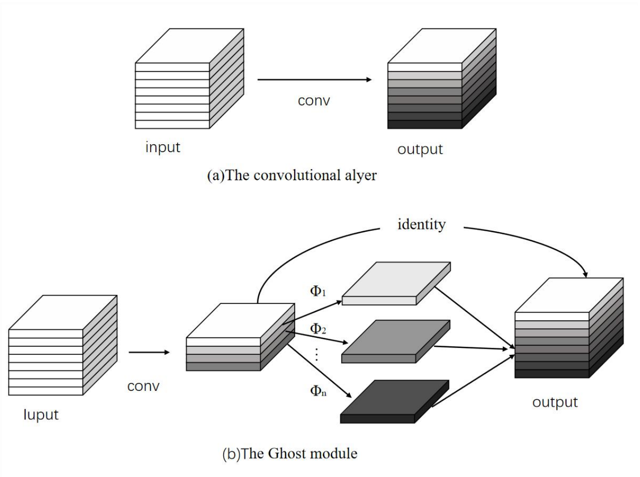
To further lightweight the algorithm, we introduced Ghostconv from GhostNet in the feature fusion part. GhostNet is also a lightweight network that transforms the heavy convolutional operations into generating a few highly diversified feature maps [14]. Then, it applies less expensive operations compared to regular convolutions to transform these feature maps and obtain similar feature maps. There are two components to the Ghost module's implementation: a regular convolution and a linear operation with fewer parameters and computations. Firstly, a portion of the feature maps is obtained through a limited regular convolution. Then, the obtained feature maps are further expanded by the linear operation to get more feature maps. Ultimately, the two feature map sets are combined along the designated axis. Fig. 3 compares regular convolution with ghost convolution.

Assuming the given input data X $\in R^{c\times h\times w}$, where h and w stand for the height and breadth of the information being input, respectively, and c is the number of input channels, and $Y \in R^{h'\times w'\times n}$ represents n feature maps with heights h' and widths w', the quantity of convolutional filters is n, with k as the kernel size, and the linear transformation has a kernel size of d and a transformation count of s. In the absence of bias terms b, the parameter compression ratio achieved by replacing traditional convolutions with Ghost convolutions can be derived as shown in equation (1).

$$r_c = \frac{c\cdot n\cdot k\cdot k}{c\cdot k\cdot k\cdot \frac{n}{s}\cdot +(s-1)\cdot d\cdot d\cdot \frac{n}{s}} \approx \frac{c\cdot s}{c+s-1} \approx s \quad (1)$$

The acceleration ratio can be derived as shown in equation (2):

$$\begin{aligned} r_s &= \frac{c\cdot n\cdot k\cdot k\cdot h'\cdot w'}{c\cdot k\cdot k\cdot \frac{n}{s}\cdot h'\cdot w' + (s-1)\cdot d\cdot d\cdot \frac{n}{s}\cdot h'\cdot w'} \\ &= \frac{c\cdot k\cdot k}{c\cdot k\cdot k\cdot \frac{1}{s} + d\cdot d\cdot \frac{(s-1)}{s}} \approx \frac{c\cdot s}{c+s-1} \approx s \end{aligned} \quad (2)$$



(a)The convolutional alyer

(b)The Ghost module

Fig. 3. GhostNet network structure.

From equations (1) and (2), it is evident that the computational expense of Ghost convolution is around s times greater than that of ordinary convolution, and the computational cost for an equivalent set of parameters is also roughly s times. The benefits of computational acceleration and parameter compression are influenced by the transformation count. Specifically, the more "ghost" feature maps generated, the better the acceleration effect, but it may lead to a decrease in detection accuracy. The transformation count is often set at 1/2 in order to achieve a compromise between speed and precision.

