Journal of Electronic Research and Application, 2025, Volume 9, Issue 4
http://ojs.bbwpublisher.com/index.php/JERA
ISSN Online: 2208-3510
ISSN Print: 2208-3502

Research on Image Perception Technology of Autonomous Driving Vehicles Based on Deep Learning

Guangyin Xiong*

Shenzhen Desai Xiwei Automotive Electronics Co., LTD., Shenzhen 518000, China

**Author to whom correspondence should be addressed.*

Abstract: This paper introduces autonomous driving image perception technology, including deep learning models (such as CNN and RNN) and their applications, analyzing the limitations of traditional algorithms. It elaborates on the shortcomings of Faster R-CNN and YOLO series models, proposes various improvement techniques such as data fusion, attention mechanisms, and model compression, and introduces relevant datasets, evaluation metrics, and testing frameworks to demonstrate the advantages of the improved models.

Keywords: Autonomous driving; Image perception; Deep learning

Online publication: August 7, 2025

1. Introduction

With the rapid development of autonomous driving technology, the demand for vehicle image perception technology is increasing day by day. The "Intelligent Vehicle Innovation Development Strategy" released in 2020 emphasizes the importance of autonomous driving technology and strict requirements for its safety and reliability. Deep learning plays a crucial role in image perception in autonomous driving, with convolutional neural networks and recurrent neural networks providing fundamental technical support for image perception. The combination of object detection, semantic segmentation, and instance segmentation technologies provides critical information for vehicle decision-making. However, traditional image processing algorithms have many bottlenecks, and commonly used deep learning models also have limitations. Therefore, improving the image perception technology of autonomous vehicles is crucial, including research on integrating multi-sensor data, optimizing model structure, improving real-time and accuracy, etc., to meet policy requirements and practical application needs.

2. Theoretical basis of autonomous driving image perception technology

2.1. Basic principles of deep learning

Deep learning plays an important role in image perception technology for autonomous driving. Convolutional Neural Network (CNN) is a deep learning model specifically designed for processing data with grid structures, such as images. It performs convolution operations by sliding the convolution kernels in the convolutional layer on the image, automatically extracting local features of the image, greatly reducing the number of model parameters and improving the efficiency of feature extraction while reducing computational complexity [1]. The pooling layer further compresses and abstracts features, enhancing their robustness. Recurrent neural networks (RNNs) and their variants, such as long short-term memory networks (LSTM) and gated recurrent unit (GRU), are primarily used for processing sequential data. In autonomous driving image perception, such as processing video streams, RNN can utilize historical information to better understand the image content of the current frame. By modeling image features in the temporal dimension, it provides more comprehensive information support for subsequent decisions.

2.2. Key technologies of image perception

Object detection technology aims to identify the location and category of specific targets in an image. In autonomous driving, it is used to detect objects such as vehicles, pedestrians, and traffic signs, providing key information for vehicle decision-making [2]. Semantic segmentation is the process of classifying each pixel in an image and determining its semantic category, such as roads, sky, vegetation, etc., to help vehicles understand the layout of their surrounding environment. Instance segmentation combines the characteristics of object detection and semantic segmentation, not only identifying different object instances, but also accurately segmenting the pixels of each instance. In the context of autonomous driving, these technologies work together. For example, object detection first identifies potential targets, semantic segmentation further clarifies the surrounding environment of the target, and instance segmentation processes individual targets more finely, jointly providing accurate image perception information for autonomous vehicles and ensuring driving safety.

3. Limitations analysis of existing image perception methods

3.1. The bottleneck of traditional image processing algorithms

Traditional image processing algorithms face many bottlenecks in the image perception of autonomous vehicles. In scenes with changing lighting conditions, the brightness, contrast, and color features of the image will undergo significant changes. Traditional algorithms based on fixed thresholds and feature extraction are difficult to adapt to these dynamic changes, resulting in a significant decrease in target recognition accuracy [3]. For target occlusion situations, traditional algorithms often rely on complete target contours and features for recognition. When the target is partially occluded, the effective features that can be extracted decrease, making it difficult to accurately determine the category and position of the target. These limitations result in poor image perception performance of traditional image processing algorithms in complex scenes, making it difficult to meet the high-precision and high-reliability image perception requirements of autonomous vehicles.

