Water area refuse identification Algorithm Based on Improved YOLOv4

ABSTRACT

Quantifying plastic refuse in water area helps to understand how plastic refuse accumulates in water area and is essential for targeted cleanup efforts. Currently, the most common methods for quantifying plastic in water area are human visual counting and sampling using nets, but such methods are costly and labor-intensive. This study proposes a watershed refuse identification algorithm based on an improved YOLOv4. Lightweight improvements to YOLOv4. EfficientNetB1 is used to replace the backbone network of YOLOv4, and the Depthwise Convolution is used to replace the original convolution to reduce the number of model parameters and computation. The anchors are re-clustered using k-means algorithm to improve the accuracy. The experimental results show that the improved algorithm improves the detection speed by 11.2% and reduces the number of parameters by 76.54% compared with YOLOv4 at the expense of 0.69% recognition accuracy.

Key words: deep learning; refuse detection; YOLOv4; EfficientNet

1.INTRODUCTION

Quantitative testing of plastic waste floating in waters is essential to understand how the plastic waste in question collects and to identify areas where cleanup efforts are urgently needed. Most plastic waste takes hundreds of years to degrade, and because it degrades so slowly (400 to 1,000 years under natural conditions), plastic waste that is discarded into rivers, lakes and oceans spreads throughout the waters with the current. If neglected, the plastic waste on the shore will become a home for mosquitoes, which not only emits unpleasant odor but also seriously affects the human living environment [1].

A large amount of watershed litter originates from land-based sources, most commonly plastic bags and bottles. Studies have shown that removing plastic litter from waters will benefit watershed ecosystems exponentially. In order to better carry out the development of a watershed plastic removal plan, it is necessary to understand the distribution of plastic litter in watersheds. Traditional methods generally use human visual counts [2-4], sampling with the help of nets [5], etc., but these methods are either labor-intensive or have additional requirements for sampling equipment. Therefore, we need to use more accurate, reliable, and low-cost methods to accomplish this.

The development of deep learning has brought a new change to quantitative detection of watershed garbage, which can extract the features of targets in images through massive training, and use the extracted features to make judgments and achieve classification and recognition of targets. Target detection with the help of this method is the trend of computer vision development. Recently, researchers have proposed methods to quantify plastics in waters using computer vision and deep learning techniques.Van Lieshout C [6] et al. used a two-stage Faster R-CNN model to actively monitor and identify plastics floating on river surfaces and showed that, on average, the automated method detected 34.6% more plastics than manual visual counts. A research group at the University of Minnesota [7] developed a computer vision model specifically for marine plastic detection in deep-sea environments. Remote sensing monitoring of plastic litter in water provides a new, less labor-intensive method for quantifying and characterizing marine plastic pollution. In this paper, we propose an improved YOLOv4-based method for waterside litter detection by introducing EfficientNet network into the YOLOv4 model; using Depthwise Convolution to replace a large number of Standard convolutions in the Neck of YOLOv4 to significantly reduce the number of parameters of the model and improve the detection speed; and using k-means clustering algorithm to obtain a new pre-selected box to improve the recognition accuracy. Finally, a practical algorithm for watershed trash detection and recognition is constructed.

2. ALGORITHM OPTIMIZATION BASED ON YOLOV4

2.1 YOLOv4 Algorithm IntroductionYOLOv4

The YOLOv4 model is a new generation of the YOLO family of algorithms proposed by Bochkovskiy in 2020 based on YOLOv3 [8]. yOLOv4 has a significant improvement in effectiveness compared to the original YOLOv3.

In terms of network architecture.

a. YOLOv4 improves Backbone and introduces the CSP module in Darknet53, the backbone network of YOLOv3. The use of CSP structure in the target detection network can reduce the computation and memory consumption of the network while keeping the capacity of the network unchanged or even slightly increased.

