Frontiers in Computing and Intelligent Systems
ISSN: 2832-6024 | Vol. 3, No. 1, 2023

Road Traffic Small Target Detection Algorithm Based on Lightweight YOLO V5

Chanyi Liu [1, 2 *], Dan Huang [1, 2], Tao Wang [1, 2], Tao Zhu [1, 2]

[1] School of Automation and Information Engineering, Sichuan University of Science & Engineering, Yibin, 644002, China
[2]Artificial Intelligence Key Laboratory of Sichuan Province, Sichuan University of Science & Engineering, Yibin, 644002, China;
[*] **Corresponding author**: Chanyi Liu (Email: 2720806204@qq.com)

Abstract: For the current problem of large number of parameters and large computation of the target detection network, the lightness improvement is carried out on the basis of yolov5s, and the dynamic convolution of OD is introduced to compensate for the accuracy degradation caused by the lightness, and the Map0.5 of the improved Yolov5s- EfficientNetV2-OD network reaches 82.8 through comparison experiments, and the FPS is 59 frames per second. The experiment proves that the network maintains the detection accuracy and detection speed of the network for small targets on the basis of light weighting. It is suitable for deployment on vehicle-mounted devices.

Keywords: Deep learning; Lightweight; Small target detection.

1. Introduction

With the rise in per capita household income in China, motor vehicle ownership has increased dramatically, and the increase in vehicle ownership has brought great pressure on the roads and the ensuing traffic accidents. In addition to the weather and the vehicle itself mechanical failure and other objective factors, traditional car driving requires the driver to always observe the surrounding road information, if fatigue, distracted easily caused by traffic accidents. With the development of artificial intelligence, the technology of automatic driving is becoming more and more mature. There is a big difference between traditional manual driving vehicles and self-driving vehicles, and automatic driving can be achieved without human intervention and precise automatic control, and when there is an unexpected situation, the automatic driving system will intervene in advance as well as emergency braking when necessary.

As a basic component of an autonomous driving system, the detection capability of the target detection is one of the reference indicators for judging the level of autonomous driving. Common detection networks are divided into two-stage detection networks such as RCNN, Fast-Rcnn,Faster-Rcnn,single-stage detection networks SSD, and Yolo[5-9] series. The two-stage network has high detection accuracy but slow detection speed, and the single-stage network has fast detection speed but slightly lower detection accuracy. For target detection, many scholars have been improving the capability of target detection networks to achieve good detection results, Ren et al improved the SSD algorithm to enhance the detection capability of the network for obscured objects or small objects. Li et al proposed Quality Focal Loss to improve the error suppression that occurs during NMS operation to improve the detection accuracy of the network. Huang et al designed a variable residual convolutional branching module with RetinaNet as the base framework to improve the detection accuracy of various types of targets. However, with the improvement of detection capability, the network structure becomes more and more complex, the number of parameters, and the model size also become larger, leading to the decrease of detection speed.

Based on the above problems, this paper performs light weighting on the basis of yolov5s, and introduces OD dynamic convolution to improve the feature extraction capability of the network by in order to compensate for the loss of model accuracy after light weighting. The improved network can guarantee the detection capability on the basis of lightweighting. It is beneficial to be deployed on in-vehicle devices.

2. Related Work

YOLOv5 algorithm is the latest version of the current YOLO series detection algorithm, proposed by Ultralytics Yolov5 main structure is shown in Figure 1, the model consists of four parts: input side (Input), backbone network (Backbone), multi-scale feature fusion network (Neck), and detection network (YOLOHead).

