Research on Traffic Target Detection Algorithm Based on Yolov5

Runjie Liu, Yipeng Duan*, Shuguang Li and Lei Shi

National Supercomputing Center in Zhengzhou, Zhengzhou University, Zhengzhou 450001, Henan, China

Abstract

Traffic target detection is a research hotspot in the fields of autonomous driving, intelligent transportation and so on. In recent years, with the rapid development of deep learning technology, it brings new opportunities and challenges to the field of traffic target detection. Aiming at the problems of missing target detection, low detection accuracy and large amount of network parameters and calculations in the traffic target detection algorithm, this paper proposes an improved YOLOv5-Ghost-CA target detection algorithm. Firstly, the CA (Coordinate Attention) module is introduced into the YOLOv5 Backbone network to enhance the receptive field of the Backbone network and the ability to capture location information. The detection accuracy is also improved. Then, GhostConv module is used to replace the convolution with step size of 2 in the whole network, so as to reduce the amount of network parameters and calculation. Experiments show that the improved algorithm has better detection accuracy, meets the real-time requirements and is easy to deploy on the equipment with insufficient resources than the benchmark algorithm on the KITTI dataset.

Key words

Object Detection, YOLOv5, CA, GhostConv

1.Introduction

In recent years, with the continuous increase of car ownership, the traffic burden and road risk are increasing, and traffic target detection has become a hot research direction. Traffic target detection is mainly divided into two directions, the traditional target detection algorithm and the target detection algorithm based on deep learning. The traditional target detection algorithm, highly affected by the environment, has poor stability and robustness. The traffic target detection algorithm based on deep learning has made a great breakthrough in detection accuracy and speed[1]. Target detection algorithms based on deep learning can be divided into two categories[1]: two-stage target detection algorithm and one-stage target detection algorithm. The two-stage detection algorithm will generate a region proposal that may contain the target to be detected after extracting the features in the first stage. In the second stage, it will locate and classify through convolutional neural network, which is characterized by high detection accuracy, such as R-CNN[2], Fast R-CNN[3], Faster R-CNN[4], SPPNet[5], etc. The one-stage detection algorithm does not need to generate candidate regions, but directly carries out positioning and classification after extracting features. Although the detection accuracy is not as good as the two-stage algorithm, the detection speed is faster, such as YOLO[6], SSD[7], FCOS[8].

In terms of improving the detection accuracy, Li Shanshan[9] replaced the basic network framework VGG-16 in SSD network with residual network ResNet-26, and trained it under KITTI dataset, which improved the detection accuracy and real-time performance. Li Xuan[10], et al. proposed a targeted occlusion regression loss function on the basis of YOLOv3, which can effectively prevent the phenomenon of target missed detection, and the accuracy is improved by 2.12%. Liu Changhua[11] proposed an improved non maximum suppression algorithm Soft-DIoU-NMS and introduced the Focal loss function to realize the effective detection of occluded targets and solve the problem of missing detection of small targets. Yin Yuhang[12] proposed the multi-scale channel attention module (MS-CAB)

ICBASE2022@3rd International Conference on Big Data & Artificial Intelligence & Software Engineering, October 21-23, 2022, Guangzhou, China

corresponding author's e-mail: 1539685209@qq.com (Yipeng Duan)

and attention feature fusion module (AFFB) on the basis of YOLOv5. On the premise of ensuring real-time performance, the mAP(mean average precision) of all targets is increased by 0.9%, while the mAP of small targets is increased by 3.5%. In terms of reducing the complexity of the algorithm, Cao Yuanjie[13], et al. proposed a lightweight target detection network (YOLO-GhostNet) with GhostNet residual structure as the Backbone to solve the problem that YOLOv4 tiny network cannot be deployed on platforms with less resources due to its large amount of parameters and calculations, which greatly reduces the amount of network calculations and parameters without loss of accuracy and speeds up the reasoning speed. Li Yushi[14] et al. introduced the lightweight network MobileNet as the Backbone feature extraction network based on YOLOv5, and embedded the CBAM module to compensate the loss of accuracy.