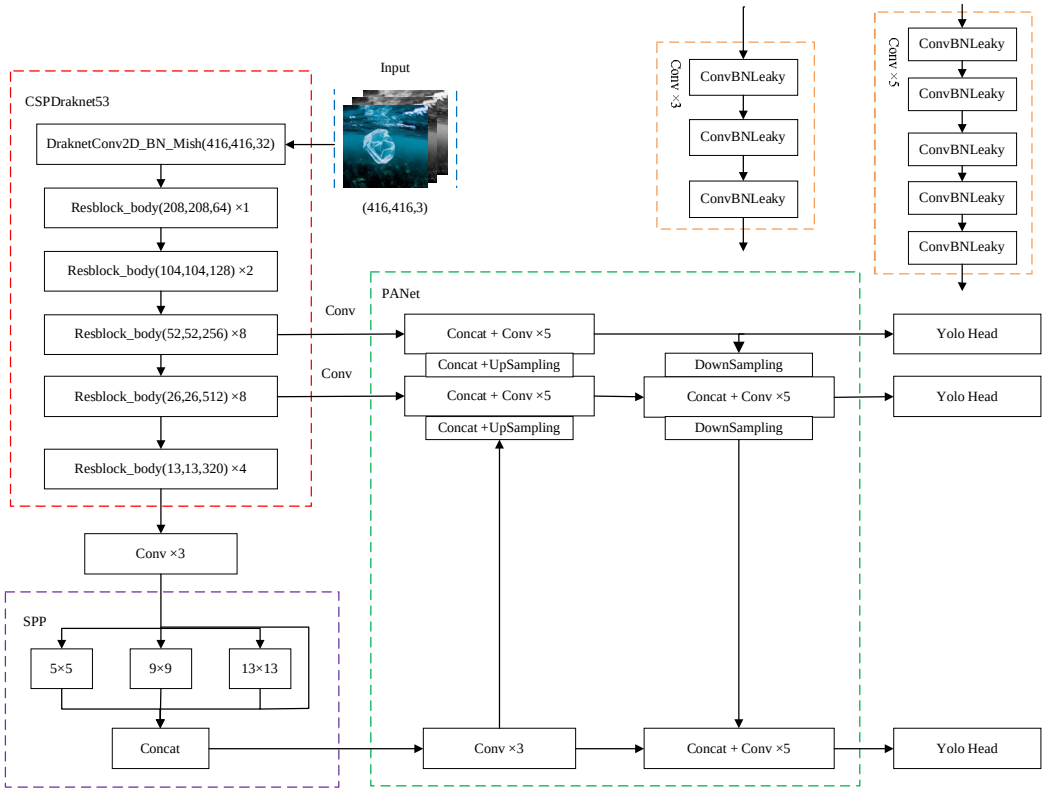
b. Use SPP module and PAN module. SPP module is to let the feature layer pass through a pooling kernel size of 5×5, 9×9, 13×13 maximum pooling layer respectively, and then Concat splicing in the channel direction for further fusion, which can increase the perceptual field and can separate the most significant contextual features, solving the problem of target multi-scale to some extent. The PAN structure is added to the top-to-bottom information fusion of FPN, which aggregates the parameters of different detection layers of different backbone layers and further improves the feature extraction capability of the network.

In terms of optimization strategies.

a. CIOU. CIOU loss is used and DIOU_NMS is used instead of NMS to make the speed and accuracy of the prediction frame regression higher.

b. Mosaic data enhancement. Four images are read at a time to increase the diversity of the learning samples by flipping, scaling, and changing the color gamut of the four images, and arranging them in four directions to form a single image.

The Mish activation function is used to replace the LeakyReLU function in the backbone. c. The Mish activation function has better performance compared to other commonly used functions such as ReLU and Swish.



Note: Concat stands for Tensor Splicing, Conv stands for Convolution, ConvBNLeaky stands for Conv plus Batch Normalization plus Leaky ReLU activation function module.

Fig.1 YOLOv4 Network Model

The YOLOv4 model can be divided into three parts: Backbone, Neck and Head. the CSPDarketnet53 structure is the Backbone of the model; the Neck structure is mainly composed of SPP and PAN; and Head is the prediction part of the model. the network structure of YOLOv4 is shown in Figure 1. The backbone feature extraction network first convolves the input image, and then inputs the obtained feature layers into the residual network for training. After the training of the backbone feature extraction network, three effective feature layers with sizes of 52×52×256 (feature layer 1), 26×26×512 (feature layer 2), and 13×13×1024 (feature layer 3) are output. The output effective feature layer 3 is subjected to 3 convolution operations and the output results after the operations are input to SPP. The pooled output is stacked and subjected to another 3-time convolution. The output after 3 convolutions is upsampled and stacked with the effective feature layer 1 and effective feature layer 2 of the backbone feature extraction network output, which increases the feature characterization capability. After that, downsampling is performed in the second stage. The purpose of the process of continuous upsampling and downsampling is to obtain better features [9]. Finally Head uses the extracted features to make predictions on the image.

