Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

A modified YOLOv4 detection method for a vision-based underwater garbage cleaning robot*&

Manjun TIAN[1,2], Xiali LI[2], Shihan KONG[3], Licheng WU[2], Junzhi YU‡[3,4]

[1]*First Research Institute of the Ministry of Public Security of PRC, Beijing 100048, China*

[2]*School of Information Engineering, Minzu University of China, Beijing 100081, China*

[3]*Department of Advanced Manufacturing and Robotics, College of Engineering, Peking University, Beijing 100871, China*

[4]*State Key Laboratory of Management and Control for Complex Systems, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*

E-mail: tianmanjun2018@163.com; xiaer_li@163.com; kongshihan@pku.edu.cn; wulicheng@tsinghua.edu.cn; junzhi.yu@ia.ac.cn

Received Oct. 1, 2021; Revision accepted Mar. 7, 2022; Crosschecked July 15, 2022

Abstract: To tackle the problem of aquatic environment pollution, a vision-based autonomous underwater garbage cleaning robot has been developed in our laboratory. We propose a garbage detection method based on a modified YOLOv4, allowing high-speed and high-precision object detection. Specifically, the YOLOv4 algorithm is chosen as a basic neural network framework to perform object detection. With the purpose of further improvement on the detection accuracy, YOLOv4 is transformed into a four-scale detection method. To improve the detection speed, model pruning is applied to the new model. By virtue of the improved detection methods, the robot can collect garbage autonomously. The detection speed is up to 66.67 frames/s with a mean average precision (mAP) of 95.099%, and experimental results demonstrate that both the detection speed and the accuracy of the improved YOLOv4 are excellent.

Key words: Object detection; Aquatic environment; Garbage cleaning robot; Modified YOLOv4
https://doi.org/10.1631/FITEE.2100473 **CLC number:** TP242

1 Introduction

Polluted water causes 1.7 million deaths from curable diseases every year. It is also conservatively estimated that a large amount of wetland will be lost every year, and that wetland loss is equivalent to the value created by 800 000 hectares of wetland. The poor state of the marine environment is directly or indirectly affecting about 275 million people's access to cheap protein (i.e., fish), and will also cause 3.1 billion people around the world to be unable to obtain 20% of their supply of animal protein. In addition, it will cause serious damage to the global tourism and fishing industries (Ekins and Gupta, 2019). In particular, the pollution of plastic waste is quite severe. Due to the extremely slow degradation, plastics have become ubiquitous and have been associated with marine health impacts such as entanglement, ingestion, potential dispersal of invasive species and toxicity, and contamination through trophic levels (Ostle et al., 2019). The ocean is suffering from

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 61725305, U1909206, T2121002, and 62073196), the Postdoctoral Innovative Talent Support Program (No. BX2021010), and the S&T Program of Hebei Province, China (No. F2020203037)

& A preliminary version was presented at the 40th Chinese Control Conference, Shanghai, China, July 26–28, 2021

ORCID: Junzhi YU, https://orcid.org/0000-0002-6347-572X

increasing pollution, especially from plastics; eight million tons of plastic products enter the ocean every year, equivalent to a truckload of plastic garbage being dumped into the ocean every minute (Jambeck et al., 2015). According to a recent research report released by the marine conservation organization Oceans Asia, due to the sudden coronavirus outbreak, at least 1.56 billion masks were discarded into the ocean in 2020. Most of the masks are disposable and take 400 to 500 years to degrade in the ocean. They may carry germs, and are "hotbeds" for germs to multiply, posing a potential threat to the survival of marine life. Therefore, it is urgent to protect water resources.

In recent years, with the gradual optimization of control technology and the continuous improvement in software and hardware performance, intelligent robot research has become a research hotspot, and the development of intelligent robots has advanced by leaps and bounds (Albitar et al., 2016; Laschi et al., 2016; Mahler et al., 2016, 2017, 2018, 2019; Xu et al., 2017; Bai et al., 2018; Prabakaran et al., 2018; Kim et al., 2019; Li CY et al., 2019). In this context, an intelligent robotic platform has been developed in our laboratory to clean up underwater garbage.

Robotic garbage cleaning can improve work efficiency, reduce labor, maintain safety, and enhance reliability. The robot's main task is to detect, capture, and collect objects. The detection algorithm is indispensable when the robot works to discover, approach, and collect targets. A superior detection method can provide the robot with real-time and reliable target information and assist the robot in completing various tasks.

Object detection includes two primary tasks, i.e., the classification and positioning of targets, which are two most basic and challenging research topics in the field of computer vision. Recently, with the advancement and development of computer science and technology, object detection has been widely implemented in areas such as intelligent surveillance (Fu et al., 2019), intelligent transportation (Mhalla et al., 2019), intelligent agriculture (Horng et al., 2020), military (Astapov et al., 2014), and medical fields (Tschandl, 2020).

With respect to object detection, accuracy and speed are the two most important indicators (Pu et al., 2021). Specifically, when the target is moving at a high speed or the application scene is complex and changeable, the speed and accuracy of the object detection method will face huge challenges. Among traditional object detection algorithms, histogram of oriented gradient (HOG) (Dalal and Triggs, 2005) uses a histogram to count the edges of objects, and performs satisfactorily when expressing features. Local binary pattern (LBP) (Ojala et al., 2002) is good at expressing object texture information. Harr (Whitehill and Omlin, 2006) can rapidly extract features, and is widely applied due to its strong ability to express object edge information. Scale invariant feature transform (SIFT) (Lowe, 2004) is a local image feature that is unaffected by rotation, scaling, and brightness changes, and is robust in viewing angle changes, affine transformations, and noise. In 2001, Viola's cascade + Harr scheme achieved outstanding results in face detection (Viola and Jones, 2001). The deformable part module (DPM) is also a popular detector. However, the performance of DPM is ordinary, and it cannot adapt to images with sharp rotations. Thus, its stability and robustness are inadequate. In addition, the workload is relatively heavy (Felzenszwalb et al., 2008). Therefore, traditional methods have insufficient capability of feature extraction, which results in limited accuracy. Manual feature design is very troublesome and laborious.

