Research Progress on Autonomous Driving of Automobiles Based on Deep Learning

Zibo Yuan

Sino-European School of Technology of Shanghai University, Shanghai University, Shanghai, 201900, China

Abstract. With the wide application of deep learning in fields such as image processing and computer vision, research on autonomous driving based on deep learning has become a research hotspot in autonomous driving technology. In deep learning, the application of Convolutional Neural Network models can significantly enhance the performance of autonomous driving in environmental perception. Combined with multi-sensor fusion technology, it has considerably improved the accuracy of target detection and semantic segmentation. Based on the information output from the perception layer, at the decision-making and control level, the Recurrent Neural Network model and the Transformer model further play a role in optimizing path decision-making and vehicle control strategies. Finally, through an end-to-end algorithmic framework, imitation learning and reinforcement learning are integrated to achieve the synergy of perception, decision-making and control, thereby significantly enhancing the overall performance of the autonomous driving system. This article will conduct a review of the application of the above-mentioned models and point out the challenges faced by deep learning based on the actual application scenarios. On this basis, it concludes that future research on autonomous driving based on deep learning will develop in the directions of model lightweighting, model refinement, and multi-technology integration.

Keywords: Deep Learning; Autonomous Driving; Perception and Decision-Making; Vehicle Control; End-To-End.

1. Introduction

Autonomous driving is a typical application scenario of deep learning. People are prone to safety issues such as fatigue and improper operation during driving, while autonomous driving has the advantages of reducing driving pressure and ensuring driving safety. In recent years, the rapid development of computer vision technology has brought about a huge breakthrough in the research and development of autonomous driving systems. Through multiple functional modules such as environmental perception, path decision-making, and vehicle control, autonomous driving systems have effectively improved the driving performance of vehicles in complex urban roads, highways, and adverse weather conditions. This cutting-edge technology is attracting extensive attention from both the academic and automotive industries.

In terms of environmental perception, the Convolutional Neural Network (CNN) algorithm architecture demonstrates extraordinary performance. This algorithm can accurately identify various key targets on the road, such as vehicles, pedestrians, and road traffic signs and markings, and deeply understand the semantic information of the scene. The Recurrent Neural Network (RNN) model and the Transformer model have the ability to process time series data. By analysing the temporal sequence features of consecutive frames, they provide a more scientific basis for vehicle trajectory prediction and behavior control command generation. To ensure system reliability, a multi-sensor fusion solution based on sensors such as cameras, millimeter-wave radars, and Lidars has been formed. Additionally, an end-to-end learning framework that directly generates control signals from raw sensor data further optimizes the system's generalization ability.

This article aims to sort out the various research progress of automotive autonomous driving based on deep learning, to demonstrate the current application of deep learning in the field of automotive autonomous driving. By focusing on three major neural networks and other deep learning technologies, it comprehensively describes the technologies that have been used in today's autonomous driving technology. This article can provide readers with basic information about the

research field of automotive autonomous driving and play a role in summarizing and organizing theoretical research in this field.

2. Key Technologies of Deep Learning Applications

2.1. Application of Convolutional Neural Network in Object Detection

Object detection technology, as one of the core technologies in computer vision, aims to locate and identify specific targets in images. Traditional object detection methods have the limitation that manually extracted features are difficult to cope with complex scenarios. Therefore, this technology has gradually developed into the application of deep learning models [1].

2.1.1. One-stage method

The Convolutional Neural Network consists of three parts. The convolutional layer is responsible for extracting image features, the pooling layer is responsible for dimensionality reduction and preventing overfitting, and the fully connected layer is responsible for outputting results. CNN is a deep learning model existing in the field of computer vision. The design inspiration of this model comes from the visual system in biology, aiming to simulate human visual processing.

Compared with two-stage object detection algorithms such as the Region Convolutional Neural Network (R-CNN), the Single Shot Detector (SSD), You Only Look Once (YOLO) and RetinaNet, which are one-stage object detection algorithms that directly classify and regress based on convolutional features, effectively increasing the speed of real-time detection and achieving a transition from two-stage methods to one-stage methods 错误!未找到引用源。2-5]. Therefore, Convolutional Neural Networks are often applied in scenarios with an excessive amount of data. They can reduce a large number of parameters to a small number of parameters while retaining the discriminative features of images, thereby simplifying complex problems [6].

