Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33954

3D Object Detection for Autonomous Driving via Multi-Sensor Fusion and Deep Guidance

Yanzhe He

Chang'an Dublin International Transportation College, Chang'an University, Xi'an, China
p33708790@gmail.com

Abstract. Autonomous driving 3D object detection is a core technology of on-board perception systems. Single-modal sensors are easily constrained by the environment and are difficult to meet the application requirements of complex scenarios. This paper focuses on the research of multi-sensor fusion and depth-guided autonomous driving 3D object detection methods, sorting out the technical status of data layer, feature layer, decision layer fusion and depth-guided monocular detection, point cloud fusion, and comparing the performance and applicable scenarios of different fusion strategies. It is analyzed that the current technology faces challenges such as the difficulty of multi-sensor temporal and spatial synchronization, insufficient robustness of depth extraction, imbalance between accuracy and real-time performance, difficulty in small sample training, and lack of engineering standardization. Looking to the future, multi-modal adaptive fusion, lightweight depth-guided algorithms, and large model empowerment will become the core development directions, which can provide theoretical references for the technical optimization and mass production of autonomous driving 3D object detection.

Keywords: Autonomous driving, 3D object detection, multi-sensor fusion, depth guidance

1. Introduction

Autonomous driving stands as a frontier interdisciplinary field of automotive engineering and artificial intelligence, acting as a core driver for the upgrading and innovation of the global transportation industry. Its large-scale application helps improve traffic efficiency, reduce accident risks, and enhance travel experience. As the fundamental module of autonomous vehicles, the perception system ensures environmental awareness, intelligent decision-making, and autonomous control, in which 3D object detection plays a pivotal role. It is required to quickly and accurately obtain the 3D position, size, category, and motion state of obstacles, vehicles, pedestrians, and other targets in dynamic driving scenarios, providing reliable data support for subsequent planning and control.

At present, 3D object detection mainly relies on single-modal sensors such as cameras, LiDAR, and millimeter-wave radars [1]. However, these sensors are easily affected by illumination changes, bad weather, occlusion, and limited detection range, leading to perception blind spots, low accuracy, and poor robustness, which can hardly meet the demands of complex urban and highway scenarios. In contrast, multi-sensor fusion integrates complementary information from heterogeneous sensors

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33954

to overcome the limitations of single-modal perception. Meanwhile, depth-guided detection combined with point cloud augmentation provides an effective solution to few-shot training and complex-scene recognition.

Given the above background, this paper focuses on 3D object detection for autonomous driving based on multi-sensor fusion and deep guidance. It reviews the state-of-the-art of data-level, feature-level, and decision-level fusion, as well as depth-guided monocular detection and depth-point cloud fusion. It also analyzes key challenges including spatio-temporal synchronization, depth extraction robustness, accuracy-real-time trade-off, small-sample training, and engineering standardization. Finally, this paper prospects future trends such as adaptive multi-modal fusion, lightweight depth-guided algorithms, and large-model empowerment [1]. This paper aims to provide a theoretical reference for performance optimization and mass production of autonomous driving 3D object detection technology.

2. Technical basics of autonomous driving 3D object detection

The technical foundation of autonomous driving 3D object detection consists of on-board perception sensors and core fusion algorithms. Vision cameras, LiDAR and millimeter-wave radars are the mainstream sensors, with complementary performance but inherent defects: cameras are cost-effective and semantically rich but lack depth perception; LiDAR delivers high-precision spatial data but is expensive with sparse point clouds; millimeter-wave radars are weather-resistant but suffer from low resolution [2]. These characteristics determine that single-modal perception cannot meet complex scenario requirements, making multi-sensor fusion and depth guidance indispensable technical routes [3].

Multi-sensor fusion is divided into data-level, feature-level and decision-level fusion according to the integration stage. Data-level fusion retains complete raw information but has ultra-high hardware and spatio-temporal synchronization requirements; feature-level fusion balances accuracy and efficiency by extracting and fusing cross-modal features, which has become the mainstream scheme; decision-level fusion is easy to deploy with strong fault tolerance, but cannot fully exploit sensor complementarity [4]. Depth-guided detection makes up for the 3D perception defects of vision sensors via monocular depth estimation and point cloud depth completion, which is a critical supplement to LiDAR-based detection and lays a technical foundation for low-cost and high-robustness perception.

