Overview of Object Detection and Recognition in Vehicle Autonomous Driving

Qing Shi

Chang'an Dublin International College of Transportation, Chang'an University, Xi'an, 710018, China

2023902839@chd.edu.cn

Abstract. In recent years, autonomous vehicles have garnered significant attention. As a core technology, object detection and recognition enable vehicles to perceive surrounding obstacles and traffic information, facilitating safer and more convenient operations. This paper offers a comprehensive overview of object detection and recognition methods in autonomous driving, encompassing key sensors (cameras, LiDAR, radar), traditional techniques, fusion approaches, and uncertainty estimation. It analyzes the strengths and limitations of methods like VoxelNet and CenterNet, and discusses future challenges from complex, diverse environments. Finally, it explores current deficiencies, difficulties, and potential research directions. This review aims to inform researchers and practitioners in autonomous vehicle perception systems.

Keywords: 3D object detection, sensors, KITTI dataset, point cloud-based methods for object detection.

1. Background

In recent years, in the field of intelligent transportation, autonomous driving technology has been evolving at an unprecedented rate. As a core technology of autonomous driving, object detection and recognition technology enable vehicles to identify obstacles or traffic signals, thereby better processing complex traffic information. This technology has long transcended single application scenarios, from unmanned delivery vehicles shuttling through urban streets to inspection robots operating accurately in closed parks. With the significant advantages of reducing human error rates and greatly improving road and operational safety, it has been deeply integrated into diverse scenarios such as logistics transportation and industrial operation and maintenance [1]. Currently, this technology has shifted from initial single-sensor, 2D detection to multimodal, 3D detection [2]. However, there are still some challenges in this field, such as the impact of harsh environments [2], complex traffic information [3], and the inherent uncertainty of detection [4] on results. Therefore, given the rapid iteration of this technology and the existing deficiencies in its future development, it is of great significance to sort out its achievements.

2. Brief Description of 3D Detection Technology

Object detection technology initially developed as 2D object detection. Although this method has well-established datasets and detection frameworks, and facilitates positioning, it ignores 3D positional information such as object depth or pose information [5]. In contrast, 3D object detection can better present detailed information about the size and position of objects in the surrounding environment, and essentially involves identifying 3D bounding boxes of objects [5]. For example, image-based 3D detection can derive an object's x, y, z (positional information, i.e., position coordinates in a specific coordinate system), h, w, l (dimensional information, including height, width, and length), θ (orientation information), and class (object category) from RGB images [6]. Specifically, it can be expressed as the following formula: $B = [x_c, y_c, z_c, h, w, l, \theta, \text{class}]$ [5].

3. Sensors and Datasets

3.1 Sensors

In the autonomous driving technology system, sensors are core components for perceiving the external environment, playing a significant role in ensuring driving safety and enabling accurate decision-making. Currently, there is a wide variety of sensor types, which can be classified into active (e.g., cameras) and passive (e.g., radars) based on information sources [5]. Different sensors have distinct advantages in different scenarios; thus, a multimodal approach using multiple sensors simultaneously is sometimes adopted. Even when only one type of sensor is used, additional sensors are often deployed as backups to prevent failures of the primary one [1]. The specific characteristics of different sensor types are shown in Table 1 below:

Table 1. Features of Sensors

Types	Advantages	Disadvantages
Monocular Camera	Low cost, no fixed standards used, and multiple digital object detection methods	High detection uncertainty and lack of depth information in special weather conditions
Stereo Camera	Low cost, with in-depth information	Easy to be affected by weather conditions
LiDAR	Provide three-dimensional spatial data; Less affected by external light	High cost; Its output is a 3D point called a point cloud (PCL), which cannot recognize semantic objects and texture information
4D Radar	Enduring the impact of adverse weather conditions	The existing dataset limits development and has not been further developed

Among them, cameras, due to their optical principles, are highly susceptible to light changes caused by weather, which not only reduces the reliability of detection systems but also shrinks the detection range [3]. Radar, on the other hand, has multiple types based on differences in wavelength, such as short-wave radar and long-wave radar. LiDAR is a more common type among them, and their principle is to obtain 3D information of external objects on external objects by recording the time between transmitted and detected pulses [1].

