A pedestrian detection algorithm for low light and dense crowd Based on improved YOLO algorithm

Yitong Mao*

China University of Mining and Technology-Beijing, Beijing, China

Abstract. The real-time pedestrian detection algorithm requires the model to be lightweight and robust. At the same time, the pedestrian object detection problem has the characteristics of aerial view Angle shooting, object overlap and weak light, etc. In order to design a more robust real-time detection model in weak light and crowded scene, this paper based on YOLO, raised a more efficient convolutional network. The experimental results show that, compared with YOLOX Network, the improved YOLO Network has a better detection effect in the lack of light scene and dense crowd scene, has a 5.0% advantage over YOLOX-s for pedestrians AP index, and has a 44.2% advantage over YOLOX-s for fps index.

Keywords: Weak light detection, Crowd detection, YOLO Network, Pedestrian detection.

1 Introduction

In variety of realistic application of multi-object detection, accurate statistics of traffic flow play a crucial role in the self-driving system and intelligence transportation, which includes the real time detection task on the intersection. The traditional design detection algorithms mainly adopted the manually designed Histogram of Oriented Gradients, Aggregated Channel Features (HOG)[1], Wavelet Transform (Haar)[2], and Aggregated Channel Features (ACF)[3] are used to extract Features of pedestrians. And use support vector machines (SVM)[4], Adaptive Boosting (Adaboost)[5] and other classifiers to judge whether there is a object in the region. This kind of detection method is based on sliding window region. Due to the high time complexity and the lack of pertinacity, the traditional algorithms cannot achieve satisfactory results in terms of both speed and accuracy.

With the popularity of convolutional neural networks (CNNs), the pedestrian detection task has been profoundly developed. Currently, there are based on candidate regions two stage target detection algorithms Faster-RCNN (Faster region-based Convolutional Neural Networks)[6], based on the regressive one-stage target detection algorithm SSD (Single Shot Detection)[7], and YOLO(You Only Look Once)[8], etc. As the classic algorithm of one-stage target Detection, YOLO outputs Location Prediction and Class Confidence at one time, which greatly improves the performance of real-time target detection. For real time,

* Corresponding author: 1810480225@student.cumtb.edu.cn

YOLO also offers lightweight models such as YOLOv4-Tiny and YOLOv5s. This is likely to be affordable on resource constrained platforms such as mobile devices, wearables or Internet of Things (IoT) devices.

However, there are still two difficulties in pedestrian detection which have not been completely solved by existing algorithms. First, in the face of complex and changeable natural environment, it is difficult to achieve satisfactory results either through visible light information or infrared information. Such as in fog, rain, fog, weak light, pedestrians in visible light image target invisible or ambiguous situations, and although the infrared image can improve the quality of such a case the image [9], but because of its rich cannot describe the pedestrian characteristics of contour and the characteristics of color information, in a crowded scenarios can lead to more residual and checked by mistake. The second is the multi-scale variation of pedestrian targets and the detection of small pedestrian targets. When there are pedestrians of different sizes in the scene at the same time, or there are individuals of small size and low resolution in the pedestrian object, the pedestrian features extracted by the detector will be more vulnerable to the noise interference in the environmental background, which will lead to missed detection and false detection and bring great challenges to the accuracy of the detection results.

Based on the aforementioned study, this paper focuses on the difficult problems of weak light and dense pedestrian target detection in the traffic scene. Through the open data set, the latest real-time target detection model is trained. The performance of different models under weak light and dense crowd was compared and analyzed. The experimental results show that compared with similar models such as YOLOX[13], the improved YOLO5s[14] is more robust under weak light scene and dense crowd scene.

2 Pedestrians detect challenge and preliminary

2.1 Problem statement

This chapter gives a brief introduction to original YOLO algorithm, including its principle of classification and target box regression, darknet53 for feature extraction, and analyzes the challenges of pedestrian detection task when applying YOLO.

