All You Need for Object Detection: From Pixels, Points, and Prompts to Next-Gen Fusion and Multimodal LLMs/VLMs in Autonomous Vehicles [1]

Sayed Pedram Haeri Boroujeni[a] (shaerib@g.clemson.edu)
Niloufar Mehrabi[a] (nmehrab@g.clemson.edu)
Hazim Alzorgan[a](halzorg@g.clemson.edu)
Mahlagha Fazeli[a] (mfazeli@clemson.edu)
Abolfazl Razi[a*] (arazi@clemson.edu)

[a]School of Computing, Clemson University, Clemson, SC 29632, USA

Corresponding Author:
Abolfazl Razi
School of Computing, Clemson University, Clemson, SC 29632, USA
Email[1]: arazi@clemson.edu
Email[2]: shaerib@g.clemson.edu

[1]This material is based upon the work supported by the National Science Foundation (NSF) under Grant Numbers 2008784 and 2204721.

All You Need for Object Detection: From Pixels, Points, and Prompts to Next-Gen Fusion and Multimodal LLMs/VLMs in Autonomous Vehicles

Sayed Pedram Haeri Boroujeni[a], Niloufar Mehrabi[a], Hazim Alzorgan[a], Mahlagha Fazeli[a], Abolfazl Razi[a*]

[a]*School of Computing, Clemson University, Clemson, 29632, SC, USA*

Abstract

Autonomous Vehicles (AVs) are transforming the future of transportation through advances in intelligent perception, decision-making, and control systems. However, their success is tied to one core capability, reliable object detection in complex and multimodal environments. While recent breakthroughs in Computer Vision (CV) and Artificial Intelligence (AI) have driven remarkable progress, the field still faces a critical challenge as knowledge remains fragmented across multimodal perception, contextual reasoning, and cooperative intelligence. This survey bridges that gap by delivering a forward-looking analysis of object detection in AVs, emphasizing emerging paradigms such as Vision-Language Models (VLMs), Large Language Models (LLMs), and Generative AI rather than re-examining outdated techniques. We begin by systematically reviewing the fundamental spectrum of AV sensors (camera, ultrasonic, LiDAR, and Radar) and their fusion strategies, highlighting not only their capabilities and limitations in dynamic driving environments but also their potential to integrate with recent advances in LLM/VLM-driven perception frameworks. Next, we introduce a structured categorization of AV datasets that moves beyond simple collections, positioning ego-vehicle, infrastructure-based, and cooperative datasets (e.g., V2V, V2I, V2X, I2I), followed by a cross-analysis of data structures and characteristics. Ultimately, we analyze cutting-edge detection methodologies, ranging from 2D and 3D pipelines to hybrid sensor fusion, with particular attention to emerging transformer-driven approaches powered by Vision Transformers (ViTs), Large and Small Language Models (SLMs), and VLMs. By synthesizing these perspectives, our survey delivers a clear roadmap of current capabilities, open challenges, and future opportunities, highlighting underexplored avenues such as multimodal reasoning, cooperative perception, and foundation-model integration. We aim to establish this work as a definitive reference for researchers, practitioners, and developers, fostering accelerated innovation toward safer and more intelligent autonomous driving systems.

Keywords: Autonomous Vehicle (AV), Computer Vision (CV), Vision Transformers (ViTs), Large Language Models (LLMs), Vision Language Models (VLMs), Sensor Fusion.

1. Introduction

Object detection is a cornerstone task in autonomous driving systems and stands as a critical component for environmental understanding and safe navigation. It enables Autonomous Vehicles (AVs) to identify surrounding vehicles, pedestrians, cyclists, and obstacles using data captured from various sensors such as cameras, LiDAR, and Radar. Beyond traditional sensing methods, emerging multimodal models, including LLMs and VLMs are being integrated into AV perception systems to enhance scene interpretation and contextual reasoning. Figure 1 illustrates examples of object detection across these sensor modalities, highlighting the complementary roles of multimodal AI in advancing AVs' perception.

AVs have surfaced as a disruptive technology that can revolutionize transportation, surveillance, logistics, agriculture, environmental monitoring, and public safety. Underneath such systems lie sophisticated sensing, decision-making, and control infrastructures that enable their autonomous operation without human oversight [1]. Their optimal execution depends on the full integration of the perception, localization, planning, and control modules, whose performance is determined by innovations in sensing technologies and intelligent algorithms. The rapid advance of technology in AI, Machine Learning (ML), and sensors has greatly influenced the rapid development and fielding of these robotic systems [2]. Despite the great progress, autonomous systems still encounter many technical and practical challenges, particularly in ensuring reliable performance in complex, dynamic, and unpredictable environments. These systems must robustly handle inaccurate sensors, hardware faults, and the computational challenge of not just high-dimensional but also high-frequency real-time data. In parallel with these technical developments, AVs have also drawn great attention from industry and academia

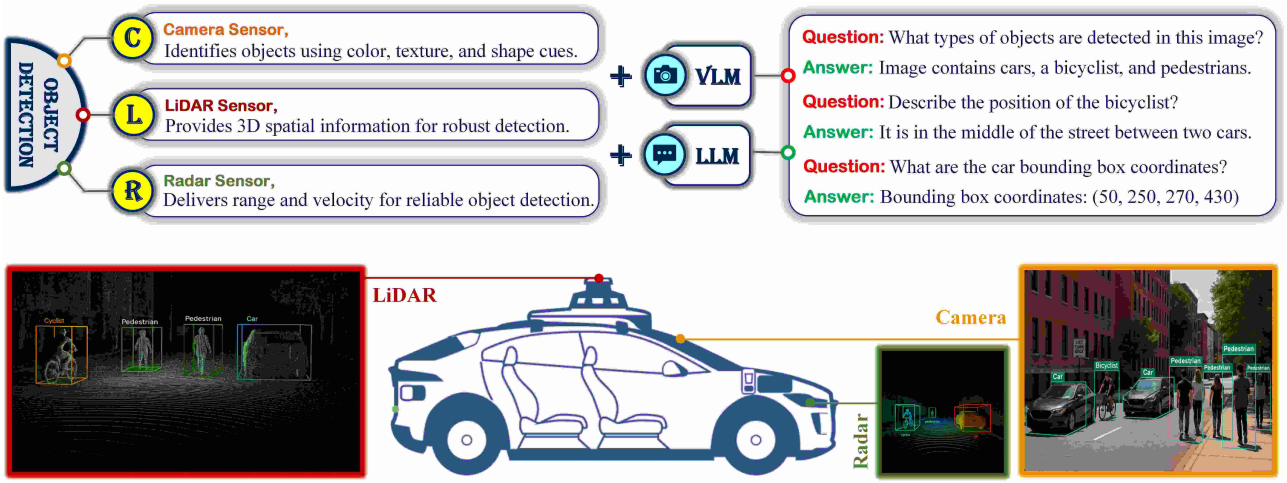

Figure 1: Visualization of object detection across multiple sensor modalities in autonomous vehicles. The RGB image demonstrates 2D detection with colored bounding boxes for cars, pedestrians, and cyclists. LiDAR and Radar point cloud showcases 3D detection through spatially aligned boxes. The integration of multimodal AI, including LLMs and VLMs, supports contextual understanding and enhances detection accuracy by fusing visual and spatial cues from diverse sensor inputs.

with regard to their promising road safety, traffic congestion mitigation, reduction in transportation cost, and social inclusion for elderly and disabled people [3]. Their deployment spans a broad range of applications, from self-driving taxis and delivery robots to autonomous agricultural machinery and emergency response systems [4]. Nevertheless, achieving fully autonomous driving in the open world is still an open problem. AVs must operate in a rapidly evolving environment that includes complex urban settings, diverse weather conditions, unpredictable human behavior, and rare edge cases for which limited or no available training data exists. In addition, the absence of standardized testing procedures, unclear regulatory guidelines, and ongoing public concerns about safety and accountability remain major barriers to widespread deployment. As AVs transition from controlled environments to real-world roads, it becomes increasingly important to resolve these technical and societal issues to ensure their safe, reliable, and broadly accepted integration into transportation systems.

Effective adoption and integration of AVs depend on the careful consideration and coordination of multiple interrelated factors. Robustness, scalability, reliability, real-time efficiency, accuracy, and safety are fundamental requirements for autonomous driving systems to operate effectively and gain acceptance in real-world applications [5]. First and foremost, safety is the most challenging issue, and AVs must function consistently in a dynamic and sometimes unpredictable scenario to protect passengers, pedestrians, and other road users. The ultimate goal is to significantly minimize, or ideally remove, accidents from human error. This requires AV systems to incorporate multiple layers of safety mechanisms, fail-safe redundancies, and rigorous testing to reduce operational risks. Reliability is closely linked to safety, as it ensures that the AV performs consistently across a broad spectrum of use cases, from typical situations to rare edge cases. Reliable behavior is essential for the public to gain confidence in AVs and support their integration into complex traffic systems. Moreover, robustness is equally important. The AV should be able to operate effectively even in the presence of noisy sensors, hardware faults, extreme weather conditions, uncertain road environments, and adversarial inputs. These disruptions can be managed by a well-designed system that not only endures but continues to function reliably and reinforce user trust. Timely responsiveness is also a critical requirement for AV systems. AVs must process data from their sensors, interpret their surroundings, and make instantaneous decisions in rapidly evolving environments. The ability to act within milliseconds is crucial for avoiding collisions and ensuring smooth operation in dynamic. Together, these elements lay the foundation for a reliable, resilient, and publicly trusted autonomous vehicle system. At the same time, perception, localization, and prediction must achieve high levels of accuracy, as AVs are required to detect objects, estimate their positions, and anticipate the behavior of surrounding agents to navigate safely and make informed decisions. Efficiency is also important, as the system has to effectively utilize computational and energy resources [6]. Finally, the scalability of AV systems to accommodate multiple types of vehicles, levels of autonomy, and operational conditions (without sacrificing performance or increasing the complexity of the system) is required.

Another critical aspect influencing the performance of autonomous vehicles is their strong dependence on data throughout both development and deployment. Developing high-performing AV systems requires access to large-scale, high-quality, diverse, and well-labeled datasets that comprehensively represent the full spectrum of real-world driving scenarios. From crowded city streets to rare situations like road construction or extreme weather, the system must be trained on and exposed to varied conditions to develop the ability to generalize

beyond controlled test environments [7]. The size of the dataset is especially important, as complex models demand extensive training data to capture the variability of real-world driving. In addition, features extracted from multi-sensor data collected by cameras, LiDAR, and Radar, as well as mechanisms that provide high-quality annotations, are necessary to handle tasks such as object detection, tracking, and behavior prediction. Annotating this multimodal data, however, remains a labor-intensive and time-consuming task that typically involves manual work or human expert supervision of the process to maintain reliable and valid information across modalities [8]. In addition, real-world datasets typically exhibit a long-tail distribution, where critical events such as pedestrian crossings in low-light conditions or near-collision incidents are rarely captured. This lack of edge cases makes it difficult to train models that are resilient to rare but high-stakes occurrences. To mitigate these problems, several techniques for synthetic data generation and data augmentation have been developed in order to augment diversity in the data and to counterbalance the class distribution. It should be noted that transferring knowledge of learned synthetic data to real-world scenarios is still a challenge, in terms of retaining fidelity, transferability, and semantic consistency. It must also be consistently sustained to maintain AV systems in tune with dynamic infrastructure, traffic laws, and driver actions, ultimately underscoring the need for scalable and adaptable data pipelines. In the absence of enough coverage or representative samples, models become biased or too specialized, which decreases their generalization capabilities in unseen conditions. Ultimately, a data-driven technique is necessary for designing intelligent AV systems that can perform reliably in real-world driving environments, not just in theoretical settings [9].

Autonomous vehicles must perceive and operate with high precision as they navigate environments enriched with diverse sensors, each designed to fulfill distinct functional roles. Cameras acquire high-resolution visual information for subsequent recognition and interpretation of the environment. LiDAR systems provide high-resolution three-dimensional depth measurements, enabling precise spatial mapping of objects and structural elements within a scene [10]. In contrast, Radar is particularly effective under conditions of severe visual degradation, such as rain, fog, or low-light environments, offering consistent and robust object detection and tracking in adverse weather and limited visibility scenarios [11]. Ultrasonic sensors assist with close-range obstacle detection, which is especially useful during parking or low-speed maneuvers. These sensors are often supplemented by high-resolution maps with precise information about road trajectories, lane boundaries (lines), and traffic signs [12]. Furthermore, vehicle-to-everything (V2X) communication increases the situational awareness of vehicles, as they can share data with surrounding infrastructure, other vehicles, and pedestrians [13]. In particular, vehicle-to-vehicle (V2V) communication allows the AVs to exchange their positions, velocities, accelerations, and intended trajectories, and to perform cooperative actions to reduce the risk of collision. Vehicle-to-infrastructure (V2I) communication, frequently enabled by roadside units (RSUs), serves as a crucial source of environmental and regulatory information, such as traffic signal timings and phases, work zones, and speed limits [14]. Moreover, Infrastructure-to-infrastructure (I2I) communication is a crucial part of this ecosystem that serves as the link between fixed infrastructure nodes, such as traffic lights, and RSUs. In addition to external sensing, AVs depend on ego-centric information: their own position, orientation, velocity, and acceleration, as retrieved from Global Positioning System (GPS) [15], Internal Measurement Units (IMUs) [16], and wheel odometry. The states of these internal variables are important for localization, path planning, and motion control of the vehicle. Novel technologies such as vehicle-to-language (V2L) are also under development, aiming to transform multimodal sensory data into natural language descriptions [17]. This enables AVs to produce interpretable explanations of their environment, internal states, and behavioral choices, thereby enhancing transparency, user trust, and system accountability.

Effective processing of rich sensorimotor data in AVs depends on sophisticated perception algorithms to retrieve, interpret, and aggregate scene descriptions. These algorithms are targeted to address key perception tasks, such as object detection, classification, and segmentation, that are necessary to develop a comprehensive understanding of the operational environment. These perception pipelines include 2D vision-based models, 3D LiDAR-driven methods, and increasingly, hybrid 2D–3D fusion strategies that combine complementary features across modalities. The latest deep learning models, such as convolutional neural networks (CNNs) and transformer-based models, can learn automatic discriminative feature representations from multimodal sensor measurements. Object detection in real time, for instance, is achieved by detection systems such as Faster R-CNN [18], and Single Shot MultiBox Detector (SSD) [19], while pixel-level scene understanding is obtained through segmentation models such as DeepLab [20] and Mask R-CNN [21]. In addition to detecting objects in a single frame, such models have now been adapted to understand scene dynamics over time. This allows the models to follow object motion and anticipate their near-future positions. They have also begun to integrate multiple tasks (detection and segmentation) into a single model, which improves processing efficiency and contributes to richer and more semantically meaningful representations of the surrounding environment. To further improve the performance and robustness of such systems, sensor fusion methods are used systematically to integrate heterogeneous data from diverse sensors. These fusion techniques, which can operate at raw data, feature, or decision stages, frequently utilize probabilistic filters, attention models, or deep nets to map

and reconcile heterogeneous inputs. This sophisticated algorithmic pipeline is tailored to ensure real-time responsiveness, precision, and resilience, thereby enabling AVs to operate reliably in unpredictable conditions.

Large Language Models (LLMs) [22] have received widespread interest over the past few years as they can process and generate human-like language within extensive contextual settings. Their incorporation in autonomous and assistive systems changes the way machines perceive, understand, and interact in the environment. In AVs, LLMs are investigated for improving perception and reasoning through natural language interpretation of sensor data, scene descriptions as well as high-level decision-making [23]. This results in better semantic comprehension and an increased ability to deal with complex or ambiguous situations, which traditional models can misinterpret. Additionally, LLMs support natural language querying of sensory inputs and map-based information, enabling intuitive human-vehicle interaction and enhancing explainability in decision pipelines. Concurrently, LLMs are also being adopted in assistive technologies to help people with disabilities, where they can help implement functionalities such as voice-controlled navigation, real-time description of the environment, and personalized interaction. Building on LLMs, Vision-Language Models (VLMs) [22] integrate visual and textual context for a better understanding of an environment [24]. In AVs, this combination enables self-driving cars to describe traffic scenes in human-like language, better understand road conditions, and explain why they perceive and react as they do. Connecting visual input with semantic understanding allows for more efficient and flexible operations of AVs in complex and dynamic environments, making interactions with human users clearer and more intuitive. The multimodal nature of this multi-source integration promotes informed and context-aware decision-making, which is crucial for reliably and safely dealing with real-world traffic scenes. For automated aids, they increase accessibility by allowing machines to verbally describe visual content during use, providing support for low-vision and cognitively impaired users. VLMs can also help close the gap between vision and language, making intelligent systems more interpretable, interactive, and user-oriented.

1.1. Contributions of This Survey

This survey provides a comprehensive and structured review of object detection in autonomous vehicles, emphasizing the integration of diverse sensor technologies and the evolution of detection methods. After examining over seven hundred research and survey articles, the study identifies key trends, innovations, and research gaps across multiple facets of AV perception systems. The survey is uniquely organized into three core components: AV sensor technologies, AV datasets, and object detection methods. In the first part, we analyze different sensor types, including camera, ultrasonic, LiDAR, and Radar sensors, as well as their fusion strategies. The second part introduces a novel taxonomy of AV datasets categorized by ego-vehicle, infrastructure, and communication paradigms (e.g., V2L, V2V, V2I, V2X, I2I). Finally, the third part presents a detailed review of object detection techniques, including 2D camera-based, 3D LiDAR-based, 2D–3D fusion, and LLM/VLM-based approaches. By synthesizing these diverse elements, our survey not only highlights the current state of object detection in AVs but also offers insights into underexplored areas and emerging opportunities, setting a foundation for future research directions in this rapidly advancing field.

The key contributions of this review paper can be summarized as follows:

- Reviews significant and recently published survey papers in the context of object detection in autonomous vehicles, summarizing their contents, strengths, and limitations. We highlight the key topics covered in each study, as well as the missing or underexplored areas that remain unaddressed
- Conducts an extensive review of state-of-the-art detection methods in AVs, including 2D camera-based approaches, 3D LiDAR-based techniques, and hybrid 2D–3D sensor fusion strategies. The review covers core methodologies, network architectures, performance benchmarks, and application contexts.
- Explores and evaluates emerging AI techniques, particularly Vision Transformers (ViTs), Large Language Models (LLMs), and Vision-Language Models (VLMs), as well as their integration into AV perception systems for enhanced object detection capabilities. We analyze how these models enable multimodal reasoning, spatial understanding, and instruction-following in autonomous driving tasks.
- Conducts a comparative evaluation of object detection algorithms across all four categories, including 2D (camera-based), 3D (LiDAR-based), 2D–3D fusion, and LLM/VLM-based methods, by analyzing their performance on widely used benchmark datasets. We also identify and discuss the top 10 representative algorithms in each category, examining their key innovations, architectures, strengths, and limitations.
- Examines various types of sensor technologies used in AVs and their applications in object detection. We outline the strengths and weaknesses of each sensor modality, including camera, ultrasonic, LiDAR, and Radar sensors, along with a discussion of their integration and utilization within AV perception systems.

- Proposes a unique and structured categorization of AV datasets developed in recent years, followed by a systematic analysis of each dataset. This includes ego-vehicle, roadside, and cooperative perception datasets (V2V, V2I, V2X, I2I). We highlight key characteristics such as sensor modalities, data volume, annotation types, collection environments, and application focus.
- Discuss the simulation platforms commonly used for synthetic AV dataset generation and testing. We compare these simulators based on key features such as supported sensor modalities (e.g., camera, LiDAR, Radar), environmental realism, weather and traffic control, annotation capabilities, computational requirements, and compatibility with AV development frameworks.
- We highlight open problems and future directions at the end of each major section of our survey paper to guide future advancements in AV perception and object detection. This survey paper aims to to serve as a comprehensive and practical reference for researchers, practitioners, and developers in the field of autonomous driving, enabling better understanding, benchmarking, and design of AV detection systems.

1.2. Organization of This Paper

The general structure of the paper is illustrated in Figure 2. Section 2 reviews existing survey literature on object detection in autonomous driving systems, highlighting their covered topics and identifying potential gaps. Section 3 describes the details of AV technologies and sensor specifications used in autonomous driving applications. The existing AV datasets are provided in section 4, with a focus on their sensing modalities, annotation types, coverage diversity, and applicability to various perception tasks. Section 5 presents a comprehensive review of object detection methods in autonomous vehicles, covering 2D camera-based approaches, 3D LiDAR-based techniques, 2D–3D fusion strategies, and emerging methods based on LLMs and VLMs. Lastly, Section 6 contains conclusions and future directions.

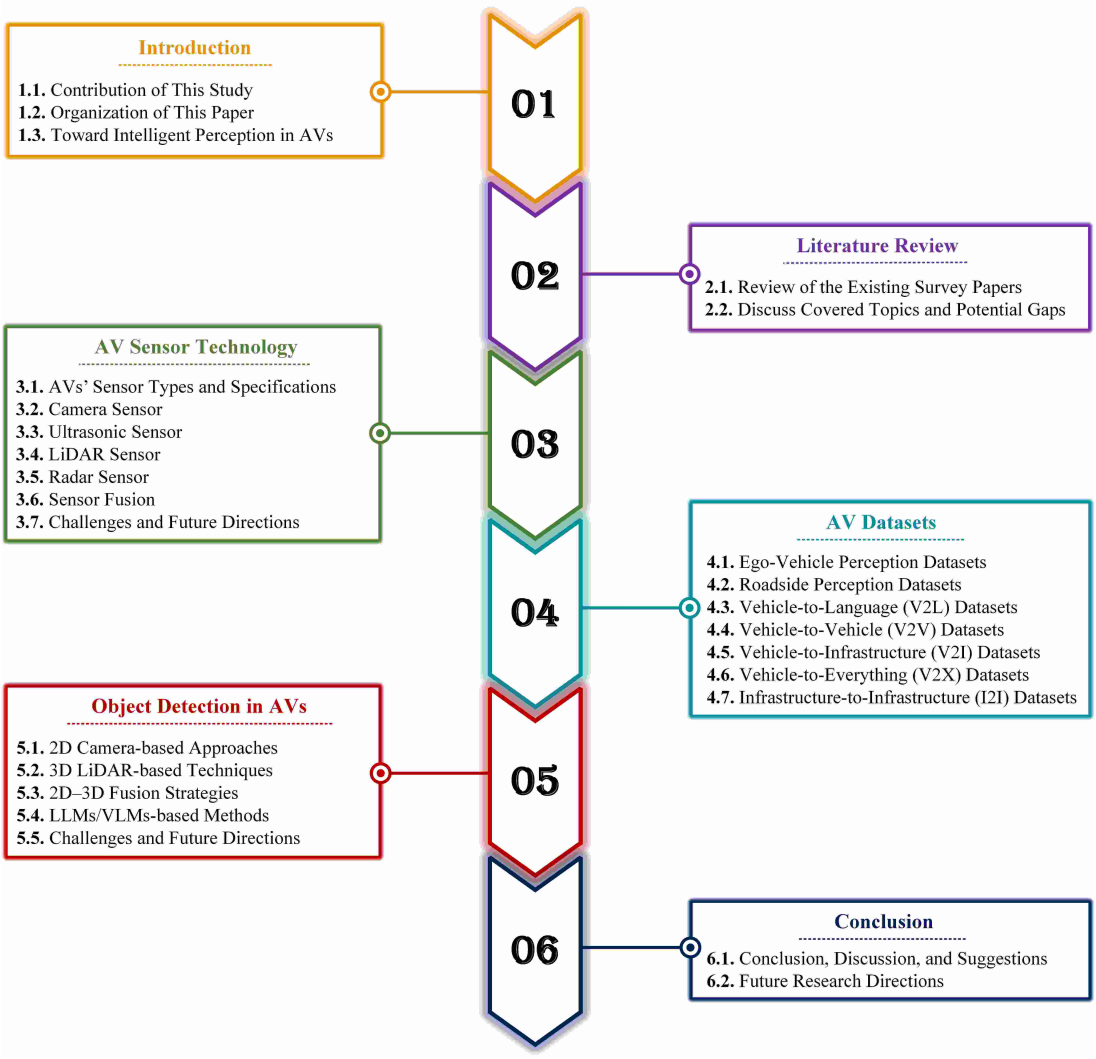

Figure 2: The organization of this survey paper.

2. Background and Related Literature

In the related literature section, we conducted a systematic search across major academic databases to collect the most influential and relevant survey papers on Autonomous Vehicles (AVs) published between 2018 and 2025. Our search included keywords such as "autonomous vehicle," "AV sensors," "object detection," "computer vision," "Vision Language Models (VLMs)," "Large Language Models (LLMs)," "Cooperative Perception (CP)," "communication," "multimodal fusion," "deep learning," "3D object detection," "LiDAR," "point cloud," "camera," "Radar," "autonomous driving," and "sensor fusion." We then critically analyzed the identified papers based on their titles, abstracts, methodologies, findings, and citation counts. This approach allowed us to select survey papers that align closely with our focus on comprehensive AV research, including perception, decision-making, and multimodal sensor fusion. Table 1 summarizes the strengths and weaknesses of each selected survey, highlighting their limitations and open research challenges in the field.

In 2025, three recent survey papers [25, 26, 27] provide a comprehensive review of advancements in 3D object detection and sensor fusion technologies for AVs. The first paper, titled "Developments in 3D Object Detection for Autonomous Driving: A Review", presents an overview of 3D object detection methods using different sensor modalities such as camera, LiDAR, and multi-sensor fusion approaches. It categorizes detection architectures and discusses transformer-based models and real-time deployment challenges. However, it lacks a detailed review of AV datasets, sensor configurations, and hardware platforms, as well as the integration of foundation models, VLMs, or LLMs for advanced multimodal perception in AVs. The second paper, titled "A Survey of the Multi-Sensor Fusion Object Detection Task in Autonomous Driving", focuses on fusion-based object detection architectures and compares several fusion strategies such as early, middle, and late fusion. It also discusses Transformer-based fusion techniques and popular datasets like KITTI and nuScenes. Nonetheless, the study fails to address recent advancements in 2D and 3D detection methods, Multimodal models (VLMs and LLMs), and a comprehensive review of AV datasets and sensor technologies. The third paper, titled "A Review of 3D Object Detection Based on Autonomous Driving", investigates point-based, voxel-based, and projection-based 3D detection methods with their benchmarking. While it provides an architectural taxonomy, it does not review CP frameworks, omits deep analyses of sensors and datasets, and lacks exploration of large-scale language or vision-language models for AV perception.

In 2024, two review papers [28, 29] focus on breakthroughs in 3D object detection and the latest trends in multimodal fusion for autonomous driving perception tasks. The first paper, titled "Robustness-Aware 3D Object Detection in Autonomous Driving: A Review and Outlook", provides an in-depth analysis of robustness-aware techniques across camera-only, LiDAR-only, and multi-sensor detection methods. It discusses adversarial vulnerabilities and surveys benchmark datasets such as KITTI and nuScenes. However, the paper does not explore all the existing datasets, CP frameworks, AV sensors, or VLMs and VLMs that are becoming increasingly relevant in AV systems. The second paper, titled "Emerging Trends in Autonomous Vehicle Perception: Multimodal Fusion for 3D Object Detection", reviews fusion techniques across early, intermediate, and late fusion stages and highlights various sensor modalities, including LiDAR, cameras, Radar, and thermal imaging. However, not only are many existing AV 2D detection methods and significant AV datasets missing, but this paper also does not include any CP frameworks, VLM methods, or segmentation approaches.

In 2023, paper [30] presents a broad historical survey of object detection methods over the past two decades, spanning traditional techniques like HOG and DPM to modern deep learning models such as YOLO, SSD, and Transformer-based detectors. However, this survey focuses on general-purpose object detection and lacks attention to autonomous vehicle-specific challenges such as sensor integration, cooperative perception, or communication-aware detection tasks. Similarly, the paper [31] provides the first dedicated review of image-based 3D object detection for autonomous driving. Although this paper covers existing AV detection methods and some popular datasets, a few key techniques (multi-modal fusion) in AV technologies as well as sensor modalities such as LiDAR and Radar are missing. Additionally, it does not explore all the potential methods including cooperative and language-driven frameworks. A comprehensive review of 3D object detection methods for autonomous driving is provided in [32]. This paper emphasizes various detection architectures categorized by input modalities, such as LiDAR-based, image-based, and multi-modal approaches, and includes detailed comparisons across major datasets like KITTI, nuScenes, and Waymo. It also highlights sensor characteristics, fusion strategies, and detection challenges under real-world driving conditions. Although this paper covers a wide range of detection algorithms and sensor fusion techniques, it lacks discussion on CP frameworks, a structured dataset taxonomy based on their types, and leaves out emerging topics such as VLMs, foundation models, and LLM-based perception frameworks in AV systems.

Paper [33] presents a exhaustive review of deep learning-based fusion techniques for combining image and point cloud data in autonomous driving. It categorizes fusion methods based on various stages (early, middle, and late), highlights sensor configurations, and the various possible fusion architectures in object detection

Table 1: Summary of existing survey papers on different aspects of autonomous vehicles.

Year	Survey	Content Included	Potential Gaps	Application Domain
2025	[25]	• 3D object detection methods that are based on LiDAR or fusion approaches. • Discusses practical deployment challenges in real-world AV driving applications. • Provides research directions for temporal, predictive, and cooperative perceptions.	• Only considers 3D detection methods, missing 2D-based detection approaches. • Lack of detailed review of AV sensor hardware and configurations in AVs. • Does not explore LLMs, VLMs, and foundation models in autonomous driving.	☑ 3D Object Detection
2025	[26]	• Explores camera–LiDAR fusion algorithms for object detection tasks in AVs. • Reviews sensor characteristics (camera, LiDAR, Radar) used in sensor fusion. • Compares 11 benchmark datasets (KITTI, nuScenes, Waymo, etc.) in AV driving.	• Most of the recent 2D and 3D object detection approaches in AVs are missing. • Lack of systematic categorization of AV datasets into meaningful taxonomies. • Does not review and compare LLMs, VLMs, and foundation models in AVs.	☑ Sensor Fusion Detection
2025	[27]	• Compares state-of-the-art point-based, voxel-based, and point-voxel methods. • Reviews 7 popular datasets (e.g., KITTI, nuScenes, Waymo, ApolloScape). • Highlights multi-modal fusion architectures (early, deep, and late fusion) in AVs.	• Does not include sensor-level analysis and reviews of AV sensing hardware. • There is no discussion of LLMs, VLMs, and foundation models used in AVs. • Only mentions a limited number of AV datasets, most of them not considered.	☑ 3D Object Detection
2024	[28]	• Summarizes 3D object detection methods in AVs categorized by sensor modality. • Reviews robustness-aware methods in camera, LiDAR, and multi-modal detection. • Discusses transformer-based, graph-based, and point-voxel models in 3D detection.	• LLM-based, VLM-based, and foundation models are not addressed at all. • Focuses solely on 3D object detection, leaving out 2D detection techniques. • Does not consider cooperative perception frameworks and datasets in AVs.	☑ 3D Object Detection
2024	[29]	• Provides a review of multimodal fusion techniques for 3D object detection in AVs. • Discusses sensor modalities (camera, LiDAR, Radar, thermal, ultrasonic, RGB-D). • Includes a comparative analysis across 15 widely used autonomous vehicle datasets.	• Only considers a few number of AV datasets, CP and 2D datasets are missing. • does not address LLM-based or VLM-based models for autonomous vehicles. • Lack of 2D detection information, while focuses exclusively on 3D detection.	☑ Multimodal 3D Detection
2023	[30]	• Provides a thorough historical and technical evolution of object detection methods. • Analyzes detection challenges such as multiscale variation and object rotation. • Covers 8 object detection datasets and metrics (PASCAL VOC, MS-COCO, etc).	• Focuses on general-purpose object detection, not specialize in AV scenarios. • Lacks exploration of Vision–Language Models or LLM-based perception tasks. • Does not include sensor-specific detection methods and hardware details.	☑ AV Object Detection
2023	[31]	• Reviews image-based 3D object detection methods applied within AV scenarios. • Introduces two taxonomies of the state-of-the-art for organizing existing methods. • Reviews 11 AV datasets (KITTI, nuScenes, Waymo) and discusses their details.	• There is no discussion of CP (V2V, V2X) frameworks, protocols, or datasets. • Does not touch on LLMs, VLMs, or foundation models applied in AV driving. • Does not review LiDAR-only, Radar-only, or multi-modal fusion methods.	☑ 3D Object Detection
2022	[32]	• Offers a taxonomy of 3D object detection methods, categorized by input modality. • Presents a discussion on sensors (passive vs. active) in AV object detection. • Discusses diversity, annotation quality, and benchmarking details of 7 AV datasets.	• Does not consider VLM, LLM, CP, and foundation models in object detection. • Although a couple of datasets are reviewed, many AV datasets are missing. • Focuses mostly on 3D detection, with little to no coverage of 2D detection.	☑ 3D Object Detection
2021	[33]	• Provides an extensive review of deep learning methods for camera-LiDAR fusion. • Offers benchmarking and model comparisons for KITTI dataset for various tasks. • Identifies open challenges and future directions in autonomous vehicle systems.	• 2D object detection and traditional sensor reviews are not included in paper. • Does not cover cooperative perception frameworks and datasets such as V2V. • Lacks a detailed review of AV datasets including scopes and characteristics.	☑ AV Multimodal Fusion
2020	[34]	• Offers a review of deep multi-modal object detection and semantic segmentation. • Proposes a structured methodology for multi-modal fusion and fusion operations. • Discusses key open challenges such as sensor misalignment and label quality.	• Despite the validity of the discussed approaches, they are quite outdated. • Lack of detailed and technical information about AV datasets and sensors. • There is no information about CP, LLMs, VLMs, foundation models in AVs.	☑ Detection & Segmentation
2019	[35]	• Provides an overview for assessing drivability in autonomous vehicle driving. • Reviews a taxonomy of 54 public driving datasets categorized by characteristics. • Highlights open challenges in metric design and joint attention modeling in AVs.	• Despite the validity of the discussed methods, they are quite outdated. • focuses heavily on dataset and risk assessment, but not sensors and hardware. • Does not review or analyze object detection methods (2D, 3D, fusion-based).	☑ AV Object Detection
Our Survey		• **Reviews state-of-the-art 2D (Camera) object detection techniques in AVs.** • **Reviews state-of-the-art 3D (LiDAR) object detection approaches in AVs.** • **Surveys Camera–LiDAR fusion techniques for object detection in AVs.** • **Provides an extensive overview of LLMs, VLMs, and ViTs in AV systems.** • **Explores LLM/VLM-based multimodal methods for object detection in AVs.** • **Analysis CP frameworks (V2V, V2I, V2X, I2I), as well as their datasets.** • **Discusses AV sensor types, architectures, technologies, and modalities.** • **Reviews AV dataset information, types, characteristics, and applications.** • **Outlines existing simulation platforms used for AV dataset generation.** • **Highlights challenges and open problems for each discussed topic in AVs.**	• **N/A**	☑ **AV Object Detection**

tasks. Moreover, the paper provides valuable insights into the strengths and limitations of fusion strategies across multiple benchmarks. However, it does not cover recent advancements in multimodal foundation models, a detailed analysis of large-scale AV datasets, and segmentation frameworks. In 2020, a paper titled "Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving" [34] provides a detailed review of multi-modal methods for object detection and semantic segmentation tasks in AVs. It highlights the advantages of combining data from different sensors such as cameras, LiDAR, and Radar, and presents an overview of datasets, fusion strategies. However, the paper does not sufficiently address recent advances in sensors modalities, all the existing datasets, foundation models, vision-language approaches, and cooperative driving. In 2019, the survey paper [35] provided a comprehensive review of object detection methods based on deep learning, focusing on the evolution of CNN-based models and benchmark datasets. It sheds light on both one-stage and two-stage detectors, as well as key improvements in detection performance and computational efficiency. It also reviews a taxonomy of 54 public driving datasets categorized by their types and characteristics. Nevertheless, multi-modal fusion techniques for object detection in autonomous driving are not addressed. Additionally, the integration of language-driven models for scene understanding and the discussion of large-scale AV-specific datasets and sensor analysis remain incomplete.

3. AV Technology and Sensor Specifications

This section provides an overview of existing AV technologies, with a particular focus on recent advancements in AV sensors and device specifications relevant to autonomous driving. In the domain of autonomous vehicles, sensors play a critical role in ensuring robust perception and accurate environment understanding. Modern computer vision systems heavily rely on a wide range of sensors such as cameras, LiDAR, Radar, or ultrasonic devices to precisely detect and monitor surrounding objects. The subsequent subsection will particularly discuss the details of various sensors deployed in AVs for efficient perception, environment understanding, and robust decision-making.

3.1. AVs' Sensor Types

AV sensors are generally categorized into various groups based on various factors such as functionality, range, resolution, energy, design, and purpose. In this survey, we have categorized them into two major types based on their operational principles: proprioceptive and exteroceptive sensors. Proprioceptive sensors, also known as internal state sensors, capture the vehicle's internal conditions such as velocity, acceleration, angular rate, tire pressure, or voltage levels. Common examples of sensors used in this category include IMUs (Inertial Measurement Units), GPSs (Global Positioning Systems), GNSSs (Global Navigation Satellite Systems), gyroscopes, and magnetometers [36]. On the other hand, exteroceptive sensors, often referred to as external state sensors, collect data from the vehicle's surrounding environment. This category includes cameras, LiDAR (Light Detection and Ranging), RADAR (Radio Detection and Ranging), and ultrasonic sensors, which are commonly used for perception tasks such as object detection, obstacle avoidance, and environmental mapping [37].

In addition to this classification, sensors can be grouped as passive or active based on their signal emission behavior. Passive sensors rely only on surrounding signals without emitting any energy, such as RGB and Thermal cameras. However, active sensors emit energy into the environment and then measure the reflected signals to determine the output, such as LiDAR and Radar [38]. Figure 3 provides an overall perspective and statistical analysis of various sensors used in AV. This figure visually presents the performance analysis of each AV sensor across multiple parameters, providing a clear assessment of the strengths and limitations of each sensor type. For convenience in comparison, each sensor attribute is ranked on a scale from 1 (Very Low) to 5 (Very High). Additionally, the possible positions of each sensor on the vehicle are illustrated using their corresponding colors (Yellow for the camera, Blue for the ultrasonic, Green for LiDAR, and Red for Radar).

To further elaborate on the characteristics of each AV sensor, the following Table 2 provides a detailed specification on the four major exteroceptive sensors commonly integrated in autonomous vehicles: camera, ultrasonic, LiDAR, and Radar sensors. A thorough understanding of the specifications and capabilities of each sensor is crucial for more efficient and reliable autonomous driving, enabling accurate perception and timely decision-making. Given that the primary focus of this survey is on the vision and perception aspects of AV systems, we exclude proprioceptive sensors and concentrate exclusively on exteroceptive sensing technologies. Each subsection will highlight the operational principles, technological foundations, and role of these sensors in AV systems, as well as their potential strengths and weaknesses.

Table 2: Detailed specifications of the most significant AV sensors employed in autonomous driving systems.

Criteria	Metric	Camera	Ultrasonic	LiDAR	Radar	Additional Note
Lateral Resolution	Pixels	Very High	Very Low	High	Low	☑ Indicates the level of detail the sensor captures across the image plane (X and Y).
Angular Resolution	Degrees (°)	High	Very Low	Very High	Low	☑ Determines the minimum angle between two objects that a sensor can measure.
Device Speed	Latency	High	Low	Moderate	Moderate	☑ Defines how quickly sensor can capture and update the data from environment.
Sensor Dimension	Size (cm^3)	Medium	Very Small	Large	Medium	☑ Impacts sensor feasibility, aerodynamics, and design constraints in AVs.
Day Operation	Visibility	Excellent	Good	Very Good	Good	☑ Measures sensor performance under the bright lighting or sunlight glare.
Night Operation	Visibility	Poor	Good	Very Good	Very Good	☑ Indicates the sensor effectiveness in low-light or entirely darkness situations.
Adverse Weather	Performance	Poor	Moderate	Poor	Very Good	☑ Reflects the sensor robustness in heavy rain, fog, snow, or dust conditions.
Device Cost	USD ($)	Low	Very Low	Very High	Moderate	☑ One of the major factor for commercial deployment and large-scale AV integration.
Sensor Range	Meters (m)	Moderate	Very Short	Long	Very Long	☑ Specifies the maximum range at which detection performance remains reliable.
Color Detection	RGB	Yes	No	No	No	☑ A valuable ability for applications such as traffic sign and light recognition.
Distance Measurement	Accuracy (m)	Moderate	Low	Very High	High	☑ Measures object distance which is crucial for collision avoidance and navigation.
Velocity Measurement	Speed (m/s)	No	No	Limited	Very High	☑ Enables to estimate the speed of nearby moving objects surrounding the vehicle.
Depth Measurement	Resolution (mm)	Low	No	Very High	Moderate	☑ Describes the sensor's capability to capture distance from the sensor to surfaces.
Field of View	Degrees (°)	90–120°	30°	360°	60-120°	☑ Specifies the angular coverage the sensor can observe in a single frame.
Frame Rate	Rate (Hz)	30–60 Hz	<10 Hz	10-20 Hz	10-30 Hz	☑ Determines how frequently the sensor captures data over the time intervals.
Object Classification	Capability	Very High	No	Moderate	Low	☑ Indicates whether the sensor supports object classification task.
Noise Sensitivity	Level	High	Moderate	High	Low	☑ Reflects sensor's vulnerability to signal interference and false alarms.
Power Consumption	Watts (w)	Low	Very Low	High	Moderate	☑ Indicates the amount of electrical power required to operate the sensor.
Computational Load	Load	High	Very Low	High	Low	☑ Shows the required processing resources to handle sensor data.
Calibration Complexity	Level	High	Low	High	Low	☑ Defines the complexity of setup and ongoing calibration tasks of the sensor.
Fusion Compatibility	Level	High	Low	High	Moderate	☑ Shows how easily the sensor integrates into sensor fusion frameworks.
Maintenance Frequency	Frequency	Low	Moderate	Moderate	Low	☑ Suggests how often the sensor needs cleaning, recalibration, or maintenance.
Installation Complexity	Level	Low	Very Low	High	Low	☑ Reflects effort and constraints involved in mounting the sensors on vehicle.
Sensor Durability	Lifespan	Moderate	Low	High	Very High	☑ Indicates sensor robustness to vibration, impact, and environmental degradation.
Additional Detail	☑ Specifies how each parameter is measured (e.g., degrees, pixels, meters), enabling objective sensor comparison across various key factors such as resolution, range, cost, environmental robustness, and performance in AV perception tasks.	☑ Captures high-resolution 2D (RGB or grayscale) image data, enabling object classification, traffic sign recognition, and lane detection. It should be noted that the performance degrades in low-light, glare, fog, or rain.	☑ Uses high-frequency sound waves to measure distance to nearby objects. It is suitable for parking assistance, blind spot detection, and short-range obstacle avoidance. Extremely low-cost and compact, but limited to close-range sensing.	☑ Emits laser pulses to generate high-density 3D point clouds, enabling precise depth estimation, surface reconstruction, and shape recognition. It offers 360° scanning, however, it is expensive and performance degrades under adverse weather.	☑ Transmits radio waves to detect object range, angle, and velocity. It is a valuable tool for tracking moving targets and operating in darkness, rain, or fog. However, it suffers from low spatial resolution and difficulty distinguishing object shapes.	☑ Here are the most potential applications of each sensor that align with their characteristics: **Camera:** Object Recognition, Pedestrian Detection, Traffic Sign Detection. **Ultrasonic:** Parking assistance, blind Spot Detection, Obstacle Avoidance. **LiDAR:** Environment Mapping, Obstacle Detection, Distance Estimation. **Radar:** Lane Change Assistance, Moving Object Detection, Speed Measurement.

