An improved network for pedestrian-vehicle detection based on YOLOv7

Zhihui Chen
School of Electronic Information and Automation, Tianjin University of Science and Technology,
1038 Dagu Nanlu, Hexi District, Tianjin, China

Xiaoyan Chen
School of Electronic Information and Automation, Tianjin University of Science and Technology,
1038 Dagu Nanlu, Hexi District, Tianjin, China

Keying Ren
School of Electronic Information and Automation, Tianjin University of Science and Technology,
1038 Dagu Nanlu, Hexi District, Tianjin, China
E-mail:1594838831@qq.com, cxywxr@tust.edu.cn, renkeying@mail.tust.edu.cn
https://www.tust.edu.cn/

Abstract

With the continuous development and improvement of information technology, target detection has gradually attracted people's attention. Therefore, object detection has become more important. In this paper, a large number of experiments have been made to improve the detection accuracy of pedestrians, vehicles and license plates in cities. The improved algorithm based on YOLOv7 was used to conduct a large number of experiments on the urban pedestrian vehicle dataset. Experiments show that new improvements have improved detection accuracy.

Keywords: YOLOv7, target detection, Swin Transformer, CNeB

1. Introduction

With the sequential development of computer technology, the task of target detection in the transportation system is gradually transformed from manual to machine. Object detection is a basic research direction in the field of computer vision technology, depth learning and image recognition. It is the technical premise of more complex high-level computer vision tasks, such as predicting the behavior of objects in images after the detection task. Traditional detection methods mainly use HOG[1] or vector machine algorithm. They mainly use sliding Windows to detect targets. This method has the disadvantages of long time redundancy and poor robustness, so it is not suitable for large-scale applications. In recent years, target detection algorithm based on deep learning is gradually used in all walks of life. The algorithm based on deep learning has the characteristics of strong real-time and high accuracy in detection. Therefore, target detection algorithm based on depth learning has become the mainstream algorithm in target detection. This paper uses the depth learning algorithm to detect vehicles and pedestrians.

2. Detection algorithm

The main goal of pedestrian and vehicle detection algorithm is to find the target from a series of images or videos, and determine the size and location of the target according to the feature information. But in the actual scene of the city, there will be a lot of interferences. For example, occlusion, illumination and other conditions lead to false detection and missing detection, resulting in inaccurate detection. At present, the detection algorithms based on depth learning are mainly divided into two categories, namely, the two-stage target detection model based on region extraction and the one-stage target detection model for direct position regression.

2.1 *Two-stage object detection algorithm*

The core of the two-stage target detection model based on region extraction is the CNN, which has the advantages of local connection and weight sharing, and has good robustness in object classification applications. Firstly, the region extraction operation uses the CNN backbone to extract image features, then finds out possible foreground objects (candidate regions) from the feature map, and finally performs sliding window operation on the candidate regions to further judge the target category and location information, which greatly reduces the time complexity of calculation. The classical algorithms include R-CNN, SPP-NET, Fast R-CNN, Faster R-CNN and Mask R-CNN. R-CNN (Region CNN) was proposed by Ross-Girshick and was the earliest algorithm based on depth learning target detection. The mAP (mean average precision) of the algorithm on the VOC2007 dataset reaches 66%, breaking through the bottleneck that R-CNN has been unable to overcome for many years, and significantly improving the detection rate. But it has some disadvantages. The intermediate data generated by this algorithm will consume too much storage resources, and the input image must be forcibly scaled to a fixed size, which will lead to deformation of the target object and affect the accuracy of detection. SPP-NET solves the problems of repeated calculation and fixed size based on R-CNN.

Fast R-CNN uses the feature pyramid in SPP-NET for reference to improve the R-CNN network, greatly reducing the training time and improving the model detection performance. However, Fast-CNN generation of candidate boxes is time-consuming, and can not be used for end-to-end training and GPU acceleration. To solve this problem, Ren et al. designed RPN (Region Proposal Network,) in 2015 and proposed Faster R-CNN[2] model. Faster R-CNN achieves end-to-end training, but because of its structure, it has the disadvantages of large parameters, slow detection speed and poor real-time performance. The above target detection methods based on region extraction generate the frame of the region of interest, and then classifies and regresses the frame. Although the detection accuracy has been continuously improved, the detection speed is generally slow, which is not suitable for application scenarios with high real-time requirements.

