ISSN: 2414-1895 DOI: 10.6919/ICJE.202605_12(5).0013

Research on Obstacle Detection and Distance Measurement for the Blind based on Improved YOLOv7-Tiny

Wei Shen, Liang Yang, Yuan Tao, Wanjun Yao

Anhui Xinhua University, School of Big Data and Artificial Intelligence, Hefei, Anhui 230088, China

Abstract

In the daily walking process of visually impaired individuals, the accuracy of obstacle detection and distance measurement is crucial to their safety. This paper proposes an improved obstacle detection and recognition algorithm based on YOLOv7-Tiny. The algorithm adds a three-dimensional attention mechanism module, SimAM, to the Neck-to-Head part of the lightweight YOLOv7-Tiny network to enhance obstacle category features and improve the recognition accuracy of obstacle categories. At the same time, a monocular ranging method is incorporated during obstacle detection, allowing the distance of the detected obstacles to be determined simultaneously. Experimental results show that the improved model achieves an mAP@0.5 value of 91.4% in evaluation metrics, which is 2% higher than before improvement, with shorter detection time and faster detection speed. The improved model demonstrates better detection and recognition performance, validating its effectiveness.

Keywords

YOLOv7-Tiny; SimAM; Monocular Distance Measurement.

1. Introduction

Nowadays, there are a large number of visually impaired people in the world, including a certain proportion of blind people. For people with blind disabilities, it is difficult to independently identify obstacles on the road, such as steps, pillars, bicycles parked at will, and cars stationary at intersections. Especially in the complex road environment of the city, these greatly limit their range of movement and independence, and may even cause safety accidents, endangering their lives. At present, the commonly used travel aids for blind cane people are blind canes and guide dogs, although blind canes can help perceive close obstacles, but the detection range is small, and it is difficult to detect distant obstacles in advance. The use of guide dogs has problems with long training and training cycles and high costs. In addition to blind canes and guide dogs, electronic aids can also be used, and in complex environments, the recognition ability of the equipment will be seriously reduced, and the system response delay will cause the positioning deviation distance to be large, which cannot provide safe navigation. With the rapid development of computer vision, deep learning and other fields, it provides new possibilities for solving the problem of visual assistance for blind people. Deep learning algorithms can accurately classify various targets in the face of complex environments.

Deep learning algorithms, especially from convolutional neural networks (CNNs) to deep convolutional neural networks (DCNNs), have made significant advancements in the field of image detection and recognition. In 1998, Lecun[1] et al. proposed the first true convolutional neural network - LeNet5, which can be applied to handwritten number recognition, with the continuous development of CNN, more and more convolutional neural networks have emerged, although the recognition effect of LeNet5 is not optimal, but it is the cornerstone of the development of deep convolutional neural networks. Since then, deep convolutional neural networks have developed

ISSN: 2414-1895 DOI: 10.6919/ICJE.202605_12(5).0013

rapidly, and there has been a steady stream of networks used in the field of target recognition, such as AlexNet[2], VGGNet[3], DenseNet[4], ResNet[5], etc. For example, in terms of monitoring and security, deep learning-based models can accurately identify walking vehicles and people in complex and crowded street surveillance images or videos, and can quickly and accurately capture abnormal behaviors in the image picture, such as sudden retrograde of vehicles, running, red light runs at intersections, etc., and can quickly alarm and prompt and deal with them in a timely manner when an emergency is detected, ensuring road traffic safety. For example, in 2020, He Yu of Northeastern University proposed a research on the classification and detection of textured surface defects focusing on the limited number of samples, and proposed a supervised defect detection algorithm for the insufficient number of samples in defect detection. By fine-tuning the pre-trained model ImageNet, the model can be reused on a small amount of annotated data, and on the other hand, it integrates multi-level features to improve the positioning and detection accuracy, which has good applications[6]. In the field of medical health, this technology can achieve precise identification and detection of medical image lesions. At the same time, current AI (artificial intelligence) systems can quickly assist doctors in making diagnostic judgments. For example, in 2022, Dong Li from East China Normal University proposed an automated detection system for CT image lung nodules. According to the need for early lung screening and combined with deep learning technology, a medical image recognition system was established. A false-positive nodule suppression algorithm and a fine segmentation algorithm for lung nodules were proposed, which can effectively identify and locate lung nodules in CT images, providing doctors with certain effective and reliable assistance, greatly improving the efficiency and accuracy of clinical detection and recognition[7]. In the field of automated retail, some large public places such as schools, shopping malls, and hospitals have added unmanned self-service vending cabinets, which use cameras to capture and detect items in real time and perform automatic checkout without staff supervision, saving a large amount of labor costs. In agriculture, an agricultural image processing platform can be established to process and identify images of crops. By using deep learning algorithms, the system can identify types of crop diseases and pests and take timely pest control measures, reducing losses and increasing crop yield. In daily life, mobile phone unlocking, app verification, access control, and face payment are all authenticated through facial recognition[8]. Therefore, deep learning technology has penetrated every aspect of daily life and has become an indispensable part of it.