2.2 YOLOv4 algorithm improvement

The YOLOv4 algorithm has excellent performance after adopting numerous optimization strategies. However, when it needs to be deployed in devices with poor performance, the number of parameters of the model has to be controlled by the arithmetic power and memory resources of the device, and the algorithm model of YOLOv4 still seems too large to perform well on devices with poor performance.

Therefore, the original network structure of YOLOv4 needs to be improved by compressing the number and size of parameters of the model to improve the prediction speed while ensuring a certain accuracy.

2.2.1 Replacement of backbone feature extraction network

EfficientNet is a new convolutional neural network model designed by Google in 2019 using NAS [10]. A new idea of model scaling method is investigated for EfficientNetB0, the base network model in EfficientNet, and a series of EfficientNetB1-B7 models are obtained by a formula containing composite coefficients to uniformly weigh the network depth, width, and input image resolution after increasing and decreasing the network structure[11].

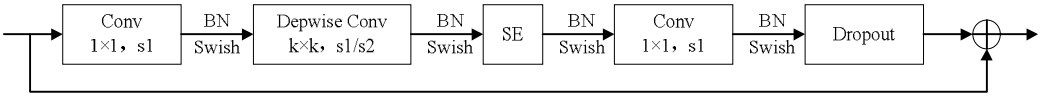
EfficientNet balances the improvement of image size, network depth and network width, and increases from all three aspects to improve the recognition accuracy of the network.

Table1 **EfficientNetB0** network

Stage	Operator	Resolution	Channels	Layers
1	Conv3×3	224×224	32	1
2	MBConv1, k3×3	112×112	16	1
3	MBConv6, k3×3	112×112	24	2
4	MBConv6, k5×5	56×56	40	2
5	MBConv6, k3×3	28×28	80	3
6	MBConv6, k5×5	14×14	112	3
7	MBConv6, k5×5	14×14	192	4
8	MBConv6, k3×3	7×7	320	1
9	Conv1×1&Pooling&FC	7×7	1280	1

As shown in Table 1, the core structure of the EfficientNet model is the MBConv module (shown in Figure 2). the MBConv module mainly consists of a 1×1 normal convolution (up-dimensioning role); a k×k Depthwise Conv, the exact value of k is given in the network framework; an SE module; a 1×1 normal convolution (down-dimensioning role); and a Droupout layer. The SE module is a lightweight attention mechanism module that can be easily added to the network model with a small increase in model complexity and computational expense. In the detection task, as the data stream deepens in the network, the target features

become weaker and weaker, which can easily cause small and weak targets to be missed. The SE module is added at appropriate locations in the network to enhance the extraction of channel association information and improve the recognition rate [12-13].



Note: Depwise Conv stands for Depthwise Separable Convolution.

Fig.2 MBConv module

2.2.2 SPP improvements

The SPP pyramid pooling [14] is used in the YOLOv4 algorithm network with three maximum pooling layers with pooling kernel sizes of 5 × 5, 9 × 9, and 13 × 13. As shown in Figure 3, the SPPF structure, on the other hand, uses three maximum pooling layers of size 5 × 5, and the output after each pooling becomes the input for the next pooling. After testing, the computation results of SPP and SPPF are identical, and the computation speed of SPPF is faster than that of SPP.