2.2 *One stage target detection*

Different from the two-stage target detection algorithm, the one-stage detection algorithm pays more attention to the improvement of detection speed. Among the single-stage detection algorithms, SSD (Single Shot Multi Box Detector)[3] and YOLO (You Only Look Once) series are the main methods.

The one-stage target detection algorithm does not use the middle layer to extract candidate regions, but performs feature extraction, target classification and position regression in the entire convolution network, and then obtains the target position and category through a backward calculation. The accuracy is slightly lower than that of two-stage detection, but its speed is greatly improved, so that the detection algorithm based on One stage target detection can be used in many scenes that require reasoning speed.

This experiment uses YOLOv7 algorithm. YOLOv7 is faster and more accurate than most known target detectors. Among the known real-time target detectors above 30 frames per second of GPU V100, YOLOv7 has the highest detection accuracy. According to the different code running environments (edge GPU, common GPU and cloud GPU), YOLOv7 has designed three models respectively: YOLOv7 tiny, YOLOv7 and YOLOv7-W6. Compared with other YOLO series network models, the detection idea of YOLOv7 is similar to that of YOLOv4 and YOLOv5[4], and its network structure is shown in the Figure 1 below.

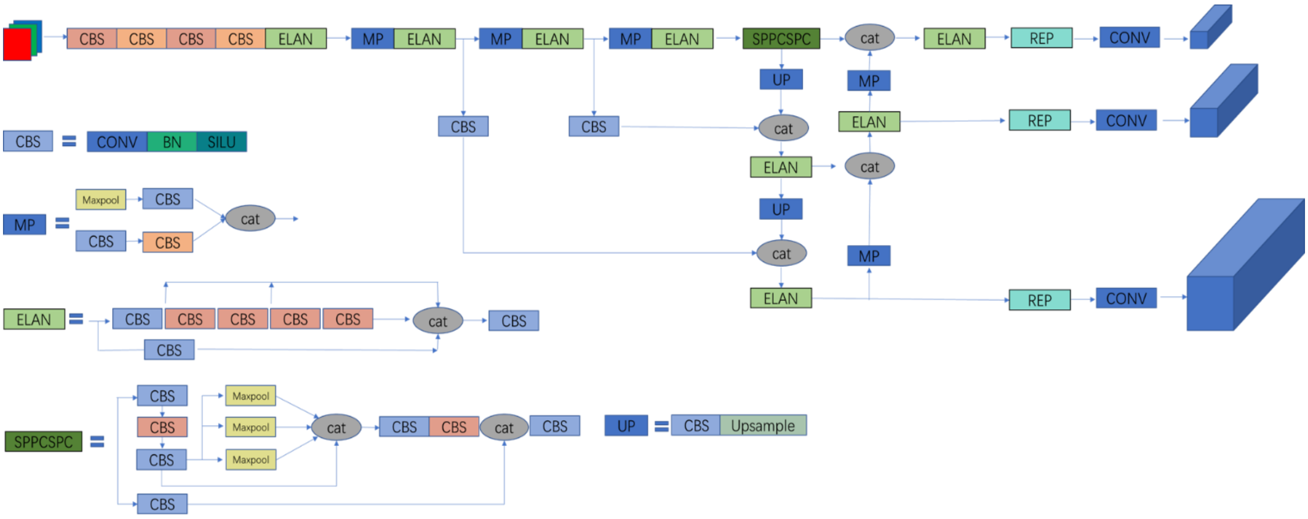

Fig. 1 YOLOv7 network structure diagram

2.3 *YOLOv7*

YOLOv7[5] network is divided into input, backbone and head. The input picture is preprocessed by data enhancement. The preprocessed images enter the backbone for feature extraction. Then, the extracted features are fused by the head to obtain features of large, medium and small sizes. Finally, the fused features are sent to the detection head and the results are output. The backbone network of YOLOv7 network model is mainly composed of convolution, ELAN module, MPConv module and SPPCSPC module. In addition to architecture optimization, the method proposed in this study also focuses on the optimization of the training process, focusing on some optimization modules and optimization methods.

2.3.1 *Re-parameterization*

The model re-parameterization technology combines multiple calculation modules into one in the reasoning stage. The weights of the models under different iterations are weighted and averaged. This method divides a module into several module branches during training, and integrates multiple branch modules into a completely equivalent module during reasoning.