3. Research status of multi-sensor fusion

Multi-sensor fusion is the core solution to break through single-modal perception limitations, and its technical routes show clear differentiation in performance and engineering implementation.

3.1. Data-level fusion

Data-level fusion focuses on the accurate registration and integration of LiDAR point cloud and vision image original data, mainly through camera-LiDAR calibration to realize pixel and point cloud coordinate mapping [3]. Although this fusion method retains the most comprehensive original information, its huge computation and strict spatio-temporal synchronization requirements make it difficult to meet the real-time demand of vehicle-mounted terminals, so it is only suitable for high-precision test platforms and static environment modeling scenarios with low real-time requirements [5].

DOI: 10.54254/2755-2721/2026.GU33954

3.2. Feature-level fusion

Feature-level fusion achieves cross-modal information interaction through early or late feature fusion, attention mechanism and multi-scale feature fusion, which can balance detection accuracy and real-time performance effectively [6]. Representative methods such as F-PointNet and MSF-3D have verified the advancement of feature fusion, but the fixed fusion paradigm is difficult to adapt to heterogeneous multi-sensor information, and the fusion effect is easily affected by spatio-temporal misalignment, which limits the detection performance in extreme weather and occlusion scenarios.

3.3. Decision-level fusion

Decision-level fusion integrates the independent detection results of each sensor through simple fusion rules, with the advantages of low computation, easy deployment and strong fault tolerance. When a single sensor fails, the system can still maintain basic perception functions, so it is widely used in L2 and below assisted driving systems and redundant detection links [6]. However, this method cannot exert the complementary advantages of multi-sensor information, and its detection accuracy is limited by the performance of single-modal detection, so it is difficult to meet the perception requirements of high-level autonomous driving.

3.4. Performance and applicable scenario

The three fusion strategies show significant differences in accuracy, real-time performance and engineering difficulty. Data-level fusion has the highest accuracy but poor real-time performance; feature-level fusion is the optimal solution for L3 and above autonomous driving in complex dynamic scenes; decision-level fusion is suitable for low-cost and high real-time demand scenarios. The selection of fusion strategy must be matched with the vehicle-mounted hardware computing power, sensor configuration and actual application scenarios, which is the core principle to realize the engineering deployment of fusion perception technology [7].

4. Research status of depth-guided detection

Depth-guided detection optimizes 3D perception performance by introducing depth information constraints, which solves the core pain points of monocular vision lack of depth and LiDAR point cloud sparsity, and has become a key research direction for reducing the cost of perception systems and improving robustness.

4.1. Depth-guided monocular 3D detection

Depth-guided end-to-end monocular 3D detection realizes synchronous depth estimation and 3D target regression, which is divided into single-branch and dual-branch network models. Single-branch models have small parameters and fast inference speed, which are suitable for vehicle-mounted terminals with limited computing power; dual-branch models have higher depth estimation accuracy and stronger detection robustness, but the model complexity is high and the real-time performance is insufficient. At present, the accuracy of depth estimation in complex environments such as strong light and occlusion is still difficult to guarantee, which leads to unstable detection results and becomes the main bottleneck restricting the practical application of this technology.

DOI: 10.54254/2755-2721/2026.GU33954

4.2. Depth-point cloud fusion detection

3D object detection method integrating depth information and point cloud is an important research direction to improve the performance of LiDAR 3D detection. This method takes LiDAR point cloud as the core, uses the depth estimation results of vision images to complete, enhance and semantically constrain sparse point clouds, solving the sparsity problem of LiDAR point clouds in long-distance, occluded and small-target scenarios, and improving the spatial integrity and semantic information richness of point clouds [8].