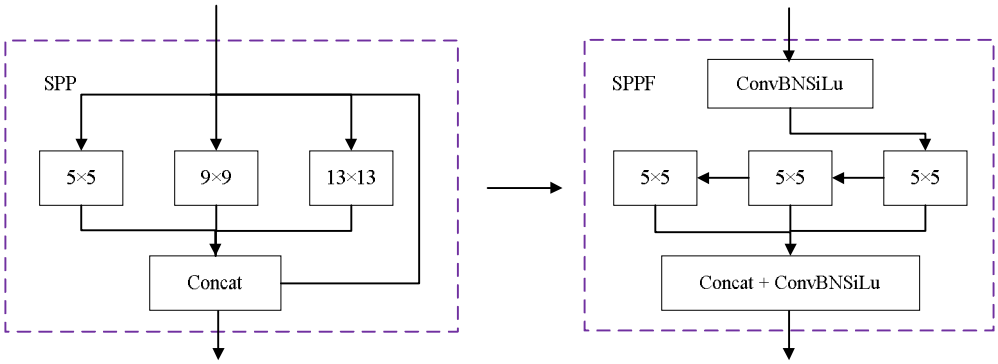


Fig.3 SPP AND SPPF module

2.2.3 Neck part improvement

The YOLOv4 algorithm network, which makes extensive use of Standard convolution in Neck, i.e., the ConvBNLeaky structure in Fig. Using Depthwise Convolution [15-16] instead of the above structure can effectively reduce the number of parameters in the network, and the number of parameters in the neural network directly affects the computational effort during the network model computation. The deep separable convolution is in fact the deep convolution first and then the point-by-point convolution. Figure 5-6 shows the comparison between normal convolution and deep separable convolution.

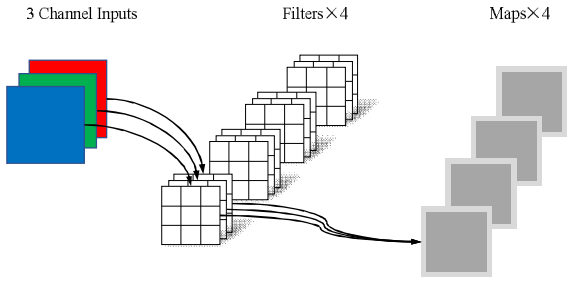
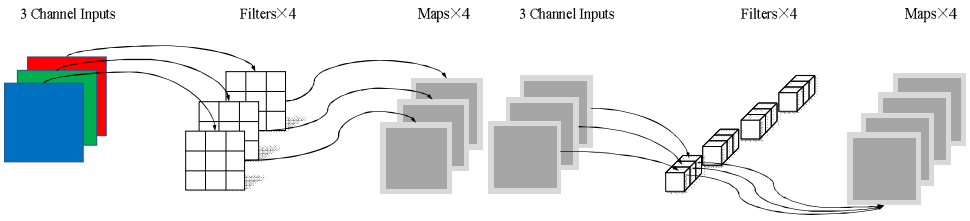


Fig.4 Standard convolution



a) b)

Note: a) for Depthwise Convolution, b) for Pointwise Convolution

Fig.5 Depthwise Convolution

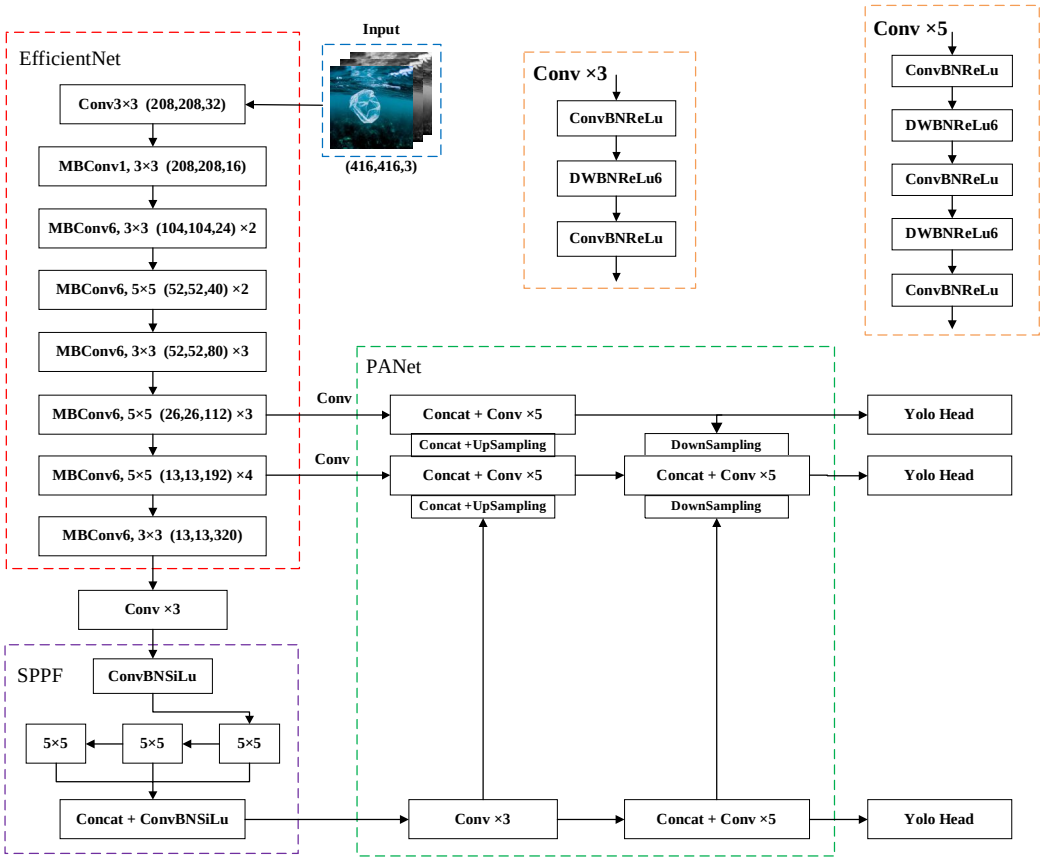
2.2.3 Re-clustering anchors using k-means