Similarly, applying deep convolutional neural networks to vision assistance systems for the blind can enable the detection of obstacles on roads for visually impaired individuals. By constructing a dataset of obstacle images in real walking environments for the blind, during the training process, a deep convolutional neural network is used to automatically extract and learn the key features of different target obstacles. For example, common obstacles encountered by visually impaired people walking on tactile paths or sidewalks at intersections, such as electric bicycles, cars, bicycles, and poles, allowing the model to focus on the features of different categories of obstacles. Once the model training is completed, the model can capture environmental images in real time and recognize various target objects in a short time, providing visually impaired individuals with quick and accurate environmental information, thus requiring a high detection speed. This paper uses the lightweight YOLOv7-Tiny[9] model and improves it by adding the SimAM[10]attention mechanism module. This increases accuracy without adding parameters, ensuring detection speed. At the same time, in the process of detecting and recognizing target obstacles, a monocular ranging method is added to measure distance. In addition to detecting and recognizing target objects, it can also measure the distance between a person and the target obstacle. By combining the target recognition method with the ranging function, it can provide visually impaired individuals with a "perceivable at any time" living environment, improving their mobility, while also to some extent enhancing their travel experience and quality of life.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202605_12(5).0013

2. Obstacle Detection and Distance Measurement Method

2.1 Improved YOLOv7-Tiny Network Structure

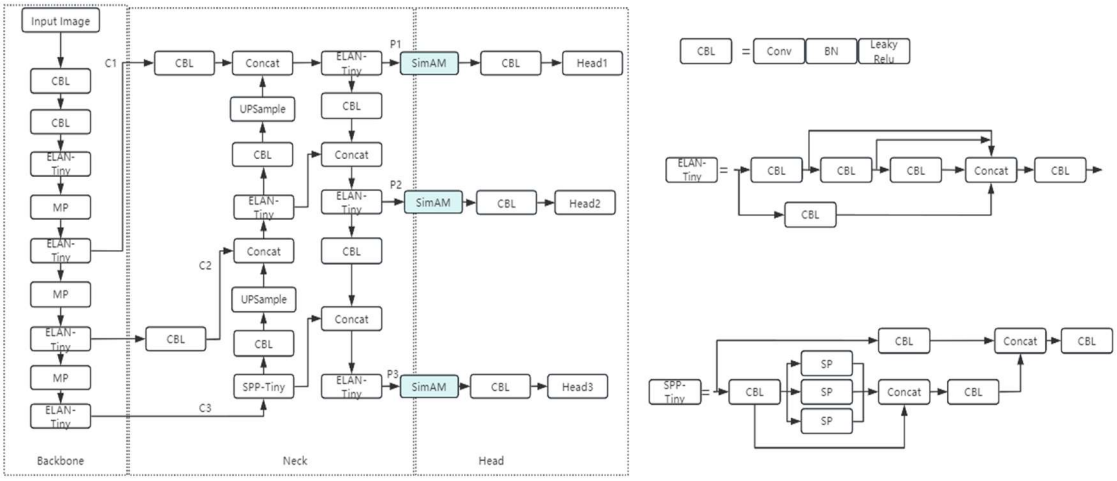

Figure 1. Improved YOLOv7-Tiny Network Structure

Figure 1 shows the overall structure of the improved YOLOv7-Tiny network, which mainly consists of three parts: backbone, neck, and head.

