Road Object Detection of YOLO Algorithm with Attention Mechanism

Jiacheng Li, Huazhang Wang*, Yuan Xu, Fan Liu

College of Electrical Engineering, Southwest Minzu University, Chengdu, China
Email: 180852082013@stu.swun.edu.cn

Abstract. In auto-driving cars, incorrect object detection can lead to serious accidents, so high-precision object detection is the key to automatic driving. This paper improves on the YOLOv3 object detection algorithm, and introduces the channel attention mechanism and spatial attention mechanism into the feature extraction network, which is used to autonomously learn the weight of each channel, enhance key features, and suppress redundant features. Experimental results show that the detection effect of the improved network algorithm is significantly higher than that of the YOLOv3 algorithm.

Keywords: autonomous driving, object detection, YOLOv3, channel attention mechanism, spatial attention mechanism

1 Introduction

Vehicle object detection is currently a focus of intelligent transportation research. Object detection of vehicles, pedestrians, etc. is an important task in the automatic driving system, which requires high accuracy of object detection and sufficient real-time performance.

The task of object detection is to find all the objects of interest in the image and determine their positions. Object detection algorithms based on convolutional neural networks can be divided into two categories: the first category is a two-stage object detection algorithm, represented by the Faster R-CNN [1] series, which generates a series of candidate frames as samples, and then classify the samples through the convolutional neural network, the performance is superior in detection accuracy and positioning accuracy; the second type is a one-stage object detection algorithm, based on the SSD[2] series and YOLO[3- 5]. The series is representative, without generating candidate frames, the problem of positioning the object frame is directly converted into a regression problem, and the algorithm speed is superior.

From the driver's perspective, the road ahead is complex and changeable, with many types of objects, large and small. The object in the long-distance image area is usually small in size and features not obvious, and it is difficult to accurately detect and locate. The YOLOv3 object detection algorithm has fast detection speed, strong comprehensive performance, and the ability to detect small objects. However, due to its insufficient detection ability for multiple long-distance small objects, missed detection is easy to occur.

In the field of computer vision, the attention mechanism is an application that simulates the human visual observation mechanism. For example, the human visual system tends to focus on effective information that can assist judgment, and ignores unimportant information. There are also a lot of invalid features in the process of neural network learning. The attention mechanism module can help network training tend to train with effective features. SENet[6] (Squeeze-and-Excitation Networks) models the correlation between feature channels and strengthens important features to improve accuracy; CBAM[7] (Convolutional Block Attention Module) is embeddable, Lightweight general convolutional attention mechanism module. This module is different from the SEnet module that only considers the channel attention mechanism. It proposes an attention mechanism structure that combines the two dimensions of channel and space, which has better performance than the benchmark model. It can be seen that the attention mechanism can improve the performance of computer vision tasks.

2 Improved YOLOv3 Algorithm Related Theory

2.1 Overview of YOLOv3 Algorithm

YOLOv3 uses the Darknet-53 network structure with the fully connected layer removed as the backbone network. It draws on the practice of the Residual network [8] and sets up shortcut connections between some layers. The output layer draws on FPN[9] (feature pyramid networks), and uses multi-scale prediction to detect objects of different sizes. The network structure is shown in Figure 1.