YOLOv4 is trained on large datasets such as ImageNet, and three sets of nine prior frames (anchors) are obtained by clustering, which are the default prior frames for small, medium and large objects. During training, the a priori frames are first roughly selected at the possible locations of the target, and then adjusted on top of them. If the selected prior frame fits the distribution of the dataset, then the network can learn more easily and get good detectors. In this paper, the waterside trash detection algorithm is mostly applied to the waterside camera platform, and the trash size is relatively uniform under the guarantee of a certain distance, which is completely different from the situation in large datasets with multiple types and huge size differences. If the default a priori frame is used, trash of similar size will be assigned to different layers for prediction, and there is a mismatch between the prediction scale and the perceptual field.

To solve this problem, the k-means clustering algorithm[17] is used to obtain the anchors applicable to watershed garbage detection, and the detection accuracy is improved by adjusting the anchors to make full use of the three output layers of large, small and medium size.

2.2.4 Improved YOLOv4 network model

The structure of the improved YOLOv4 network model is shown in Figure 6. The input layer image size is adjusted to 416×416×3. EfficientNetB1 is used to replace the original backbone feature extraction network of YOLOv4, and the convolution is adjusted to ensure that the three effective feature layers extracted can be transmitted to the SPP module and PANet module. In Neck, the SPPF structure is used to replace the SPP structure to improve the computational speed without changing the effect, and the Depthwise Convolution is used to replace the normal convolution to reduce the number of parameters in the network.



Note: SPPF stands for improved Spatial Pyramidal Pooling, ConvBNReLu stands for Conv plus Batch Normalization plus ReLU activation function module, DWBNReLu6 stands for Deep Separable Convolution plus Batch Normalization plus ReLu6 activation function module, and ConvBNSiLu stands for Conv plus Batch Normalization plus SiLu activation function module.

Fig.6 Improved YOLOv4 network model

3. RESULTS AND DISCUSSION

3.1 Test environment and watershed litter dataset

The experiment uses Pytorch-GPU 1.7.1 deep learning framework, running deep learning workstation with Intel Xeon E5-2699 v4 processor, 2.20 GHz, NVIDIA Quadro M6000 graphics card, Ubuntu 20.04.4 operating system, NVIDIA driver version 470.103.01. CUDA version 11.3, CUDNN version 8.2.1.

The DeepTrash shared by Gautam et al [18] was added to their own collected images of watershed litter as the experimental dataset. The experimental dataset has two categories, for plastics and bottles. In order to increase the number of datasets, the ratio and brightness of the images in the original dataset were randomly adjusted, and the datasets were scaled and flipped. The expanded dataset has a total of 3354 images with a size of 416×416. There are 3065 plastic annotations and 3253 bottle annotations in all images. Before training, 3017 images were randomly selected in the test dataset according to the ratio of 9:1 as the training validation set and 336 images as the test set. In the training validation set, another 2715 images were randomly selected as the training set and 302 images as the validation set according to the ratio of 9:1.

3.2 Model training methods and evaluation metrics for improved algorithms

3.2.1 Model training methods

In this experiment, the pre-trained model trained on the VOC dataset with the help of YOLOv4 is trained by migration learning. The

number of iterations (Epoch) is set to 300, and the backbone network is frozen for the first 60 training sessions with a batch_size of 64, and only the pretrained weights are loaded. The backbone network is unfrozen for the next 240 training sessions, with a batch_size of 32, and the whole network is trained together. The network model training parameters are set as follows: the momentum factor is 0.937 and the decay coefficient is 0.0005. The momentum factor is the tendency of the value of the loss function to decrease during training. Using SGD stochastic gradient descent, the maximum learning rate of the training process is set to 0.04, and the minimum learning rate is limited to 0.0016, using the Mosaic data enhancement and cosine annealing algorithms.