The YOLO model is widely used in one-stage methods. As a single CNN architecture, the YOLO model can perform real-time object detection with high speed and accuracy. Taking YOLOv5 as an example, this model uses CSPDarknet as the backbone network, supports SPP modules and PANet networks, and provides multiple model sizes to meet the hardware requirements of various devices. At the same time, YOLOv5 has adaptive anchor boxes and the ability to automatically learn the best input size [7]. Combining the above advantages, YOLOv5 is applied to L2-level ADAS, capable of supporting functions such as traffic sign recognition and forward collision warning. However, although one-stage detectors have a fast detection speed, they still have the problem of relatively low detection accuracy [5].

2.1.2. Two-stage method

Due to the low accuracy issue of the one-stage method, in target detection, the two-stage target detection algorithm Faster Region Convolutional Neural Network (Faster R-CNN) is often used in combination and multi-sensor input is added to ensure that the vehicle can complete high-precision 3D target detection and have the ability to recognize small targets in complex scenes during autonomous driving.

However, point cloud datasets obtained from sensors such as cameras and LiDAR have characteristics such as a non-structured, unordered and sparse nature, which cannot be directly used as input information in Faster R-CNN. To apply point cloud data to the input end of Faster R-CNN, there are currently three popular methods: one is point cloud voxelization, which discretises the three-dimensional space into small cubes; the second is the planar projection method, which maps three-dimensional point cloud data onto a two-dimensional plane and converts it into a structured image format, such as top view and front view; the third is to use point cloud deep learning architectures, such as PointNet++ and PointCNN, to extract features and perform pattern recognition on point clouds. After special processing of point cloud data using the above methods, combined with Sparse Convolutional Network and Point Convolution, the purpose of avoiding computational waste caused

by excessive empty voxels can be achieved. At the same time, the local features of point cloud data in three-dimensional space are also retained [8].

Ultimately, Faster R-CNN will output relatively precise 3D bounding boxes, providing parameters for the surrounding environment of the vehicle and a theoretical basis for subsequent path planning and vehicle control.

2.1.3. Semantic Segmentation

The features learned by deep learning methods based on CNN have the advantage of being more accurate in terms of target semantic information. Through multi-level convolutional structures and skip connections, deep learning methods can incorporate global context information into the semantic segmentation process [9]. So as to achieve the purpose of consistency and accuracy in the segmentation results [10].

Up to now, CNN has been widely applied in image segmentation and semantic analysis. In the field of autonomous driving, semantic segmentation has typical application methods such as road scene understanding and dynamic target analysis. At the level of road scene understanding, algorithms like Deeplabv3+ and BiseNetV2 can precisely divide road, sidewalk and obstacle areas, providing a theoretical basis for subsequent path planning. At the level of dynamic target analysis, for dynamic targets such as vehicles and pedestrians, instance segmentation methods like Mask R-CNN and YOLOv8-Seg demonstrate advantages in efficiency and accuracy. Meanwhile, to strengthen the detection ability of small targets, FPN feature pyramid and Convolutional Block Attention Module (CBAM) are widely adopted.

In conclusion, the application of CNN has become one of the core technologies in the field of semantic segmentation. For different segmentation scenarios, scholars have proposed diverse network architectures and optimization strategies. It can be said that the progress of semantic segmentation technology has laid a solid foundation for the future intelligence of autonomous driving.

2.2. Application of Convolutional Neural Network in Object Detection

2.2.1. The Application of Recurrent Neural Networks and Transformer Models

After using CNN to solve the problem of object detection in autonomous driving, Recurrent Neural Networks (RNN) and Transformer models will be employed to address the issues of path decision-making and vehicle control on this basis.

RNN, through its temporal modelling capability, encodes the historical states of vehicle movement with an encoder-decoder architecture and utilizes hidden states to remember key temporal correlation features, extracting the vehicle's position information and motion characteristics to achieve the purpose of vehicle trajectory prediction. The specific process is shown in Figure 1 [11]. However, when dealing with long time series, RNN is greatly affected by gradient explosion and gradient vanishing [11]. To address this issue, there are two types of developments and extensions of RNN: one is Long Short-Term Memory (LSTM) and the other is Gate Recurrent Unit (GRU) [12].

