Received: 3 April 2025 | Revised: 9 June 2025 | Accepted: 19 June 2025 | Published online: 21 July 2025

REVIEW

Journal of Computational and Cognitive Engineering
2025, Vol. 00(00) 1–20
DOI: 10.47852/bonviewJCCE52025836

BON VIEW PUBLISHING

Advances in Deep Learning for Autonomous Vehicle Perception: A Comprehensive Review

Namitha Kalakunnath[1,*], Aneesh Varghese[2], Abekaesh Prakash Anuradha[1], Dhanush Kumar Girish[1] and Renjith Sasidharan[1]

[1]*Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham Amritapuri, India*

[2]*Amazon Web Services Texas, USA*

Abstract: Autonomous vehicle (AV) perception tasks are critical for enabling self-driving cars to navigate complex environments, relying on advanced technologies to interpret and understand the surrounding world through sensors, deep learning models, and sensor fusion techniques. This review paper provides a comprehensive overview of deep learning architectures applied to AV perception tasks, with a particular focus on recent advancements from 2019 to 2024. The paper begins by examining 3D object detection techniques, exploring the state-of-the-art methods developed during the past six years. Moreover, object localization innovations are discussed, pointing out certain key advancements in that area. The paper also discusses sensor fusion techniques and how they are central to improving performance in 3D object detection. Finally, the discussion encompasses various environmental perception methods such as road and lane detection and traffic sign recognition, all of which are crucial for the safe and efficient operation of AVs. This paper aims to provide insights into the evolving landscape of AV perception and its applications in intelligent transportation systems.

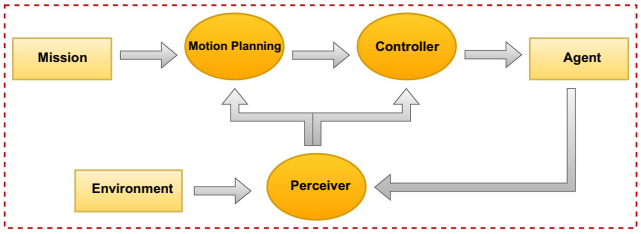
Keywords: autonomous vehicle, perception, object detection, sensor fusion, object localization, deep learning

1. Introduction

Autonomous vehicles (AVs) are reshaping transportation by offering enhanced safety, reduced traffic congestion, and improved energy efficiency [1]. For safe navigation, AVs must accurately perceive and interpret their surroundings. Deep learning has emerged as a critical technology enabling AVs to reliably perform core tasks, including precise 3D object detection, accurate object localization, effective sensor fusion, and robust environmental perception [2, 3]. The deep learning innovations for AV perception are essential for addressing challenges posed by complex driving conditions, such as dynamic environments, varying weather, and the demand for real-time processing. However, the rapid evolution and diversity of these methodologies have complicated the knowledge landscape, making it challenging for researchers and practitioners to effectively summarize, integrate, and build upon recent developments.

The autonomy of vehicle systems heavily relies on robust environmental perception, sophisticated motion planning, and accurate control mechanisms [4]. Figure 1 illustrates the architecture of such autonomous systems. Utilizing a perception module [5], these systems interpret both their internal states and external environments through techniques like visual localization, mapping, and environmental analysis. They integrate the results of perception [6] with motion planning to generate suitable control commands, enabling precise navigation and the autonomous completion of complex tasks. Effective understanding and interpretation of the surrounding environment are essential for autonomous systems [7]. Visual simultaneous localization and mapping (vSLAM) significantly enhances autonomy by allowing robots to determine their position and create environmental models using visual sensors. With vSLAM, autonomous systems can effectively gather environmental data, build accurate environment models, and assess their internal status, thus increasing their intelligence and operational autonomy.

Figure 1
Architecture of autonomous systems



Accurate perception and understanding of the environment are essential for autonomous systems. Visual simultaneous localization and mapping (vSLAM) algorithms are widely employed to build various environmental representations, including sparse, semi-dense, and dense maps, depending on specific application requirements [8]. The AV's state, comprising its position and orientation, must be continuously monitored, as real-time state estimation is foundational for autonomous control [9]. Despite their effectiveness in self-localization and ego-motion estimation, contemporary

***Corresponding author:** Namitha Kalakunnath, Department of Computer Science and Engineering, Amrita Vishwa Vidyapeetham Amritapuri, India. Email: namithak@am.amrita.edu

vSLAM methods often rely on restrictive assumptions, such as static scenes and photometric consistency [10].

This review paper provides a thorough information resource for scholars and practitioners by examining cutting-edge methodologies and their practical applications. This review's primary contributions are as follows:

1) This article examines the latest advances (2019–2024) in deep learning architectures employed in AV perception systems.
2) It highlights state-of-the-art 3D object detection methods for understanding vehicle environments, combined with an analysis of object localization approaches designed to improve spatial awareness and accuracy in autonomous systems.
3) It provides a detailed overview of sensor fusion strategies that integrate data from Light Detection and Ranging (LiDAR), cameras, and radar to support robust perception capabilities.
4) It evaluates key environmental perception techniques, including road and lane detection and traffic sign recognition, which are critical for reliable and safe autonomous navigation.

1.1. Search strategy and inclusion criteria

This study follows a four-step procedure for identifying and selecting articles, as illustrated in Figure 2. The preliminary search process began by identifying papers by keywords through standard electronic databases such as Google Scholar, Scopus, ACM, Springer Link, Elsevier, IEEE Explore, MDPI, and Wiley, including journal articles, book chapters, and conference papers published within the last six years (2019–2024). The primary keywords utilized to identify relevant articles were "autonomous vehicle perception," "deep learning," "3D object detection," "object localization," "sensor fu sion," "lane detection," and "traffic sign recognition" with logical operators (AND/OR) between each keyword. The articles are initially evaluated using the criteria outlined in Figure 3. In all, only 116 articles met the inclusion criteria for this review. Table 1 outlines the key acronyms used in the article.

