

Multi-task panoramic driving perception algorithm based on improved YOLOv5

Xin Zeng *, Yong Qi, Mingjun Liu

Nanjing University of Science and Technology, Nanjing, China

* Corresponding Author Email: 911876230@qq.com

Abstract. Panoramic driving perception algorithm is becoming more and more important in the field of autonomous driving. A multi-task panoramic driving perception algorithm can quickly help vehicles make reasonable decisions. Aiming at the real-time high-precision problem of a multi-task panoramic driving perception algorithm, a multi-task panoramic Driving perception algorithm based on YOLOv5 structure is proposed. The algorithm consists of a backbone network for feature extraction and two branch networks for specific tasks. The C3 module in the original backbone network of YOLOv5 is replaced with an inversion residual bottleneck module to reduce the computation amount of the network and improve the network recognition accuracy. This algorithm proposes a new branch network, which can simultaneously train the segmentation of driveable area and lane detection, and improve the speed of feature extraction. In BDD100K, the public data set of Berkeley AI Laboratory, the detection speed of the algorithm proposed in this paper reaches 110FPS, which is 19FPS higher than that of the algorithm before improvement. The detection accuracy of traffic target detection reaches 77.2%, and the detection accuracy of lane line detection reaches 71.1%. It can well meet the real-time requirements of panoramic driving perception.

Keywords: Deep learning, YOLOv5 algorithm, traffic object detection, driveable area segmentation, lane line detection.

1. Introduction

1.1. Background

In recent years, with the rapid development of deep learning, especially in autonomous driving. However, many deep learning algorithms are used in the sensing systems of autonomous vehicles. The perception system extracts visual information from images taken by a monocular camera placed in the front of the vehicle, and helps the decision-making system of the vehicle to make good driving decisions, so as to achieve the purpose of controlling the vehicle behavior. In order for the vehicle to drive safely on the road under the premise of obeying the traffic rules, the visual perception system should be able to process the surrounding scene information in real time, and then help the decision-making system to make judgments, including: the location of obstacles, whether the road is drivable, the location of the lane, etc. Therefore, the panoramic driving sensing algorithm must ensure the detection of the three most critical tasks: traffic object detection, driveable area segmentation and lane line detection.

In published work, Each individual task achieved advanced level, for example, SSD [1] and YOLOv4[2] deal with object detection, and UNet[3], SegNet[4] and PSPNet[5] propose a semantic segmentation algorithm. LaneNet[6], SAD-ENet [7] and SCNN[8] are used for lane detection. However, in the application of autonomous driving scenarios, these tasks are not separate.

So many researchers have proposed multi-task networks, which can process multiple tasks simultaneously to speed up the image analysis process, the network can share information between multiple tasks, which may improve the performance of each task, because many tasks usually share the same trunk network is used for feature extraction. Among them, Mask R-CNN [9] is proposed to jointly detect objects and segment instances, and each task achieves state of the art performance. But this cannot be directly applied to intelligent transportation, because the network cannot detect driveable areas and lanes. MultiNet[10] uses an Encoder-decoder structure. The network consists of a shared backbone network and three separate branch networks for classification, object detection,

and semantic segmentation. It performs well on these tasks, and the drivable region segmentation task is state of the art on the KITTI dataset. However, the classification task is not as important as lane detection in panoramic driving sensing systems. Dlt-net [11] combines traffic object detection, driveable area segmentation and lane line detection, and proposes a context tensor to fuse feature maps between branch networks to share mutual information. Although it has competitive performance, it cannot achieve real-time performance. YOLOP [12] constructed an efficient multi-task network for the panoramic driving sensing system, which includes target detection, driveable area segmentation and lane line detection tasks. The network can be deployed on the embedded device Jetson TX2 through TensorRT to achieve real-time performance, although it has reached an advanced level in real-time performance and high accuracy. However, its three-branch network is used to handle three different tasks, which increases the inference time of the network.

In conclusion, since both driveable region segmentation and lane detection belong to semantic segmentation, there is some room for improvement in branch networks for different tasks. Therefore, this paper designs a simple and efficient detection algorithm (Driving Perception-YOLO), DP-YOLO. The main contributions of this paper are as follows:

1. An inversion residual bottleneck structure is designed to replace the original C3 module of YOLOv5 backbone network with an inversion residual bottleneck module to improve the recognition accuracy of the network model.
2. A branch network is designed, which can train the two tasks of driveable area segmentation and lane line detection at the same time, and improve the speed of feature extraction.