In practical applications, the two-stage algorithm is difficult to meet the real-time requirements. In the one-stage algorithm, the YOLO series algorithm has developed rapidly in recent years and outperforms SSD and FCOS on the popular data sets. Compared with the previous version, YOLOv5 has greatly improved the detection speed. Therefore, this paper uses YOLOv5 algorithm to realize real-time detection of targets in traffic scenes.

2.Improvement of YOLOv5 network

2.1.YOLOv5 network model

The network model of the original YOLOv5 is composed of three parts: Backbone, Neck and Head. According to the size of the model, it is divided into four versions: YOLOv5, YOLOv5m, YOLOv5l and YOLOv5x. This paper not only wants to improve the detection accuracy, but also takes into account two factors: detection speed and algorithm complexity. Therefore, this paper selects YOLOv5s as the benchmark algorithm.

The components of the network are shown in Figure 1. Backbone network of YOLOv5 is composed of Fous, Conv, C3, SPP(spatial pyramid pooling) and other modules. The Fous module assembly, in the channel dimension, vertical and horizontal intervals slicing section of the input. Compared with convolution down sampling, the Fous output depth is increased by 4 times, more information is retained, and the cost of convolution is reduced and the speed is improved. Conv module performs convolution, regularization, activation and other operations on the input in turn. C3 module, refers to the structural characteristics of CSPNet[15] , divides the input feature map into two parts. The splitting and merging strategies are used across stages to reduce the duplicated gradient information, so that the network can have better learning ability and less calculation The SPP module passes the input through three different sizes of the largest pool layer, and then connects with the input, this structure can greatly improve the receptive field, and the speed loss is small.

The Neck part is mainly composed of FPN[16](feature pyramid networks) and PAN[17](path aggregation networks). The FPN layer transfers strong semantic features downward while the PAN layer transfers strong positioning features upward, The two parts complement each other to enhance the feature extraction ability of the model. Head is the detection layer, and the traditional CNN network only inputs the highest level features into the detection layer. Therefore, the information loss of small target features after multi-layer transmission makes it difficult to identify. YOLOv5 inputs three different sizes of features into the detection layer to detect large, medium and small targets respectively, overcoming the limitations of the traditional detection layer.