Figure 2
PRISMA model used in the literature review

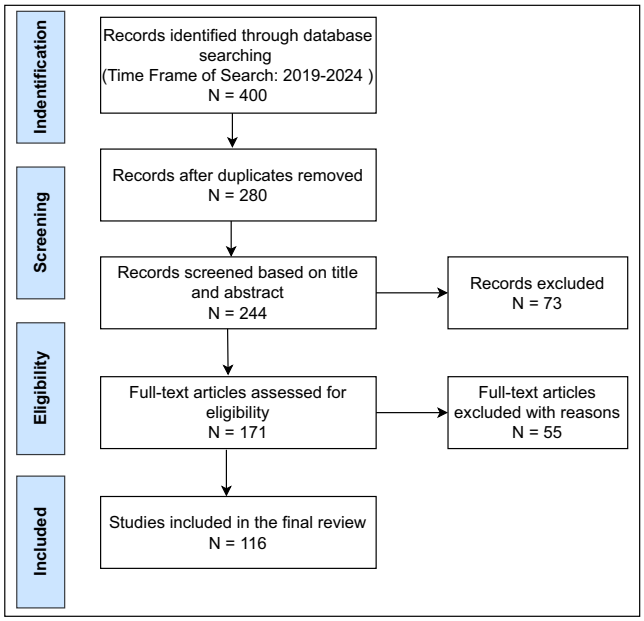


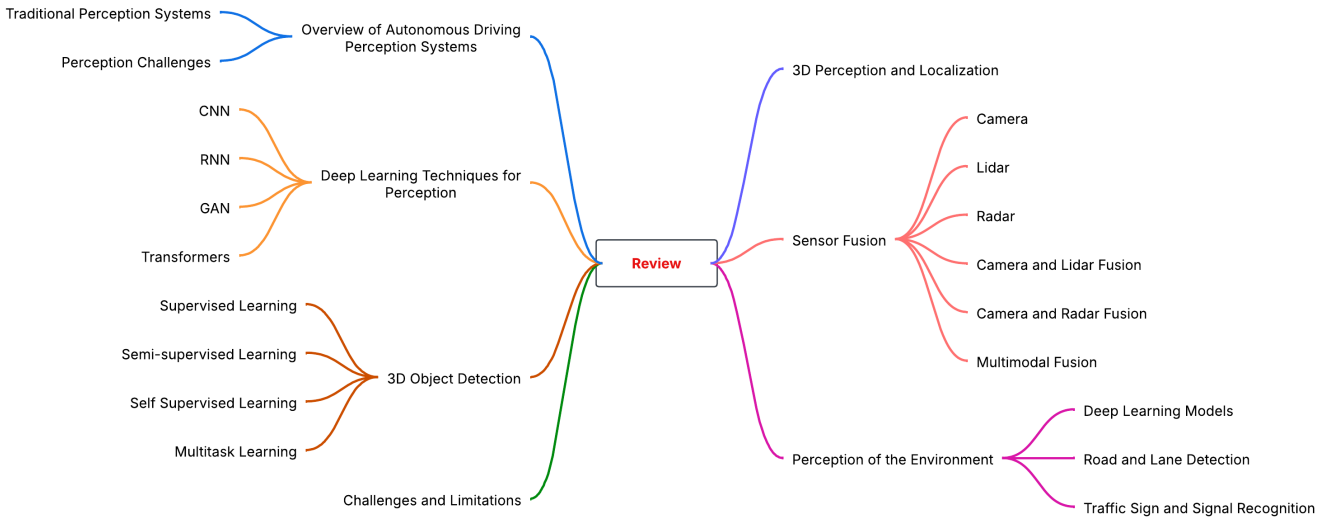
Figure 3
Inclusion criteria for the paper selection process Figure_01

Table 1
Acronyms and descriptions

Acronyms	Description
AV	Autonomous Vehicle
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
GAN	Generative Adversarial Networks
vSLAM	Visual Simultaneous Localization and Mapping
AP	Average Precision
GRU	Gated Recurrent Unit
MLP	Multilayer Perceptron
IMU	Inertial Measurement Unit
LiDAR	Light Detection and Ranging
ROI	Region of Interest
SDTM	Simultaneous Dynamic Triangulation Mapping
LFA	Local Feature Aggregation
BEV	Bird's Eye View
ADS	Autonomous Driving Systems
TSSR	Traffic Sign and Signal Recognition

The article is structured into several sections: Section 2 reviews AV perception systems, including traditional systems and their key components. Section 3 analyzes deep learning approaches used in perception tasks, while Section 4 summarizes 3D object detection methods. Section 5 covers techniques for the 3D perception and localization of AVs. Section 6 explores sensor fusion strategies, and Section 7 addresses specific challenges related to environmental perception. Section 8 highlights the limitations and ongoing challenges, and Section 9 suggests directions for future research. The whole structure of this study is depicted in Figure 4.

Figure 4
Structure of our study



2. Overview of Autonomous Driving Perception Systems

Perception involves gathering data from sensory inputs and the environment, interpreting it to achieve meaningful comprehension, and subsequently employing this information to inform actions. In AVs, perception relies on several onboard sensors that collect diverse data types, which are analyzed using sophisticated computational algorithms and machine learning models to assess the vehicle's environment and guide driving decisions. Comprehending perception in AVs necessitates knowledge of the sensors incorporated inside these vehicles and the unique methodologies utilized for interpreting sensory data.

2.1. Traditional perception systems

The development of AVs began with the establishment of sophisticated perception systems designed to ensure safe navigation. The primary objective of these systems is to assist vehicles in identifying and comprehending critical environmental components, such as other vehicles, pedestrians, roadways, and traffic signals. Essential perception tasks encompass localization and mapping, object detection and tracking, lane marker recognition, traffic sign identification, and comprehensive scene understanding, necessitating sophisticated sensing technologies and dependable data processing techniques.

Traditionally, AVs had various sensor arrays, each executing distinct functions. Cameras proficiently recorded visual elements like lane markings, traffic signs, and adjacent objects, although they were susceptible to disruptions caused by fluctuating light conditions, shadows, and visibility challenges [11]. Radar devices offered reliable readings of object distances and velocities irrespective of weather conditions, crucial for monitoring vehicles and barriers; nevertheless, they were deficient in the spatial resolution characteristics of cameras and LiDAR. LiDAR sensors utilize laser pulses to create intricate three-dimensional maps, providing accurate spatial measurements essential for mapping and object detection; nevertheless, they are expensive and less effective under adverse weather conditions. Ultrasonic sensors, defined by limited range and low resolution, were predominantly appropriate for parking assistance and close-range object avoidance.

2.2. Perception challenges in autonomous driving

Despite their innovative designs, traditional perception systems in AVs encountered several significant challenges, the most critical being limited precision and robustness. These systems often failed under variable lighting conditions, adverse weather, or visual obstructions [12]. For instance, a lane detection system relying on cameras might misinterpret shadows cast by vehicles or trees as lane markings, leading to incorrect navigation. Such inconsistencies emphasized the need for more reliable and adaptive solutions.

Another major hurdle was the complexity involved in sensor fusion. Merging data from multiple sensors proved to be error-prone, especially when the temporal or spatial alignment of sensor outputs was not perfectly synchronized. Misaligned data could result in flawed interpretations and unreliable decisions, complicating the development of a coherent and accurate perception framework. The intricacies of this integration process posed a serious obstacle to achieving smooth and dependable performance.

Additionally, the high computational demands of traditional systems created further limitations. Processing data in real-time from high-resolution sensors, such as LiDAR, requires powerful and expensive computing resources. The vast datasets generated needed to be analyzed rapidly to avoid delays in decision-making, but this often restricted the complexity of algorithms that could be used effectively. Moreover, these systems lacked flexibility and scalability, making it difficult to adapt to new environments or evolving traffic regulations. For example, a traffic sign recognition system might struggle to accommodate newly introduced signs or regional variations without significant reconfiguration, highlighting the urgent need for more adaptable and future-ready perception systems.

3. Deep Learning for Autonomous Perception

This section categorizes 3D object detection methods according to their training strategies, addressing the training processes of various deep learning models.

3.1. Convolutional neural networks (CNNs)

The automobile's vision system, or perception system, is where the quest for accuracy and safety starts at the forefront of self-driving car technology, when time is of the essence. Imagine a car navigating the streets of a city while identifying other cars, pedestrians, and traffic signs. Convolutional neural networks (CNNs) hold the key to this miracle [13–24]. They break down images into their important elements with the aid of several layers that resemble those seen in human visual cortices. For example, they are able to distinguish between various types of things based just on their shapes and recognize stop signs and lane markers. The basis for an AV's understanding of its surroundings is its capacity to recognize spatial hierarchies in pictures.

3.2. Recurrent neural networks (RNNs)

On the other hand, the world is a dynamic entity. As a result, driving presents a dynamic situation that necessitates anticipating not just what is currently happening but also what might happen in the future. Recurrent neural networks are useful in this situation. These networks are adept at managing sequences, recalling past inputs to predict subsequent events [25, 26]. Consider a self-driving car that can identify a cyclist's path or the movements of nearby cars. When presented with complex scenarios, the vehicle uses RNNs to analyze temporal trends during its journey and make informed decisions.

3.3. Generative adversarial networks (GANs)

However, what happens if there isn't enough training data to cover every scenario? Generative adversarial networks (GANs) are useful in this situation. GANs provide realistic simulations of wide-ranging driving conditions by using a dual network architecture, one of which generates data and the other assesses its authenticity [27]. The development of perception systems that can handle uncommon but significant events, such as odd road designs and strange climates, depends on such synthetic datasets.

