2023 IEEE International Conference on Image Processing and Computer Applications (ICIPCA) | 979-8-3503-1467-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICIPCA59209.2023.10257755

Research on Multi-class Road Obstacle Recognition and Decision Based on YOLOP Combined YOLOV5 Algorithm

Ganrong Dong[#]
Jiaxing University
Jiaxing, China
dgr2358312156@163.com

Tiancong Han[#]
Tiangong University
Tianjin, China
hantiancong2002@163.com

Chenxiao Feng[*,#]
Tianjin University of Science and Technology
Tianjin, China
f506609@163.com
[#]These authors contributed equally.

***Abstract*—The paper proposes a visual real-time sensing system. It carries YOLOv5 and YOLOP models and can be carried on the vehicle embedded device. The system simultaneously performs vehicle, pedestrian, traffic light, traffic sign detection, driveable area segmentation and lane detection. The paper not only considers the high accuracy, but also considers the computational cost of running on the vehicle when selecting the model to get the sensing information. The paper chooses YOLOP, which enables FPS to reach 23 (after TRT acceleration) on embedded devices. YOLOv5 also has extremely fast execution speed and accuracy on embedded devices. The combination of the two well meets the speed and accuracy required by the automatic driving decision control. And upload all sensing information to the embedded device. At the same time, a priority decision system for automatic driving is also proposed: pedestrian and vehicle distance information(divided into high, medium and low risk) > information of signal lights and traffic signs > distance information of lane line > information of driving area. Finally, this paper combines perception information with decision control and carries out visual output. Through experiments, it can be seen that the decision has achieved good results and the automatic driving function can be realized in most Chinese real street scenes.**

Keywords*—yolop, yolov5, decision priority, distance risk system

I. Introduction

Autonomous driving is a popular trend at present, but its development still has some problems. For example, the degree of intelligence is not high, and the recognition complexity of dynamic and solid roadblocks is relatively high. This is a difficult bottleneck for autonomous driving. Therefore, for the future development of more intelligent and faster automatic driving, The paper carried out the research on roadblock recognition.

At present, many scholars have studied roadblock recognition. For lane detection: Peng Hong et al. proposes a lane detection algorithm which combines the information of road structure model with extended Kalman filter (EKF). However, when there is a large length of interference line parallel to the lane line in the road at the same time, the recognition error will occur. For traffic sign detection: Liu Shunmin proposed a traffic sign target detection method based on SSD algorithm by conducting target detection on traffic signs. For traffic light detection: Using deep learning and OpenCV's vision library, Yu Chuli and Zhu Qiang solved the problem of a self-driving vehicle recognizing traffic lights during the day. For pedestrian detection: Wang Chenglong et al. adopted the pedestrian detection method based on HOG+SVM to establish the pedestrian detection model. The measurement accuracy reached 90.6%. But their processing time for each frame of video image is not ideal. Du Xuefeng et al. designed a forward vehicle recognition system based on OpenCV technology, and used Adaboost algorithm to train the cascade classifier to recognize vehicles.

However, they only identify single-class roadblocks, and the existing algorithms are not universal. Therefore, the paper carries out the research in this paper to identify multi-class roadblocks and conduct a more systematic and comprehensive study. This can reduce the difficulty of algorithm integration and realize a lightweight multi-class road block recognition.

II. Dataset

A. Introduction to datasets

Traffic light dataset S2TLD was published by Shanghai Jiao Tong University. It contains four categories with nearly 10,000 instances (including red, yellow, green, off). Scenes cover sunny, rainy days, morning, noon, evening and night environments, as well as a variety of classic road scenes [1].

The traffic sign data set CCTSDB was produced by Teacher Zhang Jianming from Changsha University of Science and Technology. A total of 13,859 data sets were randomly selected, including nearly 18,000 examples, which are divided into three categories: compulsory, prohibition and warning [2][3][4].