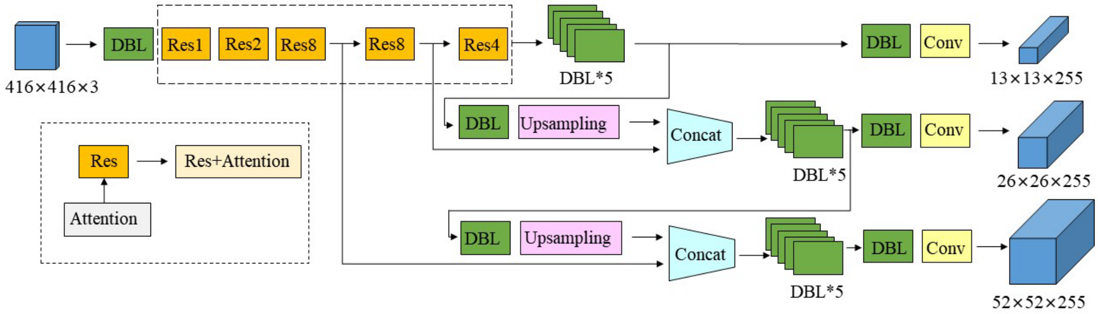


Figure 1. YOLOv3 network structure

The input size of the network is 416×416, and DBL in the backbone network is the basic component of YOLOv3, which is composed of convolution, BN (Batch Normalization)[10] and leaky relu[11] activation functions. Res(n) is the number of residual units in the residual block. Concat is tensor splicing, which splices the Upsampling of the Darknet middle layer and a later layer. Finally, the prediction result is output, including the category, confidence information and adjustment parameters of the prediction frame, and the prediction frame is screened using the confidence threshold and non-maximum value suppression to obtain the final detection result. The dotted box in the figure is a schematic diagram of the introduction of the attention mechanism on the YOLOv3 residual network in this article.

2.2 Improved Network Introducing Channel Attention Mechanism

On the basis of the YOLOv3 network structure, the channel attention mechanism of SEnet is introduced. Figure 2 is the Block unit of SENet. Ftr in the figure is the traditional convolution structure, and X and U are the input and output of Ftr. The added part of SENet is the structure behind U: First do a global average pooling [12] for U, and then output 1×1×C data through two levels of full connection, and finally use the sigmoid function to limit the output to the range of 0 to 1, take this value as the scale multiplied by U on the C channels as the input data of the next stage. The specific steps are as follows:

Squeeze operation: Global average pooling is performed on the feature values of all channels, and each two-dimensional feature channel becomes a real number, namely z_c, which characterizes the global distribution of the response on the feature channel to obtain the global receptive field. The calculation formula of z_c is as follows:

$$z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i,j) \tag{1}$$

where H and W are respectively the height and width of the picture, u represents the feature map obtained after the standard convolution operation of the input image, $u_c$represents the c-th two-dimensional matrix in u, c represents the number of feature channels, $u_c(i,j)$ represents the element in the i-th row and j-th column in the channel feature map matrix u_c.

Excitation operation: By capturing the non-linear interaction relationship between the channels, the weight of each channel is evaluated, so as to rescale the original feature in the channel dimension. The calculation formula for the result of the Excitation operation is as follows:

$$\mathrm{s} = \sigma\big(W_2\delta(W_1 z)\big) \tag{2}$$

where s represents the set of each channel weight s_c, σ represents the ReLU function, and σ represents the Sigmoid activation function. z represents the set of real numbers zc obtained by the Squeeze operation of each channel. The two weights of W_1 and W_2 are obtained through learning. The dimensions of W_1 and W_2 are $\frac{C}{r} \times C$ and $\mathrm{C} \times \frac{C}{r}$ respectively, C is the number of channels, and r is the scaling parameter to reduce the amount of calculation. In order to achieve the balance between propagation speed and detection accuracy, refer to [6], the value of r in this paper is 16.

Reweight operation: The weight of the Excitation output is weighted to the previous feature channel by channel through multiplication, and the original feature recalibration in the channel dimension is completed, thereby enhancing the attention to the key channel domain. The calculation formula of the output after the recalibration of the input feature map u combined with the weight s is as follows:

$$\tilde{\mathrm{x}} = s_c \cdot u_c \tag{3}$$

where s_c represents the corresponding weight of each channel, and u_c is the channel feature matrix corresponding to each feature map.

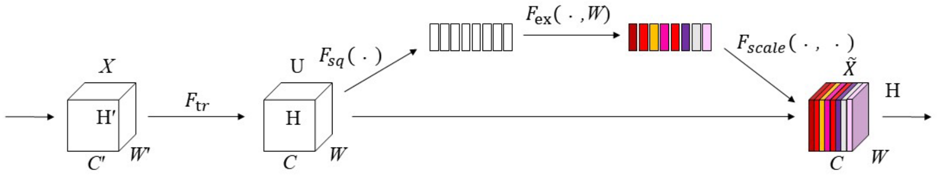


Figure 2. SE Block

The principle of this structure is to enhance the important features and weaken the unimportant features by controlling the size of the scale, thereby making the extracted features more directional. In this paper, by transforming the YOLOv3 backbone network Darknet-53, adding SENet to ResNet for optimization, the optimized structure is shown in Figure 3. Compared with the ResNet residual network, the SE-ResNet attention residual combination structure enhances the nonlinear characteristics of the network and improves the generalization ability of the model without changing the output dimension. The SE-ResNet attention residual combination structure is used as the basic structural unit to realize the information interaction between channels, suppress the influence of useless features on the model, and improve the detection accuracy.