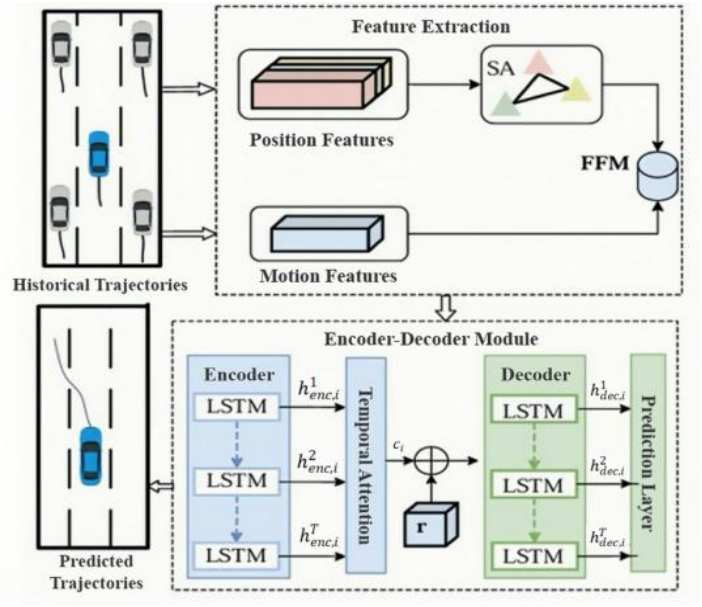


Figure 1. Trajectory prediction route based on RNN[11].

Transformer, with its self-attention mechanism and multi-head attention structure, along with its unique encoder-decoder architecture, excels in context information acquisition, long-distance dependency, and parallel computing [11, 13]. The structure of the Transformer model is shown in Figure 2 [13], which mainly consists of two components: a self-attention mechanism and a feedforward network. The self-attention mechanism establishes associations by calculating the dot product similarity between elements and then performs a weighted sum of context information based on normalized attention weights of other elements, thereby dynamically updating the representation at each position. This mechanism can autonomously capture the internal dependencies within a sequence, effectively addressing the limitations of traditional encoders in modeling long-distance dependencies. The feedforward neural network, on the other hand, performs nonlinear transformations on the representations output by self-attention to further enhance the expressive power of features. As a core component of feature extraction [14], the attention mechanism can precisely identify important nodes and patterns in trajectory changes. The core advantage of this architecture lies in its dynamic weight allocation mechanism, which can automatically adjust the degree of attention to different features based on input characteristics.

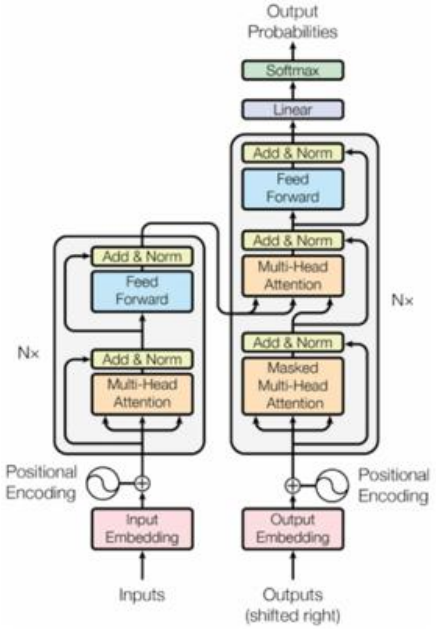


Figure 2. Overview of Transformer [13].

In the path planning and vehicle control of autonomous driving, RNN and Transformer architectures each have their own advantages. RNN, with its ability to model time series, can process continuous sensor input data (such as vehicle speed and steering angle) frame by frame and pass historical information through hidden states, making them suitable for short-term trajectory prediction and low-latency control decisions. However, the serial computing nature of RNN limits its ability to model long sequences, and it is prone to issues like vanishing or exploding gradients. In contrast, the self-attention mechanism of Transformers enables parallel computation of global dependencies, allowing for more efficient modeling of the long-term space-time relationships between vehicles and complex environments, especially demonstrating stronger feature extraction capabilities in multimodal sensor fusion. In complex motion scenarios, Transformers can dynamically focus on key targets through attention weights, while RNNs are more suitable for generating low-level control instructions with high real-time requirements. Based on their different characteristics, in the field of autonomous driving, RNN is often used to handle high-frequency low-level control signals for vehicle control purposes, while Transformers mainly handle high-level semantic understanding to optimize current path planning. Ultimately, the two frameworks form a hierarchical decision-making effect, jointly enhancing the robustness and adaptability of autonomous driving systems in dynamic environments.