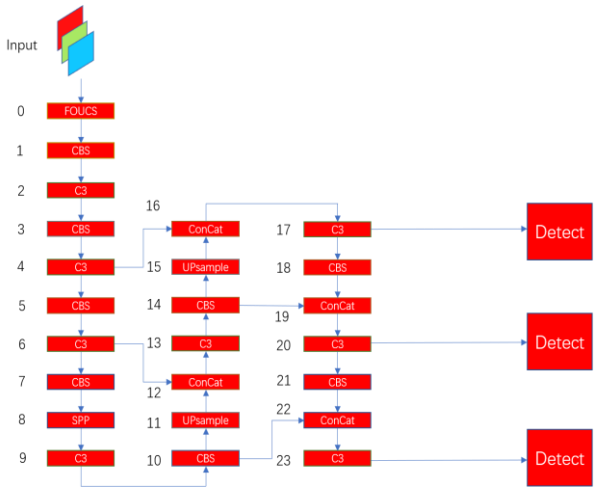

Figure 1. Yolov5 Network

Yolov5 input consists of SPP structure, FOUCS structure, and C3 structure, which are used to extract the image network features.Neck side consists of multi-scale pyramid FPN+PAN structure, through such a pyramid structure to increase the resolution of the image, so that the image can be divided into more size pyramids. The more layers of the pyramid, the more

details can be extracted. In addition, a batch normalization layer is used, which can effectively prevent the occurrence of overfitting. The prediction side consists of three detection heads, corresponding to large, medium and small size targets. the Yolov5 structure is shown in Figure 1.

3. Method

In this paper, we optimize YOLOv5s from two aspects: lightweight and attention. Firstly, the yolov5s backbone network is replaced with EfficientNetV2 network to reduce the computational complexity and make the detection model more lightweight, and secondly, the conventional convolution is replaced with ODConv dynamic convolution in the output layer, which uses a multidimensional attention mechanism to learn complementary attention along four dimensions of the kernel space through a parallel strategy.

3.1. Lightweight

EfficientNetv2 is a lightweight network that combines training perceptual neural architecture search and scaling, which better balances the relationship between accuracy, training speed and number of parameters.EfficientNetV2-S structure table is shown in Table X.

Table 1. EfficientNetV2- S structure table

Stage	Operator	Stride	Channel	Layers
0	Conv3x3	2	24	1
1	Fused-MBConv1, K3x3	1	24	2
2	Fused-MBConv4, K3x3	2	48	4
3	Fused-MBConv4, K3x3	2	64	4
4	MBConv4,K3x3,SE0.25	2	128	6
5	MBConv6,K3x3,SE0.25	1	160	9
6	MBConv6,K3x3,SE0.25	2	272	15
7	Conv1x1&Pooling&FC	-	1792	1

To address the problem of slow speed due to DW convolution in the MBConv module in EfficientNet, the efficientNetv2 network addresses the problem of accuracy degradation and slower training by using the Fused-MBConv structure instead of the MBConv structure in the efficientNet network.The Fused- MBConv replaces the 1x1 normal convolution and 3x3 deep separable convolution in MBConv by 3x3 normal convolution, and then uses the NAS technique to search for the best combined structure, the network structures of MBConv and Fused-MBConv are shown in Figure 2, Figure 3

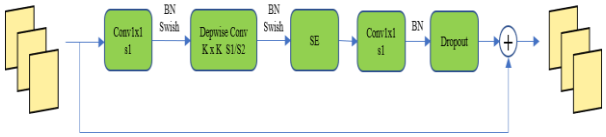

Figure 2. MBConv Module

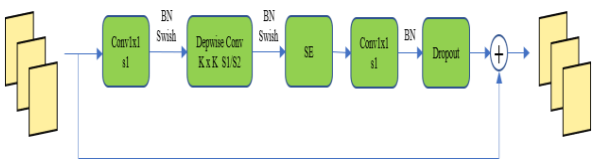

Figure 3. Fused-MBConv Module

The structure of the Yolov5s backbone network after replacing it with efficientNetv2 is shown in Figure4.