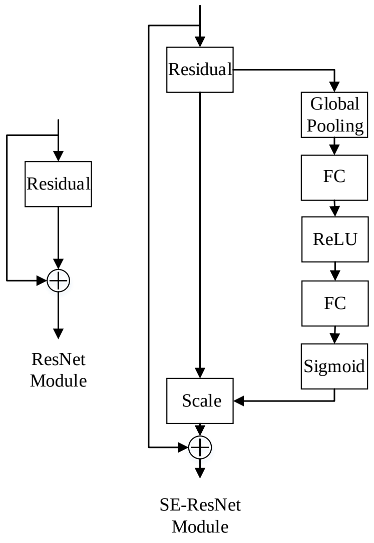


Figure 3. ResNet Module and SE-ResNet Module

2.3 Introducing an Improved Network that Combines Channels and Spatial Attention Mechanisms

In addition to global average pooling, Global Max Pooling can also help in feature selection. It can also make up for the information lost by global average pooling to a certain extent. Therefore, based on the CBAM module, the introduced channel attention mechanism needs to consider two pooling operations at the same time. The calculation formula is as follows:

$$F' = M_c(F) \otimes F \tag{4}$$

where $F \in R^{H\times W\times C}$ is the input feature map, $F' \in R^{H\times W\times C}$ is the feature map after attention enhancement, and H, W, and C respectively represent the feature map The length, width and number of channels, ⊗ means multiplication element by element. $M_c(F)$ represents the attention extraction operation for F in the channel dimension.

The calculation formula of $M_c(F)$ is as follows:

$$\begin{aligned} M_c(F) &= \sigma\Big(MLP\big(AvgPool(F)\big) + MLP\big(MaxPool(F)\big)\Big) \\ &= \sigma(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \end{aligned} \tag{5}$$

where F^c_{avg} and F^c_{max} are respectively one-dimensional vectors obtained by global average pooling and maximum pooling of each channel of the feature map *F*.

Global average pooling has feedback for each feature value on the feature map, which can effectively learn image background information, and maximum pooling can collect prominent feature value information. MLP (Multilayer Perceptron) is a three-layer perceptron network shared by F^c_{avg} and F^c_{max}. W_0 and W_1 are the weights of the three-layer perceptrons, and σ is the Sigmoid activation function. The process is shown in Figure 4.

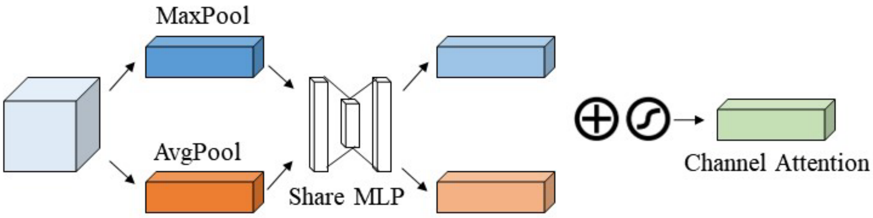


Figure 4. Channel attention mechanism of CBAM

In addition, the spatial relationship of features can also be used for modeling to supplement the positional relationship information that the channel attention mechanism cannot obtain. On this basis, the spatial attention mechanism is further added, and the channel and space attention mechanisms are combined to screen the channel and space feature information at the same time. The calculation formula is as follows:

$$M_S(F) = \sigma\Big(f^{7\times7}\big(AvgPool(F); MaxPool(F)\big)\Big) = \sigma(f^{7\times7}(F^s_{avg}; f^s_{max})) \tag{6}$$

The feature map output by the channel attention mechanism is used as the input feature map of this module. First, a channel-based global maximum pooling and global average pooling are performed, and the two results are concat based on the channel, and then pass a standard The convolutional layer is connected and convolutional mixed to generate a spatial attention map, and then activated by a sigmoid function to generate a feature map of the spatial attention mechanism. The process is shown in Figure 5.

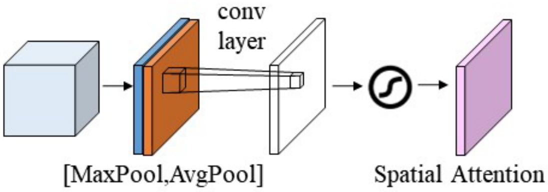


Figure 5. Spatial attention mechanism of CBAM

The CBAM module that combines the channel and spatial attention mechanism is introduced into the residual branch of the ResNet residual network, and the CBAM-ResNet attention residual combined structure can be obtained. The structure is shown in Figure 6.

