YOLOMH:You Only Look Once for Multi-task Perception Network High-efficiency

Fang Liu
Tianjin Polytechnic University
Jianxi Miao
miaojxi_up@163.com

Tianjin Polytechnic University
Bowen Sun
Tianjin Polytechnic University
Weixing Su
Tianjin Polytechnic University

Research Article

Keywords: Panoptic Driving Perception, Multi-Task Network, HDASPP, YOLOMH

Posted Date: December 12th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-3724985/v1

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Machine Vision and Applications on March 29th, 2024. See the published version at https://doi.org/10.1007/s00138-024-01525-3.

YOLOMH:You Only Look Once for Multi-task Perception Network High-efficiency

Liu Fang[1†], Sun Bowen[2†], Miao Jianxi[3*], Su Weixing[4†]

[1]School of Software Engineering, Tiangong University, Tianjin, China.
[2]School of Software Engineering, Tiangong University, Tianjin, China.
[3*]School of Software Engineering, Tiangong University, Tianjin, China.
[4]School of Software Engineering, Tiangong University, Tianjin, China.

*Corresponding author(s). E-mail(s): miaojxi_up@163.com;
Contributing authors: lf@tiangong.edu.cn; 1304473797@qq.com;
suweixing@tiangong.edu.cn;
†These authors contributed equally to this work.

Abstract

Aiming at the requirements of high accuracy, lightweight and real-time performance of the panoptic driving perception system, this paper proposes an efficient multi-task network(YOLOMH). The network uses a shared encoder and three independent decoding heads to simultaneously complete the three major panoptic driving perception tasks of traffic object detection, road drivable area segmentation and road lane segmentation. Thanks to our innovative design of the YOLOMH network structure: first, we design an appropriate information input structure based on the differences information requirements between different tasks, and secondly, we propose a Hybrid Deep Atrous Spatial Pyramid Pooling(HDASPP) module to efficiently complete the feature fusion work of the neck network, and finally effective approaches such as anchor-free detection head and Depthwise Separable Convolution(DCN) are introduced into the network, making the network more efficient while being lightweight. Experimental results show that our model achieves competitive results in both accuracy and speed on the challenging BDD100K dataset, especially in terms of inference speed, The model's inference speed on NVIDIA TESLA V100 is as high as 107 Frames Per Second(FPS), far exceeding the 49 FPS of the YOLOP network under the same experimental settings. The final visualization shows that YOLOMH can excellently complete the panoptic driving perception tasks, which is conducive to the safe and reliable autonomous driving of autonomous vehicles.

Keywords: Panoptic Driving Perception, Multi-Task Network, HDASPP, YOLOMH

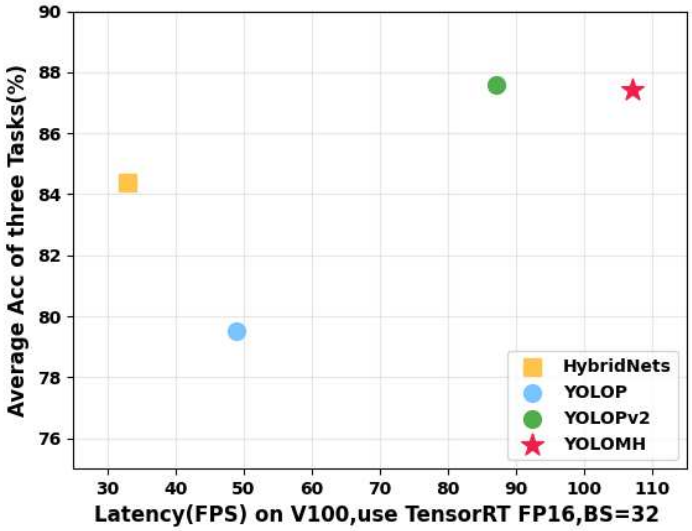

Fig. 1 Speed and accuracy comparison between YOLOMH and other advanced multi-task networks

1 Introduction

In recent years, people have been trying to build a powerful panoptic driving perception system to serve autonomous driving. The tasks of the panoptic driving perception system include object detection, drivable area segmentation and lane line segmentation. It primarily utilizes onboard sensors to gather information regarding the size and location of obstacles surrounding the vehicle, the drivable area of the road, as well as the length and location of road lanes. These environmental cues provide essential foundational support for higher-level tasks such as decision planning and behavioral control in autonomous driving, ensuring safe and reliable autonomous driving of autonomous vehicles on the road [1]. Different from other application scenarios, the environment perception system plays a very important role in the field of autonomous driving. In addition to accuracy, low-cost, high-efficiency and lightweight models are also the goals pursued by researchers.

Currently, the main onboard sensors applied to environmental perception systems in the field of autonomous driving include cameras, lidar, etc. [2]. Compared to cameras, lidar is insensitive to color and light, but is expensive. On the contrary, the RGB image captured by the camera has rich texture and color information, which is suitable for object detection and segmentation tasks. In addition, cameras also have the characteristics of low cost and easy on-board installation [3]. Therefore, using images captured by cameras as input and combining with deep learning models to achieve a panoptic driving perception system is currently a competitive solution for low-cost Advanced Driver Assistance Systems(ADAS), as it can meet the requirements for high efficiency and low cost in autonomous driving [2].