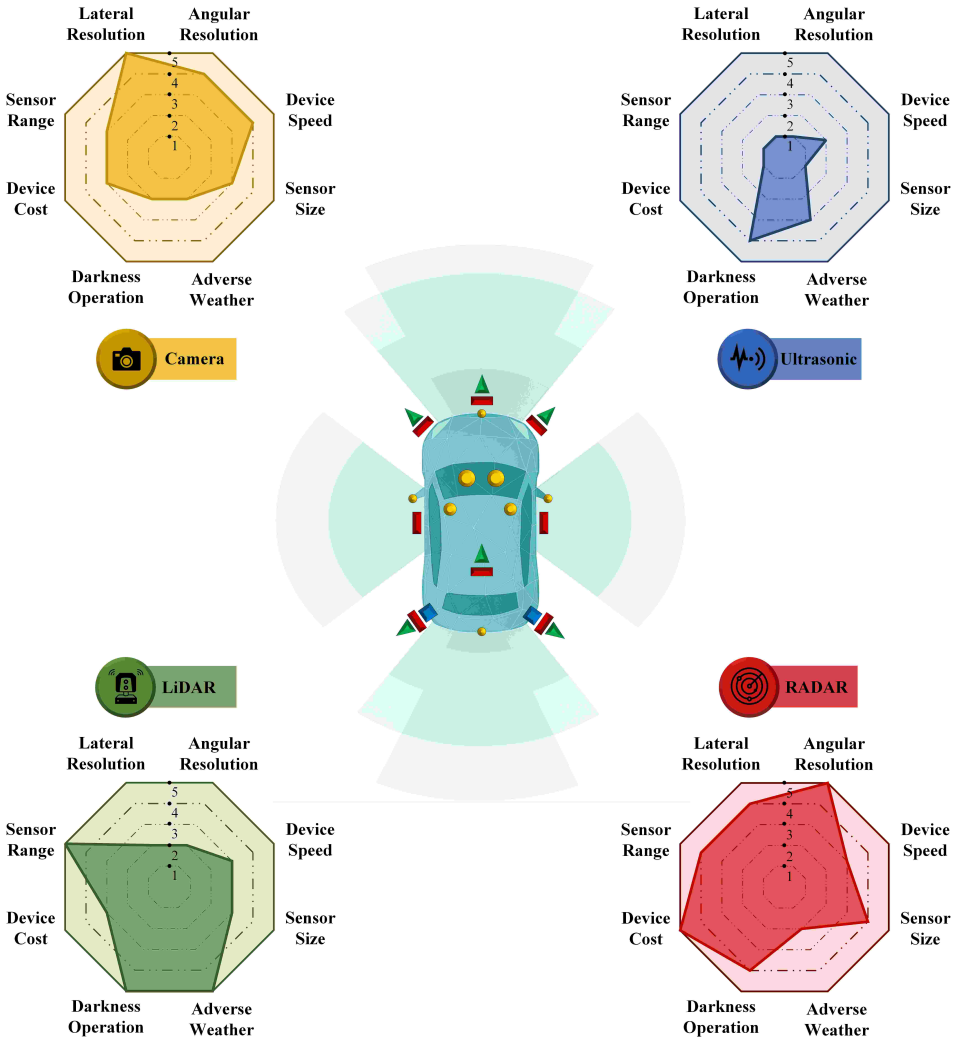

Figure 3: Overview of major sensors used in AVs based on their types and perception performance. Sensor performance is evaluated on a scale from 1 to 5, where 1=Very Low, 2=Low, 3=Medium, 4=High, and 5=Very High

3.2. Camera Sensor

Autonomous driving systems rely mainly on cameras as the primary sensors, since they provide crucial visual data for a wide range of perception tasks. There are four major types of cameras used in AVs: monocular, stereo, depth, and thermal cameras. Monocular cameras use a single lens to capture 2D images of the environment, much like standard digital cameras. Their ability to produce high-resolution images, along with their low-cost design, makes them a favorable choice for many perception tasks in AVs. Additionally, they enable the extraction of detailed texture patterns, rich color information, and overall visual appearance. In modern systems, monocular vision is often paired with deep learning models such as CNNs and ViTs to infer depth information and extract semantic content from source images. However, since these systems rely on learned context and visual patterns to estimate depth, they require robust training data and well-designed algorithms to perform reliably under varying driving conditions. Monocular cameras suffer from several limitations, including vulnerability to light and weather conditions, as well as an inability to provide depth information. Despite these drawbacks, they are widely used in lane detection, traffic sign recognition, and object classification.

The stereo camera is another vision system widely used in autonomous driving. This sensor imitates human binocular vision through two horizontally spaced lenses to calculate depth measurements by analyzing disparities between the left and right image views. This disparity information enables the AV to create real-time 3D spatial reconstructions of the surrounding environment. However, precise stereo performance depends heavily on suitable calibration to ensure accurate alignment and consistent depth measurement. Moreover, stereo cameras demand significant computational resources, especially when operating in low-light conditions, on low-texture surfaces, and at long-range distances. By considering the range and lighting conditions, stereo systems are particularly effective for various tasks such as object detection, obstacle avoidance, and motion tracking [39]. On the other hand, depth cameras provide distance information directly by using active sensing methods

such as structured light or time-of-flight technology. Unlike stereo cameras that rely on visual disparity, depth cameras emit signals (infrared patterns or light pulses) and measure the time of the reflected signal to estimate depth. This direct measurement makes depth cameras more robust in low-texture environments and scenes with limited visual features. Since they need to operate at short ranges from objects, they are often used in low-speed driving and indoor navigation scenarios. Lower performance under strong sunlight, limited effective range, and sensitivity to reflective or transparent surfaces are some of the key limitations of depth camera sensors. Despite these drawbacks, they remain a valuable sensor in AV systems, especially when fused with other sensors.

The thermal camera is the last type of vision sensor commonly used in autonomous driving systems. It captures infrared radiation emitted by objects rather than relying on visible light for the perception of the surrounding environment. This makes them a reliable choice in situations where standard cameras often fail to perform very well, such as nighttime driving or conditions involving fog, heavy rain, smoke, or direct sunlight. They play a crucial role in enhancing safety by enabling pedestrian and wildlife detection, as well as emergency braking in low-visibility conditions. However, thermal cameras usually offer lower spatial resolution and are incapable of capturing color or fine texture details. They also require specialized processing algorithms, which make them more computationally expensive. Despite these limitations, thermal cameras provide an important complementary sensing modality, enhancing the overall robustness and safety of AV perception systems [40]. In summary, AVs can benefit from various camera sensors, as each camera type offers unique strengths and limitations depending on environmental conditions, object types, and specific driving tasks. Table 3 highlights the key advantages and disadvantages of each camera type, providing a comprehensive comparison to assist researchers in choosing the most appropriate camera for particular objectives.

Table 3: Summary of the advantages and disadvantages of the various types of camera sensors.

Sensors	Advantages	Disadvantages
Monocular Cameras	• Low-cost solution for visual scene understanding. • Lightweight and easy to integrate with AV system. • Captures rich texture and detailed color information. • Supports classification and object detection tasks.	• Lacks natural depth perception for 3D estimation. • Performance notably degrades in low-light conditions. • Highly sensitive to weather and illumination changes. • Relies heavily on the high-quality training dataset.
Stereo Cameras	• Provides real-time depth info from image disparity. • Enables 3D reconstruction of nearby environment. • Passive sensing without emitting external signals. • Captures color and rich texture visual information.	• High computational cost for disparity matching. • Struggles on texture-less and uniform surfaces. • Requires precise calibration between camera lenses. • Performance drops in adverse lighting conditions.
Depth Cameras	• Directly captures depth from a single viewpoint. • Offers precise short-range 3D scene understanding. • Compact and highly suitable for close-range tasks. • Provides texture, color, and appearance information.	• Limited distance range compared to stereo or LiDAR. • Frequently struggles in direct and bright sunlight. • Performance affected by reflective surface materials. • Less effective in consistent long-distance depth sensing.
Thermal Cameras	• Detects heat signatures in complete darkness. • Performs very well in fog and glare environments. • Enhances safety in low-visibility conditions. • Useful for night-time pedestrian detection.	• Expensive compared to standard vision cameras. • Lacks color and texture visual information. • Lower spatial resolution than RGB cameras. • Difficult to interpret without contextual data.

3.3. Ultrasonic Sensor

Ultrasonic sensors are commonly used in modern vehicles for short-range proximity sensing. They measure distances by emitting high-frequency sound waves and analyzing their echoes. These sensors operate by generating sound waves ranging from 20 kHz to several hundred kHz (depending on the sensor type), which exceed the human hearing range. Then, they measure the distance between the vehicle and the object by calculating the time it takes for the emitted sound wave to reflect off the object and return as an echo. Generally, ultrasonic sensors can be categorized into one of the following three types: short-range, mid-range, and long-range. Short-range sensors (<2 m) provide high-resolution distance measurements, which are suitable for parking assist and blind-spot detection at low speeds. Mid-range sensors (2–5 m) offer moderate resolution and distance range, making them ideal for obstacle avoidance and curb detection. Lastly, long-range sensors (>5 m) extend detection capabilities for high-speed collision warning and closing-speed estimation on freeways or highways.

Nowadays, the use of ultrasonic sensors in modern vehicles has increased, as they serve as essential components for Advanced Driver-Assistance Systems (ADAS). These sensors are typically embedded in the front and rear bumper corners for parking assistance and blind-spot detection, or in the vehicle's front grille or upper bumper to enable high-speed collision warnings. Ultrasonic sensors offer several advantages for AVs by enhancing

safety and reliability through close-range environmental monitoring. Additionally, unlike optical sensors, they perform very well in complete darkness, which makes them useful for night-time driving. They are also highly cost-effective compared to alternatives like LiDAR and Radar. However, their limited detection range makes them less effective not only for detecting long-distance objects but also in high-speed situations. Also, their performance can be decreased in the presence of intense traffic noise due to interference with high-frequency sound waves. Furthermore, environmental factors such as rain, humidity, and temperature fluctuations can affect their accuracy. Furthermore, they are highly sensitive to environmental factors such as rain, humidity, and temperature. As a result, ultrasonic sensors are well-suited for short-range tasks such as parking assistance and nearby object detection at low speeds.

3.4. LiDAR Sensor

LiDAR sensors are among the most fundamental components in modern AV perception systems. These devices calculate distances by transmitting laser pulses and measuring the time it takes for the reflected signals to return. As a result, the system is able to construct 3D point cloud data of the environment for accurate spatial awareness and understanding of the surroundings. Unlike 1D (one-dimensional) sensors that capture only distance information (x-axis) or 2D (two-dimensional) sensors that measure angular position (y-axis), 3D LiDAR sensors measure vertical coordinate (z-axis) and provide detailed three-dimensional representations of the environment. LiDAR technology is typically categorized into three main types: mechanical LiDAR, solid-state LiDAR, and hybrid LiDAR. Mechanical LiDAR, including rotating mirror and spinning sensor subtypes, uses rotating components to provide 360° scanning with high-resolution data. Fully Solid-State LiDAR includes two main types, where Flash LiDAR captures the entire scene with a single light pulse, and Optical Phased Array (OPA) uses electronic beam steering for rapid scanning [41]. Hybrid LiDAR, including Micro-Electro-Mechanical Systems (MEMS) and Risley Prism subtypes, incorporates minimal moving parts like micro-mirrors or prisms to steer laser beams. Table 4 presents the major pros and cons of each LiDAR type to provide valuable information for researchers to choose the most suitable system for their specific application.

Autonomous driving systems significantly benefit from LiDAR technologies due to the several advantages they offer in AVs. First of all, they are able to provide exact distance measurements not only during the day but also at night. Additionally, they perform very accurately and reliably in adverse weather conditions, such as heavy rain, snow, and fog. They are able to capture detailed geometric information about the surroundings, which enhances object recognition and environmental mapping. In contrast, LiDAR sensors face several limitations in autonomous driving, including high cost and mechanical vulnerability. Moreover, the data generated by LiDAR requires substantial processing power and storage capacity due to their high volume and complexity.

Table 4: Summary of the advantages and disadvantages of the various types of LiDAR sensors.

Sensors	Advantages	Disadvantages
Mechanical LiDAR	• Provides full 360° environmental coverage. • Offers high spatial scanning resolution. • well-established in the automotive industry. • Enables accurate long-range object detection.	• Limited vehicle integration due to bulky design. • High manufacturing and maintenance costs. • Mechanical components degrade over time. • Requires complex alignment for AV integration.
Solid-state LiDAR	• High reliability due to non-mechanical parts. • Offers high durability over the long term. • Compact structure simplifies system design. • Resistant to shock and vibration damage.	• Lower resolution than other LiDAR types. • Limited detection range compared to mechanical. • Limited availability in current market offerings. • Less effective in complex terrain environments.
Hybrid LiDAR	• Compact size enables easier sensor placement. • Lower cost than mechanical LiDAR systems. • Faster scanning speed than other types. • Easier integration into modern AV designs.	• Limited field of view coverage range. • The Calibration process may be complicated. • Performance may degrade in bright sunlight. • Calibration process can be technically complex.

3.5. Radar Sensor

RADAR (Radio Detection and Ranging) sensors are another primary technology widely used in modern vehicles. These sensors emit radio waves to detect the distance and speed of surrounding objects. They measure object speed using the Doppler effect, which analyzes frequency shifts caused by moving objects. The two main types of Radar sensors broadly used in AVs are Impulse Radar and FMCW (Frequency-Modulated Continuous

Wave) Radar. The Impulse transmits discrete radio pulses, while FMCW continuously emits modulated signals to provide superior depth perception and range. In AVs, Radar sensors are typically mounted on the front and rear bumpers to ensure full environmental awareness. Based on their range capabilities, they serve as an integral part of ADAS in modern vehicles. Short-Range Radar (SRR) is particularly effective for detecting objects during low-speed maneuvers, while Medium-Range Radar (MRR) and Long-Range Radar (LRR) are ideal for monitoring vehicles and obstacles at higher speeds on highways and busy roads [39]. It is worth mentioning that the integration of Radar sensors with other sensor data has made it a rising trend in automotive safety.

Table 5 outlines the key features of each Radar type, as well as providing major pros and cons of Radar sensors commonly used in autonomous vehicles to guide researchers in choosing the most appropriate Radar for particular objectives. Using Radar sensors in AV applications provides multiple benefits. These sensors operate reliably under adverse weather conditions and low-light environments, where optical sensors often struggle. These sensors also provide precise speed and distance measurements, which are highly crucial for collision avoidance and adaptive cruise control systems. Furthermore, Radar technologies are significantly more cost-effective than alternatives such as LiDAR sensors. However, the spatial resolution of Radar systems is lower than LiDAR and modern camera-based systems. The reflection of radar signals from metal objects such as road signs and guardrails sometimes results in incorrect detection. Despite these limitations, the reliable performance and cost-effectiveness of Radar sensors continue to support their important role in autonomous driving systems.

Table 5: Summary of the key features, advantages, and disadvantages of various Radar sensor types.

Sensors	Range	FOV	Advantages	Disadvantages
Short-Range Radar	Short ∼0.2-30 m	Wide ∼120°	• Fast response in close range. • Wide-angle object detection. • Effective in tight spaces. • Low power consumption.	• Lower resolution than cameras. • Limited detection distance. • Poor performance in heavy rain. • Affected by surface interference.
Medium-Range Radar	Medium ∼30-80 m	Medium ∼60°-90°	• Balanced range and FOV. • Good for mid-speed driving. • Reliable in poor weather. • Works very well in traffic.	• Not ideal for highway use. • May misclassify static objects. • Resolution not very high. • Limited vertical coverage.
Long-Range Radar	Long ∼80-250 m	Narrow ∼20°-30°	• Perfect long-distance tracking. • Precise velocity measurement. • Robust in adverse weather. • Crucial for highway driving.	• Cannot detect nearby objects. • Expensive and larger size. • Narrow field of view. • Complex to integrate.

3.6. Sensor Fusion

Sensor fusion plays a foundational role in AVs' perception by combining data from multiple sensor types to enhance the accuracy and reliability of environmental understanding. It leverages the strengths of each sensor to address the limitations inherent in individual sensing modalities. AV systems require three essential factors for effective sensor fusion integration. First of all, a detailed understanding of each sensor's characteristics is necessary for determining its strengths, limitations, and optimal role within the fusion framework. Secondly, appropriate sensor calibration is required to ensure optimal performance by accurately aligning sensor placements and synchronizing timing data across all devices. Last but not least, selecting and evaluating the right sensor fusion algorithm is highly critical for effectively combining multi-sensor data to achieve robust and accurate environmental perception. Consequently, an effective sensor fusion system for AVs is built upon three key principles: comprehensive sensor understanding, precise calibration, and careful algorithm selection [38].

One of the most common and well-known applications of sensor fusion in AVs is object detection, where data from multiple sensors is combined to identify and classify objects in the driving environment, such as vehicles, pedestrians, and road signs. AVs require fast and precise object detection capability for more informed driving decisions to enhance road safety and reduce the risk of harm to people inside and outside the vehicle. Radar sensors perform reliably in adverse weather conditions and provide precise velocity measurements. Although they are valuable devices for dynamic environment monitoring, they face significant challenges in spatial resolution and object classification. In other words, Radar cannot accurately recognize and distinguish objects that are either non-moving or have complex shapes [42]. Radar data is often combined with inputs from cameras and LiDAR to overcome these limitations. The camera sensor provides color and texture information, and LiDAR offers high-resolution 3D spatial data. Therefore, Radar improves detection performance in low-visibility

conditions, while the camera enhances object recognition and spatial accuracy. Integrating multiple sensors results in a more robust object detection system, supporting safer and more reliable autonomous navigation.

Table 6 presents a comprehensive overview of sensor fusion methods categorized by fusion level: low-level, mid-level, and high-level fusion. In low-level fusion, raw data from multiple sensors (e.g., pixel values, point clouds, or raw Radar signals) are combined to provide rich sensory information. This level requires precise calibration and high computational resources due to the complexity of processing raw sensor data. Mid-level fusion integrates intermediate features from individual sensors (e.g., edges, depth maps, or motion cues) to offer a balanced trade-off between accuracy and efficiency. High-level fusion combines the final decisions from separate processing modules, such as detected objects or classified regions. This approach makes the system more robust and modular, but may result in the loss of some finer details in the process.

Table 6: Summary of sensor fusion methods categorized by processing level, highlighting their sensor pairs, complementarity, applications, as well as their pros and cons.

Fusion Level	Sensor Pair	Complementarity	Use Case	Pros (✓) / Cons (✗)
Low-Level Fusion	Lidar + Camera	Depth + Texture	• 3D object detection	✓ Detailed spatial information
	Lidar + Radar	Dense + Motion	• Obstacle mapping	✓ Rich sensory information
	RGB + Thermal	Visual + Thermal	• Low-light vision	✗ Requires precise calibration
	Stereo Cameras	Dual Perspective	• Thermal detection	✗ High processing cost
Mid-Level Fusion	Camera + Lidar	Feature + Geometry	• Obstacle detection	✓ Accuracy-cost efficiency
	Camera + Radar	Visual + Velocity	• Lane assistance	✓ Simpler than low-level
	Camera + Ultrasonic	Context + Proximity	• Scene classification	✗ Feature alignment needed
	Lidar + Ultrasonic	Shape + Proximity	• Traffic sign recognition	✗ Moderate latency
High-Level Fusion	All 4 Sensor Types	Safety Redundancy	• Trajectory planning	✓ Easy to implement
	Camera+Lidar+Radar	Decision Aggregation	• Localization support	✓ Robust to missing data
	V2X + Camera/Lidar	Communication	• Object tracking	✗ Inconsistency in decisions
	IMU + GPS	Sensor + Localization	• Mission planning	✗ Low spatial precision

In summary, multi-sensor fusion technology offers significant advantages for AV driving systems by enhancing environmental perception and operational safety. Fusing data from various sensors, including camera, LiDAR, RADAR, and ultrasonic sensors, improves the system's accuracy and reliability across diverse conditions. Using multiple sensors makes the system more secure and resilient by addressing individual sensor limitations and providing backup in case of sensor failure. On the other hand, there are several challenges in multi-sensor fusion that should be considered for achieving efficient autonomous driving systems. First, the system results in a high computational cost to process large volumes of heterogeneous data in real time, as well as maintaining synchronization and alignment. Additionally, the system's performance is highly dependent on accurate calibration and precise spatial and temporal alignment. Lastly, achieving effective sensor fusion requires developing more advanced algorithms to integrate and interpret data from different sensor types, which increases the overall system complexity. Despite these challenges, sensor fusion remains essential for building scalable and dependable autonomous driving systems that prioritize safety.

3.7. Challenge, Discussion, and Future Directions

Despite significant advancements in AV sensor technologies and fusion frameworks, several technical and practical challenges remain in real-world deployment. A primary challenge is integrating and calibrating heterogeneous sensors with different resolutions, response times, and data formats. Real-time processing of massive, multimodal sensory data imposes a high computational cost, affecting decision-making in time-critical scenarios. Additionally, individual sensor limitations, such as the inability of cameras to perform well in poor lighting or LiDAR's performance degradation in adverse weather, directly affect overall system robustness. Sensor fusion, although powerful, introduces complexity in synchronization, feature alignment, and temporal consistency, making the system prone to drift or failure if not finely tuned. Furthermore, complex scenarios such as uncommon road geometries or unexpected human behavior are often not well represented in existing training datasets, which limits the ability of perception algorithms across unseen conditions.

Looking ahead, several novel and forward-thinking directions show great promise for the next generation of AV perception systems. One emerging idea is context-aware sensor fusion, where the vehicle intelligently adjusts which sensors to prioritize based on the current driving scene—for example, relying more on radar during heavy

fog or emphasizing camera input in urban environments. Another exciting direction is using neuromorphic event-based cameras, which mimic the human retina and offer faster response times with lower data bandwidth. We also see potential in collaborative learning across AV fleets, where vehicles share knowledge about sensor calibration, unusual objects, or environmental conditions without transferring raw data to preserve privacy while improving the cooperative perception system. Lastly, using self-supervised cross-sensor learning, where one sensor helps teach another without labeled data, could dramatically reduce the cost and time needed to train reliable perception systems. These forward-looking ideas offer fresh possibilities for building smarter, safer, and more adaptable autonomous vehicles in the years ahead.

4. Autonomous Vehicle Datasets

In the realm of autonomous vehicle research, the accessibility and diversity of high-quality datasets are fundamental to advancing perception systems, decision-making models, and overall driving safety. This section presents a comprehensive review of existing AV-related datasets, emphasizing their context, characteristics, applications, and strengths and limitations. These datasets differ significantly in terms of collection methods, sensor modalities, data granularity, and annotation depth, which can directly influence their effectiveness for various research objectives such as object detection, tracking, behavior prediction, and scene understanding. Recognizing these differences is critical for researchers seeking to identify suitable benchmarks for developing and evaluating AV algorithms. To facilitate a clear and systematic overview, we have categorized the existing AV datasets into seven major categories based on their specifications and purposes. Every AV dataset can be reasonably classified into one of the following types: Ego-Vehicle Perception, Roadside Perception, Vehicle-to-Language (V2L) Datasets, Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), Vehicle-to-Everything (V2X) and Infrastructure-to-Infrastructure (I2I). Each category represents a distinct communication and perception framework within the AV domain, shaped by specific design constraints and research objectives. Figure 4 provides a comprehensive insight into the overview of various dataset categories used in AVs.

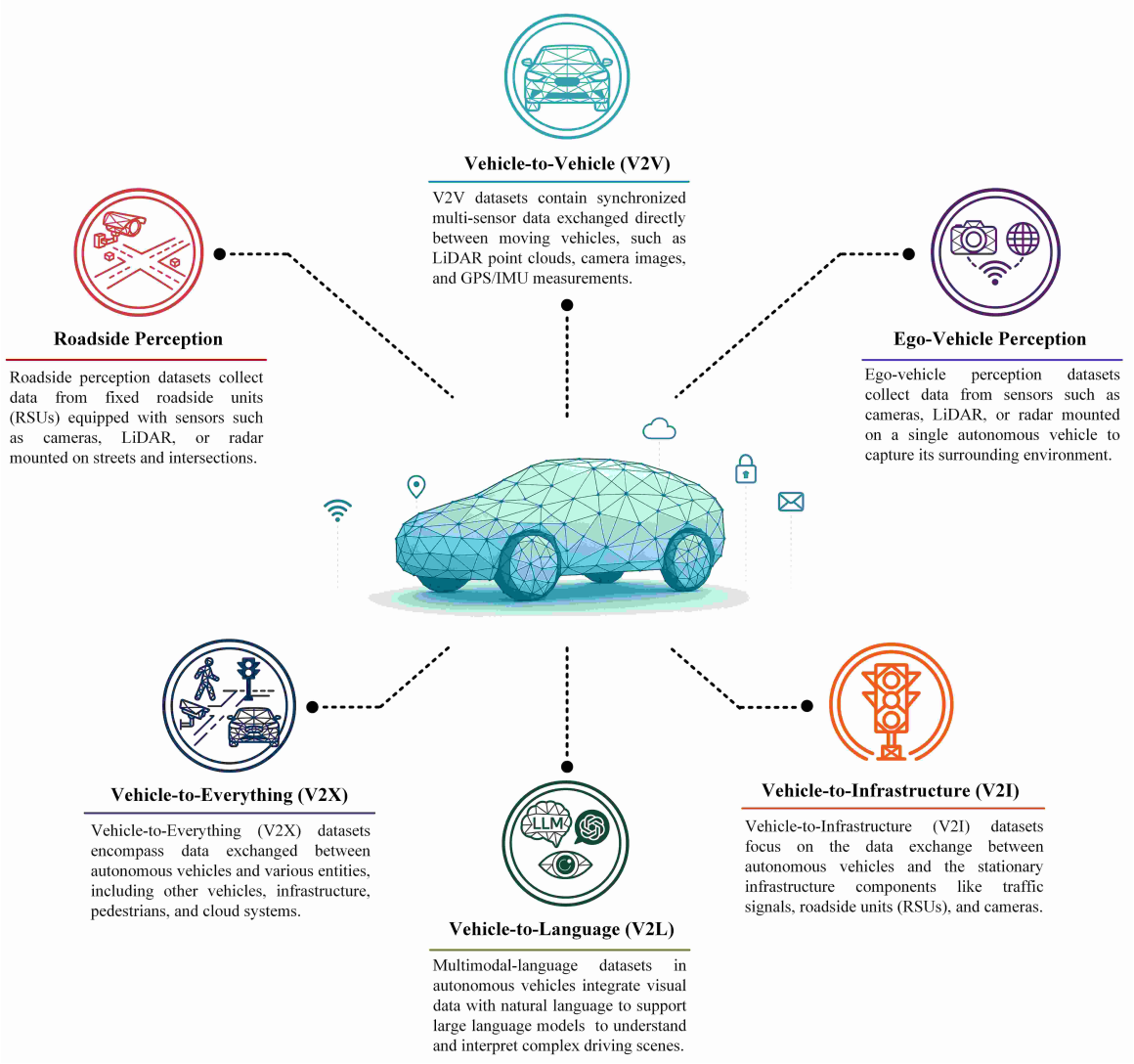

Figure 4: Overview of major AV datasets based on their specifications and applications.

Ego-Vehicle Perception datasets are collected directly from sensors mounted on the AV itself, capturing the vehicle's first-person view of its surroundings. Roadside Perception datasets originate from fixed infrastructure, such as roadside cameras or LiDAR units, offering complementary, third-party observations of traffic environments. Vehicle-to-Language (V2L) Datasets combine visual inputs with natural language descriptions or queries to enable semantic understanding and reasoning within AV systems. Vehicle-to-Vehicle (V2V) datasets capture communication and interaction data exchanged between vehicles to enhance cooperative perception and coordination [43]. Vehicle-to-Infrastructure (V2I) datasets focus on data shared between vehicles and static infrastructure elements like traffic lights or roadside units for improved navigation and safety. Vehicle-to-Everything (V2X) encompasses broader connectivity scenarios, integrating V2V, V2I, and interactions with other road users such as pedestrians or cyclists. Finally, Infrastructure-to-Infrastructure (I2I) datasets involve the communication and synchronization of data between multiple infrastructure nodes to create a cohesive traffic management framework.

The subsequent subsections will delve into further detail on each AV dataset category by assessing the key features of each dataset, such as the number of samples, data modalities, labeling schemes, and the specific methodologies adopted for data acquisition and annotation. Through a detailed analysis of these critical aspects, this work offers a comprehensive and structured resource for researchers and practitioners, allowing informed decisions when selecting datasets for training, testing, or benchmarking AV models. Ultimately, this collection contributes to the broader landscape of autonomous driving research by supporting reproducibility, robustness, and innovation in the development of intelligent transportation systems.

4.1. Ego-Vehicle Perception Datasets

Ego-vehicle perception datasets are based on sensor data collected directly from the autonomous vehicle itself, such as cameras, LiDARs, Radars, or GPS modules. These datasets capture the vehicle's perspective as it navigates dynamic environments, and are crucial for developing core AV capabilities like object detection, semantic segmentation, motion planning, and trajectory prediction. Popular examples include KITTI, nuScenes, and Waymo Open Dataset, which enable benchmarking of perception and decision-making models based on real-time sensory input. Table 7 provides a comprehensive summary and comparative analysis of the ego-vehicle perception datasets. In the following, we categorize the datasets discussed based on the perception tasks they support (detection, segmentation, and tracking). Specifically, we divide them into five categories: (1) datasets supporting all three tasks; (2) datasets that support detection and segmentation; (3) datasets that support detection and tracking; (4) datasets that support detection only; and (5) datasets that support tracking only.

4.1.1. Detection, Segmentation, and Tracking Datasets

OmniHD-Scenes [45], Zenseact Open [46], and BDD100k [61] datasets support all three major perception tasks, making them ideal for end-to-end autonomous driving pipelines. The OmniHD-Scenes dataset [45] is a large-scale multimodal dataset developed and released by 2077AI in 2024 to provide complete omnidirectional high-definition data for autonomous driving. The dataset is collected from real-world roads across urban, suburban, and rural areas in China under varying weather and lighting conditions. OmniHD-Scenes integrates 360° high-definition LiDAR sensors, 4D imaging Radar sensors, and ultra-high-resolution RGB cameras. The 128-beam LiDAR sensor generates ultra-dense point clouds, providing precise geometric and depth information. Six high-resolution cameras capture front, rear, and surrounding views at resolutions of 3840×2160 and 1920×1080 pixels at 30 frames per second. Additionally, the vehicle is equipped with six 4D Radar sensors to enable omnidirectional environmental perception. Other onboard devices include a GNSS/IMU for measuring location and angular velocity, an ADCU for processing sensor data and making driving decisions, and an IPC for handling heavy data workloads. The advantages of this dataset are its extensive video clips, frames, and data points, combined with an advanced 4D annotation pipeline. It provides comprehensive omnidirectional high-definition coverage and a detailed multimodal annotation system. Moreover, the OmniHD-Scenes dataset offers strong generalization capabilities by capturing data from multiple cities under different weather conditions. This enables researchers to conduct long-range perception and robust dynamic object tracking under adverse conditions, as well as achieve complete 360° scene understanding without major occlusions or missing data regions. However, the OmniHD-Scenes dataset is limited by partial annotation coverage (only 200 of 1,501 clips are fully labeled) and requires substantial computational and storage resources due to the high resolution and large volume of its multimodal data.

The Zensact Open Dataset (ZOD) [46] was gathered over two years across fourteen European nations, from Sweden to Italy, to include diverse road conditions and environmental settings. It was designed to address existing dataset limitations by offering extensive sensing capabilities and diverse geographical coverage, as well as supporting multiple research applications such as detection, tracking, segmentation, sensor fusion, and weather

Table 7: Summary and comparative analysis of the Ego-Vehicle Perception Datasets in autonomous vehicles.

						Sensor			Application			
Ref	**Dataset**	**Year**	**Region**	**Frame**	**Format**	**Camera**	**Lidar**	**Radar**	**Detection**	**Segmentation**	**Tracking**	**Access**
[44]	MSU-4S	2024	USA	+100K	JPEG, PCD, YOLO	Yes	Yes	Yes	✓	✗	✗	Web
[45]	OmniHD-Scenes	2024	China	+450K	RGB, LiDAR, Radar	Yes	Yes	Yes	✓	✓	✓	Web
[46]	Zenseact Open	2023	Europe	+100K	LiDAR, RGB, Radar, GPS	Yes	Yes	Yes	✓	✓	✓	Web
[47]	Ithaca365	2022	USA	7K	LiDAR, RGB, GPS	Yes	Yes	No	✓	✓	✗	Web
[48]	ONCE	2021	China	8M	LiDAR, RGB	Yes	Yes	No	✓	✗	✗	Git
[49]	Leddar PixSet	2021	Canada	29K	LiDAR, RGB, Radar, IMU	Yes	Yes	Yes	✓	✗	✓	Web
[50]	Pandaset	2020	USA	64K	LiDAR, RGB	Yes	Yes	No	✓	✓	✗	Web
[51]	A2D2	2020	Germany	+40K	LiDAR, RGB	Yes	Yes	No	✓	✓	✗	Web
[52]	Udacity	2020	USA	15K	RGB, GPS	Yes	No	No	✓	✗	✗	Git
[53]	Waymo	2020	USA	+10M	LiDAR, RGB, TFRecord	Yes	Yes	No	✓	✗	✓	Web
[54]	A*3D	2020	Singapore	+20K	LiDAR, RGB	Yes	Yes	Yes	✓	✗	✗	Git
[55]	InD	2020	Germany	+50K	RGB, Trajectories	Yes	No	No	✗	✗	✓	Web
[56]	exiD	2020	Germany	+30K	RGB, Trajectories	Yes	No	No	✗	✗	✓	Web
[57]	nuScenes	2019	USA	1.4M	LiDAR, RGB, Radar, GPS	Yes	Yes	Yes	✓	✓	✗	Web
[58]	ArgoVerse	2019	USA	3M	LiDAR, RGB, HD Maps	Yes	Yes	No	✓	✗	✓	Web
[59]	BLVD	2019	USA	+100K	LiDAR, RGB	Yes	Yes	No	✓	✗	✓	Git
[60]	Comma2K19	2019	USA	+50K	RGB, GPS, IMU	Yes	No	No	✓	✗	✗	Git
[61]	BDD100k	2018	USA	120M	RGB, GPS	Yes	No	No	✓	✓	✓	Web
[62]	ApolloScape	2018	China	+140K	LiDAR, RGB	Yes	Yes	No	✓	✓	✗	Web
[63]	CityScape	2016	Germany	+25K	RGB	Yes	No	No	✓	✓	✗	Web
[64]	Oxford RobotCar	2015	UK	+20M	LiDAR, RGB, GPS, IMU	Yes	Yes	No	✓	✗	✗	Web
[65]	KITTI	2012	Germany	+40K	LiDAR, RGB, GPS, IMU	Yes	Yes	No	✓	✗	✓	Web

* For the reader's convenience, a direct hyperlink to each dataset's original source—whether a website or GitHub repository—is provided in the 'Access' column to enable easy access and download.

classification. The Zensact Open Dataset used multiple vehicles with multiple sensor types to obtain high-resolution data across various modalities. The camera system contains an 8MP RGB sensor, which provides a 120° field of view at 10.1 Hz. Additionally, it provides point cloud data from one Velodyne VLS128 and two Velodyne VLP16 LiDAR units, which extend their maximum range to 245 meters. Meanwhile, the Continental ARS513 B1 Radar system provides medium-range spatial data at 60ms intervals, and GNSS/IMU units deliver high-precision vehicle localization. The ZOD dataset contains detailed annotations across different sensor modalities, including 2D and 3D bounding boxes for both dynamic and static objects, semantic segmentation of lanes, road surfaces, and ego-road labeling, as well as road surface conditions (e.g., wet, snowy) and classification of 156 traffic sign types. The ZOD dataset demonstrates strong potential for model generalization across different environments, as it covers fourteen countries with diverse weather conditions, road structures, and traffic patterns. It combines high-resolution camera, LiDAR, and Radar sensors with detailed annotations, supporting advanced perception, fusion, and autonomous driving challenges. On the other hand, the ZOD dataset has several weaknesses. Although the dataset covers many areas in Europe, it has a lower volume, lacks data from other continents, and includes only short sequences (up to a few minutes).

The BDD100K (Berkeley DeepDrive 100K) dataset [61] is one of the largest and most diverse open datasets available for autonomous driving research. The video clips were captured from various cities across the United States, including San Francisco, Oakland, San Jose, and New York City. The video clips include over 1,000 hours of driving footage, 120 million frames, and GPS/IMU trajectory data for each video. This dataset offers several advantages, including coverage of a wide range of real-world driving scenarios such as different geographic locations, climate conditions, and daytime periods, which makes it well-suited for training robust and generalizable perception models. Additionally, it contains a broad range of dense annotations, including bounding boxes for object detection of cars, buses, trains, and pedestrians. Semantic segmentation and instance segmentation are also included to annotate every single pixel in the image. Notable drawbacks of this dataset

are lacking Lidar or Radar information, which limits its applicability for sensor fusion tasks, 3D object detection, and perception system development.

4.1.2. Detection and Segmentation Datasets

Ithaca365 [47], Pandaset [47], A2D2 [51], nuScenes [57], CityScape [63], and ApolloScape [62] datasets are useful for detection and segmentation tasks. The Ithaca365 [47] was collected through a 15km route in New York over the course of a year, focusing on temporal variability by recording daily scenes across all four seasons. This dataset used six RGB cameras, 3D LiDAR, and GNSS/INS to ensure accurate localization. These sensor modalities offer a detailed view of the environment suitable for multiple tasks such as 2D and 3D object detection, semantic segmentation, and depth estimation. A key strength of the Ithaca365 dataset is its ability to support strong modeling for long-term scene understanding and perception consistency at a low computational cost. In addition, it provides high-resolution RGB images, dense point clouds, and geo-referencing for both urban and suburban environments, accompanied by precise calibration files and metadata logs indicating the weather conditions at the time of capture. Regarding the limitations of the dataset, it contains only 7,000 frames, which is significantly fewer than other datasets and affects its generalizability. Furthermore, the dataset lacks Radar data collection, which restricts its applicability in sensor fusion tasks.

The PandaSet dataset [47] was collected using a Chrysler Pacifica minivan equipped with advanced sensors, capturing data across urban environments in California, USA. The sensor suite includes six cameras, a 360-degree spinning LiDAR (Pandar64), a forward-facing long-range LiDAR (PandarGT), a GPS unit, and IMU sensors. It contains 28 object classes (e.g., cars, cyclists, pedestrians) annotated with 3D bounding boxes, along with 37 semantic categories that provide point-level annotations. A key strength of this dataset is its high-resolution LiDAR point clouds from both the Pandar64 and PandarGT sensors, making it well-suited for sensor fusion tasks. However, the PandaSet dataset faces limited generalizability due to its lack of geographical diversity, as the data was collected from specific urban areas in California with less variation in road types, regions, and weather conditions. Furthermore, the camera-recorded scenes are relatively short (about 8 seconds per scene), which limits their effectiveness for some tasks such as long-term behavior modeling or route prediction. Lastly, the dataset does not include Radar data, which reduces its applicability for depth perception tasks. The A2D2 dataset [51] was collected from various urban, suburban, and rural areas in southern Germany to provide a multimodal dataset suitable for semantic segmentation, 3D object detection, and sensor fusion. The data was recorded using six camera sensors and five LiDAR units, all equipped with various hardware and software to ensure time synchronization for supporting real-time processing, multi-sensor data fusion, and offline analysis. It includes semantic segmentation annotations across 38 classes, 12,497 frames with 3D bounding box annotations, and approximately 392,556 unannotated sequential frames for self-supervised learning. The combination of multiple cameras and LiDAR units is one of the major advantages of the A2D2 dataset, enabling multimodal support for various perception tasks. However, the annotation process was not sequential, which may pose challenges for researchers requiring temporal continuity. Additionally, generalizability may be limited since the data were collected exclusively in Germany.

The Cityscapes dataset [63] was developed to serve two main purposes: providing a benchmark and offering a large-scale dataset for training and testing various algorithms in pixel-level and instance-level semantic labeling. The data were collected over a period of seven months (spring, summer, and fall) in 50 cities across Germany and neighboring countries. The dataset contains 5,000 finely annotated and 20,000 coarsely annotated images, with annotations spanning 30 classes grouped into 8 categories. For example, the "vehicle" category includes classes such as car, truck, bus, on rails, motorcycle, bicycle, caravan, and trailer. The Cityscapes dataset also provides metadata that includes GPS coordinates, ego-motion data, and outside temperature. The strength of the dataset lies in its large scale and high-quality annotations. Additionally, its comprehensive metadata enhances its utility for context-aware modeling. However, a key limitation is that it focuses exclusively on urban scenes, which reduces its applicability to highway or rural environments. The nuScenes dataset [57], released in March 2019, provides a comprehensive multimodal dataset for autonomous driving research. Data was collected from dense urban environments in Boston and Singapore, with 1,000 driving scenes, each lasting 20 seconds, totaling 5.5 hours of driving data. Each scene includes synchronized data from multiple sensors. Two Renault Zoe electric vehicles were equipped with six cameras (360° view), one LiDAR, five Radars, and a GPS/IMU. The dataset contains 1.4 million annotated 3D bounding boxes across 23 object classes, with additional attributes like visibility, pose, and activity. It also includes metadata such as velocities, accelerations, and tracking information.

Lastly, the ApolloScape dataset [62], released in 2018, was designed to support research in autonomous driving through extensive, semantically annotated data. It includes 143,906 video frames with pixel-level annotations for semantic segmentation collected from various urban environments in China under diverse traffic and weather conditions. A mid-size SUV was equipped with two VUX-1HA laser scanners, two front-facing VMX-

CS6 cameras, and an IMU/GNSS unit for accurate positioning and orientation. ApolloScape comprises multiple sub-datasets covering scene parsing, car instances, lane segmentation, self-localization, trajectory, tracking, and stereo tasks. It offers high-resolution RGB images with pixel-wise annotations for 28 object classes, dense LiDAR scans with point-level labels, and lane marking annotations categorized by color and type. Manually annotated trajectories further support behavior prediction research. Data is divided into training, validation, and testing sets. ApolloScape's strengths lie in its diversity, detailed annotations, and real-world complexity. However, its geographic limitation to China may affect generalizability, and manual labeling introduces potential human error. The dataset's size also poses challenges for storage and processing.

4.1.3. Detection and Tracking Datasets

Leddar PixSet [49], Waymo [53], ArgoVerse [58], BLVD [59], and KITTI [65] datasets provide annotated sequences that support object detection, as well as short-term and long-term tracking, with a strong emphasis on object localization and temporal consistency. Leddar PixSet [49] is a publicly available dataset for autonomous driving research and development released in 2021. It includes full-waveform LiDAR data to support developing and testing sensor-fusion algorithms. Data were collected during the summer months across urban and suburban areas of Quebec City and Montreal, Canada. The PixSet dataset contains 29,000 frames distributed over 97 sequences, with 1.3 million annotated 3D bounding boxes. The sensor suite mounted on a Toyota RAV4 includes a Leddar Pixell LiDAR capable of full-waveform data acquisition, a mechanical scanning LiDAR, three cameras, a Radar unit, and an integrated IMU/GPS. The PixSet dataset offers several notable strengths. It is the first dataset to include full-waveform data from a flash LiDAR, offering richer information compared to conventional datasets. The combination of LiDAR, Radar, cameras, and IMU/GPS supports thorough sensor fusion research. Data collected across varying weather and lighting conditions contributes to the robustness of algorithms trained on it. Moreover, with over 1.3 million annotated 3D bounding boxes, the dataset is well-suited for developing advanced tracking algorithms. However, a key drawback is its limited geographic diversity.

The Waymo dataset [53] is designed to offer a large-scale, diverse, and high-quality resource for object detection and tracking. It features 2D (camera-based) and 3D (LiDAR-based) bounding box annotations. Labeled object categories include vehicles, cyclists, pedestrians, and traffic signs. In the 2D data, objects are annotated with tightly fitting, axis-aligned bounding boxes with four degrees of freedom (x, y, w, h). In the 3D data, bounding boxes include seven degrees of freedom $(x, y, z, l, w, h, \theta)$ to capture spatial information precisely. Each object in 2D and 3D data is given a unique ID maintained across frames to support tracking. Each object is assigned a consistent ID across frames in both 2D and 3D modalities, facilitating robust tracking. The Waymo dataset includes 1,150 scenes (6.4 hours) with synchronized camera and LiDAR data, split into 798 training, 202 validation, and 150 test scenes. Data was captured using five cameras and five LiDAR sensors, providing full 360° coverage in several cities, including San Francisco, Mountain View, and Phoenix, offering a variety of urban and suburban scenes. A primary benefit of the Waymo dataset is its comprehensive 2D and 3D annotations for object detection and tracking using synchronized camera and LiDAR data. It also enables research into camera-only systems through its 2D annotations. While the dataset covers several U.S. cities with urban and suburban scenes, it lacks representation of rural areas and adverse weather conditions. Thus, for AV applications focused on those scenarios, alternative datasets may be more appropriate.

The AgroVerse dataset [58], released in October 2019, supports 3D tracking and motion forecasting for autonomous driving, with data collected in Pittsburgh and Miami across different seasons and times. It includes 113 scenes for 3D tracking and over 324,000 five-second scenarios for motion forecasting, using LiDAR, 360° cameras, and 3D bounding box annotations. HD maps offer lane-level geometry and semantic data for map-based learning. AgroVerse2, released in 2021, expanded to six U.S. cities with diverse traffic, road layouts, and weather, adding stereo and ring cameras for a richer multimodal dataset. Despite its strengths, the dataset suffers from class imbalance, limited environmental diversity, and short scenario durations, hindering long-term behavior prediction. The BLVD dataset [59], released in 2019, supports dynamic 4D tracking (3D + time), 5D interactive event recognition (object interactions), and 5D intention prediction. It includes 3D bounding boxes with object IDs for temporal tracking and annotations for 28 predefined interactive events: 13 for vehicles, 8 for pedestrians, and 7 for riders. Future object states, such as location, orientation, and event type. Data was collected in Changshu, China, using multi-view cameras, a Velodyne HDL-64E LiDAR, and GPS/IMU under varied traffic, lighting, and road conditions. The dataset comprises 654 videos (120,000 frames), 250,000 3D boxes, 4,902 tracked objects, 6,004 event fragments, and 4,900 intention predictions. It is well-balanced across different conditions and object types. While segmentation data is not included, BLVD offers rich annotations for various tasks and evaluation metrics, focusing on relevant objects within a 50-meter range to support decision-making. Its dependence on both camera and LiDAR data, as well as its collection in China, should be considered when applying it to broader AV contexts.

KITTI dataset [65] aims to develop novel benchmarks that reflect the complexities of real-world scenarios for use in the development of AVs. The researchers equipped a Volkswagen Passat B6 with one inertial navigation system (OXTIS-RT-3003), one laser scanner (Velodyne-HDL-64E), two grayscale cameras (FL2-14S3M-C), two color cameras (FL2-14S3C-C) and four varifocal lenses (Edmund optics NT59-917). The strength of the KITTI dataset is that it provides data obtained from a variety of sensors and captures real-world driving scenarios, which can be applied to various practical algorithms in autonomous driving research. The downside is that all the data was collected in a single city, which reduces geographic diversity and limit the algorithms' generalizability.