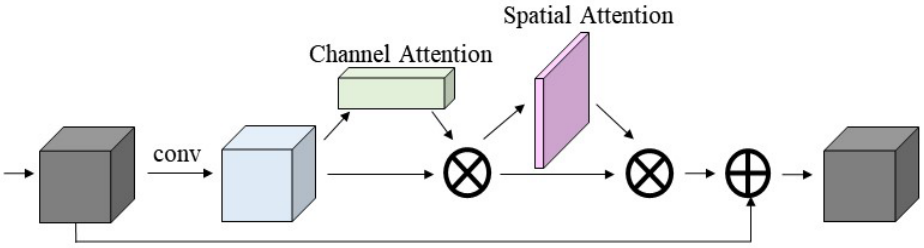


Figure 6. ResNet+CBAM

3 Experimental Results and Analysis

3.1 Experimental Software and Hardware Configuration

The hardware configuration used in the experiment is: CPU is Intel(R) Xeon(R) Silver 4110 CPU @ 2.10GHz, GPU is GeForceRTX2080Ti, video memory is 11GB, and memory size is 132GB. The software system configuration is: the operating system is Ubuntu 16.04, the CUDA version is 11.0, the programming language is Python 3.6, and the deep learning framework uses PyTouch.

3.2 Dataset and Training Process

The original algorithm YOLOv3, the improved algorithm YOLOv3-SE and the improved algorithm YOLOv3-CBAM are compared and tested on the BDD100K[13] dataset. BDD100K was released by the AI Lab of the University of Berkeley. The data set contains 100,000 high-definition videos. The key frames are sampled at the 10th second of each video to obtain 100,000 pictures (picture size: 1280*720), and 10 Objects are labeled. The marked contents are Bus, Traffic light, Traffic sign, Person, Bike, Truck, Motor, Car, Train, Rider, and there are about 1.84 million calibration frames in total.

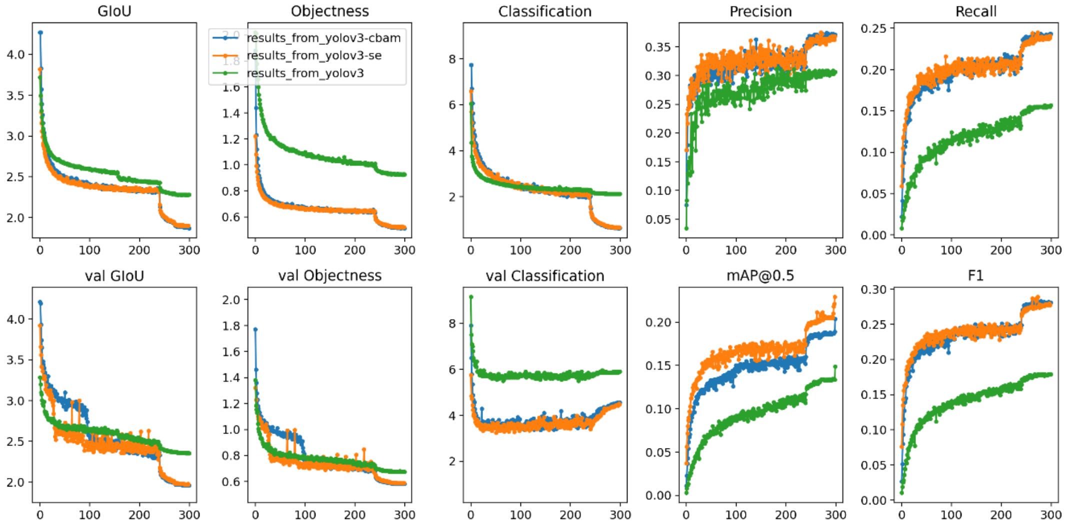
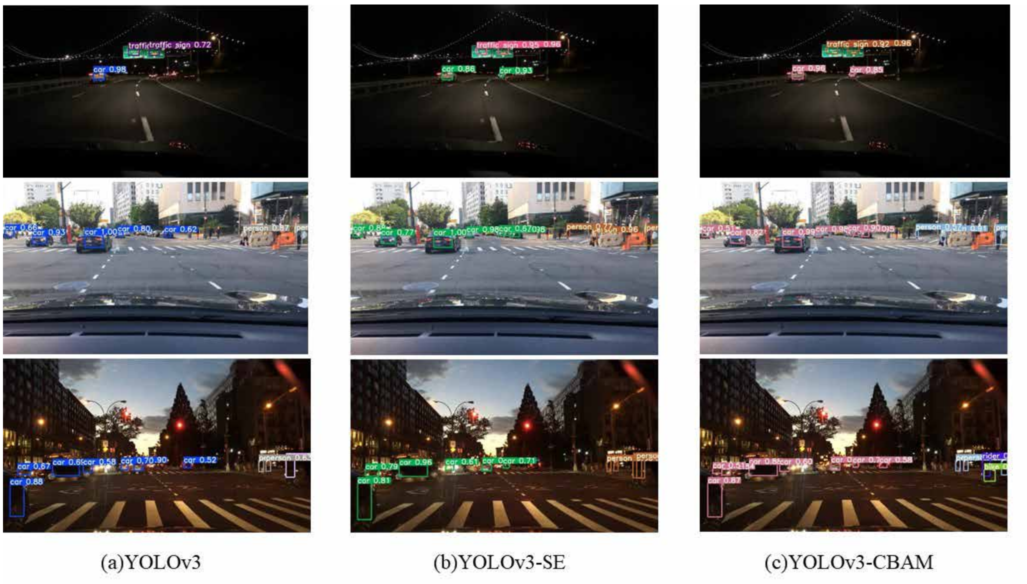