C. Squeeze-and-Excitation

The attention mechanism is a visual focus mechanism that simulates the rapid acquisition of key information and filtering of irrelevant information in the human brain. It aims to quickly extract the crucial features from an image. Attention mechanisms are often employed in computer vision to improve neural networks' feature extraction performance. By assigning weights to the input, attention mechanisms can amplify or emphasize the important feature information in the image, making it a parameterized pooling method. Multiple experiments have shown that the SE attention mechanism, with fewer parameters and computational requirements, significantly enhances our optimization approach compared to frequently used attention mechanisms such as CBAM, CA, and ECA [15]. Fig. 4 depicts the SE attention mechanism's structure.

It primarily involves two steps: squeeze and excitation. During the squeeze, the SE attention mechanism lowers the feature maps' dimensionality through worldwide average pooling, capturing the global average for each channel. By doing this, the feature maps' spatial dimensions are decreased, yet the information unique to a specific channel is retained. In the excitation step, the global average values are processed through two fully connected layers, introducing non-linear transformations. These completely linked layers pick up on each channel's weight significance [16]. Typically, the *ReLU* activation function is employed to introduce non-linearity, thereby enabling the model to

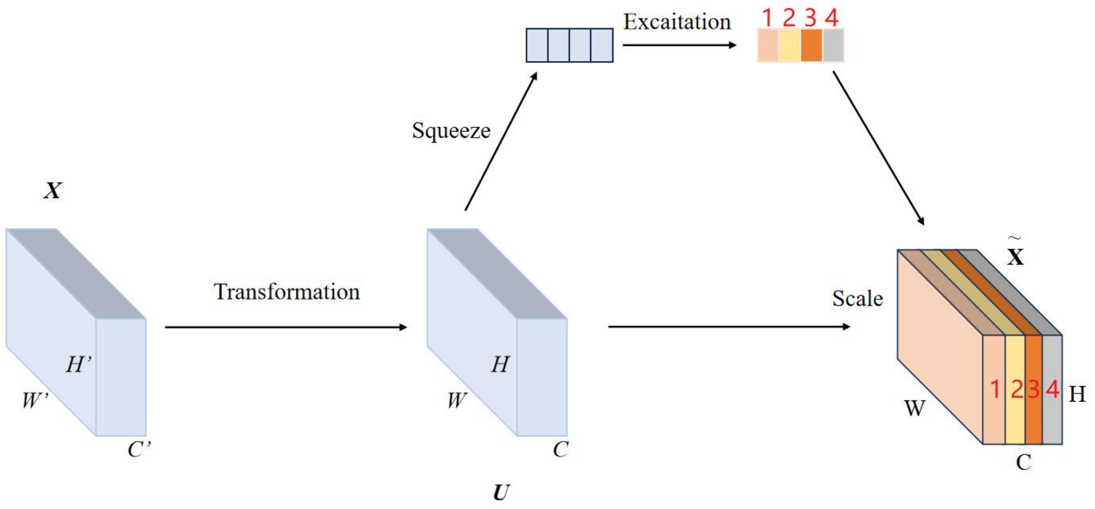


Fig. 4. Squeeze-and-Excitation

capture complex relationships between channels. To give major feature channels extra weight, the computed channel weights are finally multiplied by the original feature maps. This enhances the representation capability of these channels and suppresses unimportant channels.

D. Optimization of the Loss Function

The loss function of the YOLOv5s model includes classification loss, localization loss, and target confidence loss. Yolov5s employs the binary cross-entropy loss function for the class probability score, the C*IoU* loss for the bounding box, and the BCEclsloss loss function for the loss of objectivity score. The *CIoU* loss equation is shown in equation (3), where υ and α are shown in equations (4) and (5), respectively. w^{gt} and h^{gt}.

$$CIoU = IoU - (\frac{\rho^2}{C^2} + \alpha\upsilon) \tag{3}$$

$$\upsilon = \frac{4}{\pi^2}(tan^{-1}\frac{w^{gt}}{h^{gt}} - tan^{-1}\frac{w}{h})^2 \tag{4}$$

$$\alpha = \frac{\upsilon}{\upsilon + (1 - IoU)} \tag{5}$$

Equation (3) shows the crossing point of the combination of the expected box and the ground truth box. The Euclidean separation between the two boxes' centers is denoted by the symbol ρ, while the diagonal distance of the smallest enclosing rectangle that contains both boxes is represented by the symbol c. The width and height of the surface of the truth box are denoted by w^{gt} and h^{gt} in equation (4), while the width and height of the predicted box are represented by w and h.

