Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **9**, Number **1**, February **2024**

YOLOv5-LSMA: A Detection Algorithm for Deep-sea Plastic Garbage

Jin Han*, Shengxia Li

College of Computer Science and Technology
Shandong University of science and technology, Qingdao 266590, China
shnk123@163.com, 3504380998@qq.com

Chunhui Liu

Graduate School of Fisheries and Environmental Sciences
Nagasaki University, Nagasaki 852-8521, Japan
liuchunhuinanjing@163.com

*Corresponding author: Jin Han

Received September 12, 2023, revised November 16, 2023, accepted December 24, 2023.

Abstract. *With the development of industry and the passage of time, the problem of deep-sea garbage has become more and more obvious. To more effectively detect deep-sea garbage, this paper proposes a detection algorithm YOLOv5-LSMA based on YOLOv5 for deep-sea plastic garbage. Firstly, the loss-reduction down-sampling (LRDS) module is proposed to reduce the feature loss in the process of down-sampling in the neck network. Secondly, to improve the detection ability of the deep network, a multi-scale channel attention mechanism (MCAM) module is proposed, and the dilated convolution is introduced to make the deep network have a wider receptive field. Finally, to make full use of the detailed information on shallow features, a simplified bi-directional feature pyramid network (S-BiFPN) structure is constructed based on the simplification and improvement of the BiFPN structure. According to the experimental results, the detection precision of the YOLOv5-LSMA algorithm proposed in this paper is 4.7% higher than that of YOLOv5 algorithm, and the model detection capability is effectively improved. It provides an effective algorithm for the identification of deep-sea plastic garbage.*
Keywords: Underwater Plastic Garbage, Loss Reduction Down-sampling, Simplified BiFPN, Multi-Scale Channel Attention Mechanism, Dilated Convolution

1. **Introduction.** With the continuous development of industry, urbanization, and science and technology, the pollution of marine and lake environments is becoming more and more serious, especially the marine environment, where the garbage pollution in rivers and lakes will eventually flow into the ocean and aggravate marine environment pollution [1]. Deep-sea trash includes discarded fishing nets, woven bags, disposable lunch boxes, randomly discarded drinking bottles, discarded ropes, plastic bags, and various packaging bags, some of which float on the surface of the water and some of which are sunk at the bottom. According to statistics from the government's marine litter test conducted in 2021 in 51 regions of China, plastic litter is the most abundant among floating litter on the sea surface, beach litter, and bottom litter, accounting for 92.9%, 75.9%, and 83.3% of the total respectively, which also reflects the seriousness of the harm of plastic litter to the marine environment from the side. Since plastic waste is difficult to degrade naturally

and can remain in the water for at least hundreds of years, it not only pollutes the marine and lake environment, affects the marine landscape and marine navigation, but also gradually enters the food chain and threatens biodiversity [2, 3, 4, 5]. In recent years, there have been cases of marine organisms accidentally eating plastic bags by mistake and starving to death, which shows that plastic garbage poses a great threat to the life safety of marine organisms.

At present, the detection and cleaning of deep-sea garbage mainly rely on manual labor, which requires a lot of manpower and material resources, and the workable range is very limited. Compared with manual processing, machine vision-based detection methods can save a lot of labor and can work on a wider range of tasks with higher efficiency. However, traditional machine vision detection algorithms need to construct effective features extraction algorithms based on different features of the target to be detected, with relatively poor flexibility and generality. Deep learning algorithms are used to detect deep-sea garbage, which not only provides better flexibility and generality, but also achieves higher detection accuracy. So far, deep learning-based detection algorithms have not been widely used in deep-sea garbage detection. Improving the detection accuracy of the detection algorithm is the main goal of this study.

2. **Related Work.**

2.1. **Research status of deep-sea garbage detection based on deep learning and the history of YOLO development.**

In recent years, deep learning has been developing continuously, and the neural network has been widely used in various fields, such as medicine [6], industry [7], intelligent transportation [8], video salient region detection [9], image description [10], anomaly detection [11], etc. The existing deep-sea garbage target detection algorithms are divided into traditional and deep learning-based detection algorithms. Among them, neural network target detection algorithms are mainly divided into two categories: two-stage and one-stage detection algorithms [12]. In 2014, the R-CNN [13] two-stage detection algorithm was proposed by Girshick et al.. Subsequently, representative two-stage detection algorithms, such as Fast R-CNN [14] and FasterR-CNN [15], have been proposed. However, two-stage detection algorithms have poor real-time performance. Therefore, since 2016, Redmon et al. proposed the YOLO series of representative one-stage detection algorithms, such as SSD [16], YOLOv1 [17], YOLOv2 [18], YOLOv3 [19], and YOLOv4 [20]. The real-time performance of the detection algorithm is improved without reducing the detection accuracy. The details of the above YOLO series of algorithms are shown in Table 1.

The related research is not hot due to the small number of early large underwater visible image target detection datasets. Since 2017, the 3 most popular generalized target detection models have been directly applied to fish detection, outperforming traditional algorithms. In 2017, Sung et al. performed real-time vision-based fish detection using the YOLOv1 algorithm [21]. In 2018, Christensen et al. utilized the SSD algorithm for the detection, localization, and classification of fish and fishes under harsh conditions [22]. In 2018, Mandal et al. evaluated fish abundance in underwater videos using the Faster R-CNN algorithm [23]. Last few years, researchers have been using these detection algorithms to progressively work in the field of deep-sea garbage detection. Xue proposed a one-stage deep-sea garbage detection network ResNet50-yolov3 [24]. Tang and Gao proposed a surface floating garbage detection algorithm based on an improved convolution neural network [25]. Yuan and Zang proposed an underwater trash target detection algorithm based on attention mechanism Ghost-YOLOv5 [26].