The backbone part is the backbone network of YOLOv7-Tiny, which is also the feature extraction part, which mainly includes the CBL module, ELAN-Tiny module, and the maximum pooling operation MP (MaxPooling) part in Figure 1. The CBL module extracts features through convolutional operations, and then performs feature fusion through the ELAN-Tiny module. In the ELAN-Tiny module, the channels are adjusted by 1×1 convolution, features are extracted by 3×3 convolutions, and finally the features on the four branches are stitched (Concat), so as to enhance the features of the feature map. In the backbone part, three different key scale feature maps (C1, C2, and C3 in Fig. 1) are finally formed for the feature fusion of the neck part, which not only ensures the computational efficiency, but also constructs rich multi-scale features.

The Neck part draws on the architecture idea of PANet[11], adopts two paths, upsampling and downsampling, and the feature map C3 received by the backbone part performs upsampling with the feature map C2 in the middle and the feature map C1 in the shallower part for Concat splicing. The Concat splicing operation is different from the fusion method of FPN[12] networks, which is not feature addition but feature splicing. In the downsampling path, the features output by the ELAN-Tiny module and the SPP-Tiny module are stitched and fused with Concat, and finally the feature maps of different scales used by the detection head are output P1, P2 and P3 in the neck part.

Head1, Head2 and Head3 respectively correspond to three different scale feature maps, and these three scale feature maps are decoded and predicted at the same time, and finally the position, confidence and category probability of different targets will be predicted at their respective scales, realizing efficient prediction of targets at different scales.

In this study, because the image data in the dataset is collected in the real road environment of blind people walking on the road, the collected images will be affected by the complexity of the surrounding environment, weather, lighting, etc., resulting in a decrease in the recognition rate of obstacles in the blind road in the image, especially in foggy, cloudy, rainy and other low-light environments, the recognition rate is low, and the detection effect will be biased. It can enhance the target features without adding additional parameters, which enhances the obstacle features in the image while ensuring the detection speed and improving the target recognition rate. As shown in Figure 1, SimAM modules are added to the output P1, P2, and P3 of the neck of the YOLOv7-Tiny network.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202605_12(5).0013

2.2 SimAM Module

The attention mechanism is like the human eye focusing on a certain point or area. When a person's attention is concentrated on a specific region to focus on the important information in that area, the brain's memory for the important information seen in that region is also strengthened. Similarly, applying the attention mechanism to model training can focus on the important information in feature maps, enhance the feature information of image regions, and thereby improve the model's recognition rate. The SimAM module is a three-dimensional, parameter-free attention mechanism module. As shown in Figure 2, based on channel attention mechanisms and spatial attention mechanisms, it establishes a three-dimensional attention mechanism, assigning weights to each neuron without increasing additional parameters.

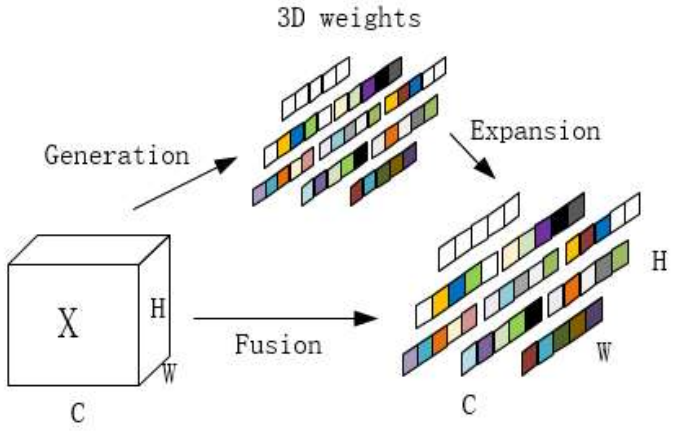

Figure 2. SimAM Attention Mechanism

In neuroscience research, activating a certain neuron may inhibit surrounding neurons. Inspired by this, the SimAM attention mechanism establishes the linear separability between neurons through an energy function. The minimum energy function is shown in equation (1). The smaller the energy value e_t^*, the greater the difference between the target neuron t and the surrounding neurons, and the higher the importance of that neuron. Equation (2) is the formula for enhancing feature processing using the SimAM attention mechanism based on equation (1).