2.2 Problem analysis

Different from RCNN, YOLO series is a one-stage target detector, which does not have a separate area recommendation network (RPN) and relies on anchor frames of different scales. Based on this feature, YOLO algorithm is more suitable for the detection of regular objects with similar length and width, while pedestrian targets have problems such as multi-scale, multi-angle, and poor image quality. It causes instability and difficulty in detection. This paper mainly discusses the pedestrian detection in low light and dense crowd.

Fig. 1. Object Density.

Fig. 2. Object occlusion.

According to research, the difficulties faced by pedestrian detection are as follows:

Table 1. Difficulties of pedestrians detection.

View	CCTV is top-mounted, so pedestrians at the picture in depression angle.
Crowd	It can be very crowded at certain times.
Background	Billboards, tree stump, signal lights, cars and other obstructs will affect the final test results.
Light	The changeable weather outside, and low light at night make it difficult to detect pedestrians at the intersection.
Image quality	Video from CCTVs has poor frame resolution and motion blur.

2.3 YOLO algorithm preliminary

YOLO scales the image to 640*640*3 and then unpacks the image into a grid of size. Each grid region is responsible for detecting up to one target. Feed images into backbone feature extracting network. When the center point of the object to be detected falls into a grid, the grid will predict B predicted boxes for it. If there are objects of class C to be measured, the dimension of the border output vector is C+5, where 5 represents T= (x, y, w, h, S), (x, y) represents the center point coordinate of the prediction box, (w, h) represents the width and height of it, and S represents the confidence degree of it, which is a number ranging from 0 to 1. The calculation method is as follows:

$$S = P(C_i) \times P(O) \times I \tag{1}$$

When there is an object in the prediction box, $P(O) = 1$, otherwise, $P(O) = 0$, represents the IoU of predicted bounding box and Ground truth bounding box, and$P(C_i)$represents the probability that the object belongs to the ith object in class C when the object exists.

Then the predicted bounding boxes were graded according to the confidence threshold, and repeated with the same target were deduplicated by the NMS algorithm.

2.4 YOLO algorithm preliminary

Darknet53 is the backbone network for feature extraction and test proposed by Redmon J in YOLOv3. The basic unit of the network is composed of the convolution layer, Batch Normolization and Leaky ReLU activation function.

In order to avoid model degradation caused by deepening network layer, Darknet53 uses residual network for reference and sets 5 residual blocks in the network: {Block1, Block2, Block3, Block4, Block5}, each residual blocks are denoted as Res N, which contains N residual units. Network parameters are shown in Figure. 4:

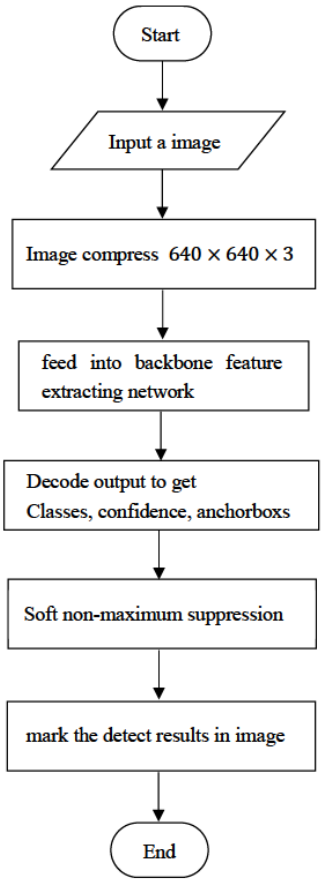

Fig. 3. YOLO inference flow chart.

		Type	Filters	Size	Output
		Convolutional	32	3×3/1	416×416×32
		Convolutional	64	3×3/2	208×208×64
Res1:	1×	Convolutional	32	1×1/1	208×208×32
		Convolutional	64	3×3/1	208×208×64
		Residual			208×208×64
		Convolutional	128	3×3/2	104×104×128
Res2:	2×	Convolutional	64	1×1/1	104×104×64
		Convolutional	128	3×3/1	104×104×128
		Residual			104×104×128
		Convolutional	256	3×3/2	52×52×256
Res8:	8×	Convolutional	128	1×1/1	52×52×128
		Convolutional	256	3×3/1	52×52×256
		Residual			52×52×256
		Convolutional	512	3×3/2	26×26×512
Res8:	8×	Convolutional	256	1×1/1	26×26×256
		Convolutional	512	3×3/1	26×26×512
		Residual			26×26×512
		Convolutional	1024	3×3/2	13×13×1024
Res4:	4×	Convolutional	512	1×1/1	13×13×512
		Convolutional	1024	3×3/1	13×13×1024
		Residual			13×13×1024

Fig. 4. Structure of darknet53.