Table 1. History of development of YOLOv1 toYOLOv4.

Network	Proposed timing	Proposer	Strengths	Weaknesses
YOLOv1	2016	Redmon et al.	Low computational volume, fast running speed.	Poor detection of dense targets
YOLOv2	2016	Redmon et al.	Favorable for small target detection	Poor detection of dense targets
YOLOv3	2018	Redmon et al.	The problem of small target detection is basically solved.	Computational volume has increased
YOLOv4	2020	Bochkovskiy et al.	Higher detection accuracy.	Computational volume has increased

Because the light attenuation degree is higher in the deep-sea environment than in the land environment, it affects the visibility of the target and makes the image blurring intensify, which in turn affects the detection performance of the model in the deep-sea environment and increases the detection difficulty. To address this problem, this paper proposes an improved deep-sea plastic waste detection algorithm YOLOv5-LSMA based on YOLOv5. Designing the loss reduction down-sampling module to reduce the feature loss caused by ordinary convolution down-sampling; use the improved BiFPN (bi-directional feature pyramid network) structure S-BiFPN (Simplify bi-directional feature pyramid network) to balance the contribution of each feature map to the network and extract more detailed information in the feature map; the multi-scale channel attention mechanism is proposed to improve the network attention, and dilated convolution is introduced to expand the deep network receptive field and improve detection performance.

2.2. **The Introduction of YOLOv5 Algorithm.** The YOLOv5 algorithm was proposed by Glenn Jocher in 2020 and is one of the typical representatives of the one-stage target detection algorithm. The algorithm directly extracts features on the input image and performs classification and regression on the feature map. Compared with the two-stage algorithm, its detection speed is faster and can better meet the real-time requirements of the network. According to the size of the model, YOLOv5 is divided into four different magnitude networks: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x, which differ only in network depth and width.

A 640×640 standard RGB image is passed through the backbone network Darknet53, which is down-sampled 4 times, 8 times, 16 times, and 32 times in turn to obtain effective feature layers of sizes 160×160, 80×80, 40×40 and 20×20, respectively. Different feature layers are equivalent to dividing the image into three (R, G, and B channels) SS (S is the number of grids) grids of different scales. Each grid node is responsible for predicting the target in its lower right corner and if the center of the detected target falls within a grid, the node responsible for detecting this grid will detect the target. Three anchors of

different sizes and N (N is the number of categories) categories probabilities are generated for each grid node. Each anchor consists of five parameters, namely the offset (x, y) of the center point of the prediction box relative to the center point of the grid, the width and height scaling ratio w, h of the prediction box relative to the grid, and the confidence. Finally, the redundant prediction boxes are filtered out by setting the IoU threshold and non-maximum suppression.

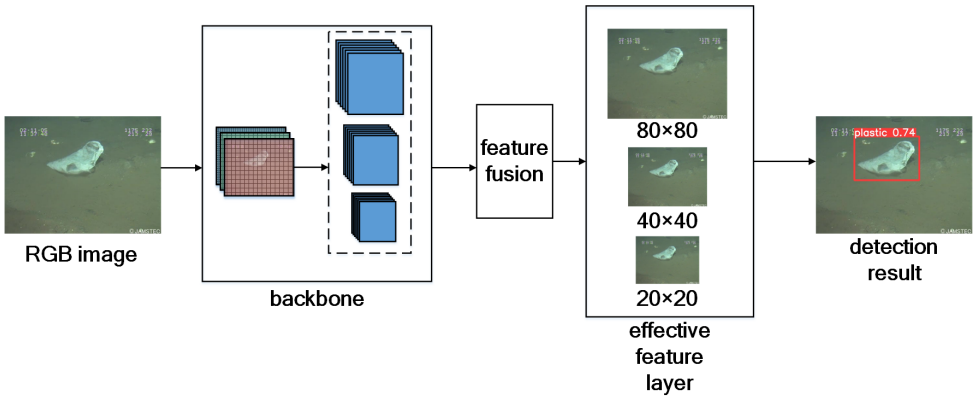


FIGURE 1. Detection process of YOLOv5 algorithm.

The effective feature layers with sizes of 80×80, 40×40, and 20×20 are taken into the feature fusion network, and the feature maps are fused and extracted using BiFPN to obtain the semantic information and positioning information of the feature map. In the end, the valid feature layers output from the feature fusion network are taken as input to the detection head for detection and regression, so that the prediction box tries to fit the position, width, and height of the real box. The algorithm detection process is shown in Figure 1.

3. Improved Algorithm:YOLOv5-LSMA.

3.1. Overall architecture of YOLOv5-LSMA. In order to improve the detection precision of deep-sea plastic waste, this paper proposes an improved deep-sea waste detection algorithm based on YOLOv5. The major improvements include: designing a loss reduction down-sampling (LRDS) module and replacing the ordinary convolution down-sampling module in the original network module with this module, optimizing the down-sampling process; proposing the Multi-scale Channel Attention Mechanism (MCAM) and introducing dilated convolution; finally using the S-BiFPN for secondary feature extraction of the shallow feature map to enhance the feature expression capability of the feature map. The overall structure is shown in Figure 2.