3.4. Transformers and attention mechanisms

The car needs to focus on key elements and ignore irrelevant details in order to comprehend complex pictures. Transformers and attention processes convert this by giving distinct input information parts with varying degrees of relevance. By helping automobiles comprehend context, these attention techniques allow them to distinguish between a pedestrian and a background advertising board, resulting in more accurate and sophisticated decisions [18, 28–30]. Table 2 summarizes the major deep learning architectures used for performing various tasks in autonomous driving perception.

4. 3D Object Detection

The growth of AV perception has moved from 2D to more complex 3D object identification, with depth information playing an important role. This review categorizes contemporary deep learning

Table 2
Summary of deep learning architectures in autonomous driving

Article	Year	Article type	Architecture type	Perception tasks	Dataset
[13]	2023	Journal	CNN	Vehicle dynamics modeling and control	CarSim
[28]	2024	Journal	Attention-based Transformer	Egocentric gesture recognition	EgoGesture, Interactive Museum, Ego-Driving (to be released)
[14]	2024	Journal	CNN	Instance and drivable segmentation	BDD100K
[15]	2021	Journal	Attention-based CNN	Panoramic semantic segmentation	Cityscapes, Mapillary Vistas, IDD20K, PASS, WildPASS
[16]	2024	Journal	CNN	Free road segmentation	KITTI, R2D
[17]	2020	Journal	CNN	Lane detection	CULane, TuSimple, CurveLanes
[18]	2023	Journal	Attention-based CNN	Object detection	Custom Dataset of 800 Objects
[19]	2022	Journal	CNN	Object detection	MSDataset
[29]	2023	Journal	Attention-based Transformer	Object detection	–
[20]	2023	Journal	CNN	Object detection	Local Bandung Dataset (15,717 images)
[21]	2022	Journal	CNN	Uncertainty-aware decisions	PeSOTIF
[22]	2022	Journal	CNN	Semantic segmentation	–
[30]	2024	Journal	Transformer, CNN	Lane detection	CULane, TuSimple, CurveLanes
[27]	2021	Conference	GAN	Visual data generation	KITTI, CityScapes
[25]	2021	Journal	RNN, LSTM, GRU	Motion prediction, Motion planning	NGSIM data
[23]	2020	Conference	CNN	Object detection	KITTI
[24]	2019	Conference	CNN	Object detection	KITTI

Figure 5
Classification of 3D object detection methods based on training strategies

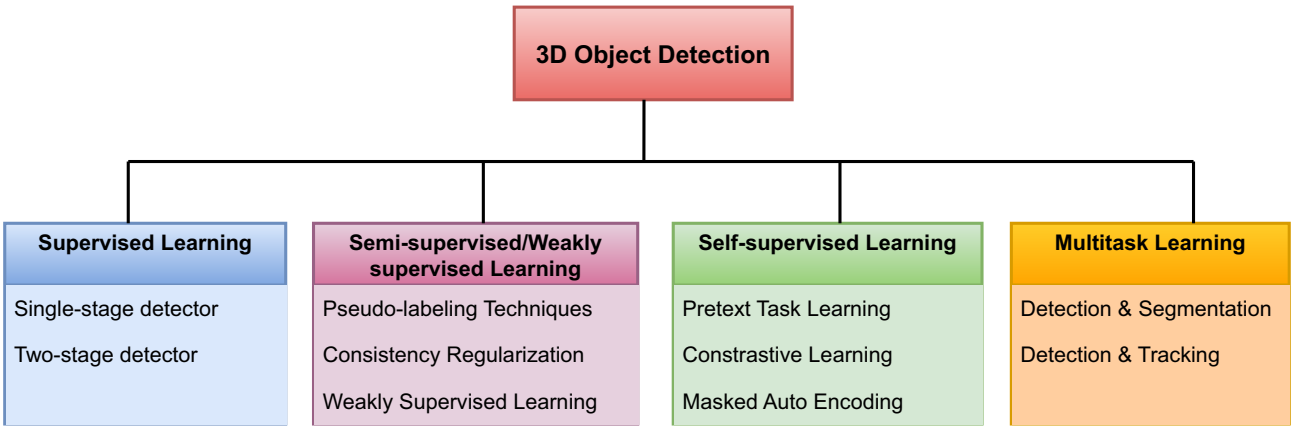


Table 3
Comparison of techniques in 3D object detection

Learning technique	Representative methods	Advantages	Limitations
Single-stage detectors	CenterNet3D, 3DSSD, SMOKE	Real-time inference, simpler pipeline, anchor-free architectures	Lower precision, struggles with occlusions and small objects
Two-stage detectors	PV-RCNN, Voxel RCNN, CT3D	High accuracy, better refinement of proposals, strong multi-scale features	Higher latency, computationally expensive
Semi-supervised learning	Mix-Teaching, 3DIoUMatch	Reduces the need for labeled data, enables scalable training	Quality of pseudo-labels critical, risk of error propagation
Weakly supervised learning	WeakM3D, VS3D, MonoGRNet	Leverages cheap annotations (e.g., 2D boxes), useful in early data stages	Lower generalization, often task- or modality-specific
Self-supervised learning	MonoRUn, BEVSOC, GeoMAE	Utilizes unlabeled data, enhances pretraining and robustness	Pretext tasks must be carefully designed, not always task-specific
Multitask learning	HDMapNet	Joint detection, segmentation, tracking; improves resource efficiency	Training complexity, risk of negative transfer across tasks

algorithms for 3D object detection based on the learning strategies used. The training strategies used to develop the models are the basis for the categorization of 3D object detection methods in Figure 5. Learning paradigms are classified into four primary categories: supervised learning trains models with labeled data, semi-supervised or weakly supervised learning uses both labeled and unlabeled data, self-supervised learning learns representations from data without annotations, and multitask learning trains a single model to solve multiple related tasks at the same time, allowing for shared representations that can improve 3D object detection performance. Each of these learning methods provides distinct benefits and drawbacks in terms of data requirements, model complexity, and total detection accuracy, as summarized in Table 3.

4.1. Supervised learning

4.1.1. Single-stage detectors

Single-stage 3D object detectors have made major advances in AV perception, with an emphasis on balancing accuracy and real-time performance. CenterNet3D [31] pioneered anchor-free keypoint estimation for direct 3D box regression, whereas 3DSSD [32] optimized performance using a lightweight design and fusion sampling. FCOS3D [33] improved monocular detection by converting 7-DoF targets to the image domain, whereas IA-SSD [34] added instance-aware point selection algorithms. SMOKE [35] improved detection with keypoint estimation and 3D variable regression. SA-SSD [36] enhanced localization, Voxel-FPN [37] merged multi-scale features, HV-Net [38] tackled voxel size difficulties, and HDNet [39] used HD maps as priors. These approaches jointly exhibit the advancement of anchor-free methods and efficient point selection while preserving real-time performance.