The emergence of the convolutional neural network (CNN) solves the problems of the traditional object detection methods. It uses the convolution operation to extract target information characteristics, which significantly improves the accuracy of object detection. With the improvement of computing power in recent years, the development of graphics processing units (GPUs), and the maturity of big data technology, deep neural networks based on CNNs have been developed in many fields (Choi, 2018; Hannun et al., 2019; Fei et al., 2020; Song et al., 2020; Hussain et al., 2021). For instance, in the ecotoxicology field, the deep learning method is used to train a predictive model that can support hazard assessment and eventually the design of safer engineered nanomaterials (ENMs) (Karatzas et al., 2020). In the field of astronomy, a class of advanced machine learning techniques has been applied to autonomously confirm and classify potential meteor tracks in video imagery. Deep learning has been shown to perform remarkably well, even surpassing the human. Note that it might supplant human visual inspection and review in meteor imagery

collection tasks (Gural, 2019).

Since 2014, the region-based CNN (R-CNN) (Girshick et al., 2014; Girshick, 2015; Ren et al., 2015) series of networks have achieved unprecedented results in speed and accuracy. All of them divide object detection into two stages called the region proposal network (RPN) and object classification. However, the RPN branch in the network causes the computation to be too large to meet the needs of high-speed detection.

To solve the problem of the two-stage network being relatively slow, researchers proposed single-stage networks, such as the YOLO series network (Redmon et al., 2016; Redmon and Farhadi, 2017, 2018; Bochkovskiy et al., 2020; Hsu and Lin, 2020) and single shot multibox detector (SSD) (Liu W et al., 2016; Ming et al., 2022). The single-stage network merges the location problem and the classification problem into a regression problem. The target's position information and category information can be directly obtained in the output layer, which significantly improves the detection speed. Notably, YOLOv3 was compared with Faster R-CNN and SSD (Benjdira et al., 2019; Park et al., 2019), and the results showed that YOLOv3 is superior in both speed and accuracy. YOLOv4 is improved especially with respect to speed and accuracy as compared to YOLOv3. Among the YOLO series networks, YOLOv4 is particularly outstanding in terms of speed and accuracy, and the YOLOv4 algorithm is therefore adopted in our work as the basic detection algorithm. As an extension of our previous work (Tian et al., 2021), this paper adds a new architecture of YOLOv4 with four scales to improve the detection accuracy, a more detailed analysis of the underwater object detection experiments, and a robotic grasping application case.

The primary contributions of this paper are as follows:

1. To better achieve high-precision target detection, YOLOv4 is converted into a four-scale detection network, from 13, 26, and 52 to 13, 26, 52, and 104. Thanks to this improvement, targets in various sizes can be better considered.

2. Because the robot has higher requirements for real-time detection, the YOLOv4 model is pruned to reduce a lot of unnecessary calculations. Finally, because the weight file is only 9.455% of the original weight file, the frame rate can reach 66.67 frames/s and the mean average precision (mAP) is 95.099%.

2 Underwater garbage cleaning robot

Traditional underwater garbage cleanup relies mainly on manual salvage, which is time-consuming and labor-intensive. Workers need to return to the water surface to rest and change equipment regularly. Furthermore, due to the perilous underwater environment, workers may be attacked by aggressive marine organisms and may be hooked by corals. Working in water with chemical pollution or radioactive materials may affect workers' health. In this work, we design an underwater garbage collection robot that can independently identify and search for underwater garbage targets based on a vision system, accurately locate, and finally approach and capture targets.

In recent years, robots have become more and more mature, and their functions have been improved continuously. For example, a robot has been designed that efficiently and autonomously captures garbage on the water surface (Li XL et al., 2020; Kong et al., 2021). With the increasing awareness of environmental protection, target detection for underwater environments has also become a research hotspot (Valdenegro-Toro, 2019; Hong et al., 2020).

Our underwater garbage cleaning robot is illustrated in Fig. 1. The robot has a binocular camera to achieve object detection and position measurement. Four propellers are installed at its four bottom corners to assist the robot in moving forward, backward, left, and right. Two propellers installed on the left and right sides control the robot's motion in heave, and depth-keeping movement can be achieved by controlling these two propellers. Control of these six propellers allows the robot to realize flexible movements. There is a manipulator at the front of the robot for grasping the target.

3 Underwater object detection

3.1 Underwater garbage dataset

A survey by the World Wildlife Fund (WWF) shows that nearly one million tons of worn-out fishing nets are discarded and left in the ocean every year. Almost all of these fishing nets are made of

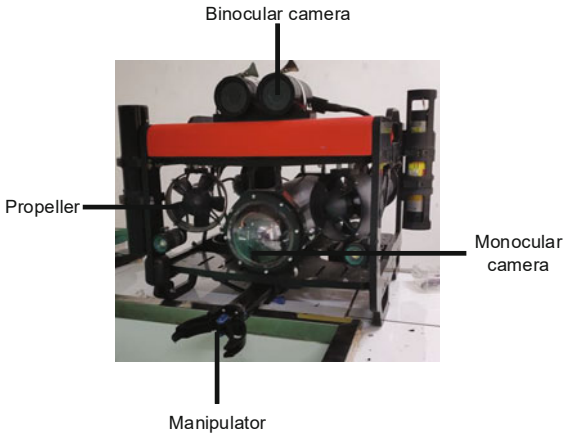

Fig. 1 Illustration of the developed vision-based garbage capture robot

non-degradable plastic, and will remain in the ocean for hundreds of years. Marine organisms are often entwined in fishing nets, which makes it impossible to move freely for food, and eventually leads to entrapment and death. Every year, countless whales, sea turtles, and dolphins die from plastic bags or plastic fragments, and the number is increasing year by year. It is not uncommon for marine creatures to be killed by accidentally eating a large number of plastic bags. During dissection, a large amount of non-degradable plastic waste was found in the stomachs of dead marine life. As a result, the populations of many rare marine species are seriously threatened.

In our work, the main targets to be detected include plastic bags and damaged fishing nets underwater (Fig. 2). Table 1 shows the specific configuration of the proposed dataset. Our work includes 6600 images: the training set consists of 6200 images and the other 400 pieces are used as the test set. The images in the dataset were taken mainly by a camera in a real underwater environment from different viewing angles and moments, taking into consideration different illumination conditions and relative orientations between the robot and targets. The brightness and clarity of some images have been adjusted to simulate different water quality environments. There are many obstacles in the underwater environment, and the reason for detecting stones in our work is to realize the obstacle avoidance operation of the robot. The images are labeled with LabelImg, and the target in the image is enclosed in a rectangular box. The coordinate information (the upper left and lower right corners) and object box category information are recorded in an XML file.

