Research on Vehicle Object Detection Algorithm Based on Deep Learning

Zhengkun Shen, Huaixian Yin

Abstract— **Aiming at the low detection efficiency and accuracy of existing deep learning vehicle target detection algorithms, a VOD-YOLOv5 vehicle target detection algorithm was proposed based on the YOLOv5 model. In this paper, a new lightweight convolutional neural module (cf) is proposed to improve the feature extraction capability of the network. At the same time, the attention mechanism of space and channel fusion is integrated into the network model, which improves the detection accuracy of small and medium-sized targets in fuzzy images. The experimental results show that compared with the original YOLOv5 model, the VOD-YOLOv5 model proposed in this paper has a 4% increase in average accuracy (mAP), and the average accuracy (AP) of detection of different target classes has been improved, and the detection speed meets the real-time requirements, effectively improving the detection performance of the vehicle target detection model.**

Index Terms—**deep learning; YOLOv5; object detection**

I. INTRODUCTION

With the rapid development of computers in the automotive field, automatic driving has attracted wide attention. At present, the realization of automatic driving is divided into three steps: perception, decision [1], and execution. Target detection is one of the most important components of autonomous driving. Vehicle target detection is an important module of front-end data acquisition, perfect target detection can quickly and accurately complete the vehicle detection and can provide effective data guarantee for the realization of the entire automatic driving. Therefore, it is of great significance to design a vehicle target detection algorithm with excellent robustness and real-time performance

In recent years, deep learning technology has developed rapidly, and object detection algorithms based on deep learning have also developed rapidly. Compared with traditional object detection algorithms, the object detection algorithm based on deep learning has strong generalization ability and robustness, and can meet the real-time detection of vehicle targets. Object detection algorithms based on deep learning are divided into two categories: first, one-stage detection algorithm. Typical networks include YOLO[2] and SSD[3]. Second, two-stage detection algorithm. Representative networks include R-CNN[4], Faster R-CNN[5], Mask R-CNN[6], and Spp[7]. The one-stage detection algorithm does not produce candidate boxes, and directly transforms the problem of target border positioning into a regression problem, so the speed is relatively fast and can meet the needs of real-time detection. The two-stage detection algorithm first generates a series of sample candidate boxes by the algorithm, and then classiifies the samples through the convolutional neural network, so the detection accuracy and positioning accuracy are high.

In recent years, researchers have continued in-depth research on the basis of the original research. Pang et al. [8] proposed the Libra R-CNN network, which uses an effective balance learning framework for target detection, and the detection performance is improved compared with the detection accuracy of Faster-RCNN. Luo et al. [9] proposed an image adaptive correction algorithm based on Faster-RCNN and combined NAS optimization and feature enrichment to extract multi-layer features more effectively. The target feature rich set combined with multi-layer feature information and cross-layer connection was used to enrich vehicle target information, which improved the robustness of the model. Jamiya et al. [10] proposed the Little YOLO-SPP model based on the YOLOv3-Tiny network. The paper mainly modified the feature extraction network of the tiny target detection network of the YOLOv3 model and added the space pyramid pool to realize the tiny target detection and improve the detection speed and accuracy. Based on YOLOv4, Cai et al. [11] proposed a new feature fusion module pa++ and adopted five scale detection layers to improve the detection accuracy of small targets. In addition, an optimized network pruning algorithm was proposed to improve the computational efficiency, and the detection accuracy and reasoning speed of the improved model were greatly improved compared with the original YOLOv4.

In this paper, the YOLOv5 algorithm is used as the base model to make further changes and improvements. (1) A new lightweight convolutional neural module cf is proposed to replace the c3 module in the original network to improve the feature extraction capability of the network; (2) The CA attention mechanism is added to enhance the expression ability of mobile network learning features and improve the detection ability of small and medium-sized targets in fuzzy images.

II . YOLOv5 NETWORK STRUCTURE

The YOLO algorithm divides the image into N×N grid cells of the same size, each of which is responsible for the detection and positioning of the object containing the grid. Moreover, each grid should predict B bounding boxes, and each bounding box should not only return its own position (x, y, w, h), but also predict a confidence value, a total of 5 values. At present, YOLOv5 includes YOLOv5s, YOLOv5m,

Manuscript received November 12, 2023

Shen Zhengkun, male, is a master student majoring in Vehicle Engineering at Qingdao University. His research interest is deep learning object detection.