4.1.2. Two-stage detectors

Through the region proposal and refining stages, two-stage 3D object detectors greatly improved AV perception. PV-RCNN [40] integrated voxel CNN and PointNet-based features, whereas Voxel RCNN [41] obtained good accuracy with voxel-only representations. Fast Point RCNN [42] combined voxel and point cloud processing with attention mechanisms, while Part-A^2 net [43] introduced part-aware supervision to enhance proposal quality. Recent improvements include HPV-RCNN [44] with multi-attentive voxel encoding, and CT3D [45] with channel-wise transformers for proposal refinement. These methods highlight the field's progress toward multi-representation learning and complex architectures while maintaining excellent accuracy for self-driving applications. Table 4 shows the comparison of models evaluated with the KITTI test set. The performance comparisons in Table 4 are derived directly from the official KITTI 3D object detection benchmark leaderboard

Table 4
Benchmark results of supervised learning models on KITTI 3D object detection test set

Model	Environment	Speed	Stage	Car 3D AP (%)			Pedestrian 3D AP(%)			Cyclist 3D AP(%)		
		(fps)		Easy	Moderate	Hard	Easy	Moderate	Hard	Easy	Moderate	Hard
CenterNet3D [31]	GPU @ 1.5 Ghz	–	One	88.23	79.23	75.34	53.32	49.42	41.63	–	–	–
3DSSD [32]	GPU @ 2.5 Ghz	–	One	88.36	79.57	74.55	54.64	44.27	40.23	–	–	–
IA-SSD (single) [34]	1 core @ 2.5 Ghz	85	One	88.87	80.32	75.10	49.01	41.20	38.03	80.78	66.01	58.12
IA-SSD (multiclass) [34]	1 core @ 2.5 Ghz	83	One	88.34	80.13	75.04	46.51	39.03	35.60	78.35	61.94	55.70
SA-SSD [36]	1 core @ 2.5 Ghz	25	One	95.03	91.03	85.96	–	–	–	–	–	–
Voxel-FPN [37]	GPU @ 2.5 Ghz	50	One	85.64	76.70	69.44	85.25	64.36	61.00	68.77	61.86	56.40
HV-Net [38]	1 core @ 2.5 Ghz	31	One	92.83	88.82	83.38	54.84	48.86	46.33	83.97	71.17	63.65
PV-RCNN [40]	1 core @ 2.5 Ghz	–	Two	90.25	81.43	76.82	52.17	43.29	40.29	78.60	63.71	57.65
Fast Point RCNN [42]	GPU @ 2.5 Ghz	–	Two	84.28	75.73	67.39	–	–	–	–	–	–
Part-A^2 net [43]	GPU @ 2.5 Ghz	–	Two	77.86	85.94	72.00	44.50	54.49	42.36	62.73	78.58	57.74
HPV-RCNN [44]	RTX 4060Ti	–	Two	89.33	80.61	75.53	52.54	43.86	41.56	84.24	69.56	61.42
CT3D [45]	1 core @ 2.5 Ghz	–	Two	87.83	81.77	77.16	–	–	–	–	–	–

[46]. The models were assessed using the KITTI dataset using a standardized benchmark process, which has predetermined training and testing divisions (typically 7481 training images and 7518 test images). All methodologies adhere to the KITTI evaluation parameters for 3D average precision (AP) across varying difficulty levels (Easy, Moderate, Hard).

4.2. Semi-supervised learning

4.2.1. Pseudo-labeling technique

Semi-supervised learning with pseudo-labeling has emerged as a potential solution to 3D object detection for autonomous cars, minimizing the need for manual annotations. Mix-Teaching [47] pioneered instance-level picture patch merging with uncertainty-based filtering, whereas 3DIoUMatch [48] presented a teacher-student framework utilizing 3D IoU estimation for pseudo-label refining. Zhang et al. [49] focused on decreasing annotation workload through teacher-student label generation, while DetMatch [50] employed multimodal fusion of RGB and point cloud data to construct robust pseudo-labels. Recently, 3DSimDet [51] proposed a high-quality pseudo-label generator that leverages predictions from its novel BIoU Head module. The generator demonstrated improved detection performance within a semi-supervised learning framework while maintaining real-time capabilities crucial for autonomous driving applications. These approaches show that labeled and unlabeled data may be combined well using complex pseudo-labeling algorithms, outperforming fully supervised methods while addressing label quality and multimodal integration issues.

4.2.2. Consistency regularization

In the field of AV 3D object detection, consistency regularization has become a crucial semi-supervised learning technique. A privacy-preserving method using de-identified intermediate features was presented by Upcycling [52]. For efficient supervision, hybrid pseudo-labels were combined with feature-level Ground Truth sampling (F-GT). By combining uncertainty-based filtering with instance-level image patching, Mix-Teaching [47] enhanced this field and indirectly enforced consistency through mixed images with high-quality pseudo-labels. In particular, these methods handle privacy issues and computational efficiency while preserving strong detection performance, illustrating the development of consistency regularization in semi-supervised learning.

4.2.3. Weakly supervised learning

A promising method for lowering the annotation burden in 3D object detection for driverless cars is weakly supervised learning. For efficient detection, WeakM3D [53] uses 3D alignment loss and RoI LiDAR points in conjunction with 2D bounding boxes as weak supervision. VS3D [54] uses cross-modal knowledge distillation from pre-trained image networks and presents an unsupervised 3D proposal module based on point cloud densities. Despite not being precisely weakly supervised, MonoGRNet [55] makes a contribution by using geometric reasoning under sparse-depth supervision. These methods address the problem of limited fully annotated 3D data in autonomous driving applications by showing how to use less accurate annotations and 2D information to create effective 3D detection systems.

4.3. Self-supervised learning

4.3.1. Pretext task learning

An inventive method for enhancing 3D object detection in autonomous driving is self-supervised learning via pretext tasks.

Using unlabeled point cloud data, Shi and Rajkumar [56] showed how geometric pretext tasks may effectively pre-train neural networks, resulting in enhanced object heading accuracy and AP. This field was advanced by MonoRUn [57], which introduced a framework that learns geometry and dense correspondences through a variety of pretext tasks, such as regional reconstruction with uncertainty awareness and dense 2D-3D correspondence mapping. These methods show how well-crafted pretext tasks can improve 3D object detection skills while lowering the need for copious manual annotations, offering supplementary features to those discovered from labeled data for reliable autonomous car perception.

4.3.2. Contrastive learning

A potent self-supervised method for 3D object identification in driverless cars is contrastive learning. In order to represent 3D-to-2D relationships without the need for explicit camera calibration, BEVSOC [58] presented a calibration-free technique that makes use of group-equivariant convolutional networks and projection modules. For point cloud object detection, Shi and Rajkumar [56] showed how well contrastive learning works when combined with geometric pretext tasks. They found that while contrastive loss alone increases AP, integrating it with geometric tasks improves both precision and heading accuracy. These methods demonstrate how contrastive learning may be used to improve detection robustness and generalization skills in autonomous driving applications while lowering the need for annotation.

4.3.3. Masked autoencoding

Masked autoencoding (MAE) has become a potent self-supervised pretraining method for autonomous driving's 3D object identification. While GD-MAE [59] suggested a more straightforward paradigm utilizing a generative decoder for flexible masking methods, Occupancy-MAE [60] presented range-aware random masking for voxel-based LiDAR data. Voxel-MAE [61] tackled sparse point cloud issues by masked voxel reconstruction, whereas GeoMAE [62] utilized geometric features through centroid prediction, normal estimation, and curvature prediction targets. Across the main autonomous driving benchmarks, these methods show how MAE can effectively reduce the amount of labeled data needed while enhancing detection performance and generalization capabilities, especially in difficult situations like small object detection and variable point cloud densities.

4.4. Multitask learning

4.4.1. Detection and segmentation

With a shared emphasis on creating effective and precise multitask learning frameworks that can complete both tasks at once, the papers showcase the developments in joint 3D object detection and semantic segmentation for autonomous driving applications [63–66]. Leveraging the synergy between the two tasks and enhancing autonomous systems' total perceptual capabilities is the main driving force behind this. In order to achieve state-of-the-art results on the difficult BDD100K dataset and show real-time performance on embedded devices, Wu et al.[63] present a comprehensive panoptic driving perception network that simultaneously performs lane detection, drivable area segmentation, and traffic object detection. Using a Spatial Embeddings (SEs) strategy to aggregate foreground points, Zhou and Tuzel provide a novel detection framework that outperforms existing methods on the KITTI dataset and jointly predicts 3D bounding boxes and instance segmentation [64]. In order to achieve state-of-the-art results on the nuScenes dataset, Xie et al.[65] resent a unified framework that uses multi-camera image inputs to perform 3D object detection and map segmentation in the Bird's-Eye View (BEV) space. They propose several important design choices that greatly enhance the performance of camera-based 3D perception tasks. In order to achieve state-of-the-art results on a large benchmark dataset, Meyer et al.[66] present an extension to an effective LiDAR-based 3D object detector by introducing a sensor fusion approach that incorporates image data with LiDAR input. This improves detection performance, particularly at long ranges, and expands the model's capabilities to perform 3D semantic segmentation in addition to 3D object detection [66]. All things considered, these studies indicate how multitask learning works well for 3D object detection and semantic segmentation in autonomous driving, presenting a range of cutting-edge methods that help develop complete and real-time perception systems for autonomous cars.