3.2. Loss reduction down-sampling module. The ordinary convolution down-sampling process is shown in Figure 3, in which the feature map is down-sampled by convolution, the width and height are changed to 1/2 of the original size, and the number of channels remains the same. Therefore, the process of ordinary convolution is a process accompanied by the loss of feature map information, and the loss of information in this process will lead to a reduction in the effective information available to the network detection part, which in turn will affect the detection precision of the module.

To reduce the above loss, we designed a loss-reducing down-sampling module with the focus structure and dilated convolution to improve the detection performance of the

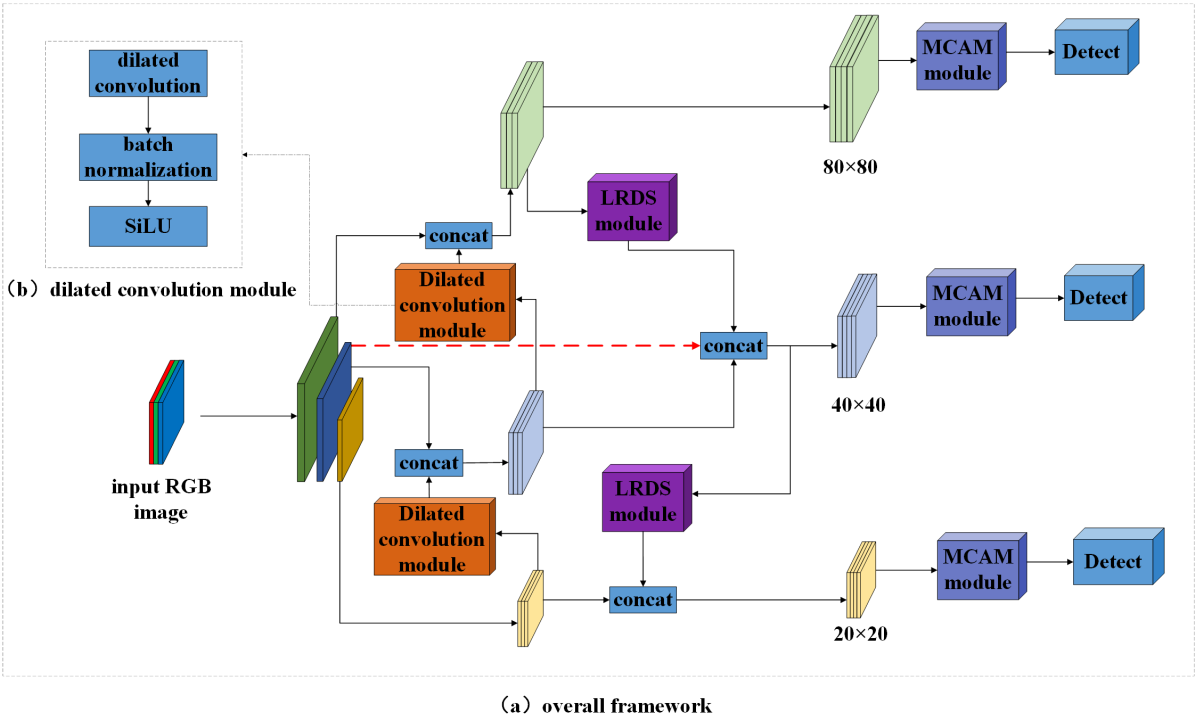


FIGURE 2. Whole structure of YOLOv5-LSMA model.

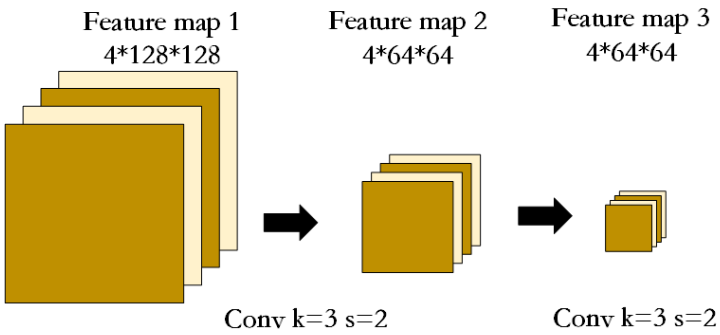


FIGURE 3. Process of convolution down-sampling.

module. The Focus structure is used to reduce the loss of information during the down-sampling, and the dilated convolution is used to increase the receptive field of the feature map and capture the multi-scale context information. The specific method is as follows.

3.2.1. *Focus structure.* The Focus structure is shown in Figure 4. In the input feature map, a value is taken every other pixel point, and an equal interval of 2 times down-sampling is performed. The sampled pixel points are spliced according to the original relative position, and the large input feature map is divided into 4 small feature maps whose height and width are half of the original feature map. Then the 4 small feature maps are spliced in the channel dimension, so that the height and width of the feature map obtained are half of the original feature map, but the number of channels is expanded to 4 times the original feature map. This process changes the width, height, and number of channels of the feature map, which not only completes the 2 times down-sampling but also reduces the loss of information in the down-sampling process.

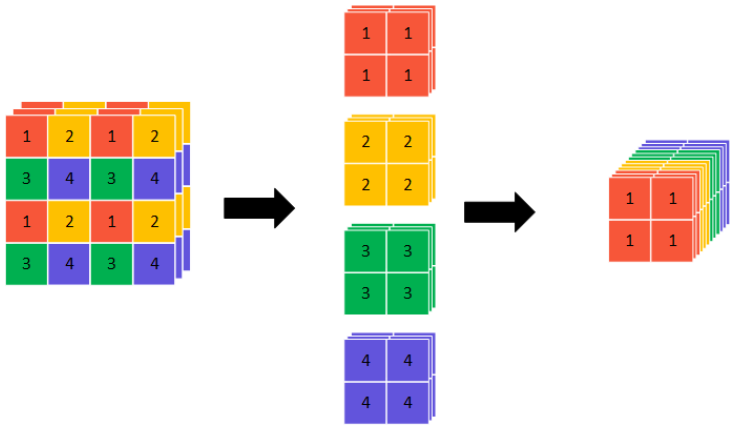


FIGURE 4. Structure of Focus.