However, Yolov5 uses *CIoU* as a loss function, which will have some shortcomings. For example, the aspect ratio's description is a relative value, and there is a certain degree of ambiguity. Also, the calculation of *CIoU* is more complicated. It takes into account parameters such as the union and intersection ratios, the height and breadth of the bounding box, and the center point. The advantage of this is that it makes the prediction box's direction of convergence more precise, thereby further speeding up the convergence. However, when the aspect ratio factor υ in the *CIoU* calculation equation calculates the gradient for w and h, it will be found that the two have opposing gradient directions., as shown in equations (6) and (7):

$$\frac{\partial\upsilon}{\partial w} = \frac{8}{\pi^2}(tan^{-1}\frac{w^{gt}}{h^{gt}} - tan^{-1}\frac{w}{h}) * \frac{h}{w^2 + h^2} \tag{6}$$

$$\frac{\partial\upsilon}{\partial h} = -\frac{8}{\pi^2}(tan^{-1}\frac{w^{gt}}{h^{gt}} - tan^{-1}\frac{w}{h}) * \frac{w}{w^2 + h^2} \tag{7}$$

This results in the illogical consequence that every time the parameters are changed using the random gradient descent approach, the neural network changes w and h in the opposite ways. This increases computational complexity, especially in large-scale object detection tasks, resulting in higher computational costs. This does not meet the goals of our lightweight model. Therefore, the *EIoU* function is deployed as the new loss function, and its calculation is shown in equation (8):

$$EIoU = IoU - \left((\frac{\rho}{C})^2 + (\frac{\rho_w}{w^c})^2 + (\frac{\rho_h}{h^c})^2\right) \tag{8}$$

In equation (8), w^c and h^c stand for the height and breadth of the smallest enclosing rectangle that includes the actual and expected frames. respectively, ρ_w and ρ_h represent the width difference and height difference between the two frames. *EIoU* calculates the loss independently for width and height, so ideal convergence can be achieved to improve localization accuracy.

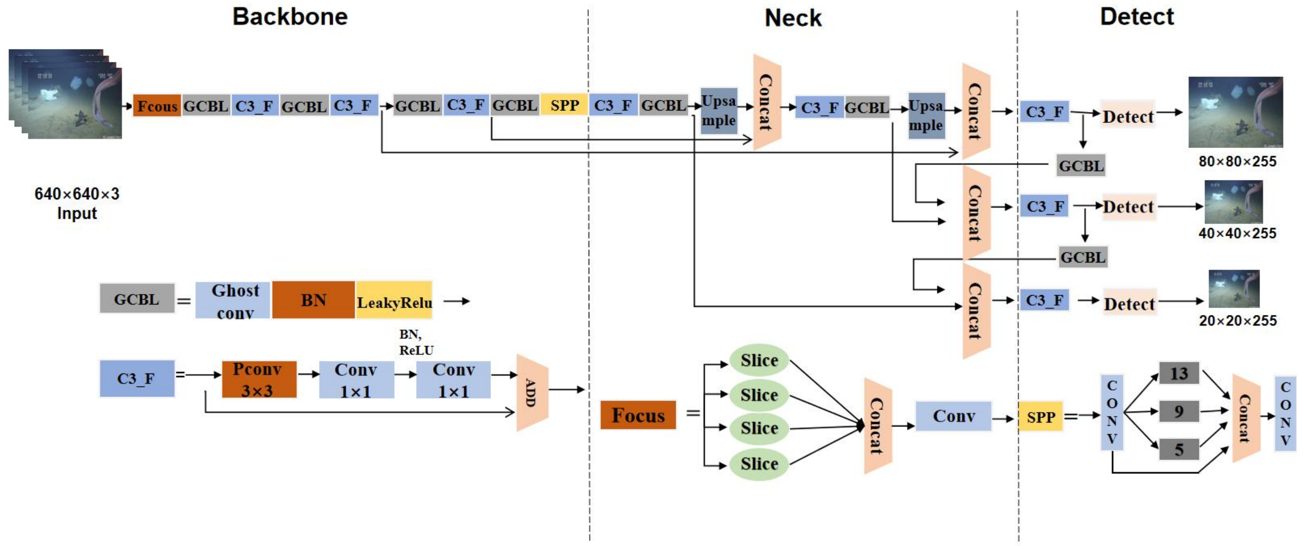


Fig. 5. Improved YOLOv5s network structure.