4.4.2. Detection and tracking

Using multitask learning frameworks, the articles demonstrate advances in jointly learning 3D object detection and tracking for autonomous driving [67–71]. These studies create effective models that can complete both tasks at once, making use of the synergies to enhance perceptive abilities overall. The creation of spatiotemporal representations to facilitate collaboration between detectors and trackers [69], end-to-end joint monocular 3D detection and tracking with spatial-temporal feature aggregation [67], camera-LiDAR fusion for robust multi-object detection and tracking [68], using 3D box depth-ordering and motion modeling for monocular 3D vehicle tracking [70], and incorporating historical object tracking data to improve 3D object detection in cooperative perception settings [71] are some of the key innovations. These combined tracking and detection techniques show cutting-edge results on benchmark datasets while preserving real-time inference speeds, helping to create complete and effective perception systems for self-driving cars.

5. 3D Perception and Localization

Localization is an essential task for AVs, enabling them to accurately determine their position and orientation within an environment. This capability is vital for safe navigation, lane-keeping, and overall vehicle control. There are many different methods for localization, each well-known for its own advantages and disadvantages. These are GPS-IMU fusion, SLAM, and a priori map-based localization.

Perception is the process of extracting certain information from the data generated by various sensors, allowing for a more thorough understanding of the robot's perceived surroundings. This is important for object identification, obstacle recognition, and environmental interaction. A summary of the 3D perception and localization techniques is provided in Table 5. Table 6 provides a summary of the datasets used for training and evaluating AV perception tasks.

5.1. Localization strategies

Odometry technique uses data from motion sensors, such as wheel encoders and inertial measurement units (IMUs), to estimate changes in position over time. While odometry provides a continuous estimate, it suffers from cumulative errors like drift, which can degrade accuracy over extended distances [1]. By using cameras to capture images, visual localization identifies and tracks environmental features or landmarks. This method can be highly effective

Table 5
Detailed summary: 3D perception and localization

Component	Methods/techniques	Advantages	Limitations
Localization Strategies	Odometry, visual localization, SLAM	Effective real-time estimation, feature robustness	Odometry drift errors; sensitivity to lighting and occlusion (visual localization); computational complexity (SLAM)
Perceptual Sensors	LiDAR, cameras, ultrasonic sensors, sensor fusion	High resolution, comprehensive environmental understanding, robustness to single-sensor failures	LiDAR cost, weather sensitivity, camera lighting limitations, low resolution of ultrasonic sensors
Deep Learning	PointNet, PointNet++, VoxelNet, Frustum PointNet, 3D spatial feature extraction	Improved accuracy through direct spatial and structural data handling	Computational complexity, extensive dataset requirement, real-time application challenges
PointNet & PointNet++	Direct point cloud processing, shared MLPs, symmetric max pooling, multi-scale hierarchical learning (PointNet++)	Efficient, low computational cost, robustness to transformations, multi-scale detail capture (PointNet++)	Difficulty capturing local detail (PointNet); increased computational demands (PointNet++)
VoxelNet	Voxelization, end-to-end 3D convolution, Structured 3D data handling, simplified PointNet feature extraction	Real-time capability, improved spatial coherence, structured data handling suitable for AV applications	High computational needs for dense 3D data, performance limitations in cluttered environments
Frustum PointNets	2D detection integration, ROI generation, 3D bounding box estimation, instance segmentation, spatial attention	High accuracy in complex and occluded scenarios, ROI-focused computational efficiency	Dependence on accurate initial 2D detection; effectiveness reduced with inaccurate 2D detection

Table 6
Summary of open source datasets for autonomous vehicle perception tasks

Dataset	Modality	Tasks	Size	Notes
KITTI [46]	Camera, LiDAR, GPS	3D detection, tracking, segmentation	7481 images	Real-world street scenes
nuScenes [72]	Camera, LiDAR, Radar, IMU	Detection, tracking, segmentation, fusion	40157 images	Full sensor suite, weather variations
Waymo Open [73]	Camera, LiDAR	3D detection, tracking, lane detection	1,950 segments of 20 s each, collected at 10 Hz (390,000 frames)	High-res LiDAR, multiple vehicles
Lyft [74]	Camera, LiDAR	Motion prediction and forecasting	170,000 scenes, where each scene is 25 seconds long	Urban driving scenes
Cityscapes [75]	Camera	Semantic segmentation	5000 annotated images with fine annotations, 20000 annotated images with coarse annotations	Urban German roads with pixel labels
BDD100K [76]	Camera	Detection, segmentation, lane detection	100,000 videos. Each video is about 40 seconds long, 720 p, and 30 fps	Multitask, diverse weather and time
CRUW Radar [77]	Radar, Camera	Object detection, tracking	400K frames	Radar fusion benchmark, low-light data
CARLA [78]	Camera and pseudo-sensors	Detection, segmentation	–	Used for training and safety testing, simulated using Unreal Engine 4

in visually distinct environments but is susceptible to variations in lighting and occlusion [79].

Simultaneous localization and mapping (SLAM) algorithms enable robots to generate a map of an unfamiliar environment while simultaneously determining their position within it. This process utilizes sensor fusion and probabilistic models such as Kalman filters or particle filters to address uncertainties in the acquired data, focusing on precision in localization [1].

5.2. Perceptual sensors

LiDAR uses laser beams to measure distances, providing high-resolution 3D maps of the environment. It is highly effective in both indoor and outdoor settings, offering precise distance measurements that aid in navigation and obstacle avoidance [79]. Cameras capture visual information that can be processed for object recognition, scene understanding, and depth estimation. Visual data is crucial for tasks like identifying landmarks and detecting moving objects [1]. Ultrasonic sensors emit sound waves to measure the distance to nearby objects, making them useful for basic obstacle detection, particularly in low-cost applications [79]. To enhance the precision of perception, sensor fusion combines data from multiple sources, such as LiDAR, cameras, and IMUs. This approach mitigates the drawbacks of individual sensors, resulting in a more comprehensive and reliable understanding of the environment [79].

5.3. Deep learning techniques for 3D perception

Advancements in the field of perception were achieved mainly through deep learning techniques, allowing the handling of raw point clouds directly [80]. Unlike previous traditional feature-based or 2D projected procedures, depth learning models utilize 3D data's spatial and structural characteristics for significant hope for improvements in object detection, segmentation, and localization tasks [81].

Deep learning methods seek to characterize 3D worlds by directly employing spatial hierarchies and establishing local-global feature relationships from the point clouds. These developments underscore the porosity of PointNet and PointNet++ networks, which directly conduct operations on an unordered set of points and largely reflect permutation invariance, ensuring integrity in maintaining spatial information [82].

5.3.1. PointNet and PointNet++

PointNet was an innovative model that pioneered a new way of processing point clouds by working directly on the raw point data without needing voxelization or image projections [80]. The model employs certain shared MLPs and the symmetric max pooling function for the extraction of global features, which allows it to deal with unordered data with robustness against transformations like rotation and translation [81].