Vehicle detection, driveable zone division, lane line data set BDD100K covers different time, different weather conditions (including sunny, cloudy and rainy days, as well as different times of day and night) and driving scenarios. Road target detection is marked with 2D boundary boxes on 100,000 pictures of various vehicles, people and traffic signs. Driveable areas are complex driveable decisions learned from 100,000 images. Lane markings are multiple lane markings on 100,000 driving guide images. Solid lines, dotted lines, double lines and single lines are marked in the lane marking pictures [5].

Pedestrian dataset nuScenes is a public large-scale dataset for autonomous driving developed by the Motional team. It collects data from different places, marking them at different locations, weather conditions, vehicle types, vegetation, roads, as well as traffic conditions. The complete dataset includes about 1.4 million camera images. The data set in this paper comes from more than 9000 images of pedestrians separately extracted in part1 of nuScenes [6][7].

B. Reasons for choosing these data sets

S2TLD data set is selected because it has a better target for environment and road scenes, so that the model trained by it has stronger anti-interference and higher recognition accuracy. The reason for selecting CCTSDB is that if a sign is sorted into a data set, the recognition rate and recall rate will be very low due to the small number of data sets. This has a high probability of missing the mark. Therefore, only three categories are selected, so that the recognition accuracy and recall rate of the trained model will be much higher. Many of the pedestrians in this data are small targets that are extremely difficult to identify. nuScenes was selected because it can make the model more anti-jamming, so as to identify the target more effectively in the real scene. Comprehensive scene test data are from CCTSDB and S2TLD. This paper realizes decision control based on Chinese traffic rules, so all tests are from real driving scenes in China. CCTSDB and S2TLD just fit this characteristic.

III. Method

A. Introduction and result analysis of YOLOP

YOLOP segments three tasks simultaneously through a network. YOLOP builds a multi-task that can share information between multiple tasks. The overall schematic structure of the YOLOP network is shown in the Fig. 1 below. [11]

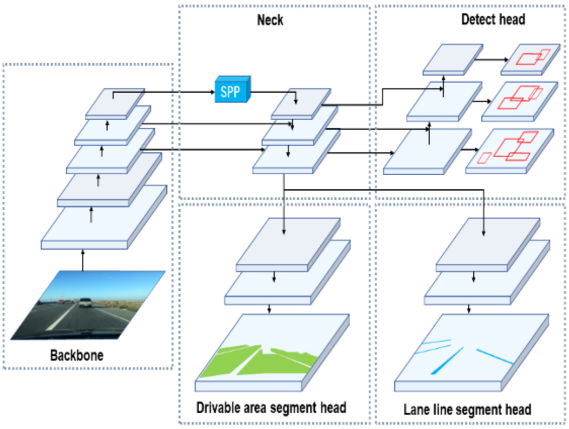

Fig. 1. The YOLOP network model

1) Model introduction

The network consists of one Encoder and three separated Decoders. Among them, Encoder includes two parts: backbone and neck. Three Decoders complete car detection, lane detection, drivable area segmentation task. The following parts are introduced one by one.

a) Encorder

This is the part shared by the three tasks in the network. Backbone network adopts CSP-Darknet structure to extract the features of input images.The Neck network consists of space pyramid pool (SPP) module and feature pyramid network (FPN) module. SPP modules generate and fuse features of different scales. FPN modules fuse features of different semantic levels.

b) Decorders

The YOLOP contains three decoders for the three tasks.

Detect Head: YOLOP adopts the multi-scale detection technology based on the anchor frame Anchor. This structure consists of "path aggregation network". YOLOP combines the semantic features of top-down transmission of FPN and the image features of bottom-up transmission of PAN to obtain better feature fusion effect. The fusion feature map obtained from this is used for detection.