1.2. Related Works

At present, the mainstream object detection algorithms can be divided into one-stage method and two-stage method. The SDD [1] series and YOLO series algorithms are milestones in single-stage methods, which perform bounding box regression and object classification simultaneously. So far, researchers have released five versions of the YOLO series. YOLOv1[13], YOLOv2[14], YOLOv3[15], YOLOv4[2]. Compared with YOLOv4, YOLOv5 is the state of the art in all known YOLO implementations. The term YOLOv5 used in the following paragraphs refers to YOLOV5S-P5 unless otherwise specified. The two-stage method completes the detection task in two steps. Firstly, the region proposal is obtained, and then the features in the region proposal are used to locate and classify the target. Regional proposals have gone through several stages of development. R-cnn [16] creatively attempts to extract region proposals on the original image using selective search instead of sliding Windows, while Fast R-CNN [17] performs this directly on feature maps. The RPN network proposed in Faster RCNN [18] greatly reduces the time consumption and achieves higher accuracy. Based on the former, R-FCN [19] proposed a fully convolutional network, which replaced the fully connected layer with the convolutional layer to further speed up the detection speed.

The task of driving area segmentation belongs to the direction of semantic segmentation. Due to the rise of deep learning, the accuracy of semantic segmentation algorithm has been better improved. FCN [20] introduces fully convolutional network into semantic segmentation for the first time. This method replaces fully connected network with convolutional network. Despite the improvement of skip connection, its performance is still limited by low-resolution output. UNet [3] constructs the Encoder-decoder architecture, which combines the information of the bottom layer and the top layer to output higher resolution. DeepLab [21] uses the dilated convolution algorithm to expand the receptive field and obtain more contextual information while maintaining similar computational costs. In this method, a fully connected conditional random field (CRF) is used to improve the ability of the model to capture details and improve the output quality. PSPNet [5] proposed the pyramid pool module, which can aggregate the context information of different regions, so as to improve the ability to obtain global information.

In recent years, deep learning methods use network models to automatically learn target features, which has strong generalization ability and can effectively improve the accuracy of target detection. Thanks to the powerful feature extraction ability of convolutional neural networks, the performance

is also constantly improving. SCNN [8] proposed a new neural network structure to extract lane lines, and this method extended the traditional deep layer-by-layer convolution to slice-by-slice convolution in feature mapping, so as to realize message passing between pixels between rows and columns in the layer. Enet SAD [7] proposed a self-attention distillation module. The self-attention distillation module of this method solves the influence of large backbone network on speed and can aggregate text information. This approach allows the use of a more lightweight backbone to achieve high performance while maintaining real-time efficiency. UFAST [22] defines lane line detection as the task of finding the set of lane line positions in a specific row of an image. This row-based classification uses global features, and this method achieves the fastest detection speed at present.

The goal of multi-task learning is to learn better representations by integrating multiple tasks into a unified network and sharing information between tasks. MultiNet [10] proposed an encoder-decoder structure. This network can be trained end-to-end using multi-task learning algorithms. DLT-NET [11] also inherits the encoder-decoder structure. This network proposes context tensors to fuse feature maps between branch networks and share specified information between tasks. YOLOP [12] also inherits the encoder-decoder structure, and the three sensing tasks of the network (vehicle detection, driveable area segmentation and lane line detection) have achieved an advanced level on the BDD100K dataset [23]. Both DLT-NET and YOLOP are network structures consisting of a shared backbone network and three separate branch networks, each of which handles three tasks. Since both driveable area segmentation and lane line detection belong to semantic segmentation, the branch network for different tasks has some room for improvement. Therefore, this paper uses a branch network to train the two tasks of driveable area segmentation and lane line detection.

2. Methodology

2.1. Network Architecture

This paper proposes a multi-task panoramic driving sensing algorithm framework DP-YOLO based on YOLOv5 network structure, which uses end-to-end network to realize real-time traffic object detection, driveable area segmentation and lane line detection.