Yin Huaixian, female, doctor, senior experimentalist, master tutor, main research interests are automotive reliability research, electric vehicle intelligent power integration technology.

YOLOv5l and other models, which increase successively in network width and depth.

The network structure of YOLOv5 consists of four parts, namely, input layer, backbone network, neck network and prediction layer. As shown in Figure 1.

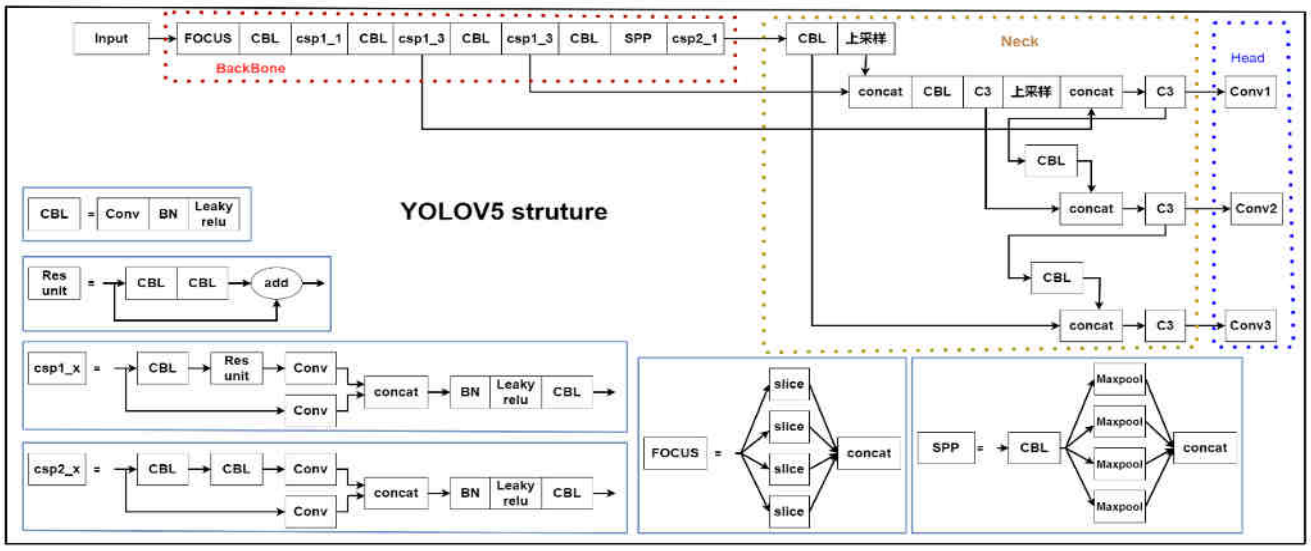

Fig 1. Structure diagram of YOLOv5

In the input layer, the input image size is 640×640, the number of channels is 3, and the input image is mainly used for adaptive anchor frame calculation, adaptive picture scaling, and Mosaic[12] data enhancement. Through the above operations, the detection effect of small targets is improved, and the training speed and detection ability of the model are also improved.

The backbone network consists of four parts: Focus module, SPP (space pyramid pool) module, CBL and CSP[13]. The Focus module mainly performs slicing and convolution operations. After slicing operations, the number of channels becomes 4 times, and the memory used for parameter calculation and cuda is also reduced. The CSP structure divides the input into two branches and performs convolution operations to halve the number of channels, then performs Bottleneck*N operations on one of the branches, and concat the last two branches, which bottleneck and output sizes do not change, but allow the model to learn more features. In YOLOv5, CSP has two designs, namely CSP1_X structure and CSP2_X structure. SPP, also known as space pyramid Pool, parallel input through multiple Max pools of different sizes, and then further fusion, in order to solve the problem of multi-scale targets.

The neck network structure mainly realizes the fusion of shallow graphic features and deep semantic features. It mainly includes top-down FPN[14] module and bottom-up PAN[15] module. FPN module transfers deep semantic features to shallow layer to enhance semantic information. The PAN module complements the FPN by adding a bottom-up pyramid structure behind the FPN that transmits the strong positioning features of the shallow layer to the deep layer.

In the output layer, three detection heads are used to downsample the original image in different multiples to generate three feature vectors of different sizes. It is used to output the position, size and category of the object in the image.