$$e_t^* = \frac{4(\hat{\sigma}^2 + \lambda)}{(t - \hat{u})^2 + 2\hat{\sigma}^2 + 2\lambda} \tag{1}$$

$$\tilde{X} = \text{sigmoid}\left(\frac{1}{E}\right) \odot X \tag{2}$$

In equation (1), $\hat{u} = \frac{1}{M}\sum_{i=1}^{M} X_i$, $\hat{\sigma}^2 = \frac{1}{M}\sum_{i=1}^{M}(X_i - \hat{u})^2$.Here, t is the target neuron of the input feature x, X_i represents the neuron at the i-th spatial location, and M = H × W is the number of neurons per channel. In equation (2), E is the energy map composed of the energy values at all positions (different channels and spatial locations), and ⊙represents element-wise multiplication.

2.3 Monocular Distance Measurement Method

This article incorporates monocular distance measurement in the process of detecting target obstacles. In addition to being able to detect target obstacles, it can also determine the distance between the visually impaired person and the target obstacle during walking. The monocular distance measurement method calculates the distance based on formula conversion, with the specific formula shown in Equation (3). By adding the monocular distance measurement method to the process of

ISSN: 2414-1895 DOI: 10.6919/ICJE.202605_12(5).0013

detecting and recognizing obstacles, it can simultaneously provide the distance to the target obstacle. Determining the distance to obstacles can create a perceptible and safe environment for visually impaired individuals, allowing them to take timely actions and adjust their movement trajectory according to the distance, thereby ensuring their travel safety. This approach can improve their frequency of travel.

$$\mathbf{D} = (\mathbf{F} * \mathbf{W})/\mathbf{P} \tag{3}$$

In equation (3), D represents the distance from the target to the camera, F is the camera focal length, W is the width or height of the target (based on the height of a typical target), and P refers to the number of pixels the target occupies in the image.

3. Experimental Environment and Dataset

3.1 Experimental Environment

The environment was set up and the lightweight model YOLOv7-Tiny, as well as the improved YOLOv7-Tiny, were constructed for experiments. During the experiments, different models were trained under the same environment and parameter configuration. The configuration of the model training environment is shown in Table 1.

Table 1. Environmental Configuration for Model Training

Environment Configuration	
CPU	12th Gen Intel(R) Core(TM) i5-12600KF 3.70GHz
GPU	NVIDIA GeForce RTX 4060 Ti
Python	3.8.9
Torch	2.4.1+cu121

3.2 Dataset Introduction

The dataset is a self-constructed dataset, collected by using a camera to photograph obstacles on blind paths in real-life scenes. A total of 413 images were collected, each containing one or more target objects. The dataset is divided into training, validation, and test sets according to a certain ratio. Examples of the images are shown in Figure 3. The target obstacles are categorized into four classes: car, bicycle, e-bike, and other. Annotation tools were used to label the collected image dataset, saved as .xml files, and a VOC dataset was created.

Figure 3. Dataset Display

ISSN: 2414-1895 DOI: 10.6919/ICJE.202605_12(5).0013

4. Analysis of Experimental Results

4.1 Experimental Evaluation Metrics

The model evaluation metrics used in this experiment include Precision (P), Recall (R), and mean Average Precision (mAP). The formulas for these evaluation metrics are shown in Equations (4), (5), and (7). mAP@0.5 refers to the mAP value when the IOU (Intersection over Union) threshold is 0.5.

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \tag{4}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}} \tag{5}$$

$$\text{AP} = \int_0^1 \text{P(r)d}_\text{r} \tag{6}$$

$$\text{mAP} = \frac{\text{sum(AP)}}{\text{classes}} \tag{7}$$

4.2 Comparison before and after Model Improvement

The model with the added SimAM module after improvement was trained on the same training set under the same environment and parameters as the original YOLOv7-Tiny model. After training, the generated weight file best.pt was used to test the test set. The P-R curves before and after the improvement are shown in Figure 4, and the test results are shown in Table 2. From Table 2, it can be seen that both the precision and recall of the model improved to some extent after the improvement. The mAP@0.5 value of the improved model tested was 2% higher than the original, and at the same time, the detection time per image did not increase, ensuring detection speed while improving accuracy.