3.2. Defects in existing deep learning models

Faster R-CNN and YOLO series models are commonly used deep learning image perception models, but they have certain limitations. Although Faster R-CNN has high accuracy, its complex network structure and multi-stage detection process result in slow detection speed, making it difficult to meet the real-time requirements of autonomous driving [4]. The YOLO series models have certain advantages in real-time performance, but their accuracy may decrease when dealing with small or dense targets. For example, YOLOv3 may experience missed or false detections when detecting small-sized objects. This is because the resolution of its feature maps is limited, making it difficult to capture detailed information about small targets. Meanwhile, as the density of targets increases, the model may confuse the bounding boxes of different targets, thereby affecting the accuracy of detection.

4. Construction of an improved image perception model

4.1. Design of multi-modal data fusion architecture

4.1.1. Camera and radar data fusion mechanism

In the research of image perception technology for autonomous vehicles, a cross-modal perception framework based on spatiotemporal alignment and feature-level fusion is proposed for the fusion mechanism of camera and radar data. By precise spatiotemporal alignment, the consistency of camera and radar data in both temporal and spatial dimensions is ensured, providing an accurate foundation for subsequent fusion. On this basis, feature-level fusion is performed to fully explore the feature information of the two data modalities. Using deep learning algorithms to learn and analyze the fused features improves the perception ability of the environment. This fusion mechanism can effectively compensate for the shortcomings of a single sensor, improve the accuracy and reliability of image perception, and provide strong guarantees for the safe driving of autonomous vehicles [5].

4.1.2. Attention-enhanced feature extraction network

Building a multi-scale feature fusion module based on channel attention is a key component of an attention-enhanced feature extraction network. This module aims to effectively integrate feature information from different scales, while utilizing channel attention mechanisms to highlight important feature channels. By designing multi-scale convolution kernels, it is possible to capture targets and detailed information of different sizes in the image, and then fuse these multi-scale features. During this process, the channel attention mechanism can adaptively learn the importance weights of each channel and perform weighted fusion of features from different channels [6]. This can make the model pay more attention to the feature channels related to the target, suppress the interference of irrelevant information, thereby improving the effectiveness and accuracy of feature extraction, and further enhancing the performance of autonomous vehicle image perception.

4.2. Optimization strategy for lightweight models

4.2.1. Compression method for knowledge distillation model

In the knowledge distillation model compression method, model parameter level compression can be achieved by designing a teacher-student network architecture [7]. Teacher networks are typically large and well-performing pre-trained models with rich knowledge. The student network is relatively small and structurally simple. During the training process, students learn to mimic the output of the teacher's network through online learning. The soft labels of teacher networks contain more information between categories, which helps student networks better learn the distribution of data. At the same time, by minimizing the difference between the outputs of the student network and the teacher network, the student network can learn as much knowledge and features as possible from the teacher network while significantly reducing the number of parameters, thereby improving the performance and efficiency of the model to meet the demand for lightweight models in autonomous vehicle image perception.

4.2.2. Dynamic calculation path selection mechanism

In the construction of improved image perception models, lightweight model optimization strategies and dynamic computation path selection mechanisms are crucial. For lightweight model optimization, various methods can be used to compress model parameters, reduce computational complexity while maintaining performance. For example, quantization techniques can be used to compress the weights of the model [8]. In terms of dynamic calculation path selection mechanism, the complexity of the scene needs to be considered. Design an adaptive computing resource allocation algorithm based on scene complexity, dynamically adjusting the computing path according to different scene characteristics. In simple scenarios, choose a more streamlined computing path to improve efficiency; In complex scenarios, switching to more complex but accurate calculation paths ensures the accuracy of image perception, thereby improving the overall performance of autonomous vehicle image perception technology.