Driable Area Seg Head & Lane Det Head: These two parts adopt the same network structure.YOLOP feeds the underlying size feature of the FPN into the split branch. It is designed with a segmented branch that outputs the feature size of by three upsampling processes. The three branches represent the probability that each pixel is the driving area/lane line/background.

c) Loss Function

Vehicles Det: The detection task consists of three parts. Detection losses ($\mathrm{L_{det}}$) includes classification losses, target losses, and frame losses. The weighted sum of these three losses is shown in the following equation .

$$L_{det} = \alpha_1 L_{class} + \alpha_2 L_{obj} + \alpha_3 L_{box} \tag{1}$$

Focal loss was adopted for $\mathrm{L_{class}}$ and $\mathrm{L_{obj}}$.

CloU loss was adopted for $\mathrm{L_{box}}$.

Lane/AreaSeg: Both split tasks adopt .Lane line segmentation has additional IoU losses due to its effectiveness in predicting reserve classes.

d) Conclusion

In conclusion, the expressions for all Loss are as follows:

$$L_{all} = \gamma_1 L_{det} + \gamma_2 L_{da-seg} + \gamma_3 L_{ll-seg} \tag{2}$$

$\alpha_{1,2,3}$ and$\gamma_{1,2,3}$ are both adjustable parameters to balance the loss of the various parts.

2) Model outcome analysis

TABLE I. Car Detection Result

Network	Recall(%)	mAP50(%)	Speed(fps)
Faster R-CNN	77.2	55.6	5.3
DLT-Net	89.4	60.2	8.6
YOLOP	89.2	76.5	41

TABLE II. Drivable Area Seg Result

Network	mIoU(%)	Speed(fps)
DLT-Net	71.3	9.3
PSPNet	89.6	11.1
YOLOP	89.2	41

TABLE III. Lane Detection result

Network	IoU(%)	Accuracy(%)
SCNN	15.8	35.79
Enet-SAD	16.0	36.56
YOLOP	26.2	70.5

The data in TABLE I,II,III above are quoted from YOLOP[11]. By comparing the effect of YOLOP and mainstream target detection models in TABLE I, we found that YOLOP had a significant improvement in recall rate and mAP50. More importantly, its execution speed can reach 41fps, which makes it competent for decision-making control in autonomous driving scenarios. Compared with mainstream lane detection networks and mainstream segmentation networks in TABLE II and TABLE III, YOLOP also outperforms other networks.

B. Introduction and result analysis of YOLOPv5

1) Model introduction

Compared with other classical target detection networks, YOLOv5 network has achieved a good balance between detection accuracy and detection speed, making it one of the most advanced target detection networks at present. The YOLOv5 network can be divided into four versions according to the size of the model. The YOLOv5s model is the smallest and the fastest. The YOLOv5s network structure is consistent with other types. The network result of YOLOV5 is shown in the following Fig2[10].

2) Model Structure

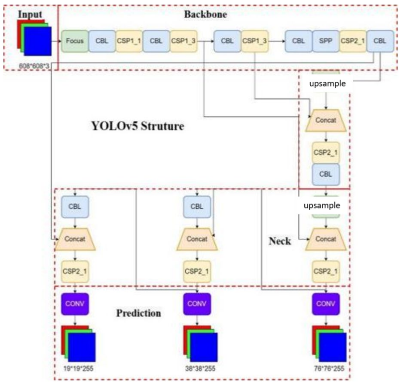

Fig. 2. The YOLOv5 structure

The network structure of YOLOv5 algorithm includes four parts: input end (Input), backbone network (Backbone), multi-scale feature fusion module (Neck) and prediction end (Prediction), as shown in the figure 2 [8].

Mosaic data enhancement mechanism was adopted at the input end of YOLOv5, which improved the training speed of the model and the accuracy of the network. The speed of YOLOv5 has been greatly improved by the addition of an adaptive zooming mechanism. At the same time, YOLOv5 also adds adaptive anchor frame calculation. During each training, the position of the optimal anchor frame can be calculated adaptively according to the name of the data set and marked[9].

Bockbone is the backbone network of YOLOv5 network, including Focus structure and CSP structure. This structure can ensure that floating point computation is reduced on the premise of no missing information, thus improving the reasoning speed.