2.5 Soft non-maximum suppression (soft-NMS)

YOLO backbone feature extracts multiple overlapping boxes from the same object. In order to remove redundant boxes, Traditional NMS calculates the IoU of all boxes, then rank each box with confidence. Then iterate over the boxes list, and delete the highly overlapping boxes each time while holding the ones with the highest score.

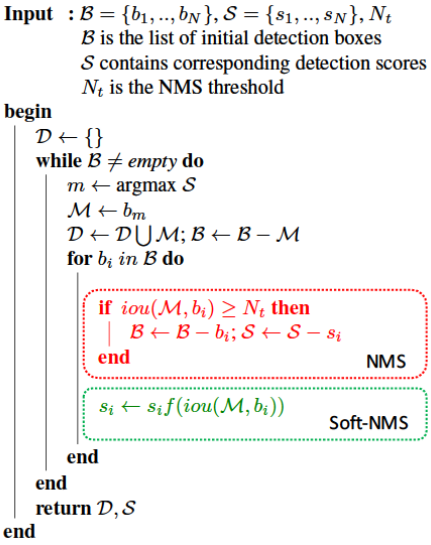

Fig. 5. NMS inference[12].

b_i is the BBox box to be processed, B is the set of BBox boxes to be processed, s_i is the b_i box update score, N_t is the hyperparameter of NMS, the D set is used to put the final BBox, f is the reset function of confidence score. The larger the IoU of b_i and M, the score of b_i is greater and the s_i is more largely dropping.

The method of Traditional NMS filtering redundancy is to directly delete all boxes whose IoU is larger than the threshold, resulting in information loss. In the algorithm execution, soft NMS no longer simply deletes boxes whose IoU is greater than the threshold, but uses the reset function to reduce the score of the box. The linear weighted reset function is shown as follows:

$$s_i = \begin{cases} s_i, & IoU(M, b_i) < N_t \\ s_i\left(1 - IoU(M, b_i)\right), & IoU(M, b_i) > N_t \end{cases} \tag{2}$$

3 Method

3.1 YOLOv5 algorithm introduction

YOLOv5 proposed by Ultralytics LLC is an improved version based on YOLOv4, and it is currently an excellent one-stage detection network in terms of accuracy and detection speed[10]. YOLOv5 provides s, m, l, x network models of different sizes, which can be selected according to different accuracy and real-time requirements. The network model is divided into four parts: Input, Backbone, Neck (multi-scale feature fusion module) and Prediction. The network structure is shown in Figure. 6.

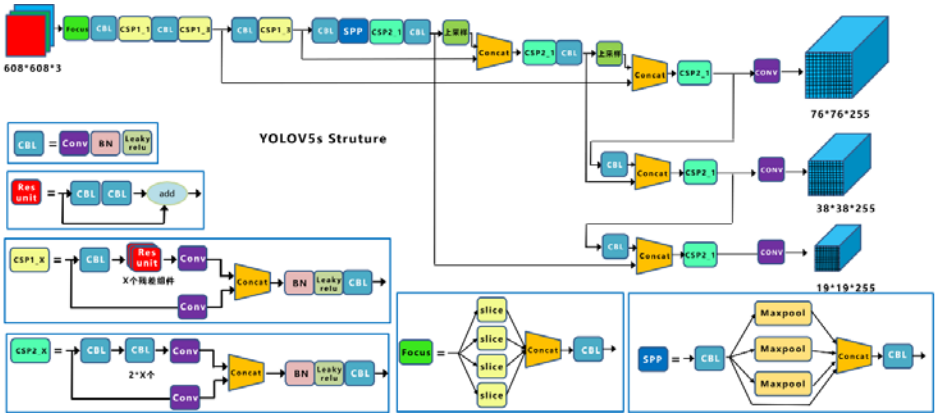

Fig. 6. Network structure diagram of YOLOv5-s.