In deep learning-based object detection, there are two mainstream approaches. The first one is two-stage approach, represented by the RCNN series of algorithms [5–7]. These two-stage approaches prioritize detection accuracy, usually extracting high-quality candidate frames first, and then complete the classification and detection tasks. This approach often sacrifices computational efficiency as a prerequisite, which is not conducive to the practical deployment of perception models on vehicles. Single-stage detectors are increasingly popular in the industry due to their efficient performance on embedded devices [4]. The You Only Look Once(YOLO) series of algorithms [8–12] are typical representatives. It completes the detection tasks of three types of objects (large, medium, and small) through a multi-scale approach. Due to its good balance between speed and accuracy, it has become the most popular detection framework in practical applications [13]. The recently proposed YOLOv8 network achieves the best balance between speed and accuracy, but each model can only perform a single task, making it difficult to meet the real-time operational requirements of multi-task panoptic driving perception systems. In the field of low-cost ADAS, multi-task detection networks based on a single model are considered to be an efficient solution for panoptic driving perception systems. Of course, segmentation models have also developed rapidly in recent years. In the fields of road drivable area segmentation and lane line segmentation, the mainstream algorithms include UNet [14], SegNet [15], PSPNet [16], SCNN [17] and ENet-SAD [18] etc.

Unfortunately, although the above approaches have achieved good results in solving their respective tasks, in actual autonomous driving systems, especially low-cost ADAS, in addition to focusing on accuracy, limited computational resources and cost issues are also typically considered. Therefore, it is usually unrealistic to run a separate model for each individual task in an actual autonomous driving system, because this requires the on-board computing platform of the intelligent driving vehicle to have high computing power, which is not only detrimental to network deployment but also fails to meet the low-cost requirement [1].

Against this background, this paper considers using low-cost camera sensors combined with multi-task networks based on deep learning to efficiently solve the above problems. The YOLOMH structure proposed in this paper is an encoder-decoder mode, which uses three task heads to simultaneously complete the tasks of object detection, drivable area segmentation and lane line segmentation in the panoptic driving perception system. We designed the corresponding detection and segmentation task heads in a decoupled manner so that they can share the image feature information extracted by the encoder, thereby avoiding the time and cost overheads associated with using single-task networks separately. Figure 2 shows the inference results of YOLOMH, where the red boxes indicate the model's predicted road vehicle obstacle information, the green area represents the safe drivable area, and the blue lines denote the road lane information.

In the practical application of panoptic driving perception systems, real-time performance is a crucial factor, faster response speed increases the possibility of avoiding accidents and ensuring personnel survival. A key metric for real-time is inference time or FPS [2]. This paper evaluates the proposed YOLOMH on the BDD100K [19] dataset. The results show that the proposed network exhibits competitive results in

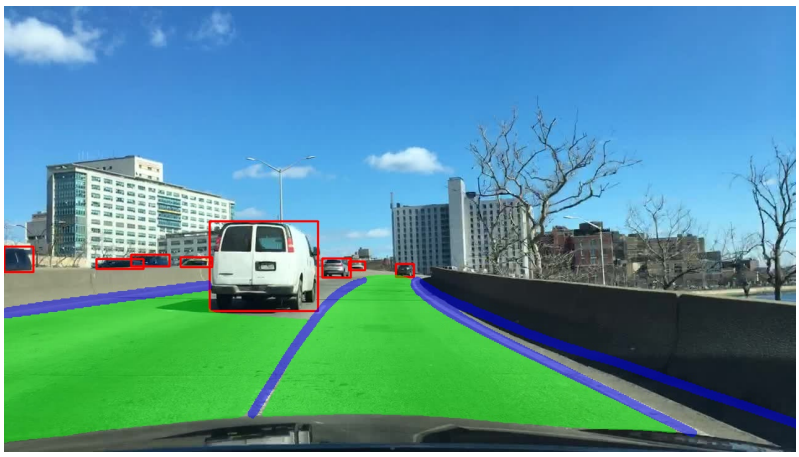

Fig. 2 The inference results of YOLOMH

both accuracy and speed. Specifically, the mAP50 value for object detection is 81.3%, the mIoU value for drivable area segmentation is 92.7%, and the IoU value for lane line segmentation is 29.4%. In terms of inference speed, our model can reach an amazing 107 FPS on V100 GPU, which is far higher than the 49 FPS of the YOLOP [1] model under the same experimental settings. Figure 1 shows a comparison of the proposed multi-task network and other advanced multi-task networks in terms of average accuracy (average accuracy of traffic object detection, drivable area segmentation and lane line segmentation tasks, shown on the vertical axis) and inference time (shown on the horizontal axis). It can be observed that the proposed YOLOMH achieves an optimal fusion of accuracy and speed, effectively meeting the requirements of autonomous driving vehicles for high-accuracy and low-latency systems.

In summary, the main contributions of this paper are: (1) Within the field of autonomous driving, we innovatively design an efficient perceptual multi-task network (YOLOMH), The network can simultaneously complete the three major panoptic driving perception tasks of traffic object detection, road drivable area segmentation and road lane line segmentation, and It has excellent real-time conditions. (2) The proposed HDASPP module not only replaces the original FPN [20] structure for efficient neck network feature fusion, but also has high adaptability, which can take into account the receptive field information of different scales by adjusting the number and dilation rate of Hybrid Dilated Convolutions(HDC) [21]. (3) We introduce the anchor-free idea into the object detection head to avoid complex Non-Maximum Suppression(NMS) operations, and combining the HDASPP and DCN [22] modules to make the network more efficient while realizing lightweight. We will verify the effectiveness of the proposed network in subsequent experiments. Especially in terms of visualization, the model shows impressive results.

2 Related Work

This section will review some classic network models in panoptic driving perception systems, including object detection in single-task areas, road drivable area segmentation, road lane segmentation and panoptic driving perception multi-task network models. We focus on deep learning-based approaches.