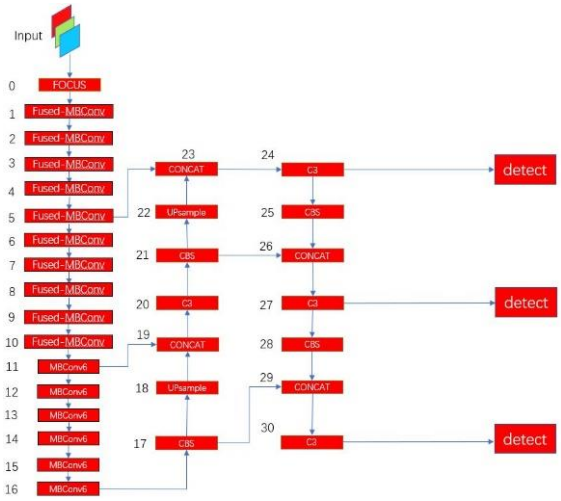

Figure 4. Yolov5s- EfficientNetV2 structure

3.2. ODConv

ODConv convolution takes into account the dynamics of the null field, input channel, output channel and other dimensions, so it is called full-dimensional dynamic convolution. oDConv adopts multidimensional attention mechanism to learn complementary attention in four dimensions of convolutional kernel: input channel, output channel, kernel space and number of kernels through parallel strategy, which enhances the feature extraction ability of convolution in all aspects. The structure of ODConv comparison is shown in Figure 5.

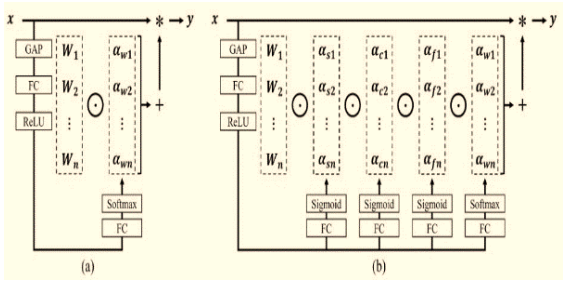

Figure 5. DyConv and ODConv structure comparison diagram

The above figure shows a schematic comparison of the structures of DyConv (a),ODConv (b). Unlike CondConv and Dyconv, which compute a single attention scalar, for the convolution kernel ODConv utilizes a new multidimensional attention mechanism that computes four types of attention in parallel along all four dimensions of the kernel space. The four types of attention convolution are illustrated in Figure 6

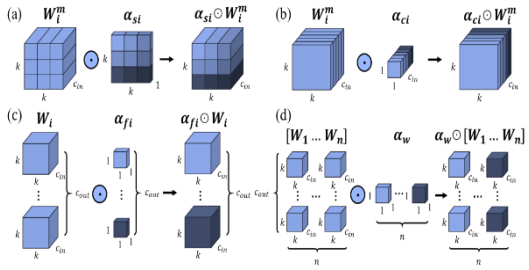

Figure 6. ODConv illustration diagram

(a), (b) (c) (d) represent the position multiplication along the spatial dimension, the channel multiplication along the input channel dimension, the filter multiplication along the output channel dimension, and the kernel multiplication along the kernel dimension of the convolution kernel space,

respectively. These four types of attention are complementary, and by progressively multiplying the convolution with different attentions along the position, channel, filter, and kernel dimensions will make the convolution operation differ for each dimension of the input, providing better performance to capture rich contextual information. Therefore, ODCOnv can significantly improve the feature extraction capability of convolution. The structure of the network after adding dynamic OD convolution is shown in Figure 7.

Figure 7. Yolov5s- EfficientNetV2-OD structure diagram

3.3. Dataset

The KITTI dataset, co-founded by Karlsruhe Institute of Technology and Toyota Institute of Technology, is the largest international dataset for evaluating computer vision algorithms in autonomous driving scenarios. The KITTI includes real image data from urban, rural, and highway scenes, with up to 15 vehicles and 30 pedestrians per image, and various levels of occlusion and truncation. The original dataset is classified and subdivided into car, van, truck, pedestrian, pedestrian(sitting), cyclist, tram, and misc. In this paper, the original KITTI dataset consists of car, truck, van as car, pedestrian, person_sitting as person, and cyclist as rider, and the final KITTI dataset consists of car, person, and rider. These three categories.