E. *Improved YOLOv5s Network Structure*

In this article, the backbone of YOLOv5 is replaced with the lightweight C3-Faster module, which initially reduces the computational load and complexity of the model. With fewer parameters in the new model, this substitution improves inference time without sacrificing performance. Further improvements involve substituting Ghost convolutions for the regular convolution modules in the YOLOv5s network. Ghost convolutions maintain model performance to a certain extent while decreasing computational costs, particularly in mobile devices or resource-constrained environments, aiming to further reduce the model's parameter count. Embedding SE attention modules in the network structure helps enhance the model's focus on crucial features, improving its performance in object detection tasks. By increasing the emphasis on learning and leveraging key information, the model is able to enhance detection accuracy. Lastly, utilizing the *EIoU* loss helps better reflect differences in the position and dimensions of target boxes, improving the model's prediction accuracy for overlapping objects or those with significant size variations in object detection, thereby enhancing overall detection performance. This process results in an optimized algorithmic model. Fig. 5 displays the enhanced YOLOv5s network structure.

IV. EXPERIMENT

A. *Datasets and Experimental Environment*

This research employs the Trash-ICRA19 dataset for marine debris detection. The images within this dataset are derived from the J-EDI dataset, which is overseen by the Japan Agency for Marine-Earth Science and Technology (JAMSTEC). The videos comprising this dataset exhibit significant variations in terms of quality, depth, scene objects, and the cameras used. They include photos of several kinds of marine trash taken in actual settings. Furthermore, the clarity of the water and the quality of the lighting vary across the videos. In total, the dataset includes over 7600 images, encompassing organisms, garbage, and ROVs. Prior to conducting experiments, the sample photos were split into training, validation, and test sets at random in an 8:1:1 ratio.