4.1.4. Detection-only Datasets

MSU-4S [44], ONCE [48], Udacity [52], A*3D [54], Comma2K19 [60], and Oxford RobotCar [64] datasets offer annotations solely for object detection tasks, typically in the form of 2D or 3D bounding boxes, without additional support for segmentation or tracking. The MSU-4S (Michigan State University Four Seasons) dataset [44], released in 2024, provides multimodal data for evaluating autonomous driving performance across seasonal and environmental variations. The dataset is collected around the Michigan State campus using a modified 2017 Chevrolet Bolt EV. It contains over 100,000 high-resolution frames from cameras, LiDAR, and Radar, supporting tasks such as 2D/3D object detection, sensor fusion, and domain adaptation. The test vehicle is equipped with three FLIR cameras, two LiDARs (Ouster OS-1 and Velodyne VLP-32), six Continental Radar sensors, and an IMU, capturing robust data under a wide range of conditions, including heavy rain and snow. Annotations include 2D YOLO-format bounding boxes, 3D YAML-format boxes, and detailed metadata (e.g., weather, date, and time), with outputs in JPEG and PCD/YAML formats for easy integration. MSU-4S is particularly strong in representing real-world, seasonally diverse driving scenes suitable for cross-modal learning and sensor redundancy research. Despite its strengths, the dataset has several limitations: it is geographically constrained to a single region, lacks annotations for semantic segmentation and object tracking, and presents compatibility challenges for systems using different sensor configurations. Additionally, some sequences remain only partially annotated, although future updates are anticipated. Overall, it serves as a valuable resource, especially when combined with datasets from other regions.

The One Million Scenes (ONCE) dataset [48], released in 2021, supports 3D object detection and semi/self-supervised learning for autonomous driving. It includes one million LiDAR frames, seven million camera images, and 144 hours of driving data across 200 km^2 of urban and suburban areas in China under various weather and lighting conditions. Sixteen thousand scenes are fully annotated with 2D and 3D bounding boxes for five object classes. Data was collected using a vehicle equipped with a 40-beam LiDAR and seven cameras for full 360° coverage, with annotations provided in three coordinate systems. The dataset offers rich, synchronized multimodal data and extensive unlabeled samples, ideal for sensor fusion and addressing the long-tail problem. However, it is limited to only five annotated object classes. The Udacity dataset [52], released in 2016, was collected in Mountain View, California, during daylight to support autonomous driving tasks like object detection, tracking, and behavioral cloning. It includes around 404,916 training frames, 5,614 test frames, and 10 hours of driving data. The vehicle used a front-facing camera (1920×1200 at 2Hz), LiDAR, GPS, and IMU. Over 65,000 labels across 9,423 frames were annotated and split into training and testing sets. The dataset offers real-world multimodal sensor data for experimentation. However, the dataset lacks key labels (e.g., pedestrians, traffic lights), and being limited to daytime urban scenes with low temporal resolution, has reduced generalizability.

The A3D dataset [54], released in 2019, was designed to offer more diverse driving conditions than existing datasets, covering variations in weather, traffic density, and environments. It includes 39,000 frames with 230,000 3D object annotations for vehicles, cyclists, pedestrians, and blocked objects. Data was collected in Singapore using two PointGrey Chameleon3 cameras and a Velodyne HDL-64ES3 LiDAR, spanning urban, suburban, and diverse road types under various lighting and weather conditions. A3D supports 2D and 3D detection and is valuable for real-world algorithm development. However, its use of costly sensors, Singapore-specific scenes, and limited support for segmentation or tracking should be considered. The comma2k19 dataset [60], released in 2018, was designed to support mapping and localization algorithms using GNSS and commodity sensors. It includes 33 hours (2,019 one-minute segments) of video data collected along a 20 km stretch of California's Highway 280 between San Jose and San Francisco, using a Sony IMX2984 camera. The dataset also provides raw GNSS data, CAN messages, and inertial measurements (gyroscope, accelerometer, magnetometer). It enables tasks like future lane estimation and global pose computation using low-cost sensors. However, it lacks annotations for object detection, segmentation, or tracking, limiting its use to localization-related applications or testing pre-trained models in highway scenarios.

The Oxford RobotCar dataset [64], released in 2015 by the University of Oxford, supports long-term autonomous vehicle research in urban environments. It features over 100 traversals of a 10 km route in central Oxford, UK, recorded across different seasons, times of day, and weather conditions using a Nissan LEAF

EV. The dataset spans 1,000 km and over 20 million frames. It enables research in SLAM, localization, map building, and sensor fusion, equipped with six cameras, 2D/3D LiDAR, GPS/INS, wheel odometry, and IMU. Its strength lies in capturing repeated, real-world environmental variations for long-term autonomy evaluation. While it lacks Radar data and semantic labels and is geographically limited to one route, it remains highly valuable for studying localization robustness, environmental drift, and sensor fusion. Due to its size, significant computational resources are required, but it is well-documented and widely used in the research community.

4.1.5. Tracking-only Datasets

InD [55] and exiD [56] datasets focus exclusively on object tracking, often using trajectory-level annotations from top-down perspectives. The Intersection Drone (InD) dataset [55], collected in Germany in 2020 by RWTH Aachen University, captures detailed trajectories of road users at unsignalized urban intersections using drone-mounted cameras. Data was collected at four intersections in Aachen and Heinsberg, totaling over 50,000 annotated frames with 8,200 vehicles and 5,300 vulnerable road users (pedestrians and cyclists). Recorded at 25 FPS, the drone's bird's-eye view ensures high-accuracy, unobstructed trajectory data ideal for interaction modeling, behavior prediction, and safety validation in autonomous driving. The dataset includes semantic metadata (e.g., movement types, intersection maps) and tools for parsing and visualization. While its strength lies in capturing natural behavior in complex, mixed-traffic environments, limitations include the absence of LiDAR/Radar, its geographic specificity to Germany, a relatively small size, and a lack of support for perception tasks like detection or segmentation. However, InD remains a valuable resource for studying road user interactions and is best used alongside larger, multimodal datasets.

The Extended Interaction (exiD) dataset [56], released in 2020 by RWTH Aachen University, captures trajectory data from drones observing highway driving in Germany. It focuses on behaviors like merging and lane changing across seven challenging 420-meter highway sections. The dataset includes 16 hours of footage, over 30,000 frames, and data on 69,430 road users covering 27,300 km. Drone-based bird's-eye views provide accurate, time-stamped trajectories with annotations, semantic metadata (e.g., vehicle types, lane markings), and HD maps. Data is delivered in CSV format with Python tools for parsing and visualization. exiD excels in clean, high-resolution data for interaction modeling and planning in high-speed environments. However, it lacks LiDAR/Radar data, semantic segmentation, and object classification, and is limited in size and geographic diversity. Despite these limitations, it is a valuable resource for studying real-world highway interactions.

4.2. Roadside Perception Datasets

Roadside perception datasets are acquired from fixed infrastructure such as traffic surveillance cameras, LiDARs mounted on poles, and Radars installed at intersections. These static sensors provide an overhead or third-person view of traffic flow, enabling wide-area situational awareness and continuous monitoring of road segments. Such datasets are particularly valuable for augmenting ego-vehicle data, studying traffic behavior, and developing infrastructure-assisted safety systems. Table 8 presents a detailed overview and comparison of the prominent roadside perception datasets. In the following, we categorize the datasets discussed in Table 8 based on the perception tasks they support (detection, segmentation, and tracking). Specifically, we divide them into two categories: (1) datasets that support detection and tracking, (2) datasets that support detection using both camera and LiDAR sensors, and (3) detection datasets rely only on camera sensor.

4.2.1. Detection and Tracking Datasets

IPS300+ [69], Interaction [73], and CityFlow [74] datasets enable both object detection and temporal tracking, making them ideal for trajectory prediction, multi-agent behavior modeling, and traffic flow analysis from a stationary infrastructure perspective. Data was collected at a single intersection in Beijing using an Intersection Perception Unit (IPU) with an 80-layer LiDAR, two 5.44 MP cameras, and GPS. It contains 14,198 frames with an average of 319.84 labeled objects per frame, making it ideal for dense traffic scene analysis. Annotations include 3D bounding boxes for seven classes (e.g., pedestrian, cyclist, car) in KITTI format. The top-down view offers unobstructed perception, which is impossible from vehicle-mounted sensors, and the fusion of LiDAR and camera data supports sensor fusion research. Limitations include data from only one fixed intersection, a stationary sensor setup, and a 5 Hz annotation rate, which may be insufficient for high-temporal-resolution tasks. Still, IPS300+ is a strong resource for intersection perception in dense urban traffic.

The Interaction dataset [73], released in September 2019, captures over 100 hours of video from 11 locations across the U.S., China, Germany, and Bulgaria using drones and fixed cameras. It includes diverse driving scenarios, such as urban intersections, highway merges, roundabouts, lane changes, and U-turns, as well as

Table 8: Summary and comparative analysis of the Roadside Perception Datasets in autonomous vehicles.

						Sensor			Application			
Ref	**Dataset**	**Year**	**Region**	**Frame**	**Format**	**Camera**	**Lidar**	**Radar**	**Detection**	**Segmentation**	**Tracking**	**Access**
[66]	RoScenes	2024	China	+4.8K	RGB, BEV, GPS	Yes	No	No	✓	✗	✗	Git
[67]	TUMTraf	2023	Germany	+4.8K	RGB, LiDAR	Yes	Yes	No	✓	✗	✗	Web
[68]	A9-Dataset	2022	Germany	+1K	RGB, LiDAR, GPS	Yes	Yes	No	✓	✗	✗	Git
[69]	IPS300+	2022	China	+14K	RGB, LiDAR	Yes	Yes	No	✓	✗	✓	Web
[70]	Rope3D	2022	China	+50K	RGB, LiDAR, GPS	Yes	Yes	No	✓	✗	✗	Web
[71]	LUMPI	2022	Germany	+200K	RGB, LiDAR, GPS	Yes	Yes	No	✓	✗	✗	Web
[72]	WIBAM	2021	UK	+33K	RGB, BEV	Yes	No	No	✓	✗	✗	Git
[73]	Interaction	2019	Global*	+16Hr	Video, Trajectories	Yes	Yes	No	✓	✗	✓	Web
[74]	CityFlow	2019	USA	+229K	MP4, GPS, Trajectories	Yes	No	No	✓	✗	✓	Git

∗ For the reader's convenience, a direct hyperlink to each dataset's original source—whether a website or GitHub repository—is provided in the 'Access' column to enable easy access and download.
∗ '**Global**' indicates that the data was originally collected from different countries including USA, China, Germany, and Bulgaria.

various road users (cars, trucks, buses, bikes, and pedestrians). Each agent is annotated with position, velocity, heading, object type, and interaction type (e.g., cooperative merge, aggressive overtake). The dataset supports behavior prediction, trajectory forecasting, multi-agent motion planning, imitation learning, and reinforcement learning. Its strengths include rare interactive behaviors, international scene diversity, and precise bird's-eye tracking. However, it lacks LiDAR/Radar data, has limited weather and lighting conditions, and requires manual map alignment for some planning frameworks. CityFlow dataset [74], released in 2019, was designed to advance research in multi-agent multi-camera (MTMC) vehicle tracking and re-identification (ReID) in urban environments. It contains 3.25 hours of synchronized HD footage from 40 cameras across 10 intersections in a mid-sized U.S. city, covering up to 2.5km between cameras. The dataset includes 229,680 annotated bounding boxes for 666 unique vehicle identities and provides camera calibration for spatio-temporal analysis. Strengths include realistic traffic scenarios, high-quality annotations, and precise tracking. Limitations include a lack of long-term or rare-event coverage, geographic restriction, and fixed viewpoints that limit algorithm robustness.

4.2.2. Detection-only Datasets (Camera + LiDAR)

TUMTraf [67], A9-Dataset [68], Rope3D [70], and LUMPI [71] datasets support detection tasks using both RGB and LiDAR from roadside perspectives. TUMTraf dataset [67], released in 2023 by the Technical University of Munich, features data from a 7-meter-high gantry at a busy intersection near Munich, Germany. The setup includes two high-resolution RGB cameras and two LiDAR sensors, capturing 4,800 labeled LiDAR frames with 57,406 annotated 3D objects across 10 classes and 273,861 attributes. Manual annotations ensure accurate 3D bounding boxes and tracking. The dataset includes diverse traffic behaviors such as turns, overtaking, and U-turns, making it valuable for studying complex interactions and sensor fusion. Limitations include its single-location scope and fixed sensor setup, which may reduce generalizability and lack dynamic perspectives in vehicle-mounted systems. The A9-Dataset [68], released in 2022, was collected from the A9 highway near Munich, Germany, as part of the Providentia++ project. It supports infrastructure-based research in perception, traffic analysis, and autonomous driving, with applications in 3D object detection, sensor fusion, and traffic prediction. The dataset includes 1,000 sensor frames and 14,000 annotated objects, captured using a multi-sensor setup: four Basler cameras, an Ouster OS1-64 LiDAR, Doppler Radar, and event-based cameras. Each object is manually labeled with 3D bounding boxes, position, orientation, and size. Features include calibrated multi-sensor fusion and a bird's-eye view of highway traffic. However, its scope is limited to a single highway setting with minimal geographic or scene diversity.

The Rope3D dataset [70] uses a unique perspective to advance monocular 3D perception for roadside applications. It was collected in China under varying weather and lighting, and combines data from fixed roadside cameras and LiDAR mounted on parked or moving vehicles. The dataset has 50,000 images and more than 1.5 million labeled objects in 13 classes. Although 2D annotations are more common due to occlusions and distance, the dataset provides detailed joint 2D-3D annotations. Its key strengths are the novel roadside viewpoint and environmental diversity. However, it has limited geographic diversity and lacks dynamic perspectives due to fixed sensor setups. The LUMPI (Leibniz University Multi-Perspective Intersection) dataset [71], released in June

2022, was recorded at a busy intersection in Hanover, Germany, under various weather and lighting conditions. It includes 2D images and 3D point clouds captured from a multi-perspective setup: three surveillance cameras, four ego-perspective LiDARs, and a mobile mapping van with two RIEGL VQ-250 LiDARs and a GNSS/IMU system. Annotations were created through background subtraction, DBSCAN segmentation, tracking via Extended Kalman Filter and ICP, followed by 3D bounding box estimation, classification, and manual refinement. Its strengths lie in diverse viewpoints and environmental conditions for offering robust perception. However, its static sensor setup limits the dynamic perspective typically offered by vehicle-mounted systems.

4.2.3. Detection-only Datasets (Camera-only)

RoScenes [66] and WIBAM [72] datasets rely on monocular or BEV camera views, designed for object detection and scene understanding without additional 3D sensing or tracking. The RoScenes dataset, released in 2024, aims to enhance vision-based Bird's Eye View (BEV) approaches using data from 14 highway scenes in China [66]. It includes 1.21 million images and 21.3 million 3D bounding box labels covering 64,000m^2. The sensor setup features 6–12 roadside cameras mounted on 10m poles, two cameras (different zooms) on each side, and drones flying at 300m to capture full highway coverage without blind spots. Its strengths include a massive scale of data, diverse viewpoints that reduce occlusions, and an efficient BEV-to-3D annotation method. However, the dataset is limited to highway scenarios under sunny conditions, lacking scenes like tunnels or intersections. The WIBAM (Wide Baseline Multiview) dataset [72], released in October 2021, supports weakly supervised training of monocular 3D object detectors using traffic surveillance footage from elevated cameras at UK intersections. It addresses challenges in estimating 3D vehicle poses from elevated views. The dataset includes 33,000 images captured by four unsynchronized RGB cameras (originally 2560×1140, downsampled to 1920×1080), with 116,702 automatically generated 2D training annotations and 1,651 manually labeled 3D bounding boxes in the test set. It enables pose prediction and employs a weak supervision approach using multi-view reprojection loss. Strengths include reduced labeling costs and real-world applicability; however, its fixed elevated viewpoint limits use in autonomous driving, and auto-generated annotations may introduce noise.

4.3. Vehicle-to-Language (V2L) Datasets

V2L datasets or Multimodal-language datasets integrate visual data (images or video) with natural language inputs, including commands, descriptions, or queries. These datasets are designed to be compatible with large language models (LLMs), supporting tasks such as visual question answering, command interpretation, and semantic analysis in AV systems. They enable human-AI interaction and allow autonomous vehicles to interpret, respond, or learn from language-based cues. Table 9 below highlights the key features and differences among notable multimodal-language datasets used in AV research. To meaningfully organize the datasets discussed in Table 9, we categorize them based on their underlying sensor modalities, which play a critical role in determining the complexity, realism, and multimodal learning potential of each dataset. Specifically, we divide the datasets into two groups: (1) those that incorporate both camera and LiDAR sensors, offering rich 3D spatial context and (2) those based on camera-only data, which focus on visual-language reasoning from camera inputs.

4.3.1. Camera + LiDAR Datasets

DriveLMM.o1 [75], V2V-QA [77], MAPLM-QA [82], NuScenes-QA [84], and DriveLM [78] datasets provide both RGB imagery and point cloud data, making them suitable for spatially-grounded language tasks, 3D-aware VQA, and multi-sensor fusion in perception and reasoning. The DriveLMM.o1 dataset [75], released in March 2025 by Mohamed Bin Zayed University of AI, the DriveLMM-o1 dataset serves as a multimodal benchmark to evaluate LLM reasoning in autonomous driving. It includes over 18,000 visual question answering (VQA) samples across four subsets, covering diverse real-world driving scenarios that involve perception, prediction, and planning tasks. Each sample contains synchronized multiview images, LiDAR data, temporal cues, questions, ground-truth answers, and structured step-by-step reasoning. Designed to test LLMs' interpretability and multi-step decision-making, the dataset also provides tools for data parsing and custom evaluation metrics. It supports research in visual reasoning, sensor fusion, and end-to-end autonomous decision-making. Strengths include detailed annotations, dynamic scene diversity, and structured explanations. Limitations include limited geographic and environmental diversity, as well as high computational demands due to data complexity. DriveLMM-o1 sets a new standard for evaluating multimodal LLM reasoning in Avs.

The V2V-QA dataset [77], released in April 2025 by Carnegie Mellon University and NVIDIA, is a large-scale multimodal benchmark for enhancing collaborative autonomous driving via language-based reasoning. Built on real-world V2V4Real and V2X-Real datasets, it includes over 48,000 frames and 1.45 million question-answer

Table 9: Summary and comparative analysis of the Vehicle-to-Language (V2L) Datasets in autonomous vehicles. "QA" denotes Question Answering, "VQA" refers to the Visual Question Answering, "PPR" indicates Perception, Planning, Reasoning, and "CPPR" means Cooperative PPR.

				Data		Sensor		Application				
Ref	**Dataset**	**Year**	**Source**	**#Frame**	**#QA**	**Camera**	**Lidar**	**VQA**	**PPR**	**CPPR**	**Detection**	**Access**
[75]	DriveLMM.o1	2025	NuScenes Dataset	+1.9K	18K	Yes	Yes	✓	✓	✗	✗	Git
[76]	CODA-LM	2025	CODA Dataset	+9K	63K	Yes	No	✓	✓	✗	✓	Git
[77]	V2V-QA	2025	V2V4Real, V2X-Real	+48K	1.45M	Yes	Yes	✓	✓	✓	✗	Git
[78]	DriveLM	2024	NuScenes, Carla	+74K	375K	Yes	Yes	✓	✓	✗	✗	Git
[79]	NuScenes-MQA	2024	NuScenes Dataset	+34K	1.5M	Yes	No	✓	✓	✗	✓	Git
[80]	OmniDrive	2024	NuScenes, OpenLane-v2	+34K	450K	Yes	No	✓	✓	✗	✗	Git
[81]	TOKEN	2024	NuScenes, DriveLM	+28K	434K	Yes	No	✓	✓	✗	✓	N/A
[82]	MAPLM-QA	2024	Original Dataset	+14K	61K	Yes	Yes	✓	✓	✗	✓	Git
[83]	Lingo-QA	2024	Original Dataset	+28K	420K	Yes	No	✓	✓	✗	✓	Git
[84]	NuScenes-QA	2024	NuScenes Dataset	+34K	460K	Yes	Yes	✓	✓	✗	✓	Git
[85]	NuInstruct	2024	NuScenes Dataset	+11K	91K	Yes	No	✓	✓	✓	✓	Git
[86]	DARAMA	2023	Original Dataset	+18K	103K	Yes	No	✓	✓	✗	✓	Web

∗ For the convenience of the reader, each dataset's original source—whether a website or GitHub repository hyperlink—is provided in the 'Access' column to facilitate direct access and download.

(QA) pairs that simulate vehicle-to-vehicle communication. The QA system supports object grounding, identification, and driving planning tasks in multi-agent environments using synchronized LiDAR and camera data. Each frame contains five QA types generated through geometric templates and contextual rules. The dataset enables training and evaluation of vision-language models, like V2V-LLM, which outperform standard LLMs in cooperative perception and intention prediction. Strengths include its focus on explainability, situational awareness, and multi-agent interaction. Limitations include dependency on specific sensor setups, restricted geographic coverage, high computational demands, and uncertain usage permissions. NuScenes-QA dataset released in 2023 by Fudan University [84]. It is a large-scale VQA benchmark built on the nuScenes dataset to enhance reasoning in autonomous driving. It contains 460,000 QA pairs derived from 34,000 driving scenes with synchronized RGB images and LiDAR point clouds. Using 3D scene graphs and logic-based templates adapted from the CLEVR benchmark, researchers generated diverse questions ranging from basic attributes to complex spatiotemporal inference. Temporal questions require multi-frame reasoning to track changes and object motion. The dataset evaluates vision-language models and includes baseline results from Transformer-based models. Although NuScenes-QA advances VQA for driving tasks, limitations include a lack of background scene understanding, limited linguistic diversity due to templated questions, and shallow semantic depth.

The MAPLM-QA dataset [82], introduced in 2024 through a collaboration between Tencent's T Lab and several universities. It is a large-scale multimodal benchmark for vision-language reasoning in autonomous driving. It includes 61,000 QA pairs linked to 14,000 sensor-rich frames featuring panoramic RGB images, 3D LiDAR point clouds, and HD maps from diverse real-world environments like intersections, highways, and rural roads. To overcome limitations of prior VQA datasets, MAPLM-QA challenges LLMs to reason about road layouts, markings, and navigation cues. Data was collected using vehicles equipped with six cameras, LiDAR, and GPS/IMU, with QA annotations refined through a vision model and human verification. The dataset connects language prompts to spatial and environmental features and offers JSON-formatted tools for integration and evaluation. By evaluating LLMs like CLIP and LLaMA, this benchmark highlights their effectiveness in real-world grounding and exposes their limitations in spatial reasoning. Although it is limited by geographic scope and high computational requirements, MAPLM-QA is a pioneering resource for advancing LLM-based perception and planning. The DriveLM dataset [78], released in January 2025 by OpenDriveLab, is a large-scale benchmark for evaluating LLMs and VLMs in autonomous driving via Graph VQA. Unlike standard perception-control datasets, DriveLM emphasizes multi-step reasoning across perception, prediction, planning, and behavior tasks. It includes over 74,000 annotated frames across two subsets: DriveLM-nuScenes (4,871 frames, ∼91.4 QAs/frame) and DriveLM-CARLA (70,006 frames, ∼20.5 QAs/frame). DriveLM-nuScenes uses keyframe-based manual annotation with automatic QA generation, while DriveLM-CARLA relies on expert-driven simulation

and sensor data (RGB + LiDAR). DriveLM's strengths lie in explainability, zero-shot generalization, and structured decision flow. Limitations include reliance on simulation-based data and the high resource demands.

4.3.2. Camera-only Datasets

CODA-LM [76], NuInstruct [85], OmniDrive [80], Lingo-QA [83], and DARAMA [86] datasets rely solely on RGB visual input, often from front-facing or panoramic cameras, and are commonly used in visual-language reasoning, question answering, and instruction-following without explicit 3D spatial data. Dalian University of Technology introduced the CODA-LM dataset [76] in 2024 as a comprehensive benchmark for assessing Large Vision-Language Models (LVLMs) in practical autonomous driving systems. In order to evaluate multimodal understanding, it consists of 9,768 high-resolution scenes with uncommon driving events combined with textual prompts. CODA-LM supports three subtasks: general perception (object/event detection), regional perception (localized object description), and driving suggestion (context-based decision reasoning). It uses a hierarchical evaluation system with automated text-only LLM scoring, which offers better alignment with human judgment than multimodal evaluators. Additionally, it introduces CODA-VLM, a fine-tuned model that outperforms all open-sourced driving LVLMs and achieves GPT-4V-level results. Strengths include its focus on safety-critical corner cases, structured reasoning, and reduced reliance on human evaluation. Limitations include restricted coverage of all rare driving scenarios and high computational demands, which may limit accessibility.

In 2024, HKUST, Sun Yat-Sen University, and Huawei Noah's Ark Lab collaborated to release the NuInstruct dataset [85]. It is a large-scale instruction-based multimodal benchmark for autonomous driving. It contains more than 91,000 instruction-response pairs derived from multiview driving videos, addressing limitations in prior datasets by incorporating spatial-temporal data and diverse perspectives. The dataset is organized into 17 subtasks across four categories, including perception, prediction, risk, and planning, with reasoning to support real-world driving decision-making processes. QA pairs are generated using an SQL-based method and require reasoning across frames and views. Strengths include comprehensive coverage of driving tasks and temporal Bird's-Eye-View cross-view reasoning. However, it lacks 3D object detection and traffic light tasks, requires high computational resources, and may need domain adaptation for generalization. OmniDrive dataset [80], released in April 2025 by Hong Kong Polytechnic University and Beijing Institute of Technology. It is a state-of-the-art multimodal dataset that links 3D driving tasks with LLMs using counterfactual reasoning. Built on nuScenes, it adds 34,000 annotated frames and 434,000 QAs focused on "what if" scenarios that bridge perception, planning, and reasoning. The dataset includes factual and counterfactual QA pairs, trajectory data, and keyframes in JSON format for transformer-based pipelines. Two evaluation agents, Omni-Q (trajectory supervision) and Omni-L (vision-language alignment), show improved scene understanding and planning response. OmniDrive's strengths lie in integrating perception and decision-making while enhancing interpretability and explainability for AVs. Limitations include synthetic scene generation that reduce realism and high computational demands.

The Lingo-QA dataset [83], released in September 2024 by Wayve Technologies, is a large-scale multimodal benchmark for evaluating LLMs in autonomous driving. It includes 28,000 short video scenarios paired with 419,000 QA pairs covering tasks like hazard anticipation, traffic comprehension, and high-level route reasoning. Lingo-QA is built for fine-grained evaluation with QA annotations to evaluate visual perception and decision-making. Key strengths of the Lingo-QA dataset include robust annotation, multi-level reasoning, and standardized assessment. However, its limitations include reliance on short, front-facing video clips with no LiDAR or sensor data, limited scalability (up to 7B LLMs), and Lingo-Judge's dependence on the dataset and lack of support for diverse response formats or preference-based evaluation. DRAMA dataset (Driving Risk Assessment Mechanism with A captioning module) [86], released in 2022 by Honda Research Institute USA, is a multimodal benchmark for developing risk-aware and explainable decision-making in autonomous vehicles. The DRAMA dataset combines video and object-level QA with free-form captions explaining driver reactions, linking risk localization to natural language reasoning. It features synchronized video from two front-facing cameras, along with CAN and IMU data. Annotations include open- and closed-ended questions about risk identification, causes, and reasoning, supported by natural language explanations. DRAMA's key strength lies in integrating multimodal risk perception with interpretability for LVLMs. However, its reliance on subjective human annotations, limited coverage of failure cases, and the risks of deploying incorrect models in safety-critical settings require careful use and additional validation for real-world applications.

4.4. Vehicle-to-Vehicle (V2V) Datasets

V2V datasets capture cooperative communication between vehicles, emphasizing shared perception, intent broadcasting, and collaborative planning. These datasets help address limitations of ego-centric perception by enabling vehicles to "see" beyond their own sensors, reducing latency in reaction times, and supporting safer

maneuvering. Applications include joint trajectory planning, decentralized sensing, and early hazard detection. Table 10 outlines a comparative analysis of important V2V datasets for cooperative driving research. In the following, we categorize the datasets presented in Table 10 according to the data collection environment, distinguishing between simulation-based and real-world V2V datasets. We divide them into two categories: (1) simulation-based datasets, which are generated using platforms such as CARLA or AirSim and allow for controlled, repeatable experiments; and (2) real-world datasets, which are collected from physical vehicle deployments and offer higher environmental realism and sensor noise characteristics.

Table 10: Summary and comparative analysis of the Vehicle-to-Vehicle (V2V) Datasets in autonomous vehicles.

						Sensor			Application			
Ref	**Dataset**	**Year**	**Region**	**Frame**	**Format**	**Camera**	**Lidar**	**Radar**	**Detection**	**Segmentation**	**CP**	**Access**
[87]	Open Mars	2024	USA	+2M	RGB, LiDAR, Trajectories	Yes	Yes	No	✓	✗	✓	Git
[88]	OPV2V-H	2024	Carla	+9K	RGB, LiDAR, Depth	Yes	Yes	No	✓	✗	✓	Git
[89]	OPV2V+	2023	Carla	11,464	RGB, LiDAR, Depth	Yes	Yes	No	✓	✗	✓	Web
[90]	IRV2V	2023	Carla	42.5K*[1]	RGB, LiDAR, BEW	Yes	Yes	No	✓	✗	✓	Git
[91]	V2V4Real	2023	USA	60K	RGB, LiDAR, HDMaps	Yes	Yes	No	✓	✗	✓	Web
[92]	LUCOOP	2023	Germany	+54K	LiDAR, IMU, GNSS	Yes	Yes	No	✓	✗	✓	Web
[93]	CP-UAV	2022	AirSim	131.9K	RGB, LiDAR, BEW	Yes	Yes	No	✓	✗	✓	Git
[94]	OPV2V	2022	Carla	11,464	RGB, LiDAR, BEW	Yes	Yes	No	✓	✗	✓	Web
[95]	CODD	2021	Carla	8,783	LiDAR, Transform.txt	No	Yes	No	✓	✓	✗	Git
[96]	COMAP	2021	Carla	+4.3K	RGB, LiDAR, Seg-Mask	Yes	Yes	No	✓	✓	✓	Git
[97]	T&J Cooper	2019	USA	100	RGB, LiDAR, GPS	Yes	Yes	Yes	✓	✗	✓	N/A

∗ For the convenience of the reader, each dataset's original source—whether a website or GitHub repository hyperlink—is provided in the 'Access' column to facilitate direct access and download.
*[1] IRV2V dataset includes 34K camera images and 8.5K LiDAR frames.

4.4.1. Simulation-based Datasets

OPV2V [94], OPV2V+ [89], OPV2V-H [88], IRV2V [90], CP-UAV [93], CODD [95], and COMAP [96] datasets are generated using various simulators, offering controlled environments with customizable scenarios, ground-truth data, and ease of replication for CP tasks. The OPV2V dataset [94] is a large-scale, open simulated dataset for cooperative Vehicle-to-Vehicle perception and was released in 2021. It contains 70 scenes, 11,464 frames, and 232,913 annotated 3D vehicle bounding boxes. The data was collected from eight towns in CARLA and a digital replica of Culver City, Los Angeles, USA. A key strength of the dataset is that it provides 16 implemented models to evaluate different data fusion strategies. However, its main limitation is that it consists of synthetic data, which may not fully reflect the complexity of real-world environments. The OPV2V+ dataset [89] is the extension of OPV2V, released in 2023 and co-simulated using OpenCDA and CARLA. OpenCDA provides driving scenarios, and CARLA provides maps and weather conditions. It includes 73 Scenes, six road types, and nine cities. It contains 12,000 frames of LiDAR point clouds and RGB camera images, along with 230,000 annotated 3D bounding boxes. The benchmark includes four LiDARs and four fusion strategies, a total of 16 models. The limitation is that it is based on a simulated environment and not reflect real-world scenarios. It also depends on the original OPV2V dataset.

The OPV2V-H dataset [88] is an extension of the OPV2V dataset, which was released in 2024 to address the open heterogeneous collaborative perception problem by introducing the HEterogeneous ALliance (HEAL) framework. HEAL supports the integration of new and heterogeneous agents into a collaborative perception system with minimal retraining. The sensor configuration includes two LiDARs and four cameras per agent. HEAL enhances collaborative performance through real-world scenarios in which vehicles utilize different sensor suites and reduce training costs. The primary constraint of this dataset is that it requires BEV features and depends on the OPV2V dataset. Researchers from Shanghai Jiao Tong University, the University of Southern California, and Shanghai AI Laboratory collaborated to develop the IRegular V2V (IRV2V) dataset in 2023 [90]. It aims to help and promote research on temporal asynchrony for collaborative perception. In the real world, agents send and receive data at different times due to communication delays and interruptions, which can cause problems in collaborative perception. To solve this issue, they proposed CoBEVFLOW, a method that handles asynchronous collaborative messages sent at irregular, continuous time stamps without discretization.

With a bird's eye view (BEV) flow, it avoids additional noise by transporting the original perceptual features. The limitation is that it is a synthetic dataset, and real-world validations are limited to the DAIR-V2X dataset.

The CP-UAV (CoPerception) dataset [93], released in 2022, supports UAV-based collaborative perception using co-simulated environments from AirSim and CARLA. It includes 131,900 synchronized images collected by five UAVs flying at three altitudes in three virtual towns. Each UAV is equipped with five RGB cameras, and the dataset provides pixel-wise semantic segmentation, 2D/3D bounding boxes, Bird's Eye View maps, and occlusion labels. Strengths include multi-agent perspectives and comprehensive annotations. However, being fully simulated, it may not capture real-world complexities and is limited to RGB data without additional sensors like LiDAR. CODD (Cooperative Driving Dataset) [95] is an open-source synthetic dataset released in 2021 and created using CARLA. The dataset contains 108 sequences used for training, validation, and testing. It is used in cooperative 3D object detection, cooperative object tracking, and multi-agent SLAM. CODD's strength is in capturing interactions between multiple agents. Using CARLA allows for generating diverse and controlled scenarios and facilitates reproducibility and scalability, but may not reflect real-world environments. It focuses only on LIDAR, lacking sensor modalities such as cameras or Radars. COMAP (COllective Multi-Agent Perception) [96] is a synthetic dataset generated using the CARLA and SUMO simulators, released in 2021. It supports training and testing of 3D object and BEV detectors under varying vehicle-to-vehicle communication ranges. CARLA handles sensor data generation, while SUMO manages traffic flow and vehicle navigation using maps transformed from CARLA. The dataset allows exploration of overlapping and non-overlapping field-of-view scenarios in a controlled environment. However, it lacks real-world sensor noise and environmental complexity, and is limited to LiDAR input, without camera or Radar data.

4.4.2. Real-World Datasets

Open Mars [87], V2V4Real [91], LUCOOP [92], and T&J Cooper [97] datasets are collected from physical vehicle deployments in real-world settings, making them valuable for evaluating the robustness and scalability of autonomous driving systems. The Open MARS Dataset [87] is a large-scale multimodal dataset focused on real-world multiagent and multitraversal autonomous driving. It was collected in a 20 km^2 area in Ann Arbor, Michigan, from October 2023 to March 2024, which includes urban, highway, residential, and campus environments. It comprises two subsets: Multitraversal (5,757 traversals across 66 locations) and Multiagent (53 scenes involving 2–3 vehicles), totaling over 1.4 million frames and 15,000 multiagent image/LiDAR frames. The dataset uses a rich sensor suite: 128-channel LiDAR, IMU, GPS, three narrow-angle cameras, and three wide-angle fisheye cameras, all synchronized at 10 Hz. It provides full 3D ego-pose estimations and geometric calibration, with a structure compatible with nuScenes tools. Strengths include high-resolution, well-calibrated multimodal data and close-proximity multiagent interaction. However, it's geographically limited to a single city, lacks semantic labels for perception tasks, and is limited to a specific sensor setup, which affect generalizability.

The V2V4Real dataset [91] was released in 2023 to facilitate research on V2V cooperative perception through a collaboration between UCLA and five other institutions. The data was collected by two vehicles equipped with multimodal sensors driving together through diverse scenarios. The dataset covers 410 km and comprises 20,000 LiDAR frames, 40,000 RGB frames, 240,000 annotated 3D bounding boxes across five classes, and high-definition maps (HDMaps). V2V4Real supports cooperative 3D object detection, cooperative 3D object tracking, and sim-to-real domain adaptation. It offers real-world, multimodal data but is limited to only two vehicles and may not fully capture vehicle interactions in complex traffic scenarios. The LUCOOP (Leibniz University Cooperative Perception and Urban Navigation) dataset [92] is a time-synchronized, multimodal dataset collected by three interacting measurement vehicles. It was released in 2023 and includes approximately 54,000 LiDAR frames, around 700,000 IMU measurements, and more than 2.5 hours of 10 Hz GNSS raw data. The dataset provides 3D bounding box annotations for evaluating object detection approaches, as well as highly accurate ground-truth poses for each vehicle throughout the measurement campaign.

The Cooper (Cooperative Perception for Connected Autonomous Vehicles) system was introduced in 2019 [97]. It supports the fusion and aggregation of 3D point clouds from multiple vehicles. Each vehicle collects raw LiDAR data independently and transmits it to nearby vehicles, after which a data fusion algorithm combines the information. The authors also introduced the T&J dataset, which demonstrates higher performance for vehicle collaboration than the KITTI dataset. Their sensor setup includes cameras, GPS sensors, Radars, and LiDARs. Cooper is one of the pioneering works in cooperative perception.

4.5. Vehicle-to-Infrastructure (V2I) Datasets

V2I datasets focus on data exchange between AVs and stationary infrastructure components, such as traffic signals, roadside units (RSUs), and cameras. These interactions provide vehicles with vital contextual data

like signal phase and timing (SPaT), road conditions, or environmental alerts. V2I communication enhances decision-making and supports intelligent transportation services like adaptive signal control and emergency vehicle prioritization. Table 11 summarizes the core attributes and comparative insights of major V2I datasets. To meaningfully organize the datasets discussed in Table 11, we categorize them based on their simulation-based or real-world origin. We divide the datasets into two groups: (1) simulation-based datasets, which are generated in virtual environments; and (2) real-world datasets, which are captured from physical deployments across roads and intersections.

Table 11: Summary and comparative analysis of the Vehicle-to-Infrastructure (V2I) Datasets in autonomous vehicles.

				Frame			Sensor			Application		
Ref	**Dataset**	**Year**	**Region**	**Image**	**LiDAR**	**Format**	**Camera**	**Lidar**	**Radar**	**Detection**	**Perception**	**Access**
[98]	V2I-HD	2025	China	15,254	0	RGB, HD Maps	Yes	No	No	✓	✓	Drive
[99]	V2X-DSI	2024	Carla	14K	57K	RGB, LiDAR	Yes	Yes	No	✓	✓	N/A
[100]	V2X-Radar*[1]	2024	China	40K	20K	RGB, LiDAR, BBox	Yes	Yes	Yes	✓	✓	Git
[101]	TUMTraf-V2X	2024	Germany	5K	2K	RGB, LiDAR, BBox	Yes	Yes	No	✓	✓	Git
[102]	OTVIC	2024	China	15,045	15,045	RGB, LiDAR, GPS	Yes	Yes	No	✓	✓	N/A
[103]	DAIR-V2XReid*[2]	2024	China	2,556	0	RGB, Vehicle-ID	Yes	No	No	✗	✓	Git
[104]	HoloVIC	2024	China	+100K	+100K	RGB, LiDAR, BBox	Yes	Yes	No	✓	✓	Web
[105]	V2X-Seq	2023	China	15K	15K	RGB, BBox, Trajectory	Yes	Yes	No	✓	✓	Git
[106]	CARTI	2022	Carla	0	11K	LiDAR	No	Yes	No	✓	✓	Git
[107]	DAIR-V2X	2022	China	71,254	71,254	RGB, LiDAR, BBox	Yes	Yes	No	✓	✓	Git
[108]	COOPER	2020	Carla	5K	5K	RGB, LiDAR, Depth	Yes	Yes	No	✓	✓	Git

∗ For the convenience of the reader, each dataset's original source—whether a website or GitHub repository hyperlink—is provided in the 'Access' column to facilitate direct access and download.
*[1] V2X-DSI dataset additionally comprises 20K 4D-Radar data.
*[2] DAIR-V2XReid dataset is typically used for vehicle matching and re-identification tasks.

4.5.1. Simulation-based Datasets

V2X-DSI [99], CARTI [106], and COOPER [108] datasets are generated in synthetic environments, allowing for precise ground-truth annotations and scenario control. The V2X-DSI dataset [99] was released in 2024 using the CARLA simulator. It contains 56,984 frames in 57 diverse urban scenarios. It evaluates LiDAR sensors with 16, 32, 64, and 128 beam densities on CP scenarios to determine cost-effective sensor configurations. The strength is that it addresses the practical concern of the high cost of dense beam LiDAR sensors. It also provides benchmarking across multiple LiDAR configurations, which offers valuable insights for researchers. The simulated data may not reflect the complexity of the real world. The CARTI dataset [106] is a cooperative dataset developed using CARLA to evaluate a deep-fusion algorithm called Grid-wise Fusion (GFF). It was introduced along with PillarGrid, a feature-level cooperative 3D object detection method, in 2022. The sensor setup includes LiDAR mounted on the vehicle and roadside LiDAR. The COOPER dataset [108] was collected from simulated urban road scenarios using CARLA and was released in 2020. The sensor setup included LiDAR and a camera on the vehicle and LiDAR on the roadside infrastructure. The authors used data to evaluate a cooperative perception system for 3D object detection using early and late fusion. However, using synthetic data limits real-world scenarios. Additionally, there is no radar in the sensor setup.

4.5.2. Real-World Datasets

V2I-HD [98], V2X-Radar [100], TUMTraf-V2X [101], OTVIC [102], DAIR-V2XReID [103], HoloVIC [104], V2X-Seq [105], and DAIR-V2X [107] datasets are collected from physical roadside units and vehicles in real-world urban and highway environments. The V2I-HD dataset [98], released in 2025, is a real-world dataset and benchmark to help the development of online HD mapping for Vehicle-Infrastructure Cooperative Autonomous Driving (VICAD) by promoting vision-centric V2I systems. The dataset includes collaborative and synchronized camera frames from vehicles and roadside infrastructures, as well as HD map elements annotated by humans. The strengths of this dataset are real-world data and the utilization of camera data, which is more cost-efficient than LiDAR. However, relying solely on cameras and excluding other sensors limits its applicability.

The V2X-Radar dataset [100] was released in 2024 and was collected using a connected vehicle platform and a roadside unit equipped with 4D Radar, LiDAR, and multi-view cameras. It contains 20,000 LiDAR frames, 40,000 camera images, 20,000 4D Radar samples, and 350,000 annotated boxes across five categories: vehicles, buses, trucks, pedestrians, and bicycles. The strength of this dataset is that it integrates 4D radar to enhance perception in challenging weather conditions, such as sunny, rainy, dusk, and nighttime. The data is collected from real-world driving scenarios. However, its geographical scope is limited, and the dataset may lack exposure to rare events due to its collection from only a 15-hour driving log. The TUMTraf-V2X dataset [101] was collected in Germany, and it consists of 2,000 annotated point clouds, 5,000 annotated images, and more than 3,000 3D bounding boxes. The annotations cover eight specific classes and represent complex driving scenarios. The sensor setup for the infrastructure includes one LiDAR and four cameras, while the vehicle consists of a LiDAR, a camera, a GPS, and an IMU sensor. The strength lies in offering data from both vehicle and roadside viewpoints. It also covers complex traffic conditions. However, the dataset does not have geographic diversity.

The OTVIC dataset [102] is a multimodal, multi-view dataset featuring online transmission from real-world scenes for vehicle-to-infrastructure cooperative 3D object detection. The data was collected from a highway in China and released in 2024. The dataset includes 14,045 frames and 24,452 manually annotated vehicles. OTVIC provides a realistic benchmark but has limitations in terms of geographic coverage and sensor modalities. The HoloVIC dataset [104] was released in 2024 and contains 100,000 synchronized data frames, as well as 11.47 million annotated 3D bounding boxes. The supported tasks are monocular 3D detection, LiDAR 3D detection, multi-object detection, and multi-sensor multi-object tracking. The strengths of this dataset are its diverse sensor layout, comprehensive 3D annotations, and real-world driving scenarios. However, it is geographically limited and focuses only on short-term interactions.

The V2X-Seq dataset [105] was collected from 28 intersections with 672 hours of data and was released in 2023. The dataset contains two parts: the sequential perception dataset, which includes 15,000 frames from 95 scenarios, and the trajectory forecasting dataset, which contains 80,000 vehicle-view scenarios, 80,000 infrastructure-view scenarios, and 50,000 cooperative-view scenarios. The dataset comprises real-world sequential data. However, geographical diversity and sensor modality are limited, which can affect the generalizability and accuracy of the algorithms based on these data. The DAIR-V2X dataset [107] was introduced in 2022 to support research in autonomous driving systems. This dataset comprises 71,254 LiDAR frames, 71,254 camera frames, and 3D annotations. The data was collected from 28 intersections in Beijing, China. The strengths of the dataset are that it contains real-world data and facilitates sensor fusion between the infrastructure and the vehicle. However, the data was collected in one city, lacking scenario diversity. The data was also collected using only LiDARs, so it is limited in viewpoint diversity.