Neck network module is mainly used to generate feature pyramid. The network layer is composed of a series of mixed and combined image features to enhance the diversity of features.

The prediction terminal is the final inspection part of the YOLOv5 model. It classifies the target and returns the boundary box. The non-maximum suppression method is used to preserve the prediction box with the highest confidence, so as to complete the whole process of target detection.

C. Introduction of the principle of single visual ranging

The similar triangle method is used for the ranging of monocular camera. The method diagram is as follows:

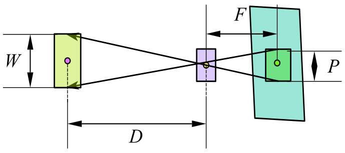

Fig. 3. Single visual distance diagram

As shown in Fig. 3, The formula for the real distance between the object and the camera is as follows:

$$D = (F * W)/P \tag{3}$$

D is the distance from the target to the camera.F is the focal length of a given camera.W is the true height of the target.P refers to the width of pixels in the x direction or the height of pixels in the y direction occupied by the target in the image (obtained from the target detection results of YOLOv5). By setting the actual height(Unit:inches) of targets such as pedestrians and cars, this formula can be used to calculate the distance of targets ahead.

D. Model summary

This paper uses the model trained by the YOLOP "End-to-end" training method to make decisions. Instead of modifying and adjusting the YOLOP model, this paper directly applies the YOLOP open source model.

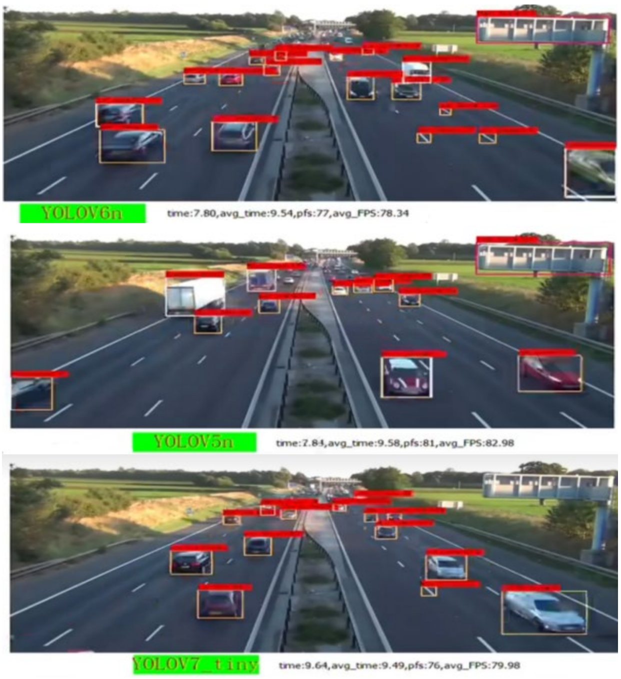
YOLOV6n
time:7.80,avg_time:9.54,pfs:77,avg_FPS:78.34
YOLOV5n
time:7.84,avg_time:9.58,pfs:81,avg_FPS:82.98
YOLOV7_tiny
time:9.64,avg_time:9.49,pfs:76,avg_FPS:79.98

Fig. 4. Multi-series model detection comparison

As shown in Fig. 4, The paper compares YOLOv5 with YOLOv6 and YOLOv7tiny horizontally, and find that YOLOv5 can achieve more stable recognition (less false detection and missing detection). It's much more balanced and stable than v6 and v7 (which are prone to false detection). Its model size is small and running speed is the fastest, suitable for embedded device target detection. To sum up, YOLOv5s model is selected for target detection in this paper.

This paper compares the traditional lane detection method with YOLOP. Conventional lane detection works well when the lane is straight and clear, but it doesn't work well in more complex environments (such as night). Therefore, YOLOP needs to be used for lane detection in actual scenarios. However, due to the poor effect of YOLOP multi-target detection, misjudgment and missing judgment are very easy to occur. Therefore, YOLOv5s, which is also suitable for embedded vehicle equipment, is selected for pedestrian, traffic lights and traffic signs target detection. YOLOv5s was used to assist YOLOP in decision control.