2.1 Object Detection

In the field of object detection, mainstream detection algorithms can be divided into two-stage detection approaches and one-stage detection approaches. Representative works of the two-stage detection approaches include RCNN [5], Fast-RCNN [6] and Faster-RCNN [7], which complete the detection task in two steps: firstly, obtaining the Regions of Interest(RoI), and then using the features in the region suggestion to classify and localize the object [4]. As autonomous driving systems have increasing requirements for object detection speed, single-stage detectors have received more and more attention from the industry because of their fast and efficient performance on embedded devices. The YOLO series of algorithms are typical representatives of single-stage detectors. It accomplishes object detection at three different scales, large, medium and small, by dividing the feature map grid with different resolutions, and then considers object detection as a regression problem for end-to-end training and inference. Due to its good balance between speed and accuracy, it has become the most popular detection framework in practical applications [13]. The early representative work of the YOLO series is YOLOv3 [9], which opened up a new path for first-level detectors by introducing a multi-scale detection head. Subsequently, YOLOv4 [10] reorganized the detection framework into several independent parts (backbone, neck, and head), and verified bag-of-freebies and bag-of-specials at the time to design a framework suitable for training on a single GPU. YOLOX [11] introduces decoupled heads and anchor-free approaches, which greatly simplifies the network training and decoding stages. As we can see, efficiency has always been a goal sought by researchers in the object recognition task of autonomous driving. With the continuous development of the YOLO series, currently YOLOv5-v8 are competing candidates for efficient detector deployment. Although the YOLO algorithm is so excellent, it can only complete one task at a time, and most of them use a anchor-base detection mechanism, which cannot be regarded as a true end-to-end detection algorithm [2, 14]. It can become a potential risk in terms of latency when deployed at the in-vehicle device side.

2.2 Drivable Area And Lane Line Segmentation

In the field of autonomous driving, semantic segmentation networks can be used to effectively divide the road drivable area and road lane line information. FCN [23] ignited the flame of the first fully convolutional segmentation network, which improved the recognition rate by 20% on the Pascal VOC2012 [24] dataset compared with the traditional approaches. However, it only focuses on local information and does not consider global information, resulting in rough segmentation results. In the field of drivable area segmentation, UNet uses the classical encoder-decoder structure. PSPNet further introduces pyramid pooling to extract features at different levels, thereby

effectively dividing drivable areas. However, it still has certain limitations when dealing with multi-scale features. SSN [25] adds conditional random field units in the post-processing stage to improve its segmentation performance, but its higher memory consumption is not conducive to practical drivable area segmentation tasks. In the field of lane line segmentation, SCNN replaces the traditional layer-by-layer convolution with the slice-by-slice convolution within the feature map, which aggregates the information of the slices of different dimensions. The message passing mechanism is utilized to capture the strong spatial associations between lanes, which significantly improves the lane detection performance, but the approach has a large delay in real-time applications. Subsequently ENet-SAD created self-attention distillation to help low-level feature mappings learn knowledge from high-level feature mappings. This approach improves the performance of the model while keeping the model lightweight, but the distillation operation will increase the training time of the model. and complexity. CurveLane-NAS [26] uses a neural architectural search technique to obtain a network with better performance, which is beneficial for the detection of curved lanes. However, NAS is computationally expensive and requires a lot of GPU time for searching. Although the results in drivable area segmentation and lane line segmentation have been better, for example, ENet-SAD is able to achieve lightweight while guaranteeing the performance of the model, it is still only able to perform a single task and cannot meet the needs of multi-task detection in panoptic driving perception.

2.3 Multi-Task Approaches

In order to ensure that the panoptic driving perception system can still operate efficiently on low-cost ADAS devices with limited computational resources, some scholars have attempted to integrate these perception networks into a single model. This integration approach saves computational resources and satisfies the real-time requirement [2]. The purpose of multi-task network is to learn better representations by sharing information between multiple tasks, especially the CNN-based multi-task learning approach also enables convolutional sharing of network structure, which is beneficial for better expression of feature information among different tasks [1]. MASK R-CNN [28] extends the Faster R-CNN by adding a branch that predicts the object masks, which combines the instance segmentation and object detection tasks effectively together to parallelize the object detection and instance segmentation tasks. Subsequently, YOLOP adopts an encoder-decoder structure and introduces two additional segmentation heads based on YOLOv5, effectively combining the three major tasks of traffic object detection, drivable area segmentation and lane line segmentation for the first time. However, two redundant segmentation heads in YOLOP leaves some room for optimization. subsequently, HyBridNets [29] considered from the perspective of feature information extraction and fusion, and introduced the BiFPN [30] in the neck layer to further improved the network performance. Shortly after, YOLOPv2 [27] optimizes for the redundancy problem that exists in the two segmentation task heads of YOLOP, and uses the E-ELAN [12] module to further lightweight the network. The approach outperforms previous similar multi-task network approaches by achieving the current optimal fusion of accuracy and speed on the BDD100K dataset. However, it