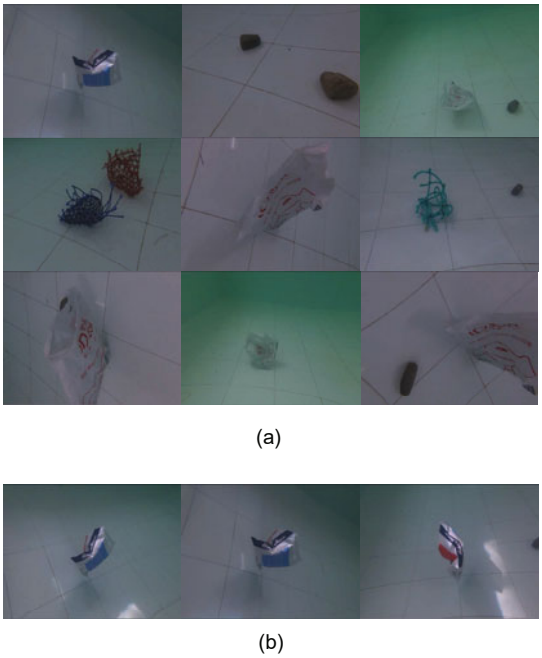

Fig. 2 Dataset of underwater garbage: (a) from different times and clarity with various conditions of brightness, to better simulate real environments; (b) from different view angles, to imitate the robot approaching the target

Table 1 Experimental parameters*

Parameter	Value
Number of training set images	6200
Number of test set images	400
Number of nets in the training set	4661
Number of nets in the test set	340
Number of bags in the training set	2784
Number of bags in the test set	255
Number of stones in the training set	6170
Number of stones in the test set	356
Size of input images	416 × 416
Learning rate	0.001
Batch	64

*GPU: GTX 1080Ti × 3

3.2 Underwater object detection based on YOLOv4

The YOLO series network is a typical single-stage detection algorithm. It extracts image features through a large number of CNN-based layers; the target's position information and category information are directly returned eventually, which improves the detection accuracy and significantly increases the detection speed.

The overall network structure of YOLOv4 is very similar to that of YOLOv3. However, YOLOv4

includes some excellent technologies to strike a balance of detection speed and accuracy. The network structure is shown in Fig. 3a. YOLOv4 can realize multi-scale and multi-size object detection. It will uniformly modify the size of any input image to an image of 416×416×3. After the image passes through the YOLOv4 network, feature maps will be generated in three sizes in the output layer: 13×13×[3×(4+1+3)], 26×26×[3×(4+1+3)], and 52×52×[3×(4+1+3)]. Note that 13, 26, and 52 are the width and height of the feature maps. The depth of the feature maps, [3×(4+1+3)], indicates that there are three anchor boxes with different sizes to predict objects of different sizes. The "4" represents the bounding box coordinate information and "1" is the object confidence information; if