The loss function of YOLOv5 prediction layer is divided into three parts: Classes loss (cls), Objectness loss (obj), and Location loss (loc), which correspond to the loss of object classification, confidence and detection frame respectively. Among them, BCE loss is adopted for both cls loss and obj loss, and CIou loss is adopted for box loss[16]. Therefore, the total loss function of YOLOv5 is:

$$Loss = \lambda_1 L_{cls} + \lambda_2 L_{obj} + \lambda_3 L_{loc} \#(1)$$

Where, λ_1, λ_2 and λ_3 are equilibrium coefficients.

III. IMPROVEMENTS TO YOLOv5

A. *Proposed Cf module*

In order to improve the overall detection performance of the network, this paper proposes the Cf module as shown in Figure 2 on the basis of C3 module in the original YOLOv5s network model.

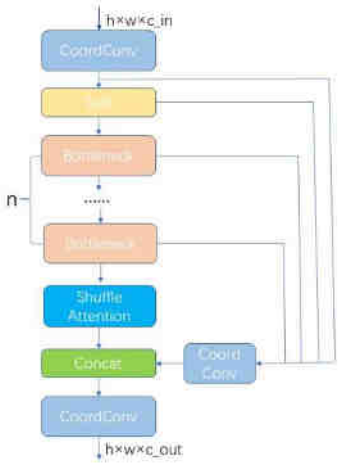

Fig2. Cf structure diagram

The Cf module in the original YOLOv5 network mainly relies on the idea of CSPNet to extract shitter, and designs C3 Block based on the idea of residual structure. The main

branch gradient module of CSP is a BottleNeck module, that is, the residual module. At the same time, the number of stacks is controlled by the parameter n, which means that the value of n is different for models of different sizes. The addition of Cf module can not only ensure the lightweight of the model, but also obtain more abundant gradient flow information.

B. Attention-increasing mechanism

In neural network learning, generally speaking, the more parameters of the model, the stronger the expression ability of the model and the more information stored by the model, but this will also bring the problem of information overload. Therefore, how to focus on the more critical information in the current task and reduce the attention to other information is a problem worthy of further investigation. Moreover, in the field of image recognition, the global information of an image is the key to computer recognition of images. However, when the YOLOv5 network performs convolution operations on input images, each convolution kernel can only correspond to a local receiving region, and only the context information of the local receptive field is considered, without considering the correlation between the global context information, which easily leads to the loss of target features. To sum up the above problems, combined with the actual situation of this paper, the introduction of attention mechanism module.

The coordinate attention encodes channel relationships and long-range dependencies through precise location information in two steps: coordinate information embedding (coordinate information embedding) and coordinate attention generation (coordinate attention generation), its specific structure is as follows.

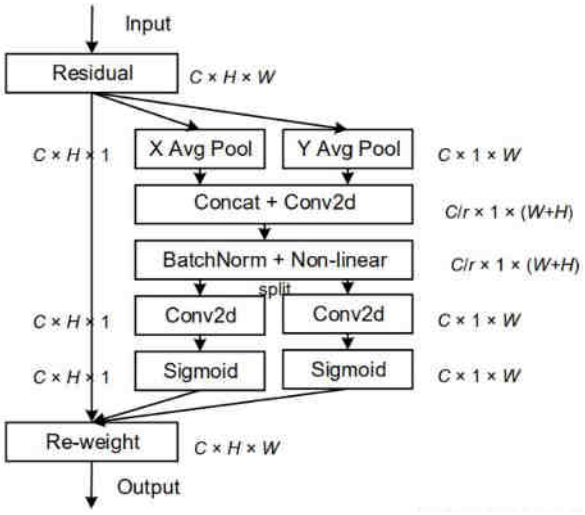

Figure 3. Structure of coordinate attention

First, looking at the coordinate embedding part, global pooling is often used in channel attention to globally encode spatial information as channel descriptors, making it difficult to store location information. In order to facilitate the attention module to capture spatial long range dependencies with accurate location information, global pooling is decomposed into a pair of one-dimensional feature coding operations. Specifically, for input X, each channel is encoded along the horizontal and vertical directions using pooled cores of dimensions (H,1) and (1, W), and a pair of direction-aware attention graphs are returned. This transformation allows the attention module to capture long-range dependencies along one spatial direction and store precise information along another spatial direction, which helps the network more accurately locate the target of interest.