At present, the research of this method mainly focuses on two core links: point cloud depth completion and depth-point cloud feature fusion. For point cloud depth completion, mainstream methods generate dense depth maps through visual depth estimation, register the depth maps with sparse LiDAR point clouds, and fill the sparse areas of point clouds with depth information to generate dense mixed point clouds, represented by Depth Completion Net, LiDAR-BEV Depth and so on. Some studies also introduce semantic segmentation results to impose semantic constraints on the completed point clouds and improve the point cloud feature representation of small targets. For depth-point cloud feature fusion, it is mainly divided into point cloud-level fusion and feature-level fusion. Point cloud-level fusion directly sends the completed dense point clouds into the 3D detection model for detection, which is simple and easy to implement. Feature-level fusion first extracts the spatial features of completed point clouds and semantic features of visual depth respectively, and then carries out cross-modal feature interaction, with better fusion effect and higher detection accuracy, represented by PC-Depth Fusion, Depth-enhanced Point Pillars and so on.

The 3D object detection method integrating depth information and point cloud greatly improves the accuracy and robustness of LiDAR 3D detection without increasing hardware costs, especially showing significant advantages in detecting small targets (such as pedestrians, non-motor vehicles) and long-distance targets. It has become an important optimization direction for LiDAR 3D object detection.

4.3. Accuracy and real-time optimization

The trade-off between accuracy and real-time performance of depth-guided detection algorithms is a core constrained problem, and also a key optimization direction in this field. Current research is mainly carried out from three dimensions: algorithm design, feature optimization and hardware acceleration, to realize the collaborative improvement of accuracy and real-time performance and meet the engineering application requirements of on-board terminals [9].

The balance between detection accuracy and real-time performance is a core constraint for depth-guided detection algorithms, and it is also a key optimization direction in related research. At present, the real-time performance of such algorithms is generally improved through model lightweight design and computation reduction.

Lightweight network backbones including MobileNet and ShuffleNet are usually adopted to replace traditional ResNet structures, which can effectively reduce model parameters and computational complexity.

Model pruning, quantization and knowledge distillation technologies are widely utilized to compress pre-trained depth-guided detection models, so that the inference speed of models can be greatly improved with minor accuracy loss.

The network structure is further optimized by eliminating redundant feature extraction layers, and the lightweight design of multi-scale feature fusion is adopted to balance feature expression ability and computational efficiency.

DOI: 10.54254/2755-2721/2026.GU33954

5. Key problems and challenges in research

As the core technology of the on-board perception system, the engineering implementation and large-scale application of autonomous driving 3D object detection in complex driving scenarios still face multi-dimensional core technical bottlenecks and engineering practice challenges.

5.1. Spatio-temporal synchronization

The heterogeneity of multi-sensor information and spatio-temporal synchronization problems are fundamental obstacles to fusion perception. Heterogeneous sensors such as on-board vision, LiDAR and millimeter-wave radars have essential differences in data modality, sampling frequency, spatial coordinate system and other aspects. Time synchronization errors and spatial external parameter drift easily cause spatio-temporal misalignment of perception data, seriously restricting the accuracy and reliability of multi-modal fusion perception.

5.2. Depth extraction robustness

Insufficient accuracy and robustness of depth information extraction in complex scenarios is a core performance bottleneck. In complex driving conditions such as extreme light, bad weather and target occlusion, single-modal perception is prone to missed detection and false detection. The depth estimation accuracy of multi-modal fusion models under heterogeneous noise interference is insufficient, which can hardly meet the all-scenario high-reliability perception requirements [6].

5.3. Accuracy-real-time trade-off

The trade-off contradiction between real-time performance and robustness of detection algorithms is hard to reconcile. High-precision 3D object detection models usually have complex network structures, and the inference delay is difficult to adapt to the real-time requirements of on-board systems. Lightweight models tend to sacrifice detection accuracy in complex scenarios, failing to meet the dual requirements of real vehicle application at the same time.

5.4. Small-sample training

Insufficient model generalization ability under small-sample and less-annotated scenarios is a core pain point for open road applications. Autonomous driving scenarios have significant long-tail distribution characteristics, with scarce samples of rare targets and special working conditions. Existing deep learning models rely on large-scale annotated data with limited generalization ability, making it difficult to deal with the unknown target recognition needs under open roads.