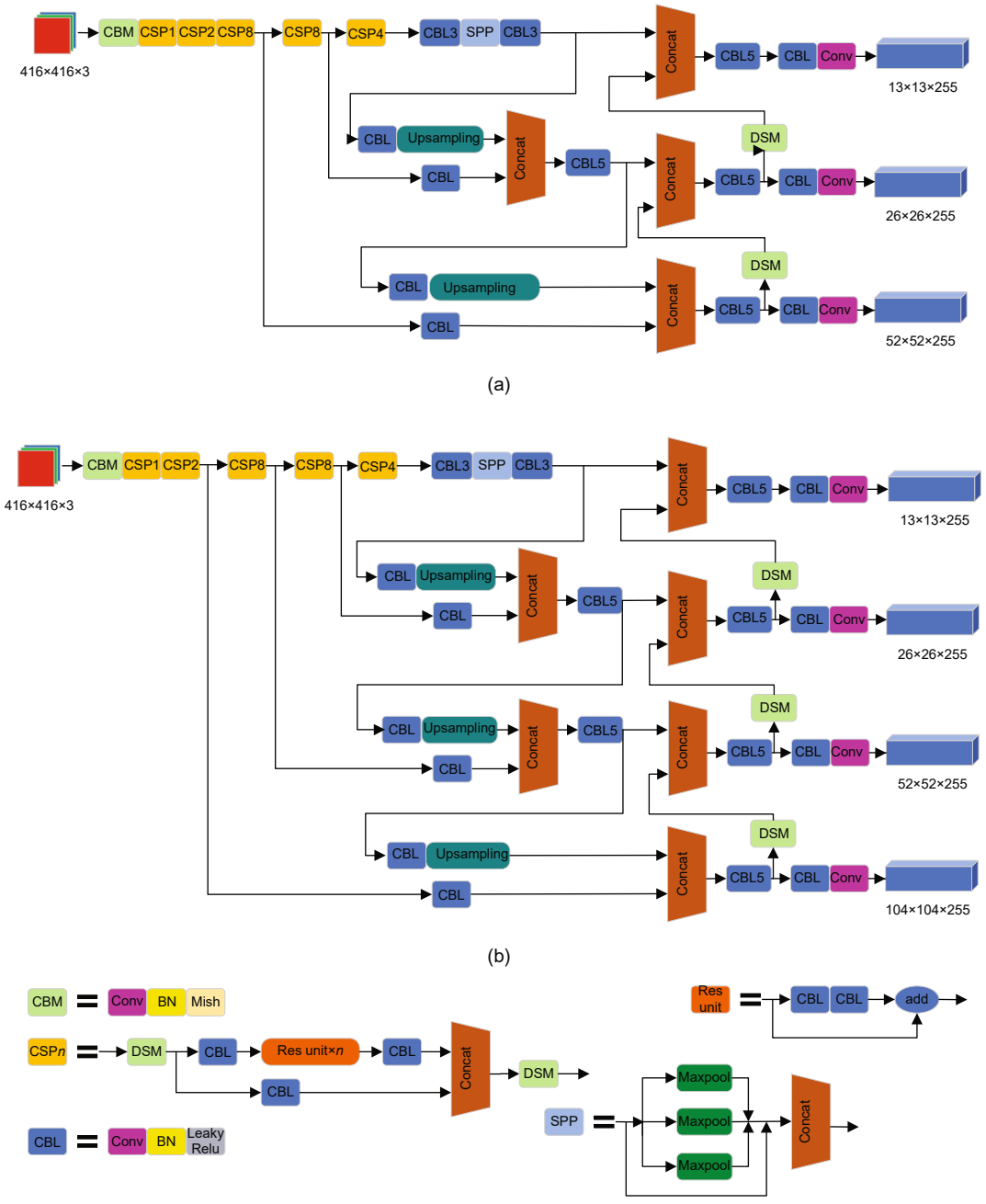

Fig. 3 Network architecture: (a) original YOLOv4; (b) 4S-YOLOv4

the bounding box contains the target, its value is close to 1; otherwise, it is close to 0. The last "3" is the category number of the target. The YOLOv4 extraction network is CSPDarknet53. Compared with Darknet53, CSPDarknet53 can extract more image features. The cross stage partial connections (CSP) (Wang et al., 2020) structure can reduce the amount of computation and enhance the gradient, strengthening the learning ability of the CNN and reducing the computational cost and memory consumption. In addition, it inserts the spatial pyramid pooling (SPP) (He et al., 2015) module between the main network and the final output layer, which effectively increases the feature acceptance range of the backbone network, significantly separates the most important context features, and greatly expands the receptive field. The network also introduces the feature pyramid network (FPN) (Lin et al., 2017) and pyramid attention network (PAN) (Li HP et al., 2018) structures. The FPN layer conveys strong semantic information by upsampling from top to bottom, while PAN conveys strong location features by downsampling from bottom to top. Consequently, feature fusion of different detection layers from different backbone layers is realized and further improves feature extraction capabilities. Furthermore, weighted residual connections (WRC), CSP, cross mini batch normalization (CmBN), and self-adversarial training (SAT) have been added to the network.

The activation function of YOLOv4 is also different from that of YOLOv3, that is,

$$\text{Mish} = x \cdot \tanh(\ln(1 + \mathrm{e}^2)). \tag{1}$$

The function image is shown in Fig. 4. It is not completely cut off when the value is negative, but allows a relatively small negative gradient to flow in, which ensures the flow of information. Additionally, because the activation function has no boundaries, the problem of gradient saturation is avoided. The Mish function can also ensure the smoothness of each point, which improves the gradient descent effect.

YOLOv4 adopts mosaic data enhancement. Its main steps are: (1) read four images at a time; (2) flip them, zoom them, or change color gamut; (3) arrange them in four directions; (4) calculate the data of four original images at one time. The process is shown in Fig. 5. This process greatly enriches the image background information and the diversity of the dataset even when there is insufficient data, ensuring better training results.

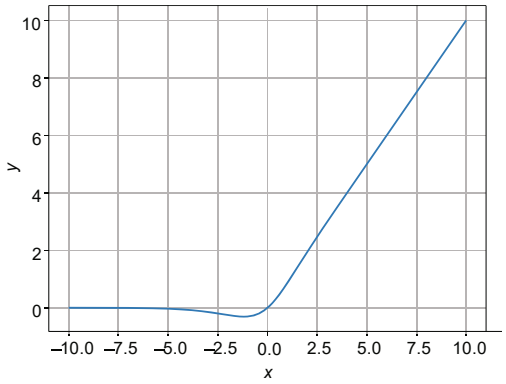

Fig. 4 Mish function image

YOLOv4 uses the complete intersection over union (CIoU) loss function. The IoU loss function does not consider the distance between the two boxes when the two boxes do not intersect. The generalized intersection over union (GIoU) solves this problem, but when the predicted bounding box is within the ground truth box, it cannot distinguish the relative position relationship. The emergence of the distance intersection over union (DIoU) resolves the drawbacks of GIoU, but DIoU does not take the width-to-height ratio between the predicted box and the ground truth box into account. CIoU solves all the above problems. It takes the overlapping area, center point distance, and width-to-height ratio between the predicted box and the ground truth box into consideration; the formula is as follows:

$$\text{CIoU} = \text{IoU} - \frac{\rho^2(b, b^{\text{gt}})}{c^2} - \alpha v, \tag{2}$$

where $\rho^2(b, b^{\text{gt}})$ is the distance of central points of two boxes, c is the diagonal length of the smallest enclosing box covering two boxes, α is a positive trade-off parameter, and v measures the aspect ratio consistency as follows:

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w^{\text{gt}}}{h^{\text{gt}}} - \arctan\frac{w}{h}\right)^2, \tag{3}$$

where w and h are the width and height, respectively. Then the loss function can be defined as

$$L_{\text{CIoU}} = 1 - \text{IoU} + \frac{\rho^2(b, b^{\text{gt}})}{c^2} + \alpha v. \tag{4}$$

The trade-off parameter α is defined as

$$\alpha = \frac{v}{(1 - \text{IoU}) + v}. \tag{5}$$

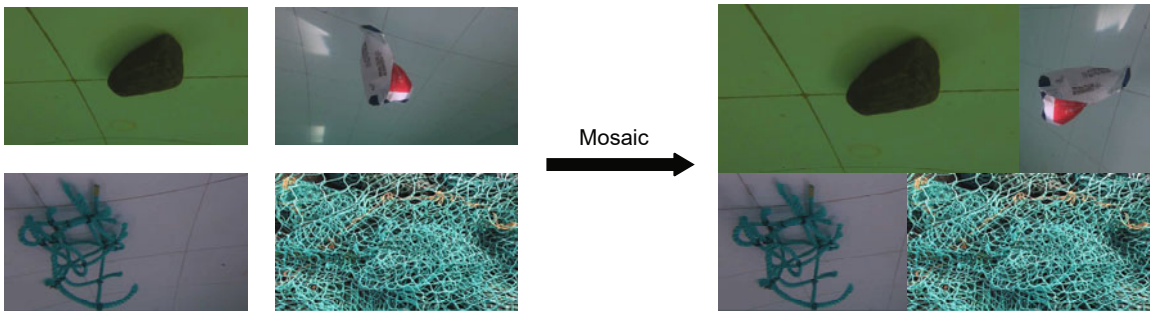

Fig. 5 Mosaic data augmentation

4 Improvements in YOLOv4

4.1 4S-YOLOv4

In the robot's actual working area, when the robot gradually approaches the target object, the object's image is gradually enlarged. In other words, there is plenty of size information about the target object. To better consider target objects of various sizes, and to further improve the accuracy of the detection network, our work adds a detection scale to the original YOLOv4, which is transformed from the original 13×13, 26×26, and 52×52. It becomes 13×13, 26×26, 52×52, and 104×104, and is called 4S-YOLOv4. Fig. 3b shows a schematic of its network structure.