4.6. Vehicle-to-Everything (V2X) Datasets

V2X datasets provide a comprehensive view of inter-connected communication between vehicles, infrastructure, pedestrians, and cloud services. This category encompasses datasets that facilitate both V2I and V2V communications. These datasets enable AVs to function cohesively within complex urban ecosystems, addressing scenarios involving vulnerable road users, real-time environmental updates, and coordinated traffic operations. V2X datasets are pivotal for building fully cooperative, connected, and automated mobility systems. Table 12 presents a structured comparison of leading V2X datasets and their application domains. In the following, we categorize the datasets presented in Table 12 according to their simulation-based or real-world origin. We divide them into two categories: (1) simulation-based datasets, which are generated using simulation platforms; and (2) real-world datasets, which are collected from physical vehicle deployments.

4.6.1. Simulation-based Datasets

DeepAccident [111], Adver-City [112], Multi-V2X [113], WHALES [114], SCOPE [117], OPV2V-N [118], V2X-Sim [119], V2X-ViT [120], and DOLPHINS [121] datasets are created using simulation platforms, offering precise ground-truth labels, synchronized multi-agent data, and diverse cooperative scenarios. DeepAccident dataset [111], released in 2024 by HKU, Huawei, and Dalian University of Technology, is a safety-focused synthetic dataset for accident prediction in V2X environments. It was generated by CARLA and features 57,000 frames and 285,000 samples of multi-agent interactions in 12 major intersection accident types. Each scene includes data from RGB cameras and LiDAR sensors on four vehicles and one roadside unit. Tasks include 3D detection, tracking, BEV segmentation, motion forecasting, and full-scene accident prediction. While highly structured and valuable for risk-aware learning, it lacks real-world sensor noise and requires heavy computation.

The Adver-City dataset [112], released in March 2025 by Queen's University, is a large-scale synthetic dataset

Table 12: Summary and comparative analysis of the Vehicle-to-Everything (V2X) Datasets in autonomous vehicles.

				Frame			Sensor			Application		
Ref	**Dataset**	**Year**	**Region**	**Image**	**LiDAR**	**Format**	**Camera**	**Lidar**	**Radar**	**Detection**	**Segmentation**	**Access**
[109]	Mixed-Signals	2025	Australia	0	45.1K	LiDAR, BBox	No	Yes	No	✓	✗	Web
[110]	V2X-ReaLO	2025	China	25,028	25,028	RGB, LiDAR, ROS-bags	Yes	Yes	No	✓	✗	N/A
[111]	DeepAccident	2024	Carla	57K	57K	RGB, LiDAR, BEW	Yes	Yes	No	✓	✓	Git
[112]	Adver-City	2024	Carla	24,087	24,087	RGB, LiDAR, IMU/GNSS	Yes	Yes	No	✓	✓	Web
[113]	Multi-V2X	2024	Carla	549K	146K	RGB, LiDAR, BBox	Yes	Yes	No	✓	✗	Git
[114]	WHALES	2024	Carla	70K	17K	RGB, LiDAR, Trajectory	Yes	Yes	No	✓	✗	Git
[115]	V2X-R	2024	China	150,908	37,727	RGB, LiDAR, BBox	Yes	Yes	Yes	✓	✗	Git
[116]	V2XPnP	2024	USA	208K	40K	RGB, BBox, Trajectory	Yes	Yes	Yes	✓	✗	Web
[117]	SCOPE	2024	Carla	17,600	17,600	RGB, LiDAR, BBox	Yes	Yes	No	✓	✓	Git
[118]	OPV2V-N	2024	Carla	11,464	11,464	RGB, LiDAR, BBox	Yes	Yes	No	✓	✗	N/A
[119]	V2X-Sim	2022	Carla	10K	10K	RGB, LiDAR, GPS, Depth	Yes	Yes	Yes	✓	✓	Git
[120]	V2X-ViT	2022	Carla	11,447	11,447	RGB, LiDAR, BBox	Yes	Yes	No	✓	✗	Git
[121]	DOLPHINS	2022	Carla	42,376	42,376	RGB, LiDAR	Yes	Yes	No	✓	✗	Web
[122]	V2X-Real	2022	USA	171K	34K	RGB, LiDAR, BBox	Yes	Yes	No	✓	✗	Web

∗ For the convenience of the reader, each dataset's original source—whether a website or GitHub repository hyperlink—is provided in the 'Access' column to facilitate direct access and download.

for evaluating cooperative perception in adverse weather. Built using CARLA and OpenCDA, it includes 24,000 frames and 890,000 annotations across 110 scenarios featuring rain, fog, glare, and varying lighting and traffic conditions. Inspired by U.S. National Highway Traffic Safety Administration (NHTSA) crash data, it focuses on dangerous driving areas like intersections and rural curves. Data comes from five agents (ego, RSUs, and vehicles) equipped with RGB, segmentation cameras, LiDAR, and GNSS/IMU. It supports tasks like detection, tracking, and segmentation, with a label format matching OPV2V. Strengths include diverse agents, weather types (including glare), and realistic layouts. Limitations include a lack of real-world sensor data. Multi-V2X [113] is a large-scale synthetic cooperative perception benchmark released in 2024 by Tsinghua University. It is the first dataset specifically designed to assess autonomous driving performance under varying CAV (Connected and Autonomous Vehicle) penetration rates. Built using CARLA and SUMO, it includes 549,000 RGB images, 146,000 LiDAR frames, and 4.2 million annotated 3D bounding boxes across six object classes. The dataset simulates both V2V and V2I scenarios across urban, suburban, and rural settings. Vehicles and RSUs are equipped with RGB cameras, LiDARs, and GNSS modules. The six included towns feature diverse road topologies and traffic conditions. Multi-V2X supports 3D object detection, tracking, and benchmarking using fusion strategies (early, late, intermediate). Meanwhile, its synthetic nature limits realism and weather variation.

WHALES (Wireless enHanced Autonomous vehicles with Large number of Engaged agentS) [114] is a large-scale synthetic dataset released in 2024 by Tsinghua University. It is generated using CARLA, and features 70,000 RGB images, 17,000 LiDAR frames, and over 2 million annotated 3D bounding boxes, with an average of 8.4 agents per scene. WHALES supports tasks like object detection and CP benchmarking under realistic communication constraints. While it offers rich multimodal annotations and novel scheduling tasks, its synthetic nature limits realism, and its high computational demands may restrict accessibility. SCOPE (Synthetic COllective PErception) [117] is a 2024 synthetic multi-agent dataset from the University of Tübingen and FZI Karlsruhe, designed to advance collective perception in AVs. It includes over 17,600 frames across 40+ scenarios with up to 24 agents, featuring realistic sensor models (RGB, solid-state LiDAR), detailed weather simulation, and urban environments. SCOPE offers annotations for object detection and semantic segmentation, supporting research on perception resilience, sensor alignment, and domain adaptation. Its strengths include high-fidelity environmental modeling and diverse road user types. Limitations include simulation-based constraints on real-world unpredictability and the high computational cost of processing multimodal data.

V2X-Sim [119] is a large-scale synthetic dataset released in 2022 by NYU, USC, and Shanghai Jiao Tong University. Built using CARLA and SUMO, it simulates V2X scenarios with synchronized multi-agent sensor data from LiDAR, RGB, semantic, and depth cameras on both vehicles and infrastructure. V2X-Sim's strengths lie in its early contribution to multimodal, multi-agent CP benchmarks and its compatibility with open-source

tools. However, its simulated nature limits realism in terms of sensor noise and unpredictable driving behavior. DOLPHINS (Dataset for cOLlaborative Perception enabled Harmonious and INterconnected Self-driving) [121] is a large synthetic benchmark released in 2022 by Tsinghua University to support both V2V and V2I collaborative perception. It features synchronized full-HD images and 64-line LiDAR point clouds collected from ego vehicles, auxiliary vehicles, and RSUs across six diverse scenarios, including urban, highway, mountain, and rural settings under weather conditions like rain and fog. While DOLPHINS excels in scope and detail, it is limited by its synthetic nature, low frame rate for high-speed scenes, and lack of radar data.

4.6.2. Real-World Datasets

Mixed-Signals [109], V2X-ReaLO [110], V2X-R [115], V2XPnP [116], and V2X-Real [122] datasets are collected from physical vehicle deployments in real-world settings. Mixed-Signals [109] is a real-world V2X dataset released in 2025 by Cornell and the University of Sydney. It features data from three vehicles and one RSU with different LiDAR setups, as well as 45,100 point clouds and 240,600 3D bounding boxes. Its main strengths are showcasing sensor heterogeneity and real-world multi-agent perception. Limitations include a single location, short recording duration, and lack of camera or radar data. V2X-ReaLO [110] is a real-world dataset released in March 2025 by UCLA. Built on V2X-Real, it offers 25,028 frames (6,850 annotated keyframes) from two connected vehicles and two RSUs, supporting real-time evaluation of early, late, and intermediate fusion strategies. Its key strength lies in practical and real-time V2X fusion under real-world conditions, with open-source support for modular experimentation. Limitations include a lack of radar data and limited geographic diversity.

V2X-R [115] is a multimodal synthetic dataset released in March 2025 to advance robust autonomous driving in adverse weather. It includes 12,079 driving scenarios with 37,727 LiDAR and radar point clouds, 150,908 RGB images, and over 170,000 3D bounding boxes. Additionally, it enables enhanced 3D object detection under fog and snow. Despite being synthetic and resource-intensive, V2X-R offers a critical platform for developing weather-resilient, multi-agent, multimodal autonomous perception systems. V2XPnP [116] is a large-scale real-world benchmark introduced by University of California researchers in December 2024. It includes around 40,000 frames across 100 vehicle-centric and 63 infrastructure-centric scenarios. Its advantages is combining temporal consistency with full V2X coverage, which is ideal for studying real-time multi-agent collaboration. However, it lacks detailed sensor specifications and demands high computational resources.

4.7. Infrastructure-to-Infrastructure (I2I) Datasets

I2I datasets encompass coordinated sensing and data sharing among multiple infrastructure nodes, such as inter-linked traffic cameras or RSUs. This communication paradigm supports large-scale traffic management, surveillance handoffs, and synchronized control across urban regions.I2I datasets help connect and coordinate multiple sensor systems that are spread out across different infrastructure components. This integration enhances redundancy, resilience, and scalability in smart city deployments. Table 13 provides an in-depth comparison of I2I datasets, highlighting their scope and infrastructure coordination capabilities.

The RCOOPER dataset [123] was collected from simulated urban scenarios using CARLA and was released in 2020. The sensor setup included LiDAR, a camera on the vehicle, and LiDAR on the roadside infrastructure. The authors used the data to evaluate a cooperative perception system for 3D object detection using early and late fusion. It is one of the premier studies on cooperative perception. However, using synthetic data limits real-world applicability. There is no radar in the sensor setup. The InScope dataset [124], released in 2024, was developed to address occlusion challenges in autonomous driving. It includes 187,787 annotated 3D bounding boxes and 303 tracking trajectories. The dataset supports four benchmark tasks: collaborative 3D object detection, multi-source fusion, domain adaptation, and multi-object tracking. Its ability to mitigate occlusion through infrastructure-side sensing is one of its main advantages; however, its drawbacks include limited geographic diversity and access requirements.

5. Object Detection in Autonomous Vehicles

Object detection plays a pivotal role in AV perception systems by enabling real-time understanding of the driving environment. Accurate detection of surrounding vehicles, pedestrians, cyclists, and static obstacles is essential for safe navigation, decision-making, and Path planning. Over the past decade, significant advances have been made in object detection techniques, driven by developments in sensor technology, deep learning, transformer-based architectures, and more recently, methods based on Large Language Models (LLMs) and Vision-Language Models (VLMs). In AV systems, object detection pipelines are typically built upon a variety

Table 13: Summary and comparative analysis of the Infrastructure-to-Infrastructure (I2I) Datasets in autonomous vehicles.

				Frame			Sensor			Application		
Ref	Dataset	Year	Region	Image	LiDAR	Format	Camera	Lidar	Radar	Detection	Perception	Access
[123]	RCooper	2024	China	50K	30K	RGB, LiDAR, BBox	Yes	Yes	No	✓	✗	Git
[124]	InScope	2024	China	0	21,317	LiDAR, BBox, Trajectory	No	Yes	No	✓	✗	Git

∗ For the convenience of the reader, each dataset's original source—whether a website or GitHub repository hyperlink—is provided in the 'Access' column to facilitate direct access and download.

of sensor modalities, such as RGB cameras, LiDAR, and radar, with details discussed in Section 3. Building on this foundation, modern object detection approaches can be broadly categorized based on the primary input modality or through the use of multi-modal fusion strategies.

Generally, object detection methods in AVs can be broadly categorized into one of the following four types: 2D camera-based detection, 3D LiDAR-based detection, 2D–3D fusion detection, and LLM/VLM-based detection. Figure 5 provides a comprehensive overview of these detection approaches along with their underlying sensor modalities, learning paradigms, and model architectures.

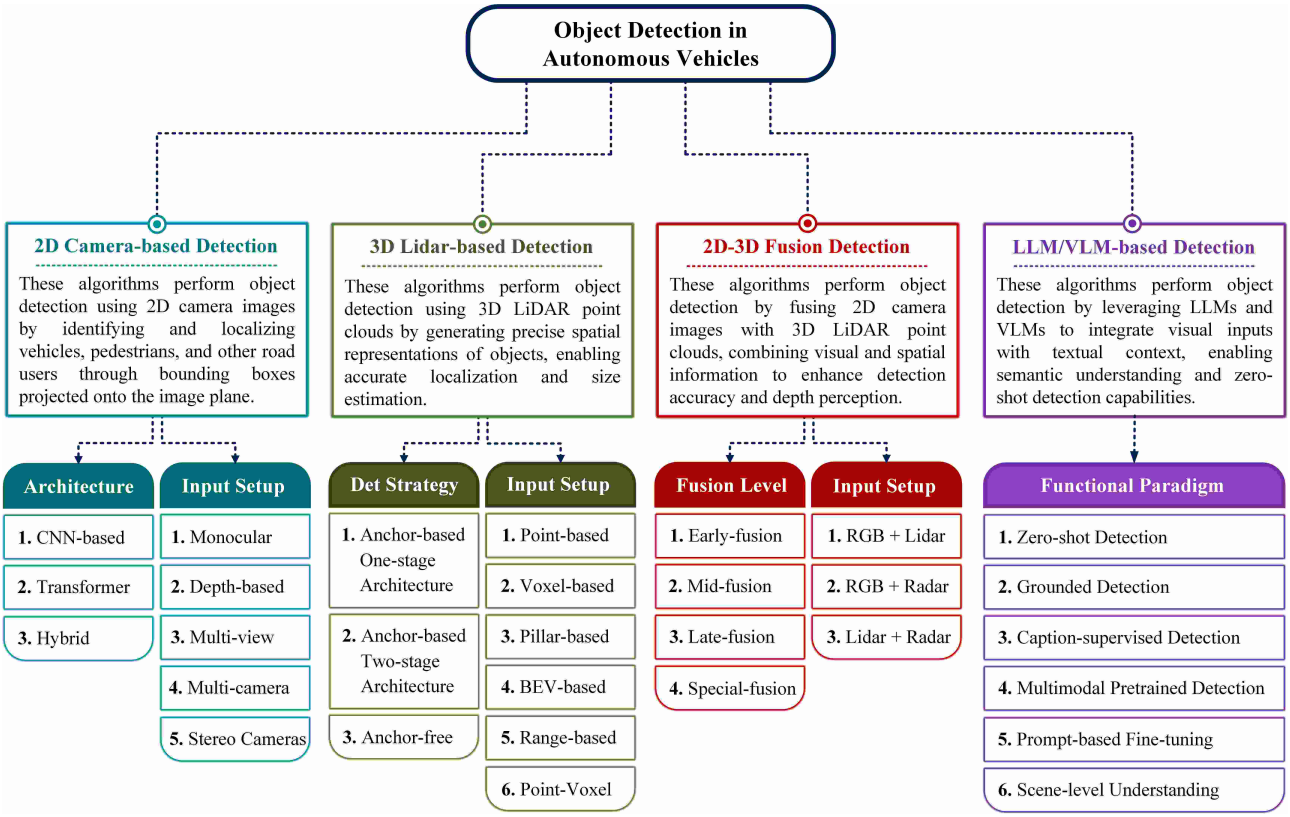

Figure 5: A comprehensive taxonomy of object detection methods in AVs, categorized into four primary types. Each category includes representative subtypes based on sensor configuration, data representation, fusion strategy, and model architecture.

In 2D camera-based detection, algorithms rely on RGB imagery to identify and localize objects using bounding boxes projected onto the image plane [125]. These methods can be further classified by their underlying architecture into three main groups: CNN-based (e.g., YOLOv4, Faster R-CNN), Transformer-based (e.g., DETR, Deformable DETR), and Hybrid models that integrate convolutional backbones with transformer modules (e.g., Swin Transformer + Faster R-CNN). At the same time, 2D detection approaches can also be categorized by their input configuration, including monocular (single-view RGB), depth-based monocular, multi-view geometry-based, multi-camera system-level, and stereo camera setups, each tailored to specific perception requirements and environmental conditions. In contrast, 3D LiDAR-based detection methods operate on raw or preprocessed point clouds to generate precise spatial representations of objects. These techniques encompass different detection strategies, including anchor-based one-stage detectors (e.g., SECOND, YOLO3D), anchor-based two-stage detectors (e.g., PV-RCNN, Voxel R-CNN), and anchor-free detectors (e.g., 3DSSD, YOLOv8-3D),

each offering trade-offs between speed, localization accuracy, and robustness to sparse data. In parallel, 3D detection approaches can also be categorized by their input setup, including point-based, voxel-based, pillar-based, BEV-based, point-voxel-based, and range-based. 2D–3D fusion detection combines visual and geometric information from both cameras and LiDAR sensors to enhance object recognition under challenging conditions. Fusion strategies are typically classified as early-fusion, mid-fusion, late-fusion, or special-fusion (not strictly fitting into the previous three). Additionally, they can be categorized based on their input modalities to three different categories. Lastly, LLM/VLM-based detection approaches utilize large language and vision-language models to incorporate textual context into visual reasoning, enabling semantic-level understanding and generalization across unseen categories. These techniques include zero-shot detection, grounded detection (region-text alignment), caption-supervised Detection (language-augmented supervision), multimodal pretrained detection (Joint vision-language models), prompt-based fine-tuning (domain-adapted queries), and scene-level understanding. This categorization provides a unique framework and valuable insight into the current landscape of object detection in AVs, guiding future research toward more robust and generalizable perception systems [126].

The following subsections highlight recent advancements across diverse sensor modalities and model architectures. Subsection 5.1 reviews 2D camera-based object detection methods, while Subsection 5.2 explores 3D object detection techniques using LiDAR (point-cloud). Subsection 5.3 examines 2D–3D fusion strategies that combine camera (image) and LiDAR (point-cloud) data, categorized by early, mid, and late fusion stages. Finally, Subsection 5.4 focuses on emerging LLMs/VLMs-based object detection methods that enable open-vocabulary and prompt-driven detection through multimodal understanding.

5.1. 2D Camera-based Approaches

Identifying and localizing key objects in the driving scene, including vehicles, pedestrians, and cyclists, is essential for the safe operation of autonomous driving systems. In the context of 2D camera-based object detection, existing algorithms can generally be classified into three categories based on their architectures: CNN-based, Transformer-based, and hybrid approaches that integrate both CNN and Transformer components. As illustrated in Figure 6, these approaches follow an overall framework where camera-captured RGB images are processed through feature extraction, region proposal, and object classification/localization stages. In CNN-based models, convolutional layers learn hierarchical spatial features, gradually refining them for detection tasks. In contrast, Transformer-based models leverage self-attention mechanisms to capture long-range dependencies and global context across the image. Hybrid approaches aim to combine the localized feature learning of CNNs with the reasoning capability of Transformers, enhancing detection accuracy and robustness in driving scenes.

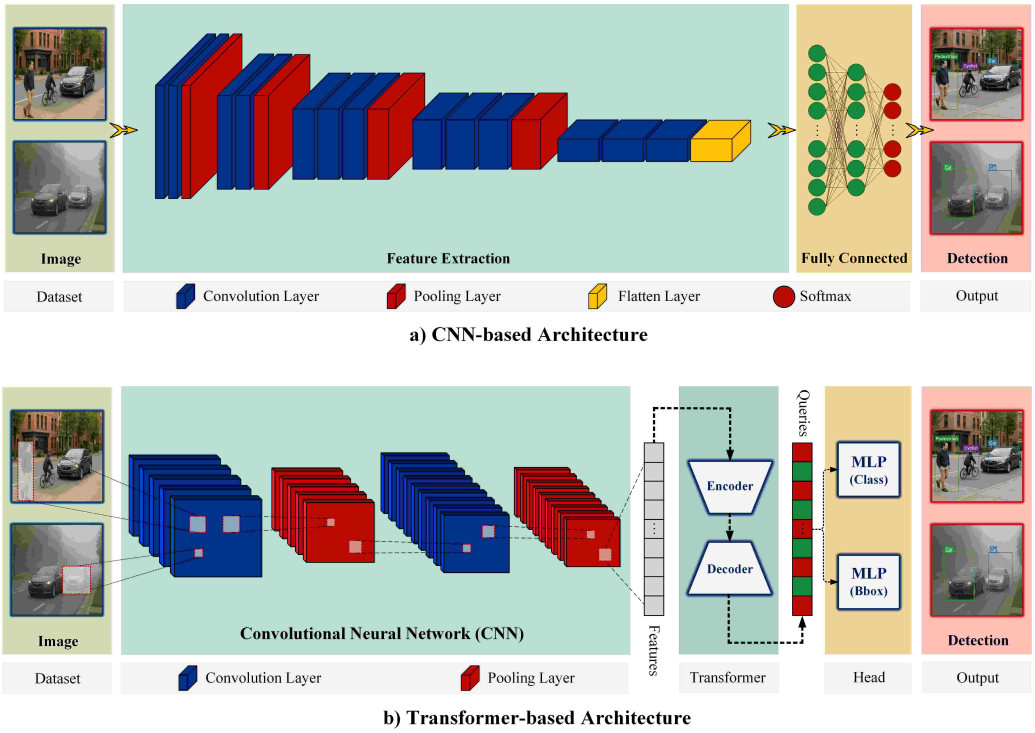

Figure 6: Overall framework of the 2D camera-based approaches in the context of autonomous driving systems. **a)** general architecture of CNN-based models, **b)** general architecture of Transformer-based models.

This section presents all existing 2D camera-based algorithms for object detection in autonomous vehicles from 2016 to 2025. To ensure a consistent and fair comparison, we organize the results into four separate tables, as these methods report performance on different datasets and use varying evaluation metrics. Table 14 provides a comparative analysis of 2D object detection methods using camera-based data from the KITTI dataset, where both the input and output are in 2D (RGB images to 2D bounding boxes). Table 15 focuses on 3D object detection methods that use camera-based KITTI data, but generate 3D bounding boxes from 2D image inputs. Table 16 presents a comparative analysis of 2D/3D object detection methods using camera-based data from the NuScenes dataset, where the input is 2D images and the output is either 2D or 3D bounding boxes. Finally, Table 17 summarizes YOLO-based methods evaluated on the camera-based KITTI dataset for 2D object detection. This structured categorization enables consistent benchmarking and highlights performance trends across architectures, datasets, and detection tasks. It should be noted that the values highlighted in red represent the best performance achieved for each evaluation metric across all listed methods.

Table 14: Comparative analysis of 2D object detection methods using the camera-based KITTI dataset.

					Car			Pedestrian			Cyclist		
Ref	**Method**	**Year**	**Backbone**	**Modality**	**Easy**	**Medium**	**Hard**	**Easy**	**Medium**	**Hard**	**Easy**	**Medium**	**Hard**
[127]	AspectNet	2022	ResNet-50	Mono	96.16	**93.87**	88.17	87.17	76.57	68.42	86.15	77.11	71.24
	Innovation:	Adds an aspect ratio prediction head with improved sampling and loss to handle imbalance.											
[128]	Gaussian YOLOv3	2019	Darknet-53	Mono	90.61	90.20	81.19	87.84	79.57	72.30	89.31	81.30	80.20
	Innovation:	Models bounding boxes as Gaussians with predicted localization uncertainty.											
[128]	Gaussian YOLOv3†	2019	Darknet-53	Mono	98.74	90.48	**89.47**	87.85	**79.96**	**76.81**	90.08	86.59	**81.09**
	Innovation:	Enhanced NMS and loss weighting for Gaussian outputs.											
[129]	SINet	2019	PVA-Net	Mono	**99.11**	90.59	79.77	**88.09**	79.22	70.30	**94.41**	**86.61**	80.68
	Innovation:	Context-aware RoI pooling with multi-branch heads to reduce scale variance.											
[130]	ASSD	2019	ResNet-101	Mono	89.28	89.95	82.11	69.07	62.49	60.18	75.23	76.16	72.83
	Innovation:	In-place self-attention for global context modeling in SSD.											
[131]	RefineDet	2018	VGG-16	Mono	98.96	90.44	88.82	84.40	77.44	73.52	86.33	80.22	79.15
	Innovation:	Two-step anchor refinement with transfer connection blocks.											
[132]	RFBNet	2018	VGG-16	Mono	87.31	87.27	84.44	66.16	61.77	58.04	74.89	72.05	71.01
	Innovation:	Introduces the Receptive Field Block (RFB), a module that enhances feature discriminability.											
[133]	SqueezeDet	2017	SqueezeNet	Mono	90.20	84.70	73.90	77.10	68.30	65.80	82.90	75.40	72.10
	Innovation:	Fully convolutional single-stage detection with ConvDet, optimized for speed and size.											
[134]	MS-CNN	2016	VGG-Net	Mono	90.03	89.02	76.11	83.92	73.70	68.31	84.06	75.46	66.07
	Innovation:	Multi-scale detection with proposal and detection sub-networks deconvolutional upsampling.											
[135]	SDP + RPN	2016	VGG-16	Mono	90.14	88.85	76.11	80.09	70.16	64.82	81.37	73.74	65.31
	Innovation:	Selects features by object scale and rejects negatives layer-wise to enhance detection accuracy.											

From Table 14, we provide an in-depth discussion of two selected models: AspectNet [127] and SINet [129], which achieved the highest performance among all evaluated methods, and Gaussian YOLOv3 [128], one of the most widely used models in the context of autonomous driving. AspectNet [127] approaches object detection tasks uniquely with an anchor-free detection scheme. It sets itself apart from comparable models by adding a special head that predicts object aspect ratios. Rather than tackline classification and regression features separately, the network combines them before reaching the aspect module, which allows geometric details to play a role in both areas. This additional branch is designed with just a few convolutional layers, ensuring it does not slow inference time too much. Another notable change is how positive and negative training samples are defined: the criteria for matching are tweaked to tackle the class imbalance that can hurt performance in single-stage models. These architectural decisions enhance the detector's ability to recognize elongated or skewed objects, as the receptive field is shaped better to fit the actual geometry of the detected objects.

SINet is a two-stage detector that uses a conventional CNN backbone (PVA) to generate proposals from multiple feature levels. Per proposal, a Context-Aware RoI pooling (CARoI) layer upsamples only small RoIs via bilinear deconvolution whose kernel size is set by the ratio of the target pooled size to the proposal size, and then applies max pooling, preserving small-object structure without upscaling entire feature maps. The pooled features from several backbone layers are concatenated and fed to a scale-aware decision head that partitions proposals by size; each branch (one conv + one fully connected layer with classification and box-regression heads) shares early features but is trained with a Gaussian-jittered threshold around the dataset's

median object scale to reduce intra-class variation. At inference, outputs from all branches are merged using a lightweight "soft-NMS" that averages the coordinates of highly overlapping, high-confidence boxes. Because deconvolution is applied only to small RoIs and the branching does not increase the number of processed proposals, these additions maintain end-to-end training while introducing essentially no extra runtime.

Gaussian YOLOv3 [128] keeps the overall layout of YOLOv3 intact, with three detection scales, feature aggregation from the backbone, and dense predictions. However, it changes the way bounding boxes are parameterized. Each coordinate is modeled as a Gaussian variable with its variance, meaning the network outputs not only position estimates but also an uncertainty term for each. These variances are fed into a modified regression loss based on negative log-likelihood, and they also influence the suppression of overlapping boxes at inference. The result is a system that can flag low-confidence localizations without discarding high-confidence detections that fall near ambiguous boundaries. Compared with AspectNet's emphasis on encoding geometric priors into a deterministic pipeline, Gaussian YOLOv3 takes a probabilistic route to preserve the speed of a one-stage detector while offering a measure of confidence that can be applied to downstream tasks.

Table 15 focuses on 3D object detection methods that use camera-based KITTI data, but generate 3D bounding boxes from 2D image inputs. From Table 15, we provide a detailed discussion of four selected models: M3D-RPN [136], recognized as one of the most widely adopted models in this field; MonoDGP [137], which is the latest released method; MonoTAKD [138], which achieved the highest results in AP_{BEV} Car and $Pedestrian$ evaluations; and MonoDiff [139] which achieved the highest results in AP_{3D} Car evaluation on the KITTI dataset. It should be mentioned that the values highlighted in red represent the best performance achieved for each evaluation metric across all listed methods.

The Monocular 3D Region Proposal Network (M3D-RPN) [136] presents an integrated approach for generating 2D and 3D object proposals from a single RGB image. Built upon a DenseNet-121 backbone modified with dilated convolutions to preserve a stride of 16, the network branches into two complementary pathways: one employing standard spatially invariant convolutions to capture global context, and another utilizing depth-aware convolutions to encode row-dependent geometric cues. In the latter, feature maps are divided into horizontal bins, with each bin processed by distinct convolutional kernels to reflect perspective variations inherent to urban driving scenes. Anchors are designed to carry 2D and 3D parameters, initialized from dataset-derived statistics to provide strong geometric priors. Predictions from the two branches are combined to regress full 3D bounding box parameters, followed by a lightweight post-optimization step that refines object orientation through projection consistency between 3D boxes and their 2D counterparts. M3D-RPN reduces complexity while maintaining accuracy by consolidating proposal generation into a single end-to-end stage.

MonoDGP [137] employs a transformer-based architecture that unites geometric priors with a novel perspective-invariant error formulation for depth refinement. A ResNet-50 backbone extracts multi-scale features, which are enhanced by a Region Segmentation Head (RSH) to emphasize foreground content and embed segment labels distinguishing object regions from background. This network's detection is decoupled into two parallel decoding processes: a 2D visual decoder for initializing queries and reference points from appearance cues and a 3D depth-guided decoder that integrates visual and depth features to model spatial relationships. Rather than relying on direct depth prediction or multi-branch fusion, MonoDGP computes geometric depth using the projection formula and introduces an explicit geometry error term to correct systematic offsets caused by perspective and object surface effects. With its stable and compact statistical distribution, this error term reduces training complexity while improving depth accuracy. Combining query decoupling allows the method to achieve improved convergence stability and precise 3D localization without requiring auxiliary data sources.

The MonoTAKD [138] framework introduces a three-part distillation strategy to enhance monocular 3D detection, comprising a LiDAR-based teacher, a camera-based teaching assistant (TA), and a camera-based student. Using a ResNet-50 backbone, the TA network integrates accurate ground-truth depth maps with visual features to produce high-fidelity BEV representations, referred to as 3D visual knowledge. These representations guide the student via intra-modal distillation, mitigating errors caused by monocular depth estimation. The student, architecturally aligned with the TA but relying on predicted depth, is structured with two feature branches: one dedicated to learning visual cues from the TA, and another designed to absorb spatial information from residual BEV features obtained by subtracting the TA's output from that of the LiDAR-based teacher. Cross-modal residual distillation transfers only these spatial cues, narrowing the modality gap without requiring the student to replicate the entire LiDAR feature space. A Spatial Alignment Module, combining dilated and deformable convolutions with channel attention, further refines spatial branch outputs. At the same time, a Feature Fusion Module merges the visual and spatial streams for final detection. This targeted multi-stage guidance enables the student to capture a richer geometric context than conventional monocular models.

Table 16 presents a comparative analysis of 3D object detection methods using camera-based data from the NuScenes dataset, where the input is 2D images and the output is 3D bounding boxes. From Table 16, we

Table 15: Comparative analysis of 3D object detection methods using the camera-based KITTI dataset.

						AP_{BEV} Car			AP_{3D} Car			AP_{3D} Pedestrian			AP_{3D} Cyclist		
Ref	**Method**	**Year**	**Venue**	**Backbone**	**Modality**	**Easy**	**Moderate**	**Hard**	**Easy**	**Moderate**	**Hard**	**Easy**	**Moderate**	**Hard**	**Easy**	**Moderate**	**Hard**
[137]	MonoDGP	2025	CVPR	Res50	Mono	35.24	25.23	22.02	26.35	18.72	15.97	–	–	–	–	–	–
	Innovation: Proposes geometry error prediction with decoupled 2D/3D queries and region segmentation to advance monocular 3D detection.																
[138]	MonoTAKD	2025	CVPR	Res50	Mono	**38.75**	**27.76**	**24.14**	27.91	19.43	16.51	**16.15**	**10.41**	**9.68**	**13.54**	**7.23**	**6.86**
	Innovation: Presents teaching assistant distillation uniting intra-modal visual knowledge transfer and cross-modal residual learning to enhance monocular 3D detection.																
[140]	FD3D	2024	AAAI	MonoDLE	Mono	34.20	23.72	20.76	25.38	17.12	14.50	–	–	–	–	–	–
	Innovation: Introduces AFOD-based feature supervision with pixel-level and distribution-level guidance to enhance monocular 3D object detection accuracy.																
[141]	MonoCD	2024	CVPR	DLA-34	Mono	33.41	22.81	19.57	25.53	16.59	14.53	–	–	–	–	–	–
	Innovation: Introduces complementary depth prediction using global clues and geometric relations to reduce coupling and improve monocular 3D detection.																
[139]	MonoDiff	2024	CVPR	Res50-FPN	Mono	–	–	–	**30.18**	**21.02**	**18.16**	13.51	8.94	7.28	8.52	4.35	3.78
	Innovation: Employs diffusion models for monocular 3D detection and pose estimation, refining predictions through iterative denoising of depth and pose.																
[142]	GUPNet++	2024	TPAMI	DLA-34	Depth	–	–	–	24.99	16.48	14.58	12.45	8.13	6.91	6.71	3.91	3.80
	Innovation: Introduces probabilistic geometry uncertainty propagation with IoU-guided confidence to mitigate projection error amplification in monocular 3D detection.																
[143]	MonoDETR	2023	ICCV	Res50	Mono	33.60	22.11	18.60	25.00	16.47	13.58	–	–	–	–	–	–
	Innovation: Introduces depth-guided Transformer with global reasoning and depth-aware queries to enhance monocular 3D object detection accuracy.																
[144]	MonoUNI	2023	NeurIPS	Res50	Mono	–	–	–	24.75	16.73	13.49	15.78	1.34	8.74	7.34	4.28	3.78
	Innovation: Proposes unified vehicle/infrastructure-side monocular 3D detection network leveraging sufficient depth clues for robust cross-perspective object localization.																
[145]	DEVIANT	2023	CVPR	Res50	Mono	29.65	20.44	17.43	21.88	14.46	11.89	–	–	–	–	–	–
	Innovation: Introduces depth-guided dynamic deformable attention to adaptively refine features for improved monocular 3D object detection accuracy.																
[146]	MonoPGC	2023	CVPR	DLA34	Mono	32.50	23.14	20.30	24.68	17.17	14.14	–	–	–	–	–	–
	Innovation: Incorporates pixel-level geometry via depth cross-attention, depth-space-aware transformer, and depth-gradient encoding for improved 3D detection.																
[147]	MonoATT	2023	CVPR	DLA-34	Mono	36.87	24.42	21.88	24.72	17.37	15.00	10.55	6.66	5.43	5.74	3.68	2.94
	Innovation: Proposes adaptive token Transformer that dynamically adjusts token count to improve efficiency and accuracy in monocular 3D detection.																
[148]	DCD	2023	ECCV	DLA34	Depth	–	–	–	23.81	15.90	13.21	10.37	6.73	6.28	4.72	2.74	2.41
	Innovation: Introduces depth-consistent distillation to transfer depth awareness from LiDAR-based teacher to monocular student for improved detection.																
[149]	MonoDDE	2022	CVPR	DLA34	Mono	33.58	23.46	20.37	24.93	17.14	15.10	11.13	7.32	6.67	5.94	3.78	3.33
	Innovation: Develops a depth solving system producing 20 diverse estimations and robustly combining them for reliable monocular 3D detection.																
[150]	MonoGround	2022	CVPR	DLA-34	Mono	30.07	20.47	17.74	21.37	14.36	12.62	12.37	7.89	7.13	4.62	2.68	2.53
	Innovation: Introduces a learnable ground plane prior with dense depth supervision, depth-align training, and two-stage depth inference.																
[151]	MonoJSG	2022	CVPR	DLA-34	Mono	32.59	21.26	18.18	24.69	16.14	13.64	11.94	7.36	6.03	8.03	3.87	3.33
	Innovation: Proposes joint semantic and geometric cost volume with adaptive depth sampling for refined monocular 3D detection.																
[152]	Kinematic3D	2022	ECCV	Dense121	Mono	26.69	17.52	13.10	19.07	12.72	9.17	–	–	–	–	–	–
	Innovation: Incorporates kinematic motion constraints to refine monocular 3D object detection in dynamic driving scenes.																
[153]	ImVoxelNet	2022	WACV	Res50	Mono	–	–	–	17.15	10.97	9.15	–	–	–	–	–	–
	Innovation: Projects posed images into a unified voxel representation enabling general-purpose 3D detection from monocular or multi-view input.																
[154]	ImVoxelNet	2022	CVPR	DLA34	Mono	–	–	–	20.10	12.99	10.50	12.47	7.62	6.72	1.52	0.85	0.94
	Innovation: Introduces homography loss aligning projected 3D boxes to ground plane to improve monocular 3D detection accuracy.																
[154]	MonoFlex	2022	CVPR	DLA34	Mono	–	–	–	21.75	14.94	13.07	11.87	7.66	6.82	5.48	3.50	2.99
	Innovation: Introduces homography loss aligning projected 3D boxes to ground plane to improve monocular 3D detection accuracy.																
[155]	CMKD	2022	ECCV	Res50	Depth	–	–	–	25.09	16.99	15.30	17.79	11.69	10.09	9.60	5.24	4.50
	Innovation: Transfers depth-aware features from a LiDAR-based teacher to a monocular student using cross-modal knowledge distillation framework.																
[156]	DID-M3D	2022	ECCV	DLA34	Depth	32.95	22.76	19.83	24.40	16.29	13.75	–	–	–	–	–	–
	Innovation: Decouples instance depth into visual and attribute depths with uncertainty modeling and enables effective affine augmentation.																
[157]	GUPNet	2021	ICCV	DLA34	Mono	–	–	–	20.11	14.20	11.77	14.72	9.53	7.87	4.18	2.65	2.09
	Innovation: Proposes Geometry Uncertainty Projection and Hierarchical Task Learning to mitigate depth error amplification in monocular 3D detection.																
[158]	MonoDLE	2021	CVPR	DLA34	Mono	24.79	18.89	16.00	17.23	12.26	10.29	5.34	3.28	2.83	4.59	2.66	2.45
	Innovation: Diagnoses localization error in monocular 3D detection and proposes center supervision, distance filtering, and IoU-based size loss.																
[159]	MonoRCNN	2021	ICCV	Res50	Mono	25.48	18.11	14.10	18.36	12.65	10.03	11.21	7.28	5.85	2.89	1.67	1.54
	Innovation: Introduces geometry-based distance decomposition with uncertainty-aware regression to improve interpretability, robustness, and accuracy in detection.																
[160]	MonoFlex	2021	CVPR	DLA34	Mono	28.23	19.75	16.89	19.94	13.89	12.07	9.43	6.31	5.26	4.17	2.35	2.04
	Innovation: Decouples truncated object prediction and adaptively ensembles multiple depth estimators using uncertainty for robust monocular 3D detection.																
[161]	MonoRUn	2021	CVPR	Res101	Mono	27.94	17.34	15.24	19.65	12.30	10.58	10.88	6.78	5.83	1.01	0.61	0.48
	Innovation: Employs self-supervised dense 3D reconstruction with uncertainty propagation and Robust KL loss for monocular 3D detection.																
[162]	DDMP-3D	2021	CVPR	Res101	Depth	28.08	17.89	13.44	19.71	12.78	9.80	4.93	3.55	3.01	4.18	2.35	2.04
	Innovation: Introduces depth-conditioned dynamic message propagation to enhance monocular 3D object detection through adaptive multi-scale feature interaction.																
[163]	MonoPair	2020	CVPR	DLA34	Mono	–	–	–	13.04	9.99	8.65	10.02	6.68	5.53	3.79	2.12	1.83
	Innovation: Leverages pairwise spatial constraints with uncertainty-aware optimization to improve detection of occluded objects.																
[164]	RTM3D	2020	ECCV	DLA34	Mono	19.17	14.20	11.99	14.41	10.34	8.77	–	–	–	–	–	–
	Innovation: Reformulates monocular 3D detection as keypoint detection of nine projected 3D box vertices plus center, using geometric constraints for real-time inference.																
[165]	UR3D	2020	ECCV	Res34	Mono	21.85	12.51	9.20	15.58	8.61	6.00	–	–	–	–	–	–
	Innovation: Learns a distance-normalized unified representation with distance-guided NMS and cascaded point regression for compact, accurate 3D detection.																
[166]	MoVi-3D	2020	ECCV	Res34	Mono	–	–	–	15.19	10.90	9.26	8.99	5.44	4.57	1.08	0.63	0.70
	Innovation: Leverages a virtual view-based training strategy to normalize object appearances across different depths, enabling a single-stage network to generalize.																
[167]	DA-3Ddet	2020	ECCV	Res101	Mono	–	–	–	16.77	11.50	8.93	8.7	7.1	6.7	14.5	11.5	11.5
	Innovation: Introduces feature domain adaptation and context-aware segmentation to align pseudo-LiDAR features with real-LiDAR representations.																
[168]	MonoDIS	2019	ICCV	Res34	Mono	17.23	13.19	11.12	10.37	7.94	6.40	–	–	–	–	–	–
	Innovation: Proposes loss disentanglement for 2D/3D detection and a self-supervised 3D-box confidence, enabling stable end-to-end monocular training.																
[136]	M3D-RPN	2019	ICCV	Dense121	Mono	–	–	–	14.76	9.71	7.42	4.92	3.48	2.94	0.94	0.65	0.47
	Innovation: Introduces a single-shot monocular 3D RPN with depth-aware convolution and shared 2D–3D anchors for accurate localization.																

provide a detailed discussion of two selected models: DETR3D [169], recognized as a well-known and widely adopted model in this field; and CorrBEV [170], which achieved the highest results on the NuScenes dataset. It should be noted that the values highlighted in red represent the best performance achieved for each evaluation metric across all listed methods.

DETR3D [169] addresses multi-view 3D object detection by establishing a direct mapping between a sparse set of learned 3D object queries and multi-camera image features, bypassing any intermediate dense geometric reconstruction. Feature extraction is performed by a shared ResNet–FPN backbone, producing multi-scale image representations for each view. Each object query contains a latent 3D hypothesis, decoded into a reference point in the scene. Through calibrated camera intrinsics and extrinsics, these points are back-projected into each image, where bilinear interpolation retrieves the corresponding feature vectors from all pyramid levels. Aggregated features are merged into the queries and fine-tuned using multi-head self-attention. This projection–refinement cycle repeats across multiple transformer decoder layers, with each stage producing updated classifications and 3D box regressions. DETR3D maintains an end-to-end pipeline that improves efficiency without sacrificing competitive detection accuracy by avoiding dense depth estimation, voxelization, and non-maximum suppression.

CorrBEV [170] improves upon transformer-based multi-view detection frameworks like BEVFormer. The task is completed by integrating a correlation-driven multi-modal approach to counteract the loss of discriminative information under occlusion. Its "Multi-modal Prototype Generator" forms category-specific visual prototypes from cropped ground-truth object patches and derives complementary language prototypes from class names processed through a lightweight BERT encoder. These visual–linguistic embeddings are combined and introduced into the backbone features using a depth-wise correlation mechanism. The Correlation-guided Query Learner leverages these correlation features in two complementary ways: first, to initialize object queries using high-confidence spatial locations, improving recall for partially occluded objects; and second, to perform dual-path mixed sampling in which 3D queries aggregate both backbone and correlation-derived features for richer semantic–geometric representation. An occlusion-aware trainer further augments the system via pseudo-occlusion, randomly masking pixels in high-visibility samples, then a contrastive alignment loss that brings occluded and obvious instances of the same category closer in the embedding space. This modular design yields consistent gains across all occlusion levels and integrates seamlessly into diverse baseline architectures with minimal computational cost.