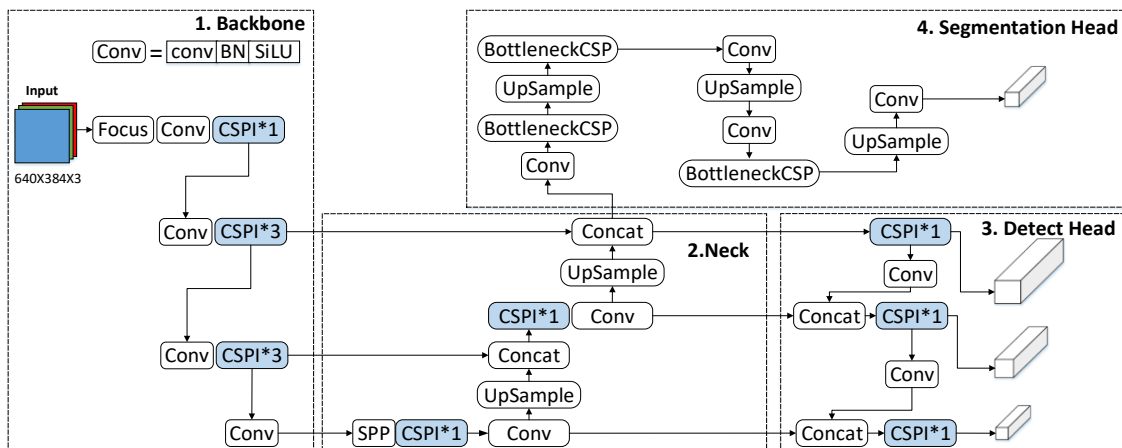


Fig 1. The architecture of DP-YOLO

As shown in Figure 1, The proposed network consists of four parts: 1) Backbone network: Backbone network is used to extract features of input images. In this paper, an improved model based on YOLOv5 is adopted as the backbone network. 2) Neck: Neck is used to fuse the features generated by the encoder. The neck of this paper is mainly composed of spatial pyramid pool (SPP) module [24], characteristic Pyramid network (FPN) module [25]. 3) Detect Head: Similar to YOLOv4, we adopt an anchor-based multi-scale detection scheme. we use a structure called Path Aggregation Network[26](PAN), a bottom-up feature pyramid network. 4) Segmentation Head: The branch network can train the two tasks of driveable area Segmentation and lane line detection simultaneously.

2.2. CSPI module

As shown in Figure 2, the CSPI module proposed in this paper, this module draws on the idea of CSPNet[27], maps the features of the base layer into two parts, and then merges them through a cross-stage hierarchical structure, which can greatly reduce the computation amount of the network and improve the running speed and detection accuracy of the network.

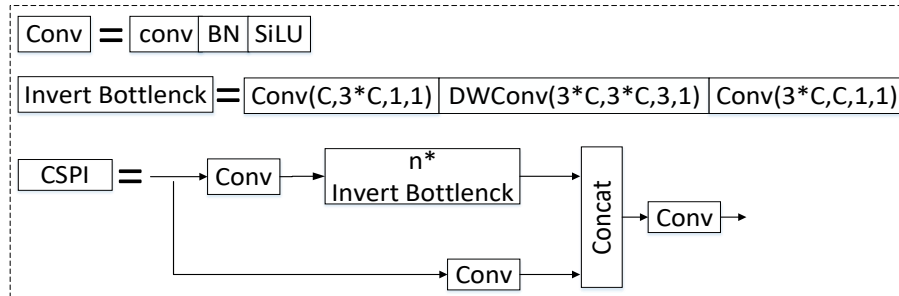


Fig 2. CSPI module

In this paper, we propose the Invert Bottleneck residual component structure, which consists of three layers. The first layer is the Conv layer, which maps the lower dimensional space to the higher dimensional space. The second layer is DWConv layer. DWConv is a depth-separable convolution layer, and depth-separable convolution is applied for spatial filtering. The third layer is the Conv layer, which maps the high-dimensional space to the low-dimensional space.

2.3. Branch network

The branch network consists of four layers of Conv convolutional layers, three layers of BottleneckCSP modules, and three layers of upper sampling layers. As shown in Figure 3, in this paper, the bottom layer of FPN is input to the segmentation branch with the size of $(W/8, H/8, 256)$, because the bottom layer of FPN has strong semantic information and high resolution information for localization. After three times of upsampling processing, this paper restores the output feature map to the size of $(W, H, 4)$, which represents the probability of each pixel in the input image for the background corresponding to the drivable area and lane line. In addition, this paper uses the nearest interpolation method in the upsampling layer to reduce the computational cost. This paper uses the BottleneckCSP module, which strengthens the capability of the network feature fusion and improves the detection accuracy. Therefore, the proposed branch network not only obtains high precision output, but also reduces its inference time.