4.2 Pruned 4S-YOLOv4

In recent years, with the improvement in computing power, CNNs have gradually become the main method in the field of computer vision. More and more types of excellent deep neural networks have been proposed, such as VGGNet (Simonyan and Zisserman, 2015), GoogleNet (Szegedy et al., 2015), ResNets (He et al., 2016), and AlexNet (Krizhevsky et al., 2017). Researchers try to improve network performance by optimizing the network structure, adjusting the loss function, and improving the training method, to continuously optimize the accuracy and speed of the network. When the depth of the neural network increases, the number of parameters inevitably increases. Although a larger and deeper neural network can more fully extract the target's feature information, which is of great help in improving the detection accuracy, the requirements for the network's hardware performance are also increasing. A ResNet network with 152 layers will have more than 60 million parameters, and it will require more than 20 gigabit floating point operations when inferring an image with a resolution of 224×224 (Liu Z et al., 2017). This may not be possible on a platform with low-level hardware, so it will not be able to perform optimally.

To achieve high-speed and high-precision detection, we prune the 4S-YOLOv4 network, and thus greatly reduce the number of parameters and calculation of the model, as well as the resource occupation of the model. Therefore, it enormously improves the inference speed of the model.

The core of the channel pruning algorithm is to delete the channel with the smallest contribution and its input-output relationship by calculating the channel in the network (Molchanov et al., 2019). This enhances the channel-level sparsity of the convolutional layer by applying L_1 regularization on the channel scaling factor. Therefore, pruning feature channels leads to a slim object detector. The main formula is

$$L=\sum_{(x,y)} l(f(x,W),y)+\lambda\sum_{\gamma\in\Gamma} g(\gamma), \tag{6}$$

where x and y denote the training input and the target respectively, W denotes the trainable weight, $g(\cdot)$ is a sparsity-induced penalty on the scaling factors, Γ is the scaling factor of the batch normalization layer, and λ balances the two terms. The schematic is shown in Fig. 6.

5 Experiments and results

5.1 Testing results

In this experiment, a detection task involving three types of target objects—fishing nets, plastic bags, and stones—was implemented on the GTX

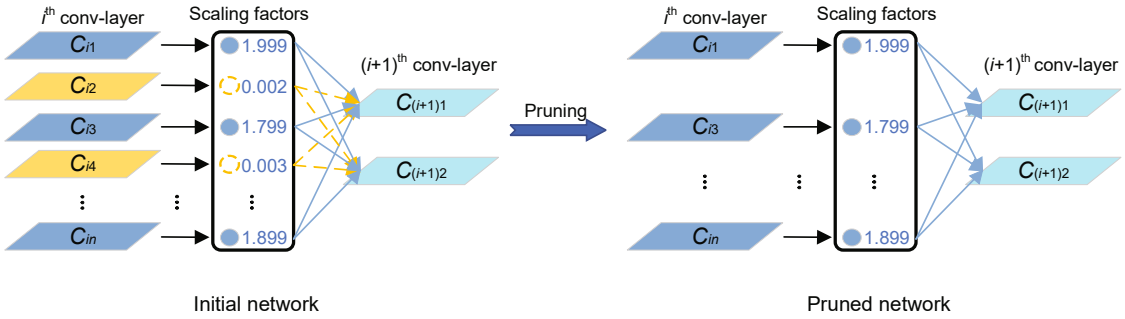

Fig. 6 Channel pruning diagram

1080Ti × 3. Then, the frame rate, mAP, and average precision (AP) for each category of the proposed 4SP-YOLOv4 (Pruned 4S-YOLOv4), YOLOv4, 4S-YOLOv4, YOLOv3, SSD, and Faster R-CNN were calculated for quantitative analysis. The weight file sizes of these networks were also compared, and the comparison results are shown in Fig. 7. The final quantitative results are listed in Table 2, and the detection performance of 4SP-YOLOv4 is shown in Fig. 8.

We also conducted a robot underwater target capturing experiment to verify whether the robot can adapt to different environments with the help of vision algorithms, and observed its state in a complex environment to determine whether the robot has the ability to solve practical problems. Underwater experiments are shown in Figs. 9 and 10. This robot carried a binocular camera, responsible for obtaining the objects' position information using the stereo principle. Note that due to the soft texture of broken plastic bags and nets, it is not necessary to measure their specific orientation to accomplish grasping. In the real underwater environment, the robot can detect objects in time and accurately. Then, according to the detected object information, the robot is able to approach the object and successfully grasp it.

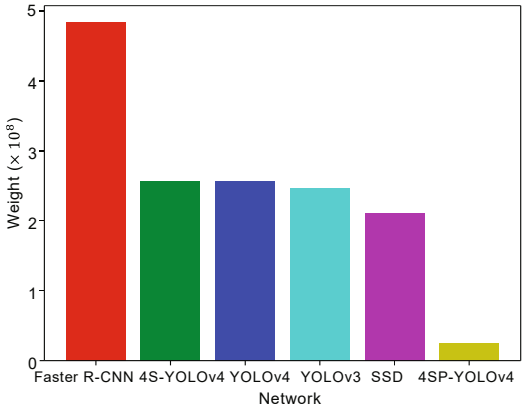

Fig. 7 Weight comparison chart

5.2 Discussion

The detection effect of 4SP-YOLOv4 is shown in Fig. 8, and the results of comparison with other detection methods are listed in Table 2. We compared the weight file sizes of different network models, and the results are shown in Fig. 7. The results showed that 4SP-YOLOv4 was remarkable in all aspects in this experimental scenario. After the conversion to 4S-YOLOv4, the accuracy was improved, which proves the correctness of this improvement. However, due to the addition of a scale of calculation, its speed dropped a little. Obviously, 4S-YOLOv4

Table 2 Detection results on the dataset

Network	FPS	AP_{net}	AP_{bag}	AP_{stone}	mAP	Size of weight (byte)
4SP-YOLOv4	**66.667**	**0.970**	0.899	0.984	0.951	**2.4323×10^7**
4S-YOLOv4	36.364	0.966	**0.913**	**0.985**	**0.955**	2.5726×10^8
YOLOv4 (Bochkovskiy et al., 2020)	43.478	0.914	0.908	0.919	0.913	2.5606×10^8
YOLOv3 (Redmon and Farhadi, 2018)	38.030	0.887	0.873	0.897	0.886	2.4635×10^8
Faster R-CNN (Ren et al., 2015)	7.143	0.878	0.857	0.903	0.879	4.8349×10^8
SSD (Liu W et al., 2016)	14.545	0.877	0.870	0.890	0.879	2.1033×10^8