3.2.2. *Dilated convolution.* As shown in Figure 5, the 3 figures represent the convolutional layers with different dilation rates and are independent of each other. The black dots in the image represent the convolution kernel of 3×3 size, and the blue shadow part represents the size of the receptive field after convolution. The dilated convolution is to fill a certain number of 0 in the middle of the standard convolution kernel. During the filling process, a hyperparameter, the dilation rate (DR), can be set to determine the number of filled 0. The dilation rate can be interpreted as the distance between two adjacent black dots, and when different DRs are set, different sizes of receptive fields can be obtained. The dilated convolution can expand the receptive field without increasing the number of parameters, but the value DR is too large to produce a grid effect, resulting in the loss of more feature information, or even the direct loss of the important pixels of the detection target, which affects the detection performance of the model. In order to avoid the above grid effect, which leads to the loss of important information about the target, three dilated convolutions are used in this paper, with DRs of 1, 2, and 3, respectively.

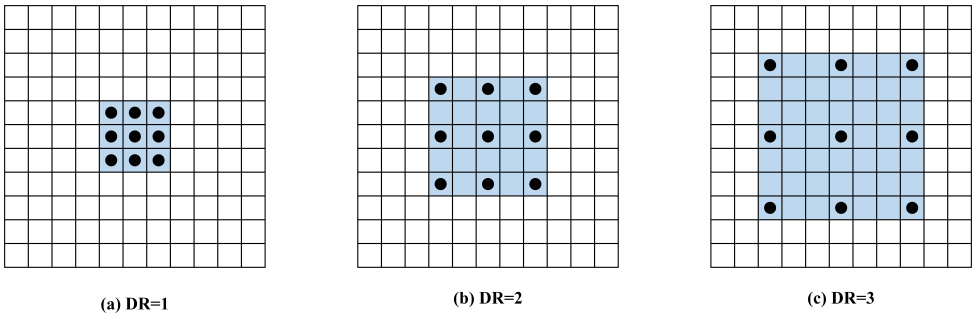


FIGURE 5. Dilated convolution.

3.2.3. *Loss reduction down-sampling module.* In this paper, the above two structures are used to design a loss reduction down-sampling module to reduce the feature loss in the down-sampling process. As shown in Figure 6, the feature map is first down-sampled by the focus structure and then passes through two branches, one branch passes through three dilated convolutions with the size of 3×3, the step of 1, and the dilation rate of 1, 2, 3, respectively. It can increase the receptive field of the feature map while avoiding the grid effect brought by the dilated convolution. The other branch is used as a residual connection to add to the output of the first branch, and the feature information before and after convolution is fused.

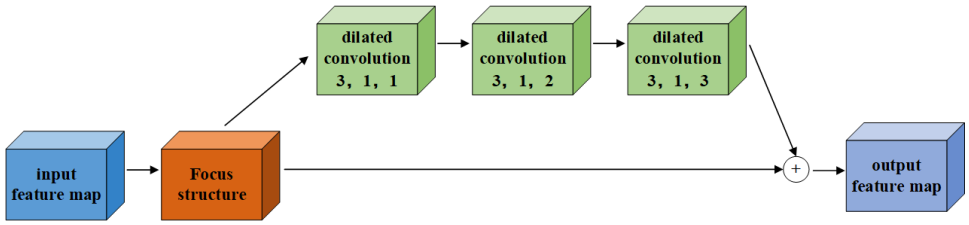


FIGURE 6. Structure of LRDS.

The mathematical expression of the loss reduction down-sampling module is shown in Formula (1):

$$OUT = DC3(DC2((DC1(LD(x))))) + LD(x) \tag{1}$$

where LD denotes the focus down-sampling, $DC1$, $DC2$, and $DC3$ denote the dilated convolution (dilation rate: 1, 2, 3), and + denotes the add operation.

3.3. **Multi-scale channel attention machanism.** A typical representative of the channel attention mechanism is SENet, whose core idea is to globally compress the feature map and score it in the channel dimension, and adaptively calibrate the weight of each channel to select important information and ignore unimportant information. Therefore, the attention of the network to channels unrelated to the detection target can be weakened and the attention of the network to channels related to the detection target can be enhanced by introducing a channel attention mechanism, thus improving the characterization ability and detection precision of the model. However, global average polling leads to a certain degree of spatial information loss.

In the proposed multi-scale channel attention mechanism, multi-scale feature polling can compensate for the loss of spatial information caused by global average polling to a certain extent, and then highlight the important channel weights by feature superposition. The structure is shown in Figure 7, where a feature map of $W \times H \times C$ (W, H, and C are the width, height, and number of channels of the feature map in order) is input, and its height and width are compressed by multi-scale adaptive average polling in the W and H dimensions to obtain $1 \times 1 \times C$, $2 \times 2 \times C$, and $4 \times 4 \times C$ features with the global receptive field in each channel. Then the features of $2 \times 2 \times C$ and $4 \times 4 \times C$ are subjected to general operations such as dimensionality reduction and then added with the features $1 \times 1 \times C$ to obtain the features compressed by multi-scale pooling. Finally, the compressed features are excited and the weights are assigned according to the importance of the channels.