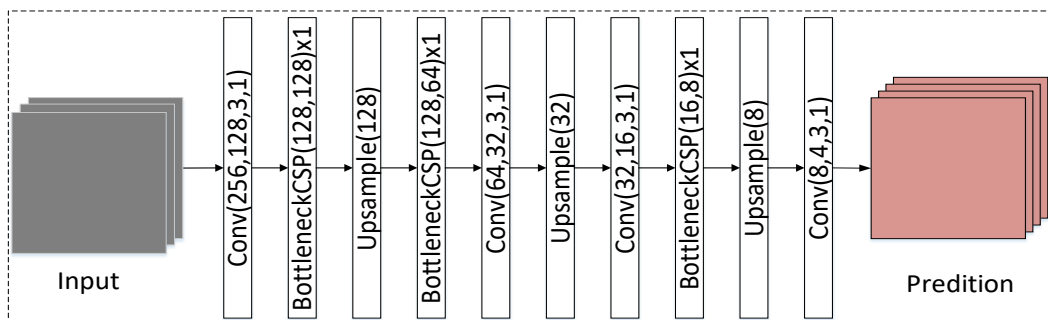


Fig 3. Segmentation branch network of DP_YOLO

2.4. Loss function

In the training process of deep learning, the loss value of the current model is obtained by calculating the difference between the real label in the image and the output result of the neural network. The loss value reflects the difference between the output result and the real label. In the process of back propagation, the network parameters are adjusted according to the loss value so that the predictable result of neural network can approach the true value.

$$loss_{all} = \alpha loss_{det} + \beta loss_{seg} \quad (1)$$

where α , β are tuning parameters to balance the total loss, $loss_{det}$ is the loss for object detection task and $loss_{seg}$ is the loss for segmentation task, the formulation can be written as follows:

$$loss_{det} = \alpha_1 loss_{class} + \alpha_2 loss_{obj} + \alpha_3 loss_{box} \quad (2)$$

$loss_{class}$ and $loss_{obj}$ are focal loss [28], which is implemented for classifying class and the confidence of objects, respectively. The focal loss reduces the slope of loss function and focuses on misclassified examples. $loss_{obj}$ is computed by CIOU Loss[29], which takes into account distance, overlap rate, scale similarity, and aspect ratio between prediction box and ground truth. The effect of prediction box regression is better. Next $loss_{seg}$ is multi-class hybrid loss that is utilized for multi-class segmentation of background, drivable area and lane line.

$$loss_{seg} = \lambda_1 loss_{da-seg} + \lambda_2 loss_{ll-seg} \quad (3)$$

Both of the loss of drivable area segmentation $loss_{da-seg}$ and lane line segmentation $loss_{ll-seg}$ contain Cross Entropy Loss with Logits, which aim to minimize the classification

errors between pixels of network outputs and the targets. It is worth mentioning that $loss_{IoU}$ is added to $loss_{ll-seg}$ as it is especially efficient for the prediction of the sparse category of lane lines. Where x is the predicted value of the network, and $sigmoid$ uses a function to map x to the space of (0,1) to reflect the current category probability. z indicates the real annotation information. The value can be 0 or 1.

$$loss_{da-seg} = -(z \log(sigmoid(x)) + (1-z)(sigmoid(1-x))) \quad (4)$$

$$loss_{IoU} = \left(\frac{TN}{TN + FP + FN} \right) \quad (5)$$

$$loss_{ll-seg} = -(z \log(sigmoid(x)) + (1-z)(sigmoid(1-x))) + loss_{IoU} \quad (6)$$

where $\alpha, \beta, \alpha_1, \alpha_2, \alpha_3, \lambda_1, \lambda_2$ can be tuned to balance all parts of the total loss.

3. Experiments

3.1. Setting

1) Dataset Setting: The BDD100K dataset [11] supports the research of multi-task learning in the field of autonomous driving. With 100k frames of pictures and annotations of 10 tasks, it is the largest driving video dataset. As the dataset has the diversity of geography, environment, and weather, the algorithm trained on the BDD100k dataset is robust enough to migrate to a new environment. Therefore, we choose the BDD100k dataset to train and evaluate our network. The BDD100K dataset has three parts, training set with 70K images, validation set with 10K images, and test set with 20K images. Since the label of the test set is not public, we evaluate our network on the validation set. In order to save memory usage, the image size in BDD100K dataset is adjusted from 1280×720×3 to 640×384×3 in this paper.

2) Implementation Details: The operating system version of the experimental machine in this paper is CentOS Linux Release 7.6.1810, the CPU model is Hygon C86 7185 32-Core Processor CPU@2.0 GHz, and the GPU model is NVIDIA Tesla T4. The video memory size is 16GB and the memory

size is 50GB. The program code is implemented using Python3.8 and Pytorch 1.9, with CUDA 11.2 and CUDNN 7.6.5 for GPU acceleration. The iteration times of the model were set to 200, the initial learning rate was set to 0.01, and the momentum and weight attenuation were set to 0.937 and 0.0005, respectively. In the training process, the learning rate was adjusted by preheating and cosine annealing to make the model converge faster and better.