2.2.2. The extended applications of the Transformer model

There are currently three main architectures of Transformer models. The first one is the Vanilla Transformer, which is the original Transformer and is based on a pure self-attention mechanism without CNN or RNN structures [15]. The second one is the Vision Transformer (ViT), which adopts an input process specifically designed for images [16]. The input image needs to be divided into fixed-size patches (e.g.16×16 or 32×32 pixels). These image patches are processed through a linear embedding layer and position encoding before all patch sequences are encoded by a standard Transformer encoder. The third one is the Multimodal Transformers, which can utilize the self-attention mechanism to treat each input as a fully-connected graph and process data by capturing global patterns. This inherent characteristic enables the Transformer to be applicable to modality agnostic processes, as each token's embedding is regarded as a node in the graph, thus being compatible with multiple modalities [13].

At present, to meet the increasingly high demands of the autonomous driving car market, researchers have conducted deeper studies based on the Transformer model. The Li team proposed a framework that combines Birds-eye-view (BEV) and Transformer, named BEVFormer. This is a typical BEV + Transformer application framework, which can effectively aggregate space-time features from multi-view cameras and historical BEV features. The input data of BEVFormer only relies on multi-view cameras, making it a pure visual solution. The BEV features generated by this framework can simultaneously support multiple 3D perception tasks such as 3D object detection and map segmentation, which is of great value to autonomous driving systems. BEVFormer has three key designs: first, grid-based BEV queries that flexibly fuse spatial and temporal features through the attention mechanism; second, a spatial cross-attention module that can aggregate spatial features from multi-camera images; third, a temporal self-attention module that extracts temporal information from historical BEV features, which is beneficial for the speed estimation of moving objects and the detection of severely occluded objects, with negligible computational overhead. Li team’s research proves that using space-time information from multi-camera inputs can significantly improve the performance of visual perception models [17].

However, although the pure visual solution has advantages such as low cost, it still has the limitation of large error in long-distance estimation. Based on this, researchers have proposed a multimodal fusion solution, that is, the combination of LiDAR and cameras. To solve the fusion problem of LiDAR and cameras, Bai team proposed Transfusion. The core of Transfusion is an innovative sequential fusion method, that is, using two Transformer decoder layers as the detection head. In Transfusion, the first layer of the decoder generates initial bounding boxes from LiDAR features using a set of sparse target queries; the second layer of the Transformer decoder adaptively

fuses the target queries with relevant useful image features through spatial and contextual relationships. This design optimizes the computational efficiency of detection and achieves the purpose of complementary enhancement of cross-modal features [18].

Transformer, with its powerful global modeling capabilities and parallel computing advantages, is comprehensively transforming the autonomous driving technology system. In the aspects of environmental perception, trajectory prediction, and decision-making planning, Transformer has demonstrated its potential for development and structural superiority. With the continuous development of lightweight Transformer models, the deployment efficiency of this model on in-vehicle computing platforms will continue to improve.

2.2.3. End-to-end algorithm framework

Li team pointed out in their research that traditional autonomous driving systems rely on three modules: environmental perception, path decision-making, and vehicle control. The end-to-end autonomous driving system architecture, as shown in Figure 3 [19], integrates the perception and decision-making modules into a single neural network process. This approach effectively reduces the accumulation of errors between different modules. At the same time, the optimization of the entire system is goal-oriented towards the ultimate realization of dynamic driving tasks, thereby significantly improving the overall performance of the system and avoiding the problem of computational waste [19].