Table 17 summarizes YOLO-based methods evaluated on the camera-based KITTI dataset for 2D object detection. Monocular 2D object detection methods have evolved to advance efficiency, representational depth, and domain-specific tuning. Early designs, such as YOLOV3-tiny [191], prioritized minimizing computation cost to improve efficiency with a precision trade-off. Later variants, including YOLOv5n [192] with its scaled CSP-based backbone (C2f) derived from CSPDarknet53 and YOLOv6n [193] employing the EfficientRep backbone, coupled backbone revisions with refined neck structures and decoupled heads to raise accuracy without undermining real-time operation. YOLOv7 [194] extended these gains by introducing re-parameterized convolution, a broadened trainable bag-of-freebies, and coarse-to-fine hierarchical label assignment. In turn, YOLOv8n [192] replaced C3 modules with C2f blocks. It adopted an anchor-free head, simplifying the pipeline while preserving detection strength, whereas YOLOv8-Lite [195] targeted resource-limited environments through compact backbone choices, lightweight feature aggregation, and attention mechanisms. ShuffYOLOX [196] modified the YOLOX framework by utilizing the lightweight ShuffDet backbone and integrating ECA attention within the PAFPN, which reduces architectural complexity without negatively affecting detection capabilities.

Domain-specific optimization has emerged as a critical trend within recent iterations. YOLO-Vehicle-v1s [197] was designed for vehicle detection in adverse visual conditions, incorporating image dehazing and multimodal image-text fusion to enhance robustness. YOLOv8-RTDAV [192] targeted road-target detection by embedding ECA-based C2f modules, DySample-based upsampling, and EIoU loss, yielding notable improvements in small-object accuracy. Transformer-CNN hybridization appeared in SwinYOLOv5s [198], which leveraged swin transformer attention with adaptive self-concat fusion to capture global context while mitigating false detections. Similarly, improved YOLOv5 [199] employed NAS-pruned branches, coordinate attention, and structural re-parameterization to achieve both speed and high precision for small-object detection. The latest designs, YOLOv9t [200] and YOLOv10n [201], represent a shift toward more fundamental architectural changes: the introduction of the GELAN backbone and programmable gradient optimization in YOLOv9t improves training stability, while YOLOv10n adopts an NMS-free consistent dual-assignment strategy to harmonize efficiency with accuracy. Collectively, these methods illustrate a clear trajectory from minimalist, speed-oriented architectures to hybrid designs that balance inference efficiency with enhanced feature representation, addressing the increasingly diverse deployment requirements of modern detection systems.

Table 16: Comparative analysis of 2D/3D object detection methods using the camera-based NuScenes dataset.

Ref	Method	Year	Venue	Backbone	Modality	mAP	NDS	mATE	mASE	mAOE	mAVE	mAAE
[170]	CorrBEV-fm	2025	CVPR	Res101	Muiti-View	41.30	50.70	0.646	0.265	0.482	0.473	0.134
	Innovation:	Introduces multimodal prototypes and correlation learning to improve occluded detection.										
[170]	CorrBEV-sp	2025	CVPR	Res101	Muiti-View	**56.20**	**64.00**	**0.488**	**0.241**	**0.303**	**0.243**	**0.123**
	Innovation:	Enhances detection of occluded objects in multi-view 3D perception.										
[171]	RoPETR	2025	ArXiv	V2-99	Muiti-View	**52.90**	**61.40**	**0.537**	**0.255**	**0.289**	**0.229**	**0.195**
	Innovation:	Enhances camera-only detection with multimodal rotary position embedding for better velocity estimation.										
[172]	BEVFormer	2024	IEEE	Res101	Muiti-View	44.50	53.50	0.631	0.257	0.405	0.435	0.143
	Innovation:	Uses spatiotemporal transformers with deformable attention to learn BEV from LiDAR-camera inputs.										
[173]	RayDN	2024	ECCV	Res50	Muiti-View	46.90	56.30	0.579	0.264	0.433	0.256	0.187
	Innovation:	Introduces depth-aware hard negative sampling along camera rays to reduce false positives in 3D detection.										
[174]	StreamPETR	2023	ICCV	Res50	Muiti-View	45.00	55.00	0.569	0.267	0.413	0.265	0.196
	Innovation:	Uses object-centric temporal modeling with motion-aware layer normalization for efficient 3D detection.										
[175]	PETRv2	2023	ICCV	Res50	Muiti-View	34.90	45.60	0.700	0.275	0.580	0.437	0.187
	Innovation:	Extends PETR with temporal 3D position embedding alignment and task-specific queries for 3D perception.										
[176]	BEVDepth	2023	AAAI	Res50	Muiti-View	35.10	47.50	0.639	0.267	0.479	0.428	0.198
	Innovation:	Improves 3D detection via explicit depth supervision, camera prediction, and depth refinement modules.										
[177]	SparseBEV	2023	ICCV	Res50	Muiti-View	44.80	55.80	0.581	0.271	0.373	0.247	0.190
	Innovation:	Proposes a sparse BEV detector with adaptive attention, spatio-temporal sampling, and adaptive mixing.										
[178]	HoP	2023	ICCV	Res50	Muiti-View	40.00	51.00	0.607	0.275	0.515	0.286	0.216
	Innovation:	Introduces method with short- and long-term temporal decoders to improve BEV feature learning.										
[179]	PolarFormer	2023	AAAI	Res101	Muiti-View	41.50	47.00	0.657	0.263	0.405	0.911	0.139
	Innovation:	Introduces Polar coordinate BEV representation with cross-attention decoding and multi-scale learning.										
[180]	FrustumFormer	2023	CVPR	Res101	Muiti-View	47.80	56.10	0.575	0.257	0.402	0.411	0.132
	Innovation:	Introduces adaptive instance-aware resampling with temporal frustum fusion to enhance object localization.										
[181]	FB-BEV	2023	ICCV	V2-99	Muiti-View	**53.70**	**62.40**	**0.439**	**0.250**	**0.358**	**0.270**	**0.128**
	Innovation:	Combines forward and depth-aware backward projection to densify BEV features.										
[182]	MV2D	2023	ICCV	V2-99	Muiti-View	46.30	51.40	0.542	0.247	0.403	0.854	0.127
	Innovation:	Generates dynamic 3D object queries from 2D detections with sparse cross-attention.										
[183]	CAPE	2023	CVPR	V2-99	Muiti-View	52.50	61.00	0.503	0.242	0.361	0.306	0.114
	Innovation:	Introduces camera-view position embeddings with bilateral attention to decouple extrinsic variations.										
[184]	3DPPE	2023	ICCV	VoV-99	Muiti-View	51.40	56.60	0.569	0.255	0.394	0.796	0.138
	Innovation:	Introduces depth-guided 3D point positional encoding with a hybrid-depth module for object localization.										
[185]	PETR	2022	ECCV	Res101	Muiti-View	37.00	45.50	0.647	0.251	0.433	0.933	0.143
	Innovation:	Encodes 3D coordinates into multi-view image features, producing position-aware representations detection.										
[186]	UVTR	2022	NeurIPS	Res101	Muiti-View	37.90	48.30	0.731	0.267	0.350	0.510	0.200
	Innovation:	Unifies voxel-based representation across modalities, enabling cross-modality fusion and knowledge transfer.										
[187]	BEVStereo	2022	AAAI	Res50	Muiti-View	37.20	50.00	0.597	0.270	0.438	0.367	0.190
	Innovation:	Proposes dynamic temporal stereo with EM-updated depth sampling and size-aware circle NMS.										
[188]	PGD	2022	CoRL	Res101	Muiti-View	38.60	44.80	0.626	0.245	0.451	1.509	0.127
	Innovation:	Combines probabilistic depth uncertainty modeling with graph-based geometric depth propagation.										
[169]	DETR3D	2022	CoRL	Res101+FPN	Muiti-View	41.20	47.90	0.641	0.255	0.394	0.845	0.133
	Innovation:	Uses sparse 3D object queries with geometric back-projection to fuse multi-view features.										
[189]	Graph-DETR3D	2022	ACM	Res101	Muiti-View	41.80	47.20	0.668	0.250	0.440	0.876	0.139
	Innovation:	Leverages dynamic 3D graph feature aggregation and depth-invariant multi-scale training.										
[190]	FCOS3D	2021	ICCV	Res101	Muiti-View	35.80	42.80	0.690	0.249	0.452	1.434	0.124
	Innovation:	Adapts anchor-free FCOS with 3D target reformulation and 2D-guided multi-level prediction.										

Table 17: Comparative analysis of Yolo-based detection methods using the camera-based KITTI dataset.

Ref	Model	Year	Venue	Backbone	Modality	Precision	Recall	mAP@50	Parameters	Model Size	FPS
[198]	SwinYOLOv5s	2025	MDPI	MobileNetV3	Mono	**96.00%**	**90.20%**	**95.70%**	50.30M	–	–
	Innovation:	Integrates swin transformer attention and adaptive fusion to improve global feature capture and reduce false detections.									
[192]	YOLOv8-RTDAV	2025	Scientific Reports	C2f	Mono	86.90%	81.10%	88.00%	2.81M	5.78MB	102.00
	Innovation:	Enhances YOLOv8n with ECA-based C2f, P2 small-object head, SPPELAN, and EIoU loss for better road-target accuracy.									
[197]	YOLO-Vehicle-v1s	2024	ArXiv	CSPDarknet53	Mono	–	–	93.50%	–	45.40MB	226.00
	Innovation:	Introduces vehicle-focused YOLO variant with image dehazing and multimodal image-text fusion to improve detection.									
[195]	YOLOv8-Lite	2024	ICCK	C2f	Mono	76.72%	–	75.32%	4.39M	–	85.00
	Innovation:	Employs FastDet backbone and CBAM attention to achieve lightweight, and high detection accuracy..									
[196]	ShuffYOLOX	2023	MDPI	ShuffDet	Mono	–	–	92.20%	35.45M	113.90MB	–
	Innovation:	Uses lightweight ShuffDet backbone and ECA-enhanced PAFPN to increase speed and maintain accuracy in AVs.									
[199]	improved YOLOv5	2023	Scientific Reports	RepNAS	Mono	92.90%	90.10%	90.10%	–	–	202.00
	Innovation:	Adds structural re-parameterization, NAS-pruned branches, small-object head, and coordinate attention for detection.									
[201]	YOLOv10n	2024	NeurIPS	CSPNet	Mono	87.20%	72.10%	80.20%	27.02M	5.50MB	103.00
	Innovation:	Redesigns YOLO with efficiency–accuracy balance and NMS-free training via consistent dual assignment strategy.									
[200]	YOLOv9t	2024	ECCV	GELAN	Mono	82.10%	69.30%	77.50%	2.01M	4.43MB	48.00
	Innovation:	Introduces GELAN backbone and programmable gradient information to improve training stability and detection accuracy.									
[192]	YOLOv8n	2023	N/A	C2f	Mono	88.00%	76.30%	85.20%	3.01M	5.97MB	148.00
	Innovation:	Employs C2f backbone and decoupled anchor-free head for stronger features and simpler inference.									
[194]	YOLOv7	2023	CVPR	ELAN	Mono	92.30%	97.10%	95.20%	–	–	132.00
	Innovation:	Adds trainable bag-of-freebies, re-parameterization, and coarse-to-fine lead-guided label assignment for better accuracy.									
[193]	YOLOv6n	2022	ArXiv	EfficientRep	Mono	85.30%	67.10%	76.00%	4.23M	8.30MB	189.00
	Innovation:	Employs EfficientRep backbone, Rep-PAN neck, and efficient decoupled head for speed with maintained accuracy.									
[192]	YOLOv5n	2020	N/A	CSPDarknet53	Mono	84.90%	70.00%	78.60%	2.51M	5.04MB	127.00
	Innovation:	Nano-scale YOLOv5 variant with CSPDarknet backbone for lightweight real-time detection.									
[191]	YOLOv3-tiny	2020	ICACCS	Tiny Darknet	Mono	82.90%	64.80%	72.50%	1.21M	23.20MB	276.00
	Innovation:	Compact YOLOv3 variant with reduced Darknet backbone trading accuracy for faster inference.									

5.2. 3D LiDAR-based Techniques

LiDAR sensors play a pivotal role in AV perception by providing accurate 3D geometric information of the surrounding environment, enabling precise object localization and size estimation under varying lighting and weather conditions. Unlike camera-based approaches that operate in the 2D image plane, LiDAR-based methods directly process point clouds, offering richer spatial representations and improved robustness to illumination changes. Over the past decade, a wide range of 3D LiDAR-based object detection architectures have emerged, each employing different spatial representations and learning strategies to balance accuracy, efficiency, and robustness [202]. In this survey, we categorize existing approaches into six main groups, including BEV-based, point-based, voxel-based, pillar-based, range-based, and point-voxel hybrid methods, providing a detailed analysis of their core principles, representative algorithms, strengths, and limitations.

BEV-based methods take LiDAR point clouds and project them onto a top-down Bird's-Eye View grid, where each cell collects information from the points that fall within it. This BEV representation gets treated like a regular 2D image and processed through convolutional neural networks. The approach is computationally efficient and works well for understanding spatial relationships like road layouts, lane structures, and where vehicles are positioned. Since it shows the scene from above, it naturally fits with the mapping and path planning systems that autonomous cars use. However, flattening everything into this top-down view can reduce depth detail and hide information about what's stacked vertically, making it harder to detect objects hidden behind others or to capture fine height variations. The output is typically a set of 3D bounding boxes in the global coordinate frame, along with class labels and confidence scores.

Figure 7 illustrates the overall architecture of point-based approaches in 3D LiDAR-based object detection. Point-based methods directly operate on raw, unordered LiDAR point clouds by applying feature sampling, grouping, and point-wise feature learning to extract meaningful representations from the data. Then, 3D bounding boxes are predicted based on the downsampled points and features. Sampling is an important step in the architecture, and several strategies are commonly employed in 3D LiDAR-based detection: Random Sampling, which selects points without geometric bias; Farthest Point Sampling (FPS), which iteratively selects the farthest points to ensure spatial coverage; Voxel Grid Sampling, which divides the space into uniform voxels

and selects representative points from each; Uniform Sampling, which ensures evenly spaced points within a defined extent; Importance Sampling, which prioritizes points with higher learned or predefined significance; and Density-based Sampling, which reduces over-representation of dense areas while preserving points in sparse regions. By processing each sampled point individually while grouping local neighborhoods, point-based methods preserve complete geometric details of the scene, making them highly suitable for fine-grained shape modeling.

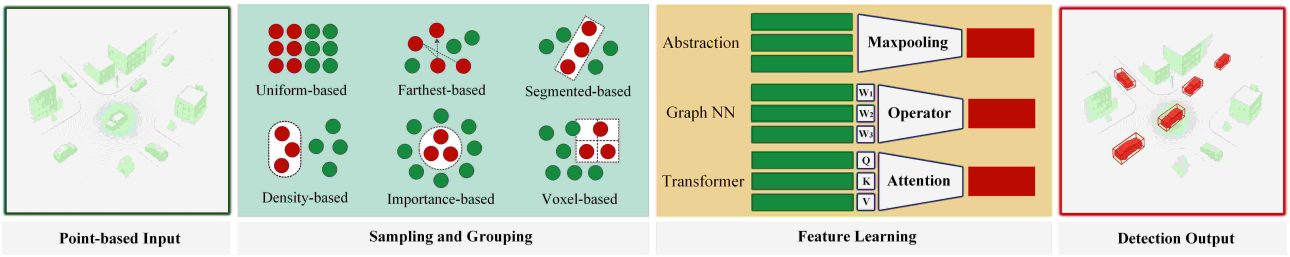

Figure 7: Overall framework of the Point-based approaches in 3D Lidar object detection.

Figure 8 illustrates the overall architecture of range-based LiDAR object detection methods. Range-based methods project LiDAR point clouds into a Range View representation using spherical coordinates, where each pixel corresponds to a single LiDAR return. Typically, the input combines geometric features, such as range, X, Y, and Z coordinates, with auxiliary features like elongation and intensity. For instance, in the Waymo dataset, each channel is remapped so that warmer colors represent the smallest values and cooler colors represent the largest values within their respective domains. This representation preserves accurate depth information and provides full 360° coverage of the environment. By treating the range image as a dense 2D grid, 2D convolutional neural networks can efficiently process the data while leveraging well-established image-based techniques. Range-based methods are particularly effective for long-range perception, as they maintain precise range and angular relationships. However, they can introduce spatial distortions, especially for far-away or steeply angled objects.

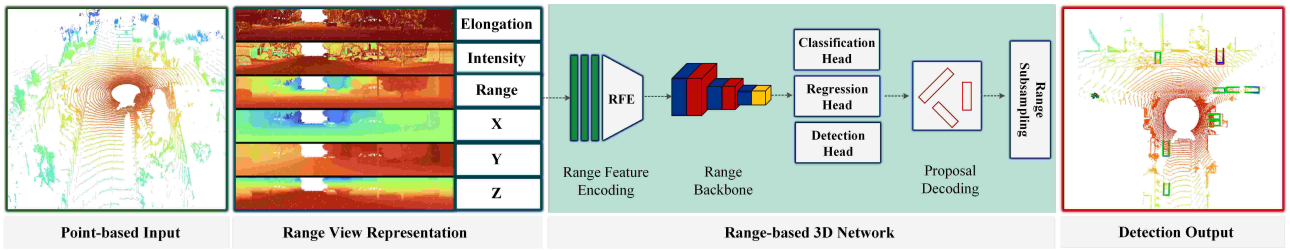

Figure 8: Overall framework of the Range-based approaches in 3D Lidar object detection.

Figure 9 presents the overall architecture of voxel-based 3D LiDAR object detection methods, which includes the voxelization process, feature extraction, and detection head. Voxel-based methods transform raw LiDAR point clouds into structured volumetric representations by dividing the 3D space into spaced voxels. In order to extract spatial features, 3D convolutional neural networks (3D CNNs) aggregate the points within each voxel. This representation makes Convolutional efficiency possible while maintaining the significance of geometric context. Voxelization, transforming point clouds into fixed-size voxel grids, is a crucial stage in voxel-based architectures. The resolution of these voxels significantly impacts performance: finer resolutions capture more geometric detail but increase memory consumption and computational cost. In comparison, rougher resolutions reduce resource demands but may lose critical structural information. Some methods also employ sparse convolution techniques to process only occupied voxels, improving efficiency. By applying convolutions over the voxelized space, these methods effectively capture local and global spatial patterns, making them suitable for large-scale autonomous driving scenes. However, voxel-based approaches suffer from quantization errors due to discretization, and high-resolution grids can lead to substantial memory overhead.

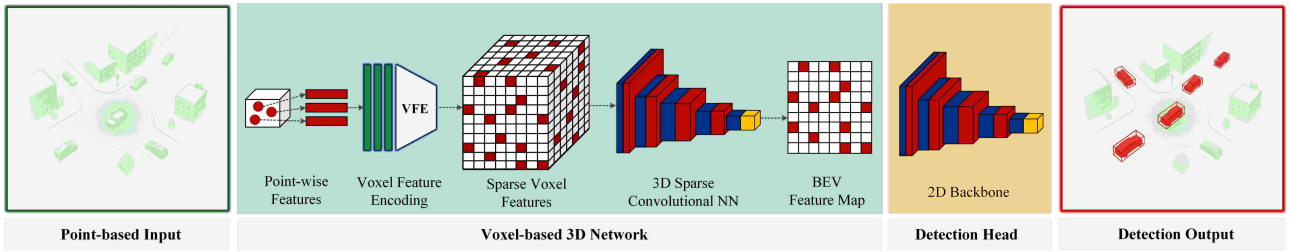

Figure 9: Overall framework of the Voxel-based approaches in 3D Lidar object detection.

Figure 10 outlines the overall architecture of pillar-based LiDAR object detection methods, including the pillar transformation process, pillar feature encoding, and detection head. Pillar-based methods organize LiDAR point cloud data by dividing the scene into vertical columns, or "pillars," and merging the height dimension. Instead of keeping the whole 3D structure, the data is flattened into a 2D layout, with each pillar storing the features from the inside points. These features are often aggregated using methods such as mean pooling or learned encoders before being passed to the detection network. This design makes the 2D convolutional networks faster and less memory-intensive, enabling high throughput on embedded automotive hardware. The advantage of this approach lies in its efficiency, making it well-suited for real-time detection in autonomous driving scenarios. However, by discarding vertical resolution, pillar-based methods may struggle with tall or vertically complex objects, where fine height details are important for accurate recognition. The final output is typically a set of 3D bounding boxes with associated class labels and confidence scores for each detected object.

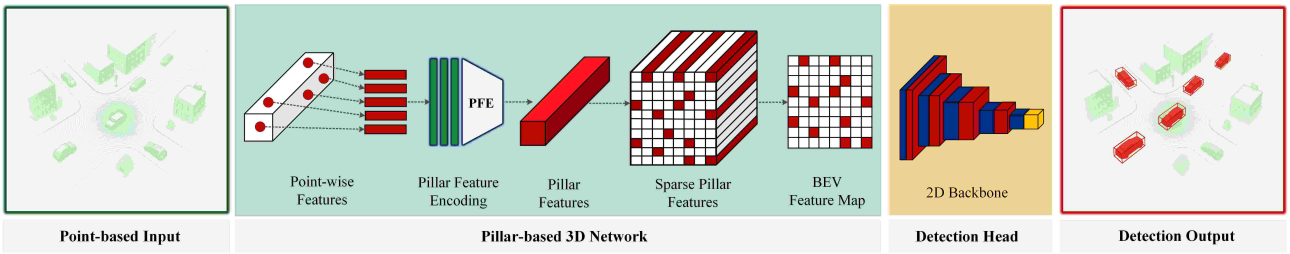

Figure 10: Overall framework of the Pillar-based approaches in 3D Lidar object detection.

Figure 11 presents the overall architecture of point–voxel hybrid LiDAR object detection methods, including the voxelization stage, point-level feature extraction, feature fusion, and detection head. Point–voxel hybrid methods combine the strengths of point-based and voxel-based approaches for better accuracy and efficiency. The key idea is to learn fine-grained point features that capture local geometric details while extracting structured voxel features that allow for efficient convolution operations on regular grids. Typically, raw point clouds are first voxelized to produce coarse spatial features, and representative points are sampled (e.g., using farthest point sampling) to extract high-resolution features. These two feature streams are then fused, either through concatenation or attention-based aggregation, to provide a richer representation for the detection head. This strategy preserves detailed geometry while maintaining the computational benefits of voxel processing. However, the architecture and training process are generally more complex, often requiring careful balancing of the contributions from each branch. The final output is typically accurate 3D bounding boxes along with what each object is and how confident the system is about those predictions.

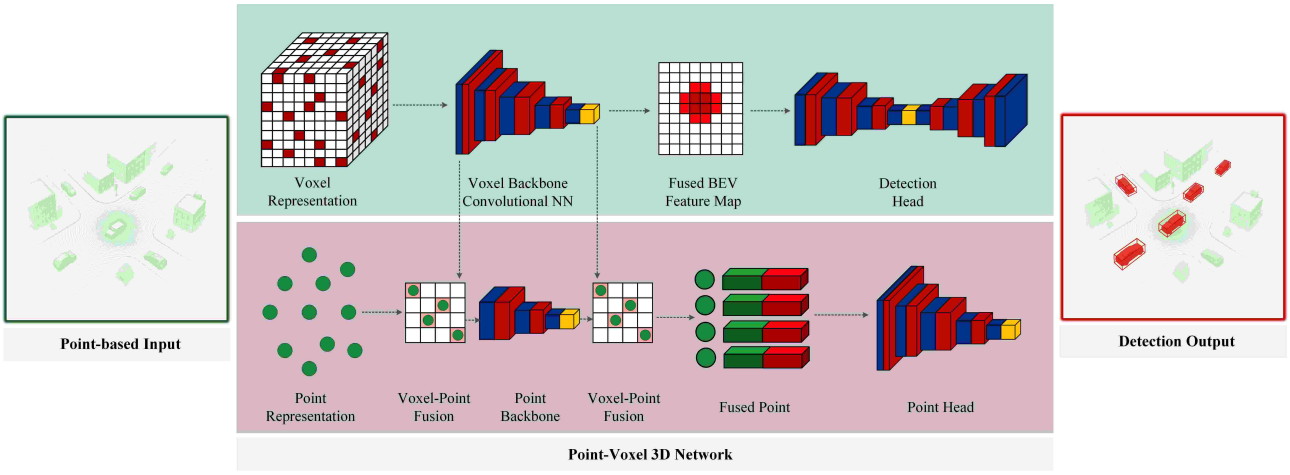

Figure 11: Overall framework of the Voxel-Point approaches in 3D Lidar object detection.

In the following, we present all existing 3D LiDAR-based algorithms for object detection in autonomous vehicles from 2018 to 2025. To ensure a consistent and fair comparison, we organize the results into three separate tables, as the performance of these methods is evaluated on different datasets. Table 18 provides a comparative analysis of 3D object detection methods using LiDAR-based data from the KITTI dataset, while Table 19 summarizes 3D object detection methods evaluated on the LiDAR-based NuScenes dataset. Finally, Table 20 presents a comparative analysis of 3D object detection methods based on LiDAR–Radar data from the VOD dataset.

Table 18: Comparative analysis of 3D object detection methods using the LiDAR-based KITTI dataset.

Ref	Method	Year	Venue	Backbone	Modality	AP_{BEV} Car			AP_{3D} Car			AP_{3D} Pedestrian			AP_{3D} Cyclist		
						Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
[203]	Fade3D	2025	T-ITS	SFENet	Point	94.86	89.28	86.55	90.92	82.00	77.49	–	–	–	–	–	–
	Innovation: Lightweight input encoder, spatially enhanced BEV backbone, and IoU-aware re-weighting for deployable 3D detection.																
[204]	AS-Det	2025	AAAI	AS	Point	95.22	91.33	86.25	90.55	82.41	77.08	–	–	–	–	–	–
	Innovation: Active sampling and multi-scale center-feature aggregation for adaptable single-stage 3D detection (LiDAR/4D Radar).																
[205]	LION	2024	NeurIPS	LION	Voxel	91.40	89.29	87.26	89.25	80.97	78.28	–	–	–	–	–	–
	Innovation: Linear Group RNN models long-range voxel dependencies efficiently.																
[206]	TSSTDet	2024	IEEE	3D CNN	Voxel	**95.80**	**92.11**	**89.23**	91.84	85.47	80.65	75.13	69.38	64.31	**95.16**	**76.24**	**71.62**
	Innovation: Models spatial shape features with a transformer inside the voxel pipeline.																
[207]	SwiftPillars	2024	AAAI	SPE	Pillar	–	–	–	88.07	79.83	77.50	55.34	50.21	45.93	86.06	64.92	60.75
	Innovation: Dual-attention pillar encoding with lightweight operators and multi-scale aggregation for high-speed deployment.																
[208]	PVT-SSD	2023	CVPR	Sparse CNN	Point-Voxel	95.23	91.63	86.43	90.66	82.29	76.85	–	–	–	–	–	–
	Innovation: Point-voxel transformer with query initialization and virtual range-image acceleration.																
[209]	DSVT	2023	CVPR	Transformer	Voxel	92.04	88.33	87.81	87.89	80.90	78.66	–	–	–	–	–	–
	Innovation: Dynamic sparse window attention and rotated set partitioning; no custom CUDA.																
[210]	Ada3D	2023	ICCV	3D/2D CNN	Voxel	93.51	89.42	88.19	89.70	85.29	82.86	–	–	–	–	–	–
	Innovation: Adaptive inference via importance prediction, density guidance, and sparsity-preserving normalization.																
[211]	VoxelNeXt	2023	CVPR	Sparse CNN	Voxel	92.66	87.87	87.04	87.86	80.02	77.34	–	–	–	–	–	–
	Innovation: Dense heads removed; direct prediction from sparse voxel features for efficient detection/tracking.																
[212]	3D HANet	2023	TGRS	VFE	Point-Voxel	94.33	91.63	86.33	90.79	84.18	77.57	–	–	–	–	–	–
	Innovation: Plug-and-play 3D heatmap auxiliary network enhancing spatial perception without added latency.																
[213]	TED-S	2023	AAAI	TESP CNN	Voxel	–	–	–	**93.05**	**87.91**	**85.81**	72.38	67.81	63.54	93.09	75.77	71.20
	Innovation: Transformation-equivariant voxel features with TeSpConv; TeBEV and TiVoxel pooling.																
[214]	RDIoU	2022	ECCV	CT-stacked	Voxel	94.90	89.75	84.67	90.65	82.30	77.26	–	–	–	–	–	–
	Innovation: Rotation-Decoupled IoU stabilizes training by mitigating rotation coupling in 3D IoU optimization.																
[215]	IA-SSD	2022	CVPR	PointNet++	Point	93.14	89.48	84.42	88.87	80.32	75.10	46.51	39.03	35.60	78.35	61.94	55.70
	Innovation: Learnable instance-aware downsampling and contextual centroid perception for efficient point-based detection.																
[216]	PillarNet	2022	ECCV	VGG-18/34	Pillar	93.66	87.85	86.64	89.94	79.01	77.31	–	–	–	–	–	–
	Innovation: Powerful sparse 2D CNN encoder with optimized neck for pillar-based detection.																
[217]	VoxSeT	2022	CVPR	VoxSeT	Voxel	92.70	89.07	86.29	88.53	82.06	77.46	–	–	–	–	–	–
	Innovation: Voxel-based set attention enabling transformer self-attention on variable-size clusters.																
[218]	BtcDet	2022	AAAI	RPN	Voxel	92.81	89.34	84.53	90.60	83.02	77.36	69.39	61.19	55.86	91.45	74.70	70.08
	Innovation: Learns occluded shape priors and integrates shape occupancy into proposal generation/refinement.																
[219]	Voxel-RCNN	2021	AAAI	RPN	Voxel	93.54	91.18	88.91	92.38	85.29	82.86	64.88	57.32	52.11	89.25	72.52	67.03
	Innovation: Voxel RoI pooling to directly extract 3D voxel features for refinement.																
[220]	CIA-SSD	2021	AAAI	SPConvNet	Voxel	93.74	89.84	82.39	89.59	80.28	72.87	–	–	–	–	–	–
	Innovation: Aligns confidence and localization via feature aggregation with IoU-aware confidence rectification.																
[221]	CT3D	2021	ICCV	RPN	Voxel	92.36	88.83	84.07	87.83	81.77	77.16	65.73	58.56	53.04	91.99	71.60	67.34
	Innovation: Channel-wise Transformer with proposal-to-point embedding and extended re-weighting.																
[222]	SPG	2021	ICCV	VFE	Pillar	92.80	89.12	86.27	90.64	82.66	77.91	–	–	–	–	–	–
	Innovation: Generates semantic points in missing regions to recover degraded LiDAR data.																
[223]	SE-SSD	2021	CVPR	SPConvNet	Voxel	95.68	91.84	87.62	91.49	82.54	77.15	67.98	59.72	54.83	91.77	72.54	68.78
	Innovation: Self-ensembling with consistency constraints, orientation-aware IoU loss, and shape-aware augmentation.																
[224]	RangeDet	2021	ICCV	FPN	Range	87.96	69.03	48.88	–	–	–	**82.20**	**75.39**	**65.74**	–	–	–
	Innovation: Pure range-view detector with Meta-Kernel, range-conditioned FPN, and weighted NMS.																
[225]	SA-SSD	2020	CVPR	RPN	Voxel	95.03	91.03	85.96	88.75	79.79	74.16	61.72	55.01	49.94	88.82	71.25	65.88
	Innovation: Improves single-stage detection with a detachable auxiliary network, point-level supervision, and part-sensitive warping.																
[226]	Part-A^2	2020	IEEE	SPConvNet	Point-Voxel	84.76	89.52	81.47	77.86	85.94	72.00	44.50	54.49	42.36	62.73	75.58	57.74
	Innovation: Part-aware segmentation and part-aggregation RoI pooling for proposal refinement.																
[227]	Point-GNN	2020	CVPR	GNN	Point	93.11	89.17	83.90	88.33	79.47	72.29	–	–	–	–	–	–
	Innovation: Fixed-radius graph with auto-registration reduces translation variance; box-merging improves accuracy.																
[228]	3DSSD	2020	CVPR	SA	Point	92.66	89.02	85.86	88.36	79.57	74.55	35.03	27.76	26.08	66.69	59.00	55.62
	Innovation: Removes FP layers and refinement using fusion sampling, anchor-free regression, and 3D center-ness labels.																
[229]	TANet	2020	AAAI	Attention	Voxel	–	–	–	83.81	75.38	67.66	54.92	46.67	42.42	73.84	59.86	53.46
	Innovation: Triple attention (channel/point/voxel) with coarse-to-fine regression.																
[230]	SPVCNN	2020	ECCV	SPVCNN	Point-Voxel	–	–	–	–	–	–	49.20	41.40	38.40	80.10	63.70	56.20
	Innovation: Sparse Point-Voxel Convolution with high-resolution point branch and NAS-optimized design.																
[231]	PV-RCNN	2020	CVPR	3D CNN	Point-Voxel	94.98	90.65	86.14	90.25	81.43	76.82	64.26	56.67	51.91	88.88	71.95	66.78
	Innovation: Voxel-to-keypoint encoding and keypoint-to-grid RoI abstraction.																
[232]	PointRCNN	2019	CVPR	PointNet++	Point	92.13	87.39	82.72	86.96	75.64	70.70	47.98	39.37	36.01	74.96	58.82	52.53
	Innovation: Bottom-up 3D proposals via point-cloud segmentation, refined in canonical coordinates using bin-based regression.																
[233]	STD	2019	ICCV	PointNet++	Point-Voxel	89.66	65.32	76.06	86.61	77.63	76.06	53.08	44.24	41.97	78.89	62.53	55.77
	Innovation: Transformer backbone on sparse voxels with Local/Dilated Attention and GPU hash–based Fast Voxel Query.																
[234]	SECOND	2018	MDPI	Sparse CNN	Voxel	88.07	79.37	77.95	83.13	73.66	66.20	45.31	35.52	33.14	75.83	60.82	53.67
	Innovation: Introduces sparse convolution with GPU-optimized rule generation, novel angle loss, and ground-truth sampling for faster LiDAR 3D detection.																
[235]	VoxelNet	2018	CVPR	VFE	Voxel	89.60	84.81	87.57	81.97	65.46	62.85	65.95	61.05	56.98	74.41	52.18	50.49
	Innovation: End-to-end voxel feature encoding learning local geometry and global context.																

From Table 18, we have selected four models for detailed discussion, focusing on their distinctive contributions for 3D object detection in KITTI dataset: TED-S [213], obtained the highest results in AP_{3D} *Car* evaluations; TSSTDet [206], which achieved the highest results in AP_{BEV} *Car* and AP_{3D} *Cyclist* evaluations; SECOND [234], noted as one of the most widely adopted models in this field; and Fade3D [203], representing the latest advancement in this domain. It should be mentioned that the values highlighted in red represent the best performance achieved for each evaluation metric across all listed methods.

TED-S [213] is a two-stage LiDAR 3D object detector that introduces transformation-equivariant processing into the Voxel-RCNN framework. The network first applies a Transformation-equivariant Sparse Convolution (TeSpConv) backbone, in which the input point cloud is transformed into multiple discrete yaw-rotated and mirrored copies, voxelized, and processed with shared sparse convolutional weights to produce multi-channel equivariant voxel features. These are aggregated at the scene level through TeBEV pooling, which collapses each transformation channel into a BEV map, aligns them via bilinear sampling, and max-pools across channels to form a compact, transformation-consistent representation for the 2D RPN proposal stage. Proposals are then refined using TiVoxel pooling, where instance-level grids are projected into each transformation channel, voxel features are sampled via Voxel Set Abstraction, concatenated across channels, and fused using cross-grid attention for the final detection head. This design preserves geometric consistency under predefined rigid transformations while leveraging an efficient RPN-refinement pipeline for high-accuracy 3D detection.

TSSTDet framework [206] utilizes a multistage design that addresses two challenges in LiDAR-based object detection: the considerable variation in object orientation and occlusion affecting geometry. In the first stage, a Rotational-Transformation Convolutional backbone (RTConv) processes multiple rotated and reflected variants of the input point cloud through a shared sparse convolutional network, producing sets of voxel features that remain equivariant to the applied transformations. These features are aligned and merged into a unified bird's-eye-view representation via a custom RT BEV pooling process, which feeds a region proposal network to generate initial 3D candidate boxes. Building on these proposals, the second stage introduces the Voxel-Point Shape Transformer (VPST). This autoregressive transformer module reconstructs the object geometry by predicting a sequence of quantized voxel features from the partial observation, thereby enriching the proposal with occlusion-compensated detail. The final stage uses a Fusion and Refinement network, which combines the original rotation equivariant features with the completed shape features for enabling high-confidence bounding box predictions.

SECOND [234] extends voxel-based 3D detection by introducing a sparse convolutional backbone that dramatically reduces the computational footprint of earlier dense 3D convolution designs. The network begins by grouping points into voxels and applying a Voxel Feature Encoding (VFE) module to produce learned descriptors for each occupied cell. Those features enter the Sparse Convolutional Middle Extractor, a series of 3D sparse convolutional layers that operate on non-empty locations to maintain spatial fidelity without unnecessary computation in space. The vertical resolution is reduced as processing progresses, enabling the feature map to be reshaped into a dense BEV representation suitable for efficient 2D convolutional processing. This representation is then handled by a region proposal network with an SSD-like multi-scale design, combining standard and transposed convolutions to generate proposals across object sizes. This design also introduces refinements such as a sine-error formulation for orientation regression, which further enhances geometric precision and detection stability while maintaining the backbone's streamlined nature.

Fade3D framework [203] aims to minimize computational latency while maintaining the essential spatial cues for 3D object detection. The processing begins with a Lightweight Input Encoder (LIE), which converts the raw point cloud into a compact Birds Eye View (BEV) image. Rather than employing computationally intensive point-based encoders, the system relies on a pilot study showing that recording only the maximum height and maximum reflectance per grid cell achieves an optimal balance between accuracy and runtime. This encoded BEV representation is passed to SFENet, a specialized 2D convolutional backbone constructed in three stages: an initial convolutional block for base feature extraction, a depth-wise shuffle block augmented with channel attention for efficient mid-level processing, and a re-parameterized block that simplifies to a single convolution at inference time to accelerate deployment. The final stage integrates an IoU aware loss re-weighting strategy toward more challenging samples without impacting runtime, which ensures an uncompromised inference speed.

Table 19 presents a comparative analysis of 3D object detection methods using the LiDAR-based NuScenes dataset. From Table 19, we provide a detailed discussion of three selected models: PointPillars [255], recognized as one of the pioneer model in this field; Voxel Mamba [236], which achieved the highest results in terms of NDS (NuScenes Detection Score), mAP, and trailer detection on the NuScenes dataset; and LION [205], obtained the highest results for car, truck, bus, pedestrian, and cyclist detection on NuScenes dataset. PointPillars [255] adopts a pillar-based representation to enable a purely 2D convolutional pipeline for real-time 3D object detection. The approach first discretizes the point cloud into vertical columns (pillars) with infinite extent along the z-axis, then uses a simplified PointNet to extract point-wise features within each pillar. These are aggregated and scattered into a dense pseudo-image in the bird's-eye-view plane. A 2D CNN backbone, composed of a top-down path and a multi-scale upsampling path, processes this pseudo-image to extract high-level spatial features. Finally, an SSD-style detection head performs classification and oriented 3D bounding box regression directly in BEV space, allowing end-to-end training without any 3D convolutions.

Voxel Mamba [236] introduces a group-free state-space backbone for 3D object detection, exploiting the linear-time sequence modeling of State Space Models (SSMs) to bypass the computational bottlenecks of

Table 19: Comparative analysis of 3D object detection methods using the LiDAR-based NuScenes dataset.

Ref	Method	Year	Venue	Backbone	Modality	NDS	mAP	Car	Truck	Bus	Trailer	Pedestrian	Motorcycle	Bicycle
[236]	Voxel Mamba	2024	NeurIPS	SSMs	Voxel	**73.0**	**69.0**	86.8	57.1	68.0	**63.2**	89.5	74.7	50.8
	Innovation:	Replaces attention with group-free state-space operators on voxels to model long-range dependencies efficiently.												
[205]	LION-RetNet	2024	NeurIPS	LION3D	Point	71.9	67.3	**89.6**	64.3	**78.7**	44.6	**89.8**	73.5	56.6
	Innovation:	Enhances BEV fusion via LiDAR–image multi-scale interaction with cross-attention and dynamic modality weighting.												
[205]	LION-RWKV	2024	NeurIPS	LION3D	Voxel	71.7	67.6	89.4	59.0	77.6	37.1	89.7	74.3	56.2
	Innovation:	Employs the same BEV fusion strategy with an RWKV backbone for recurrent state-based spatio-temporal feature modeling.												
[205]	LION-Mamba	2024	NeurIPS	LION3D	Voxel	72.1	68.0	87.9	**64.9**	77.6	44.4	89.6	**75.6**	**59.4**
	Innovation:	Employs the same BEV fusion strategy with a Mamba backbone for efficient long-range dependency modeling.												
[237]	PillarNeSt-Base	2024	T-IV	ConNeXt	Pillar	71.3	66.9	87.4	56.4	64.0	63.0	86.6	69.4	46.8
	Innovation:	Scales pillar-based backbones with large-scale pretraining to enhance spatial feature learning for LiDAR 3D object detection.												
[237]	PillarNeSt-Large	2024	T-IV	ConvNeXt	Pillar	71.6	66.9	87.4	56.4	64.0	63.0	86.6	69.4	51.5
	Innovation:	Enlarges model capacity within the same architecture to further enhance feature representation and detection performance.												
[238]	SAFDNet	2024	CVPR	Sparse CNN	Voxel	72.3	68.3	87.3	57.3	68.0	61.7	89.0	71.1	44.8
	Innovation:	Introduces adaptive feature diffusion to mitigate center feature missing while preserving efficiency in fully sparse LiDAR detection.												
[239]	FSDv2	2024	TPAMI	SparseUNet	Voxel	71.7	66.5	86.1	53.0	66.5	61.1 5	87.1	71.1	51.7
	Innovation:	Replaces instance clustering with virtual voxels, simplifying FSD and addressing center-feature missing in fully sparse detectors.												
[211]	VoxelNeXt	2023	CVPR	Sparse CNN	Voxel	70.0	64.5	84.6	53.0	64.7	55.8	85.8	73.2	45.7
	Innovation:	Eliminates dense heads by directly predicting objects from sparse voxel features, enabling efficient fully 3D detection and tracking.												
[240]	LinK	2023	CVPR	LinK-based	Voxel	71.0	66.3	86.7	55.7	65.7	62.1	85.5	73.5	47.5
	Innovation:	Introduces a linear kernel that assigns weights only to non-empty voxels, improving efficiency in sparse 3D CNNs.												
[241]	HEDNet	2023	NeurIPS	Sparse CNN	Point-Voxel	72.0	67.7	86.5	57.9	70.4	61.2	87.9	70.4	46.9
	Innovation:	Builds hierarchical encoder–decoder with semantic- and depth-guided blocks to exploit multi-scale 3D context efficiently.												
[242]	LargeKernel3D	2023	CVPR	3D CNN	Voxel	70.9	65.4	85.5	53.8	64.4	59.5	85.9	72.7	46.8
	Innovation:	Uses spatial-wise partition convolution and position embedding to realize efficient, very large 3D kernels in sparse CNNs.												
[209]	DSVT	2023	CVPR	BEV	Voxel	72.7	68.0	86.8	54.8	67.4	63.1	88.0	73.0	47.2
	Innovation:	Efficient transformer backbone using dynamic sparse attention and rotated set partitioning without custom CUDA operations.												
[243]	FSTR	2023	TGRS	VoxelNeXt	Point	71.5	67.2	86.5	54.1	66.4	58.4	88.6	73.7	48.1
	Innovation:	Introduces a fully sparse transformer with dynamic queries and Gaussian denoising for efficient long-range LiDAR 3D detection.												
[244]	Uni3DETR	2023	NeurIPS	Sparse CNN	Point-Voxel	68.5	57.3	86.1	57.8	63.5	38.2	88.7	74.6	42.2
	Innovation:	A unified DETR-style framework using point–voxel interaction to handle indoor and outdoor scenes consistently.												
[245]	FCOS-LiDAR	2022	NeurIPS	LiDAR-Net	Range	65.7	60.2	82.2	47.7	52.9	48.8	84.5	68.0	39.0
	Innovation:	Standard 2D convolutions and multi-frame MRV projection.												
[246]	AFDetV2	2022	AAAI	RPN	Voxel	68.5	62.2	86.3	52.6	62.8	59.9	85.4	68.3	34.8
	Innovation:	Single-stage anchor-free LiDAR detector with SC-Conv backbone, IoU-aware rescoring, and keypoint auxiliary supervision.												
[186]	UVTR-L	2022	NeurIPS	ResNet-101	Voxel	69.7	63.9	86.3	52.2	62.8	59.7	84.5	68.8	41.1
	Innovation:	Cross-modality interaction and a transformer decoder.												
[247]	Focals Conv	2022	CVPR	Sparse CNN	Voxel	70.0	63.8	86.7	56.3	67.7	59.5	87.5	64.5	36.3
	Innovation:	Makes sparsity learnable via position-wise importance, improving sparse CNNs for LiDAR-only and multi-modal detection.												
[216]	PillarNet	2022	ECCV	VGG-18/34	Pillar	71.4	66.4	86.1	53.1	65.3	63.1	87.3	70.1	42.3
	Innovation:	Introduces a powerful sparse 2D CNN encoder and optimized neck for efficient pillar-based 3D detection.												
[248]	TransFusion	2022	CVPR	3D CNN	Point	70.2	65.5	86.2	56.7	66.3	58.8	86.1	68.3	44.2
	Innovation:	Uses a transformer decoder to fuse LiDAR BEV features.												
[249]	VMVF	2022	CVPR	3D CNN	Range	66.3	58.3	84.4	51.5	58.3	53.9	85.3	63.0	32.5
	Innovation:	Enhances BEV-based 3D detection by fusing RV panoptic segmentation features with semantic and instance guidance.												
[250]	CenterPoint	2021	CVPR	VoxelNet	Point	65.5	58.0	84.6	51.0	60.2	53.2	83.4	53.7	28.7
	Innovation:	Reformulates 3D detection and tracking as center keypoint estimation, enabling efficient two-stage refinement.												
[251]	Pointformer	2021	CVPR	Pointformer	Point	63.8	53.8	82.3	45.1	48.3	45.4	85.1	55.7	25.8
	Innovation:	Pure Transformer backbone capturing local and global context for point cloud detection.												
[252]	CVCNET	2020	NeurIPS	3D CNN	Voxel	66.6	58.2	82.6	49.5	59.4	51.1	83.0	61.8	38.8
	Innovation:	Introduces hybrid cylindrical–spherical voxelization and cross-view transformers to enforce BEV/RV consistency.												
[228]	3DSSD	2020	CVPR	SA	Point	56.4	42.6	81.2	47.2	61.4	30.5	70.2	36.0	8.6
	Innovation:	Removes FP layers and refinement stage via fusion sampling, candidate generation, and anchor-free regression.												
[253]	3DVID	2020	CVPR	PMPNet	Pillar	53.1	45.4	79.7	33.6	47.1	43.1	76.5	40.7	19.9
	Innovation:	Introduces PMPNet and AST-GRU for spatial/temporal coherence in online 3D video detection.												
[254]	CBGS	2019	CVPR	Sparse CNN	Voxel	63.3	52.8	81.1	48.5	54.9	42.9	80.1	51.5	22.3
	Innovation:	Class-balanced sampling/augmentation and a balanced grouping head for severe class imbalance.												
[255]	PointPillars	2019	CVPR	2D CNN	Pillar	45.3	30.5	68.4	23.0	28.2	23.4	59.7	27.4	1.1
	Innovation:	Pillar-based PointNet encoder converting point clouds to dense pseudo-images for 2D CNN detection.												