For irregular point data, PointNet provides a neural network architecture that is end-to-end learnable and sufficiently robust to noise and outlier point perturbations. It sets a new standard for 3D recognition and does away with expensive and handcrafted feature engineering. PointNet is efficient because its architecture allows an efficient and effective point cloud of object instance-specific parameters in Airborne Laser Scanning (ALS) data [83]. The PointNet model reduces memory and computation costs by more than 80% and 88%, respectively, compared to volumetric models. As a result, it is frequently chosen for mobile and portable devices [83].

This multi-scale learning approach allows PointNet++ to capture intricate details of complex 3D shapes, making it more effective for tasks such as segmentation, classification, and object detection in scenes with varying point densities [81].

5.3.2. VoxelNet

Recently, there has been a growing emphasis on 3D object detection that utilizes only point clouds. Generally, these methods can be classified into three groups: BEV-based, point-based, and voxel-based. For the first time, an end-to-end learning approach has been applied in the voxel-based domain using VoxelNet, which integrates a simplified form of PointNet to automatically extract features from each voxel. It divides the points into voxels, uses PointNet to obtain features from the separate voxels, and then merges the features from these voxels in the central convolutional layers for detection [37].

VoxelNet introduces a new approach by combining feature learning and the prediction of 3D bounding boxes within a single cohesive end-to-end framework [80]. In contrast to PointNet-based models that function on discrete points, VoxelNet transforms the 3D space into uniformly distributed 3D grids (voxels), enabling the model to handle structured data that aligns better with 3D convolutional operations [81].

VoxelNet's capability to preserve spatial coherence while handling dense 3D data enables it to excel in real-time 3D object detection applications, especially in autonomous driving situations where both speed and precision are essential [82].

5.3.3. Frustum PointNets

Frustum PointNet merges the strong points of 2D object detection and 3D point cloud processing to perform 3D object detection from RGB-D data [80]. The 2D detections are back-projected into 3D space, forming frustums (3D cones extending from the camera through the 2D bounding boxes), while Frustum PointNet narrows down points inside these volumes for further processing [81]. This technique squeezes down the search space and channels the network's focus and strength on the areas of interest and improves the detection accuracy [82].

The architecture is organized into three stages: 2D object detection, Frustum Proposal Generation, and 3D object localization via PointNet [80]. In detail, a 2D detection network was used to build an image region of interest (ROI), which was then projected onto the point cloud to concoct a proposal for frustums. At this stage, the 2D ROI served as prior knowledge to evaluate the frustum point cloud proposals. The frustum proposals are then processed using an amodal 3D bounding box estimation with PointNet, followed by 3D instance segmentation [84].

It is beneficial to combine the features of object detection in 2D images with LiDAR-based SLAM in order to detect small and distant things. It narrows down the search space for particular 3D targets and lessens noise and error interference while detecting objects in 2D photos. It provides a Simultaneous Dynamic Triangulation Mapping (SDTM) framework based on the aforementioned concept [85].

A 3D object detection method that employs spatial attention, improving the detection performance for occluded objects. The SAF (spatial attention frustum) module effectively encapsulates a significant representation of the features associated with occluded objects in a limited feature space [86]. The Local Feature Aggregation (LFA) module enhances the understanding of the local arrangement of the concealed object. It allows for a better deduction of the entire structure of the hidden object using a limited amount of locally visible point clouds, even when only the surface area of the object is visible [86].

6. Sensor Fusion

Sensor fusion enhances the accuracy and robustness of AV perception by integrating data from heterogeneous sensors such as cameras, LiDARs, and radars. While traditional reviews focus on early, mid, and late fusion strategies [87, 88], this paper adopts a modality-based perspective, analyzing fusion approaches according to sensor combinations like Camera-LiDAR, Camera-Radar,

Table 7
Summary of sensor fusion strategies in recent autonomous vehicle perception models

Fusion strategy	Representative methods	Advantages	Limitations
Camera + LiDAR	FuseMODNet, BEVFusion	High accuracy; strong spatial semantic fusion; effective in all lighting	Expensive sensors; LiDAR sensitivity to adverse weather
Camera + Radar	CRF-Net, RadarFormer	Cost-effective; operates in low visibility; reliable velocity estimation	Low spatial resolution; calibration can be difficult
LiDAR + Radar	CenterRadarNet, RODNet	Complementary range and resolution; robust in harsh conditions	Sensor-disparity makes fusion challenging
Multimodal Fusion	BEVGuide, PMF, CARLA-based frameworks	Resilient and generalizable; handles edge cases; proven real-world applicability	High compute cost; latency and synchronization challenges

and LiDAR-Radar. This practical categorization reflects real-world deployment needs and highlights how different modalities complement each other. Table 7 summarizes representative methods, their advantages, and limitations across these fusion strategies, while Table 8 provides a broader overview of recent research trends, categorized by sensor combinations, datasets, and perception tasks.

6.1. Camera

Recent progress in 3D object detection for AVs has explored various camera-based methods, including stereo vision, pseudo-LiDAR, and monocular approaches. Pseudo-LiDAR techniques convert monocular images into 3D point clouds using depth estimation, showing promise in benchmarks like KITTI-3D and NuScenes [23, 89]. However, challenges like overfitting, high computational cost, and performance gaps with real LiDAR remain. Additionally, validation bias due to overlapping datasets has been addressed through innovations like 3D confidence prediction modules [90].

End-to-end monocular models such as DD3D [89] bypass depth estimation by learning 3D object detection directly from images. While efficient and effective—especially when pre-trained on large datasets—they struggle with occlusion and complex scenes due to limited depth cues. Stereo vision methods like IDA-3D [91] and Stereo RCNN [92] estimate depth from image pairs without dedicated depth networks but still face limitations in textureless or occluded regions, impacting real-world reliability.

BEV representations have also gained traction for their improved spatial understanding. Recent breakthroughs, such as BEVFormer [93], have shown that combining spatiotemporal data from several cameras can considerably improve the performance of vision-based models. BEVFormer, for example, achieves more accurate velocity estimation and recall on low-visibility objects, demonstrating the expanding utility of camera-based fusion techniques. Also, other transformer-based models like HeightFormer [94] and WidthFormer [95] leverage multi-view cameras to generate BEV maps with low latency, though challenges persist in feature quality and accurate map generation. Despite these issues, ongoing

Table 8
Comparison of articles by year, type, modality, and task

		Paper type		Sensor modality				
Paper	Year	Journal	Conference	Camera	LiDAR	Radar	Dataset	Perception task
[89]	2021		✓	✓			KITTI-3D, nuScenes	3D object detection
[23]	2020		✓	✓			KITTI-3D	3D object detection
[90]	2021		✓	✓			KITTI-3D	3D object detection
[91]	2020		✓	✓			KITTI-3D	3D object detection
[92]	2019		✓	✓			KITTI-3D	Object localization
[93]	2022		✓	✓			nuScenes, Waymo Open	3D object detection, segmentation
[94]	2024	✓		✓			NuScenes	3D object detection
[95]	2024	✓		✓			nuScenes	3D object detection
[96]	2020	✓			✓		KITTI-3D	3D object detection
[97]	2021		✓		✓		nuScenes, Waymo	3D object detection
[52]	2023		✓		✓		KITTI-3D, Waymo, Lyft	3D object detection
[98]	2021		✓		✓		KITTI-3D, Waymo	3D object detection
[99]	2022		✓		✓		Waymo	3D object detection
[100]	2019		✓		✓		ATG4D, KITTI-3D	3D object detection

(*Continued*)

Table 8
(*Continued*)