5. Experimental verification and effectiveness evaluation

5.1. Experimental dataset and training configuration

5.1.1. Construction of KITTI and Cityscapes datasets

The KITTI dataset contains rich image data of autonomous driving scenarios, covering different weather and lighting conditions. The construction process emphasizes the authenticity and diversity of data to better simulate the actual driving environment. In terms of data annotation, strict standards are followed to accurately annotate target objects in images such as vehicles, pedestrians, traffic signs, etc. The annotation information includes location, category, etc. At the same time, in order to expand the size of the dataset and improve the generalization ability of the model, data augmentation strategies such as random cropping, rotation, flipping, and other operations were adopted. The Cityscapes dataset focuses on urban street scenes and also has detailed data annotation standards. It provides high-quality semantic segmentation annotation, which is of great value for studying the perception of autonomous driving images in urban environments. By utilizing these two datasets for model training, the performance of autonomous vehicle image perception technology can be effectively improved [9].

5.1.2. Distributed training parameter configuration

In terms of distributed training parameter configuration, we have adopted a multi-GPU parallel training hyperparameter optimization scheme for autonomous vehicle image perception technology. For the learning rate, we have conducted multiple experiments [10] and determined a suitable value that can ensure the convergence speed of the model while avoiding gradient explosion. At the same time, the batch size setting fully considers the GPU's memory limitations and data diversity to improve training efficiency. The choice of optimizer is also crucial. We compared the performance of various common optimizers on this task and ultimately determined an optimizer that performs well in both convergence speed and model accuracy. In addition, in distributed training, we reasonably set the communication parameters between various GPUs to ensure efficient information transmission and collaborative completion of model training tasks, thereby improving the accuracy and robustness of the entire model's perception of autonomous vehicle images.

5.2. Evaluation indicators and comparison benchmarks

5.2.1. mAP and IoU evaluation system

MAP (Mean Average Precision) and IoU (Intersection over Union) are important indicators for evaluating image perception technology in autonomous vehicles. MAP comprehensively considers the matching degree between predicted boxes and real boxes, as well as the detection accuracy of objects of different categories. The overall performance of the model is measured by calculating the average of AP (average accuracy) for multiple categories. A higher mAP value means that the model has higher accuracy and recall when detecting various objects. IoU focuses on measuring the degree of overlap between the predicted box and the real box, and its calculation formula is the intersection area of the predicted box and the real box divided by the union area. Usually, the higher the IoU threshold, the closer the predicted result is to the true situation. In experiments, by comparing the mAP and IoU values of different algorithms or models, their performance in image perception tasks can be intuitively evaluated, providing a basis for technical improvement and optimization.

5.2.2. Real time FPS testing framework

A scientifically reasonable experimental plan needs to be designed for the real-time FPS testing framework of autonomous vehicle image perception technology. In terms of system delay measurement, it is necessary to accurately record the time interval from input image data to output perception results. By constructing a dataset that simulates real driving scenarios, inputting it into a deep learning model, and using high-precision timers to obtain timestamps for the entire processing process. Compare the FPS values of different algorithms and models on the same dataset as a key evaluation metric for real-time performance. In addition, it is necessary to consider the impact of different hardware platforms on real-time performance to ensure that the test results accurately reflect the performance of the algorithm itself. Through such a testing framework, the real-time performance of autonomous vehicle image perception technology can be comprehensively and objectively evaluated, providing strong support for further optimization and development of the technology.