3.4. Configuration

The operating system of the experimental platform in this paper is Window10, the hardware configuration is Intel(R) Core (TM) i7-10875H 2.30GHz processor, GeForceRTX2060 graphics card with 6GB of video memory size, and the software environment equipped with Anaconda, CUDA11.4.0, using Pytorch1.7.1 Deep learning and YOLOv5 target detection framework.

3.5. Evaluation Indicators

Average precision (AP): The AP value is the area enclosed by the P-R curve (consisting of the values of precision P and recall R at different confidence thresholds) and the coordinate axes, and the larger the area enclosed by the P-R curve, the better the detection performance of the network model.

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

Mean Average Precison (mAP): indicates that the AP of multiple categories is averaged and calculated by the following formula:

$$map = \frac{\sum_{i=1}^{n} AP_i}{n} \qquad (2)$$

n denotes the number of categories.

Detection speed (FPS): reflects the real-time nature of the algorithm, and being able to detect obstacle targets in real-time is a requirement for self-driving cars. The common metric for detection speed is Frame Per Second (FPS), which indicates the number of images that can be detected per second, and the larger the FPS, the faster the model detects and the better the real-time performance.

3.6. Training Strategies

This experiment was conducted on the dataset KITTI, and 5000 images in KITTI were selected and divided into training set, validation set, and test set in the ratio of 8:1:1. The divided training set is 4000 images, and the validation set and test set are 500 images each. The input image size is 640×640, the batchsize is 16, the test batchsize is 1, the initial learning rate is 0.01, the learning rate momentum is 0.937, and the training epoch is 200 rounds. Adaptive anchor is used to set the initial anchor box size, adaptive moment estimation (Adam) optimizer is used to optimize the model, and mosaic data augmentation is used to augment the dataset, which increases the diversity of the dataset and the number of targets, so that the algorithm has better generalization ability. The other parameters are set according to the official default parameters of YOLOv5s.

4. Experiments

4.1. Lightweight Improvement Experiment

In order to verify the Effectiveness of lightweight networks, the EfficientNetv2 lightweight network was replaced with the original yolov5s and Yolov5-nano and Mobilenetv3 were replaced with the original yolov5s network. Comparative experiments were conducted using the same dataset under the same configuration environment. The comparison results are shown in Table 2

Table 2. Lightweight experimental results

Models	Backbone	Size	Par	Map0.5	Fps
Yolov5s	CSP1_X	14.4	7.07	83.5	60
Ynano	CSP1_X	3.7	1.76	78.2	93
YMbv3	Mobv3	7.3	3.55	76.1	73
YEffV2	EffV2	11.2	5.43	81.4	66

Yolov5s-nano is short as Ynano, Yolov5s-MobileNetV3 is short as YMbv3 Yolov5s-EfficientNetv2 is short as YEffv2. From the results of the lightweight experiments in Table 2, we can see that the performance of all networks tested in the data set, YOLOv5s has the highest indexes, with Map@0.5 reaching 83.5; among the lightweight networks, Yolov5s-Mobilenetv3 has the lowest accuracy rate, Map@0.5 value among all networks, and The Map@0.5 of Yolov5s-EfficientNetV2 reaches 81.4, which is higher than the 76.1 of Yolov5s-Mobilenetv3 network and 78.2 of Yolov5s-nano, and is the best performance among all lightweight networks. 22.2%, 23.2%, And the FPS is higher than the 60 frames per second of the original yolov5s. It is slightly lower than the 73 frames per second of Yolov5s-Mobilenetv3.

4.2. Add ODConv Moudle

In order to compensate for the loss of accuracy after light weighting, the OD dynamic convolution is fused in the light weighting network and the comparison experiments are conducted as follows.