Transformer-based voxel encoders. Instead of partitioning the voxel grid into smaller local groups, it serializes all non-empty voxels into a single sequence using a Hilbert Input Layer, which maps 3D voxel coordinates into a 1D order while preserving spatial locality. The backbone employs a Dual-scale SSM Block (DSB), wherein a forward branch processes the high-resolution sequence. In contrast, a backward branch operates on a downsampled BEV representation, extending the receptive field without quadratic complexity. An Implicit Window Partition (IWP) injects relative position embeddings for intra- and inter-window offsets, enabling the model to recover local context cues without explicit grouping. LION [205] replaces transformer-based voxel backbones with linear recurrent neural operators to model long-range dependencies across large voxel regions while avoiding quadratic complexity. Point clouds are voxelized and organized into fixed-size 3D windows, within which LION Blocks process voxel features hierarchically. Each block contains a pair of linear group RNN operators, applied along orthogonal spatial axes, to propagate information efficiently across extended spatial extents. Before this recurrent mixing, a 3D spatial descriptor based on submanifold convolutions enriches local geometric context to offset the positional information loss during voxel flattening. Within its hierarchy, the model reduces voxel resolution at selected stages, representing features at multiple scale, and giving them higher-level layers access.

Table 20 presents a comparative analysis of 3D object detection methods using LiDAR-based data from the VOD dataset. From Table 20, we provide a detailed discussion of three selected models: ELMAR [256] and MoRAL [257], which achieved the highest results in terms of $AP\%$ for entire area and driving corridor on VOD dataset; and L4DAR [258], recognized as one of the notable model in the context of 3D object detection.

Table 20: Comparative analysis of 3D object detection methods using the LiDAR-based VOD dataset.

						AP in the Entire Annotated Area (%)				**AP in the Driving Corridor (%)**			
Ref	**Method**	**Year**	**Venue**	**Backbone**	**Modality**	**Car**	**Pedestrian**	**Cyclist**	**mAP$_{3D}$**	**Car**	**Pedestrian**	**Cyclist**	**mAP$_{3D}$**
[256]	ELMAR	2025	ArXiv	DS+SA	Point	**76.41**	69.34	78.91	**74.89**	91.74	**81.35**	93.01	**88.70**
	Innovation:	Encodes radar-derived motion and aligns LiDAR–radar predictions via uncertainty modeling for cross-modal misalignment.											
[257]	MoRAL	2025	ArXiv	Sparse CNN	Pillar	71.23	**69.67**	79.01	73.30	90.91	78.90	**96.25**	88.68
	Innovation:	Motion-compensated multi-frame LiDAR-radar enhancement via motion-aware gating to mitigate misalignment artifacts.											
[259]	MutualForce	2025	ICASSP	CNN+RPN	Pillar	71.67	66.26	77.35	71.76	**92.31**	76.79	89.97	86.36
	Innovation:	Bidirectional radar–LiDAR feature guidance with velocity cues and LiDAR-radar enhancement for improved fusion detection.											
[204]	AS-Det	2025	AAAI	AS	Point	42.46	49.02	78.03	56.50	77.45	59.41	96.12	77.73
	Innovation:	Introduces active sampling and multi-scale center feature aggregation for adaptable single-stage 3D detection.											
[258]	L4DR	2024	AAAI	PointNet++	Pillar	69.10	66.20	**82.80**	72.70	90.80	76.10	95.50	87.47
	Innovation:	Two-stage LiDAR–radar fusion with denoising, pillar-based feature sharing, and multi-scale fusion for weather robustness.											
[260]	CM-FA	2024	ICRA	SA	Point	71.39	68.54	76.60	72.18	90.91	80.78	87.80	86.50
	Innovation:	Cross-modal hallucination with spatial/feature alignment enabling modality-agnostic training and single-sensor inference.											
[261]	RLNet	2024	ECCV	2D CNN	Voxel	70.88	69.43	78.12	72.81	90.82	78.71	91.67	87.07
	Innovation:	LiDAR–radar fusion with adaptive weighting, Doppler speed compensation, and modality dropout for robust detection.											
[262]	InterFusion	2022	IROS	CNN+RPN	Pillar	67.50	63.21	78.79	69.83	88.11	74.80	87.50	83.47
	Innovation:	Self-attention interaction aligning LiDAR/radar pillar features with radar noise reduction via pitch-angle correction.											
[250]	CenterPoint	2021	CVPR	VoxelNet	Point	32.74	38.00	65.42	45.42	62.01	48.18	84.98	65.66
	Innovation:	Reformulates 3D detection as center keypoint estimation, enabling anchor-free predictions with two-stage refinement.											
[255]	PointPillars	2019	CVPR	2D-CNN	Pillar	65.55	55.71	72.96	64.74	81.10	67.92	88.96	79.33
	Innovation:	Pillar-based PointNet encoder converting point clouds to dense pseudo-images for efficient end-to-end 3D detection.											

The ELMAR algorithm[256] is designed for single-frame operation, pairing LiDAR with 4D radar without any temporal accumulation. Its radar branch incorporates a Dynamic Motion-Aware Encoding module. It learns to classify entire objects as moving or static via a dedicated motion-aware loss, avoiding the unreliability of fixed velocity cutoffs. This produces motion descriptors that are stable enough to be used directly in detection. After each branch produces its initial set of bounding boxes, the Cross-Modal Uncertainty Alignment module attempts to match boxes between modalities, using geometric differences between matched pairs as a fine-grained measure of uncertainty. Considerable disagreements often cause predictions to automatically downweigh during training, directing the model towards cases where both sensors agree. This means the detector only captures the advantages of the radar's motion sensitivity but also uses the radar's consistency to overcome the LiDAR's weaknesses in challenging conditions.

The MoRAL framework [257] was developed with the goal of tackling the distortions that often result from combining multiple 4D radar frames, especially the misalignment of fast objects. Conventionally, points from earlier frames can stretch along the direction of motion, which creates elongated "tails" that distort shape and affect detection efficiency. MoRAL avoids this by embedding a Motion-aware Radar Encoder (MRE) at the front of the radar branch. This module uses a segmentation-based approach to distinguish moving points from static ones, then shifts only the moving points into alignment with a chosen reference frame using their

absolute radial velocity and capture time. While the radar branch undergoes this correction, the LiDAR input is processed independently through a sparse encoder. The two streams meet in the Motion Attention Gated Fusion (MAGF) module, which injects motion cues into the LiDAR feature space and applies a gating mask. That mask strengthens the signal from dynamic foreground regions while naturally dampening static background responses before the network predicts final bounding boxes. Meanwhile, L4DR pipeline [258] addresses a different challenge: the steep quality gap between LiDAR and radar data, and the fact that each sensor degrades differently in poor weather. Its processing begins with Foreground-Aware Denoising, which uses learned segmentation to strip radar scans of spurious returns before any fusion occurs. Cleaned radar data is fused early with raw LiDAR through the Multi-Modal Encoder. In this step, points are grouped into shared pillars, and each pillar allows a two-way exchange of information—radar's velocity and RCS values inform the LiDAR points in that cell. Meanwhile, LiDAR contributes its high-precision geometry and reflectivity back to radar. The fused data then passes into the Inter-modal and Intra-modal backbone, which maintains three streams: one for each modality and one combined. To prevent all three from carrying redundant patterns, a Multi-Scale Gated Fusion block filters each intra-modal stream based on cues from the inter-modal branch, enabling the network to emphasize whichever sensor is more trustworthy at that moment.

5.3. 2D–3D Fusion Strategies

2D–3D fusion methods combine complementary information from camera images and LiDAR point clouds to improve detection accuracy and robustness in autonomous driving scenarios. Camera data provides rich texture and color cues, which are valuable for recognizing objects' appearance and fine details. Meanwhile, LiDAR delivers precise geometric structure and depth information, enabling accurate distance estimation and spatial localization. By integrating these two modalities, fusion strategies can overcome the limitations of single-sensor approaches, such as poor depth perception in monocular vision or reduced semantic detail in LiDAR-only systems. These methods vary in how and when they combine features, with common approaches including early fusion at the raw data level, mid-level fusion of intermediate features, and late fusion at the decision stage. The choice of fusion scheme significantly impacts performance, balancing computational cost, information richness, and real-time feasibility. Recent advances in deep learning-based fusion architectures show that well-designed pipelines outperform single-modality methods, particularly under occlusion, low lighting, and adverse weather.

Figure 12 illustrates the general architecture of an early-fusion network, where multi-sensor data (LiDAR point clouds and RGB images) are combined at the raw data stage. This approach, also called data-level fusion, aligns sensor outputs into a shared representation space, by projecting LiDAR points onto the image plane or generating dense depth maps from point clouds. By fusing early information, this method preserves the full richness of the input data. It enables the network to learn low-level correlations between geometric structure and visual appearance. However, early fusion has some limitations, including high computational cost due to large input dimensionality, sensitivity to spatial and temporal calibration errors, and reduced robustness when one modality is degraded or missing. Despite these challenges, it is widely adopted in applications such as dense 3D object detection, semantic segmentation, and bird's-eye view mapping.

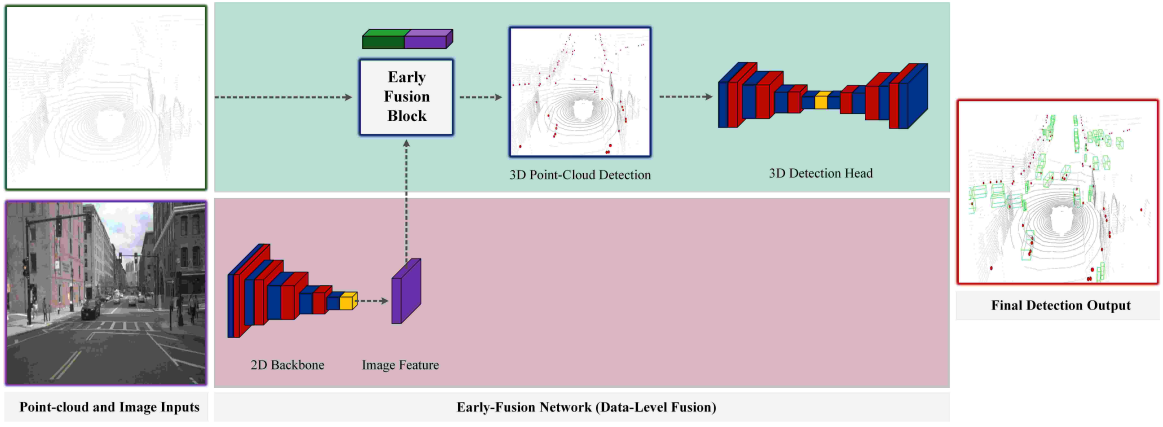

Figure 12: Overall framework of Early-Fusion 3D object detection in the context of autonomous driving systems.

Figure 13 represents the general architecture of a mid-fusion network, where LiDAR point clouds and RGB images are first processed independently through modality-specific backbones to extract intermediate feature representations before being combined. This approach, also known as feature-level fusion, merges modality-specific features, often through concatenation, element-wise operations, or attention-based mechanisms. Mid-fusion allows each backbone to specialize in processing its own modality, capturing modality-specific character-

istics such as the geometric precision of LiDAR and the rich semantic content of images. Compared to early fusion, mid-fusion is less sensitive to precise sensor calibration since the fusion occurs at a higher abstraction level, and it can be more computationally efficient by reducing the data dimensionality prior to fusion. However, it may lose some fine-grained cross-modal correlations in raw data and requires careful design to ensure feature compatibility across modalities. This strategy is widely used in AVs applications, such as 3D object detection, lane and road boundary detection, and BEV mapping.

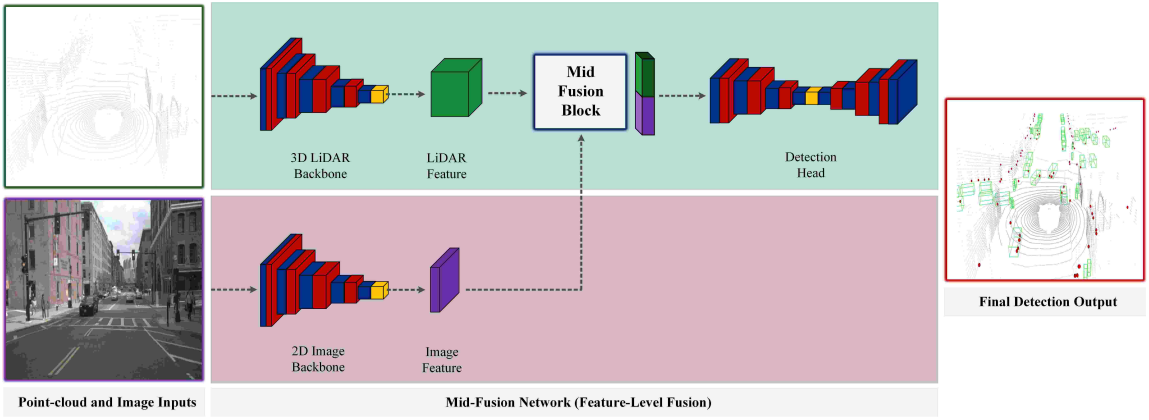

Figure 13: Overall framework of Mid-Fusion 3D object detection in the context of autonomous driving systems.

Figure 14 shows the general architecture of a late-fusion network, where LiDAR point clouds and RGB images are processed entirely separately through independent detection pipelines. The resulting outputs, such as bounding boxes, classification scores, or segmentation maps, are combined at the decision stage. This approach, also referred to as result-level fusion, integrates high-level predictions rather than raw data or intermediate features, typically using strategies such as weighted averaging, confidence-based selection, non-maximum suppression, across modalities, or rule-based decision logic. Late-fusion offers several advantages, including high modularity, robustness to partial sensor failure, and computational efficiency. However, late-fusion cannot leverage low- or mid-level cross-modal correlations, potentially limiting performance gains compared to early or mid-fusion approaches, and it relies heavily on the accuracy of individual modality-specific detectors. This strategy is often applied in safety-critical AV systems, where redundancy and fault tolerance are essential, such as in collision avoidance, multi-sensor tracking, and decision-making modules.

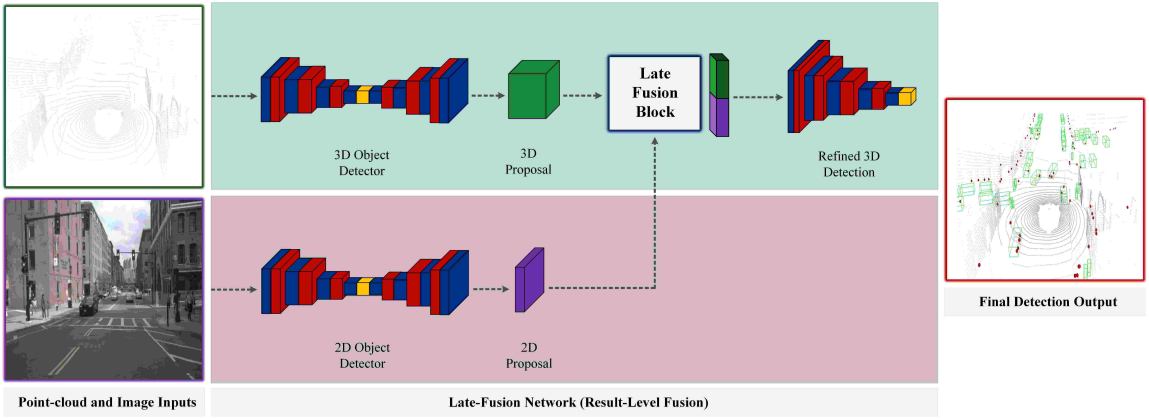

Figure 14: Overall framework of Late-Fusion 3D object detection in the context of autonomous driving systems.

In the following, we present all existing 2D–3D fusion algorithms for object detection in autonomous vehicles from 2017 to 2025. To ensure a consistent and fair comparison, we organize the results into two separate tables, as the performance of these methods is evaluated on different datasets. Table 21 provides a comparative analysis of fusion-based object detection methods using paired image–LiDAR data from the KITTI dataset, while Table 22 summarizes those evaluated on the NuScenes dataset. It should be mentioned that the values highlighted in red represent the best performance achieved for each evaluation metric across all listed methods.

Table 21: Comparative analysis of 2D–3D fusion detection methods on the paired image–LiDAR KITTI dataset.

				Backbone		AP_{BEV} Car			AP_{3D} Car			AP_{3D} Pedestrian			AP_{3D} Cyclist		
Ref	**Method**	**Year**	**Venue**	**Image**	**LiDAR**	**Easy**	**Moderate**	**Hard**	**Easy**	**Moderate**	**Hard**	**Easy**	**Moderate**	**Hard**	**Easy**	**Moderate**	**Hard**
[263]	ViKIENet	2025	CVPR	2D CNN	3D CNN	**96.31**	**93.73**	**91.64**	91.79	84.96	80.20	–	–	–	–	–	–
	Innovation:	Introduces VKI-based LiDAR–camera fusion using SKIS, VIFF, and VIRA to enhance efficiency and detection accuracy.															
[263]	ViKIENet-R	2025	CVPR	2D CNN	3D CNN	95.45	93.58	91.33	91.20	86.04	81.18	–	–	–	–	–	–
	Innovation:	Adds VIFF-R which includes rotationally equivariant features															
[264]	CLS-3D	2025	IEEE	DeepLabv3+	3D CNN	94.08	89.10	85.92	89.52	81.23	76.32	53.18	44.79	42.08	78.43	64.17	57.24
	Innovation:	Fuses LiDAR and image features via point-wise semantic augmentation with slot reweighting transformer for improved 3D detection.															
[265]	Point-Level	2025	MDPI	CNN+ECA	CNN+ECA	–	–	–	88.34	78.08	75.38	59.67	54.67	50.03	82.49	61.43	57.38
	Innovation:	ECA-enhanced PointPillars for LiDAR–camera fusion improving small-object detection and orientation.															
[266]	TinyFusionDet	2025	TCSVT	TBN	TBN	92.32	84.77	79.74	89.29	79.40	74.99	56.71	51.66	47.81	86.50	66.43	62.58
	Innovation:	HSPS, CMHA, and CMFI for hardware-efficient FPGA LiDAR–camera 3D detection.															
[267]	CSMNET	2024	TGRS	DeepLabv3+	C-FPS	90.16	85.18	80.51	82.05	72.64	67.10	48.95	42.15	39.44	82.86	67.86	60.90
	Innovation:	Class-guided sampling with PAT-based fusion to enhance small/distant object detection.															
[268]	LoGoNet	2023	CVPR	FPN	Voxel-RCNN	–	–	–	91.80	85.06	80.74	70.20	63.72	59.46	84.47	71.70	64.67
	Innovation:	Local-to-global fusion (GoF/LoF/FDA) for stronger multi-modal feature representation.															
[269]	VirConv-L	2023	CVPR	VirConvNet	VirConvNet	95.53	91.95	87.07	91.41	85.05	80.22	–	–	–	–	–	–
	Innovation:	StVD and NRConv reduce virtual point redundancy and suppress noise for efficient fusion.															
[269]	VirConv-T	2023	CVPR	VirConvNet	VirConvNet	96.11	92.65	89.69	**92.54**	86.25	81.24	73.32	66.93	60.38	90.04	73.90	69.06
	Innovation:	based on a transformed refinement scheme															
[269]	VirConv-S	2023	CVPR	VirConvNet	VirConvNet	95.99	93.52	90.38	92.48	**87.20**	**82.45**	–	–	–	–	–	–
	Innovation:	based on a pseudo-label framework															
[213]	TED-M	2023	AAAI	TeSpConv	TeSpConv	55.25	79.21	46.52	91.61	85.28	80.68	55.85	49.21	46.52	88.82	74.12	66.84
	Innovation:	Transformation-equivariant voxel features (TeSpConv), TeBEV and TiVoxel pooling.															
[270]	DFIM	2023	Elsevier	2D CNN	SA	92.61	88.69	85.77	88.36	81.37	76.71	49.91	41.97	39.27	76.99	59.12	52.97
	Innovation:	Dual-feature interaction with uncertainty-aware IoU optimization.															
[271]	MVMM	2023	IEEE	ResNet-50	3D CNN	92.17	88.70	85.47	87.59	78.87	73.78	47.54	40.49	38.36	77.82	64.81	58.79
	Innovation:	Point-cloud coloring with encoder–decoder and multi-view fusion for efficient single-stage 3D detection.															
[272]	SupFusion-v1	2023	ICCV	MVXNet	MVXNet	–	–	–	88.20	78.46	74.01	63.11	58.13	53.73	74.79	57.60	53.82
	Innovation:	Supervised fusion with Polar Sampling and deep fusion module for robust LiDAR–camera 3D detection.															
[272]	SupFusion-v2	2023	ICCV	PV-RCNN	PV-RCNN	–	–	–	92.02	85.20	82.94	74.29	66.67	61.38	91.00	**76.13**	**71.66**
	Innovation:	Implements the same supervised fusion strategy within a PV-RCNN backbone for hybrid point–voxel 3D detection.															
[273]	VPFNet	2022	IEEE	2D Backbone	Voxel-RCNN	93.02	91.86	86.94	91.02	83.21	78.20	71.47	63.59	56.85	87.77	71.99	66.14
	Innovation:	Two-stage voxel-level stereo–LiDAR fusion with mitigation of resolution mismatch and modality imbalance.															
[274]	SFD	2022	CVPR	CPConvs	Voxel-RCNN	95.64	91.85	86.83	91.73	84.76	77.92	72.94	66.69	61.59	**93.39**	72.95	67.26
	Innovation:	3D-GAF for fine-grained RoI fusion of raw/pseudo points; SynAugment and CPConv feature extraction.															
[247]	Focals Conv	2022	CVPR	2D CNN	3D CNN	95.45	91.51	91.21	90.55	82.28	77.59	–	–	–	–	–	–
	Innovation:	Learnable position-wise importance for sparse CNNs in LiDAR-only and multi-modal detection.															
[275]	VFF	2022	CVPR	ResNet-50	PV-RCNN	95.43	91.40	90.66	89.58	81.97	79.17	–	–	–	–	–	–
	Innovation:	Point-to-ray projection with ray-wise voxel fusion and aligned cross-modal augmentation.															
[275]	VFF	2022	CVPR	ResNet-50	Voxel-RCNN	95.65	91.75	91.39	89.50	82.09	79.29	73.26	65.11	60.03	89.40	73.12	69.86
	Innovation:	Optimizes fully voxel-based detection by embedding point-to-ray voxel field fusion to strengthen dense spatial reasoning and feature alignment.															
[276]	Graph-Vol	2022	ECCV	DLA-34	RPN	95.69	90.10	86.85	91.89	83.27	77.78	–	–	–	–	–	–
	Innovation:	Proposes DPA for adaptive point aggregation, RGP for robust RoI feature learning, and VFA for refinement-stage image–point fusion.															
[277]	CAT-Det	2022	CVPR	IT+CMT	PT+CMT	–	–	–	89.87	81.32	76.68	54.26	45.44	41.94	83.68	68.81	61.45
	Innovation:	Cross-Modal Transformer backbone with contrastive augmentation for 3D detection.															
[278]	MSL3D	2022	Elsevier	ResNet-18	RPN	–	–	–	87.27	81.15	76.56	52.16	43.38	38.80	76.74	62.27	56.20
	Innovation:	Aligns multi-modal receptive fields via 2D set abstraction; unified augmentation with auxiliary multi-sensor training.															
[279]	StructuralIF	2022	CVIU	SA-SSD	SA-SSD	91.78	88.38	85.67	87.15	80.69	76.26	50.80	42.42	38.35	72.54	56.39	49.28
	Innovation:	3D–2D consistent features and Lambertian MCR for attentional structural voxels.															
[280]	EPNet++	2022	IEEE	PointRCNN	PointRCNN	–	–	–	91.37	81.96	76.71	52.79	44.38	41.29	76.15	59.71	53.67
	Innovation:	CB-Fusion for cascade bidirectional interaction and MC loss for cross-modal confidence consistency.															
[281]	CLOCs-v1	2020	IROS	2D CNN	2D CNN	91.16	88.23	82.63	86.38	78.45	72.45	62.54	56.76	52.26	–	–	–
	Innovation:	Late fusion leveraging geometric/semantic consistency for efficient combination of pre-trained 2D/3D detectors.															
[281]	CLOCs-v2	2020	IROS	2D CNN	2D CNN	92.60	88.99	81.74	87.50	76.68	71.20	60.33	54.17	46.42	–	–	–
	Innovation:	Replaces SECOND with PointRCNN to leverage point-based 3D proposals within the late-fusion consistency framework.															
[281]	CLOCs-PVCas	2020	IROS	2D CNN	2D CNN	93.05	89.80	86.57	88.94	80.67	77.15	–	–	–	–	–	–
	Innovation:	Replaces SECOND with PV-RCNN to incorporate hybrid point–voxel 3D proposals within the late-fusion consistency framework.															
[282]	EPNet	2020	ECCV	2D CNN	SA + FP	94.22	88.47	83.69	89.81	79.28	74.59	52.79	44.38	41.29	76.15	59.71	53.67
	Innovation:	Point-wise correspondence weighting with CE loss enforcing classification–localization consistency.															
[283]	PointPainting-v1	2020	CVPR	DeepLabv3+	PointPillars	–	–	–	90.01	87.65	85.65	**77.25**	**72.41**	**68.53**	81.72	68.76	63.99
	Innovation:	Sequentially fuses semantic image segmentation with LiDAR by appending per-point semantic scores before PointPillars detection.															
[283]	PointPainting-v2	2020	CVPR	DeepLabv3+	VoxelNet	–	–	–	90.05	87.51	86.66	73.16	65.05	57.33	87.46	68.08	65.59
	Innovation:	Applies sequential semantic point painting prior to VoxelNet-based 3D detection.															
[283]	PointPainting-v3	2020	CVPR	DeepLabv3+	PointRCNN	92.13	87.39	82.72	90.19	87.64	86.71	72.65	66.06	61.24	86.33	73.69	70.17
	Innovation:	Applies sequential semantic point painting prior to PointRCNN-based 3D detection.															
[284]	AVOD-FPN	2018	IROS	VGG-16	VGG-16	88.53	83.79	77.90	81.94	71.88	66.38	50.80	42.81	40.88	64.00	52.18	46.61
	Innovation:	BEV–RGB feature pyramid with multimodal fusion RPN and geometry-constrained 3D encoding.															
[285]	PointFusion	2018	CVPR	ResNet-50	PointNet	–	–	–	77.92	63.00	53.27	33.36	28.04	23.38	49.34	29.42	26.98
	Innovation:	Dense fusion using 3D points as spatial anchors with learned scoring.															
[286]	F-PointNet -v1	2018	CVPR	FPN-Based	PointNet	87.28	77.09	67.90	80.62	64.70	56.07	50.88	41.55	38.04	69.36	53.50	52.88
	Innovation:	Frustum proposals from 2D detections enable efficient 3D PointNet processing.															
[286]	F-PointNet -v2	2018	CVPR	FPN-Based	PointNet++	88.70	84.00	75.33	81.20	70.39	62.19	51.21	44.89	40.23	71.96	56.77	50.39
	Innovation:	Extends F-PointNet with PointNet++ hierarchical feature learning to improve geometric detail capture in 3D detection.															
[287]	MV3D	2017	CVPR	VGG-16	VGG-16	86.02	76.90	68.49	71.29	62.68	56.56	39.48	33.69	31.51	–	–	–
	Innovation:	Multi-view region-based fusion combining BEV, FV, and RGB features for oriented 3D detection.															

From Table 21, we provide a detailed discussion of four selected models: VirConv [269], obtained the highest results in AP_{3D} *Car* evaluations; ViKIENet [263], which achieved the highest results in AP_{BEV} *Car* evaluations; PointPainting [283], got the highest results in AP_{3D} *Pedestrain* evaluations; and SupFusion [272], which attained the highest results in AP_{3D} *Cyclist* evaluations on the KITTI dataset. VirConv [269] is a multimodal 3D object detection framework that uses virtual points to strengthen LiDAR–RGB fusion while tackling two persistent issues in depth-completion-based augmentation: excessive point density and noise from inaccurate depth estimates. Its core component, the VirConv operator, combines Stochastic Voxel Discard (StVD) to filter out redundant near-range voxels while preserving geometry-rich distant points, and Noise-Resistant Submanifold Convolution (NRConv) to encode features jointly in 3D space and their 2D image projections, enabling effective suppression of boundary noise without eroding useful edge detail. From this foundation, three detector configurations were developed. VirConv-L applies an early-fusion scheme aimed at minimizing processing time. VirConv-T follows a late-fusion route and adds multi-transformation refinement to raise detection precision. VirConv-S extends the approach into a semi-supervised framework, using high-confidence pseudo-labels.

ViKIENet [263] is a multi-modal 3D object detection framework that replaces the dense, noise-prone virtual points of traditional LiDAR–camera fusion with a compact set of semantically enriched Virtual Key Instances (VKIs). Its architecture comprises three tightly integrated modules. The Semantic Key Instance Selection (SKIS) module generates VKIs by combining depth completion with semantic segmentation, converting only object-relevant pixels into 3D points augmented with class scores. These VKIs and raw LiDAR points are voxelized in separate branches, processed by sparse 3D convolutions, and aligned through the Virtual-Instance-to-Real Attention (VIRA) module, which refines VKI features using precise LiDAR depth cues. The Virtual Instance Focused Fusion (VIFF) module then performs two-stage fusion: BEV-level integration via channel and spatial attention to enhance proposal generation, and RoI-level bi-directional cross-attention to combine VKI semantic richness with LiDAR geometric accuracy for final bounding box refinement. This design reduces computation, suppresses depth-induced noise, and preserves discriminative spatial–semantic cues.

PointPainting [283] is a sequential fusion mechanism that decorates every LiDAR point with image semantics before any LiDAR-only detector processes it. An image segmentation network produces per-pixel class scores; LiDAR points are projected into that score map and "painted" by appending the corresponding class-probability vector, yielding an augmented point cloud that remains compatible with arbitrary LiDAR backbones (e.g., PointPillars, VoxelNet/SECOND, PointRCNN). The method's key property is modularity: fusion is realized as feature augmentation at the point level, enabling independent advancement of the segmentation and detection stacks and practical pipelining for low-latency deployment. SupFusion [272] is a training strategy for LiDAR–camera detectors that introduces auxiliary feature-level supervision from an assistant model trained on densified LiDAR via Polar Sampling. During training, standard 3D/2D backbones and a deep fusion module (stacked MLPs with dynamic fusion blocks) produce fusion features that are optimized not only by the detection loss but also to mimic high-quality features generated by the assistant on enhanced data; this supervision strengthens fusion layers and the upstream encoders without incurring inference overhead. The pipeline thus pairs decision-level supervision with auxiliary alignment of fusion representations, improving robustness of multi-modal feature formation and downstream detection.

Table 22 presents an analysis of 2D–3D fusion detection methods on the paired image–LiDAR NuScenes dataset. We provide a detailed discussion of four selected models: MV2DFusion [288], CLS-3D [264], and ECL3D [289], which achieved the highest results among all the methods in terms of mAP and NDS metrics on the NuScenes dataset; and TransFusion [248], recognized as one of the popular model in the context of object detection. MV2DFusion [288] employs separate 2D and 3D detection backbones, ResNet-50 with a conventional image detector and FSDv2 for LiDAR, to independently produce candidate object proposals for each modality. These proposals are transformed into modality-specific object queries by dedicated query generation modules. A six-layer transformer-based fusion decoder then processes the queries, first modeling relationships within each set via self-attention, and subsequently enabling feature exchange across modalities through cross-attention. A query calibration stage refines the fused representations before final prediction. By focusing on proposal-level fusion, this design maintains the semantic integrity to avoids the pitfalls of dense feature concatenations.

CLS-3D [264] integrates LiDAR and RGB information through a unified backbone in which each modality has its encoder but contributes to a shared feature space for joint reasoning. The image pathway applies DeepLabv3+ to obtain pixel-wise semantic probability maps geometrically aligned to the LiDAR frame. Then, each 3D point is annotated with the semantic probabilities of its corresponding image pixel, creating an enriched point set that combines geometric and class-level cues. This representation is processed by a transformer module equipped with a slot-based reweighting mechanism, allowing the network to emphasise channels that carry the most relevant spatial or temporal information. For final detection, the network employs an I3C-IoU loss considering box overlap, centre displacement, and scale, producing more precise localisation in crowded or visually challenging driving environments.

Table 22: Comparative analysis of 2D–3D fusion detection methods on the paired image–LiDAR NuScenes dataset.

Ref	Method	Year	Venue	Image Backbone	LiDAR Backbone	mAP	NDS	mATE	mASE	mAOE	mAVE	mAAE
[289]	ECL3D	2025	T-ITS	ResNet-50	VoxelNet+	**72.8**	**75.2**	**0.252**	**0.235**	**0.326**	**0.184**	**0.127**
	Innovation:	Introduces dual-scale depth supervision and deformable cross-attention for enhanced BEV feature fusion without added complexity.										
[264]	CLS-3D	2025	IEEE	DeepLabv3+	3D backbone	**75.1**	**76.8**	**0.273**	**0.244**	**0.302**	**0.214**	**0.108**
	Innovation:	Fuses LiDAR and image features via point-wise semantic augmentation with reweighted transformer for enhanced 3D detection.										
[290]	DAL	2024	ECCV	ResNet-50	VoxelNet+	71.5	74.0	0.253	0.239	0.334	0.174	0.120
	Innovation:	"Detecting As Labeling" paradigm with elegant training pipeline and instance-level velocity augmentation for robust fusion.										
[288]	MV2DFusion-v1	2024	ArXiv	ResNet-50	FSDv2	74.5	76.7	0.245	0.229	0.269	0.119	0.115
	Innovation:	Depth-guided multi-view 2D fusion aligning image features across views before BEV projection for robust image–LiDAR fusion.										
[288]	MV2DFusion-v2	2024	ArXiv	ResNet-50	FSDv2	**77.9**	**78.8**	**0.237**	**0.226**	**0.247**	**0.192**	**0.119**
	Innovation:	MV2DFusion with model ensemble and test-time augmentation for further accuracy gains.										
[291]	BEVFusion	2023	ICRA	Swin-Tiny	VoxelNet	70.2	72.9	0.261	0.239	0.329	0.260	0.134
	Innovation:	Performs sensor fusion in BEV space instead of LiDAR space, offering a simpler yet stronger fusion paradigm.										
[292]	SparseFusion	2023	ICCV	ResNet-50	VoxelNet+	72.0	73.8	0.258	0.243	0.329	0.265	0.131
	Innovation:	Instance-level sparse feature fusion with cross-modality transfer for fast, accurate LiDAR–camera 3D detection.										
[293]	CMT	2023	ICCV	VOVNet	VoxelNet	72.0	74.1	0.279	0.235	0.308	0.259	0.112
	Innovation:	End-to-end multi-modal 3D detection with direct 3D position encoding into tokens, avoiding grid sampling or voxel pooling.										
[294]	FUTR3D	2023	CVPR	VoVNet	VoxelNet	69.4	72.1	0.284	0.241	0.310	0.300	0.120
	Innovation:	Modality-agnostic feature sampler enabling end-to-end fusion across arbitrary sensor configurations and combinations.										
[295]	FocalFormer3D	2023	ICCV	ResNet-50	VoxelNet	72.9	75.0	0.250	0.242	0.328	0.226	0.126
	Innovation:	Hard Instance Probing to identify false negatives and enhance BEV features for improved 3D detection recall.										
[242]	LargeKernel3D-F	2023	CVPR	SW-LK	SW-LK	71.1	74.2	0.236	0.228	0.298	0.241	0.131
	Innovation:	Uses spatial-wise partition convolution and position embedding to realize efficient, very large 3D kernels in sparse CNNs.										
[296]	FusionFormer	2023	ArXiv	VoV-99	VoxelNet	72.6	75.1	0.267	0.236	0.328	0.226	0.105
	Innovation:	Transformer framework with uniform sampling to fuse modalities without pre-BEV voxel compression.										
[297]	EA-LSS	2023	ArXiv	2D Backbone	N/A	72.2	74.4	0.247	0.237	0.304	0.250	0.133
	Innovation:	Edge-aware, fine-grained depth modules improving accuracy and alignment in LSS-based BEV detection.										
[186]	UVTR-M	2022	NeurIPS	ResNet-101	VoxelNet+	67.1	71.1	0.306	0.245	0.351	0.225	0.124
	Innovation:	Cross-modality interaction and a transformer decoder										
[248]	TrasFusion	2022	CVPR	ResNet-50	VoxelNet	68.9	71.7	0.259	0.243	0.329	0.288	0.127
	Innovation:	Uses a transformer decoder to fuse LiDAR BEV features.										
[298]	FusionPainting	2021	ITSC	ResNet-50	VoxelNet+	66.3	70.4	0.256	0.236	0.346	0.274	0.132
	Innovation:	Semantic-level fusion with voxel-level attention to integrate multi-modal context features for improved 3D detection.										
[299]	MVP	2021	NeurIPS	DLA-34	VoxelNet+	66.4	70.5	0.263	0.238	0.231	0.313	0.134
	Innovation:	Virtual points from images to balance LiDAR density across distances and enhance multi-modal 3D detection accuracy.										
[300]	PointAugmenting	2021	CVPR	DLA-34	VoxelNet+	66.8	71.1	0.253	0.234	0.345	0.266	0.123
	Innovation:	Uses 2D detector features with cross-modal data augmentation to enhance LiDAR–camera 3D object detection.										
[301]	3D-CVF	2020	ECCV	ResNet-18	3D Sparse CNN	67.5	57.8	0.280	0.246	0.367	0.239	0.129
	Innovation:	Auto-calibrated cross-view projection with adaptive gated fusion and 3D RoI-based refinement for robust camera–LiDAR detection.										

ECL3D [289] follows a dense-to-sparse fusion pipeline with parallel image and LiDAR streams producing BEV features via ResNet-50+FPN with an LSS view transformer and a VoxelNet+SECOND backbone, respectively. The dense stage fuses BEV features to generate a category heatmap, from which top-K proposals are selected. In the sparse stage, dual-modal proposal features are extracted for final regression and classification. Two core enhancements improve accuracy: a dual-scale depth-semantic supervision module in the image branch augments LSS with an auxiliary high-resolution depth prediction branch without increasing inference cost, and an Instance BEV Feature Enhancement (IBFE) module applies deformable cross-attention between LiDAR and image BEVs to align semantic and geometric cues before proposal refinement. TransFusion [248] extracts LiDAR BEV features via a VoxelNet backbone and image features via a ResNet-50 image backbone, then applies a two-stage transformer decoder as the detection head. The first stage decodes sparse, category-aware object queries from LiDAR BEV features into initial bounding boxes. The second stage performs LiDAR–camera fusion using Spatially Modulated Cross-Attention (SMCA), which imposes a locality bias by focusing on relevant image regions for each query. An image-guided query initialization pathway seeds queries using 2D features to help detect small objects under sparse LiDAR sampling. This strategy improves robustness against sensor misalignment and degraded image conditions.

To provide a conclusive comparison and clearer interpretation of the results, Figure 15 presents a performance comparison of the top three algorithms from each detection category (2D, 3D, and 2D–3D fusion), evaluated across four key metrics: AP_{BEV} *Car*, AP_{3D} *Car*, AP_{3D} *Pedestrian*, and AP_{3D} *Cyclist*. The results highlight the strengths and weaknesses of each category, illustrating how different detection paradigms excel in specific object classes or evaluation perspectives. This comparative view not only identifies the highest-performing methods but also provides insight into the trade-offs between 2D, 3D, and fusion-based approaches.

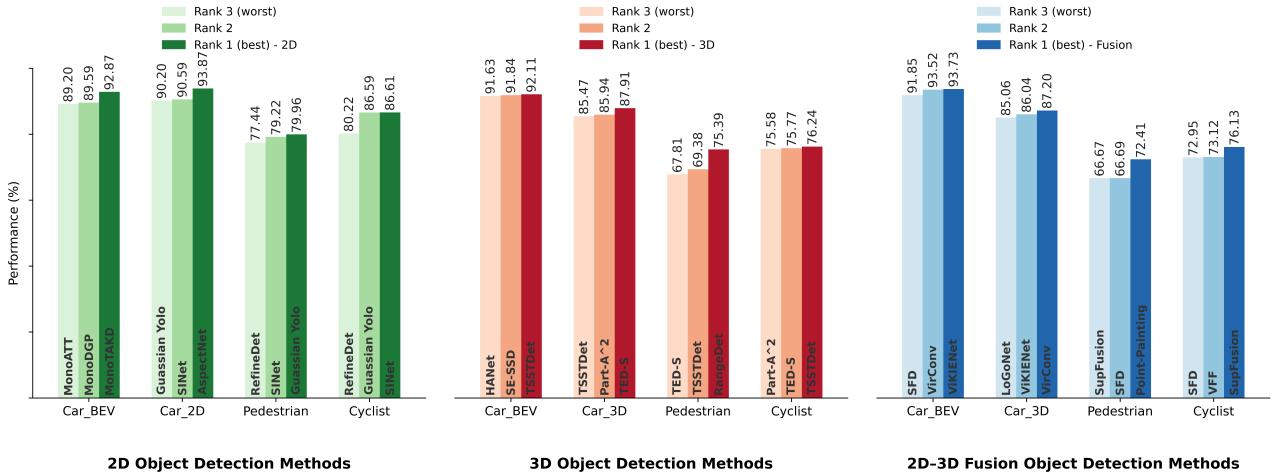

Figure 15: Performance comparison of the top three algorithms from each detection category (2D, 3D, and 2D–3D fusion object detection) across four evaluation metrics: AP_{BEV} *Car*, AP_{3D} *Car*, AP_{3D} *Pedestrian*, and AP_{3D} *Cyclist* on the KITTI dataset.

The comparative analysis across 2D, 3D, and 2D–3D fusion object detection methods highlights modality-driven performance trends. The 2D detectors, evaluated on a dedicated 2D dataset, achieve strong accuracy in the image domain, particularly for the AP_{2D} *Car* class, where the top-ranked model surpasses 93% performance. However, their capabilities are inherently constrained in reconstructing depth and precise spatial localization due to the absence of explicit 3D information. On the other hand, LiDAR-based 3D detectors directly operate in 3D space, delivering competitive spatial accuracy, especially in AP_{BEV} *Car* and AP_{3D} *Car* metrics, but often exhibit lower performance in detecting small or partially occluded objects, such as pedestrians, where point cloud sparsity and occlusion effects degrade recall. Fusion-based methods leverage complementary strengths: images' dense texture and semantic cues enhance object classification, while LiDAR provides accurate geometry and scale estimation. They achieved the highest AP_{BEV} *Car* (93.73%) and AP_{3D} *Car* (87.20%) scores, a result attributed to the combined benefits of depth perception and classification from cross-modal cues.

Despite the overall advantage, the fusion improvements are marginal in pedestrian detection. This limited gain is likely attributable to multiple factors: (1) the low point density for pedestrians in LiDAR, especially at range, reduces the geometric detail available for fusion; (2) the relatively small pixel footprint of pedestrians in images limits the discriminative value of texture cues; and (3) misalignment errors from calibration and synchronization between modalities have a greater proportional effect on small-scale targets, diminishing fusion efficacy. Furthermore, since the 2D detectors were trained on a purely 2D dataset, their high pedestrian and cyclist detection scores in the image domain do not fully translate into 3D accuracy when projected or fused, given that depth cues must be inferred indirectly. Consequently, while fusion consistently outperforms single-modality approaches in vehicle-related tasks, the marginal pedestrian improvement underscores the persistent challenge of small-object 3D detection, even under multimodal paradigms.