In terms of ranging, we choose the similar triangle method. When the focal length of our on-board camera is known, the error can be reduced to very small. The paper compares the monocular depth estimation range with the similar triangle method and the comparison result is shown in Fig. 5. The paper finds that the measurement results of monocular depth estimation are not as accurate as those of the similar triangle method when the focal length is fixed. And the most fatal shortcoming of depth estimation ranging is its extremely slow execution. Usually, it can only process 3-5 images per second, which is not suitable for decision control in real scenes.

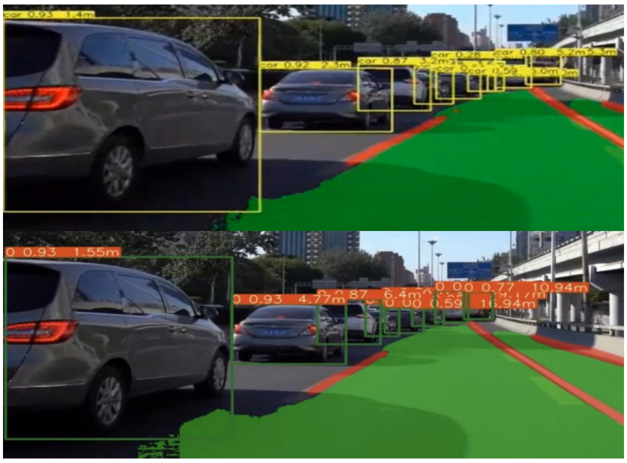

Fig. 5. Schematic diagram of similar triangle ranging method

The upper picture is measured by depth estimation; The next part is measured by our method and the following picture is measured by our method.

IV. Experiment and Result

A. Object Detection Model

The model that detects and recognizes traffic lights and judges red, green and yellow lights and no lights uses 4,636 images from 720*1280 pixels in the above data set [1].

The model for detecting and identifying pedestrians uses 9,095 images with 1600*900 pixels in Part1 of the data set [2] above.

13,859 images from the above data set were used in the model for detecting, recognizing and classifying traffic signs.

The model weight files trained by bdd100k are used for lane recognition, driveable area division and vehicle detection in this paper.

B. Experimental Procedure

First, python is used to convert the above three data set label formats, converting.xml format to yolo format or text format. Then, data set in Yolov5 format was prepared according to the 9:1 ratio of training set: test set. They were then trained in the cloud on the AutoDL website. The local host communicates with the server to transfer the data set, Yolov5s weight file and code to the AutoDL cloud disk, and then extract the data set, weight file and code from the cloud disk at the server terminal. Finally, configure a virtual environment for running Yolov5 code and implement the training model on the server.

The server configuration and model training parameters is shown in TABLE IV and V

TABLE IV. Training server parameters

GPU	RTX 3080
CPU	12 vCPU Intel(R) Xeon(R) Platinum 8255C CPU @ 2.50GHz
Memory	10GB
RAM	43GB

TABLE V. Training parameters

Weight	yolov5s.pt
Batch size	16
Workers	8
Epoch	100
Other parameters	Default

C. Training Result

The training results of traffic lights, traffic signs and pedestrian models obtained by yolov5s training are shown in TABLE VI.

TABLE VI. Traffic light, Traffic sign, Pedestrian model training results

Class	Precision (%)	Recall (%)	mAP@0.5(%)
red	95.8	97.2	97.7
yellow	99.9	87.7	93.7
green	94.3	94.9	96.4
Off	90.7	92.2	91.9
warning	98.9	99.5	99.6
prohibitory	97.7	98.2	99.2
mandatory	99.0	98.5	99.6
Pedestrian	83.2	68.2	68.0

The model obtained after training is visually output in Fig. 6, Fig. 7 and Fig. 8. The following are the labeling test results of traffic lights,traffic signs and pedestrians.