[101]	2019		✓		✓		KITTI-3D	3D object detection
[40]	2020		✓		✓		KITTI-3D, Waymo	3D object detection
[102]	2021		✓		✓		Waymo	3D object detection
[103]	2022		✓		✓		KITTI-3D, Waymo	3D object detection
[104]	2024		✓			✓	K-Radar	Object tracking
[105]	2024		✓			✓	nuScenes, Aptiv	3D object detection
[106]	2019		✓			✓	nuScenes	3D object detection
[107]	2023		✓			✓	CRUW radar	3D object detection
[108]	2021		✓			✓	CRUW	Object localization
[109]	2023	✓		✓	✓		Custom dataset	3D object detection
[110]	2019		✓	✓	✓		Dark-KITTI, KITTI-3D	3D object detection
[111]	2020	✓		✓	✓		KITTI-3D	3D object detection
[112]	2021		✓	✓	✓		nuScenes, PMF	3D object detection
[113]	2020	✓		✓	✓		Custom dataset	3D object detection
[114]	2023	✓		✓	✓		KITTI-3D	3D object detection
[115]	2019	✓		✓		✓	nuScenes, TUM (custom)	3D object detection
[116]	2020		✓	✓		✓	nuScenes	3D object detection
[106]	2019		✓	✓		✓	nuScenes	3D object detection
[117]	2022	✓		✓		✓	Custom dataset	3D object tracking
[118]	2020	✓		✓		✓	Custom dataset	Object localization
[119]	2023	✓		✓		✓	CARLA	3D object detection
[120]	2022	✓		✓	✓	✓	Custom dataset	Semantic segmentation
[121]	2023	✓		✓	✓	✓	nuScenes, ESCAPE	3D object detection
[122]	2022	✓		✓	✓	✓	CARLA	Overall perception

research continues to enhance monocular and stereo-based 3D detection, with future efforts likely to focus on scalability, occlusion handling, and real-time efficiency.

6.2. LiDAR

Recent advances in LiDAR-based 3D object detection for AVs have been driven by deep learning techniques aimed at improving accuracy, efficiency, and data utilization. SSADNet [96] enhances computation through dual-task learning for the drivable region and vehicle detection, though its accuracy remains limited. Self-supervised methods like GCC-3D [97] reduce reliance on labeled data using geometry-aware contrastive learning but may struggle in complex environments. UpCycling [52] applies semi-supervised learning to protect privacy through intermediate feature training, yet its adaptability in dynamic settings is constrained.

Transformer-based models offer improved long-range context modeling. VoTr [98] enhances point cloud detection via sparse voxel transformers but is resource-intensive. SWFormer [99] increases efficiency with windowed attention but at a computational cost. LaserNet [100] enables real-time detection by processing LiDAR's native range view, though it faces challenges with scale variation. PointPillars [101] accelerates inference through pillar encoding but is less robust in complex scenes compared to models like PV-RCNN [40], which fuses voxel- and point-based learning for high accuracy at increased complexity.

Additional methods like RangeDet [102] offer efficient range-view processing with some flexibility limitations, while VoxSeT [103] merges voxel efficiency with Transformer self-attention for powerful, though complex, detection. Despite significant progress, balancing real-time performance, accuracy, and scalability remains a core challenge. Future work will likely refine these models to better meet the demands of dynamic driving environments.

6.3. Radar

Radar-based object detection is proving vital for AVs, especially in poor visibility conditions like rain, fog, or nighttime, where vision sensors underperform. Advances in radar perception, such as CenterRadarNet [104], which fuses 4D radar data for 3D detection and tracking, have achieved state-of-the-art results on benchmarks like K-Radar. Its integrated online tracker improves object reidentification in dynamic scenes. Similarly, CenterPoint Transformer [105] enhances BEV detection using transformers, achieving a mean AP of 23.6% on NuScenes.

Other innovations focus on real-time efficiency. RRPN [106] accelerates object proposal generation directly from radar signals, outperforming older methods like Selective Search. RadarFormer [107] reduces model complexity by over 90% using a transformer design and efficient data fusion, making it ideal for real-time use. RODNet [108] leverages radar frequency data and 3D autoencoders for high-performance detection without relying on LiDAR or cameras, and when combined with camera-radar fusion, it reduces manual labeling needs.

These radar-driven approaches like CenterRadarNet, RadarFormer, and RODNet highlight radar's growing role in robust,

efficient AV perception. Their ability to operate reliably under challenging conditions, combined with modern deep learning, positions radar as a key pillar in the future of autonomous driving.

6.4. Camera and LiDAR fusion

The integration of camera and LiDAR sensors has become critical for AV perception due to their complementary strengths. Cameras give high-resolution visual data for item classification, whereas LiDAR provides exact 3D spatial and depth information, allowing for comprehensive scene interpretation. Advances in deep learning have improved sensor fusion approaches, increasing accuracy and dependability in various settings.

Several studies have proved the efficiency of camera-LiDAR fusion in object detection. A sensor fusion system that combines a thermal infrared camera and LiDAR improves detection in adverse conditions like fog and nighttime by aligning data using a 3D calibration target [109]. FuseMODNet used a CNN-based design to integrate sensor data, resulting in considerable performance improvements in low-light conditions on the Dark-KITTI dataset [110]. In another technique, LiDAR-generated region proposals were merged with a CNN for real-time detection, resulting in high accuracy on the KITTI dataset [111].

For semantic segmentation, the perception-aware multi-sensor fusion (PMF) technique fused RGB and LiDAR information using a two-stream network. It retained appearance and depth information and showed robustness in the SemanticKITTI and nuScenes datasets [112]. Low-level fusion approaches like projecting LiDAR data onto camera pictures have improved object distance estimation, especially for obstructed objects [113]. Faster-RCNN feature-level fusion improves detection accuracy and reduces computing overhead [114].

Despite these advances, issues persist in assuring data coherence, real-time performance, and resilience to environmental changes. Future research should overcome these restrictions in order to develop scalable and efficient sensor fusion methods for self-driving vehicles.

6.5. Camera and radar fusion

Radar-camera fusion is becoming a key technology in AV perception, leveraging radar's reliable distance and velocity data alongside the visual richness of cameras. Deep learning models enhance detection by integrating these modalities. CRF-Net [115] addresses radar sparsity and camera limitations for better accuracy, while RANet and BIRANet [116] use radar to guide anchor generation, improving detection in challenging conditions.

Fusion also improves tracking and localization. JPDA-based methods boost tracking in harsh weather [117], and specialized frameworks enhance lateral accuracy in occlusions [118]. Models like CR-YOLOnet [119] apply attention mechanisms to combine radar and camera data, improving detection of small and distant objects in poor conditions. Together, these advances highlight radar-camera fusion as a robust, scalable solution for AV perception.

6.6. Multimodal fusion

Multimodal fusion significantly enhances AV perception by integrating data from sensors like cameras, LiDAR, radar, and ultrasonics, boosting accuracy and resilience in complex scenarios. One advancement is the self-evolving Free Space Detection (FSD) framework, which uses online active learning to adapt in real time, outperforming models like DeepLabV3+ without requiring large datasets [120]. Similarly, a multimodal architecture incorporating image, radar, sound, and seismic data enhances detection in non-line-of-sight (NLOS) scenarios, improving safety in obstructed environments [121]. BEVGuide [123] introduces a BEV-based fusion model that merges inputs from multiple sensors using attention mechanisms, improving motion estimation and segmentation, particularly in the nuScenes dataset.

End-to-end approaches combining RGB and depth via early fusion outperform single-modality methods, as shown in CARLA simulator experiments using CIL architecture [122]. While early fusion captures detailed features, it's resource-intensive; mid-level fusion balances performance and efficiency, and late fusion supports modular design. Overall, models like BEVGuide and adaptive FSD frameworks highlight the flexibility of multimodal fusion, though further research is needed for scalable, real-time deployment in dynamic driving conditions.

7. Perception of the Environment

7.1. Deep learning models

Understanding the environment is crucial for autonomous driving systems (ADS), enabling real-time lane detection, trajectory prediction, and safe navigation. Advanced sensing and processing ensure consistent environmental perception, essential for effective decision-making on the road.