5.5. Challenges of algorithm lightweight, engineering and standardization

Algorithm lightweight under the constraint of on-board hardware, as well as engineering and standardization of perception systems, are key challenges for mass production. The computing power, power consumption and storage resources of on-board computing units are strictly constrained. Algorithm lightweight deployment needs to be realized on the premise of ensuring detection accuracy [8]. At the same time, there is a lack of unified perception performance evaluation system and functional safety standards in the industry, leading to high engineering

DOI: 10.54254/2755-2721/2026.GU33954

implementation costs and low efficiency, which seriously restrict the large-scale application of technology.

6. Future development trends and research outlook

As the core technology of the autonomous driving perception system, the future development of autonomous driving 3D object detection will closely focus on breaking through core bottlenecks such as multi-sensor fusion, depth guidance and engineering implementation, iterating towards adaptive fusion, lightweight efficiency and large model empowerment, so as to provide key technical support for the large-scale mass production of L3 and above high-level autonomous driving.

6.1. Adaptive multi-modal fusion

Adaptive fusion technology for multi-modal heterogeneous information will become a core innovation direction. Aiming at the current problems of spatio-temporal synchronization error of multi-sensors and difficult adaptation of information heterogeneity, the future will break through the fixed fusion paradigm of traditional data-level, feature-level and decision-level, and develop end-to-end adaptive fusion algorithms. Through dynamic external parameter calibration and real-time spatio-temporal alignment technology, correct the registration deviation caused by vehicle vibration and hardware aging. Relying on attention mechanism and intelligent weight distribution, realize on-demand fusion of multi-modal information of vision, LiDAR and millimeter-wave radars, maximizing the complementary advantages of sensors [9]. At the same time, fusion algorithms will extend to vehicle-cloud collaboration and road-side collaborative perception, breaking through the limitations of single-vehicle perception vision, and greatly improving the detection robustness in extreme weather and occluded scenarios.

6.2. Lightweight depth-guided algorithms

Lightweight depth-guided detection algorithms will achieve collaborative optimization of accuracy and real-time performance. Facing the strict constraints of on-board hardware computing power and power consumption, the future will focus on lightweight network architecture, model pruning, quantitative distillation and hardware collaborative acceleration technology to build efficient detection models adapted to on-board terminals [10]. For monocular depth guidance and point cloud depth completion algorithms, simplify redundant computing links, retain core depth constraints and feature extraction capabilities, and reduce the inference delay to the range required by on-board systems under the premise of controllable accuracy loss. In addition, depth guidance technology will be deeply integrated with 3D point cloud sample augmentation to solve the perception shortcomings of small and long-distance targets, enabling low-cost perception solutions to meet the perception needs of high-level autonomous driving.

6.3. Large model-enabled detection

Large models will bring a technological paradigm innovation to 3D object detection. Relying on the strong semantic understanding, few-shot learning and generalization capabilities of multi-modal large models, solve the training problems of long-tail distribution and less-annotated samples in autonomous driving scenarios. Extract general environmental features through large model pre-training to provide high-quality prior information for depth estimation and 3D object detection, and

DOI: 10.54254/2755-2721/2026.GU33954

improve the adaptability of models to unknown targets and complex working conditions. In the future, a collaborative architecture of "large model + lightweight perception terminal" will be formed. Large models will make up for the generalization defects of traditional algorithms, and terminal lightweight deployment will ensure real-time performance, promoting 3D object detection to upgrade from data-driven to cognition-driven [11].

To sum up, the future autonomous driving 3D object detection will take multi-sensor adaptive fusion as the foundation, lightweight depth guidance as the core, and large model empowerment as the breakthrough. It will simultaneously overcome technical bottlenecks and engineering problems, improve the perception performance evaluation and functional safety standards, promote the technology from test verification to large-scale commercial use, and provide solid support for the upgrading of the intelligent transportation industry.

7. Conclusion

This paper systematically reviews the technical status of 3D object detection for autonomous driving based on multi-sensor fusion and depth guidance, clarifies the performance characteristics and applicable scenarios of data-level, feature-level and decision-level fusion strategies, and summarizes the research progress of depth-guided monocular detection and depth-point cloud fusion methods. The core challenges in current technologies are identified, including multi-sensor spatio-temporal synchronization, robustness of depth extraction, accuracy-real-time trade-off, small-sample training, and algorithm lightweight and standardization. Meanwhile, the future development directions are prospected, such as adaptive multi-modal fusion, lightweight depth-guided algorithms and large model-enabled detection.