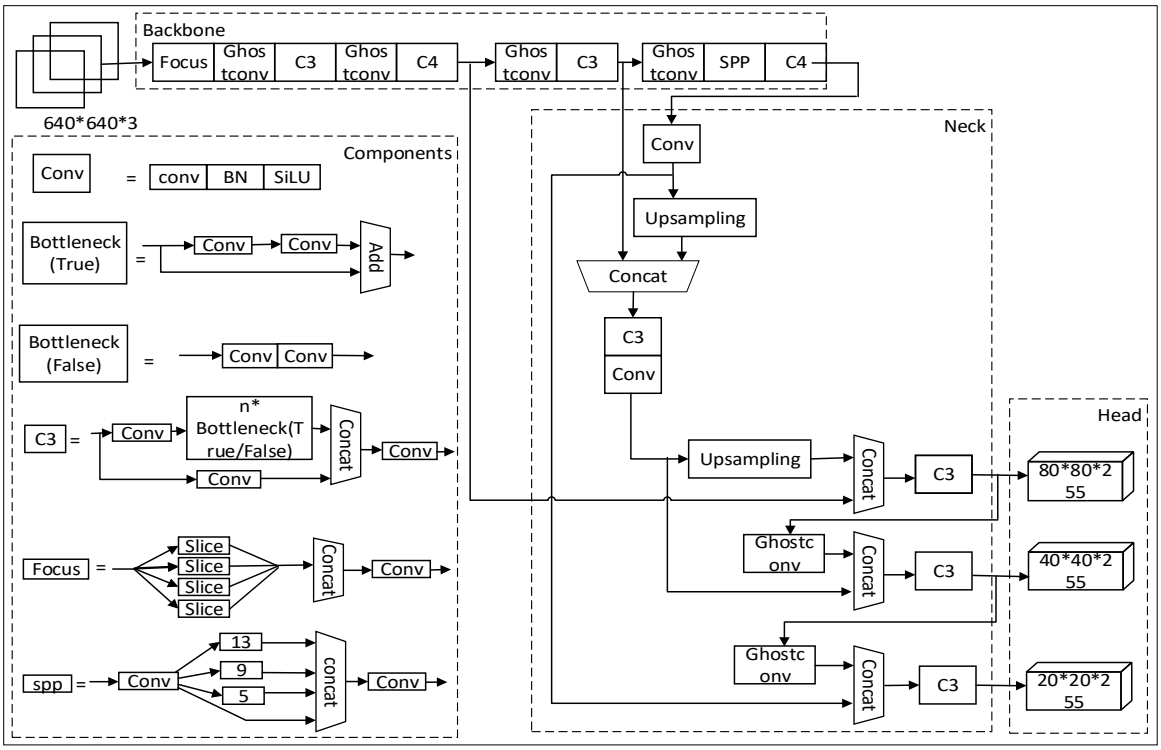

Fig 1. YOLOv5-Ghost-CA.

2.2.CA module

Hou Q[18] proposed CA(coordinate attention) in CVPR2021. For SE(Squeeze-and-Excitation)[19] attention, only the importance of each channel is measured by modeling the channel relationship, and the problem of location information is ignored.

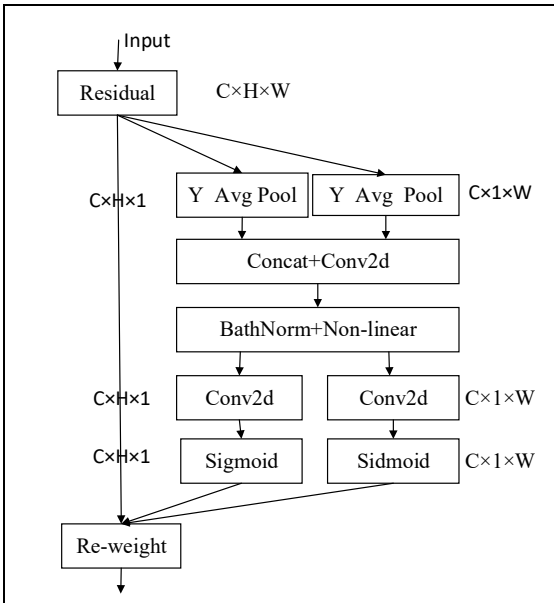

Fig 2. CA module.

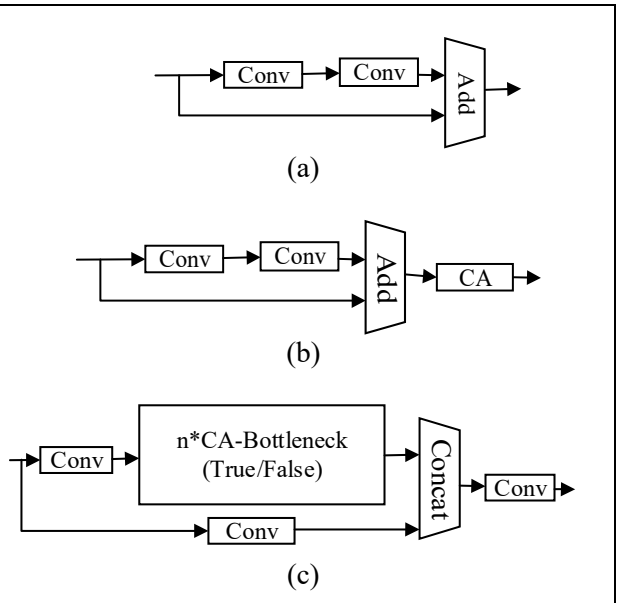

Fig 3 Bottleneck , CA-Bottleneck and CA-C3.