Additionally, Figure 16 presents a performance comparison of the top three algorithms from each detection category (2D, 3D, and 2D–3D fusion), evaluated across the mAP metric. The results highlight the strengths and weaknesses of each category, illustrating how different detection paradigms excel in specific object classes or evaluation perspectives. This comparative view not only identifies the highest-performing methods within each category but also provides insight into the trade-offs between 2D, 3D, and fusion-based approaches for AV perception. On the nuScenes benchmark, the mean Average Precision (mAP) values show a clear separation between fusion-based, LiDAR-only, and camera-only detection strategies. The highest score is achieved by MV2DFusion at 77.90%, ahead of the best LiDAR method, Voxel Mamba (69.00%), and the leading camera-based approach, CorrBEV-SP (56.20%). This margin reflects the benefit of combining complementary sensing LiDAR's precise geometry and range measurements with the rich semantic detail from cameras, allowing fusion models to localize objects better and distinguish classes under varied conditions. LiDAR-only models such as SAFDNet and LION remain strong in spatial reasoning but can miss fine visual cues. At the same time, camera-based methods, even when adapted to a BEV representation, are limited by their indirect depth estimation.

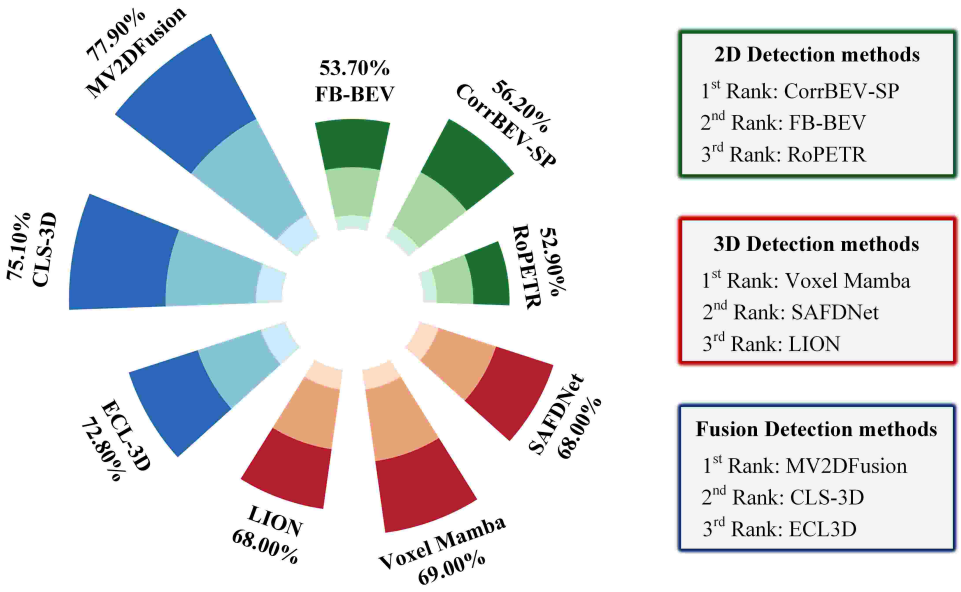

Figure 16: Performance comparison of the top three algorithms from each detection category (2D, 3D, and 2D–3D fusion object detection) across mAP evaluation metric on the NuScenes dataset.

5.4. LLMs/VLMs-based Methods

LLMs are being recognized as powerful tools for autonomous driving systems, since they facilitate semantic understanding, contextual reasoning and support the extraction of useful information from language data. Their capabilities have been demonstrated across various aspects of autonomous driving, from interpreting sensor data and generating scene summaries to assisting in trajectory planning and decision-making under dynamic traffic conditions. LLMs take a fundamentally different approach from traditional computer vision systems. While conventional CV models depend entirely on what they can "see" in images, LLMs work with human language. Meanwhile, VLMs are designed to understand and process both images and text together, allowing for more complete and meaningful interpretation of complex driving environments and scenarios. Unlike traditional CV models that only analyze image pixels or LLMs that focus on text, VLMs handle both at the same time. They can describe images, answer questions based on what they "see," and retrieve relevant visuals given a text query [302].

Figure 17 demonstrates an overview of VLM systems tailored for autonomous driving applications. The framework begins with both image and text inputs, which are first processed by image and text encoders to extract high-level semantic features. These features are then integrated through a multimodal fusion module, enabling the model to align and reason across visual and linguistic modalities. The fused representation is subsequently passed through image and text decoders, depending on the task, to generate meaningful task-specific outputs such as object detection, scene captioning, trajectory prediction, or semantic reasoning. At the heart of this architecture, VLMs enable key capabilities, including perception, interaction, prediction, reasoning, alignment, and generation, bridging the gap between low-level sensor data and high-level decision-making.

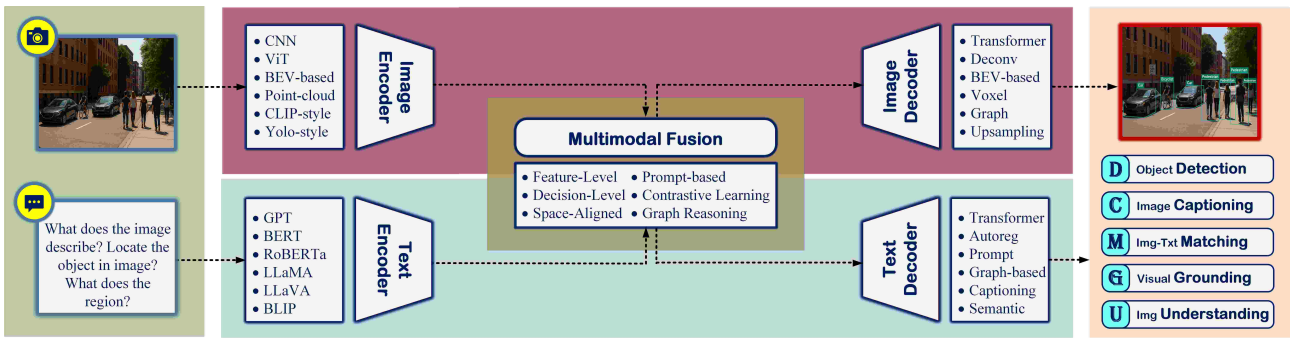

Figure 17: Overall framework of the VLMs in the context of autonomous driving systems.

To delve deeper into the context of LLMs and VLMs tailored for AVs, we now focus on recent methods explicitly developed for key AV-related tasks such as perception, reasoning, captioning, trajectory prediction,

and scene understanding. Tables 23 and 24 present a comprehensive overview of the models, highlighting their core architectures, input modalities, evaluation metrics, and specific innovations tailored for AV systems.

Table 23: Comparative analysis of LLM-based models for autonomous driving perception, reasoning, captioning, and scene understanding, highlighting their architectures, modalities, evaluation metrics, and key innovations.

Ref	Year	Name	Dataset	Architecture	Modalities	Metrics	Task
[303]	2025	LiDAR-LLM	NuScenes-QA	Encoder	Text, LiDAR	Accuracy	Reasoning
Innovation: Reframes 3D LiDAR scene understanding as a language modeling task using a position-aware transformer.							
[304]	2025	BEV-TSR	NuScenes	Encoder	Text, Vision	Accuracy	Reasoning
Innovation: Proposes BEV-TSR for text-to-BEV scene retrieval using LLMs and shared cross-modal embeddings.							
[305]	2025	LC-LLM	HighD	Encoder	Text	F1, Accuracy	Reasoning
Innovation: Introduces an LLM-based model for lane change prediction with interpretable chain-of-thought reasoning.							
[306]	2025	Traj-LLM	NuScenes	GPT-2, LoRA	Text	minADE, MR	Reasoning
Innovation: LLM-based trajectory predictor using lane-aware probabilistic learning and no prompt engineering.							
[307]	2024	DriveGPT4	BDD-X	Multimodal	Text, Video	BLEU4	Reasoning
Innovation: An interpretable end-to-end AV system using LLMs with multi-frame video and textual reasoning.							
[308]	2024	LLaVA1.5	VQA-v2	Multimodal	Text, Video	Accuracy, BLEU	Reasoning
Innovation: Demonstrates powerful LLM design with simple LLaVA modifications, achieving high performance.							
[309]	2024	GPT-3.5	Generated	Multimodal	Text	Accuracy	Captioning
Innovation: Proposes an object-level LLM that fuses numeric vectors with language for context understanding.							
[310]	2023	GAIA-1	Generated	Encoder-Decoder	Text, Video	Cross-entropy	Understanding
Innovation: Introduces a generative world model predicting future driving scenarios using video, text, and actions.							
[311]	2023	GPT-3	Generated	Multimodal	Text, Vision	CLIPort	Reasoning
Innovation: Introduces an LLM-based monitor to detect semantic anomalies in AVs through human-like reasoning.							
[312]	2023	DriveMLM	CARLA Town05	Multimodal	LiDAR, Vision	F1, BLEU4	Reasoning
Innovation: Bridges LLM decisions with vehicle control for closed-loop autonomous driving using multimodal inputs.							
[313]	2023	ADAPT	Generated	Multimodal	Text, Video	BLEU-1	Reasoning
Innovation: Introduces a real-time driving captioner that explains vehicle actions and decisions in natural language.							
[314]	2023	GPT-DRIVER	NuScenes	Multimodal	LiDAR, Vision	L2	Reasoning
Innovation: Reformulates motion planning as a language modeling task using GPT-3.5 to generate precise trajectories.							

In [303], the authors use LLMs to interpret sparse outdoor 3D LiDAR data. The proposed LiDAR-LLM reformulates 3D scene understanding tasks into a language modeling paradigm, such as captioning, grounding, and question answering. The authors design a three-stage training pipeline involving cross-modal alignment, object-centric perception, and high-level instruction tuning. The Position-Aware Transformer is central to the model, which integrates bird's-eye view positional embeddings to align 3D LiDAR features with LLM token space. They also construct large-scale synthetic datasets (420K captioning and 280K grounding pairs) using nuScenes and show that LiDAR-LLM outperforms 2D vision-language baselines in multiple 3D tasks, achieving strong results on both their synthetic benchmarks and real-world datasets like NuScenes-QA. This work marks a significant step toward general-purpose, instruction-following LLMs for 3D scene understanding in AVs. The method proposed in the [306] proposes a trajectory prediction in autonomous driving using pre-trained Large Language Models without explicit prompt engineering. The model captures high-level semantic interactions and scene context by encoding spatial-temporal features of agents and traffic scenes into LLM formats. It incorporates a novel lane-aware probabilistic learning module, powered by the Mamba architecture, to simulate human-like attention to lane structures. Moreover, a multi-modal Laplace decoder generates diverse and scene-compliant future trajectories. The results show that this method outperforms state-of-the-art baselines across key metrics, including in few-shot settings. It highlights its robustness, generalizability, and the untapped potential of LLMs for structured prediction in complex driving environments. [78] presents a method that integrates VLMs into end-to-end autonomous driving by modeling human-like reasoning through Graph Visual Question Answering. This method captures multistep, multiobject dependencies by structuring perception, prediction, and planning tasks as interconnected Question Answering pairs in a graph. Moreover, the authors propose two datasets, DriveLM-nuScenes and DriveLM-CARLA, and a baseline model, DriveLM-Agent, which combines graph prompting with trajectory tokenization to repurpose general VLMs for driving tasks. Results demonstrate that this method performs well with driving-specific models and excels in zero-shot generalization, especially with unseen sensor configurations. This work highlights Graph VQA as a promising approach for enabling explainable, generalizable, and interactive AV system.

Table 24: Comparative analysis of VLM-based models for autonomous driving perception, reasoning, captioning, and scene understanding, highlighting their architectures, modalities, evaluation metrics, and key innovations.

Ref	Year	Name	Backbone	Dataset	Modalities	Metrics	Task
[315]	2025	DynRsl-VLM	ViT+ResNet+Transformer	NuInstruct	Text, Image	Accuracy, BLEU	Reasoning
Innovation: Using dynamic resolution and efficient image-text alignment to preserve crucial visual details for driving VQA.							
[316]	2025	AV-VLM	GPT-4V	Generated	Text, Vision, LiDAR	Accuracy, Latency	Reasoning
Innovation: Introduces a real-world test platform for evaluating VLM-driven autonomous systems under controlled, diverse driving scenarios.							
[317]	2025	VLC Fusion	BEVFusion + CLIP	Waymo, ATR	Text, Vision, LiDAR	IR, mAP	Detection
Innovation: Using VLMs to adaptively weight sensor modalities based on environmental context cues for robust multimodal perception tasks.							
[318]	2025	COOOL	LLaMA-2 + CLIP	COOOLER	Video, Text	BESM, SAM	Captioning
Innovation: Introduces a multimodal VLM-LLM pipeline with CLIP for zero-shot hazard detection and explanation in traffic scenes.							
[319]	2025	NuPrompt	Transformer	NuPrompt	Text, Vision	AMOTA, ADE	Prediction
Innovation: An object-centric language prompt benchmark for 3D multi-view driving scenes with trajectory prediction.							
[320]	2024	Llama-Adapter	Multimodal-7B	Generated	Text, Vision	Accuracy, BLEU	Reasoning
Innovation: Extends VLMs with YOLOS-based object detection and camera ID-separators for improved multi-view scene understanding.							
[321]	2024	VLM-RL	LLaMA+CLIP	Generated	Vision, Text	AS, RC	Reward Design
Innovation: Introduces VLM-RL, using language goals as semantic rewards to replace manual reward engineering in RL-based driving.							
[322]	2024	VLM-AD	LLaMA-2	NuScenes	Vision, Text	L2 Error, Collision Rate	Reasoning
Innovation: Proposes VLM-AD, using VLMs as teachers to inject reasoning supervision into E2E driving without requiring them at inference.							
[323]	2024	EM-VLM4AD	Multiple	DriveLM	Image, Text	BLEU-4, METEOR	Reasoning
Innovation: Presents a lightweight multi-frame VLM optimized for real-time autonomous driving with low memory and fast inference.							
[78]	2023	DriveLM-Agent	BLIP-2	NuScenes	Text, Image	Accuracy	Reasoning
Innovation: Introduces Graph VQA to enable multi-step, graph-structured reasoning in end-to-end driving via VLMs.							
[324]	2023	BLIP-2	LLaMA-2-7B	COCO, VQAv2	Text, Vision	Accuracy, BLEU	Captioning
Innovation: A two-stage strategy using frozen image and language models with a lightweight transformer for efficient vision-language pretraining.							
[325]	2024	BEV-CLIP	CLIP+Llama2	NuScenes	Text,	BEV Accuracy,	Captioning
Innovation: A retrieval method using LLMs and knowledge graphs for rich semantic scene matching and contextual reasoning capabilities.							
[326]	2024	Talk2BEV	BEVFormer+GPT-2	NuScenes	Text, Image, LiDAR	Accuracy, IoU	Reasoning
Innovation: Introduces a LVLM interface for BEV maps enabling zero-shot visual and spatial reasoning without BEV-specific training.							
[327]	2023	MSSG	GPT-2	NuScenes	LiDAR, Text	BEV AP, 3D AP	Reasoning
Innovation: Proposes LiDAR Grounding with MSSG for efficient 3D language grounding in point clouds using joint LiDAR-language learning.							
[328]	2023	RMOT	Swin-L+BERT	KITTI	Vision	HOTA, DetA,	Tracking
Innovation: A framework for tracking multiple objects in video using language-guided expressions and contextual scene-level understanding.							
[329]	2023	CLIP2Scene	CLIP+PointNet++	NuScenes, ScanNet	Text, Vision	mIoU, Accuracy	Reasoning
Innovation: A framework to transfer CLIP knowledge to 3D scene understanding via semantic-driven contrastive learning.							
[330]	2023	CLIP	ViT+Transformer	Waymo	Text, LiDAR	3D AP Detection,	Reasoning
Innovation: Proposes an auto-labeling pipeline for open-set 3D perception using motion cues and vision-language distillation without human labels.							

Wu et al. [328] present a vision language task called Referring Multi-Object Referring (RMOT), which enables queries in natural language to refer to and track multiple objects over time in video sequences. RMOT supports flexible object cardinality and dynamic temporal variations in object status. Authors develop a benchmark, Refer-KITTI, based on the KITTI dataset, providing high-quality annotations with low labeling cost and supporting multi-object, temporally-aware queries. Additionally, their proposed TransRMOT uses an early-fusion cross-modal encoder and a query-decoupled decoder for simultaneous detection and tracking. TransRMOT performs better than CNN-based and Transformer-based baselines on the Refer-KITTI benchmark, highlighting the RMOT formulation's effectiveness and the proposed architecture for real-world, language-guided multi-object tracking. The authors in [319] propose a large-scale benchmark to combine natural language prompts into 3D perception for autonomous driving. NuPrompt is based on the nuScenes dataset, which contains more than 40,000 finely annotated language-object pairs, with each prompt referring to multiple object trajectories across 3D, multiview, and temporal frames. For simplicity, the authors define a new task for predicting object motion based on descriptive prompts and propose a baseline model, PromptTrack, incorporating a novel prompt reasoning module into a transformer-based tracking architecture. The results show that PromptTrack outperforms several strong baselines and validates the effectiveness of prompt-guided object tracking and forecasting. Their method is the foundation for research at the intersection of vision-language modeling and AV systems.

Chen et al. [329] propose a framework, Contrastive Language Image Pretraining (CLIP) models, for 3D scene understanding tasks, particularly for semantic segmentation of point clouds. They address the challenges of limited 3D annotations and domain mismatch between 2D and 3D data. Also, the authors propose a Semantic-driven Cross-modal Contrastive Learning strategy combining semantic consistency and spatial-temporal consistency. This method enables effective annotation-free segmentation, achieving state-of-the-art results on datasets like nuScenes and ScanNet, and also shows better performance in few-shot learning scenarios. This work is among the first to bridge 2D vision language pre-training and 3D scene understanding through cross-modal contrastive learning. Taparia et al. [317] present a fusion method that uses VLMs to improve object

detection under various environmental conditions. Some fusion methods often apply static fusion rules and have challenges in unseen scenarios. At the same time, VLC Fusion addresses this limitation by conditioning the fusion process on environmental context extracted using large VLMs such as GPT-4o. The authors introduce an automated pipeline to extract and query application-specific ecological cues, which are then used to modulate multimodal feature fusion via Feature-wise Linear Modulation (FiLM). Experiments on Waymo Open (RGB + LiDAR) and ATR (visible + infrared) datasets show that VLC Fusion outperforms other baselines.

6. Conclusion

This survey has provided a comprehensive and structured review of object detection in autonomous vehicles, bridging the domains of sensor technologies, benchmark datasets, and detection methodologies. By systematically analyzing 2D camera-based, 3D LiDAR-based, 2D–3D fusion, and emerging LLM/VLM-based approaches, we have synthesized their strengths, limitations, and optimal use cases of each method. Our unique dataset categorization, covering ego-vehicle, roadside, and cooperative perception datasets, offers a clear framework for evaluating and selecting suitable datasets based on application needs and purposes. Moreover, this work outlines the evolving landscape of AV perception by integrating insights from traditional CV pipelines and recent advances such as Vision Transformers and multimodal reasoning. The results underscore that robust object detection is inherently multimodal, requiring balanced trade-offs between computational efficiency, environmental adaptability, and semantic richness. Ultimately, this survey serves as a reference point for both academic and industry researchers seeking to design, benchmark, and deploy next-generation AV perception systems.

Future research in AV object detection should explore dynamic, context-aware sensor fusion pipelines that adaptively reweight multimodal inputs (camera, LiDAR, Radar) in real time based on driving context, environmental conditions, and system uncertainty, which is still largely unexplored in practice. Integrating foundation models trained on massive, cross-domain multimodal datasets could enable AVs to reason about rare and unseen scenarios beyond current benchmark coverage. Another promising yet underdeveloped direction is cross-vehicle collaborative perception enhanced by LLM/VLM reasoning, where vehicles exchange compressed semantic representations instead of raw sensor data to reduce bandwidth while preserving scene understanding. In addition, simulation-to-reality domain adaptation leveraging generative models and neural rendering could close the performance gap between synthetic and real-world data. Finally, a critical but underexplored frontier lies in uncertainty-aware perception systems that can self-assess detection reliability and dynamically adjust planning and control strategies, moving toward safer, more interpretable, and trustworthy AV decision-making.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This material is based upon the work supported by the National Science Foundation (NSF) under Grant Numbers 2008784 and 2204721.

References

[1] B. B. ELallid, N. Benamar, M. Bagaa, N. Mrani, Secure and efficient vehicle control of autonomous vehicles using federated deep reinforcement learning, Applied Soft Computing (2025) 113924.

[2] A. Khosravian, M. Masih-Tehrani, A. Amirkhani, S. Ebrahimi-Nejad, Robust autonomous vehicle control by leveraging multi-stage mpc and quantized cnn in hil framework, Applied Soft Computing 162 (2024) 111802.

[3] H. Jiang, K. Xia, Y. Zhao, Z. Yao, Y. Jiang, Z. He, Environmental impacts and emission reduction methods of vehicles equipped with driving automation systems: An operational-level review, Transportation Research Part C: Emerging Technologies 173 (2025) 104996.

[4] K. Grosse, A. Alahi, A qualitative ai security risk assessment of autonomous vehicles, Transportation Research Part C: Emerging Technologies 169 (2024) 104797.

[5] K. Wang, C. Shen, X. Li, J. Lu, Uncertainty quantification for safe and reliable autonomous vehicles: A review of methods and applications, IEEE Transactions on Intelligent Transportation Systems (2025).

[6] X. Chen, X. Wang, W. Zhao, C. Wang, S. Cheng, Z. Luan, Hierarchical deep reinforcement learning based multi-agent game control for energy consumption and traffic efficiency improving of autonomous vehicles, Energy 323 (2025) 135669.

[7] L. Zha, C. Gong, K. Lv, Real-time localization and navigation method for autonomous vehicles based on multi-modal data fusion by integrating memory transformer and ddqn, Image and Vision Computing 156 (2025) 105484.

[8] W. Sun, H. Shao, J. Li, T. Wu, E. Z. Fainman, Multi-type traffic sensor location problem for origin–destination estimation considering spatiotemporal correlation and sensor failure, Transportation Research Part C: Emerging Technologies 179 (2025) 105288.

[9] X. Chen, S. P. H. Boroujeni, X. Shu, H. Li, A. Razi, Enhancing graph neural networks in large-scale traffic incident analysis with concurrency hypothesis, in: Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems, 2024, pp. 196–207.

[10] Y. Chang, W. Xiao, B. Coifman, Using spatiotemporal stacks for precise vehicle tracking from roadside 3d lidar data, Transportation research part C: emerging technologies 154 (2023) 104280.

[11] A. D. Beza, Z. Xie, M. Ramezani, D. Levinson, From lane-less to lane-free: Implications in the era of automated vehicles, Transportation Research Part C: Emerging Technologies 170 (2025) 104898.

[12] K. Wang, J. Guo, K. Chen, J. Lu, An in-depth examination of slam methods: Challenges, advancements, and applications in complex scenes for autonomous driving, IEEE Transactions on Intelligent Transportation Systems (2025).

[13] Y. Zha, W. Shangguan, J. Chen, L. Chai, W. Qiu, A. M. López, Heterogeneous multiscale cooperative perception for connected autonomous vehicles via v2x interaction, IEEE Internet of Things Journal (2025).

[14] M. Salari, L. Kattan, M. Gentili, Optimal roadside units location for path flow reconstruction in a connected vehicle environment, Transportation Research Part C: Emerging Technologies 138 (2022) 103625.

[15] R. Praveen, S. Hundekari, P. Parida, T. Mittal, A. Sehgal, M. Bhavana, Autonomous vehicle navigation systems: Machine learning for real-time traffic prediction, in: 2025 International Conference on Computational, Communication and Information Technology (ICCCIT), IEEE, 2025, pp. 809–813.

[16] A. Mohammadi, R. Ahmari, V. Hemmati, F. Owusu-Ambrose, M. N. Mahmoud, P. Kebria, A. Homaifar, Detection of multiple small biased gps spoofing attacks on autonomous vehicles using time series analysis, IEEE Open Journal of Vehicular Technology (2025).

[17] S. D. RS, S. D. Varshni, Embedded large language models for enhanced human-machine interface in autonomous vehicles, in: 2025 International Conference on Multi-Agent Systems for Collaborative Intelligence (ICMSCI), IEEE, 2025, pp. 1143–1150.

[18] H. Kumar, P. Mamoria, D. K. Dewangan, Improving faster r-cnn for vehicle detection under varying conditions with domain adaptation technique, in: 2025 Fourth International Conference on Power, Control and Computing Technologies (ICPC2T), IEEE, 2025, pp. 1–6.

[19] A. Shrivastava, V. Kansal, A. Nagpal, K. K. Dixit, K. V. Rajkumar, et al., Ai-powered object detection for autonomous vehicles: A comparative study of machine learning models, in: 2025 International Conference on Computational, Communication and Information Technology (ICCCIT), IEEE, 2025, pp. 612–617.

[20] J. Subhedar, M. R. Bachute, Insights of semantic segmentation using the deeplab architecture for autonomous driving, MethodsX (2025) 103387.

[21] S. Chen, X. Li, K. Wang, J. Sun, B. Yang, Ranging research on telematics based on mask r-cnn dual eye stereo vision ranging algorithm, in: The International Conference Optoelectronic Information and Optical Engineering (OIOE2024), Vol. 13513, SPIE, 2025, pp. 884–889.

[22] S. P. H. Boroujeni, N. Mehrabi, F. Afghah, C. P. McGrath, D. Bhatkar, M. A. Biradar, A. Razi, Toward ai-driven fire imagery: Attributes, challenges, comparisons, and the promise of vlms and llms, Machine Learning with Applications (2025) 100763.

[23] Y. Tian, F. Lin, Y. Li, T. Zhang, Q. Zhang, X. Fu, J. Huang, X. Dai, Y. Wang, C. Tian, et al., Uavs meet llms: Overviews and perspectives towards agentic low-altitude mobility, Information Fusion 122 (2025) 103158.

[24] Z. Guo, Z. Yagudin, A. Lykov, M. Konenkov, D. Tsetserukou, Vlm-auto: Vlm-based autonomous driving assistant with human-like behavior and understanding for complex road scenes, in: 2024 2nd International Conference on Foundation and Large Language Models (FLLM), IEEE, 2024, pp. 501–507.

[25] Y. Wang, S. Wang, Y. Li, M. Liu, Developments in 3d object detection for autonomous driving: A review, IEEE Sensors Journal (2025).

[26] H. Wang, J. Liu, H. Dong, Z. Shao, A survey of the multi-sensor fusion object detection task in autonomous driving, Sensors 25 (9) (2025) 2794.

[27] H. Wang, X. Chen, Q. Yuan, P. Liu, A review of 3d object detection based on autonomous driving, The Visual Computer 41 (3) (2025) 1757–1775.

[28] Z. Song, L. Liu, F. Jia, Y. Luo, C. Jia, G. Zhang, L. Yang, L. Wang, Robustness-aware 3d object detection in autonomous driving: A review and outlook, IEEE Transactions on Intelligent Transportation Systems (2024).

[29] S. Y. Alaba, A. C. Gurbuz, J. E. Ball, Emerging trends in autonomous vehicle perception: Multimodal fusion for 3d object detection, World Electric Vehicle Journal 15 (1) (2024) 20.

[30] Z. Zou, K. Chen, Z. Shi, Y. Guo, J. Ye, Object detection in 20 years: A survey, Proceedings of the IEEE 111 (3) (2023) 257–276.

[31] X. Ma, W. Ouyang, A. Simonelli, E. Ricci, 3d object detection from images for autonomous driving: a survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 46 (5) (2023) 3537–3556.

[32] R. Qian, X. Lai, X. Li, 3d object detection for autonomous driving: A survey, Pattern Recognition 130 (2022) 108796.

[33] Y. Cui, R. Chen, W. Chu, L. Chen, D. Tian, Y. Li, D. Cao, Deep learning for image and point cloud fusion in autonomous driving: A review, IEEE Transactions on Intelligent Transportation Systems 23 (2) (2021) 722–739.

[34] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, K. Dietmayer, Deep

multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges, IEEE Transactions on Intelligent Transportation Systems 22 (3) (2020) 1341–1360.

[35] J. Guo, U. Kurup, M. Shah, Is it safe to drive? an overview of factors, metrics, and datasets for driveability assessment in autonomous driving, IEEE Transactions on Intelligent Transportation Systems 21 (8) (2019) 3135–3151.

[36] L. Hu, J. Zhang, J. Zhang, S. Cheng, Y. Wang, W. Zhang, N. Yu, Security analysis and adaptive false data injection against multi-sensor fusion localization for autonomous driving, Information Fusion 117 (2025) 102822.

[37] S. P. H. Boroujeni, A. Razi, S. Khoshdel, F. Afghah, J. L. Coen, L. O'Neill, P. Fule, A. Watts, N.-M. T. Kokolakis, K. G. Vamvoudakis, A comprehensive survey of research towards ai-enabled unmanned aerial systems in pre-, active-, and post-wildfire management, Information Fusion 108 (2024) 102369.

[38] H. Du, L. Ren, Y. Wang, X. Cao, C. Sun, Advancements in perception system with multi-sensor fusion for embodied agents, Information Fusion 117 (2025) 102859.

[39] Y. Wu, Fusion-based modeling of an intelligent algorithm for enhanced object detection using a deep learning approach on radar and camera data, Information Fusion 113 (2025) 102647.

[40] Y. Wu, J. Liu, M. Gong, Q. Miao, W. Ma, C. Xu, Joint semantic segmentation using representations of lidar point clouds and camera images, Information Fusion 108 (2024) 102370.

[41] S. Li, X. Li, H. Wang, Y. Zhou, Z. Shen, Multi-gnss ppp/ins/vision/lidar tightly integrated system for precise navigation in urban environments, Information Fusion 90 (2023) 218–232.

[42] N. Mehrabi, S. P. H. Boroujeni, Age estimation based on facial images using hybrid features and particle swarm optimization, in: 2021 11th International Conference on Computer Engineering and Knowledge (ICCKE), IEEE, 2021, pp. 412–418.

[43] A. Sarlak, H. Alzorgan, S. P. H. Boroujeni, A. Razi, R. Amin, Enhanced cooperative perception for autonomous vehicles using imperfect communication, in: 2024 20th International Conference on Distributed Computing in Smart Systems and the Internet of Things (DCOSS-IoT), IEEE, 2024, pp. 700–707.

[44] D. Kent, M. Alyaqoub, X. Lu, H. Khatounabadi, K. Sung, C. Scheller, A. Dalat, A. bin Thabit, R. Whitley, H. Radha, Msu-4s-the michigan state university four seasons dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22658–22667.

[45] L. Zheng, L. Yang, Q. Lin, W. Ai, M. Liu, S. Lu, J. Liu, H. Ren, J. Mo, X. Bai, et al., Omnihd-scenes: A next-generation multimodal dataset for autonomous driving, arXiv preprint arXiv:2412.10734 (2024).

[46] M. Alibeigi, W. Ljungbergh, A. Tonderski, G. Hess, A. Lilja, C. Lindström, D. Motorniuk, J. Fu, J. Widahl, C. Petersson, Zenseact open dataset: A large-scale and diverse multimodal dataset for autonomous driving, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 20178–20188.

[47] C. A. Diaz-Ruiz, Y. Xia, Y. You, J. Nino, J. Chen, J. Monica, X. Chen, K. Luo, Y. Wang, M. Emond, et al., Ithaca365: Dataset and driving perception under repeated and challenging weather conditions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21383–21392.

[48] J. Mao, M. Niu, C. Jiang, H. Liang, J. Chen, X. Liang, Y. Li, C. Ye, W. Zhang, Z. Li, et al., One million scenes for autonomous driving: Once dataset, arXiv preprint arXiv:2106.11037 (2021).

[49] J.-L. Déziel, P. Merriaux, F. Tremblay, D. Lessard, D. Plourde, J. Stanguennec, P. Goulet, P. Olivier, Pixset: An opportunity for 3d computer vision to go beyond point clouds with a full-waveform lidar dataset, in: 2021 ieee international intelligent transportation systems conference (itsc), IEEE, 2021, pp. 2987–2993.

[50] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang, et al., Pandaset: Advanced sensor suite dataset for autonomous driving, in: 2021 IEEE international intelligent transportation systems conference (ITSC), IEEE, 2021, pp. 3095–3101.

[51] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung, L. Hauswald, V. H. Pham, M. Mühlegg, S. Dorn, et al., A2d2: Audi autonomous driving dataset, arXiv preprint arXiv:2004.06320 (2020).

[52] Roboflow, Self-driving car dataset, accessed: 2025-02-28 (2025).
URL `https://public.roboflow.com/object-detection/self-driving-car`

[53] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al., Scalability in perception for autonomous driving: Waymo open dataset, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 2446–2454.

[54] Q.-H. Pham, P. Sevestre, R. S. Pahwa, H. Zhan, C. H. Pang, Y. Chen, A. Mustafa, V. Chandrasekhar, J. Lin, A* 3d dataset: Towards autonomous driving in challenging environments, in: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2020, pp. 2267–2273.

[55] J. Bock, R. Krajewski, T. Moers, S. Runde, L. Vater, L. Eckstein, The ind dataset: A drone dataset of naturalistic road user trajectories at german intersections, in: 2020 IEEE Intelligent Vehicles Symposium (IV), 2020, pp. 1929–1934. `doi:10.1109/IV47402.2020.9304839`.

[56] T. Moers, L. Vater, R. Krajewski, J. Bock, A. Zlocki, L. Eckstein, The exid dataset: A real-world trajectory dataset of highly interactive highway scenarios in germany, in: 2022 IEEE Intelligent Vehicles Symposium (IV), 2022, pp. 958–964. `doi:10.1109/IV51971.2022.9827305`.

[57] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, O. Beijbom, nuscenes: A multimodal dataset for autonomous driving, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11621–11631.

[58] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al., Argoverse: 3d tracking and forecasting with rich maps, in: Proceedings of the IEEE/CVF conference on

computer vision and pattern recognition, 2019, pp. 8748–8757.

[59] J. Xue, J. Fang, T. Li, B. Zhang, P. Zhang, Z. Ye, J. Dou, Blvd: Building a large-scale 5d semantics benchmark for autonomous driving, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 6685–6691.

[60] H. Schafer, E. Santana, A. Haden, R. Biasini, A commute in data: The comma2k19 dataset, arXiv preprint arXiv:1812.05752 (2018).

[61] F. Yu, W. Xian, Y. Chen, F. Liu, M. Liao, V. Madhavan, T. Darrell, et al., Bdd100k: A diverse driving video database with scalable annotation tooling, arXiv preprint arXiv:1805.04687 2 (5) (2018) 6.

[62] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, R. Yang, The apolloscape dataset for autonomous driving, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2018, pp. 954–960.

[63] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, B. Schiele, The cityscapes dataset for semantic urban scene understanding, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 3213–3223.

[64] D. Barnes, M. Gadd, P. Murcutt, P. Newman, I. Posner, The oxford radar robotcar dataset: A radar extension to the oxford robotcar dataset, in: 2020 IEEE international conference on robotics and automation (ICRA), IEEE, 2020, pp. 6433–6438.

[65] A. Geiger, P. Lenz, R. Urtasun, Are we ready for autonomous driving? the kitti vision benchmark suite, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE, 2012, pp. 3354–3361.

[66] X. Zhu, H. Sheng, S. Cai, B. Deng, S. Yang, Q. Liang, K. Chen, L. Gao, J. Song, J. Ye, Roscenes: A large-scale multi-view 3d dataset for roadside perception, in: European Conference on Computer Vision, Springer, 2024, pp. 331–347.

[67] W. Zimmer, C. Creß, H. T. Nguyen, A. C. Knoll, Tumtraf intersection dataset: All you need for urban 3d camera-lidar roadside perception, in: 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2023, pp. 1030–1037.

[68] C. Creß, W. Zimmer, L. Strand, M. Fortkord, S. Dai, V. Lakshminarasimhan, A. Knoll, A9-dataset: Multi-sensor infrastructure-based dataset for mobility research, in: 2022 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2022, pp. 965–970.

[69] H. Wang, X. Zhang, Z. Li, J. Li, K. Wang, Z. Lei, R. Haibing, Ips300+: a challenging multi-modal data sets for intersection perception system, in: 2022 International Conference on Robotics and Automation (ICRA), IEEE, 2022, pp. 2539–2545.

[70] X. Ye, M. Shu, H. Li, Y. Shi, Y. Li, G. Wang, X. Tan, E. Ding, Rope3d: The roadside perception dataset for autonomous driving and monocular 3d object detection task, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21341–21350.

[71] S. Busch, C. Koetsier, J. Axmann, C. Brenner, Lumpi: The leibniz university multi-perspective intersection dataset, in: 2022 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2022, pp. 1127–1134.

[72] M. Howe, I. Reid, J. Mackenzie, Weakly supervised training of monocular 3d object detectors using wide baseline multi-view traffic camera data, arXiv preprint arXiv:2110.10966 (2021).

[73] W. Zhan, L. Sun, D. Wang, H. Shi, A. Clausse, M. Naumann, J. Kummerle, H. Konigshof, C. Stiller, A. de La Fortelle, et al., Interaction dataset: An international, adversarial and cooperative motion dataset in interactive driving scenarios with semantic maps, arXiv preprint arXiv:1910.03088 (2019).

[74] Z. Tang, M. Naphade, M.-Y. Liu, X. Yang, S. Birchfield, S. Wang, R. Kumar, D. Anastasiu, J.-N. Hwang, Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 8797–8806.

[75] A. Ishaq, J. Lahoud, K. More, O. Thawakar, R. Thawkar, D. Dissanayake, N. Ahsan, Y. Li, F. S. Khan, H. Cholakkal, et al., Drivelmm-o1: A step-by-step reasoning dataset and large multimodal model for driving scenario understanding, arXiv preprint arXiv:2503.10621 (2025).

[76] K. Chen, Y. Li, W. Zhang, Y. Liu, P. Li, R. Gao, L. Hong, M. Tian, X. Zhao, Z. Li, et al., Automated evaluation of large vision-language models on self-driving corner cases, in: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), IEEE, 2025, pp. 7817–7826.

[77] H.-k. Chiu, R. Hachiuma, C.-Y. Wang, S. F. Smith, Y.-C. F. Wang, M.-H. Chen, V2v-llm: Vehicle-to-vehicle cooperative autonomous driving with multi-modal large language models, arXiv preprint arXiv:2502.09980 (2025).

[78] C. Sima, K. Renz, K. Chitta, L. Chen, H. Zhang, C. Xie, J. Beißwenger, P. Luo, A. Geiger, H. Li, Drivelm: Driving with graph visual question answering, in: European Conference on Computer Vision, Springer, 2024, pp. 256–274.

[79] Y. Inoue, Y. Yada, K. Tanahashi, Y. Yamaguchi, Nuscenes-mqa: Integrated evaluation of captions and qa for autonomous driving datasets using markup annotations, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2024, pp. 930–938.

[80] S. Wang, Z. Yu, X. Jiang, S. Lan, M. Shi, N. Chang, J. Kautz, Y. Li, J. M. Alvarez, Omnidrive: A holistic vision-language dataset for autonomous driving with counterfactual reasoning, arXiv preprint arXiv:2504.04348 (2025).

[81] R. Tian, B. Li, X. Weng, Y. Chen, E. Schmerling, Y. Wang, B. Ivanovic, M. Pavone, Tokenize the world into object-level knowledge to address long-tail events in autonomous driving, arXiv preprint arXiv:2407.00959 (2024).

[82] X. Cao, T. Zhou, Y. Ma, W. Ye, C. Cui, K. Tang, Z. Cao, K. Liang, Z. Wang, J. M. Rehg, et al., Maplm: A real-world large-scale vision-language benchmark for map and traffic scene understanding, in: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 21819–21830.

[83] A.-M. Marcu, L. Chen, J. Hünermann, A. Karnsund, B. Hanotte, P. Chidananda, S. Nair, V. Badrinarayanan, A. Kendall, J. Shotton, et al., Lingoqa: Visual question answering for autonomous driving, in: European Conference on Computer Vision, Springer, 2024, pp. 252–269.

[84] T. Qian, J. Chen, L. Zhuo, Y. Jiao, Y.-G. Jiang, Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 4542–4550.

[85] X. Ding, J. Han, H. Xu, X. Liang, W. Zhang, X. Li, Holistic autonomous driving understanding by bird's-eye-view injected multi-modal large models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 13668–13677.

[86] S. Malla, C. Choi, I. Dwivedi, J. H. Choi, J. Li, Drama: Joint risk localization and captioning in driving, in: Proceedings of the IEEE/CVF winter conference on applications of computer vision, 2023, pp. 1043–1052.

[87] Y. Li, Z. Li, N. Chen, M. Gong, Z. Lyu, Z. Wang, P. Jiang, C. Feng, Multiagent multitraversal multimodal self-driving: Open mars dataset, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22041–22051.

[88] Y. Lu, Y. Hu, Y. Zhong, D. Wang, Y. Wang, S. Chen, An extensible framework for open heterogeneous collaborative perception, arXiv preprint arXiv:2401.13964 (2024).

[89] Y. Hu, Y. Lu, R. Xu, W. Xie, S. Chen, Y. Wang, Collaboration helps camera overtake lidar in 3d detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 9243–9252.

[90] S. Wei, Y. Wei, Y. Hu, Y. Lu, Y. Zhong, S. Chen, Y. Zhang, Asynchrony-robust collaborative perception via bird's eye view flow, Advances in Neural Information Processing Systems 36 (2023) 28462–28477.

[91] R. Xu, X. Xia, J. Li, H. Li, S. Zhang, Z. Tu, Z. Meng, H. Xiang, X. Dong, R. Song, et al., V2v4real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13712–13722.

[92] J. Axmann, R. Moftizadeh, J. Su, B. Tennstedt, Q. Zou, Y. Yuan, D. Ernst, H. Alkhatib, C. Brenner, S. Schön, Lucoop: Leibniz university cooperative perception and urban navigation dataset, in: 2023 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2023, pp. 1–8.

[93] Y. Hu, S. Fang, Z. Lei, Y. Zhong, S. Chen, Where2comm: Communication-efficient collaborative perception via spatial confidence maps, Advances in neural information processing systems 35 (2022) 4874–4886.

[94] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, J. Ma, Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication, in: 2022 International Conference on Robotics and Automation (ICRA), IEEE, 2022, pp. 2583–2589.

[95] E. Arnold, S. Mozaffari, M. Dianati, Fast and robust registration of partially overlapping point clouds, IEEE Robotics and Automation Letters 7 (2) (2021) 1502–1509.

[96] Y. Yuan, M. Sester, Comap: A synthetic dataset for collective multi-agent perception of autonomous driving, The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences 43 (2021) 255–263.

[97] Q. Chen, S. Tang, Q. Yang, S. Fu, Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds, in: 2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS), IEEE, 2019, pp. 514–524.

[98] M. Fan, S. Yu, S. Xu, K. Jiang, H. Xiong, X. Liu, A benchmark for vision-centric hd mapping by v2i systems, arXiv preprint arXiv:2503.23963 (2025).

[99] X. Liu, B. Li, R. Xu, J. Ma, X. Li, J. Li, H. Yu, V2x-dsi: A density-sensitive infrastructure lidar benchmark for economic vehicle-to-everything cooperative perception, in: 2024 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2024, pp. 490–495.

[100] L. Yang, X. Zhang, J. Li, C. Wang, Z. Song, T. Zhao, Z. Song, L. Wang, M. Zhou, Y. Shen, et al., V2x-radar: A multi-modal dataset with 4d radar for cooperative perception, arXiv preprint arXiv:2411.10962 (2024).

[101] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, A. C. Knoll, Tumtraf v2x cooperative perception dataset, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 22668–22677.

[102] H. Zhu, Y. Wang, Q. Kong, Y. Wei, X. Xia, B. Deng, R. Xiong, Y. Wang, Otvic: A dataset with online transmission for vehicle-to-infrastructure cooperative 3d object detection, in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2024, pp. 10732–10739.

[103] H. Wang, Y. Niu, L. Chen, Y. Li, M. A. Sotelo, Z. Li, Y. Cai, Dair-v2xreid: A new real-world vehicle-infrastructure cooperative re-id dataset and cross-shot feature aggregation network perception method, IEEE Transactions on Intelligent Transportation Systems (2024).

[104] C. Ma, L. Qiao, C. Zhu, K. Liu, Z. Kong, Q. Li, X. Zhou, Y. Kan, W. Wu, Holovic: Large-scale dataset and benchmark for multi-sensor holographic intersection and vehicle-infrastructure cooperative, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22129–22138.

[105] H. Yu, W. Yang, H. Ruan, Z. Yang, Y. Tang, X. Gao, X. Hao, Y. Shi, Y. Pan, N. Sun, et al., V2x-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 5486–5495.

[106] Z. Bai, G. Wu, M. J. Barth, Y. Liu, E. A. Sisbot, K. Oguchi, Pillargrid: Deep learning-based cooperative perception for 3d object detection from onboard-roadside lidar, in: 2022 IEEE 25th International Conference on Intelligent

Transportation Systems (ITSC), IEEE, 2022, pp. 1743–1749.

[107] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, et al., Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 21361–21370.