3.2 Datasets

Current datasets are diverse, mainly divided into those for traditional 2D detection (e.g., BDD100K, MOT17) and 3D detection [2]. Take the common KITTI dataset: it applies not only to 3D object detection but also to tasks like 3D tracking [5]. It includes point cloud data and RGB images covering eight categories, each classified into "easy", "moderate" and "hard" based on object size, occlusion and other features [2]. KITTI collects data in scenarios such as rural roads, urban roads and highways via devices like laser scanners and high-resolution grayscale cameras [3]. However, KITTI has a drawback: all its data were collected using the same sensor set under a single weather condition (sunny) [3]. The table below summarizes currently widely used datasets.

Table 2. The Simple Conclusion of Datasets

Datasets	Year	Size (GB)	Scenes	Cities	360 Cameras	LiDAR Scans	Night	Location
KITTIT [3][6]	2012	43.67	50	5011	NOT	YES	NOT	Germany
NuScence [3][6]	2019	75.75	1000	225	YES	YES	YES	China
Waymo Open [3] [6]	2020	336.62	1000	31	YES	YES	YES	USA
ApolloScape [3] [6]	2019	206.18	103	78	-	-	-	-
H3D [3] [6]	2019	86.02	160	31	NOT	-	YES	USA

Argoverse [3] [6]	2019	97.81	113	88	YES	YES	YES	USA

4. Object detection method

4.1 Camera-based Methods

Camera-based object detection is primarily divided into monocular detection methods (similar to 2D detection, still lacking depth information) and stereo vision methods [3], with monocular methods featuring low cost [2]. To address their insufficient depth perception, algorithmic remedies are employed, mainly categorized into template matching-based methods and geometric properties-based methods. For instance, the Deep MANTA method obtains 3D object information by comparing 2D detected bounding boxes with existing 3D object libraries [3], but its detection is limited to vehicles, lacking other categories [4]. M3D-RPN draws on the anchor box mechanism of 2D detection, predefines 3D anchors on monocular images, and directly regresses 3D bounding box parameters [3]. While improving inference efficiency, dense anchors cause a surge in computational load, and depth estimation errors degrade localization accuracy for distant targets (e.g., on the KITTI dataset, vehicle detection accuracy drops by 27% for targets >50m away) [1]. Deep3Dbox uses convolutional networks to quantize orientation into categories and regress fine-grained orientation values, enhancing estimation accuracy [3]. CenterNet adopts an anchor-free design, predicts target centers and depth offsets, and derives 3D bounding box coordinates [3]; it simplifies the model but relies on accurate center prediction (miss rate of centers increases by 15% in occlusion scenarios) [5]. Among the two categories, template matching methods are limited by the scene adaptability of predefined templates; Geometric properties methods must overcome depth estimation errors and generalization challenges of prior assumptions (e.g., shape regularity). These jointly restrict accuracy breakthroughs in monocular 3D detection.

4.2 Point Cloud-based Methods

Point cloud data, with its rich geometric information, occupies a core position in 3D object detection for autonomous driving. Compared to unimodal data, it offers superior detection stability in complex scenarios. However, point cloud processing still faces multiple technical challenges: first, high computational costs in data processing demand high hardware computing power; second, inherent data sparsity easily causes target miss-detection, affecting accuracy; third, such data are highly sensitive to weather, with feature distortion likely in rain, snow, fog, etc. [2].

These point cloud-based methods typically start with progressive sampling of raw point clouds (e.g., selecting points via "farthest point sampling"), then learn features through point cloud operators, and finally predict 3D bounding boxes using sampled points and features [1]. While inheriting deep learning's strengths in point cloud processing, their feature learning is constrained by "number of contextual points" and "ball query radius"—too few points miss details, while an overly large range loses fine 3D information, requiring careful balancing to achieve both performance and efficiency [1].

Common point cloud-based methods are diverse, with several outlined below.