3.2.2 Evaluation Indicators

In this paper, mAP (mean value of AP value under all categories), detection speed, model size and number of parameters are used as evaluation metrics.

a. The better the performance of the general target recognition algorithm, the higher its accuracy. In multi-category object detection, each category can be plotted with a corresponding P-R curve, AP is the area under that curve, and mAP is the average value of AP for multiple categories. Where AP and mAP are calculated as follows.

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

$$mAP = \frac{1}{C}\sum_{c \in C} AP(c) \tag{2}$$

P is the accuracy rate; R is the recall rate; and C is the number of species.

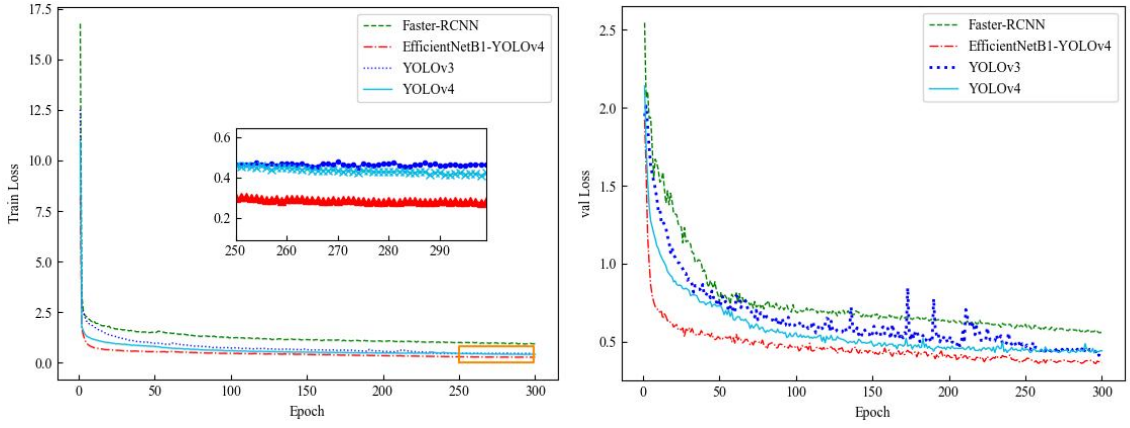
b. Detection speed. In this paper, the number of spam images detected per second (fps·s^(-1)) is used as the evaluation index of the detection speed of the improved algorithm.

c. Complexity of the algorithm is measured using the number of model parameters

.3.3 Experiments and analysis of results

In order to verify the effectiveness of the algorithms in this paper, the training and performance evaluation experiments are conducted with YOLOv4, YOLOv3, and Faster RCNN under the same hardware and software environment and data set, with the input image size fixed at 416×416, and the four algorithms are analyzed.3.3.1 改进YOLOv4 算法的性能分析

One of the indicators to judge the effectiveness of model training is the loss value. When the decreasing trend of the loss value tends to be stable, the model can be considered to have converged. In theory, for the same model, the smaller the loss value is, the better the model training effect is. In order to verify the convergence speed and stability of the algorithm in this paper, the improved algorithm is trained with YOLOv4, YOLOv3 and Faster-RCNN, and the loss value curve is plotted according to the saved log information. As shown in Figure 7.



a) b)

Note: a) for Training set curves, b) for Validation set curves

Fig.7 Loss trend of model training

The loss function curves of the four algorithms show that the loss value gradually decreases as the number of

iterations increases. all four algorithms eventually reach the convergence state, but the algorithm in this paper has a smoother loss function curve, faster convergence, and smaller final loss value compared with the other three algorithms.