The best results are in bold

Fig. 8 Detection results of 4SP-YOLOv4

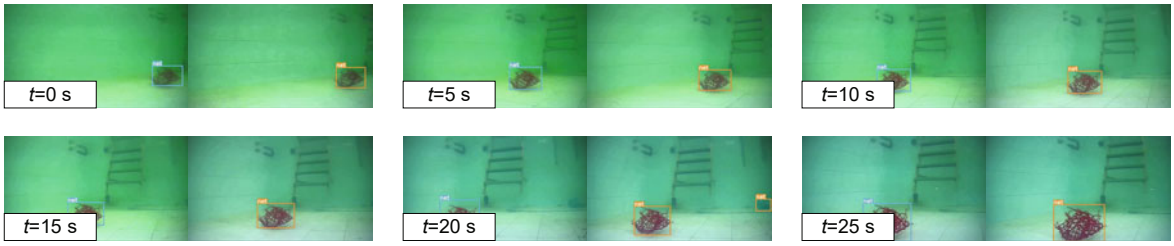

Fig. 9 Schematic of robot underwater detection

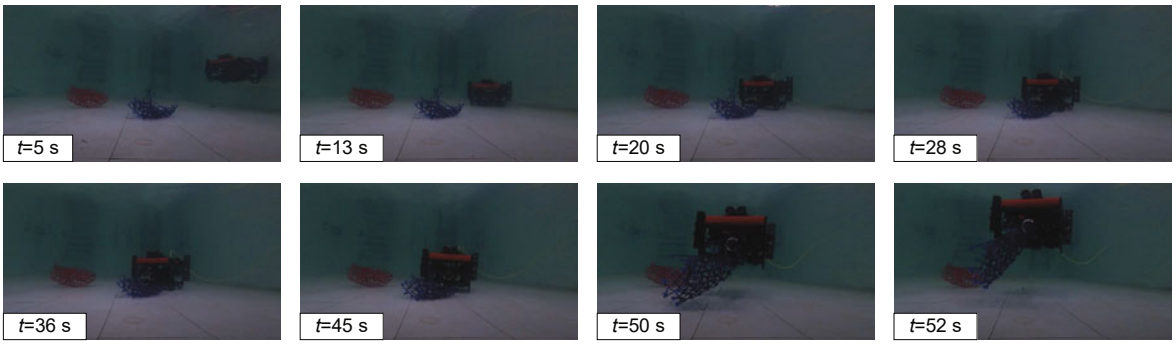

Fig. 10 Schematic of robot underwater capturing

cannot cope with high-speed application scenarios, so it prunes the model channels that contribute little to classification and positioning to minimize the loss of model accuracy. It also reduces a lot of unnecessary calculations and improves the inference speed of the model, so it can realize high precision and achieve high speed.

Because the single-stage structure of the YOLO series network allows them to achieve high speed and high precision, YOLOv3 has achieved a perfect balance between accuracy and speed. YOLOv4 is improved based on YOLOv3, and has made certain improvements in accuracy and speed. Compared with YOLOv3 and YOLOv4, the SSD feature extraction network was not as good as Darknet-53 for feature extraction. Darknet-53 achieved higher precision than ResNet-101 and was 1.5 times faster. The performance of Darknet-53 was similar to that of ResNet-152 and it was 2 times faster. Darknet-53 also achieved the highest measured number of floating-point operations per second (Redmon and Farhadi, 2018). However, to further improve the accuracy, our work added a detection scale to YOLOv4. To improve the detection speed, the 4S-YOLOv4

model was pruned. Finally, compared with the original YOLOv4, the accuracy of 4SP-YOLOv4 was 4.162% higher, the frame rate was 53.335% larger, and the weight was only 9.499% of that of the original model.

The experimental results indicated that 4SP-YOLOv4 can provide the robot with real-time and accurate object detection. High-speed detection can process the image in real time and provide the object position information for the robot in time, in response to the changeable and complex environment. Even though Faster R-CNN made a significant breakthrough in accuracy, it cannot achieve real-time detection due to the computational burden from the two-stage network. Briefly, 4SP-YOLOv4 exhibits better performance than the others in both speed and accuracy, and can effectively improve the capability of autonomous cleaning robots. Based on the aforementioned results, the proposed algorithm outperforms those in previous studies and is suitable for autonomous cleaning robots. However, for better robustness in complex aquatic environments, a more adequate dataset is indispensable. Additionally, the core contribution of this work is about the pruning strategy, which is effective in datasets with a small number of classes, like the underwater dataset we have proposed. However, this method performs slightly worse when tested on datasets with a larger number of classes, such as COCO, because the pruning process reduces the scale of the network. When there are more categories of objects in the dataset, the pruned network will be insufficient. Further research is therefore needed to improve the performance.

6 Conclusions and future work

A good detection method is essential for intelligent cleaning robots, and accuracy and speed are critical for detection algorithms. Traditional methods cannot simultaneously meet the requirements of speed and accuracy. In this paper, we propose an improved YOLOv4 algorithm. First, we convert the original YOLOv4 to 4-scale YOLOv4; then we perform model pruning on 4S-YOLOv4. Compared with other detection algorithms, 4SP-YOLOv4 can achieve 0.951 mAP in 15 ms at GTX 1080Ti × 3, and the number of parameters is only 9.499% of that of the original model, ensuring high-precision and high-speed object detection.

With only 9.499% model parameters, the hardware configuration requirement of 4SP-YOLOv4 is not as high as that of the original YOLOv4 model. In other words, it is possible to achieve high-speed and high-precision detection on devices with low configurations. The core contribution of this work is about the pruning strategy, which is effective for datasets with a small number of classes like the underwater dataset that we have proposed. In the future, we will try to transfer the model to mobile devices, such as mobile phones and cameras.

Contributors

Manjun TIAN designed the research. Manjun TIAN, Xiali LI, and Shihan KONG proposed the methods. Manjun TIAN and Shihan KONG conducted the experiments. Licheng WU and Junzhi YU processed the data. Manjun TIAN and Shihan KONG drafted the paper. Xiali LI, Licheng WU, and Junzhi YU helped organize the paper. Shihan KONG and Junzhi YU revised and finalized the paper.

Compliance with ethics guidelines

Manjun TIAN, Xiali LI, Shihan KONG, Licheng WU, and Junzhi YU declare that they have no conflict of interest.