Figure 7. Loss value and performance indicators during training

The BDD100K data set is divided into training set and test set according to 8:2. Both use YOLOv3 to train the model parameters pre-trained on the MS-COCO data set. During training, the momentum is 0.9 and the initial learning rate is 0.001. The rate reduction parameter is 0.0001, the attenuation coefficient is 0.0005, the training picture input size is 412×412, the test picture input size is 512×512, and the training is set to 300 epoch. The loss value and performance indicators during training are shown in Figure 7.

3.3 Detection Performance Evaluation

Figure 8 shows the detection results of the three methods on some pictures of the BDD100K data set, including time scenes of day, evening, and night, and light scenes of strong light, low light and extremely low light. In the first set of pictures, the YOLOv3 algorithm missed the small object car (car), while the YOLOv3-SE and YOLOv3-CBAM algorithms can both detect it; the second set of pictures introduces the attention mechanism to the small object pedestrian (person) is also better than the YOLOv3 algorithm; the third group of pictures YOLOv3-CBAM can detect more categories of small objects. It can be seen that the YOLOv3 algorithm with the introduction of the attention mechanism has significantly improved the detection effect of small objects.



(a)YOLOv3 (b)YOLOv3-SE (c)YOLOv3-CBAM

Figure 8. Test results

To evaluate the test set, use the mean Average Precision (mAP), F1 score and detection speed (Frames per second, FPS) as indicators to evaluate the detection performance. Their calculation formula is as follows:

$$P = \frac{TP}{TP + FP} \tag{7}$$

$$R = \frac{TP}{FP + FN} \tag{8}$$

$$mAP = \frac{1}{N}\sum AP \tag{9}$$

$$F_1 = \frac{2PR}{P + R} \tag{10}$$

where TP (True Positive) is the number of correctly detected objects, FP (False Positive) is the number of falsely detected objects, FN (False Negative) is the number of missed objects, AP (Average Precision) is the average accuracy of each type of object, Its value is equal to the area under the

Precision-recall curve, and N is the total number of categories. The accuracy rate P and the recall rate R are mutually restricted in practice, and there will be an imbalance in the separate comparison. Therefore, the F_1 score is used as the comprehensive evaluation index. The F_1 score takes into account both the precision rate and the recall rate, and is the harmonious average of the two.

Table 1 lists the average accuracy AP of the 3 methods for 10 types of objects, and Table 2 lists the mAP, F1, and FPS results of the 3 methods. The results show that the introduction of attention mechanism is used on the BDD100K autopilot dataset. The average accuracy of the YOLOv3 algorithm mAP and F1 scores are higher than the YOLOv3 algorithm. In terms of the detection effect of small objects, the YOLOv3 algorithm, which introduces the attention mechanism, has significantly improved the detection accuracy of small objects such as Bike, Motor, Rider, and Traffic light.