2.3.2 *Label assignment*

At the same time, the YOLOv7 proposes a new label allocation method, called coarse to fine. The researchers call the head responsible for the final output the lead head, and the head used for auxiliary training the auxiliary head. The coarse to fine hierarchy of tags is generated, which are used for auxiliary head and lead head learning, respectively. YOLOv7 proposed a deep supervision label allocation strategy, as shown in the Figure 2 below.

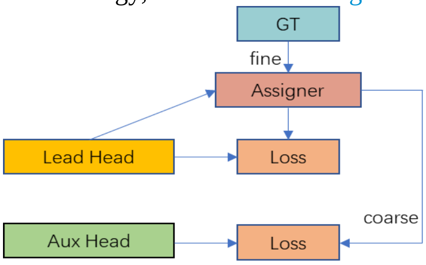

Fig. 2 Label assignment

3. Improvements

3.1 *Backbone improvement*

In this paper, The first and fourth ELANs in the backbone are replaced by the combination of C3 structure and CNeB (ConvNeXt Block). Experiments show that CNeB has reached the limit of pure Conv network design. CNeB uses large core 7*7 DWCONV to replace ordinary convolution. At the same time, although BN layer can improve convergence and reduce over fitting, it still has many complexities that will adversely affect the performance of the model, so LayerNorm is used to replace BN. Finally, replacing ReLU with GELU makes FLOPs reduce to 2.66G during training while improving accuracy. The structure diagram of CNeB is shown in Figure 3.

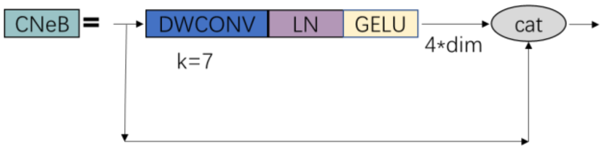

Fig. 3 CNeB structure diagram

3.2 *Head improved*

Attention mechanism has been widely used in the research of deep learning in recent years. This article uses Swin Transformer Block[6] to replace ELAN structure in the head part of YOLOv7. For the target detection task, many studies have shown that the addition of attention modules in the network can improve the representation ability of the network model, effectively reduce the interference of invalid targets, improve the detection effect on the target of interest, and achieve high effect of the network model. When the traditional Vision Transformer detects the whole image, it will face the problem of too many parameters and too high video memory usage. Therefore, the Swin Transformer Block method is adopted. Divide the incoming data into several windows, and pay attention to the small windows to reduce the video memory occupation of Q * K operation, and reduce the computational complexity from O(N^2) becomes O(N). The specific implementation will be divided into n 7*7 patches when the data is transferred to the W-MSA. So before passing through the W-MSA, data must be filled to make it an integer multiple of 7. Swin Transformer uses window self-attention to reduce computational complexity. In order to ensure the connection between non overlapping windows, it uses shifted window self-attention to recalculate the self-attention after window offset. The Swin structure diagram is shown in Figure 4.

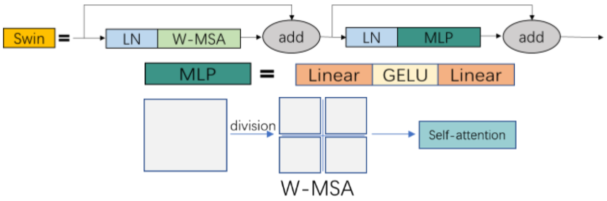

Fig. 4 Swin structure diagram

4. Experiment and software platform

4.1 *Experimental equipment*

The experimental platform is configured in Table 1.

Table. 1 Device-related configuration

Name	Version
Ubunt	20.04.3 LTS
CUDA	11.2
Graphics	GeForce GTX 1080 Ti
Frame	Torch

4.2 *Data collection*

The dataset used in the experiment was collected by ourselves. Each image contains multiple targets types. The dataset contains four types of target. They are people, faces, cars and license plates in cities. A labelimg software is used for manual marking. There are 20,000 pictures in the experimental dataset. After selected, 5000 training sets and 500 verification sets are finally set.

4.3*Experimental setup*

The experimental parameters are set as shown in Table 2.

Table. 2 Experimental parameters

Name	Parameter
Pre-training	Yolov7_training.pt
hyp	hyp.scratch.p5.yaml
batch_size	16
epochs	300

4.4 *Experimental result*

In this paper, mAP (mean Average Precision) is selected as the evaluation index of the model. mAP is the average of all categories of AP (Average Precision). It usually reflects the performance of the model.