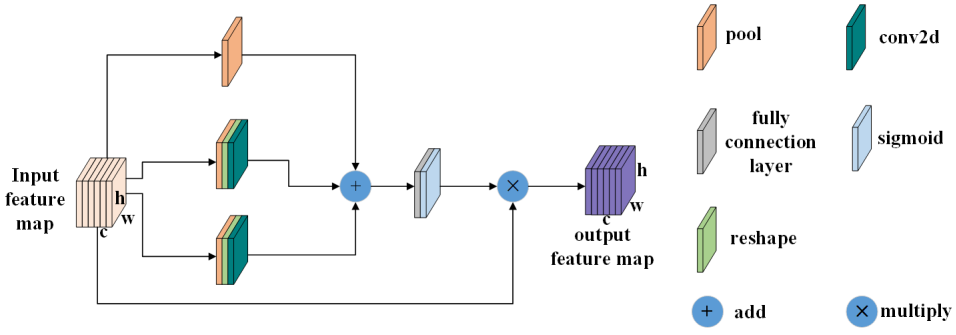


FIGURE 7. Multi-scale channel attention mechanism.

3.4. **S-BiFPN.** Figure 8(a) shows the structure of BiFPN. Unlike the PANet, the Google team believes that the amount of available effective information provided by different feature maps to the network is inconsistent, and proposes the BiFPN structure and balance the degree of their contribution to the network by assigning weights to the feature maps, so that the network obtain more effective feature information.

Inspired by the BiFPN, the weighted bi-directional feature fusion idea of the BiFPN network is combined with the FPN+PAN feature fusion network and applied to the YOLOv5 module. Since only the last three effective feature layers output by the YOLOv5 backbone network are utilized for feature fusion, the number of input nodes of BiFPN is changed to 3 to make it consistent in structure. Drawing on the core idea of BiFPN, a shortcut connection from the input nodes to the output nodes is introduced between the nodes of the effective feature layers with the same resolution, and the shallow feature map with more detailed information is fully utilized to better balance the contribution of each effective feature layer to the network, thereby improving the detection performance of the model.

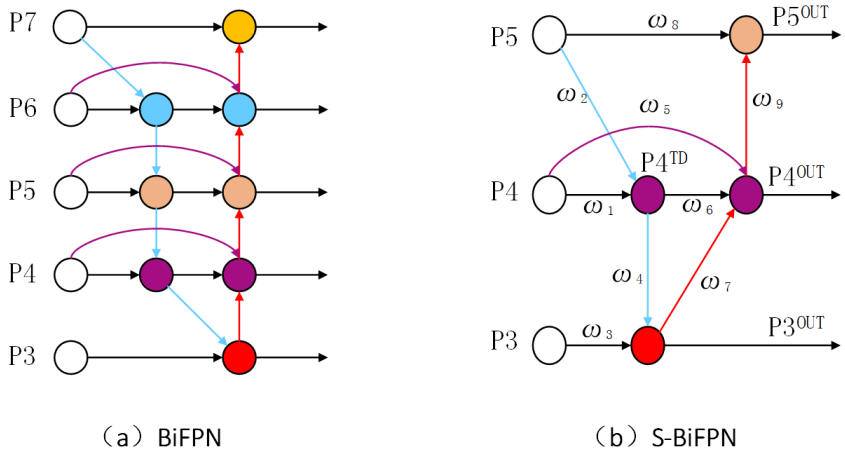


FIGURE 8. Structure of BiFPN and S-BiFPN.

As is shown in Figure 8(b), P3, P4 and P5 are the input nodes, $P4^{TD}$ is the middle layer node, $P3^{out}$, $P4^{OUT}$ and $P5^{OUT}$ are the output nodes. The mathematical expressions of the output nodes and the middle layer node are shown in the Formula (2)-(5):

$$P4^{TD} = Conv\left(\frac{\omega_1 \cdot P4 + \omega_2 \cdot Upsampling(P5)}{\omega_1 + \omega_2 + \epsilon}\right) \tag{2}$$

$$P3^{OUT} = Conv\left(\frac{\omega_3 \cdot P3 + \omega_4 \cdot Upsampling(P4^{TD})}{\omega_3 + \omega_4 + \epsilon}\right) \tag{3}$$

$$P4^{OUT} = Conv\left(\frac{\omega_5 \cdot p4 + \omega_6 \cdot P4^{TD} + \omega_7 \cdot DownSampling(P3^{OUT})}{\omega_5 + \omega_6 + \omega_7 + \epsilon}\right) \tag{4}$$

$$P5^{OUT} = Conv\left(\frac{\omega_8 \cdot P5 + \omega_9 \cdot Downsampling(P4^{OUT})}{\omega_8 + \omega_9 + \epsilon}\right) \tag{5}$$

where P_i is each input node; P_i^{OUT} is each output node; ω_i is each path weight coefficient, which is obtained by learning; *Upsampling* and *Downsampling* denote up-sampling and down-sampling operations, respectively, which serve to make the resolution of the feature map for feature fusion consistent; *Conv* means convolution operation; ϵ denotes learning rate with a value of 0.0001, which serve to avoid the value instability.