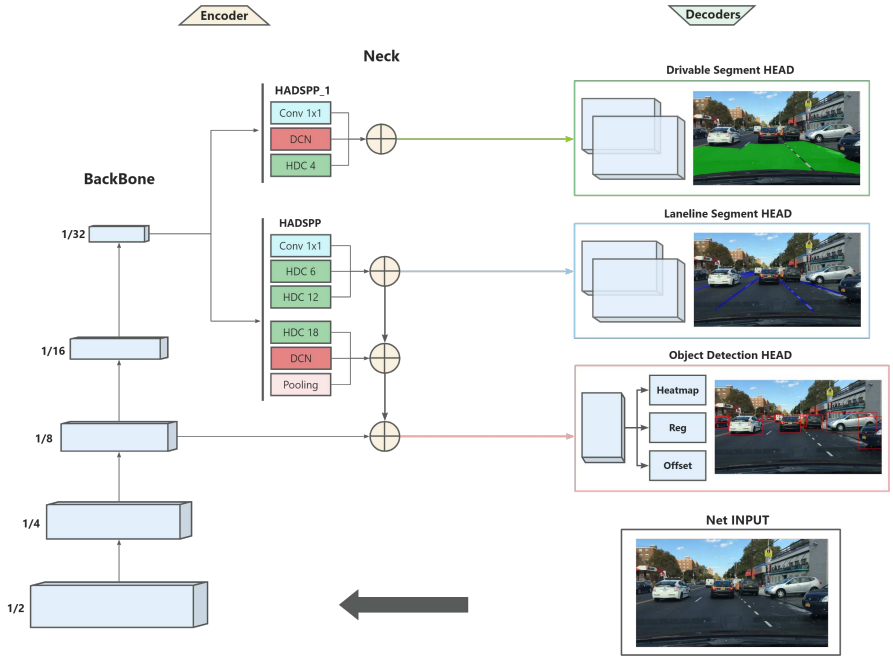

Fig. 3 The architecture of YOLOMH, end-to-end encoder-decoder architecture

does not further address the information differences between heads with different difficulty tasks, and the design of the detection head based on the sight frame still leaves room for time-consuming optimization, which is difficult to meet the time-consuming needs of environmental perception in low-cost ADAS.

3 PROPOSED METHODOLOGY

In this section, we introduce the proposed YOLOMH network in detail. Discuss how to implement an efficient multi-task network to simultaneously accomplish the three major panoptic driving perception tasks of traffic object detection, drivable area segmentation and lane line segmentation. Figure 3 shows the structure of our proposed multi-task network. In general, YOLOMH follows the mainstream structure of multi-task networks: the encoder-decoder structure. However, unlike YOLOP, we use a more lightweight module to perform feature extraction operations on the input images. Moreover, in further experiments, we found that different tasks have different feature information requirements due to their completely different detection characteristics, as the drivable area usually covers a large area while the lane lines tend to be elongated in shape. It is not reasonable for the previous multitask network to use the deepest feature information of the neck uniformly, so we design three independent decoders to perform the three major detection and segmentation tasks one by one. Finally, we apply the idea of anchor-free to the traffic object detection task, which helps improve the network detection speed and ensures the efficiency of the model in practical applications.

3.1 Encoder

The YOLOMH encoder structure is shown in Figure 3, including the backbone network for image feature extraction and the neck feature fusion network. As the shared backbone of multi-task models, the importance of feature extraction is self-evident. An excellent backbone network can help multi-task networks achieve excellent performance in all tasks [29]. The recently proposed YOLOv8 absorbs the excellent ideas of ELAN [13] in yolov7 in the backbone network, and replaces the C3 structure with the gradient-rich C2F structure to achieve further lightweighting [4]. These improvements have enabled yolov8 to show strong strength in object detection tasks. Therefore, we use YOLOv8s backbone network to efficiently accomplish the encoder's feature extraction. It is worth mentioning that, in order to achieve anchor-free and end-to-end networks, we did not copy the previous work of multi-task networks, but utilized the proposed HDASPP to perform multi-scale fusion work.

The HDASPP structure is shown in Figure 4. The module utilizes three different sizes of HDC to obtain different sizes of receptive field information, and then concatenates with 1x1 convolution and pooling layer to achieve multi-scale feature fusion information. The DCN module helps to further reduce the computational burden on the network. The reason for this design is that we found that the FPN based approach does not consider the differences between multi-task and is not friendly to computational resources, which is not conducive to the wide application of panoptic driving perception systems. In subsequent experiments, we found that for the drivable area segmentation task, using deeper features not only failed to improve the model prediction performance, but also increases the difficulty of model convergence during the training. Therefore, we use the HDASPP_1 structure(partial HDASPP) to accomplish this task. For the lane line segmentation task, we found that a larger receptive field is unnecessary, so we did not use the deeper layers(HDC, r=18) in the HDASPP structure. For the object detection task, it requires both rich deep and shallow feature information in the network [37]. Therefore, we concatenate HDASPP to the L3 layer of the backbone network to obtain the receptive field information required by traffic objects at different scales, so as to effectively accomplish the object detection task.

3.2 Decoders

As shown in Fig 3, we design different feature information sources and decoder structures for these three tasks of different difficulty. Inspired by DeepLabv3 [31] and CenterNet [32] network structures, we innovatively use HDASPP and anchor-free approach to efficiently accomplish the decoding of YOLOMH.

3.2.1 Drivable Area And Lane Line Segmentation Heads

YOLOP designs the same decoder for drivable area segmentation and lane line segmentation tasks, and uses the same feature information source. This approach, which does not distinguish the differences between tasks, cannot lead to good performance to drivable area segmentation task. We concatenate the deepest layer of the backbone network with HDASPP_1 to obtain the feature information required for drivable area segmentation. For lane line segmentation, we found that in the input image, lane

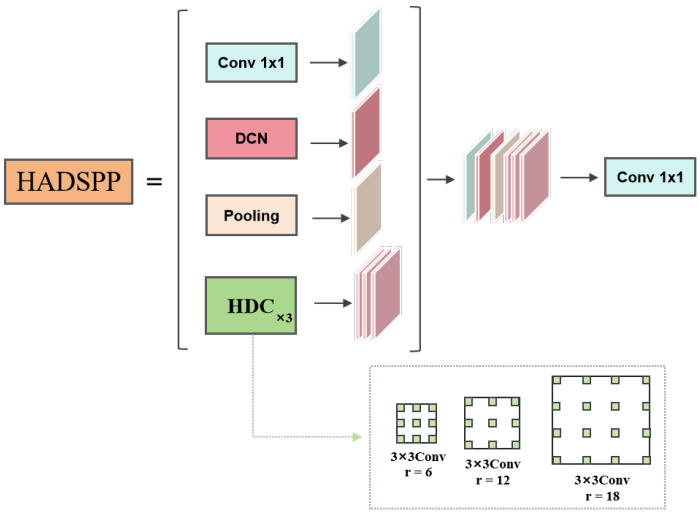

Fig. 4 Proposed structure of HDASPP, use different dilation rates to obtain rich receptive field information in the network

lines usually have a small area but are difficult to detect. This means that this task usually does not require large receptive field information and pays more attention to local contextual feature information. Therefore, we abandon the work of using the module(HDC, r=18) that contains large receptive field to support this task. Finally, YOLOMH's two segmentation tasks classify the pixel level and restore the original image feature map of (H, W, 2) after a series of deconvolutions. Each feature point represents the drivable area or lane at the pixel level, with 1 representing the object and 0 representing the background.