3.2. Evaluation Metrics

In order to verify the performance of the improved model, appropriate evaluation indexes should be selected to evaluate the network model. In this paper, precision (P), Recall (R), average precision (AP), mean average precision (mAP) and Frames Per Second were selected. FPS) and other evaluation indicators [30] to evaluate the improved network model. In the semantic segmentation task, IoU is used to evaluate the driveable area and lane segmentation, and mIoU is used to evaluate the segmentation performance of different models.

1) IoU, mIoU

IoU is a measure of pixel overlap between the predicted mask map and the true mask map. TN refers to the negative samples predicted by the model as negative class, FP refers to the negative samples predicted by the model as positive class, FN refers to the positive samples predicted by the model as negative class. mIoU is calculated by summing and averaging the IoU calculated for each category.

$$IoU = \left(\frac{TN}{TN + FP + FN} \right) \quad (7)$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^k \frac{TP}{FN + FP + TP} \quad (8)$$

Where k represents the category, and $(k+1)$ represents the addition of the background class. TP refers to the positive samples predicted by the model to be of positive class.

2) Detection speed

Generally, the detection speed is measured by Frames Per Second (FPS), which indicates the number of images that the target detection network can process Per Second. The higher the FPS value, the faster the network model can process images.

3.3. Result

In order to verify the detection performance of DP-YOLO network, this paper selects some excellent multi-task networks and networks focusing on a single task for comparison. MultiNet, DLT-net, and YOLOP all handles multiple panoramic driving awareness tasks and show good performance on BDD100k data set for object detection and drivable region segmentation. Among them, YOLOP is the most advanced network, and Faster-RCNN is one of the excellent two-level target detection networks. YOLOv5 achieves state of the art performance on COCO datasets. PSPNet achieves excellent performance on semantic segmentation task by virtue of its superior global information aggregation ability. In this paper, the above network is retrained on the BDD100K dataset, and the above method is compared with the proposed method on the task of object detection and driveable region segmentation. Since there is no suitable multi-task network to handle the lane line detection task on the BDD100K dataset, the proposed method is compared with three advanced lane line detection networks, Enet[31], SCNN and Enet SAD.

The computational cost comparison between the proposed method and other multi-task networks is shown in Table 1. Although the proposed method has more parameters (8.7m) than YOLOP (7.9m), the backbone network in this paper uses depth-separable convolution, so the number of calculations of the proposed method (14.0GFLOPs) is less than that of YOLOP (15.5GFLOPs). In addition, the branch network in this paper can train the two tasks of lane line detection and driveable area segmentation at the same time, so the reasoning speed of the method in this paper can reach 110 FPS

on the Tesla T4 video card, which is 19 frames per second more than YOLOP. Therefore, the proposed DP-YOLO network can be run in real time on both standard and embedded devices.

Table 1. Cost calculation results of various networks

Model	Params	GFLOPs	FPS
YOLOP	7.9M	15.5	91
DP-YOLO	8.7M	14.0	110

Experiments show the results of three tasks, including traffic object detection, driveable area segmentation and lane line detection. Table 2 shows the comparison results of the proposed method on the BDD100K dataset with the five network models on the vehicle detection task.

Table 2. Comparison results of traffic object detection tasks

Model	Recall(%)	mAP50(%)
MultiNet	81.3	60.2
DLT-Net	89.4	68.4
Faster R-CNN	77.2	55.6
YOLOv5s	86.8	77.2
YOLOP	89.2	76.5
DP-YOLO	89.3	77.2

As shown in Table 2, the proposed method outperforms the previous network model on the BDDK100 dataset. The proposed method achieves a recall rate of 89.3% and an average precision rate of 77.2% when the IoU is set to 0.5. In this paper, an inversion residual bottleneck module is designed, which can reduce the computation amount of the network and improve the accuracy of network recognition. Figure 4 shows the visualization results of traffic target detection under day and night conditions.