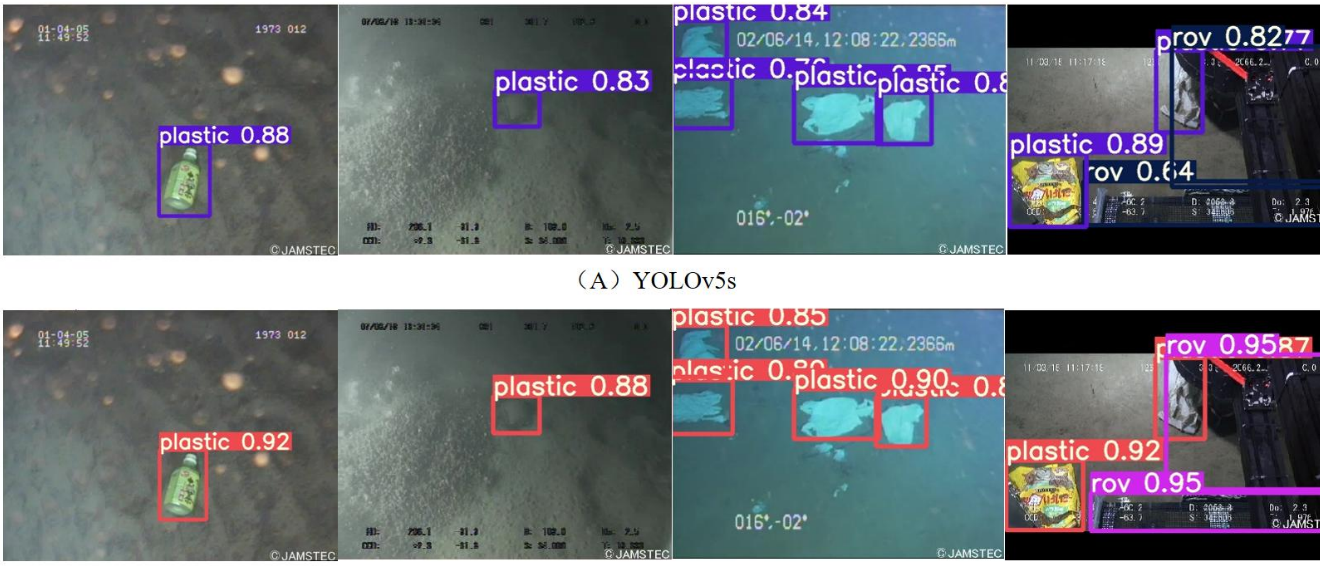
TABLE I
THE COMPUTER CONFIGURATION

Project	Context
CPU	Intel(R) Core(TM) i7-9750H
RAM	16GB
GPU	NVIDIA GeForce GTX 1660ti
Operating System	Windows10
Cuda	Cuda10.1
Data Processing	Python3.8
DL Framework	Pytorch1.7

Table I lists all of the computer configurations for the experimental setup. In order to train the model, the batch size is set to 16, the initial learning rate is adjusted to 0.01, the momentum is adjusted to 0.937, the training runs for 100 epochs, and the weight decay is adjusted to 0.0005.

B. *Evaluating Metrics*

In this paper, we will evaluate the lightweightness of a model based on its FLOPs and parameter count. The FLOPs refer to the total number of multiplication and addition operations required during the forward inference of a model, reflecting the model's demand for hardware



（A） YOLOv5s

（B） Improved YOLOv5s

Fig. 6. Detection results chart of yolov5s and improved algorithm

computational units, and they are commonly utilized to calculate the model's size. Parameter count refers to the total sum of parameters in a model, which affects both memory usage and program initialization time.

Additionally, the experiment also included accuracy, recall, average precision (AP), and mean average precision (mAP) as assessment criteria. The following formulas can be used to determine these metrics:

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

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

$$AP = \int_0^1 P(R)dR \tag{11}$$

$$mAP = \frac{1}{C}\sum_{c\in C} AP(c) \tag{12}$$

The number of marine debris samples that are accurately identified is represented by True Positives (TP) in the equations above. The quantity of marine debris samples that were mistakenly identified is called False Positives (FP). The number of samples that are marine debris but were missed by the model is represented by False Negatives (FN). For marine debris samples, assessment criteria like AP (Average Precision) and mAP (Mean Average Precision) are frequently employed. The number of samples that are marine debris but were missed by the model is represented by False Negatives (FN). Evaluation measures like Average Precision (AP) and Mean Average Precision (mAP) are frequently employed in object identification systems. The area under the precision-recall curve is used to compute average precision, or AP. With C standing for the total number of classes, mAP is the average of AP values across all classes or categories.

FPS was another metric used to evaluate the model's detection speed. In the field of video surveillance, a higher FPS is typically required to achieve real-time object detection and tracking in order to promptly detect and respond to abnormal events. Typically, the frame rate of real-time cameras is between 24 fps and 30 fps. Therefore, it is desirable to have a range of 30FPS or higher to facilitate real-time garbage detection.

C. Comparison Results and Analysis

Yolov5's detection results and the enhanced method are displayed in Fig. 6. Comparative tests were carried out in the same environment to evaluate the performance of this technique against the existing mainstream object identification models in order to confirm the efficacy of the underwater rubbish detection method suggested in this study. This experiment featured the lightweight network YOLOv5n, the YOLOv5s network with its backbone network changed to ShuffleNetV2, the standard single-stage object detection technique SSD, and the classic two-stage object detection method Faster R-CNN. Table II displays the results that were achieved.