3.2.2 Traffic Object Detection Head

In order to achieve optimal detection performance, anchor-base needs to cluster specific datasets before training and perform complex post-processing operations after inference. It has a series of shortcomings such as low generalization and complex detection heads. Anchor-free detectors have developed rapidly in the past few years. While having performance comparable to anchor mechanism detectors, they significantly reduce the number of design parameters, such as Anchor Clustering [9] and Grid Sensitivity [35]. Accelerating up the training and decoding stages of the detector [32]. The proposed YOLOMH uses an anchor-free approach based on center point prediction on the detection task, thereby avoiding complicated NMS operations on the prediction box and achieving a true end-to-end network structure. Specifically, we concatenate the HDASPP structure to the high-resolution feature layer(L3) of the backbone network to obtain contextual feature information containing different scales to simultaneously complete the detection task of traffic objects of different sizes. The network performs a series of convolution operations on the 1/8 downsampled feature map to obtain a feature map of (80, 80, 4(2)(2)) dimensions, where 4, 2, 2 respectively represent the category, center point, and width and height prediction information.

3.3 Loss Function

The proposed multi-task network loss consists of the sum of the individual task losses. Formula(1) shows our loss function:

$$\mathcal{L}_{\text{Mul-all}} = \alpha_1 \mathcal{L}_{\text{det}} + \alpha_2 \mathcal{L}_{\text{da-seg}} + \alpha_3 \mathcal{L}_{\text{ll-seg}} \tag{1}$$

where $\mathcal{L}_{det}$ is the traffic object loss, $\mathcal{L}_{da-seg}$ and $\mathcal{L}_{ll-seg}$ are the drivable area segmentation loss and lane line segmentation loss respectively. And each task is considered equally important, that is, $\alpha 1 = \alpha 2 = \alpha 3$. In the $\mathcal{L}_{da-seg}$ approach, we used cross-entropy loss to minimize the classification result between GT bounding box and predicted pixel values, and for the more difficult $\mathcal{L}_{det}$ and $\mathcal{L}_{ll-seg}$, we introduced Focal Loss [33] to deal with traffic objects and lane lines that are difficult to classify, so as to bring out the best performance of the network.

In traffic object detection, we introduce Gaussian kernel to determine positive and negative samples to obtain more positive samples for regression training. Specifically, if the prediction point falls within the Gaussian circle, it is marked as a positive sample. In order to enable the prediction point close to the ground truth(GT) bounding box center to learn better information, we allocate different losses according to the distance between the prediction point and the kernel center. The closer the distance, the better the prediction performance and the greater the weight. The maximum weight is 1. The centerness [32] loss is shown in formula(2):

$$\text{centerness} = \sqrt{\frac{\min(l,r)}{\max(l,r)} \times \frac{\min(t,b)}{\max(t,b)}} \tag{2}$$

where t, b, l, r is used to predict the position of the box. It can be found that when the center of the regression box is closer to the real box, the centerness value is closer to 1. The traffic object detection loss is further represented as shown in formula(3):

$$\mathcal{L}_{\text{det}} = \gamma_1 F \mathcal{L}_{\text{heatmap}} + \gamma_2 \mathcal{L}_{\text{offset}} + \gamma_3 \mathcal{L}_{\text{box}} \tag{3}$$

where $\gamma 1, \gamma 2$ are adjustable weight parameters and $\gamma 3$ is the centerness loss. For each location the network predicts the output of (C+4) results, (C) denotes the probability that the location is the center point of each type of object, which partially uses the Focal Loss to balance the learning of difficult and easy samples. (+4) denotes the width and height of the contained object($\mathcal{L}_{box}$), and the offset of the object center($\mathcal{L}_{offset}$). We use L1 Loss [34] to calculate the predicted center point offset, and use LCIoU [36] to calculate the object prediction box loss. The CIoU loss comprehensively considers factors such as overlap, distance, and aspect ratio between the two boxes, allowing for more precise object shape localization.

4 Experiments

This section describes the dataset and parameter configuration for our experiments. All experiments in this paper were conducted in the configuration environment of NVIDIA TESLA V100 graphics card and torch 1.10.

4.1 Experimental Data And Configuration

In the field of visual multitasking networks, the BDK100 dataset has received a lot of attention within the field of automated driving because of its features such as large data volume(10W frames), wide coverage of scenarios(weather conditions, geographic location, lighting conditions, etc.) and full range of tasks (more than ten tasks, such as detection and segmentation, etc.) Therefore, it can be easily migrated to new environments. The BDD100K dataset consists of three parts , a training set of 70K images, a validation set of 10K images, and a test set of 20K images, but the test set is not yet fully public, so we consider evaluating the performance of our network on the validation set. Part of the data display of BDD100K is shown in Figure 5.