VoxelNet converts point clouds into 3D voxels, uses a Vector Feature Extractor to extract feature vectors, employs convolutional layers to explore contextual shapes of voxel features, and finally predicts targets via a Region Proposal Network [3].

Projection methods project 3D point clouds into 2D images (e.g., planar, cylindrical, or spherical projections) and reuse mature 2D detection models for indirect 3D detection. For example, Complex-YOLO, a single-stage detector extended from YOLO, predicts additional dimensions and yaw angles; though slightly inferior in performance (achieving 50 frames/second), it is more efficient than previous methods [5].

Pseudo-LiDAR generates dense depth maps from images via depth or disparity estimation, back-projects pixels into 3D coordinates using camera parameters to form pseudo-LiDAR point clouds

(composed of pixel-corresponding 3D coordinates), and inputs these into LiDAR-based models. This method breaks the barrier between image-based and LiDAR-based approaches, enabling direct reuse of mature LiDAR detection technologies [6].

Anchor-based methods rely on predefined anchors; dense anchors require Non-Maximum Suppression (NMS) to process numerous potential targets [2].

Anchor-free methods avoid complex anchor design and flexibly adapt to multi-views (e.g., BEV, point view). For instance, PointRCNN extends Faster R-CNN to 3D, uses PointNet++ for foreground-background segmentation to generate 3D candidates bottom-up, and refines local features by converting candidate points to canonical coordinates. However, its two-stage process and some Anchor-free methods' reliance on NMS limit inference efficiency [3].

In summary, point clouds play a pivotal role in autonomous driving 3D detection, with various methods each having unique traits and limitations despite existing technical challenges.

4.3 Fusion Methods

Safety-critical systems such as in-vehicle object detection require accurate and robust perception, thus integrating cameras, millimeter-wave radars, and LiDAR are often integrated. Fusion reduces information redundancy and enhances perception capabilities, with multimodal methods categorized into "multimodal fusion" (direct fusion between modalities) and "cross-modal interaction" (linked learning between modal features) [2].

Fusion approaches mainly fall into three types: early fusion, late fusion, and deep fusion [5]. MV3D adopts a deep fusion strategy, combining LiDAR bird's-eye view, front-view projections, and camera RGB channels to optimize 3D object detection through hierarchical interaction of multi-view features [5]. AVOD pioneered early fusion, merging LiDAR bird's-eye view and camera RGB features for region proposal, and uses feature pyramid upsampling to compensate for lost details of small targets, adapting to harsh environments like rain and snow [5]. Frustum PointNet constructs a frustum proposal, 3D instance segmentation, and modality-free 3D box estimation network: It first generates 3D search space via 2D bounding box regression to align point clouds, then completes instance segmentation and 3D box regression based on frustum-internal point clouds, achieving efficient and accurate detection on COCO and KITTI datasets while better preserving 3D geometric properties [3]. The Pseudo-LiDAR method encodes depth maps from stereo or monocular images into pseudo-LiDAR point clouds, adapting to existing LiDAR target detection technologies, significantly improving detection accuracy on KITTI, and reducing reliance on expensive LiDAR systems [3].

In summary, existing multimodal fusion methods explore the trade-off between efficiency and accuracy, demonstrating the potential of cross-modal collaboration while exposing performance shortcomings due to strategy differences, driving continuous technical iteration.

4.4 Detection Uncertainty

To date, most methods fail to generate calibrated confidence accuracy, leading to hazardous accidents in practical scenarios [5]. This issue can be addressed via probabilistic object detection, which corelies in uncertainty estimation based on traditional target detection to quantify the reliability of prediction results [4]. Sources of uncertainty typically fall into two categories: epistemic uncertainty, related to model parameters, reflecting the model's ability to fit training data; and aleatoric uncertainty (also called data uncertainty), reflecting inherent noise in sensor observations [4].

Existing estimation methods roughly include four types: Monte Carlo Dropout approximates prediction distributions via multiple dropout-augmented inferences, suitable for large-scale data but time-consuming; Deep Ensembles fuses outputs from multiple independently trained networks, with high accuracy but computational costs increasing with network count; Direct Modeling directly predicts probability distribution parameters (e.g., mean and variance of Gaussian distribution) via the

network output layer, efficient but may produce poorly calibrated probabilities; Error Propagation approximates and propagates layer-wise variances, adapting to real-time scenarios [4].