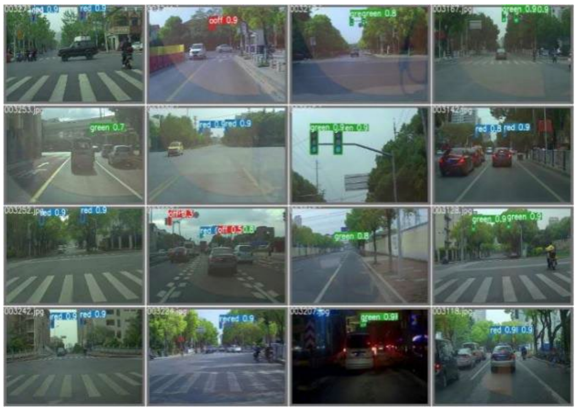

Fig. 6. Result of a traffic lights

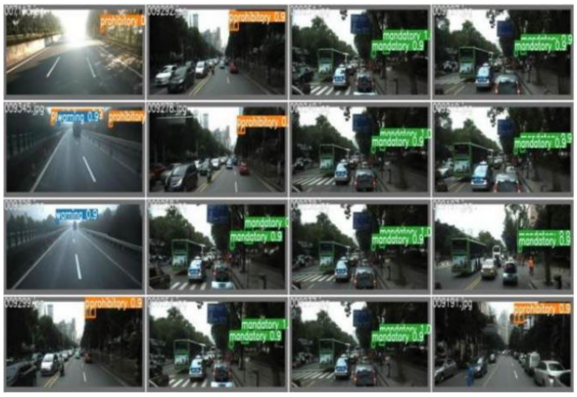

Fig. 7. Result of traffic signs

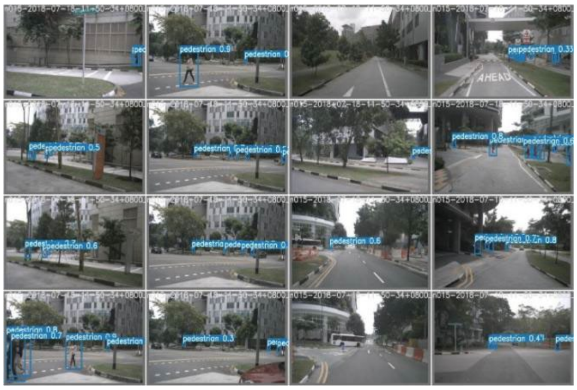

Fig. 8. Result of pedestrians

There is also a unified visual effect of driveable area,lane line division and vehicle detection,as shown in Fig. 9.

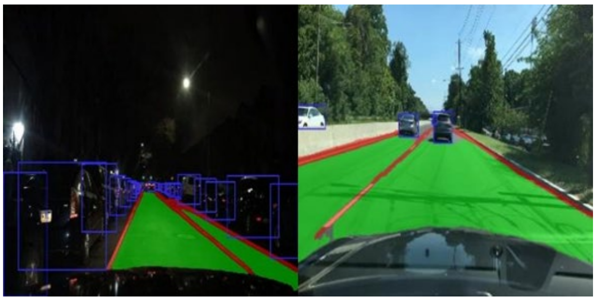

Fig. 9. The yolop driving area division, lane line detection, and vehicle detection effect diagram

D. Summary of the Experimental Results

The traffic light model and traffic sign model have great effect and realize the recognition in the real scene, which is not easy to miss the judgment.

The pedestrian model is not as fine as the default model in terms of parameters. But it enables pedestrian recognition in more complex environments. It is also not easy to miss a wrong judgment, and the pedestrian recognition effect is also very fine in the real scene.

E. Decision-Making

The system architecture of automatic driving is composed of four modules: sensing information, uploading information, decision-making and control. Among them, behavioral decision plays an significant role in connecting the preceding and the following. This paper uses the above model to acquire information about road targets (pedestrians, traffic lights, traffic signs, cars), lane lines and driveable areas. The information is then uploaded to the driving system and prioritized for decision-making and control. Finally, visual output of decision information is carried out. Vehicle speed decisions include maintaining speed, slowing down, and stopping. Direction decisions consist of straight, left, and right turns.