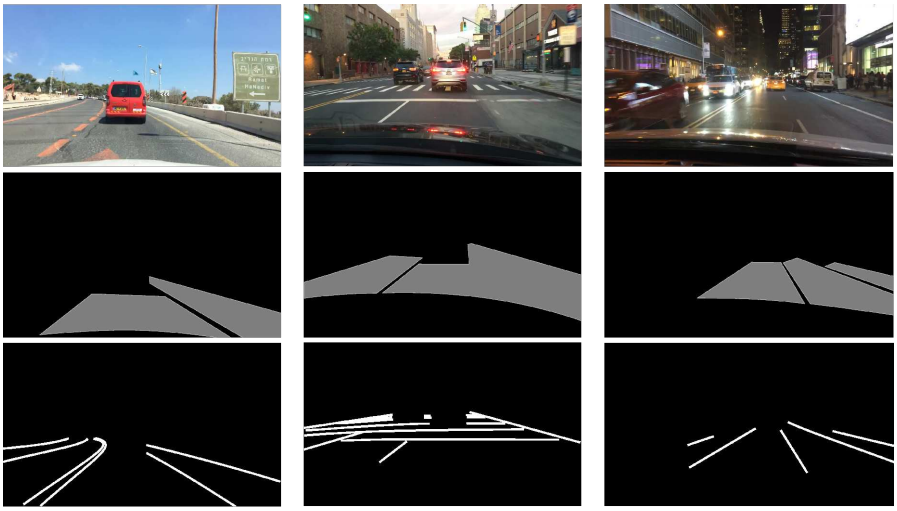

Fig. 5 The partial data and annotation of BDD100K dataset

The total number of training times for the network is 300 epochs. We adjusted the image size from 1280×720 to 640×640 for training, and then adjusted it to 640×384 for operation during the evaluation phase. This is beneficial to the training cycle and inference time of the multi-task network. We introduce cosine annealing [11] approach to adjust the learning rate. This approach has been beneficial in many neural network tasks. Similarly, we set the initial learning rate to 0.01, and trigger warm up after the 2nd epoch to set the momentum and weight decay to 0.937 and 0.005, respectively. In addition, we also set a training steep slope for the learning rate, that is, linearly reducing the learning rate to 1/10 of the original value at epoch 100, 200, and 270 to prevent missing the best performance of training. In order to improve the performance

Fig. 6 YOLOMH can capture rich receptive field information at different scales. The red, white and yellow circles in the figure represent large, medium and small traffic object information respectively

of the model, we tried to perform data augmentation on the input data. In addition to standard approaches such as random brightness, contrast, flipping and cropping of images, we also used random Mosaic [9] data augmentation operations, and in the last 20 epochs of training Mosiac enhancement operation was turned off, which proved to be effective in subsequent experiments.

4.1.1 Cost Computation Performance

Faster inference speed has always been one of the goals pursued by deep learning. Especially in autonomous driving tasks, the network needs to be installed and deployed on onboard devices with limited computing resources, so inference time is particularly important. Table 1 shows the time-consuming comparison between several excellent multi-task models and our model. We conducted comparative tests using the same experimental settings and evaluation metrics. The results show that compared with YOLOPv2 and HybridNets multi-task network models, our model has fewer parameters and inference delays. In terms of inference speed, YOLOMH is 58 FPS faster than the mainstream YOLOP and 20 FPS faster than the current best YOLOPv2.

Table 1 Network parameters and inference time results, batch size is 32.

Mul-Nteworks	Size(Pixel)	Parameters(M)	Speed(FPS)
YOLOP	640	7.9	49
HybridNets	640	12.83	33
YOLOPv2	640	38.9	87
YOLOMH	640	11.91	107

4.1.2 Traffic Object Detection Performance

Table 2 shows the comparison between several commonly used object detection models and our model . Since multi-task networks such as YOLOP are only concerned with the four vehicle categories(car, bus, truck and train) in the BDD100K dataset for traffic object detection, we will compare and analyze the results of these vehicle detections. Same as YOLOP, we use mAP50 and Recall as evaluation metrics. The results show that our network performance outperforms mainstream multi-task networks such as

YOLOP and HybridNets in both precision and recall, because of the excellent receptive field fusion work of the proposed HDASPP. YOLOMH can better detect objects of different sizes, even small objects of 5 to15 pixels (shown by the yellow circle in Fig 6), which is favorable to the reliability of panoptic sensing systems.

Table 2 Traffic object detection evaluation results.

Network	mAP50(%)	Recall(%)
YOLOV5s	77.2	86.8
YOLOP	76.5	88.2
HybridNets	77.3	89.7
YOLOMH	81.3	91.6

4.1.3 Drivable Area Segmentation Performance

Table 3 shows the evaluation results of drivable area segmentation, using mIoU to evaluate the segmentation performance of different models. Our model achieved an effect of 92.7%, which is 2.2% and 1.2% higher than HybridNets and YOLOP respectively. This is amazing in the subsequent visualization results.

Table 3 Drivable area segmentation evaluation results.

Network	mIoU(%)
MultiNet	71.6
PSPNet	89.6
YOLOP	91.5
HybridNets	90.5
YOLOMH	92.7

4.1.4 Lane Line Segmentation Performance

Table 4 shows the evaluation results of lane line segmentation, and Figure 5 shows the visualization results of lane labels. Since the lane lines in the BDD100K dataset are marked with two lines, the annotation information needs to be converted. We use the lane center as the origin and draw an 8-pixel lane mask for training. The lane width of the test set remains 2 pixels. We use pixel-level accuracy and lane IoU as our evaluation metrics. In terms of accuracy performance: Our model has achieved the best results, with a stunning accuracy improvement compared to the YOLOP network. In terms of IOU performance: Compared with the best-performing HybridNets, our model does not suffer much loss. Compared with the currently popular YOLOP network, our model improves by 3.2%.

4.2 Visualization Performance And Analysis

Fig 7, Fig 8 and Fig 9 show the visualized comparison results between the proposed network and YOLOP network on the BDD100K dataset. In order to illustrate the

Table 4 Lane line segmentation evaluation results.

Network	Acc(%)	IoU(%)
ENet	34.12	14.64
SCNN	35.79	15.84
ENet-SAD	36.56	16.02
YOLOP	70.50	26.20
HybridNets	85.40	36.60
YOLOMH	87.72	29.40

effectiveness of the improvement, we compare them under several scenarios such as daytime, dusk, night, rainstorm and backlight.