In order to make better use of the coordinate information embedding module to generate a representation with global sensitivity field and accurate position information, coordinate attention generation is designed. First, two feature graphs generated by the previous module are cascades, and then a shared 1*1 convolution is used to transform F1. The generated f is the intermediate feature graph of spatial information in the horizontal and vertical directions. Used to control module size. Then, f is cut into two separate tensors fh and fw along the spatial dimension, and two 1*1 convolution Fh and Fw are used to transform the feature graph fh and fw to the same number of channels as the input X. The following results are obtained.

$$g^h = \sigma\left(F_h\left(f^h\right)\right)\#(2)$$

$$g^w = \sigma\left(F_w\left(f^w\right)\right)\#(3)$$

Then gh and gw are extended, and as attention weights, the final output of the CA attention mechanism can be expressed as follows.

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j)\#(4)$$

At this point, the CA attention mechanism completes both horizontal and vertical attention, and it is also a channel attention.

IV. EXPERIMENTAL PROCESS AND ANALYSIS

A. *Experimental data set*

The UA-DETRAC Vehicle detection dataset [17] is 10 hours of video captured using Canon EOS 550D cameras at 24 different locations in Beijing and Tianjin, China, recorded at 25 frames per second (fps) with a resolution of 960×540 pixels. The dataset has more than 140,000 frames and 8,250 manually labeled vehicles, with a total of 1.21 million labeled object bounding boxes, and the labels of the dataset are subdivided into four categories: car, van, bus, and others. The data set used in this paper is part of the UA-DETRAC vehicle detection data set, with a total of 10,870 data images, which are divided into training set, test set and verification set according to the ratio of 7:2:1.

In order to make the model have higher detection accuracy and improve the generalization ability and robustness of the model. Data is normalized using Linear normalization, also known as Linear normalization or "max-min" normalization, which allows the difference between the maximum and minimum x values in the data to be analyzed and a cardinality to be established. Linear normalization formula:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}\#(5)$$

This method has the advantage of being easier to realize and more flexible, so that the data can reach the effect of conforming to the standard normal distribution, so as to effectively avoid the problems such as the slow convergence speed and the decrease of detection accuracy caused by too large data differences.

The K-means clustering method is used to update network anchors to avoid anchors mismatch due to changes in the data set. The training sample is increased by adjusting the hue, saturation and exposure of the image data.

B. Experimental environment and evaluation index

The hardware configuration and software version of the experiment in this paper are shown in Table 1. The environment configuration of all experiments conducted in this paper did not change during the whole process. During the model training process, the Stochastic Gradient Descent (SGD) method is used for model training, and the network hyperparameters set are shown in Table 2.

Table 2. List of hyperparameters

Training parameter name	Parameter value
Initial learning rate	0.01
Learning rate decline parameter	0.0001
Weight attenuation coefficient	0.0005
Momentum	0.937
Batch size	16
Epoch	200
Image input size	640×640

In the experiment, R (Recall), P (Precision), AP (average precision), mAP (average precision rate) of each category and Frames per second (FPS) were used as evaluation indexes to evaluate the real-time performance of the model. The formula for calculating the recall rate R and accuracy rate P is as follows.

$$recall = \frac{TP}{TP + FN} \#(6)$$

$$precision = \frac{TP}{TP + FP} \#(7)$$

Among them, TP (True Positives) refers to the number of targets correctly identified by positive sample data, namely vehicle targets correctly identified in this paper; FN (False Negatives) is the number of targets that are not correctly identified by positive sample data, namely, vehicle targets that are missed in this paper. FP (False Positives) is the number of targets that identify non-positive sample data as positive samples, that is, false vehicle targets.

Average accuracy AP and the average accuracy mAP of each category can be used to measure the pros and cons of the network model, which should be calculated according to the recall rate and accuracy of the model. The calculation formula is as follows.

$$AP = \int_0^1 P(R)dR \#(8)$$

$$mAP = \frac{1}{N}\sum_{i=1}^{N} AP_i \#(9)$$

C. Experimental results and comparative analysis

In order to explore the effect of attention mechanism on model performance and the addition of different attention mechanisms on model performance. In this paper, through comparative experiments, two frequently used Attention mechanisms (SE attention mechanism and CBAM attention mechanism) are added to the same position as Coordinate Attention(CA for short). The models added to the attention mechanism were named YOLOv5_SE, YOLOv5_CBAM, and YOLOv5_CA. After 200 rounds of iterative training, the model performance comparison of different attention mechanisms is shown in Table 4.