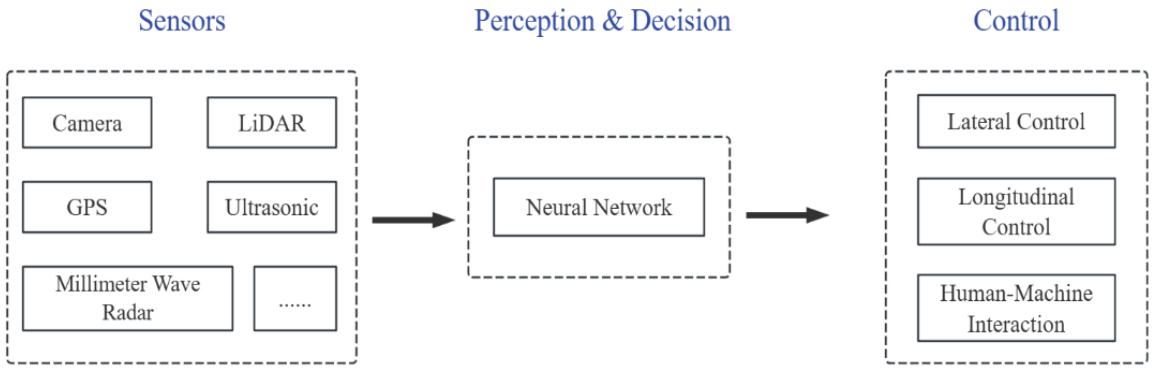


Figure 3 End-to-End Autonomous Driving System Architecture[19].

In end-to-end autonomous driving systems, reinforcement learning (RL) and imitation learning (IL) are two core autonomous learning techniques that enhance system performance by optimizing decision-making strategies and learning human driving behaviors, respectively. The core operation principle of reinforcement learning is as follows: the agent first perceives the current environment state, then makes a decision and takes an action. The environment responds to this action, updates the state, and provides a reward. The agent continuously improves the decision-making mechanism using these interaction data, with the goal of maximizing cumulative rewards. The basic principle of imitation learning is similar to supervised learning, where the task execution method is mastered by analyzing and replicating the decision-making approach of experts. This learning method can directly adopt an overall imitation strategy to approach the operation level of experts and then test the learning effect in actual application scenarios and adjust the focus and direction of imitation according to the actual situation [20].

The end-to-end autonomous driving technology, as an integrated solution, integrates multiple modules such as environmental perception and path planning into a single intelligent system, dramatically improving the overall performance of autonomous driving. What's more, the end-to-end technology can continuously improve itself through training and gradually master the ability to handle complex traffic scenarios on this basis. With the advancement of computing technology and the accumulation of road data, this intelligent driving system is rapidly approaching the level of human driving.

3. Challenges Faced in the Application of Deep Learning

Through the analysis of various models, the development of the autonomous driving field has revealed a series of key technical challenges that need to be urgently addressed. Although complex deep learning models have superior performance, they have a large number of parameters and huge computational demands, which may be limited by the limited resources of on-board devices. At the same time, autonomous driving requires a large amount of data and extreme situations that have not appeared in the training samples may occur during real driving, which may cause safety hazards in the decision-making and control of the vehicle's autonomous driving system. On the safety level, deep learning is a typical black box system, and its decision-making logic faces difficulties in certification in the highly safety-demanding autonomous driving scenarios. On the legal and ethical levels, factors such as accident liability determination and privacy protection also affect the application process of deep learning in the autonomous driving field. The above challenges require the coordinated solution of technological progress and policy regulations.

4. Future Research Directions

Autonomous driving systems are confronted with the complexity of deep learning algorithms and the limited computing resources on board. Therefore, the lightweighting of on-board model architectures is particularly important. On the basis of model optimization, to achieve the continuous evolution of autonomous driving, the same system's autonomous driving vehicles should share the learning system, so that the model optimization in real scenarios can be carried out synchronously. In the face of extreme weather, more diverse sensors are needed as input information to enable the system to make more reasonable driving choices. Currently, multi-modal fusion solutions are being widely tested, that is, on the basis of the fusion of lidar and cameras, infrared sensors are added. In terms of safety verification, the black box nature of deep learning still poses safety certification challenges, and scholars need to prove the safety boundaries of neural networks from both theoretical and practical aspects. Finally, to avoid the high cost of road testing in complex real road conditions, building a simulation platform with a huge amount of data is the trend of future development.

5. Conclusion

Deep learning technology has made remarkable progress in the field of autonomous driving and has become the core driving force for the intelligent development of autonomous driving systems. This paper systematically reviews the application of deep learning models such as CNN, RNN, and Transformers in key links like environmental perception, decision-making planning, and control, revealing the huge potential of deep learning in improving the performance of autonomous driving.