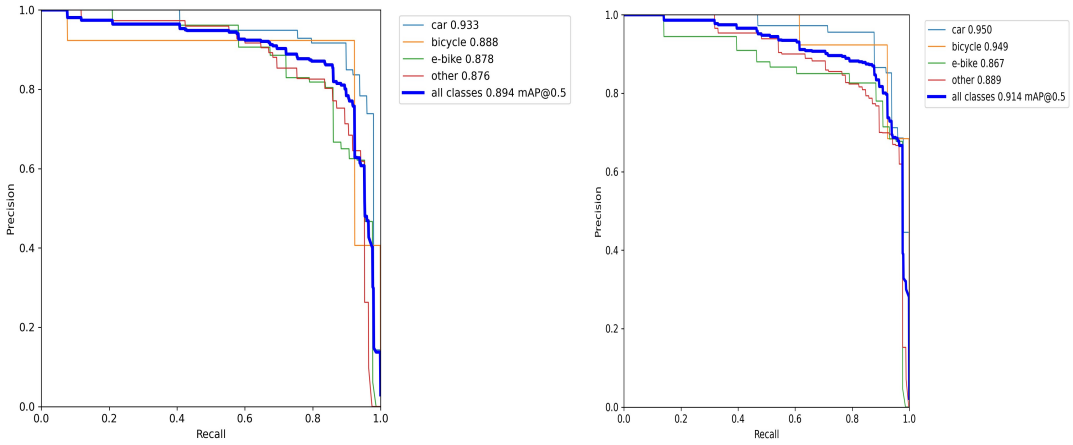

Figure 4. P-R curves before and after improvement

Table 2. Comparison results of the model before and after improvement

Model	P(%)	R(%)	mAP@0.5(%)	Single image detection time(ms)
YOLOv7-Tiny	83.7	87	89.4	7.9
YOLOv7-Tiny+SimAm	83.9	90.4	91.4	7.1

ISSN: 2414-1895 DOI: 10.6919/ICJE.202605_12(5).0013

4.3 Comparison of Experiments with Different Models

Different models, including YOLOv7-Tiny, YOLOv8n, YOLOv7-Tiny-silu, and the improved YOLOv7-Tiny model, were trained under the same environment and parameter settings. Similarly, after training, the best weight file best.pt from the training results was used to test the same test set. The comparison of the test results for different models is shown in Table 3. From the test results in Table 3, it can be seen that the YOLOv7-Tiny-silu model achieved a mAP@0.5 of 90.5% and a total processing time of 8.2 ms per image. Its recognition accuracy is slightly higher than that of YOLOv7-Tiny, but its detection speed is lower than YOLOv7-Tiny. Training was also conducted on the YOLOv8n model and tested on the same test set. The results show that the mAP@0.5 value is 86.5%, and the detection time per image is 8.3 ms. Its recognition accuracy is lower than YOLOv7-Tiny, and its detection speed is also slower compared to YOLOv7-Tiny. The YOLOv7-Tiny model with the added SimAm module is an improvement based on YOLOv7-Tiny. For the obstacle dataset for visually impaired walking used in this study, compared to the original model, the mAP@0.5 value increased by 2%, and the detection time per image was reduced by 0.8 ms. Compared with other models, YOLOv7-Tiny-silu and YOLOv8n, the improved model achieves better recognition accuracy and image detection speed to a certain extent. The experimental data indicate the effectiveness and real-time performance of the improvements.

Table 3. Comparison Results of Experiments with Different Models

Model	P(%)	R(%)	mAP@0.5(%)	Single image detection tim(ms)
YOLOv7-Tiny-silu	85	84.1	90.5	8.2
YOLOv8n	83.5	81.1	86.5	8.3
YOLOv7-Tiny	83.7	87	89.4	7.9
YOLOv7-Tiny+SimAm	83.9	90.4	91.4	7.1

4.4 Adding Monocular Distance Measurement Detection Results

Based on the formula, the code for the monocular distance measurement algorithm was established. Using the improved and better-performing model for detection, the monocular distance measurement algorithm was added to the object detection code, and a test was conducted on a specific test image. Through experiments, the test results are shown in Figure 5. From the figure, it can be seen that the detection results not only include the labeled name of the target obstacle - e-bike, and the recognition accuracy, but also the distance between the person on the road and the target obstacle, with an error range within 0.5 meters.