Model	Average precision	Weight
YOLOv5	72.62	29.3
YOLOv5_SE	72.93	29.5
YOLOv5_CBAM	73.59	30.6
YOLOv5_CA	73.65	29.9

Through the analysis of Table 4, it can be seen that the detection accuracy of the network model with added attention mechanism is improved compared with that of the network model without added attention mechanism, and the average accuracy of the three models with added attention mechanism is 63.36%. Compared with the original model, the average accuracy of model detection with SE attention mechanism and CBAM attention mechanism increased by 0.31% and 0.97%. The CA attention mechanism added in this paper has the most obvious improvement on the detection accuracy, increasing by 1.03%. Compared with the original model, the weights after the addition of attention mechanism are increased, but the increase is small, and the impact on network performance is small. The addition of CBAM attention mechanism also significantly improved the average accuracy of model detection, but the weight increased more, and the effect was better without SA attention mechanism. The SE attention mechanism may only take into account the importance of channel pixels and lack of location information in the channel due to the reason that the SE attention mechanism only has channel attention, so the effect achieved after the addition is not good enough. Therefore, after comprehensive analysis, this paper decides to add CA attention mechanism to YOLOv5 model.

In order to verify the influence of different improvement modules on the network performance, this paper conducts a separate experiment on the improvement of each module, and finally integrates all the improvements before the experiment. The experimental results are shown in Table 3. The I-YOLOV5 is the result of adding the CA attention mechanism, and only the neck part of the network is added. The detection accuracy of the model for small targets in fuzzy images is improved, and the mAP value of the model is increased by 0.59% compared with that of the YOLOv5s network. Ⅱ-YOLOv5 is the experimental result after adding cf on the basis of YOLOv5. The mAP value increases by 1%, and the network's perception ability of spatial information is improved.

Table 3. Comparison of experimental results

Network	mAP/%	Inference time /s
YOLOv5s	0.7262	0.012
Ⅰ-YOLOv5s	0.7365	0.015
Ⅱ-YOLOv5s	0.7421	0.014
ours	0.7604	0.017

It can be seen from the experimental data in Table 3 that with the gradual increase of improvement modules in the network, the value of mAP will gradually increase. Although the inference time has a small increase, it still meets the requirements of real-time detection.

D. YOLOv5 compared to our model

The comparison of the network performance of the two network models and the average accuracy of each category are shown in Table 4 and Table 5:

Table 4. Network performance comparison

Network model	precision	recall	mAP	Inference time/s
YOLOv5	0.7235	0.6197	0.7262	0.012
Ours	0.7563	0.6439	0.7604	0.015

Table 5. Average precision comparison

Target class	YOLOv5	VOD-YOLOv5
car	0.798	0.814
bus	0.841	0.867
van	0.571	0.592

According to the analysis of Table 4, the mean accuracy (mAP) of our model reaches 76.04%, which is 3.42% higher than that of the YOLOv5 model. Recall rate increased by 2.42%; Accuracy increased by 3.28%; The inference time is increased by 0.003 seconds, but it still meets the real-time requirement. Table 5 is a comparison

table of the average accuracy of the three detection categories (car, bus, van) after the detection categories others are excluded. The AP values of our model on the three targets have been improved by 1.6%, 2.6%, 2.1% respectively.

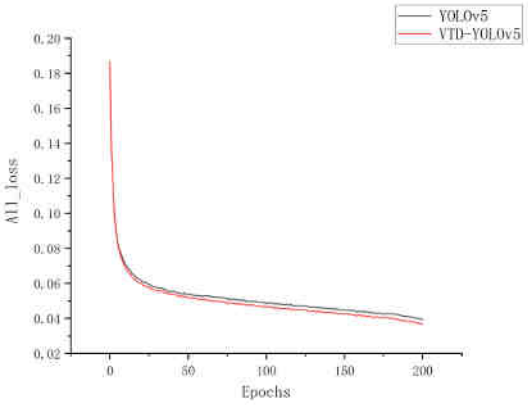

Figure 7. Total loss comparison chart

According to the comparison diagram of total loss in FIG. 7, the total loss shows a decreasing trend. With the increase of iterations, the decreasing trend becomes slower and becomes stable after 50 iterations, and the loss value of our model is lower than that of YOLOv5 network on the whole.