3.2 Input

The Input side includes Mosaic data enhancement, adaptive Anchor boxes calculation and adaptive image scaling. Mosaic data enhancement is improved in CutMix enhancement method, using random cutting, random scaling and four images in random distribution to put four pictures together, achieving the result of the rich data set. it can enhance the robustness of the network, significantly improve the network training speed, reduce the training phase of the demand for memory. Adaptive Anchor Box calculation sets initial Anchor boxes for different data sets. According to experience, feasible super parameters are taken. The anchor boxes parameters of the three grids are [116,90,156,198,373,326]; [30,61,62,45,59,119]; [10,13,16,30,33,23]. Adaptive image scaling is to scale images uniformly to a uniform size.

3.3 Backnone

Backbone includes the Focus structure and Cross Stage Partial Network (CSPNet) structure. Focus structure slice the input image to obtain the feature map.YOLOv5 draws on the CSP structure of YOLOv4backbone network and designs CSP1_X and CSP2_X CSP structures, as shown in Fig. 6. Backbone uses CSP1_X module. The CSP2_X module is used in Neck. Concat method can be used to merge the shallow features containing physical information with the deep features containing semantic information.

3.4 Neck

Neck uses FPN (Feature Pyramid Networks) +PAN (Pyramid Attention Network) structure, as shown in FIG Neck Structure diagram, FPN transmits and merges high-level feature information from top to bottom through upsampling to convey strong semantic features, while PAN is a feature pyramid from bottom to top to convey strong positioning features. The simultaneous use of the two features strengthens the ability of network feature fusion.

3.5 Prediction

Prediction includes bounding box loss function and NMS. YOLOv5 uses GIoU_Loss as the loss function of Bounding box, which effectively solves the problem that IoU cannot be calculated when BBox does not coincide, and improves the speed and accuracy of

prediction BBox regression. Weighted NMS is used in the detection phase to enhance the recognition capability of multiple targets and overlapping targets.

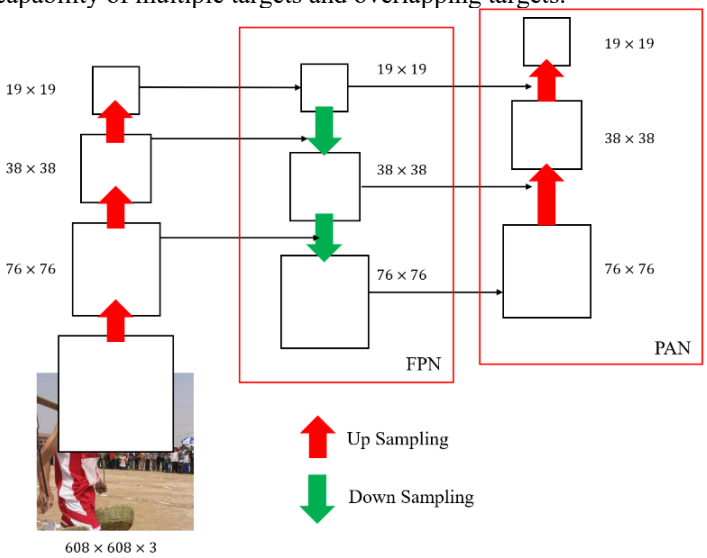

Fig. 7. Structure of Neck layer.

4 Experiment's result and evaluation

4.1 Experiment's environment

The hardware configuration of the experiment in this paper is a Laptop with AMD Ryzen 9 5900HX processor, NVIDIA RTX3070 Laptop graphics card and 32GB RAM. The software environment is Windows10 system, cuda11.1 cudnn8.0, Python3.8.8 pytorch1.8.1 framework.