In terms of environmental perception, the visual perception algorithm based on CNN has significantly improved the accuracy of target detection and semantic segmentation, providing an effective solution for understanding complex traffic scenarios. At the decision-making and control level, RNN and Transformer models have achieved more precise path planning and vehicle control strategy optimization through temporal modeling and attention mechanisms. Finally, the end-to-end learning framework has realized the full-process collaborative optimization of perception, decision-making and control through the organic combination of imitation learning and reinforcement learning, greatly improving the overall performance of the system.

References

[1] Li, C., Zhang, Z., Liang, Z., et al.: 'Review of Object Detection Models', J. Comput. Res. Dev., 2025, pp. 1-35

[2] Zhai, S., Shang, D., Wang, S., Dong, S.: 'DF-SSD: An improved SSD object detection algorithm based on DenseNet and feature fusion', IEEE Access, 2020, 8, pp. 24344-24357

[3] Muhammad, H.: 'YOLO-v1 to YOLO-v8, the rise of YOLO and its complementary nature toward digital manufacturing and industrial defect detection', Machines, 2023, 7, (11), pp. 677-697

[4] Cheng, M., Bai, J., Li, L., et al.: 'Tiny RetinaNet: a one-stage detector for real-time object detection', Proc. Int. Conf. Graph. Image Process., 2020, 11373

[5] Yu, X.: 'Research on Key Issues of Road Object Detection for Autonomous Driving in Complex Traffic Scenarios', PhD Thesis, Univ. Chin. Acad. Sci., 2024

[6] Marinó, G.C., Petrini, A., Malchiodi, D., Frasca, M.: 'Deep neural networks compression: A comparative survey and choice recommendations', Neurocomputing, 2023, 520, pp. 152-170

[7] Ning, Q., Wang, Y.: 'Research and Application of YOLO Object Detection Technology', Sci. Technol. Innov., 2025, (9), pp. 26-35

[8] Zhou, Z.: 'Research on LiDAR Point Cloud Object Detection Algorithm Based on Deep Learning', PhD Thesis, Xi'an Univ. Technol., 2024

[9] Chen, L.C., Zhu, Y., Papandreou, G., et al.: 'Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation', Lect. Notes Comput. Sci., 2018, 11211, pp. 801-818

[10] Ye, Z., Zheng, K., Zhang, L., et al.: 'Review of Deep Learning Methods for Image Semantic Segmentation', China Equip. Eng., 2025, (10), pp. 23-25

[11] Liu, X., Chen, H., Yu, D., et al.: 'Review of Trajectory Prediction for Autonomous Vehicles Considering Long-Term and Short-Term Characteristics', Front. Comput. Sci. Technol., 2025, pp. 1-23

[12] Lee, D., Kwon, Y.P., McMains, S., et al.: 'Convolution neural network-based lane change intention prediction of surrounding vehicles for ACC', Proc. IEEE Int. Conf. Intell. Transp. Syst., 2017, pp. 1-6

[13] Xu, P., Zhu, X., Clifton, D.A.: 'Multimodal Learning With Transformers: A Survey', IEEE Trans. Pattern Anal. Mach. Intell., 2023, 45, (10), pp. 12113-12132

[14] Hao, Z., Huang, X., Wang, K., et al.: 'Attention-Based GRU for driver intention recognition and vehicle trajectory prediction', Proc. CAA Int. Conf. Veh. Control Intell., 2020, pp. 86-91

[15] Vaswani, A., et al.: 'Attention is all you need', Adv. Neural Inf. Process. Syst., 2017, 30, pp. 5998-6008

[16] Dosovitskiy, A., et al.: 'An image is worth 16x16 words: Transformers for image recognition at scale', arXiv Prepr., 2020, arXiv:2010.11929

[17] Li, Z., et al.: 'BEVFormer: Learning Bird's-Eye-View Representation from Multi-camera Images via Spatiotemporal Transformers', Lect. Notes Comput. Sci., 2022, 13669, pp. 1-21

[18] Bai, X., et al.: 'TransFusion: Robust LiDAR-Camera Fusion for 3D Object Detection with Transformers', Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2022, pp. 1080-1089

[19] Li, Z., Zhang, X., Zhang, Z.: 'Research on Standardization of End-to-End Autonomous Driving Systems', China Automot., 2024, (11), pp. 3-8

[20] Xue, H.: 'Research on End-to-End Strategy for Autonomous Driving Based on Deep Imitation Reinforcement Learning', PhD Thesis, Xidian Univ., 2024