1) Classification of Information for Decision Making

a) Traffic Signal Light Information

YOLOv5 model identifies four types of signal lights: red, yellow, green and off.

b) Lane Line Information

The YOLOP model identifies the current lane-line category.

c) Pedestrian Information

The YOLOv5 model recognizes whether a pedestrian is passing ahead and outputs the distance to him or her.

d) Vehicle Information

The YOLOP model recognises that there is a vehicle ahead and outputs the distance from it.

e) Signage Information

The YOLOv5 model recognizes three categories of signs: indicating signs, prohibiting signs， and warning signs. They are shown in the visualization as 1,2, and 3 respectively.

f)Driveable Area Decision

The YOLOP model determines the drivable area.

The comprehensive effect of the above perception information is shown in Fig. 10.

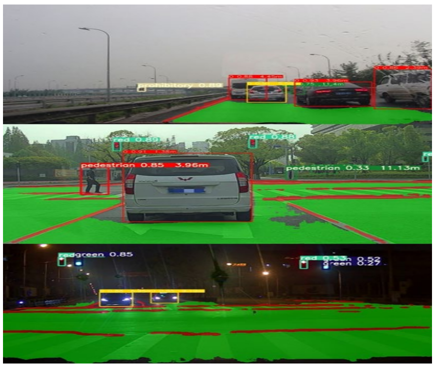

Fig. 10. The comprehensive picture of the perception information

2) Introduction to Decision Priorities

All the identified information is uploaded to the driving system for unified decision-making. Priority in descending order is: distance (high/medium risk), traffic lights or signs, lane lines, feasible area, as shown in Table VII.

TABLE VII. PRIORITIES AND DECISIONS

Priority	Decision
Distance	The distance between the self-driving car and the car or person in front of it is divided into three risk levels. Less than 5 meters is high risk, between 5 meters and 10 meters is moderate risk, and greater than 10 meters is low risk. According to different risk levels, make different decisions on speed. For high risk, you must slow down to a stop. If the risk is medium, it can be slowed down appropriately. If the risk is low, the speed can be maintained.
Traffic light	When the red light is detected, the vehicle stops and waits. If the light is green, keep going.
Traffic sign	There are three kinds of signs that can be recognized by the driving system: indicating signs, prohibiting signs and warning signs. The system adjusts the driving state of the vehicle according to the identification result. For example, if the no left turn sign is recognized, the vehicle is forbidden to turn left on the current road.
Lane line	Based on the type of lane identified by the driving system, plan the next ride. For example, if the lane line is identified as a sidewalk, reduce the speed to pass. If the dotted line is recognized, the vehicle can change lanes.
Driveable area	The driveable zone ensures that vehicles remain in a safe zone. It replaces lane lines when there are no lane lines. According to the results identified by the system, constantly revise the driving route.

3) Demonstration of Partial Effect of Decision Control

a) Example 1

When the distance to the vehicle or person in front is less than 5 meters, the distance priority is the largest. Therefore, the decision information is determined by the distance factor. As shown in the figure below, although the green light is recognized at this time, the distance from the car in front is less than 5 meters, so stop is selected.Example1 decision information is shown in Fig. 11.

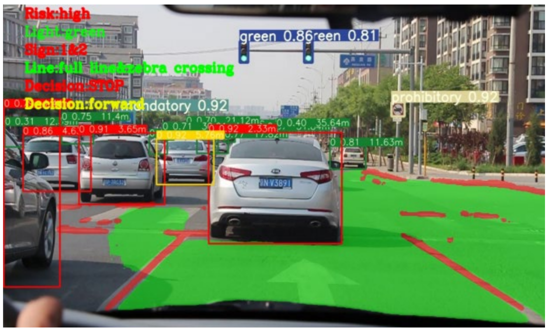

Fig. 11. Effect display 1

b) Example 2

When the distance between the vehicle and the vehicle in front is greater than 5 meters and less than 10 meters, it is considered as medium risk. Moreover, zebra crossing is recognized, so the driving system chooses to go straight and pass slowly. Example2 decision information is shown in Fig.12.