4.2 Dataset and evaluation criteria

This paper uses WilderPerson to expose the data set. It includes 13,382 images and label about 400K annotations with various kinds of occlusions. The WiderPerson dataset randomly select 8000/1000/4382 images as training, validation and testing subsets. And putting people into 5 different labels: put riders, partially visible persons, ignore regions, crowd. The data set includes a variety of common life and traffic scenes such as campuses and streets, as well as a rich variety of weather environments during the day and night, which can represent complex conditions such as multi-scale, occlusion and insufficient light conditions.

4.3 Training parameter Settings

Stochastic gradient descent SGD[11] was used in the training process, and momentum was 0.937. Cosine annealing learning rate decay is used, initial learning rate is 0.01, weight_decay is 0.0005, warmup_epochs is 3.0, warmup_momentum is 0.8.epoch is 120 times, Batch size is 16, images size is $640 \times 640 \times 3$.

Before the evaluation of the model, it is necessary to select appropriate model evaluation indexes. In order to reflect the comprehensive performance of the model, this paper adopts Mean Average Precision (mAP) index to reflect the accuracy of target detection. Frame per second (FPS) reflects the inference speed of the model. Formula (3) (4) and (5) are the calculation formula of mAP.

$$Precision = \frac{TP}{TP+FP}(4-1) \tag{3}$$

$$Recall = \frac{TP}{TP+FN}(4-2) \tag{4}$$

$$AP = \int_0^1 Precision\, d(Recall)(4-3) \tag{5}$$

TP is an example that the classifier considers to be a positive sample and is indeed a positive sample, FP is an example that the classifier considers to be a positive sample but is not actually a positive sample, FN is an example that the classifier considers to be a negative sample but is not actually a negative sample, and AP represents the detection accuracy of a single category.

First, the IoU threshold of TP, FN and FP was set to 0.95, and the IoU of the predicted boxes and ground truth of a single category was calculated one by one, thus the confidence corresponding to positive and negative samples could be obtained.

When different confidence degrees are taken, a series of Precision and Recall can be obtained. A line can be obtained after it is plotted into the two-dimensional coordinate axis recollect-precision. The area enclosed by the line and the coordinate axis is AP, and the corresponding AP of each class is calculated and the average value is calculated to obtain mAP.

4.4 Model training and comprehensive performance analysis

In order to objectively select an appropriate detection model, this paper carried out 120 rounds of iterative training on the latest SOTA model YOLOX and Advanced YOLO under the same parameter setting and based on the same data set, and selected the real-time lightweight version of the model for evaluation. The training results are as follows:

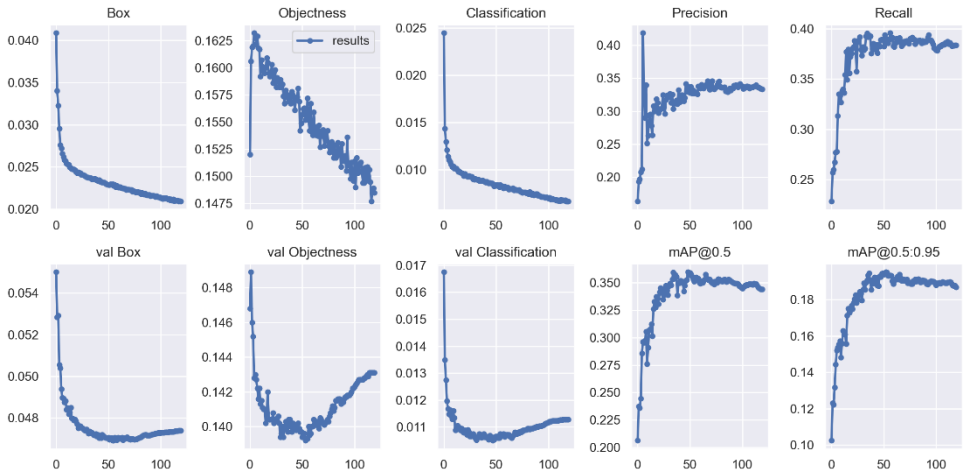

Fig. 8. Results of Advanced YOLO training.