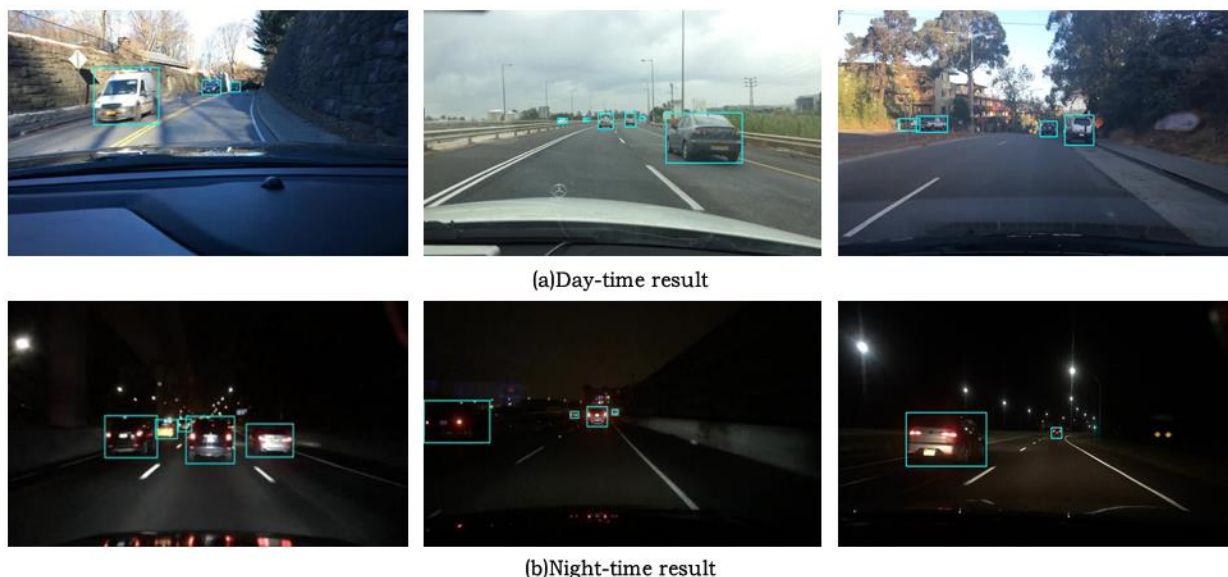


Fig 4. Traffic detection results of DP-YOLO

Table 3. Performance comparison of drivable area segmentation task

Model	Drivable mIoU(%)
MultiNet	71.6
DLT-Net	71.3
PSPNet	89.6
YOLOP	91.5
DP-YOLO	91.5

Table 3 shows the segmentation results of the proposed method and five network models in the task of drivable area. The proposed method achieves 91.5% mIoU (mIoU is used to evaluate the segmentation performance of different models). This paper combines driveable area segmentation and lane line detection into the same task, while YOLOP divides the driveable area segmentation and lane line detection into two tasks. Therefore, compared with YOLOP, this network is more lightweight. Figure 5 shows the visualization results of driving area segmentation under day and night conditions.



Fig 5. Drivable area segmentation results of DP-YOLO

Table 4. Performance comparison of lane line detection tasks

Model	Accuracy (%)	Lane Line IoU(%)
Enet	34.12	14.64
SCNN	35.79	15.84
Enet-SAD	36.56	16.02
YOLOP	70.50	26.2
DP-YOLO	71.10	26.0

Lane line detection is a major challenge in autonomous driving. In this method, detection accuracy and interaction ratio (IoU) of lane lines are used as evaluation indexes. As shown in Table 4, the detection accuracy and interaction ratio of the proposed method achieved 71.1% and 26.0% results, respectively. Figure 6 shows the visualization results of lane line detection under day and night conditions.



Fig 6. Lane line detection results of DP-YOLO

In this paper, the trained DP-YOLO network is used to test the dataset BDD100K, and the obtained detection results are shown in Figure 7. The first line shows the YOLOP mismatch problem, and the second line shows the DP-YOLO results.



Fig 7. Comparison of YOLOP and DP-YOLO Multi-tasking

Figure 8 shows the detection results of DP-YOLO, where the purple line is the lane line, the green line is the zone and the drivable zone, and the cyan bounding box is the traffic object. The DP-YOLO network in this paper has achieved good results in various detection tasks. The DP-YOLO works well in most complex scenarios, such as heavily reflective scenarios and extreme scenario weather conditions. However, the model in this paper cannot adapt to intersecting roads, lane line detection is interrupted and the driveable area is misjudged on the other side of the highway in the use case.