The CA module performs global average pooling of inputs in the height and width directions respectively, the advantage is that it can capture remote dependencies along one spatial direction, retain accurate location information along the other spatial direction, and then encode the generated feature

map respectively to form a pair of direction aware and position sensitive feature map, They can complementarily enhance the representation of the object of interest. The network structure of CA module is shown in Figure 2.

As shown in Figure 3, this paper adds the CA module to the Bottleneck module in the C3 module of the Backbone network to enhance the feature extraction ability of the backbone network and reduce the missed detection rate of the target detection algorithm. After introducing CA module, the original Bottleneck module is changed into CA- Bottleneck module, and C3 module is changed into CA-C3 module.

2.3.GhostConv module

After adding CA attention to the benchmark algorithm, this paper continues to introduce GhostConv module to reduce the parameter quantity and complexity of the network model under the condition of ensuring the accuracy, so that it can be deployed on devices with insufficient computing power. In general, in order to ensure that the network a comprehensive understanding of the input data, the deep neural network will contain rich or even redundant feature map, and many of the feature layers are similar. These similar feature layers are like ghost to each other GhostConv module comes from GhostNet[20], which does not try to remove these redundancies, but obtains them with a lower cost budget.

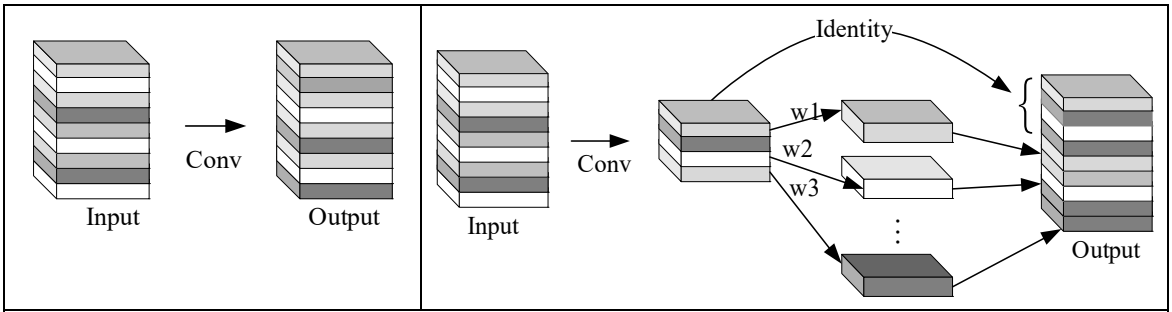

Fig 4.convolution and GhostConv.

GhostConv integrates standard volumes into two steps. The first step is to use fewer convolution kernels to perform convolution operations and generate a small number of feature map, the second step is to generate more feature map with simple linear operations, Finally, the two-step feature map is concatenated, and the size and number of the feature map are unchanged, but the overall amount of parameters and calculations are significantly reduced. Its structure is shown in Figure 4. The convolution can be expressed as:

$$Y=X*w+b \tag{1}$$

Where $X\in R^{c\times h\times w}$ is convolution input，c is the number of input channels, h and w represent the height and width of the input feature map respectively, $Y\in R^{h'\times w'\times n}$is convolution output, n represents the number of output channels, h' and w' Indicates the height and width of the output, $w\in R^{c\times k\times k\times n}$ indicates that the convolution operation is $c\times n$ convolution kernels with size $k\times k$ convolution kernel, b is the offset term.

GhostConv：

$$Y'=X*w' \tag{2}$$

$$Y_{ij}=\varphi_{ij}(Y_i'),i\in[1,m],j\in[1,s] \tag{3}$$

$m\leq n$, the offset term b is omitted. GhostConv module finally needs to generate n feature map, equation (3) needs to generate n-m feature map. Suppose that each feature graph in Y'needs to generate s feature map, Y_i' is the ith feature map in Y', φ_{ij} is the linear operation used by each Y_i' to generate the jth. Finally, as shown in Figure 4, the two parts of the feature map are directly spliced to complete the GhostConv module.