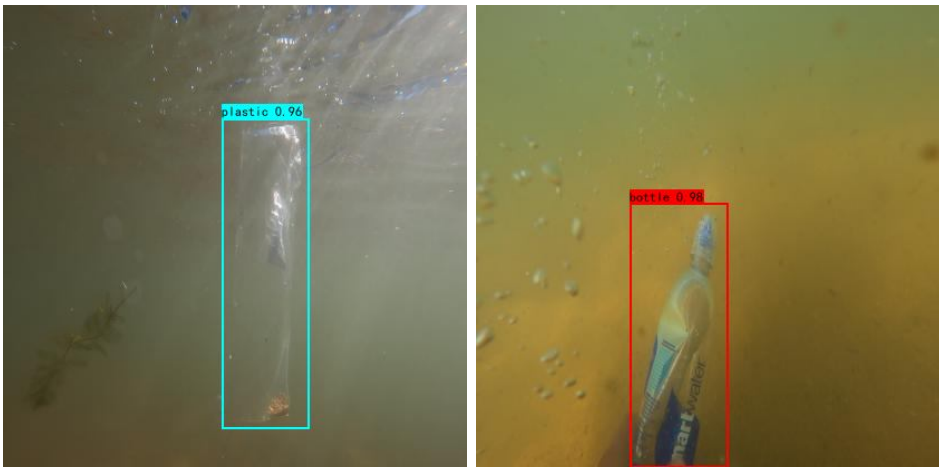
3.3.2 Analysis of the results of the improved algorithm

In order to verify the performance of this paper's algorithm for watershed spam identification, the four networks are compared in terms of mAP, detection speed, number and size of model parameters, as shown in Table 2.

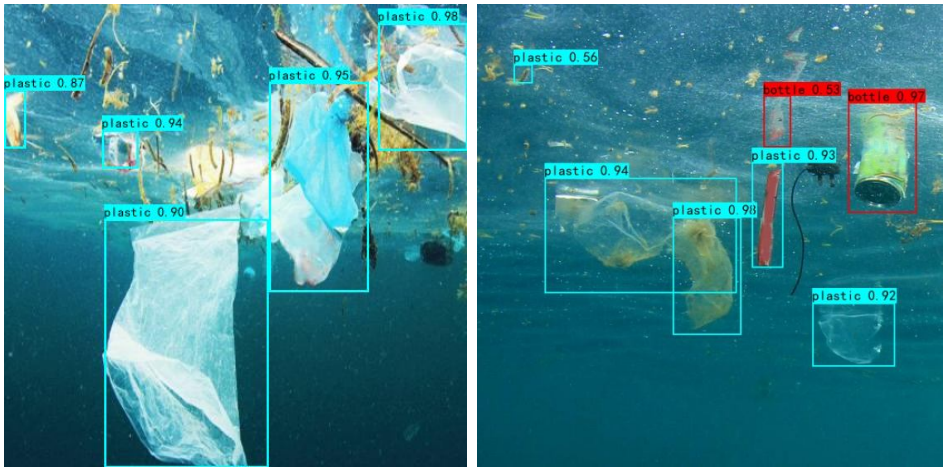
Table 2 Watershed litter detection results of different network models

Model	Backbone Network	mAP/%	Speed ($fps \cdot s^{-1}$)	Parameters
EfficientNetB1-YOLOv4	EfficientNetB1	86.35	30.47	15.1M
YOLOv4	CSPDarketnet53	87.04	27.40	64.36M
Faster-RCNN	ResNet50	65.67	8.92	28.47M
Yolov3	Darketnet53	85.03	29.07	61.95M

As can be seen from Table 2, the mAP of the algorithm in this paper is 86.35%, which is 20.98% and 1.32% higher than Faster-RCNN and Yolov3, respectively, and slightly lower than YOLOv4. In terms of the recognition speed of the algorithm, this paper is higher than all the other three, reaching 30.47 fps/s, compared with YOLOv4 and Yolov3, which are improved by 11.2% and 4.8% compared to YOLOv4 and Yolov3, respectively. In terms of the number of model parameters of the algorithm, the number of model parameters of the algorithm in this paper is 15.1M, which is only 23.46%, 53.04% and 24.36% of YOLOv4, Faster-RCNN and Yolov3. Overall, the algorithm in this paper greatly reduces the model size and the number of parameters to improve the recognition speed of waterside garbage while ensuring the recognition accuracy, thus realizing the fast and effective detection of waterside garbage.



a) b)



c) d)

Fig.8 Model testing results

Some of the recognition results of the algorithm in this paper are shown in Figure 8. It can be seen that all the garbage floating in the water can be identified correctly, even in the turbid water environment as shown in Figs. a) and b). As shown in c) and d), the presence of multiple garbage in the environment can also be detected more completely. Overall, the algorithm in this paper is effective in recognizing garbage in waters.