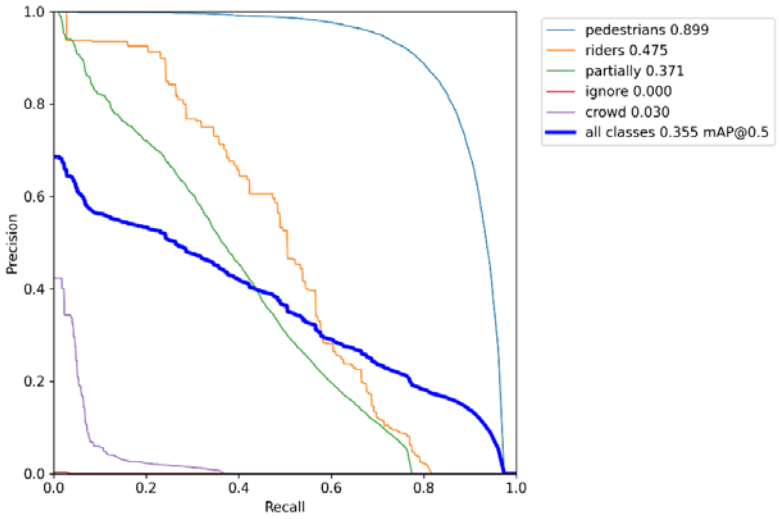

Fig. 9. Advanced YOLO training results.

Table 2. Test result of detection.

	Image size	mAP	Pedestrians AP	Riders AP	Partially AP	Params	Gflops	fps
Advanced YOLO	608 × 608	0.355	0.899	0.475	0.371	7.3M	17.0	94.3
YOLOX-s	640 × 640	0.368	0.856	0.505	0.327	8.94M	26.64	65.4

In terms of Pedestrians AP index, Advanced YOLO has a 5.0% advantage over YOLOX-s, and in FPS index, Advanced YOLO has a 44.2% advantage over YOLOX-s. In conclusion, the Advanced YOLO network model is more effective for pedestrian detection tasks.

4.5 Comparison of detection effects under weak light and dense population

In order to intuitively show the advantages of Advanced YOLO in personnel detection efficiency in low-light and crowded scenes, this paper selected typical images in low-light environment and crowded scenes as comparison in the data set.

Fig. 10. Weak light.

Fig. 11. Dense object.

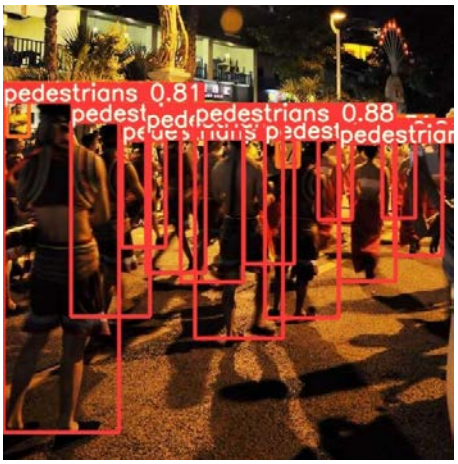

Fig. 12. Advanced YOLO performance.

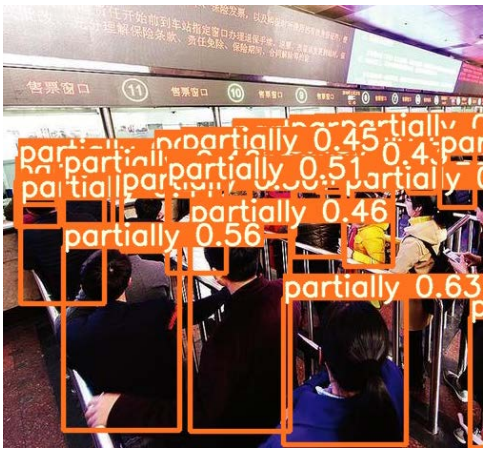

Fig. 13. Advanced YOLO performance.

Fig. 14. YOLOX-s performance.