References

Albitar H, Dandan K, Ananiev A, et al., 2016. Underwater robotics: surface cleaning technics, adhesion and locomotion systems. *Int J Adv Robot Syst*, 13(1):7. https://doi.org/10.5772/62060

Astapov S, Preden JS, Ehala J, et al., 2014. Object detection for military surveillance using distributed multimodal smart sensors. Proc 19th Int Conf on Digital Signal Processing, p.366-371. https://doi.org/10.1109/ICDSP.2014.6900688

Bai JQ, Lian SG, Liu ZX, et al., 2018. Deep learning based robot for automatically picking up garbage on the grass. *IEEE Trans Consum Electron*, 64(3):382-389. https://doi.org/10.1109/TCE.2018.2859629

Benjdira B, Khursheed T, Koubaa A, et al., 2019. Car detection using unmanned aerial vehicles: comparison between faster R-CNN and YOLOv3. Proc 1st Int Conf on Unmanned Vehicle Systems-Oman, p.1-6. https://doi.org/10.1109/UVS.2019.8658300

Bochkovskiy A, Wang CY, Liao HYM, 2020. YOLOv4: optimal speed and accuracy of object detection. https://arxiv.org/abs/2004.10934

Choi H, 2018. Deep learning in nuclear medicine and molecular imaging: current perspectives and future directions. *Nucl Med Mol Imag*, 52(2):109-118. https://doi.org/10.1007/s13139-017-0504-7

Dalal N, Triggs B, 2005. Histograms of oriented gradients for human detection. Proc IEEE Computer Society Conf on Computer Vision and Pattern Recognition, p.886-893. https://doi.org/10.1109/CVPR.2005.177

Ekins P, Gupta J, 2019. Perspective: a healthy planet for healthy people. *Glob Sustain*, 2:1-9. https://doi.org/10.1017/sus.2019.17

Fei Y, Wang KCP, Zhang A, et al., 2020. Pixel-level cracking detection on 3D asphalt pavement images through deep-learning-based crackNet-V. *IEEE Trans Intell Transp Syst*, 21(1):273-284. https://doi.org/10.1109/TITS.2019.2891167

Felzenszwalb P, McAllester D, Ramanan D, 2008. A discriminatively trained, multiscale, deformable part model. IEEE Int Conf on Computer Vision and Pattern Recognition, p.24-26.

Fu ZH, Chen YW, Yong HW, et al., 2019. Foreground gating and background refining network for surveillance object detection. *IEEE Trans Image Process*, 28(12):6077-6090. https://doi.org/10.1109/TIP.2019.2922095

Girshick R, 2015. Fast R-CNN. Proc IEEE Int Conf on Computer Vision, p.1440-1448. https://doi.org/10.1109/ICCV.2015.169

Girshick R, Donahue J, Darrell T, et al., 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.580-587. https://doi.org/10.1109/CVPR.2014.81

Gural PS, 2019. Deep learning algorithms applied to the classification of video meteor detections. *Mon Not R Astron Soc*, 489(4):5109-5118. https://doi.org/10.1093/mnras/stz2456

Hannun AY, Rajpurkar P, Haghpanahi M, et al., 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat Med*, 25(1):65-69. https://doi.org/10.1038/s41591-018-0268-3

He KM, Zhang XY, Ren SQ, et al., 2015. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans Patt Anal Mach Intell*, 37(9):1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

He KM, Zhang XY, Ren SQ, et al., 2016. Identity mappings in deep residual networks. Proc 14th European Conf on Computer Vision, p.630-645. https://doi.org/10.1007/978-3-319-46493-0-38

Hong J, Fulton M, Sattar J, 2020. A generative approach towards improved robotic detection of marine litter. Proc IEEE Int Conf on Robotics and Automation, p.10525-10531. https://doi.org/10.1109/ICRA40945.2020.9197575

Horng GJ, Liu MX, Chen CC, 2020. The smart image recognition mechanism for crop harvesting system in intelligent agriculture. *IEEE Sens J*, 20(5):2766-2781. https://doi.org/10.1109/JSEN.2019.2954287

Hsu WY, Lin WY, 2020. Ratio-and-scale-aware YOLO for pedestrian detection. *IEEE Trans Image Process*, 30:934-947. https://doi.org/10.1109/TIP.2020.3039574

Hussain E, Hasan M, Rahman A, et al., 2021. CoroDet: a deep learning based classification for COVID-19 detection using chest X-ray images. *Chaos Sol Fract*, 142:110495. https://doi.org/10.1016/j.chaos.2020.110495

Jambeck JR, Geyer R, Wilcox C, et al., 2015. Plastic waste inputs from land into the ocean. *Science*, 347(6223):768771. https://doi.org/10.1126/science.1260352

Karatzas P, Melagraki G, Ellis LJA, et al., 2020. Development of deep learning models for predicting the effects of exposure to engineered nanomaterials on *Daphnia magna*. *Small*, 16(36):2001080. https://doi.org/10.1002/smll.202001080

Kim J, Mishra AK, Limosani R, et al., 2019. Control strategies for cleaning robots in domestic applications: a comprehensive review. *Int J Adv Robot Syst*, 16(4):1-21. https://doi.org/10.1177/1729881419857432

Kong SH, Tian MJ, Qiu CL, et al., 2021. IWSCR: an intelligent water surface cleaner robot for collecting floating garbage. *IEEE Trans Syst Man Cybern Syst*, 51(10):6358-6368. https://doi.org/10.1109/TSMC.2019.2961687

Krizhevsky A, Sutskever I, Hinton GE, 2017. ImageNet classification with deep convolutional neural networks. *Commun ACM*, 60(6):84-90. https://doi.org/10.1145/3065386

Laschi C, Mazzolai B, Cianchetti M, 2016. Soft robotics: technologies and systems pushing the boundaries of robot abilities. *Sci Robot*, 41(1):eaah3690. https://doi.org/10.1126/scirobotics.aah3690

Li CY, Guo CL, Ren WQ, et al., 2019. An underwater image enhancement benchmark dataset and beyond. *IEEE Trans Image Process*, 29:4376-4389. https://doi.org/10.1109/TIP.2019.2955241

Li HP, Xiong PF, An J, et al., 2018. Pyramid attention network for semantic segmentation. Proc British Machine Vision Conf, p.285.