Yolov5s-EfficientNetv2-OD is abbreviated as YEffV2-OD. From Table 3, we can see that after adding ODConv convolution to the lightweight network Yolov5s-EfficientNetV2, the Map@0.5 value improved by 1.4% with

better detection capability for small targets, although the FPS decreased slightly by 7 frames per second. The original image and the detection effect are shown in Figure 8 and Figure 9

Table3. Add ODConv experimental results

Models	ODConv	Map0.5	FPS
Yolov5s		83.5	60
Ynano		78.2	93
YMbv3		76.1	73
YEffV2		81.4	66
YEffV2-OD	√	82.8	59

Figure 8. Original Image

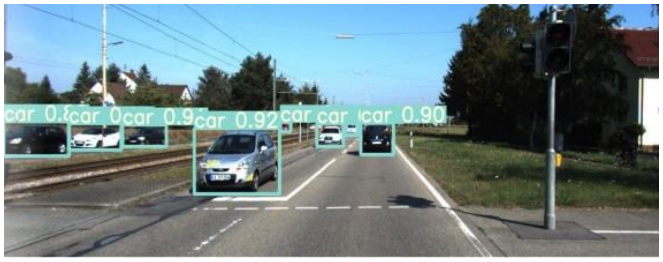

Figure 9. Detection effect picture

5. Conclusion

In this paper, firstly, the EfficientNetV2 lightweight network replaces the yolov5s backbone network, which reduces the number of network model parameters and computation, and secondly, the ODconv dynamic convolution is added to make the model feature extraction ability stronger, after experimental comparison, the improved Yolov5s-EfficientNetV2-OD network model Map@0.5 reaches 82.8 and the PFS is 59, which reduces the requirements for the performance of on-board devices while ensuring the lightweight and detection accuracy.

References

[1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation [J]. 2013.

[2] GIRSHICK R. 2015 IEEE International Conference on Computer Vision (ICCV); proceedings of the IEEE International Conference on Computer Vision, F, 2015 [C].

[3] REN S, HE K, GIRSHICK R, et al. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks; proceedings of the NIPS, F, 2016 [C].

[4] WEI L, DRAGOMIR A, DUMITRU E, et al. SSD: Single Shot MultiBox Detector [J]. 2016.

[5] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-Time Object Detection; proceedings of the Computer Vision & Pattern Recognition, F, 2016 [C].

[6] REDMON J, FARHADI A. YOLO9000: Better, Faster, Stronger; proceedings of the IEEE Conference on Computer Vision & Pattern Recognition, F, 2017 [C].

[7] REDMON J, FARHADI A J A E-P. YOLOv3: An Incremental Improvement [J]. 2018.

[8] BOCHKOVSKIY A, WANG C Y, LIAO H. YOLOv4: Optimal Speed and Accuracy of Object Detection [J]. 2020.

[9] GE Z, LIU S, WANG F, et al. YOLOX: Exceeding YOLO Series in 2021 [J]. 2021.

[10] REN J C X H, LIU J B, ET AL. Accurate single stage detector using recurrent rolling convolution [J]. IEEE Conference on Computer Vision and Pattern Recognition, 2017, Honolulu, HI, USA. New York: IEEE Press,: 752-60.

[11] AL L X W W W L E. Generalized focal loss： learning qualified and distributed bounding boxes for dense object detection. Advances in Neural Information Processing Systems, 2020.

[12] HUANG WENHAN Y G, GENG KEKE, ET AL. Target detection in complex driving scenes based on adaptive fusion of dilated convolutional features [J]. Journal of Southeast University Natural Science Edition, 2021, (51(06)): 1076-83.

[13] LI CHAO Z A Y A. Omni-dimensional dynamic convolution [EB/OL] https://arxivorg/pdf/220907947pdf, [2022-09-11].

[14] LIN T Y, DOLLáR P, GIRSHICK R, et al. Feature Pyramid Networks for Object Detection [J]. 2016.

[15] LIU S, QI L, QIN H, et al. Path Aggregation Network for Instance Segmentation [J]. 2018.

[16] Q T M L. Efficientnetv2: Smaller models and faster training [J]. International Conference on Machine Learning, 2021: 10096-106.

[17] C T M L Q E R M S F C N N. EfficientNet: Rethinking model scaling for convolutional neural networks [J]. Proceedings of the 36th International Conference on Machine Learning, 2019: 6105-14.r