Figure 5. Detection results of adding a monocular ranging algorithm

ISSN: 2414-1895 DOI: 10.6919/ICJE.202605_12(5).0013

5. Conclusion

This paper proposes an improved obstacle detection method based on YOLOv7-Tiny. It uses a camera to capture datasets of obstacles on sidewalks and at intersections in daily life, and analyzes the images in the dataset. Considering the complex backgrounds of the captured images and the low accuracy of obstacle recognition under different weather and lighting conditions, the lightweight network YOLOv7-Tiny is improved by integrating the parameter-free attention mechanism SimAM module to enhance feature extraction and improve recognition accuracy. Additionally, a monocular distance measurement method is added during detection to measure the distance between people and obstacles. The addition of this monocular distance measurement method can help visually impaired individuals perceive their surrounding environment. Compared with the original model, the improved model increases the mAP@0.5 value by 2 percentage points and has faster detection speed. Compared with other networks such as YOLOv7-Tiny-silu and YOLOv8n, it shows better recognition accuracy and detection speed, providing significant assistance for the walking safety of visually impaired individuals.

Although the effectiveness of the model improvements has been verified through experiments, in daily life, the larger the city, the more complex the environment and road conditions become. The traveling range of visually impaired people will be relatively wider, and the types of obstacles will also increase. Therefore, in the future, further research is needed. In terms of datasets, the scope of data collection can be further expanded, increasing the types and quantities of datasets under different scenarios, lighting conditions, and road conditions on tactile paths and intersections. In terms of model optimization, lightweight models can continue to be optimized, further enhancing the model's feature extraction ability in complex road scenarios for visually impaired walkers, improving the model's recognition accuracy and stability. At the same time, distance measurement algorithms should be optimized, and these algorithms should be integrated with sensor technology for positioning, improving measurement accuracy, reducing the impact of environmental complexity on distance measurement results, and establishing more stable detection models and distance measurement methods.

Acknowledgments

Anhui Xinhua University School-Level Quality Engineering Project (2024jy036); Anhui Xinhua University School-Level Scientific Research Projects (2024zr018, 2025zrzdi03); Key Project of Natural Science Research by Anhui Provincial Department of Education (2025AHGXZK30196); 2024 Anhui Province College Students' Innovation Training Program Projects (S202412216202, S202412216225).

References

[1] Lecun Y., Bottou L.. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11):2278-2324..

[2] Russakovsky O., Deng J., Su H.,et al. Imagenet large scale visual recognition challenge[A]. International Journal of Conflict and Violence[C]. 2015:211–252.

[3] SIMONYAN K, ZISSERMAN A.Very deep convolutional networks for large-scale image recognition[C]. Proceedings of the International Conference on Learning Representations, 2015.

[4] Huang G , Liu Z , Maaten L V D ,et al.Densely Connected Convolutional Networks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2017:4700-4708.

[5] HE K, ZHANG X, REN S,et al ,Deep residual learning for image recognition[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016:770–778.

[6] He Yu. Research on Texture Surface Defect Classification and Detection Methods under Limited Sample Conditions [D]. Dongbei: Northeastern University, 2020.

[7] Dong Li. Automatic Detection System for Lung Nodules in CT Images Based on Deep Learning [D]. Shanghai: East China Normal University, 2022.

ISSN: 2414-1895 DOI: 10.6919/ICJE.202605_12(5).0013

[8] Zhang Mengzhen. Research on the Application of Deep Learning-Based Face Recognition Technology in Smart Home [J]. Information and Computers, 2025, 37(13):5-7.

[9] Cheng P , Tang X , Liang W ,et al.Tiny-YOLOv7: Tiny Object Detection Model forDrone Imagery[C]. International Conference on Image and Graphics. Cham: Springer Nature Switzerland, 2023:53-65.

[10] YANG L, ZHANG R Y, LI L, et al. Simam: a simple, parameter-free attention module for convolu-tional neural networks[C]. Proceedings of the Inter-national Conference on Machine Learning, 2021:11863-11874.

[11] Liu S , Qi L , Qin H ,et al.Path Aggregation Network for Instance Segmentation[J]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2018:8759-8768.

[12] Lin T. Y., Dollár P., Girshick R., et al. Feature pyramid networks for object detection[A]. Conference on computer vision and pattern recognition[C]. IEEE, 2017:2117-2125.