Ⅴ.CONCLUSION

This paper makes improvements on the basis of YOLOv5s model. Compared with the original VOD-YOLOv5 model, a new lightweight convolutional neural module (cf) is proposed, which improves the feature extraction capability of the network. Integrate Coordinate Attention to improve the detection accuracy of small and medium-sized targets in fuzzy images. Experiments on UA-DETRAC vehicle detection data set show that compared with YOLOv5s model mAP, our proposed model has improved, and the loss has decreased to a certain extent. Besides, the detection speed meets the real-time requirements, and it has strong robustness and generalization ability, which can effectively complete the detection of vehicles. Compared with the YOLOv5s model, the overall detection performance of our improved YOLOv5 model has been improved.

REFERENCES

[1] TASAKI T, IEEE. Perception and Decision Making for the Autonomous Driving System; proceedings of the International Symposium on Micro-NanoMechatronics and Human Science (MHS) / Symposium on Understanding Brain Plasticity on Body Representations to Promote their Adaptive Functions - Grant-in-Aid for Scientific Research on Innovative Areas, Nagoya, JAPAN, F 2018，Dec 09-12, 2018 [C]. 2018.

[2] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection; proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, F 2016，Jun 27-30, 2016 [C]. 2016.

[3] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector; proceedings of the 14th European Conference on Computer Vision (ECCV), Amsterdam, NETHERLANDS, F 2016，Oct 08-16, 2016 [C]. 2016.

[4] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation; proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, F 2014，Jun 23-28, 2014 [C]. 2014.

[5] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks; proceedings of the 29th Annual Conference on Neural Information Processing Systems (NIPS), Montreal, CANADA, F 2015，Dec 07-12, 2015 [C]. 2015.

[6] HE K, GKIOXARI G, DOLLAR P, et al. Mask R-CNN [J]. Ieee Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 386-397.

[7] HE K, ZHANG X, REN S, et al. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition; proceedings of the 13th European Conference on Computer Vision (ECCV), Zurich, SWITZERLAND, F 2014，Sep 06-12, 2014 [C]. 2014.

[8] PANG J M, CHEN K, SHI J P, et al. Libra R-CNN: Towards Balanced Learning for Object Detection; proceedings of the 32nd IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, F Jun 16-20, 2019 [C]. 2019.

[9] LUO J-Q, FANG H-S, SHAO F-M, et al. Multi-scale traffic vehicle detection based on faster R-CNN with NAS optimization and feature enrichment [J]. Defence Technology, 2021, 17(4): 1542-1554.

[10] JAMIYA S S, RANI P E. LittleYOLO-SPP: A delicate real-time vehicle detection algorithm [J]. Optik, 2021, 225.

[11] CAI Y F, LUAN T Y, GAO H B, et al. YOLOv4-5D: An Effective and Efficient Object Detector for Autonomous Driving [J]. Ieee Transactions on Instrumentation and Measurement, 2021, 70.

[12] YUN S, HAN D, OH S J, et al. CutMix: Regularization Strategy to Train Strong Classifiers with Localizable Features; proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, SOUTH KOREA, F 2019，Oct 27-Nov 02, 2019 [C]. 2019.

[13] SCHMIDHUBER J. Deep learning in neural networks: An overview [J]. Neural Networks, 2015, 61: 85-117.

[14] LIN T-Y, DOLLAR P, GIRSHICK R, et al. Feature Pyramid

Networks for Object Detection; proceedings of the 30th IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, F 2017，Jul 21-26, 2017 [C]. 2017.

[15] LIU S, QI L, QIN H, et al. Path Aggregation Network for Instance Segmentation; proceedings of the 31st IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, F 2018，Jun 18-23, 2018 [C]. 2018.

[16] YU J, JIANG Y, WANG Z, et al. UnitBox: An Advanced Object Detection Network [J]. ACM, 2016.

[17] WEN L, DU D, CAI Z, et al. UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking [J]. Computer vision and image understanding: CVIU, 2020, 193.

[18] BOTTOU L. Stochastic Gradient Descent Tricks BT - Neural Networks: Tricks of the Trade [J]. 2012.