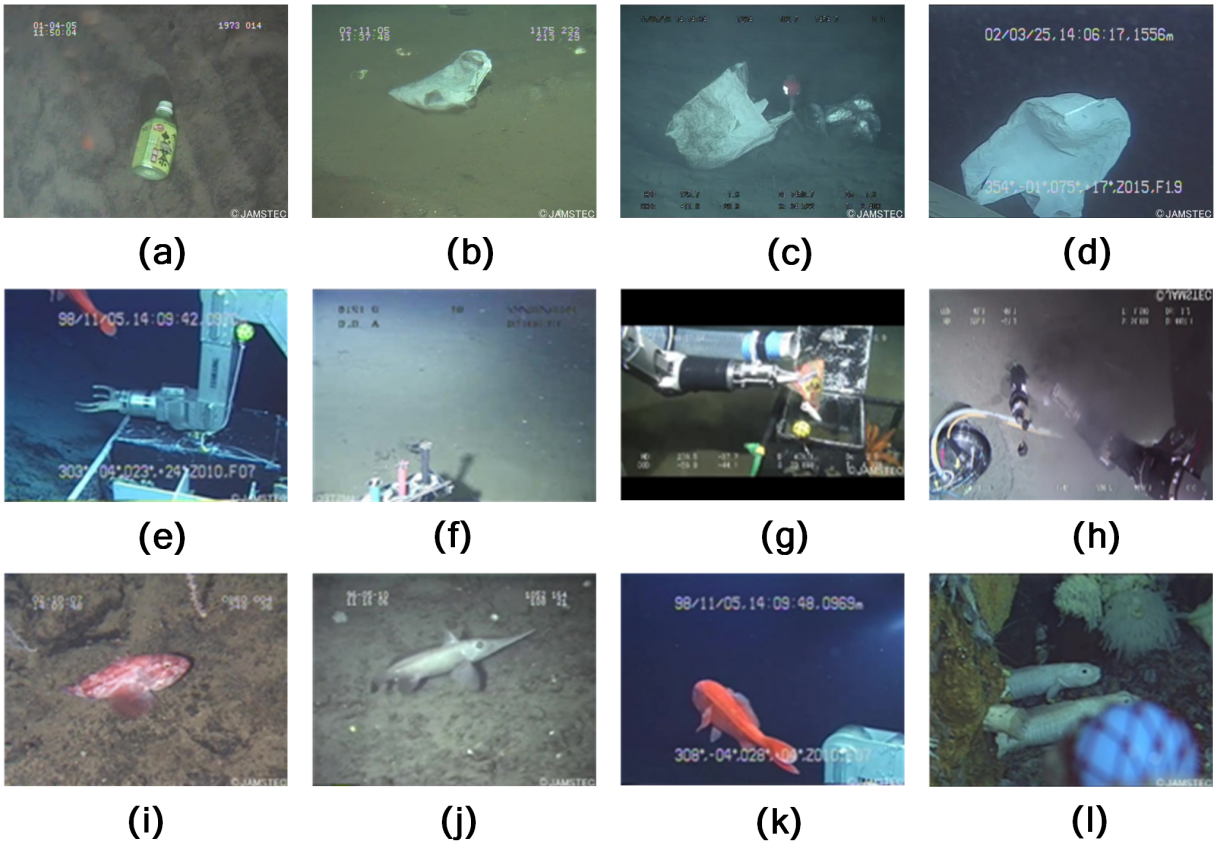


FIGURE 9. Partial dataset example.

4. **Experiment and Analysis.**

4.1. **Dataset and experiment environment.** In this paper, the experiments mainly focus on deep-sea plastic debris as the detection target, and the dataset uses the plastic deep-sea debris dataset published by Fulton et al. [27]. There are 7666 images in the dataset, including plastic, bio, and rov labels, which represent plastic, biological, and machine respectively, and the detection precision is represented by AP_0, AP_1, and AP_2 respectively. Before the experiment, the dataset was divided into training sets, test sets, and validation sets, and the proportions were 0.75, 0.1, and 0.15 respectively.

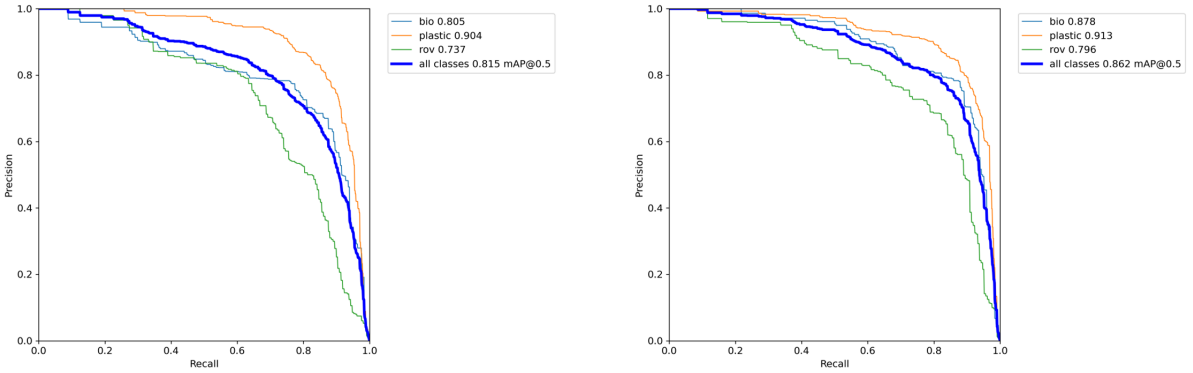
A partial dataset sample is shown in Figure 9, the first row is the plastic part, the second row is the rov part, and the third row is the bio part.