In this paper, under the same hardware conditions, we set the same parameters to conduct a comparative experiment between YOLOv7, YOLOX-s[7], YOLOv5-s and the improved YOLOv7. The experimental results are shown in Table 3.

Table. 3 Detection results of different algorithm

Method	mAP@.5	mAP@.5:.95
YOLOv5-s[4]	0.832	0.502
YOLOX-s[7]	0.885	0.511
YOLOv7[5]	0.909	0.516
our	0.949	0.537

Experiments show that the accuracy of the improved model is improved. After adding CNeb and Swin Transformer Block, the accuracy of the model is improved from 0.516 to 0.537. The inspection effect is shown in the Figure 5 below.

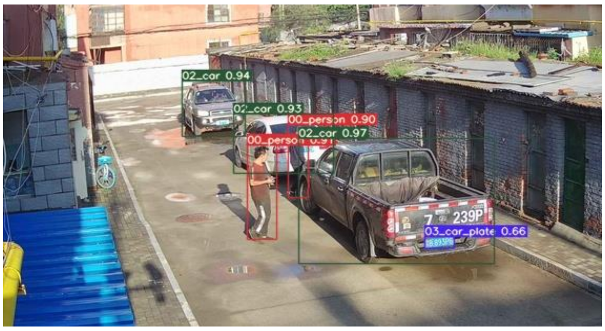

Fig. 5 detection results

5. Concluding

In order to improve the detection accuracy of people, faces, cars and license plates in the city, this paper chooses YOLOv7 as the main framework in its CNeB structure is added to Backbone to reduce model size and improve feature extraction capability, and Swin Transformer Block is added to head to ensure model parameters and improve detection accuracy. The experiment shows that the improved YOLOv7 has good performance in all aspects.

Acknowledgment

This work was supported by “Shenzhen Softsz Co. Ltd., China".

References

1. T. Zhang et al., "HOG-ShipCLSNet: A Novel Deep Learning Network With HOG Feature Fusion for SAR Ship Classification," in IEEE Transactions on Geoscience and Remote Sensing, vol. 60, pp. 1-22, 2022, Art no. 5210322, doi: 10.1109/TGRS.2021.3082759.
2. X. Xiao and X. Tian, "Research on Reference Target Detection of Deep Learning Framework Faster-RCNN," 2021 5th Annual International Conference on Data Science and Business Analytics (ICDSBA), 2021, pp. 41-44, doi: 10.1109/ICDSBA53075.2021.00017
3. V. R. A G, M. N and D. G, "Helmet Detection using Single Shot Detector (SSD)," 2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC), 2021, pp. 1241-1244, doi: 10.1109/ICESC51422.2021.9532985.
4. A. Albayrak and M. S. Özerdem, "Gas Cylinder Detection Using Deep Learning Based YOLOv5 Object Detection Method," 2022 7th International Conference on Computer Science and Engineering (UBMK), 2022, pp. 434-437, doi: 10.1109/UBMK55850.2022.9919478.
5. S. Liu, Y. Wang, Q. Yu, H. Liu and Z. Peng, "CEAM-YOLOv7: Improved YOLOv7 Based on Channel Expansion and Attention Mechanism for Driver Distraction Behavior Detection," in IEEE Access, vol. 10, pp. 129116-129124, 2022, doi: 10.1109/ACCESS.2022.3228331.
6. A. Yueyuan and W. Hong, "Swin Transformer Combined with Convolutional Encoder For CephalometricLandmarks Detection," 2021 18th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2021, pp. 184-187, doi: 10.1109/ICCWAMTIP53232.2021.9674147.
7. L. Feng and Y. Jia, "Traffic sign recognition based on YOLOX in extreme weather," 2022 Global Conference on Robotics, Artificial Intelligence and Information Technology (GCRAIT), 2022, pp. 299-303, doi: 10.1109/GCRAIT55928.2022.00070.

Authors Introduction

Mr. Zhihui Chen

He received his bachelor's degree from the school of electronic information and automation of Tianjin University of science and technology in 2021. He is acquiring for his master's degree at Tianjin University of science and technology.

Prof. Xiaoyan Chen

She, professor of Tianjin University of Science and Technology, graduated from Tianjin University with PH.D (2009), worked as a Post-doctor at Tianjin University (2009.5-2015.5).

Mr. Keying Ren

He is an Master of Control Science and Engineering, Tianjin University of Science and Technolog. The research topic is target detection and tracking based on deep learning.