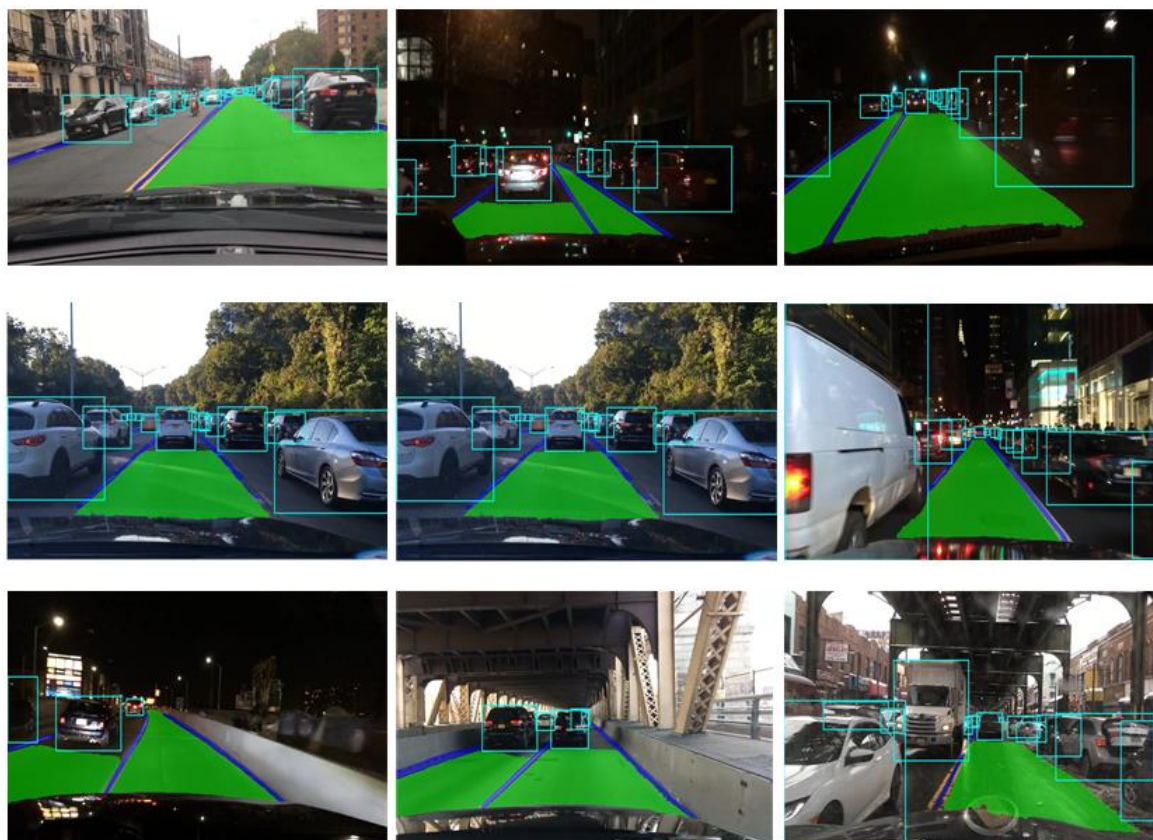


Fig 8. Multi-task panoramic driving perception detection results of DP-YOLO

3.4. Ablation Studies

Through the analysis of the above experiments, the network structure based on YOLOv5 is improved step by step in this paper, and the improved network model is trained and tested. The ablation experimental results are shown in Table 5. Recall, AP, mIoU, Accuracy, IoU is in (%), and Speed is in (MS /frame).

Table 5. Comparison results of ablation experiments

	Branch network	CSPI	Recall	AP	mIoU	Accuracy	IoU	Speed
1			89.2	76.5	91.5	70.5	26.2	11.0
2	✓		88.2	76.2	91.3	68.8	26.6	6.6
3		✓	89.7	77.3	91.6	70.9	26.5	10.8
4	✓	✓	89.3	77.2	91.5	71.1	26.0	9.1

From the experimental results, the experiment 2 in YOLOv5 network structure, on the basis of increased division of the proposed branch network, the network the lane line detection and driving regional segmentation merged into two separate tasks a task training, further enhance the reasoning speed, but test results have less to do with the three tasks. In Experiment 3, CSPI module was used to replace all C3 modules in YOLOv5 model, and its segmentation branch network was the two segmentation branch networks in YOLOP network structure. According to the experimental results of Experiment 3, the detection results of this model in three tasks were better than those of experiment 1. But in order to further enhance the speed, the experiment on the basis of experiment 3 4 YOLOP two segmentation branch network in the cost of replacement, the proposed segmentation branch network, from the perspective of the experimental results of the experiment 4, 4 in each task in the identification precision has better promotion, Recall rate in target detection task (Recall) reached 89.3%, AP was 77.2%, mIoU 91.5%, Accuracy 71.1%, IoU 26.0%, and detection speed 9.1 ms/frames. The experimental results show that the method proposed in this paper has a better role in improving the panoramic driving perception task, which provides a new idea for the research of panoramic driving perception system, and meets the real-time requirements.