Table.II RESULTS OF COMPARATIVE TEST

Network	AP（%）	FPS	GFLOPS	Paramenter
SSD	96.1	26	1.6	3941314
Faster R-CNN	98.9	0.6	300	28295818
Yolov5-shufflenet	92.6	32	7.8	860813
Ours	97.9	40	10.7	4636584

Table II shows that although the SSD network has a low processing need for devices and a straightforward framework, its device identification speed is somewhat slow, making it unable to satisfy real-time demands. Faster R-CNN achieves the highest detection accuracy, yet its massive network parameters make it challenging to apply to portable devices with constrained processing power and storage resources. YOLOv5n demonstrates good real-time detection speed, but its detection accuracy is lower among

Table III RESULTS OF ABLATION STUDY

Model	mAP（%）	FPS	Inference	GFLOPS	Paramenter
YOLOv5s	96.0	37	24.3	16.5	7068936
FasterNet	95.5	38	24.1	13.1	5816136
GhostNet	95.4	38	24.2	14.1	5856616
Fasternet+GhostNet+SE	96.4	40	23.9	10.7	4603816
Ours	98.3	40	23.9	10.7	4636584

the aforementioned algorithms. In contrast, the current mainstream improved method, ShuffleNetV2, is replacing the YOLOv5s backbone network method, which, in terms of detection speed and accuracy, is slower than our suggested approach. In conclusion, the method presented in this study offers better advantages in terms of overall performance when compared to other object detection techniques.

In this study, several enhancements are made to the YOLOv5s model: the use of C3-Faster; the incorporation of the SE attention mechanism; the replacement of the conventional convolution with Ghost convolution; and the optimization of the loss function. An ablation test was carried out and compared with the original model in order to confirm the contribution of each component, as indicated in Table III.

Table III indicates that YOLOv5s achieves very high detection accuracy, but it also has a lot of network parameters and requires substantial computational resources. It is not conducive to use on embedded mobile terminals. But, through the improvement of methods such as FasterNet and GhostNet, the model's parameter count can be greatly compressed and its computational speed can be increased, but this may also lead to a slight reduction in the mAP value.Therefore, we modified the loss function and added the SE attention mechanism with fewer parameters, which has improved the accuracy of detection. Fig. 7 shows the P-R curve of the improved Yolov5s algorithm.

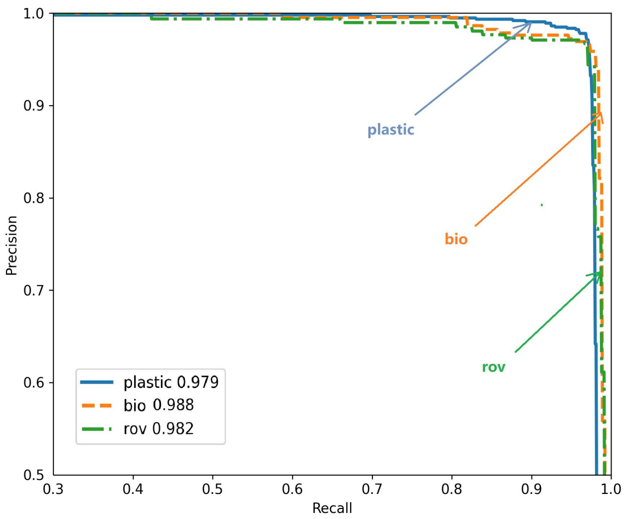


Fig. 7. P-R curve of the improved YOLOv5s algorithm

V. CONCLUSIONS

A lightweight and high-precision optimization study was conducted on the YOLOv5s object detection algorithm. The enhanced YOLOv5s approach considerably lowers the processing needs and parameter count of the network, according to the study's findings, thus meeting the computational and memory constraints of underwater mobile devices while maintaining excellent detection accuracy. This method demonstrates its capability to fulfill the accuracy requirements of complex underwater environments in detection systems, finding a medium ground between speed and precision. As a result, it successfully accomplishes the goals of this study and improves the functionality of underwater trash detection systems.

Nonetheless, there are several drawbacks to this strategy. Future research ought to focus on the following problems: (1) Implementing detection before image augmentation, such as using next-generation networks to expand the dataset, followed by a deep unsupervised quality assessment method to evaluate and select excellent pictures for use as training examples. (2) Accurately positioning targets for underwater robots involves combining multidimensional localization data with multifaceted object recognition.