Figure 7 shows the results in the daytime scene. Scene 1 shows large false negatives in the drivable area. It missed the drivable area ahead of the road and misidentified one vehicle object on the left side of the image as two vehicle objects. Scene 2 shows false detections and omissions on the traffic object task, and the detection results for lane lines are not continuous. Scene 3 has some false negatives on the drivable area and lane line detection, it misses to detect part of the drivable area and lane line information, and misses to detect the black vehicle object on the left side of the picture.

Figure 8 shows the effect of the night scene. Scene 1 has false positives on the traffic object task, which falsely detects one vehicle as two, and has lane line miss-detection. Scene 2 has the omission of vehicles with lights on on the left side of the picture, and the model omits some lane line information. Scene 3 has a large-scale traffic vehicle and lane line omission detection problem, and it can be inferred that the YOLOP network has not learned effective feature information of traffic vehicles and road lanes in the dark night scene.

Figure 9 shows the effect in dusk, rainstorm and backlight scene. Scene 1 is dusk, and the model has traffic object and lane line leakage detection problem. Scene 2 is a rainstorm, which is a good example of the drawbacks of YOLOP, as the model misses relatively fuzzy information about vehicle objects and lane lines, and there is also a misdetection problem in the drivable area task. Scene 3 is a dark backlight environment, YOLOP misdetects distant traffic objects and handles the segmentation task imperfectly.

Fig. 7 The day-time results

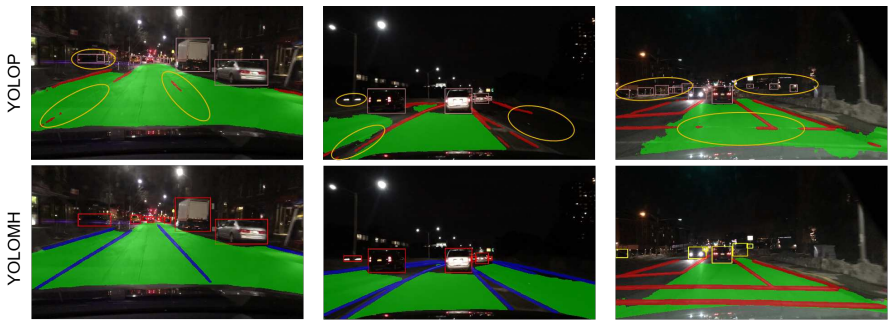

Fig. 8 The night-time results

Fig. 9 The dusk, rainstorm and backlight results

After visual analysis, it is not difficult to find that when the YOLOP network performs panoptic driving perception tasks, there are misdetections, missed detections and inaccurate prediction boxes of road objects, missed detection and disconnection of road lane lines, over and under detection of drivable areas on the road, etc. question. On the contrary, the proposed YOLOMH can well improve the above problems and achieve excellent detection results in various complex scenes such as dusk, heavy rain and backlight. As shown in the second picture in Figure 7 and Figure 9, the model can accurately identify the vehicle objects transported on the car and better complete the panoptic perception task in the blurred environment of heavy rain. In contrast, YOLOP fails to achieve this, which indicates that the proposed model provides better performance in various scenes.

4.3 Ablation Studies

We designed some experiments to verify the effectiveness of our work, including the effectiveness of data enhancement and structural improvement. All experimental configurations and indicators are consistent with the previous ones.

4.3.1 Origin VS Mosaic

As shown in Table 5, we introduced random Mosaic data augmentation in the experiment to improve the accuracy and generalization ability of YOLOMH. Origin is

standard configuration and uses image enhancement approaches such as random brightness and contrast, and more. Taking the traffic object detection task as an example, after the model was enhanced with Mosaic data, the accuracy increased by 1.1% and the recall increased by 3.1%. This experiments show that using data enhancement approaches is beneficial to the learning of multi-task networks.

Table 5 Mosaic data enhanced ablation experiments.

Origin	Mosaic	mAP50(%)	Recall(%)
✓		80.2	88.5
✓	✓	81.3**(+1.1)**	91.6**(+3.1)**

4.3.2 Origin VS HDASPP&Anchor-free

To highlight the effectiveness of our network, we compare the Origin(FPN&anchor-base) approach with the newly proposed approach(HDASPP&anchor-free), and both experiments use yolov8s as the feature extraction network. Since the improvements are mainly focused on traffic object detection, we perform the comparison on the traffic object detection task. Table 6 shows the results of our ablation experiments. It can be found that HDASPP can obtain receptive fields with different scales required for different objects (as shown in Fig 6), and the combination of the anchor-free detection idea allows the network to have a better recall, to find more positive examples of detection, and to reduce the leakage of detection in the detection task. The model also has further improvement in inference speed, which is favorable to the practical application of panoptic driving network.

Table 6 HDASPP&anchor-free ablation experiments.

Origin	HDASPPAnchor-free	mAP50(%)	Recall(%)	speed(FPS)
✓		80.9	87.1	875
	✓	81.3**(+0.4)**	91.6**(+4.5)**	921**(+46)**

5 Conclusion

In the field of autonomous driving, this paper optimizes the problems existing in the current YOLOP network and proposes an efficient end-to-end multi-task perception network YOLOMH. It uses a shared encoder and three independent and simultaneous decoding heads to complete the three major sensing tasks of traffic object detection, drivable area segmentation and lane detection in panoptic driving perception. Through experiments, this paper first considers the lightweight and practicality of the multi-task model, uses yolov8s as the backbone network to complete the image feature extraction work, uses the proposed HDASPP module to efficiently complete the feature fusion work of the neck network, and uses anchor-free idea to avoid redundant NMS to further lightweight the network. Secondly, this paper reveals that in the process of

multi-task network learning, there are differences in the feature information required for different perception tasks, which usually requires experimental adjustments to obtain the optimal structure. More importantly, the network proposed in this paper has lower latency than previous multi-task network models, which greatly improves the possibility of using panoptic driving networks in autonomous driving scenes.