4. Summary

Aiming at the problem of real-time high accuracy of an end-to-end network panoramic driving sensing system, this paper proposes a real-time high-precision multi-task panoramic driving sensing network DP-YOLO based on YOLOv5 structure. By replacing the C3 module of the original backbone network of YOLOv5 with the CSPI module, the algorithm uses the branch network to train the two tasks of driveable area segmentation and lane line detection at the same time, which improves the running speed of the network and reduces the computational load of the whole network. The proposed method is tested on dataset BDD100K and achieves better accuracy and efficiency than the existing techniques. Most importantly, compared with the previous multi-network model, the proposed DP-YOLO network achieves relatively high accuracy with fewer FLOPS.

In the future, the feature extraction network will be further optimized. We hope to propose a robust network, which can improve the detection accuracy of the three detection tasks and improve the network parameters and FLOPs under the condition of satisfying the real-time performance of the network. More specifically, our work will focus on dealing with problems in autonomous driving, such as building a network of decoders that can be trained for three tasks simultaneously. We will try to improve the performance of lane line detection and driveable area segmentation.

Acknowledgements

This work was supported by the National Key Research and Development Program (2019YFE0123800), Eu Horizon 2020 Scientific Research Programme (LC-GV-05-2019).

References

- [1] Liu W, Anguelov D, Erhan D, et al. Ssd: Single shot multibox detector [C]//European conference on computer vision. Springer, Cham, 2016: 21-37.
- [2] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object detection [J]. arXiv preprint arXiv:2004.10934, 2020.
- [3] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation [C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.
- [4] Badrinarayanan V, Kendall A, Cipolla R. Segnet: A deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 39(12): 2481-2495.
- [5] Zhao H, Shi J, Qi X, et al. Pyramid scene parsing network [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2881-2890.
- [6] Wang Z, Ren W, Qiu Q. Lanenet: Real-time lane detection networks for autonomous driving [J]. arXiv preprint arXiv:1807.01726, 2018.
- [7] Hou Y, Ma Z, Liu C, et al. Learning lightweight lane detection cnns by self attention distillation [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1013-1021.
- [8] Pan X, Shi J, Luo P, et al. Spatial as deep: Spatial cnn for traffic scene understanding[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).
- [9] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.
- [10] Teichmann M, Weber M, Zoellner M, et al. Multinet: Real-time joint semantic reasoning for autonomous driving[C]//2018 IEEE Intelligent Vehicles Symposium (IV). IEEE, 2018: 1013-1020.
- [11] Qian Y, Dolan J M, Yang M. DLT-Net: Joint detection of drivable areas, lane lines, and traffic objects[J]. IEEE Transactions on Intelligent Transportation Systems, 2019, 21(11): 4670-4679.
- [12] Wu D, Liao M, Zhang W, et al. Yolop: You only look once for panoptic driving perception [J]. arXiv preprint arXiv:2108.11250, 2021.
- [13] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, real-time object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
- [14] Redmon J, Farhadi A. YOLO9000: better, faster, stronger[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 7263-7271.
- [15] Redmon J, Farhadi A. YOLOv3: An incremental improvement [J]. arXiv preprint arXiv:1804.02767, 2018.
- [16] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
- [17] Girshick R. Fast r-cnn [C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
- [18] Ren S, He K, Girshick R, et al. Faster r-cnn: Towards real-time object detection with region proposal networks[J]. Advances in neural information processing systems, 2015, 28.
- [19] Dai J, Li Y, He K, et al. R-fcn: Object detection via region-based fully convolutional networks[J]. Advances in neural information processing systems, 2016, 29.
- [20] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

- [21] Chen L C, Papandreou G, Kokkinos I, et al. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs [J]. IEEE transactions on pattern analysis and machine intelligence, 2017, 40(4): 834-848.
- [22] Qin Z, Wang H, Li X. Ultra fast structure-aware deep lane detection[C]//European Conference on Computer Vision. Springer, Cham, 2020: 276-291.
- [23] Yu F, Xian W, Chen Y, et al. Bdd100k: A diverse driving video database with scalable annotation tooling [J]. arXiv preprint arXiv:1805.04687, 2018, 2(5): 6.
- [24] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition [J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(9): 1904-1916.
- [25] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.
- [26] Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation [C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 8759-8768.
- [27] Wang C Y, Liao H Y M, Wu Y H, et al. CSPNet: A new backbone that can enhance learning capability of CNN[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops. 2020: 390-391.
- [28] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection [C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
- [29] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression [C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(07): 12993-13000.
- [30] Wang Y H, Ding H W, Li B, et al. Mask wearing detection algorithm based on improved yolov3 in complex scenes [J]. Computer Engineering, 2020, 46(11): 12-22.
- [31] Paszke A, Chaurasia A, Kim S, et al. Enet: A deep neural network architecture for real-time semantic segmentation [J]. arXiv preprint arXiv:1606.02147, 2016.