The computation of GhostConv module is approximately 1/s of convolution. All convolution modules with step size of 2 in YOLOv5s are replaced by GhostConv module. The new structure

effectively reduces the amount of parameter calculation and compresses the model.

2.4.overall structure of Improved YOLOv5 network

YOLOv5s is very practical, and it is feasible to use it as a benchmark algorithm for traffic scene target detection. However, from the current application scenario, YOLOv5s can still be further improved. Combined with the improvement measures described earlier in this paper, the improved algorithm YOLOv5-Ghost-CA is obtained. The overall structure of the algorithm is shown in Figure 1.

3.Experiment and result analysis

3.1.Experimental dataset

This paper selects the KITTI 2D scene training dataset as the dataset of this paper. The KITTI dataset is widely used in the field of automatic driving, including images in a variety of actual scenes of urban areas, villages and highways. KITTI dataset contains 9 types of target: 'car', 'van', 'truck', 'pedestrian', 'person_sitting', 'cycle', 'tram', 'misc', 'dontcare', here 'misc' and 'dontcare' are ignored, and the dataset is finally changed into seven categories. At the same time, the dataset is divided into training set, verification set and test set according to 6:2:2.

3.2.Experimental environment

The experiment in this paper uses pytorch 1.10 framework, training environment: CUDA version is 10.2, GPU is Tesla V100S-PCIE-32GB, compilation language is python3.6.8, initial learning rate is 0.01, final learning rate is 0.2, batchsize is 16, and warm-up method with epoch of 3 and momentum parameter of 0.8 is used to warm up the learning rate.

Table 1 Performance indexes of different algorithm.

model	Precision(%)	Recall(%)	Parameters(M)	GFLOPs	mAP(%)	FPS
Yolov5s	91.8	82.1	7.07	16.4	88.1	100
Yolov5s-CA	92.3	84	6.5	16.8	90.9	91
Yolov5s-Ghost	92.9	82.7	5.93	14.1	90.4	95
Yolov5s-Ghost-CA	92.4	86.6	5.35	14.4	91.7	79

3.3.Measurement indicators

In this paper missed detection rate, mAP, parameters, GFLOPs and FPS are used to quantitatively evaluate the performance of the algorithm. The specific calculation formula of the above evaluation indicators is as follows:

$$P_{precision}=\frac{TP}{TP+FP} \tag{5}$$

$$R_{recall}=\frac{TP}{TP+FN} \tag{6}$$

$$L_d=\frac{FN}{FN+TP}\quad L_d=1\text{-}R_{recall} \tag{7}$$

$$AP=\int_0^1 P_{presicion}\,dR_{recall} \tag{8}$$

$$mAP=\frac{1}{Q_R}\sum_{q\in Q_R} AP(q) \tag{9}$$

Among them, TP (true positive) is a positive sample correctly detected as a positive sample, FP (false positive) is a negative sample incorrectly detected as a positive sample, and FN (false negative) is a

negative sample incorrectly detected as a positive sample. L_d is the missed detection rate, and mAP is used to measure the detection ability of the algorithm. Parameter quantity and calculation quantity are used to measure the complexity of the algorithm. FPS represents the number of detected pictures per second, which is used to show the detection speed of the algorithm.

3.4.comparison experiment

In order to verify the effectiveness of the improved algorithm in this paper, a comparative experiment was designed on the basis of YOLOv5s, and four groups of tests were conducted. The four algorithms were trained on the KITTI dataset, and each group of experiments used the same super parameters and training skills. The experimental results are shown in Table 1.

It can be seen from Table 1 that the YOLOv5-CA algorithm after the introduction of the CA module into the YOLOv5s algorithm has an increase of 2.8% compared with the benchmark algorithm YOLOv5s, which proves that the CA module can effectively enhance the feature extraction ability of the Backbone network. After introducing GhostConv module, the parameter quantity of YOLOv5s-Ghost algorithm is reduced by 16%, the calculation quantity is reduced by 14%, and the mAP is increased by 2.3%, indicating that the GhostConv module has the function of lightweight, which can effectively improve the overall performance of the algorithm. YOLOv5-Ghost-CA algorithm, which combines CA module and GhostConv module, reduces the missed detection rate by 4.5% and increases the mAP by 3.6% when the parameter quantity is reduced by 24.3% and the calculation amount is reduced by 12.2%. While effectively improving the detection accuracy, the smaller algorithm complexity is convenient for deployment on equipment with insufficient resources, and the detection speed of 79 FPS also meets the requirements of real-time.