[108] E. Arnold, M. Dianati, R. de Temple, S. Fallah, Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors, IEEE Transactions on Intelligent Transportation Systems 23 (3) (2020) 1852–1864.

[109] K. Z. Luo, M.-Q. Dao, Z. Liu, M. Campbell, W.-L. Chao, K. Q. Weinberger, E. Malis, V. Fremont, B. Hariharan, M. Shan, et al., Mixed signals: A diverse point cloud dataset for heterogeneous lidar v2x collaboration, arXiv preprint arXiv:2502.14156 (2025).

[110] H. Xiang, Z. Zheng, X. Xia, S. Z. Zhao, L. Gao, Z. Zhou, T. Cai, Y. Zhang, J. Ma, V2x-realo: An open online framework and dataset for cooperative perception in reality, arXiv preprint arXiv:2503.10034 (2025).

[111] T. Wang, S. Kim, J. Wenxuan, E. Xie, C. Ge, J. Chen, Z. Li, P. Luo, Deepaccident: A motion and accident prediction benchmark for v2x autonomous driving, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 5599–5606.

[112] M. Karvat, S. Givigi, Adver-city: Open-source multi-modal dataset for collaborative perception under adverse weather conditions, arXiv preprint arXiv:2410.06380 (2024).

[113] R. Li, X. Pei, Multi-v2x: A large scale multi-modal multi-penetration-rate dataset for cooperative perception, arXiv preprint arXiv:2409.04980 (2024).

[114] S. Chen, Z. Song, S. Zhou, et al., Whales: A multi-agent scheduling dataset for enhanced cooperation in autonomous driving, arXiv preprint arXiv:2411.13340 (2024).

[115] X. Huang, J. Wang, Q. Xia, S. Chen, B. Yang, X. Li, C. Wang, C. Wen, V2x-r: Cooperative lidar-4d radar fusion for 3d object detection with denoising diffusion, arXiv preprint arXiv:2411.08402 (2024).

[116] Z. Zhou, H. Xiang, Z. Zheng, S. Z. Zhao, M. Lei, Y. Zhang, T. Cai, X. Liu, J. Liu, M. Bajji, et al., V2xpnp: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction, arXiv preprint arXiv:2412.01812 (2024).

[117] J. Gamerdinger, S. Teufel, P. Schulz, S. Amann, J.-P. Kirchner, O. Bringmann, Scope: A synthetic multi-modal dataset for collective perception including physical-correct weather conditions, arXiv preprint arXiv:2408.03065 (2024).

[118] T. Wang, F. Lu, Z. Zheng, Z. Li, G. Chen, et al., Rcdn: Towards robust camera-insensitivity collaborative perception via dynamic feature-based 3d neural modeling, Advances in Neural Information Processing Systems 37 (2024) 22350–22369.

[119] Y. Li, D. Ma, Z. An, Z. Wang, Y. Zhong, S. Chen, C. Feng, V2x-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving, IEEE Robotics and Automation Letters 7 (4) (2022) 10914–10921.

[120] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, J. Ma, V2x-vit: Vehicle-to-everything cooperative perception with vision transformer, in: European conference on computer vision, Springer, 2022, pp. 107–124.

[121] R. Mao, J. Guo, Y. Jia, Y. Sun, S. Zhou, Z. Niu, Dolphins: Dataset for collaborative perception enabled harmonious and interconnected self-driving, in: Proceedings of the Asian Conference on Computer Vision, 2022, pp. 4361–4377.

[122] H. Xiang, Z. Zheng, X. Xia, R. Xu, L. Gao, Z. Zhou, X. Han, X. Ji, M. Li, Z. Meng, et al., V2x-real: a largs-scale dataset for vehicle-to-everything cooperative perception, in: European Conference on Computer Vision, Springer, 2024, pp. 455–470.

[123] R. Hao, S. Fan, Y. Dai, Z. Zhang, C. Li, Y. Wang, H. Yu, W. Yang, J. Yuan, Z. Nie, Rcooper: A real-world large-scale dataset for roadside cooperative perception, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 22347–22357.

[124] X. Zhang, Y. Li, J. Wang, X. Qin, Y. Shen, Z. Fan, X. Tan, Inscope: A new real-world 3d infrastructure-side collaborative perception dataset for open traffic scenarios, arXiv preprint arXiv:2407.21581 (2024).

[125] S. P. H. Boroujeni, A. Razi, Ic-gan: An improved conditional generative adversarial network for rgb-to-ir image translation with applications to forest fire monitoring, Expert Systems with Applications 238 (2024) 121962.

[126] M. Elrod, N. Mehrabi, R. Amin, M. Kaur, L. Cheng, J. Martin, A. Razi, Graph based deep reinforcement learning aided by transformers for multi-agent cooperation, arXiv preprint arXiv:2504.08195 (2025).

[127] T. Liang, H. Bao, W. Pan, X. Fan, H. Li, Aspectnet: Aspect-aware anchor-free detector for autonomous driving, Applied Sciences 12 (12) (2022) 5972.

[128] J. Choi, D. Chun, H. Kim, H.-J. Lee, Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving, in: Proceedings of the IEEE/CVF International conference on computer vision, 2019, pp. 502–511.

[129] X. Hu, X. Xu, Y. Xiao, H. Chen, S. He, J. Qin, P.-A. Heng, Sinet: A scale-insensitive convolutional neural network for fast vehicle detection, IEEE transactions on intelligent transportation systems 20 (3) (2018) 1010–1019.

[130] J. Yi, P. Wu, D. N. Metaxas, Assd: Attentive single shot multibox detector, Computer Vision and Image Understanding 189 (2019) 102827.

[131] S. Zhang, L. Wen, X. Bian, Z. Lei, S. Z. Li, Single-shot refinement neural network for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4203–4212.

[132] S. Liu, D. Huang, et al., Receptive field block net for accurate and fast object detection, in: Proceedings of the European conference on computer vision (ECCV), 2018, pp. 385–400.

[133] B. Wu, F. Iandola, P. H. Jin, K. Keutzer, Squeezedet: Unified, small, low power fully convolutional neural networks for real-time object detection for autonomous driving, in: Proceedings of the IEEE conference on computer vision

and pattern recognition workshops, 2017, pp. 129–137.

[134] Z. Cai, Q. Fan, R. S. Feris, N. Vasconcelos, A unified multi-scale deep convolutional neural network for fast object detection, in: European conference on computer vision, Springer, 2016, pp. 354–370.

[135] F. Yang, W. Choi, Y. Lin, Exploit all the layers: Fast and accurate cnn object detector with scale dependent pooling and cascaded rejection classifiers, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2129–2137.

[136] G. Brazil, X. Liu, M3d-rpn: Monocular 3d region proposal network for object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 9287–9296.

[137] F. Pu, Y. Wang, J. Deng, W. Yang, Monodgp: Monocular 3d object detection with decoupled-query and geometry-error priors, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 6520–6530.

[138] H.-I. Liu, C. Wu, J.-H. Cheng, W. Chai, S.-Y. Wang, G. Liu, H. Latapie, J.-C. Wu, J.-N. Hwang, H.-H. Shuai, W.-H. Cheng, Monotakd: Teaching assistant knowledge distillation for monocular 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 22266–22275.

[139] Y. Ranasinghe, D. Hegde, V. M. Patel, Monodiff: Monocular 3d object detection and pose estimation with diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10659–10670.

[140] Z. Wu, Y. Gan, Y. Wu, R. Wang, X. Wang, J. Pu, Fd3d: Exploiting foreground depth map for feature-supervised monocular 3d object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 6189–6197.

[141] L. Yan, P. Yan, S. Xiong, X. Xiang, Y. Tan, Monocd: Monocular 3d object detection with complementary depths, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 10248–10257.

[142] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, T. He, Y. Li, W. Ouyang, Gupnet++: Geometry uncertainty propagation network for monocular 3d object detection, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[143] R. Zhang, H. Qiu, T. Wang, Z. Guo, Z. Cui, Y. Qiao, H. Li, P. Gao, Monodetr: Depth-guided transformer for monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 9155–9166.

[144] J. Jinrang, Z. Li, Y. Shi, Monouni: A unified vehicle and infrastructure-side monocular 3d object detection network with sufficient depth clues, Advances in Neural Information Processing Systems 36 (2023) 11703–11715.

[145] A. Kumar, G. Brazil, E. Corona, A. Parchami, X. Liu, Deviant: Depth equivariant network for monocular 3d object detection, in: European Conference on Computer Vision, Springer, 2022, pp. 664–683.

[146] Z. Wu, Y. Gan, L. Wang, G. Chen, J. Pu, Monopgc: Monocular 3d object detection with pixel geometry contexts, arXiv preprint arXiv:2302.10549 (2023).

[147] Y. Zhou, H. Zhu, Q. Liu, S. Chang, M. Guo, Monoatt: Online monocular 3d object detection with adaptive token transformer, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 17493–17503.

[148] Y. Li, Y. Chen, J. He, Z. Zhang, Densely constrained depth estimator for monocular 3d object detection, in: European Conference on Computer Vision, Springer, 2022, pp. 718–734.

[149] Z. Li, Z. Qu, Y. Zhou, J. Liu, H. Wang, L. Jiang, Diversity matters: Fully exploiting depth clues for reliable monocular 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 2791–2800.

[150] Z. Qin, X. Li, Monoground: Detecting monocular 3d objects from the ground, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 3793–3802.

[151] Q. Lian, P. Li, X. Chen, Monojsg: Joint semantic and geometric cost volume for monocular 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1070–1079.

[152] G. Brazil, G. Pons-Moll, X. Liu, B. Schiele, Kinematic 3d object detection in monocular video, in: European Conference on Computer Vision, Springer, 2020, pp. 135–152.

[153] D. Rukhovich, A. Vorontsova, A. Konushin, Imvoxelnet: Image to voxels projection for monocular and multi-view general-purpose 3d object detection, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2022, pp. 2397–2406.

[154] J. Gu, B. Wu, L. Fan, J. Huang, S. Cao, Z. Xiang, X.-S. Hua, Homography loss for monocular 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 1080–1089.

[155] Y. Hong, H. Dai, Y. Ding, Cross-modality knowledge distillation network for monocular 3d object detection, in: European Conference on Computer Vision, Springer, 2022, pp. 87–104.

[156] L. Peng, X. Wu, Z. Yang, H. Liu, D. Cai, Did-m3d: Decoupling instance depth for monocular 3d object detection, in: European Conference on Computer Vision, Springer, 2022, pp. 71–88.

[157] Y. Lu, X. Ma, L. Yang, T. Zhang, Y. Liu, Q. Chu, J. Yan, W. Ouyang, Geometry uncertainty projection network for monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 3111–3121.

[158] X. Ma, Y. Zhang, D. Xu, D. Zhou, S. Yi, H. Li, W. Ouyang, Delving into localization errors for monocular 3d

object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 4721–4730.

[159] X. Shi, Q. Ye, X. Chen, C. Chen, Z. Chen, T.-K. Kim, Geometry-based distance decomposition for monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 15172–15181.

[160] Y. Zhang, J. Lu, J. Zhou, Objects are different: Flexible monocular 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 3289–3298.

[161] H. Chen, Y. Huang, W. Tian, Z. Gao, L. Xiong, Monorun: Monocular 3d object detection by reconstruction and uncertainty propagation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10379–10388.

[162] L. Wang, L. Du, X. Ye, Y. Fu, G. Guo, X. Xue, J. Feng, L. Zhang, Depth-conditioned dynamic message propagation for monocular 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 454–463.

[163] Y. Chen, L. Tai, K. Sun, M. Li, Monopair: Monocular 3d object detection using pairwise spatial relationships, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 12093–12102.

[164] P. Li, H. Zhao, P. Liu, F. Cao, Rtm3d: Real-time monocular 3d detection from object keypoints for autonomous driving, in: European Conference on Computer Vision, Springer, 2020, pp. 644–660.

[165] X. Shi, Z. Chen, T.-K. Kim, Distance-normalized unified representation for monocular 3d object detection, in: European Conference on Computer Vision, Springer, 2020, pp. 91–107.

[166] A. Simonelli, S. R. Bulo, L. Porzi, E. Ricci, P. Kontschieder, Towards generalization across depth for monocular 3d object detection, in: European Conference on Computer Vision, Springer, 2020, pp. 767–782.

[167] X. Ye, L. Du, Y. Shi, Y. Li, X. Tan, J. Feng, E. Ding, S. Wen, Monocular 3d object detection via feature domain adaptation, in: European Conference on Computer Vision, Springer, 2020, pp. 17–34.

[168] A. Simonelli, S. R. Bulo, L. Porzi, M. López-Antequera, P. Kontschieder, Disentangling monocular 3d object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1991–1999.

[169] Y. Wang, V. C. Guizilini, T. Zhang, Y. Wang, H. Zhao, J. Solomon, Detr3d: 3d object detection from multi-view images via 3d-to-2d queries, in: Conference on robot learning, PMLR, 2022, pp. 180–191.

[170] Z. Xue, M. Guo, H. Fan, S. Zhang, Z. Zhang, Corrbev: Multi-view 3d object detection by correlation learning with multi-modal prototypes, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2025, pp. 27413–27423.

[171] H. Ji, T. Ni, X. Huang, Z. Shi, T. Luo, X. Zhan, J. Chen, Ropetr: Improving temporal camera-only 3d detection by integrating enhanced rotary position embedding, arXiv preprint arXiv:2504.12643 (2025).

[172] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, J. Dai, Bevformer: learning bird's-eye-view representation from lidar-camera via spatiotemporal transformers, IEEE Transactions on Pattern Analysis and Machine Intelligence (2024).

[173] F. Liu, T. Huang, Q. Zhang, H. Yao, C. Zhang, F. Wan, Q. Ye, Y. Zhou, Ray denoising: Depth-aware hard negative sampling for multi-view 3d object detection, in: European Conference on Computer Vision, Springer, 2024, pp. 200–217.

[174] S. Wang, Y. Liu, T. Wang, Y. Li, X. Zhang, Exploring object-centric temporal modeling for efficient multi-view 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3621–3631.

[175] Y. Liu, J. Yan, F. Jia, S. Li, A. Gao, T. Wang, X. Zhang, Petrv2: A unified framework for 3d perception from multi-camera images, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3262–3272.

[176] Y. Li, Z. Ge, G. Yu, J. Yang, Z. Wang, Y. Shi, J. Sun, Z. Li, Bevdepth: Acquisition of reliable depth for multi-view 3d object detection, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 37, 2023, pp. 1477–1485.

[177] H. Liu, Y. Teng, T. Lu, H. Wang, L. Wang, Sparsebev: High-performance sparse 3d object detection from multi-camera videos, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 18580–18590.

[178] Z. Zong, D. Jiang, G. Song, Z. Xue, J. Su, H. Li, Y. Liu, Temporal enhanced training of multi-view 3d object detector via historical object prediction, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3781–3790.

[179] Y. Jiang, L. Zhang, Z. Miao, X. Zhu, J. Gao, W. Hu, Y.-G. Jiang, Polarformer: Multi-camera 3d object detection with polar transformer, in: Proceedings of the AAAI conference on Artificial Intelligence, Vol. 37, 2023, pp. 1042–1050.

[180] Y. Wang, Y. Chen, Z. Zhang, Frustumformer: Adaptive instance-aware resampling for multi-view 3d detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 5096–5105.

[181] Z. Li, Z. Yu, W. Wang, A. Anandkumar, T. Lu, J. M. Alvarez, Fb-bev: Bev representation from forward-backward view transformations, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 6919–6928.

[182] Z. Wang, Z. Huang, J. Fu, N. Wang, S. Liu, Object as query: Lifting any 2d object detector to 3d detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3791–3800.

[183] K. Xiong, S. Gong, X. Ye, X. Tan, J. Wan, E. Ding, J. Wang, X. Bai, Cape: Camera view position embedding for

multi-view 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 21570–21579.

[184] C. Shu, J. Deng, F. Yu, Y. Liu, 3dppe: 3d point positional encoding for transformer-based multi-camera 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 3580–3589.

[185] Y. Liu, T. Wang, X. Zhang, J. Sun, Petr: Position embedding transformation for multi-view 3d object detection, in: European conference on computer vision, Springer, 2022, pp. 531–548.

[186] Y. Li, Y. Chen, X. Qi, Z. Li, J. Sun, J. Jia, Unifying voxel-based representation with transformer for 3d object detection, Advances in Neural Information Processing Systems 35 (2022) 18442–18455.

[187] Y. Li, H. Bao, Z. Ge, J. Yang, J. Sun, Z. Li, Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 1486–1494.

[188] T. Wang, Z. Xinge, J. Pang, D. Lin, Probabilistic and geometric depth: Detecting objects in perspective, in: Conference on Robot Learning, PMLR, 2022, pp. 1475–1485.

[189] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, F. Zhao, Graph-detr3d: rethinking overlapping regions for multi-view 3d object detection, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5999–6008.

[190] T. Wang, X. Zhu, J. Pang, D. Lin, Fcos3d: Fully convolutional one-stage monocular 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops, 2021, pp. 913–922.

[191] P. Adarsh, P. Rathi, M. Kumar, Yolo v3-tiny: Object detection and recognition using one stage improved model, in: 2020 6th international conference on advanced computing and communication systems (ICACCS), IEEE, 2020, pp. 687–694.

[192] J. Gao, H. Li, Z. Li, C. Xie, X. Ji, Y. Zhang, An algorithm for road target detection of autonomous vehicles based on improved yolov8, Scientific Reports 15 (1) (2025) 21061.

[193] C. Li, L. Li, H. Jiang, K. Weng, Y. Geng, L. Li, Z. Ke, Q. Li, M. Cheng, W. Nie, et al., Yolov6: A single-stage object detection framework for industrial applications, arXiv preprint arXiv:2209.02976 (2022).

[194] C.-Y. Wang, A. Bochkovskiy, H.-Y. M. Liao, Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 7464–7475.

[195] M. Yang, X. Fan, Yolov8-lite: A lightweight object detection model for real-time autonomous driving systems, ICCK Transactions on Emerging Topics in Artificial Intelligence 1 (1) (2024) 1–16.

[196] Q. He, A. Xu, Z. Ye, W. Zhou, T. Cai, Object detection based on lightweight yolox for autonomous driving, Sensors 23 (17) (2023) 7596.

[197] X. Li, J. Chen, Y. Sun, N. Lin, A. Hawbani, L. Zhao, Yolo-vehicle-pro: A cloud-edge collaborative framework for object detection in autonomous driving under adverse weather conditions, arXiv preprint arXiv:2410.17734 (2024).

[198] H. An, J. Tang, Y. Fan, M. Liu, Improved vehicle object detection algorithm based on swin-yolov5s, Processes 13 (3) (2025) 925.

[199] X. Jia, Y. Tong, H. Qiao, M. Li, J. Tong, B. Liang, Fast and accurate object detector for autonomous driving based on improved yolov5, Scientific reports 13 (1) (2023) 9711.

[200] C.-Y. Wang, I.-H. Yeh, H.-Y. Mark Liao, Yolov9: Learning what you want to learn using programmable gradient information, in: European conference on computer vision, Springer, 2024, pp. 1–21.

[201] A. Wang, H. Chen, L. Liu, K. Chen, Z. Lin, J. Han, et al., Yolov10: Real-time end-to-end object detection, Advances in Neural Information Processing Systems 37 (2024) 107984–108011.

[202] N. Mehrabi, S. P. H. Boroujeni, A. Razi, Turbo-irl: Enhancing multi-agent systems using turbo decoding-inspired deep maximum entropy inverse reinforcement learning, Expert Systems with Applications 296 (2026) 128754.

[203] W. Ye, Q. Xia, H. Wu, Z. Dong, R. Zhong, C. Wang, C. Wen, Fade3d: Fast and deployable 3d object detection for autonomous driving, IEEE Transactions on Intelligent Transportation Systems (2025).

[204] Z. Ding, X. Zhang, Q. Jing, Y. Cheng, R. Feng, As-det: Active sampling for adaptive 3d object detection in point clouds, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, 2025, pp. 2762–2770.

[205] Z. Liu, J. Hou, X. Wang, X. Ye, J. Wang, H. Zhao, X. Bai, Lion: Linear group rnn for 3d object detection in point clouds, Advances in Neural Information Processing Systems 37 (2024) 13601–13626.

[206] H. A. Hoang, D. C. Bui, M. Yoo, Tsstdet: Transformation-based 3-d object detection via a spatial shape transformer, IEEE Sensors Journal 24 (5) (2024) 7126–7139.

[207] X. Jin, K. Liu, C. Ma, R. Yang, F. Hui, W. Wu, Swiftpillars: High-efficiency pillar encoder for lidar-based 3d detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 38, 2024, pp. 2625–2633.

[208] H. Yang, W. Wang, M. Chen, B. Lin, T. He, H. Chen, X. He, W. Ouyang, Pvt-ssd: Single-stage 3d object detector with point-voxel transformer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 13476–13487.

[209] H. Wang, C. Shi, S. Shi, M. Lei, S. Wang, D. He, B. Schiele, L. Wang, Dsvt: Dynamic sparse voxel transformer with rotated sets, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 13520–13529.

[210] T. Zhao, X. Ning, K. Hong, Z. Qiu, P. Lu, Y. Zhao, L. Zhang, L. Zhou, G. Dai, H. Yang, et al., Ada3d: Exploiting the spatial redundancy with adaptive inference for efficient 3d object detection, in: Proceedings of the IEEE/CVF

International Conference on Computer Vision, 2023, pp. 17728–17738.
[211] Y. Chen, J. Liu, X. Zhang, X. Qi, J. Jia, Voxelnext: Fully sparse voxelnet for 3d object detection and tracking, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 21674–21683.
[212] Q. Xia, Y. Chen, G. Cai, G. Chen, D. Xie, J. Su, Z. Wang, 3-d hanet: A flexible 3-d heatmap auxiliary network for object detection, IEEE Transactions on Geoscience and Remote Sensing 61 (2023) 1–13.
[213] H. Wu, C. Wen, W. Li, X. Li, R. Yang, C. Wang, Transformation-equivariant 3d object detection for autonomous driving, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 2795–2802.
[214] H. Sheng, S. Cai, N. Zhao, B. Deng, J. Huang, X.-S. Hua, M.-J. Zhao, G. H. Lee, Rethinking iou-based optimization for single-stage 3d object detection, in: European Conference on Computer Vision, Springer, 2022, pp. 544–561.
[215] Y. Zhang, Q. Hu, G. Xu, Y. Ma, J. Wan, Y. Guo, Not all points are equal: Learning highly efficient point-based detectors for 3d lidar point clouds, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 18953–18962.
[216] G. Shi, R. Li, C. Ma, Pillarnet: Real-time and high-performance pillar-based 3d object detection, in: European conference on computer vision, Springer, 2022, pp. 35–52.
[217] C. He, R. Li, S. Li, L. Zhang, Voxel set transformer: A set-to-set approach to 3d object detection from point clouds, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 8417–8427.
[218] Q. Xu, Y. Zhong, U. Neumann, Behind the curtain: Learning occluded shapes for 3d object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 2893–2901.
[219] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, H. Li, Voxel r-cnn: Towards high performance voxel-based 3d object detection, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 35, 2021, pp. 1201–1209.
[220] W. Zheng, W. Tang, S. Chen, L. Jiang, C.-W. Fu, Cia-ssd: Confident iou-aware single-stage object detector from point cloud, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 35, 2021, pp. 3555–3562.
[221] H. Sheng, S. Cai, Y. Liu, B. Deng, J. Huang, X.-S. Hua, M.-J. Zhao, Improving 3d object detection with channel-wise transformer, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2743–2752.
[222] Q. Xu, Y. Zhou, W. Wang, C. R. Qi, D. Anguelov, Spg: Unsupervised domain adaptation for 3d object detection via semantic point generation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 15446–15456.
[223] W. Zheng, W. Tang, L. Jiang, C.-W. Fu, Se-ssd: Self-ensembling single-stage object detector from point cloud, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 14494–14503.
[224] L. Fan, X. Xiong, F. Wang, N. Wang, Z. Zhang, Rangedet: In defense of range view for lidar-based 3d object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 2918–2927.
[225] C. He, H. Zeng, J. Huang, X.-S. Hua, L. Zhang, Structure aware single-stage 3d object detection from point cloud, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11873–11882.
[226] S. Shi, Z. Wang, J. Shi, X. Wang, H. Li, From points to parts: 3d object detection from point cloud with part-aware and part-aggregation network, IEEE transactions on pattern analysis and machine intelligence 43 (8) (2020) 2647–2664.
[227] W. Shi, R. Rajkumar, Point-gnn: Graph neural network for 3d object detection in a point cloud, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 1711–1719.
[228] Z. Yang, Y. Sun, S. Liu, J. Jia, 3dssd: Point-based 3d single stage object detector, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 11040–11048.
[229] Z. Liu, X. Zhao, T. Huang, R. Hu, Y. Zhou, X. Bai, Tanet: Robust 3d object detection from point clouds with triple attention, in: Proceedings of the AAAI conference on artificial intelligence, Vol. 34, 2020, pp. 11677–11684.
[230] H. Tang, Z. Liu, S. Zhao, Y. Lin, J. Lin, H. Wang, S. Han, Searching efficient 3d architectures with sparse point-voxel convolution, in: European conference on computer vision, Springer, 2020, pp. 685–702.
[231] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, H. Li, Pv-rcnn: Point-voxel feature set abstraction for 3d object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10529–10538.
[232] S. Shi, X. Wang, H. Li, Pointrcnn: 3d object proposal generation and detection from point cloud, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 770–779.
[233] Z. Yang, Y. Sun, S. Liu, X. Shen, J. Jia, Std: Sparse-to-dense 3d object detector for point cloud, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 1951–1960.
[234] Y. Yan, Y. Mao, B. Li, Second: Sparsely embedded convolutional detection, Sensors 18 (10) (2018) 3337.
[235] Y. Zhou, O. Tuzel, Voxelnet: End-to-end learning for point cloud based 3d object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4490–4499.
[236] G. Zhang, L. Fan, C. He, Z. Lei, Z.-X. ZHANG, L. Zhang, Voxel mamba: Group-free state space models for point cloud based 3d object detection, Advances in Neural Information Processing Systems 37 (2024) 81489–81509.
[237] W. Mao, T. Wang, D. Zhang, J. Yan, O. Yoshie, Pillarnest: Embracing backbone scaling and pretraining for pillar-based 3d object detection, IEEE Transactions on Intelligent Vehicles (2024).
[238] G. Zhang, J. Chen, G. Gao, J. Li, S. Liu, X. Hu, Safdnet: A simple and effective network for fully sparse 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14477–14486.
[239] L. Fan, F. Wang, N. Wang, Z. Zhang, Fsd v2: Improving fully sparse 3d object detection with virtual voxels, IEEE

Transactions on Pattern Analysis and Machine Intelligence (2024).
[240] T. Lu, X. Ding, H. Liu, G. Wu, L. Wang, Link: Linear kernel for lidar-based 3d perception, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 1105–1115.
[241] G. Zhang, C. Junnan, G. Gao, J. Li, X. Hu, Hednet: A hierarchical encoder-decoder network for 3d object detection in point clouds, Advances in Neural Information Processing Systems 36 (2023) 53076–53089.
[242] Y. Chen, J. Liu, X. Zhang, X. Qi, J. Jia, Largekernel3d: Scaling up kernels in 3d sparse cnns, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 13488–13498.
[243] D. Zhang, Z. Zheng, H. Niu, X. Wang, X. Liu, Fully sparse transformer 3-d detector for lidar point cloud, IEEE Transactions on Geoscience and Remote Sensing 61 (2023) 1–12.
[244] Z. Wang, Y.-L. Li, X. Chen, H. Zhao, S. Wang, Uni3detr: Unified 3d detection transformer, Advances in Neural Information Processing Systems 36 (2023) 39876–39896.
[245] Z. Tian, X. Chu, X. Wang, X. Wei, C. Shen, Fully convolutional one-stage 3d object detection on lidar range images, Advances in neural information processing systems 35 (2022) 34899–34911.
[246] Y. Hu, Z. Ding, R. Ge, W. Shao, L. Huang, K. Li, Q. Liu, Afdetv2: Rethinking the necessity of the second stage for object detection from point clouds, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 36, 2022, pp. 969–979.
[247] Y. Chen, Y. Li, X. Zhang, J. Sun, J. Jia, Focal sparse convolutional networks for 3d object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5428–5437.
[248] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, C.-L. Tai, Transfusion: Robust lidar-camera fusion for 3d object detection with transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 1090–1099.
[249] H. Fazlali, Y. Xu, Y. Ren, B. Liu, A versatile multi-view framework for lidar-based 3d object detection with guidance from panoptic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 17192–17201.
[250] T. Yin, X. Zhou, P. Krahenbuhl, Center-based 3d object detection and tracking, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 11784–11793.
[251] X. Pan, Z. Xia, S. Song, L. E. Li, G. Huang, 3d object detection with pointformer, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 7463–7472.
[252] Q. Chen, L. Sun, E. Cheung, A. L. Yuille, Every view counts: Cross-view consistency in 3d object detection with hybrid-cylindrical-spherical voxelization, Advances in Neural Information Processing Systems 33 (2020) 21224–21235.
[253] J. Yin, J. Shen, C. Guan, D. Zhou, R. Yang, Lidar-based online 3d video object detection with graph-based message passing and spatiotemporal transformer attention, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11495–11504.
[254] B. Zhu, Z. Jiang, X. Zhou, Z. Li, G. Yu, Class-balanced grouping and sampling for point cloud 3d object detection, arXiv preprint arXiv:1908.09492 (2019).
[255] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, O. Beijbom, Pointpillars: Fast encoders for object detection from point clouds, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2019, pp. 12697–12705.
[256] X. Peng, M. Tang, H. Sun, B. Kay, L. Servadei, R. Wille, Elmar: Enhancing lidar detection with 4d radar motion awareness and cross-modal uncertainty, arXiv preprint arXiv:2506.17958 (2025).
[257] X. Peng, Y. Wang, M. Tang, B. Kay, L. Servadei, R. Wille, Moral: Motion-aware multi-frame 4d radar and lidar fusion for robust 3d object detection, arXiv preprint arXiv:2505.09422 (2025).
[258] X. Huang, Z. Xu, H. Wu, J. Wang, Q. Xia, Y. Xia, J. Li, K. Gao, C. Wen, C. Wang, L4dr: Lidar-4dradar fusion for weather-robust 3d object detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, 2025, pp. 3806–3814.
[259] X. Peng, H. Sun, K. Bierzynski, A. Fischbacher, L. Servadei, R. Wille, Mutualforce: Mutual-aware enhancement for 4d radar-lidar 3d object detection, in: ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2025, pp. 1–5.
[260] J. Deng, G. Chan, H. Zhong, C. X. Lu, Robust 3d object detection from lidar-radar point clouds via cross-modal feature augmentation, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 6585–6591.
[261] R. Xu, Z. Xiang, Rlnet: Adaptive fusion of 4d radar and lidar for 3d object detection, in: European Conference on Computer Vision, Springer, 2024, pp. 181–194.
[262] L. Wang, X. Zhang, B. Xv, J. Zhang, R. Fu, X. Wang, L. Zhu, H. Ren, P. Lu, J. Li, et al., Interfusion: Interaction-based 4d radar and lidar fusion for 3d object detection, in: 2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2022, pp. 12247–12253.
[263] Z. Yu, B. Qiu, A. W. Khong, Vikienet: Towards efficient 3d object detection with virtual key instance enhanced network, in: Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 11844–11853.
[264] H. Mushtaq, S. Latif, M. S. B. Ilyas, S. M. Mohsin, M. Ali, Cls-3d: Content-wise lidar-camera fusion and slot reweighting transformer for 3d object detection in autonomous vehicles, IEEE Access (2025).
[265] J. Shen, Z. Fang, J. Huang, Point-level fusion and channel attention for 3d object detection in autonomous driving, Sensors 25 (4) (2025) 1097.
[266] Y. Li, F. Zeng, R. Lai, T. Wu, J. Guan, A. Zhu, Z. Zhu, Tinyfusiondet: Hardware-efficient lidar-camera fusion

framework for 3d object detection at edge, IEEE Transactions on Circuits and Systems for Video Technology (2025).

[267] M. Uzair, J. Dong, R. Shi, H. Mushtaq, I. Ullah, Channel-wise and spatially-guided multimodal feature fusion network for 3d object detection in autonomous vehicles, IEEE Transactions on Geoscience and Remote Sensing (2024).

[268] X. Li, T. Ma, Y. Hou, B. Shi, Y. Yang, Y. Liu, X. Wu, Q. Chen, Y. Li, Y. Qiao, et al., Logonet: Towards accurate 3d object detection with local-to-global cross-modal fusion, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 17524–17534.

[269] H. Wu, C. Wen, S. Shi, X. Li, C. Wang, Virtual sparse convolution for multimodal 3d object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 21653–21662.

[270] M. Chen, P. Liu, H. Zhao, Lidar-camera fusion: Dual transformer enhancement for 3d object detection, Engineering Applications of Artificial Intelligence 120 (2023) 105815.

[271] S. Li, K. Geng, G. Yin, Z. Wang, M. Qian, Mvmm: Multiview multimodal 3-d object detection for autonomous driving, IEEE Transactions on Industrial Informatics 20 (1) (2023) 845–853.

[272] Y. Qin, C. Wang, Z. Kang, N. Ma, Z. Li, R. Zhang, Supfusion: Supervised lidar-camera fusion for 3d object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 22014–22024.

[273] H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li, Y. Zhang, Vpfnet: Improving 3d object detection with virtual point based lidar and stereo data fusion, IEEE Transactions on Multimedia 25 (2022) 5291–5304.

[274] X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu, D. Cai, Sparse fuse dense: Towards high quality 3d detection with depth completion, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 5418–5427.

[275] Y. Li, X. Qi, Y. Chen, L. Wang, Z. Li, J. Sun, J. Jia, Voxel field fusion for 3d object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 1120–1129.

[276] H. Yang, Z. Liu, X. Wu, W. Wang, W. Qian, X. He, D. Cai, Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph, in: European conference on computer vision, Springer, 2022, pp. 662–679.

[277] Y. Zhang, J. Chen, D. Huang, Cat-det: Contrastively augmented transformer for multi-modal 3d object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 908–917.

[278] W. Chen, P. Li, H. Zhao, Msl3d: 3d object detection from monocular, stereo and point cloud for autonomous driving, Neurocomputing 494 (2022) 23–32.

[279] P. An, J. Liang, K. Yu, B. Fang, J. Ma, Deep structural information fusion for 3d object detection on lidar–camera system, Computer Vision and Image Understanding 214 (2022) 103295.

[280] Z. Liu, T. Huang, B. Li, X. Chen, X. Wang, X. Bai, Epnet++: Cascade bi-directional fusion for multi-modal 3d object detection, IEEE transactions on pattern analysis and machine intelligence 45 (7) (2022) 8324–8341.

[281] S. Pang, D. Morris, H. Radha, Clocs: Camera-lidar object candidates fusion for 3d object detection, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE, 2020, pp. 10386–10393.

[282] T. Huang, Z. Liu, X. Chen, X. Bai, Epnet: Enhancing point features with image semantics for 3d object detection, in: European conference on computer vision, Springer, 2020, pp. 35–52.

[283] S. Vora, A. H. Lang, B. Helou, O. Beijbom, Pointpainting: Sequential fusion for 3d object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 4604–4612.

[284] J. Ku, M. Mozifian, J. Lee, A. Harakeh, S. L. Waslander, Joint 3d proposal generation and object detection from view aggregation, in: 2018 IEEE/RSJ international conference on intelligent robots and systems (IROS), IEEE, 2018, pp. 1–8.

[285] D. Xu, D. Anguelov, A. Jain, Pointfusion: Deep sensor fusion for 3d bounding box estimation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 244–253.

[286] C. R. Qi, W. Liu, C. Wu, H. Su, L. J. Guibas, Frustum pointnets for 3d object detection from rgb-d data, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 918–927.

[287] X. Chen, H. Ma, J. Wan, B. Li, T. Xia, Multi-view 3d object detection network for autonomous driving, in: Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2017, pp. 1907–1915.

[288] Z. Wang, Z. Huang, Y. Gao, N. Wang, S. Liu, Mv2dfusion: Leveraging modality-specific object semantics for multi-modal 3d detection, arXiv preprint arXiv:2408.05945 (2024).

[289] M. Chen, M. Yang, Y. Zhang, T. Han, X. Li, H. Zhao, P. Liu, Multi-modal bev enhancement fusion for 3d object detection in autonomous driving, IEEE Transactions on Intelligent Transportation Systems (2025).

[290] J. Huang, Y. Ye, Z. Liang, Y. Shan, D. Du, Detecting as labeling: Rethinking lidar-camera fusion in 3d object detection, in: European Conference on Computer Vision, Springer, 2024, pp. 439–455.

[291] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, S. Han, Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation, arXiv preprint arXiv:2205.13542 (2022).

[292] Y. Xie, C. Xu, M.-J. Rakotosaona, P. Rim, F. Tombari, K. Keutzer, M. Tomizuka, W. Zhan, Sparsefusion: Fusing multi-modal sparse representations for multi-sensor 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 17591–17602.

[293] J. Yan, Y. Liu, J. Sun, F. Jia, S. Li, T. Wang, X. Zhang, Cross modal transformer: Towards fast and robust 3d object detection, in: Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 18268–18278.

[294] X. Chen, T. Zhang, Y. Wang, Y. Wang, H. Zhao, Futr3d: A unified sensor fusion framework for 3d detection, in: proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 172–181.

[295] Y. Chen, Z. Yu, Y. Chen, S. Lan, A. Anandkumar, J. Jia, J. M. Alvarez, Focalformer3d: focusing on hard instance for 3d object detection, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8394–8405.

[296] C. Hu, H. Zheng, K. Li, J. Xu, W. Mao, M. Luo, L. Wang, M. Chen, Q. Peng, K. Liu, et al., Fusionformer: A multi-sensory fusion in bird's-eye-view and temporal consistent transformer for 3d object detection, arXiv preprint arXiv:2309.05257 (2023).

[297] H. Hu, F. Wang, J. Su, Y. Wang, L. Hu, W. Fang, J. Xu, Z. Zhang, Ea-lss: Edge-aware lift-splat-shot framework for 3d bev object detection, arXiv preprint arXiv:2303.17895 (2023).

[298] S. Xu, D. Zhou, J. Fang, J. Yin, Z. Bin, L. Zhang, Fusionpainting: Multimodal fusion with adaptive attention for 3d object detection, in: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), IEEE, 2021, pp. 3047–3054.

[299] T. Yin, X. Zhou, P. Krähenbühl, Multimodal virtual point 3d detection, Advances in Neural Information Processing Systems 34 (2021) 16494–16507.

[300] C. Wang, C. Ma, M. Zhu, X. Yang, Pointaugmenting: Cross-modal augmentation for 3d object detection, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 11794–11803.

[301] J. H. Yoo, Y. Kim, J. Kim, J. W. Choi, 3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection, in: European conference on computer vision, Springer, 2020, pp. 720–736.

[302] A. Vahidi-Moghaddam, S. P. H. Boroujeni, I. Jebellat, E. Jebellat, N. Mehrabi, Z. Li, From shadow to light: Toward safe and efficient policy learning across mpc, deepc, rl, and llm agents, arXiv preprint arXiv:2510.04076 (2025).

[303] S. Yang, J. Liu, R. Zhang, M. Pan, Z. Guo, X. Li, Z. Chen, P. Gao, H. Li, Y. Guo, et al., Lidar-llm: Exploring the potential of large language models for 3d lidar understanding, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, 2025, pp. 9247–9255.

[304] T. Tang, D. Wei, Z. Jia, T. Gao, C. Cai, C. Hou, P. Jia, K. Zhan, H. Sun, F. JingChen, et al., Bev-tsr: Text-scene retrieval in bev space for autonomous driving, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, 2025, pp. 7275–7283.

[305] M. Peng, X. Guo, X. Chen, K. Chen, M. Zhu, L. Chen, F.-Y. Wang, Lc-llm: Explainable lane-change intention and trajectory predictions with large language models, Communications in Transportation Research 5 (2025) 100170.

[306] Z. Lan, L. Liu, B. Fan, Y. Lv, Y. Ren, Z. Cui, Traj-llm: A new exploration for empowering trajectory prediction with pre-trained large language models, IEEE Transactions on Intelligent Vehicles (2024).

[307] Z. Xu, Y. Zhang, E. Xie, Z. Zhao, Y. Guo, K.-Y. K. Wong, Z. Li, H. Zhao, Drivegpt4: Interpretable end-to-end autonomous driving via large language model, IEEE Robotics and Automation Letters (2024).

[308] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 26296–26306.

[309] L. Chen, O. Sinavski, J. Hünermann, A. Karnsund, A. J. Willmott, D. Birch, D. Maund, J. Shotton, Driving with llms: Fusing object-level vector modality for explainable autonomous driving, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 14093–14100.

[310] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, G. Corrado, Gaia-1: A generative world model for autonomous driving, arXiv preprint arXiv:2309.17080 (2023).

[311] A. Elhafsi, R. Sinha, C. Agia, E. Schmerling, I. A. Nesnas, M. Pavone, Semantic anomaly detection with large language models, Autonomous Robots 47 (8) (2023) 1035–1055.

[312] W. Wang, J. Xie, C. Hu, H. Zou, J. Fan, W. Tong, Y. Wen, S. Wu, H. Deng, Z. Li, et al., Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving, arXiv preprint arXiv:2312.09245 (2023).

[313] B. Jin, H. Liu, Adapt: Action-aware driving caption transformer, in: CAAI International Conference on Artificial Intelligence, Springer, 2023, pp. 473–477.

[314] J. Mao, Y. Qian, J. Ye, H. Zhao, Y. Wang, Gpt-driver: Learning to drive with gpt, arXiv preprint arXiv:2310.01415 (2023).

[315] X. Zhou, L. Shan, X. Gui, Dynrsl-vlm: Enhancing autonomous driving perception with dynamic resolution vision-language models, arXiv preprint arXiv:2503.11265 (2025).

[316] Y. Zhou, C. Cui, J. Peng, Z. Yang, J. Lu, J. H. Panchal, B. Yao, Z. Wang, A hierarchical test platform for vision language model (vlm)-integrated real-world autonomous driving, arXiv preprint arXiv:2506.14100 (2025).

[317] A. Taparia, N. Ngu, M. Leiva, J. S. Kricheli, J. Corcoran, N. D. Bastian, G. Simari, P. Shakarian, R. Senanayake, Vlc fusion: Vision-language conditioned sensor fusion for robust object detection, arXiv preprint arXiv:2505.12715 (2025).

[318] S. Shriram, S. Perisetla, A. Keskar, H. Krishnaswamy, T. E. W. Bossen, A. Møgelmose, R. Greer, Towards a multi-agent vision-language system for zero-shot novel hazardous object detection for autonomous driving safety, arXiv preprint arXiv:2504.13399 (2025).

[319] D. Wu, W. Han, Y. Liu, T. Wang, C.-z. Xu, X. Zhang, J. Shen, Language prompt for autonomous driving, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 39, 2025, pp. 8359–8367.

[320] L. He, Y. Sun, S. Wu, J. Liu, X. Huang, Integrating object detection modality into visual language model for enhanced autonomous driving agent, arXiv preprint arXiv:2411.05898 (2024).

[321] Z. Huang, Z. Sheng, Y. Qu, J. You, S. Chen, Vlm-rl: A unified vision language models and reinforcement learning framework for safe autonomous driving, arXiv preprint arXiv:2412.15544 (2024).

[322] Y. Xu, Y. Hu, Z. Zhang, G. P. Meyer, S. K. Mustikovela, S. Srinivasa, E. M. Wolff, X. Huang, Vlm-ad: End-to-end autonomous driving through vision-language model supervision, arXiv preprint arXiv:2412.14446 (2024).

[323] A. Gopalkrishnan, R. Greer, M. Trivedi, Multi-frame, lightweight & efficient vision-language models for question answering in autonomous driving, arXiv preprint arXiv:2403.19838 (2024).

[324] J. Li, D. Li, S. Savarese, S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in: International conference on machine learning, PMLR, 2023, pp. 19730–19742.

[325] Z. Jia, T. Gao, C. Cai, C. Hou, P. Jia, F. JingChen, Y. ZHAO, K. Zhan, F. LIU, Y. WANG, et al., Bev-clip: Multi-modal bev retrieval methodology for complex scene in autonomous driving (2024).

[326] T. Choudhary, V. Dewangan, S. Chandhok, S. Priyadarshan, A. Jain, A. K. Singh, S. Srivastava, K. M. Jatavallabhula, K. M. Krishna, Talk2bev: Language-enhanced bird's-eye view maps for autonomous driving, in: 2024 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2024, pp. 16345–16352.

[327] W. Cheng, J. Yin, W. Li, R. Yang, J. Shen, Language-guided 3d object detection in point cloud for autonomous driving, arXiv preprint arXiv:2305.15765 (2023).

[328] D. Wu, W. Han, T. Wang, X. Dong, X. Zhang, J. Shen, Referring multi-object tracking, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2023, pp. 14633–14642.

[329] R. Chen, Y. Liu, L. Kong, X. Zhu, Y. Ma, Y. Li, Y. Hou, Y. Qiao, W. Wang, Clip2scene: Towards label-efficient 3d scene understanding by clip, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7020–7030.

[330] M. Najibi, J. Ji, Y. Zhou, C. R. Qi, X. Yan, S. Ettinger, D. Anguelov, Unsupervised 3d perception with 2d vision-language distillation for autonomous driving, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8602–8612.