References

[1] Wu D, Liao M W, Zhang W T, et al. Yolop: You only look once for panoptic driving perception[J]. Machine Intelligence Research, 2022, 19(6): 550-562.

[2] Wang J, Wu Q M, Zhang N. You Only Look at Once for Real-time and Generic Multi-Task[J]. ar**v preprint ar**v:2310.01641, 2023.

[3] Liu L, Lu S, Zhong R, et al. Computing systems for autonomous driving: State of the art and challenges[J]. IEEE Internet of Things Journal, 2020, 8(8): 6469-6486.

[4] Zou Z, Chen K, Shi Z, et al. Object detection in 20 years: A survey[J]. Proceedings of the IEEE, 2023.

[5] 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.

[6] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.

[7] 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.

[8] 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.

[9] Redmon J, Farhadi A. Yolov3: An incremental improvement[J]. ar**v preprint ar**v:1804.02767, 2018.

[10] Bochkovskiy A, Wang C Y, Liao H Y M. Yolov4: Optimal speed and accuracy of object detection[J]. ar**v preprint ar**v:2004.10934, 2020.

[11] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. ar**v preprint ar**v:2107.08430, 2021.

[12] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[C]//Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 7464-7475.

[13] Terven J, Cordova-Esparza D. A comprehensive review of YOLO: From YOLOv1 to YOLOv8 and beyond[J]. ar**v preprint ar**v:2304.00501, 2023.

[14] Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation[C]//Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer International Publishing, 2015: 234-241.

[15] 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.

[16] 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.

[17] Parashar A, Rhu M, Mukkara A, et al. SCNN: An accelerator for compressed-sparse convolutional neural networks[J]. ACM SIGARCH computer architecture news, 2017, 45(2): 27-40.

[18] 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.

[19] Yu F, **an W, Chen Y, et al. Bdd100k: A diverse driving video database with scalable annotation tooling[J]. ar**v preprint ar**v:1805.04687, 2018, 2(5): 6.

[20] 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.

[21] Yu F, Koltun V, Funkhouser T. Dilated residual networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 472-480.

[22] Howard A, Sandler M, Chu G, et al. Searching for mobilenetv3[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2019: 1314-1324.

[23] 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. [24]Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88: 303-338.

[24] Everingham M, Van Gool L, Williams C K I, et al. The pascal visual object classes (voc) challenge[J]. International journal of computer vision, 2010, 88: 303-338.

[25] Jampani V, Sun D, Liu M Y, et al. Superpixel sampling networks[C]//Proceedings of the European Conference on Computer Vision (ECCV). 2018: 352-368.

[26] Xu H, Wang S, Cai X, et al. Curvelane-nas: Unifying lane-sensitive architecture search and adaptive point blending[C]//Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16. Springer International Publishing, 2020: 689-704.

[27] Han C, Zhao Q, Zhang S, et al. Yolopv2: Better, faster, stronger for panoptic driving perception[J]. ar**v preprint ar**v:2208.11434, 2022.

[28] He K, Gkioxari G, Dollár P, et al. Mask r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2961-2969.

[29] Vu D, Ngo B, Phan H. Hybridnets: End-to-end perception network[J]. ar**v preprint ar**v:2203.09035, 2022.

[30] Tan M, Pang R, Le Q V. Efficientdet: Scalable and efficient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 10781-10790.

[31] Chen L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[32] Zhou X, Wang D, Krähenbühl P. Objects as points[J]. ar**v preprint ar**v:1904.07850, 2019.

[33] 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.

[34] Zhao H, Gallo O, Frosio I, et al. Loss functions for image restoration with neural networks[J]. IEEE Transactions on computational imaging, 2016, 3(1): 47-57.

[35] Huang X, Wang X, Lv W, et al. PP-YOLOv2: A practical object detector[J]. ar**v preprint ar**v:2104.10419, 2021.

[36] 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.

[37] Li J, Chen J, Sheng B, et al. Automatic detection and classification system of domestic waste via multimodel cascaded convolutional neural network[J]. IEEE transactions on industrial informatics, 2021, 18(1): 163-173.

Liu Fang received the B.S. and Ph.D. degrees from Northeastern University, Shenyang, China, in 2006 and 2012, respectively. She is currently an Associate Professor with the Tianjin Key Laboratory of Autonomous Intelligent Technology and System, Tiangong University, Tianjin, China. Her research interests include computer vision, intelligent-assisted driving, modeling of complex systems, identification, and optimization.

Miao Jianxi received the bachelor's degree in engineering from the Wannan University of Technology, Anhui, China, in 2021. He is currently pursuing the master's degree in software engineeringwith the Tianjin Key Laboratory of Autonomous Intelligent Technology and System, Tiangong University, Tianjin, China. He research interests include autonomous driving, visual perception system.

Sun Bowen received the bachelor's degree in engineering from the Shenyang University of Technology, Liaoning, China, in 2023. He is currently pursuing the master's degree in software engineering with the Tianjin Key Laboratory of Autonomous Intelligent Technology and System, Tiangong University, Tianjin, China. He research interests include autonomous driving, visual perception system.

Su Weixing received the B.S. and M.S. degrees from Northeastern University, Shenyang, China, in 2003 and 2006, respectively, and the Ph.D. degree from the University of Chinese Academy of Sciences, Beijing, China, in 2014. He is currently a Professor with the Tianjin Key Laboratory of Autonomous Intelligent Technology and System, Tiangong University, Tianjin, China. His research interests include autonomous driving, intelligent-assisted driving, and smart manufacturing technology.