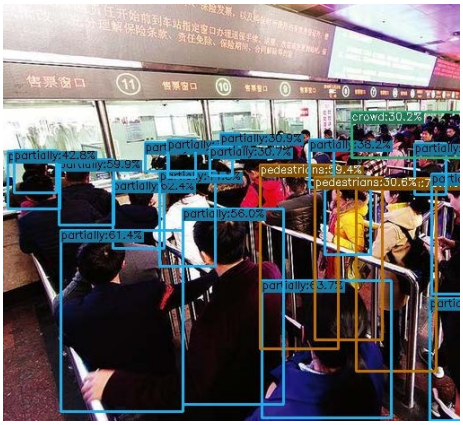

Fig. 15. YOLOX-s performance.

In Figure.10, the illumination condition is poor, and the target features are not obvious and difficult to recognize. In Figure.11, the targets are dense and the occlusion is serious. Figure. 12 and Figure. 13 are the pedestrian detection results based on Advanced YOLO; Figure. 14 and Figure. 15 are the pedestrian detection results based on YOLOX-s.

5 Conclusion

This paper compares the performance of two YOLO based target detection algorithms in pedestrian detection tasks through experiments. In the lack of lighting scenes and dense crowd scenes, Advanced YOLO has better detection accuracy than YOLOX s. Considering that the model in actual application scenarios has higher real-time requirements,

Advanced YOLO is obviously better than YOLOX-s in terms of speed, so Advanced YOLO algorithm is more conducive to the application of pedestrian detection in related fields.

During this paper design process, I would thank my school for giving me this opportunity to study, and my teacher Wu gave me carful guidance. My classmate Zelin Wang had a lot of close discussion with me and gave me many valuable suggestions. At last, I would also like to thank my family and

friends, it is because of their silent support behind me that I was able to successfully complete the paper.

References

1. Dalal N, Triggs B. Histograms of oriented radients for human detection[C]//IEEE Conference of Computer Vision and Pattern Recognition, 2005, 1:886-893.
2. Viola Paul, Jones M J. Robust teal-time face detection[J]. Journal of Computer Vision, 2004,57(2):137-154.
3. Dollár Piotr, Wojek Christian, Schiele Bernt, et al. Pedestrian detection:an evaluation of the state of the art[C]//IEEE Transactions on Pattern Analysis and Machine Intelligence, 2011, 34: 743-61(10.1109/ TPAMI.2011.155).
4. CHEN P H, LIN C J, Schölkopf B. A tutorial on v-support vector machines[J]. Appl. Stoch. Models. Bus. Ind., 2005, 21,: 111-136.
5. Y Freund, R E Schapire. Adecision-theoretic generalization of on-line learning and an application to boosting[J].Journal of Computer and SystemSciences, 1997, 55(1): 119-139.
6. Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]. neural information processing systems, 2015: 91-99.
7. LIU W, Anguelov D, et al. SSD: single shot multibox detector[C]//Proceedings of European Conference on Computer Vision, 2016: 21-37.
8. Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of CVPR, 2015: 779-788.
9. Ge J, Luo Y, Tei G. Real-Time Pedestrian Detection and Tracking at Nighttime for Driver-Assistance Systems[J]. IEEE Transactions on Intelligent Transportation Systems, 2009, 10(2): 283-298.
10. Wu D, Lv S, Jiang M, et al. Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection f apple flowers in natural environments[J]. Computers and Electronics in Agriculture,2020,178(5):174-178.
11. BOTTOU L. Stochastic gradient descent tricks[M]. Neural Networks: Tricks of the Trade. 2nd ed. Berlin Germany: Springer, 2012: 421–436. doi: 10.1007/978-3-642-35289-8_25.
12. Navaneeth Bodla, Bharat Singh, Rama Chellappa Larry, S. Davis. Improving Object Detection With One Line of Code[C]. arXiv preprint arXiv:1704.04503, 2017.
13. Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, Jian Sun. YOLOX: Exceeding YOLO Series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.
14. REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time object detection [C]// IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas: IEEE, 2016: 779-788.