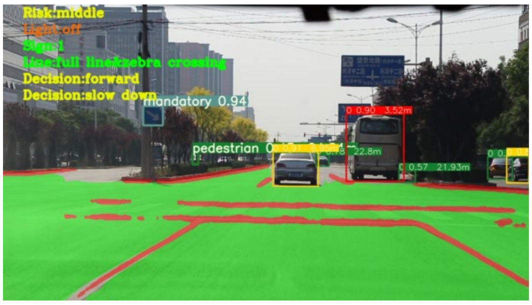

Fig. 12. Effect display 2

V. Conclusion

In order to realize the decision-making and control of automatic driving, this paper divides the system architecture of automatic driving into three modules: perception information, upload information, decision and control, which are realized in stages. For the acquisition of perceptual information, the paper finally chose the combination of YOLOP and YOLOv5 models through comparison with other methods. Finally,great results were obtained. After all the information is uploaded, this paper proposes a prioritization method based on the above information. It can realize automatic driving function in most domestic real scenes.

This paper provides a theoretical basis for the interaction between decision control and perception information in the field of autonomous driving. Next,focusing on optimizing decision control using existing algorithms so that autonomous driving can be implemented in a wider range of real-world scenarios is important.

References

[1] Yang X , Yan J , Yang X , et al. SCRDet++: Detecting Small, Cluttered and Rotated Objects via Instance-Level Feature Denoising and Rotation Loss Smoothing[J]. 2020.

[2] Zhang J, Zou X, Kuang L D, et al. CCTSDB 2021: a more comprehensive traffic sign detection benchmark[J]. Human-centric Computing and Information Sciences, 2022, 12.

[3] Zhang J , Wang W , Lu C , et al. Lightweight deep network for traffic sign classification[J]. Annals of telecommunications, 2020(7/8):75.

[4] J. Zhang, Z. Xie, J. Sun, X. Zou and J. Wang, "A Cascaded R-CNN With Multiscale Attention and Imbalanced Samples for Traffic Sign Detection," in IEEE Access, vol. 8, pp. 29742-29754, 2020

[5] Yu F , Chen H , Wang X , et al. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning[J]. 2018.

[6] Caesar H, Bankiti V, Lang A H , et al. nuScenes: A multimodal dataset for autonomous driving[J]. 2019.

[7] Fong W K , Mohan R , Hurtado J V , et al. Panoptic nuScenes: A Large-Scale Benchmark for LiDAR Panoptic Segmentation and Tracking[J]. 2021.

[8] Jinyuan Liu,Mingfeng Zhang.Automatic water depth extraction basedon YOLOv5 algorithm[J].Journal of Fujian Normal University (Natural science edition), 2023,39(01):86-92.

[9] Yuru Zhou,Dan Li,Chenyu Xiao,Zilong Zhao.Traffic sign recognitionsystem based on YOLOv5[J].Computer knowledge and technology, 2022, 18(19):97-99.

[10] Lijun Dong,Zhigao Zeng,Shenqiu Yi,Zhiqiang Wen,Chen Meng. Object detection in remote sensing image based on YOLOv5[J]. Journal of Hunan University of Technology, 2022,36(03):44-50.

[11] Wu D , Liao M W , Zhang W T , et al. YOLOP:You Only Look Once for Panoptic Driving Perception[J]. Machine Intelligence Research: English Edition, 2022, 19(6):13.

[12] S2TLD is from https://github.com/Thinklab-SJTU/S2TLD.

[13] CCTSDB 2021 is from GitHub-csust7zhangjm/CCTSDB2021.

[14] BDD100K is from http://bdd-data.berkeley.edu.

[15] Nuscenes is from https://www.nuscenes.org/nuscenes.