REFERENCES

[1] Y. W. Cheng, J. N. Zhu, M. X. Jiang, J. Fu, C. S. Pang, P. D. Wang, K. Sankaran, O. Onabola, Y. M. Liu, D. B. Liu and Y. Bengio, "FloW: A Dataset and Benchmark for Floating Waste Detection in Inland Waters," Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10953-10962, 2021.

[2] R. Coyle, G. Hardiman, Driscoll, "Microplastics in the marine environment: A review of their sources, distribution processes, uptake and exchange in ecosystems," Case Studies in Chemical and Environmental Engineering, vol. 2, pp. 100010, 2020.

[3] Z. P. Zhang and H. W. Peng, "Deeper and Wider Siamese Networks for Real-Time Visual Tracking," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, pp. 4591-4600, 2019.

[4] F. Y. Wang, J. J. Zhang, et al, "Where does AlphaGo go: from church-turing thesis to AlphaGo thesis and beyond," IEEE/CAA Journal of Automatica Sinica, pp. 113-120, 2016.

[5] D. Yuan and Y. Xu, “Lightweight Vehicle Detection Algorithm Based on Improved Yolov4,” Engineering Letters, vol. 29, no. 4, pp. 1544–1551, 2021.

[6] C. M. Wu, Y. Q. Sun, T. J. Wang and Y. L. Liu, "Underwater Trash Detection Algorithm Based on Improved YOLOv5s," Journal of Real-Time Image Processing , vol. 19, no. 5, pp. 911–920, 2022.

[7] V. T. Matias, "Submerged Marine Debris Detection with Autonomous Underwater Vehicles," 2016 International Conference on Robotics and Automation for Humanitarian Applications (RAHA), pp. 1-7, 2016.

[8] W. H. Lin, J. X. Zhong, S. Liu, T. Li and G. Li. Roimix: Proposal-Fusion Among Multiple Images for Underwater Object Detection, International Conference on Acoustics, Speech, and Signal Processing(ICASS P). pp. 2588-2592, 2020.

[9] P. F. Shi, X. W. Xu, J. J. Ni, et al, "Underwater Biological Detection Algorithm Based on Improved Faster-RCNN," Water, vol. 13, no. 17, pp. 2420, 2021.

[10] L. Wei, S. Kong, Y. Wu, et al. "Image Semantic Segmentation of Underwater Garbage with Modified U-Net Architecture Model," Sensors, vol. 22, no. 17, pp. 6546 - 6557, 2022.

[11] B. Zhang, X. X. Zhang and Z. Li, "An Efficient Face Mask Wearing Detection Algorithm Based on Improved YOLOv3," Engineering Letters, vol. 30, no. 4, pp. 1493-1503, 2022.

[12] Z. H. Zheng, P. Wang, D. W. Ren, et al. "Enhancing Geometric Factors in Model Learning and Inference for Object Detection and Instance Segmentation," In IEEE Transactions on Cybernetics, vol. 52, no. 8, pp. 8574 - 8586, 2022.

[13] J. R. Chen, S. H. Kao, H. He, W. P. Zhuo, S. Wen, C. H. Lee and S. H. Chan. "Don't Walk: Chasing Higher FLOPS for Faster Neural Networks," In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12021–12031, 2023.

[14] K. Han, Y. Wang, Q. Tian, J. Guo and C. Xu, "GhostNet: More Features From Cheap Operations," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, pp. 1577–1586, 2020.

[15] J. Hu, L. Shen and G. Sun, "Squeeze-and-Excitation Networks," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 7132–7141, 2018.

[16] X. Zhu, S. Lyu, X. Wang and Q. Zhao, "TPH-YOLOv5: Improved YOLOv5 Based on Transformer Prediction Head for Object Detection on Drone-Captured Scenarios," 2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), Montreal, BC, Canada, pp. 2778–2788, 2021.