4 CONCLUSION

1) In this paper, we propose an improved target detection model based on EfficientNet for YOLOv4, replacing the backbone CSPDarketnet53 network of YOLOv4 with the EfficientNetB1 network structure. The changed network is lighter and ensures that the accuracy of waterside trash recognition is not reduced while reducing the number of parameters.

2) According to the actual environment of waters, based on the open source dataset DeepTrash, images collected from the network are added to produce image datasets of plastic and bottles, the two most common types of waters litter, and trained with EfficientNetB1-YOLOv4, YOLOv4, Faster-RCNN and Yolov3 models, respectively. The results show that the comprehensive performance of EfficientNetB1-YOLOv4 network model is better.

(3) The performance of the EfficientNetB1-YOLOv4 network model is evaluated, and the algorithm can complete the detection task even in a turbid environment with multiple targets to be detected.

(4) In the future research work, we will further explore the lightweight improvement method of the target detection algorithm to further improve the detection accuracy and detection speed while reducing the computational volume, number of parameters and fan line size as much as possible.

REFERENCES

[1] China Ecological Environment Status Bulletin 2020 (Excerpt) [J]. Environmental Protection, 2021,49 (11):47-68. Chinese

[2] Van Emmerik T, Kieu-Le T C, Loozen M, et al. A methodology to characterize riverine macroplastic emission into the ocean[J]. Frontiers in Marine Science, 2018, 5: 372.

[3] González-Fernández D, Hanke G. Toward a harmonized approach for monitoring of riverine floating macro litter inputs to the marine environment[J]. Frontiers in Marine Science, 2017, 4: 86.

[4] Van Calcar C J, van Emmerik T H M. Abundance of plastic debris across European and Asian rivers[J]. Environmental Research Letters, 2019, 14(12): 124051.

[5] Rech S, Macaya-Caquilpán V, Pantoja J F, et al. Rivers as a source of marine litter–a study from the SE Pacific[J]. Marine pollution bulletin, 2014, 82(1-2): 66-75.

[6] van Lieshout C, van Oeveren K, van Emmerik T, et al. Automated river plastic monitoring using deep learning[J]. 2020.

[7] Fulton M, Hong J, Islam M J, et al. Robotic detection of marine litter using deep visual detection models[C]//2019 International Conference on Robotics and Automation (ICRA). Montreal:IEEE, 2019: 5752-5758.

[8] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. arXiv preprint arXiv:2004.10934, 2020.

[9] Zhang Zhao Guo,Zhang Zhendong,Li Jianian,Wang Haiyi,Li Yanbin, Li Donghao.Potato detection in complex environment based on improved YoloV4 model [J]. Transactions of the Chinese Society of Agricultural Engineering, 2021,37 (22):170-178. Chinese

[10] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. Long Beach:PMLR, 2019: 6105-6114.

[11] ZHAO Pengfei and HUANG Lijia. Target recognition method for multi-aspect synthetic aperture radar images based on EfficientNet and BiGRU[J]. Journal of Radars, 2021, 10(6): 895–904. doi:10.12000/JR20133. Chinese

[12] LIU Xue,LI Fanming,LIU Shijian. An Infrared Image Pedestrian Detection Algorithm Based on Improved SSD Algorithm [J]. Electronics Optics & Control, 2020,27 (01):42-46+59. Chinese

[13] Hu J, Shen L, Sun G. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition.Salt Lake City: IEEE,2018: 7132-7141.

[14] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.

[15] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. Salt Lake City: IEEE,2018: 4510-4520.

[16] WANG Qixuan,GUAN Shengqi,HU Luping. Industrial Bar Recognition Algorithm Based on Improved YOLOv4 [J]. Machinery & Electronics, 2022,40 (01):25-29+35. Chinese

[17] SUN Lin,LIU Meng-han,XU Jiu-cheng. K-means Clustering Algorithm Using Optimal Initial Clustering Center and Contour Coefficient [J]. Fuzzy Systems and Mathematics, 2022,36(01):47-65. Chinese

[18] Tata G, Royer S J, Poirion O, et al. A Robotic Approach towards Quantifying Epipelagic Bound Plastic Using Deep Visual Models[J]. arXiv preprint arXiv:2105.01882, 202