Li XL, Tian MJ, Kong SH, et al., 2020. A modified YOLOv3 detection method for vision-based water surface garbage capture robot. *Int J Adv Robot Syst*, 17(3):1-11. https://doi.org/10.1177/1729881420932715

Lin TY, Dollár P, Girshick R, et al., 2017. Feature pyramid networks for object detection. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.936-944. https://doi.org/10.1109/CVPR.2017.106

Liu W, Anguelov D, Erhan D, et al., 2016. SSD: single shot multibox detector. European Conf on Computer Vision, p.21-37. https://doi.org/10.1007/978-3-319-46448-0-2

Liu Z, Li JG, Shen ZQ, et al., 2017. Learning efficient convolutional networks through network slimming. Proc IEEE Int Conf on Computer Vision, p.2755-2763. https://doi.org/10.1109/ICCV.2017.298

Lowe DG, 2004. Distinctive image features from scale-invariant keypoints. *Int J Comput Vis*, 60(2):91-110. https://doi.org/10.1023/B:VISI.0000029664.99615.94

Mahler J, Pokorny FT, Hou B, et al., 2016. Dex-Net 1.0: a cloud-based network of 3D objects for robust grasp planning using a multi-armed bandit model with correlated rewards. Proc IEEE Int Conf on Robotics and Automation, p.1957-1964. https://doi.org/10.1109/ICRA.2016.7487342

Mahler J, Liang J, Niyaz S, et al., 2017. Dex-Net 2.0: deep learning to plan robust grasps with synthetic point clouds and analytic grasp metrics. https://arxiv.org/abs/1703.09312

Mahler J, Matl M, Liu XY, et al., 2018. Dex-Net 3.0: computing robust vacuum suction grasp targets in point clouds using a new analytic model and deep learning. Proc IEEE Int Conf on Robotics and Automation, p.5620-5627. https://doi.org/10.1109/ICRA.2018.8460887

Mahler J, Matl M, Satish V, et al., 2019. Learning ambidextrous robot grasping policies. *Sci Robot*, 4(26): eaau4984. https://doi.org/10.1126/scirobotics.aau4984

Mhalla A, Chateau T, Gazzah S, et al., 2019. An embedded computer-vision system for multi-object detection in traffic surveillance. *IEEE Trans Intell Transp Syst*, 20(11):4006-4018. https://doi.org/10.1109/TITS.2018.2876614

Ming X, Wei FY, Zhang T, et al., 2022. Group sampling for scale invariant face detection. *IEEE Trans Patt Anal Mach Intell*, 44(2):985-1001. https://doi.org/10.1109/TPAMI.2020.3012414

Molchanov P, Mallya A, Tyree S, et al., 2019. Importance estimation for neural network pruning. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition, p.11256-11264. https://doi.org/10.1109/CVPR.2019.01152

Ojala T, Pietikäinen M, Maenpaa T, 2002. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans Patt Anal Mach Intell*, 24(7):971-987. https://doi.org/10.1109/TPAMI.2002.1017623

Ostle C, Thompson RC, Broughton D, et al., 2019. The rise in ocean plastics evidenced from a 60-year time series. *Nat Commun*, 10(1):1622. https://doi.org/10.1038/s41467-019-09506-1

Park JH, Hwang HW, Moon JH, et al., 2019. Automated identification of cephalometric landmarks: Part 1—comparisons between the latest deep-learning methods YOLOV3 and SSD. *Angle Orthod*, 89(6):903-909. https://doi.org/10.2319/022019-127.1

Prabakaran V, Elara MR, Pathmakumar T, et al., 2018. Floor cleaning robot with reconfigurable mechanism. *Autom Constr*, 91:155-165. https://doi.org/10.1016/j.autcon.2018.03.015

Pu SL, Zhao W, Chen WJ, et al., 2021. Unsupervised object detection with scene-adaptive concept learning. *Front Inform Technol Electron Eng*, 22(5):638-651. https://doi.org/10.1631/FITEE.2000567

Redmon J, Farhadi A, 2017. YOLO9000: better, faster, stronger. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.6517-6525. https://doi.org/10.1109/CVPR.2017.690

Redmon J, Farhadi A, 2018. YOLOv3: an incremental improvement. https://arxiv.org/abs/1804.02767

Redmon J, Divvala S, Girshick R, et al., 2016. You only look once: unified, realtime object detection. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.779-788. https://doi.org/10.1109/CVPR.2016.91

Ren SQ, He KM, Girshick RB, et al., 2015. Faster R-CNN: towards real-time object detection with region proposal networks. Proc Annual Conf on Neural Information Processing Systems, p.91-99.

Simonyan K, Zisserman A, 2015. Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556

Song ZG, Zou SM, Zhou WX, et al., 2020. Clinically applicable histopathological diagnosis system for gastric cancer detection using deep learning. *Nat Commun*, 11(1):4294. https://doi.org/10.1038/s41467-020-18147-8

Szegedy C, Liu W, Jia YQ, et al., 2015. Going deeper with convolutions. Proc IEEE Conf on Computer Vision and Pattern Recognition, p.1-9. https://doi.org/10.1109/CVPR.2015.7298594

Tian MJ, Li XL, Kong SH, et al., 2021. Pruning-based YOLOv4 algorithm for underwater gabage detection. Proc 40th Chinese Control Conf, p.4008-4013. https://doi.org/10.23919/CCC52363.2021.9550592

Tschandl P, 2020. Problems and potentials of automated object detection for skin cancer recognition. *JAMA Dermatol*, 156(1):23-24. https://doi.org/10.1001/jamadermatol.2019.3360

Valdenegro-Toro M, 2019. Deep neural networks for marine debris detection in sonar images. https://arxiv.org/abs/1905.05241

Viola P, Jones M, 2001. Rapid object detection using a boosted cascade of simple features. Proc IEEE Computer Society Conf on Computer Vision and Pattern Recognition, p.511-518. https://doi.org/10.1109/CVPR.2001.990517

Wang CY, Liao HYM, Wu YH, et al., 2020. CSPNet: a new backbone that can enhance learning capability of CNN. Proc IEEE/CVF Conf on Computer Vision and Pattern Recognition Workshops, p.1571-1580. https://doi.org/10.1109/CVPRW50498.2020.00203

Whitehill J, Omlin CW, 2006. Haar features for FACS AU recognition. Proc 7th Int Conf on Automatic Face and Gesture Recognition, p.5-101. https://doi.org/10.1109/FGR.2006.61

Xu M, Karuppusamy NS, Kang BY, 2017. A novel design to improve the cooperative ability of the multi-cleaning robot in the unknown environment. *Adv Sci Lett*, 23(10):9557-9560. https://doi.org/10.1166/asl.2017.9746