Table 1. Average precision of object (%)

Class	YOLOv3	YOLOv3-SE	YOLOv3-CBAM
Bike	2.86	**19.7**	**20.7**
Bus	37.85	38.6	38.4
Car	37.63	46.5	46.8
Motor	1.90	**18.9**	**21.4**
Person	6.48	**22.9**	**23.4**
Rider	3.85	**19.8**	**19.9**
Traffic light	3.68	**7.18**	**7.63**
sTraffic sign	19.50	18.6	19.0
Train	0.00	0.00	0.00
Truck	35.13	36.9	37.2

Table 2. Comparison of performance of 3 algorithms

Detection algorithm	mAP/%	FPS	F1	Input size
YOLOv3	14.9	20.13	13.6	512*512
YOLOv3-SE	**22.9**	18.02	27.8	512*512
YOLOv3-CBAM	**23.4**	16.07	28.1	512*512

4 Conclusion

This paper improves the residual network on the YOLOv3 algorithm and introduces the channel and spatial attention mechanism to improve the object detection effect of YOLOv3 on the autonomous driving dataset BDD100K, especially the detection performance of small objects. At the same time, whether it is YOLOv3, YOLOv3-SE or YOLOv3-CBAM algorithm, the mAP value detected on the BDD100K data set is low, and further research is needed. In the field of auto-driving cars, due to the long distance of the object in the distance, the object pixel on the image is small, and after the multi-level network convolution, the feature is not obvious or even disappears, so we will continue to explore and improve the feature expression of small objects in the future, Improve the network's ability to detect small objects.

Acknowledgments. The work of this paper is supported by Southwest Minzu University Graduate Innovative Research Project (Master Program CX2020SP139). A special acknowledgement should be given to Southwest Minzu University for its experimental conditions and technical support.

References

1. REN S Q, HEK M, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks[C]// Proceedings of the 2017 International Conference on Pattern Analysis and Machine Intelligence, DC: IEEE Computer Society, 2017, 39(6):1137-1149

2. Wei Liu, Dragomir Anguelov, Dumitru Erhan, ChristianSzegedy, Scott Reed, Cheng-Yang Fu, and Alexander CBerg. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision (ECCV), 2016: 21–37.
3. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788
4. Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271
5. Redmon J, Farhadi A. Yolov3: An incremental improvement[J].arXiv preprint arXiv: 1804.02767, 2018.
6. HU J, SHEN L, SUN G.Squeeze-and-Excitation Networks[C]//Proceedings of the 2017 International Conference on Computer Vision and Pattern Recognition (CVPR), Piscataway: IEEE, 2018:7132-7141.
7. WOO S, PARK J, LeeJY, et al.CBAM: Convolutional Block Attention Module[C]// Proceedings of the 2017 European Conference on Computer Vision (ECCV), New York: ACM, 2018:3-19.
8. 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.
9. Lin T Y, Dollar P, Girshick R, et al. Feature pyramid networks for object detection[C]// Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition. Honolulu, HI, USA. IEEE, 2017:936-944.
10. Ioffe S, Szegedy C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift[J]. arXiv preprint arXiv: 1502.03167, 2015.
11. He K, Zhang X, Ren S, et al. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification[C]// CVPR. IEEE Computer Society, 2015.
12. Lin M，ChenQ，YanS. Network in network[J].arXiv: 1312.4400，2013
13. Yu F, Chen H, Wang X, et al. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning[C]//Proceedings of the 2017 International Conference on Computer Vision and Pattern Recognition (CVPR), 2018: 2633-2642
14. Everingham M, Gool L V, Williams C K I, et al. The Pascal Visual Object Classes (VOC) Challenge[J]. International Journal of Computer Vision, 2010, 88(2):303-338.
15. Wang C Y, Liao H Y M, Yeh I H, et al. CSPNet: A New Backbone that can Enhance Learning Capability of CNN[J]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020: 390-391
16. LIN T Y, DOLLAR P, GIRSHICK R, et al. Feature Pyramid Networks for Object Detection[C]//Proceedings of the 2017 International Conference on Computer Vision and Pattern Recognition (CVPR), Piscataway: IEEE, 2017:2117-2125.
17. Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.
18. Choi J, Chun D, Kim H, et al. Gaussian YOLOv3: An Accurate and Fast Object Detector Using Localization Uncertainty for Autonomous Driving[C]// The IEEE International Conference on Computer Vision (ICCV), 2019