Multimodal fusion breaks through single-sensor bottlenecks, building accurate and robust perception capabilities for in-vehicle safety systems from adapting to harsh environments to preserving geometric properties.

5. Future Research Directions

Despite significant progress in 3D object detection in terms of algorithm accuracy, modal fusion, and scenario adaptability [6], challenges remain for complex real-world scenarios (e.g., extreme weather, dynamic multi-target interactions) and industrial deployment needs, including insufficient algorithm generalization, low efficiency of multimodal collaboration, and high annotation costs. In terms of datasets, existing mainstream datasets (e.g., KITTI) are mostly collected in ideal scenarios, lacking the diversity of real complex environments. Future needs include large-scale, multi-sensor synchronously annotated datasets for complex scenarios [7]. Tasks such as pedestrian detection, lane detection, roadside camera detection, and detection in special weather scenarios are emerging as promising research directions in 3D object detection, due to the scarcity of related datasets, annotation information, and experimental support [2]. Additionally, generalization is a common issue in existing models. In complex scenarios, models rely on domain adaptation and self-supervised learning to enhance robustness, but self-supervised methods remain immature. Further exploration of unsupervised domain adaptation and pseudo-label optimization is needed to reduce reliance on manual annotation [7]. The performance of image-based 3D object detection highly depends on accurate depth estimation. Though depth estimation and 3D detection have long developed independently, explorations of Pseudo-LiDAR and multi-task networks have verified the potential of their joint optimization, making this a valuable future research direction [6].

6. Conclusion

This paper focuses on 3D object detection and recognition technologies in autonomous driving, discussing the main implementation paths of current 3D detection technologies, key features of sensors and datasets, and challenges in practical applications. Methodologically, it analyzes the technical evolution of 3D object detection, sensor applications, dataset characteristics, and detection methods by reviewing renowned academic literature. In summary, 3D object detection for autonomous driving has formed a relatively complete technical system: sensors evolve from single modality to multimodal fusion, and detection methods develop from single-data-based to cross-modal interaction, with significant progress in accuracy and efficiency. The technology still faces challenges such as insufficient adaptability to complex real-world scenarios, the dilemma of balancing real-time performance with accuracy, and high costs in some aspects. Future efforts should focus on constructing large-scale complex scenario datasets, developing generalization technologies such as unsupervised domain adaptation, and optimizing multi-sensor fusion strategies to advance 3D object detection toward greater reliability and efficiency.

References

[1] Mao, J., Shi, S., Wang, X., & Li, H. (2023). 3D object detection for autonomous driving: A comprehensive survey. International Journal of Computer Vision, 131, 1909–1963.

[2] Zhang, P., Li, X., Lin, X., & He, L. (2025). A new literature review of 3D object detection on autonomous driving. Journal of Artificial Intelligence Research, 82.

[3] Ghasemieh, A., & Kashef, R. (2022). 3D object detection for autonomous driving: Methods, models, sensors, data, and challenges. Transportation Engineering, 8, 100115.

[4] Feng, D., Harakeh, A., Waslander, S. L., & Dietmayer, K. (2022). A review and comparative study on probabilistic object detection in autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 23, 9961 - 9980.

[5] Arnold, E., Al - Jarrah, O. Y., Dianati, M., Fallah, S., Oxtoby, D., & Mouzakitis, A. (2019). A survey on 3D object detection methods for autonomous driving applications. IEEE Transactions on Intelligent Transportation Systems, 20, 3782 - 3795.

[6] Ma, X., Ouyang, W., Simonelli, A., & Ricci, E. (2024). 3D object detection from images for autonomous driving: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(5), 3537–3556.

[7] Wang, K., Zhou, T., Li, X., & Ren, F. (2023). Performance and challenges of 3D object detection methods in complex scenes for autonomous driving. IEEE Transactions on Intelligent Vehicles, 8(2), 1699–1716