5.3. Comparative analysis of experimental results

5.3.1. Comparison with traditional methods

In rainy and foggy weather scenarios, the deep learning based autonomous driving vehicle image perception technology proposed in this study presents significant advantages compared to traditional methods. Traditional methods often face a significant decrease in image recognition accuracy under such complex weather conditions. Through unique algorithms and model structures, this technology can more effectively extract image features and accurately recognize targets. Specifically, there is a 10.8% improvement in mean accuracy (mAP). This improvement indicates that the technology can better adapt to environmental changes in rainy and foggy weather, providing more reliable image perception information for autonomous vehicles, thereby improving the safety and stability of autonomous driving and demonstrating powerful performance beyond traditional methods.

5.3.2. Comparison with existing models

The improved model was validated on the Tesla T4 platform, achieving real-time performance of 45FPS. Compared with existing models, it has advantages in both accuracy and real-time image perception. Existing models may experience inaccurate perception in complex environments, while improved models can more accurately identify targets such as roads, vehicles, and pedestrians by optimizing algorithm structure and parameters. In terms of real-time performance, some existing models may be limited by computing resources and algorithm complexity, and cannot meet the real-time requirements of autonomous driving. Improving models can achieve higher frame rates while ensuring accuracy, better adapting to the real-time perception needs of autonomous vehicles, and providing more effective image perception solutions for the development of autonomous driving technology.

6. Conclusion

The image perception system for autonomous vehicles based on an improved deep learning architecture has multiple advantages. It can process image information more efficiently, accurately identify various road elements and obstacles, and provide a reliable decision-making basis for vehicle driving. In complex urban road scenarios, this technology can effectively cope with changing road conditions and environments, improving driving safety and comfort. For example, it can accurately identify the location and dynamics of vehicles and pedestrians during traffic congestion. In the future, further research can be conducted on how to better adapt the system to factors such as weather changes and differences in lighting in dynamic environments. Exploring more advanced algorithms to achieve precise tracking and trajectory prediction of multiple targets in order to improve the accuracy of multi-target tracking, thus promoting the development of autonomous driving technology towards a smarter and safer direction.

Disclosure statement

The author declares no conflict of interest.

References

[1] Wen LH, Jo KH, 2022, Deep Learning-Based Perception Systems for Autonomous Driving: A Comprehensive Survey. Neurocomputing, 489: 255–270.

[2] Huang Y, Chen Y, 2020, Autonomous Driving with Deep Learning: A Survey of State-of-Art Technologies. arXiv preprint, arXiv: 2006.06091.

[3] Grigorescu S, Trasnea B, Cocias T, et al., 2020, A Survey of Deep Learning Techniques for Autonomous Driving. Journal of field robotics, 37(3): 362–386.

[4] Jebamikyous HH, Kashef R, 2022, Autonomous Vehicles Perception (AVP) Using Deep Learning: Modeling, Assessment, and Challenges. IEEE Access, 10: 10523–10535.

[5] Fayyad J, Jaradat MA, Gruyer D, et al., 2020, Deep Learning Sensor Fusion for Autonomous Vehicle Perception and Localization: A Review. Sensors, 20(15): 4220.

[6] Alaba SY, Ball JE, 2023, Deep Learning-Based Image 3-D Object Detection for Autonomous Driving. IEEE Sensors Journal, 23(4): 3378–3394.

[7] Li G, Yang Y, Qu X, et al., 2021, A Deep Learning Based Image Enhancement Approach for Autonomous Driving at Night. Knowledge-Based Systems, 213: 106617.

[8] Zhang J, Cao J, Chang J, et al., 2023, Research on the Application of Computer Vision Based on Deep Learning in Autonomous Driving Technology, International Conference on Wireless Communications, Networking and Applications, Springer Nature Singapore, Singapore, 82–91.

[9] Lee DH, Chen KL, Liou KH, et al., 2021, Deep Learning and Control Algorithms of Direct Perception for Autonomous Driving. Applied Intelligence, 51(1): 237–247.

[10] Tahir NM, Bature UI, Baba MA, et al., 2020, Image Recognition Based Autonomous Driving: A Deep Learning Approach. Int. J. Eng. Manuf, 10(6): 11–19.