The experiments in this paper use Python programming language and Pytorch deep learning framework to build the model, and the optimizer is SGD, epoch is set to 100, batch is set to 64, the image size is set to 640×640, and the detection threshold is set to 0.5.

4.2. **Model evaluating indication.** In the object detection task, the correctness of the detection result is generally determined by the set value of IoU and the confidence threshold, and the precision of the detection results is judged by mAP.

IoU refers to the overlap ratio between the detection box and the real box, and its formula is:

$$IOU = \frac{Area(Detection) \cap Area(True)}{Area(Detection) \cup Area(True)} \tag{6}$$



(a) YOLOv5 **(b) YOLOv5-LMSA**

FIGURE 10. Precision recall curves.

where $Area(Detection)$ denotes the detection box area, $Area(True)$ denotes the real box area, $\cap$ denotes the intersection set, and $\cup$ denotes the union set.

The mean average precision (mAP) is generally used to measure the detection effect of the detection algorithm. The value of mAP is positively correlated with the precision, and its formula is:

$$AP = \int_0^1 P(R)d(R) \tag{7}$$

$$mAP = \frac{1}{classes} \sum_{i=1}^{classes} \int_0^1 P(R)d(R) \tag{8}$$

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

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

where P is the precision rate, R is the recall rate, TP is the true positive sample, FP is the false positive sample, FN is the false negative sample, and $classes$ is the number of categories.

The precision-recall curve can be obtained by calculating the above formula, and the area under the line indicates the AP value. As shown in Figure 10: where Recall denotes the recall rate and Precision denotes the accuracy rate. From Figure 10, it can be seen that the YOLOv5-LSMA model has a larger area under the line, the AP value has improved, and the precision rate and recall rate have also improved.

In order to show more intuitively the improvement effect of this algorithm of these two detection algorithms, as shown in Figure 11, it can be seen that the algorithm in this paper effectively improves the accuracy and reduces the leakage rate.

4.3. **Comparison and analysis of experiment results.** The detection algorithm of this paper is compared with other detection algorithms, and the comparison results are shown in Table 2. As shown in Table 2, the algorithm of this paper has the highest mAP of 86.2% compared with Faster R-CNN, SSD, YOLOv2, Tiny-YOLO, YOLOv3, YOLOv7, YOLOv7-tiny-silu, and YOLOv5. The detection precision of the detection algorithm of

FIGURE 11. Comparison detection results.

TABLE 2. Comparison of experimental results.

Methods	$AP_0/(\%)$	$AP_1/(\%)$	$AP_2/(\%)$	$mAP/(\%)$
SSD [27]	69.8	6.2	55.9	67.4
YOLOv2 [27]	82.3	9.5	52.1	47.9
Tiny-YOLO [27]	70.3	4.2	20.5	31.6
Faster R-CNN [27]	83.3	73.2	71.3	81.0
YOLOv3	89.1	80.6	69.8	79.9
YOLOv7	88.6	82.5	69.3	80.1
YOLOv7-tiny-silu	88.1	71.6	67.2	75.6
YOLOv5	90.4	80.5	73.7	81.5
YOLOv5-LSMA	**91.3**	**87.8**	**79.6**	**86.2**

this paper for each category is higher than other models, which reflects the superiority of the algorithm.

4.4. **Ablation experiment.** In order to verify the effectiveness of each module, ablation experiments were carried out in this paper. All ablation experiments use the same experimental environment, experimental parameters, and dataset, and all of them were improved with the YOLOv5 module as the baseline, and the modules were added to the baseline algorithm model in turn. The experimental results are shown in Table 3.

As can be seen from Table 3, adding S-BiFPN to the YOLOv5 model, its mAP is improved by 1.5% compared with the baseline, which better balances the contribution of each feature map to the network, and the shallow feature of the network is better utilized. Adding the loss-reducing down-sampling module to the basis of the YOLOv5 module, its mAP is improved by 3.3% compared with the baseline, which effectively reduces the loss of feature information in the down-sampling process. The introduction of the dilated convolution based on the YOLOv5 module improves the mAP by 0.4% compared with the baseline, which effectively improves the deep network receptive field size. Adding the MCAM module to YOLOv5 improves its mAP by 0.5% compared with the baseline, and the model is more reasonable in channel weight distribution. In

TABLE 3. Comparison results of ablation studies.

Baseline	Dilated convolution	MCAM	S-BiFPN	LRDS	$AP_0/(\%)$	$AP_1/(\%)$	$AP_2/(\%)$	$mAP/(\%)$
✓					90.4	80.5	73.7	81.5
✓	✓				90.3	81.8	73.6	81.9
✓		✓			91.0	82.2	72.8	82.0
✓			✓		91.1	81.3	76.4	83.0
✓				✓	91.3	84.3	78.7	84.8
✓	✓	✓	✓	✓	**91.3**	**87.8**	**79.6**	**86.2**

summary, the improvement of YOLOv5 in this paper is effective in the detection of deep-sea plastic garbage.

5. **Conclusions.** Aiming at the detection of plastic waste in deep-sea scenes, this paper proposes an improved YOLOv5-LSMA algorithm based on the YOLOv5 detection algorithm. Firstly, the S-BiFPN is used to balance the degree of contribution of each feature map to the network, so as to make full use of the detailed information of the shallow feature maps. Secondly, the loss reduction down-sampling module is proposed to reduce the feature loss of the network during the Neck part down-sampling process. Finally, the multi-scale channel attention mechanism module is proposed, and the dilated convolution is introduced to enable the model to better assign the weights of each feature channel and expand the receptive field of the deep network, which has better detection performance. The experimental results show that the YOLOv5-LSMA algorithm proposed in this paper has a better detection effect on deep-sea plastic waste detection. However, the current model is computationally intensive, has many parameters, and is very difficult to deploy to resource-constrained removable devices. Therefore, in the future, the algorithms can be considered to be lightweight so that they can be deployed to mobile devices.