This paper clarifies the technical logic and practical bottlenecks of multi-sensor fusion and depth-guided 3D object detection, which provides clear theoretical references and technical ideas for the performance optimization, engineering deployment and mass production of autonomous driving perception systems. It helps to break the limitations of single-modal perception, promote the coordinated improvement of detection accuracy, real-time performance and environmental robustness, and support the technological iteration of on-board perception for autonomous vehicles.

In the future, with the continuous innovation of adaptive fusion algorithms, lightweight network design and multi-modal large model technologies, the technical bottlenecks of 3D object detection will be gradually resolved. The in-depth integration of multi-sensor fusion, depth guidance and large model empowerment will further enhance the all-scenario perception capability of autonomous driving systems, accelerate the commercial application of L3 and above high-level autonomous driving, and provide solid technical support for the intelligent upgrading of the global transportation industry.

References

[1] Wang Liyong, Cui Ao, Su Qinghua, et al. Research on Real-time Perception System for Temporary Road of Unmanned Vehicles Based on Multi-sensor Data Fusion [J/OL].electron measure technology, 1-10 [2026-02-13].https: //link.cnki.net/urlid/11.2175.TN.20260210.1645.060.

[2] Zhang Ailing, Wang Yafei, Hua Yiding, et al. Research on the Generation of Enhanced Samples for Autonomous Driving 3D Point Clouds and the Standardization of Testing Processes [J]. China Automotive (Bilingual Edition), 2026, 36(01): 24-29.DOI: 10.27018/j.cnki.cqgz.2026.01.012.

[3] Liu Ping, Wang Shuo-han, Zhang Yi-kang, et al. Method for Extracting Motion Information from Vehicle-mounted Visual Images [J]. Journal of Chongqing Jiaotong University (Natural Science Edition), 2026, 45(01): 106-112.

[4] He Yi, Ling Zhiying, Qiu Zhijun, et al. An End-to-End Monocular 3D Object Detection Method for Autonomous Driving Based on Deep Guidance [J/OL]. China Journal of Highway and Transportation, 1-19 [2026-02-13].https:

DOI: 10.54254/2755-2721/2026.GU33954

//link.cnki.net/urlid/61.1313.U.20260112.1913.026.
[5] Li Xingbing, Xu Zhefeng. Research on Non-Motor Vehicle Target Recognition in Traffic Monitoring Images Based on YOLOv5s [J]. Intelligent City, 2025, 11(12): 1-5.DOI: 10.19301/j.cnki.zncs.2025.12.001.
[6] Wu Mengyao. Discussion on the Perception Module of Autonomous Driving Based on Computer Vision [J]. Automotive Electrical Appliances, 2025, (12): 66-68.DOI: 10.13273/j.cnki.qcdq.2025.12.015.
[7] Shen Gaowei. Research on Industrial Robot Target Recognition and Grasping Technology Based on Deep Learning [J]. Automation Application, 2025, 66(22): 45-47+53.DOI: 10.19769/j.zdhy.2025.22.012.
[8] Li Wenjie. Research on Target Recognition and Following System for Mobile Robots [D]. Qilu University of Technology, 2025.DOI: 10.27278/d.cnki.gsdqc.2025.000238.
[9] Chen Z , Zhang Z , Su Q , et al. Object detection for autonomous vehicles under adverse weather conditions [J]. Expert Systems With Applications, 2026, 296 (PB): 128994-128994. DOI: 10.1016/J.ESWA.2025.128994.
[10] Wu Wenling. Real-time Image Processing and Target Detection Technology in Automotive Driving Environments [J]. Automotive Electrical Engineering, 2025, (07): 117-119. DOI: 10.13273/j.cnki.qcdq.2025.07.018.
[11] Liang Huiqing. Research on Object Detection Algorithm for Autonomous Vehicles Based on Deep Learning [J]. Auto World, 2025, (12): 22-24.