4.Conclusion

This paper proposes an improved traffic target detection algorithm YOLOv5-Ghost-CA based on YOLOv5s. The CA module is introduced to enhance the ability to capture location information of the Backbone network, and improve the feature extraction ability of the Backbone network. Replacing the convolution with step size of 2 in the network with GhostConv module reduces the amount of parameters and GFLOPs. The experimental results show that the mAP of the improved algorithm is increased by 3.5%, the amount of parameters is reduced by 24.3%, and the amount of computation is reduced by 12.2%. It is more conducive to deploy on the equipment with insufficient computing resources, and meets the real-time requirements, which has high application value.

5.References

[1] L P Fang, H J He, G M Zhou. Overview of target detection algorithm[J]. Computer engineering and application, 2018,54(13):11-18+33.

[2] R Girshick, J Donahue,T Darrell, et al. Rich feature hierarchies for object detection and semantic-segmentation[C]//2014IEEE Conference on Computer Vision and Pattern Recognition, 2014:580-587.

[3] R Girshick, Fast R-CNN[C]//2015 IEEE International Conference on Computer Vision, 2015:1440-1448.

[4] S P Ren, K He, R Girshick, et al. Faster R-CNN: towards real time object detection with region proposal networks[J]IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6):1137-1149.

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

[6] J Redmon, S Divvala, R Girshick, et al. You only look once: unified, real-time object detection[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition, 2015:6517-6525.

[7] W Liu, D Anguelov, D Erhan, et al.SSD:single shot multi-box detector[C]//14th European Conference on Computer Vision.Cham:Springer, 2016:21-37.
[8] Z Tian, C Shen, H Chen, et al. Fcos: Fully convolutional one-stage object detection[C]// Proceedings of the IEEE/CVF international conference on computer vision, 2019:9627-9636.
[9] S S Li. Traffic scene multi-target detection based on deep learning[D]. Hunan University, 2017.
[10] X Li, J Li, W Haiyan. Research on target detection algorithm in dense traffic scenes[J]. Computer technology and development, 2020,30 (07): 46-50.
[11] C H Liu. Automatic driving road target detection in complex traffic scenes[D] Dalian University of technology, 2021.
[12] Y H Yin. Research on traffic scene target detection method based on feature fusion[D] Dalian University of technology, 2021.
[13] Y J Cao, Y X Gao. Lightweight beverage recognition network based on GhostNet residual structure[J] Computer engineering, 2022,48 (03): 310-314.
[14] Y S Li, C Y Zhang, Y K Zhao, et al. Research on lightweight obstacle detection model based on model compression[J/OL].Laser Journal: 1-7.
[15] C Y Wang, H Y LIAO, Y H Wu, et al. CSPNet: a new backbone that can enhance learning capability of CNN[C]//Proceedings of IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. USA:IEEE Press，2020:571-1580.
[16] T Y Lin, P Dollar, R Girshick, et al. Feature pyramid networks for object detection[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition(CVPR).Honolulu: IEEE Computer Society, 2017, OI 10.1109/CVPR.2017.106.
[17] S Liu, L Qi, L Qin, et al. Path aggregation network for instance segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:8759-8768.
[18] Q Hou, D Zhou, J Feng. Coordinate attention for efficient mobile network design[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021:13713-13722.
[19] J Hu, L Shen, G Sun. Squeeze-and-excitation networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018:7132-7141.
[20] K Han, Y Wang, Q Tian, et al. GhostNet:More Features From Cheap Operations[C]. /IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020:1577-1586.