Acknowledgment. The author would like to acknowledge the support of the Natural Science Foundation of Shandong Province (ZR2021MD057) and Construction of Quality Courses for Postgraduate Education in Shandong Province (SDYKC21063).

REFERENCES

[1] Z.-M. Guo, Z.-X. Chen, J. Dai, X. Gong, and G.-S. Zeng, "Design of a new deep-sea garbage salvage robot," *Journal of Changsha University*, vol. 36, no. 05, pp. 31–36, 2022.

[2] X. Liu, X. Sun, H.-N. Zhu, T. Gan, and Y. Zhao, "Pollution status and countermeasures of floating garbage in china 's offshore," *Environmental Health Engineering*, vol. 29, no. 05, pp. 23–29, 2021.

[3] D.-J. LI, "Marine plastic pollution and response," *World Environment*, vol. 2020, no. 01, pp. 71–73, 2020.

[4] Y.-T. Li and Y.-N. Qu, "Enlightenment of japan 's marine waste management laws and policies to china," *Journal of North University of China (Social Science Edition)*, vol. 38, no. 06, pp. 148–152, 2022.

[5] W. Sun, X.-C. Tang, Y.-D. Xu, H.-J. Zhang, Y.-J. Liu, and J.-X. Ma, "Distribution, composition and variation characteristics of marine debris in coastal waters of shandong province," *Science and Technology and Engineering*, vol. 16, no. 18, pp. 89–94, 2016.

[6] Y.-R. Ma, Y.-J. Peng, and T.-Y. Wu, "Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 2121–2133, 2022.
[7] B. Wang, Y. Wang, K. Qin, and Q. Xia, "Detecting transportation modes based on lightgbm classifier from gps trajectory data," in *2018 26th International Conference on Geoinformatics*, 2018, pp. 1–7.
[8] F. Zhang, T.-Y. Wu, Y. Wang, R. Xiong, G. Ding, P. Mei, and L. Liu, "Application of quantum genetic optimization of lvq neural network in smart city traffic network prediction," *IEEE Access*, vol. 8, pp. 104 555–104 564, 2020.
[9] F. Zhang, T. Wu, and G. Zheng, "Video salient region detection model based on wavelet transform and feature comparison," *EURASIP Journal on Image Video Processing*, vol. 2019, no. 1, pp. 1–10, 2019.
[10] E. K. Wang, X. Zhang, F. Wang, T.-Y. Wu, and C.-M. Chen, "Multilayer dense attention model for image caption," *IEEE Access*, vol. 7, pp. 66 358–66 368, 2019.
[11] W. Gan, L. Chen, S. Wan, J. Chen, and C.-M. Chen, "Anomaly rule detection in sequence data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12 095–12 108, 2023.
[12] J. Ning, M. Ma, L.-C. Chai, and F. Zhongying, "Overview of object detection algorithms for deep learning," *Information Recording Materials*, vol. 23, no. 10, pp. 1–4, 2022.
[13] R. B. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *CoRR*, vol. abs/1311.2524, 2013. [Online]. Available: http://arxiv.org/abs/1311.2524
[14] R. Girshick, "Fast r-cnn," in *2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.
[15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.
[16] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *Computer Vision – ECCV 2016*, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 21–37.
[17] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.
[18] J. Redmon and A. Farhadi, "Yolo9000: Better, faster, stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 6517–6525.
[19] ——, "Yolov3: An incremental improvement," *CoRR*, vol. abs/1804.02767, 2018. [Online]. Available: http://arxiv.org/abs/1804.02767
[20] A. Bochkovskiy, C. Wang, and H. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *CoRR*, vol. abs/2004.10934, 2020. [Online]. Available: https://arxiv.org/abs/2004.10934
[21] M. Sung, S.-C. Yu, and Y. Girdhar, "Vision based real-time fish detection using convolutional neural network," in *OCEANS 2017 - Aberdeen*, 2017, pp. 1–6.
[22] J. H. Christensen, L. V. Mogensen, R. Galeazzi, and J. C. Andersen, "Detection, localization and classification of fish and fish species in poor conditions using convolutional neural networks," in *2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV)*, 2018, pp. 1–6.
[23] R. Mandal, R. M. Connolly, T. A. Schlacher, and B. Stantic, "Assessing fish abundance from underwater video using deep neural networks," in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–6.
[24] X. Bing, "Research on deep learning methods for classification and detection of deep-sea garbage," 2022.
[25] T. Wei and G. Han, "Application of improved convolutional neural network algorithm in surface floating garbage detection," *China Science and Technology Paper*, vol. 14, no. 11, pp. 1210–1216, 2019.
[26] H.-C. Yuan and T.-Q. Zang, "Underwater garbage target detection based on attention mechanism ghost-yolov5," *Environmental Engineering*, vol. 4, no. 07, pp. 214–221, 2023. [Online]. Available: http://kns.cnki.net/kcms/detail/11.2097.X.20220913.1006.004.html
[27] M. Fulton, J. Hong, M. J. Islam, and J. Sattar, "Robotic detection of marine litter using deep visual detection models," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 5752–5758.