In smart cities, predicting how people perceive urban environments is a growing research area supporting human-centered design. Traditional methods relying on expert opinions are time-consuming, prompting the use of machine learning to estimate perceptual scores from single-frame street view images [124]. However, aspects like dual visual states (pleasant/unpleasant), perceptual mapping using CNNs, and spatial factors affecting perception are often underexplored [125]. CNNs are used in both urban design and Unmanned Aerial Vehicle (UAV) applications for object detection and feature extraction, underlining the importance of environmental perception frameworks for reliable, adaptive operations and human-centered outcomes [126].

7.2. Road and lane detection

Machine vision that adapts quickly utilizes machine vision techniques alongside a RALPH (Rapidly Adapting Lateral Position Handler) vision system developed by Carnegie Mellon University and Assist-Ware Technologies Inc. RALPH processes the obtained image through three main phases: capturing the image, analyzing road curvature, and finally determining the vehicle's lateral position in relation to the center of the lane.

Currently, the methods include geometric modeling and traditional techniques, while artificial intelligence (AI) involves deep learning and machine learning. Examples of deep neural networks and architectures include CNN, Fully Convolutional Network (FCN), and RNN. In the last four years, there has been an increasing amount of research focused on the use of deep learning [127].

To predict vehicle paths, the lane detection pipeline typically includes perspective transformations to convert camera images to top-down views. In this format, DBSCAN clustering identifies lane boundaries, followed by fitting quadratic polynomials to the points for smooth path estimation. This process, as demonstrated by CNN-PP, allows for lateral offset estimation and accurate curvature modeling for real-time driving decisions [128]. The CNN-PP operates at 10 fps, maintaining efficiency in dynamic driving environments.

Innovations in sensor fusion, such as radar-camera fusion, complement lane detection, especially in adverse weather conditions. CR-YOLOnet combines radar and camera data, improving the detection of small or distant objects under low-visibility scenarios, thereby strengthening overall lane detection and vehicle localization [129]. Evaluations on diverse datasets like TORCS, TuSimple, CULane, and LLAMAS demonstrate the effectiveness of these models in handling various lane configurations and traffic environments. For example, DSUNet outperforms UNet in terms of recall, while DeepLabV3 shows better robustness against occlusions [127, 128].

7.3. Traffic sign and signal recognition

Traffic Sign and Signal Recognition (TSSR) is essential for AVs to interpret road signs, signals, and markers, ensuring safe and lawful navigation. This involves three key layers: sensing, perception, and decision-making, which work together to help the vehicle understand and respond to its environment. Object detection techniques like YOLOv3 and Viola-Jones were used to identify vehicles, pedestrians, traffic lights, and signs. Among the nine detectors tested, YOLOv3 outperformed Viola-Jones in terms of precision, recall, and processing speed [130]. For comprehensive environmental understanding, including the detection of static and moving obstacles, vehicles rely on various onboard sensors, while localization and mapping determine vehicle position relative to global coordinates [131].

Advanced traffic signal control systems now integrate IoT and video processing, using tools like OpenCV and Raspberry Pi to manage lights based on vehicle count and density, though this is best suited for low-traffic intersections [132]. In the perception layer, CNNs enable accurate traffic sign and signal classification under diverse conditions. The decision-making layer converts this data into timely actions, such as stopping or rerouting, ensuring safety and real-time responsiveness [88, 133].

8. Challenges and Limitations

8.1. Data scarcity and quality issues

Data scarcity and quality difficulties are important challenges in AV perception, limiting the development of dependable and resilient systems. Rare events, such as extreme weather or nighttime driving, are difficult to capture in significant quantities, and annotating complicated datasets, such as 3D point clouds or semantic maps, is time-consuming and costly. Furthermore, datasets frequently contain sensor noise, calibration mistakes, and labeling inconsistencies, compromising the reliability of supervised learning models. Imbalanced datasets, in which typical scenarios are over-represented but rare or difficult cases are underrepresented, impose additional biases in model training. Issues with synchronization and alignment in multimodal data exacerbate these obstacles, lowering sensor fusion performance. Increased data diversity and quality via data augmentation, synthetic data generation in simulation environments, self-supervised learning, and active learning can help mitigate these challenges. These are important considerations to make sure that AV perception systems work in different and dynamic real-world scenarios.

8.2. Computational complexity and efficiency

Real-time performance requirements present significant challenges with respect to computational complexity and efficiency for deep learning-based AV perception. These methods have millions of parameters and demand a lot of processing to interpret high-dimensional data such as LiDAR point clouds and video streams for tasks like 3D object detection or semantic segmentation or multimodal fusion. Low latency is essential for real-time decision-making, as delays could endanger safety in ever-changing scenarios. Additionally, processing high frame rates and varied driving scenarios within the AV's constrained power budget adds another layer of complexity. These models must be optimized further for deployment on edge devices to bridge the gap between computation requirements and performance. The challenges of AV perception motivate innovations in model pruning, quantization, efficient designs, and hardware accelerators to ensure AV perception systems remain fast, scalable, and energy-efficient enough to operate reliably in the real world.

8.3. Real-time processing and latency

Deep learning-based perception in AVs faces major challenges with real-time processing and latency. While many models prioritize accuracy, they often overlook the time needed for inference—critical for making timely decisions in dynamic environments. Tasks like 3D object detection and semantic segmentation are computationally intensive, leading to delays that can compromise safety.

Current methods typically focus on metrics like accuracy and recall, but lack standardized benchmarks for inference time, making it difficult to assess suitability for real-world applications. This issue worsens with multimodal sensor fusion (e.g., combining data from cameras, LiDAR, and radar), which adds processing complexity and potential latency.

AV systems must operate within tight temporal limits—processing data in milliseconds—to respond effectively to road conditions. Excessive latency can result in missed detections or delayed responses. Addressing this trade-off between speed and performance requires optimized model architectures, hardware acceleration, and techniques like pruning and quantization to develop fast, accurate, and responsive perception systems.

8.4. Ethical and safety considerations

The ethical and safety issues are always very important when talking about AV perception systems that use deep learning. The problem is that deep learning models often present enormous challenges that make it difficult to reach an optimal balance between AV decision-making and the acceptable level of risk toward human safety.

8.4.1. Lack of explainability and transparency

Decision-making in the case of deep learning models such as deep neural networks is hard to fathom, which is one of the reasons these models are often referred to as "black boxes." This "black box" problem is indeed a serious issue, especially when one considers the vehicle scenario where an explanation is needed as to why a decision was made, for example, in the case of a vehicle collision with a pedestrian or vice versa. It is one thing to state that the model followed a particular sequence of reasoning resulting in a decision and another thing to be able to understand that reasoning. And so the question of how trustworthy the model's choice is does come to the forefront, especially if the ethical constraint in this specific example implies saving a human's life in the case of an emergency.

8.4.2. Bias and fairness

The performance of deep learning models is highly dependent on the data they are taught. If the training datasets were to be biased or unrepresentative, it is highly probable that the AV perception system would have similar biases. For instance, if the model were trained on a narrow set of data, which has people, but not enough of different ethnicities and genders, the model would not be able to recognize particular individuals, which means there could be consequences in the real world. While it is important to ensure fairness, the more pressing issue is the ethical issues of AV development, especially concerning safety-critical aspects such as object identification or traffic sign recognition.

8.4.3. Accountability and liability

The liability of an accident is heavily debated, as no one knows if the software developers, the manufacturer, or even the vehicle itself is to be blamed. All parties share partial responsibility, and legal and ethical dilemmas often ensue, while deep learning models do add to the ambiguity. For instance, if a model's prediction leads to damage, the legally critical question emerges: who is to take the blame? This is particularly burdensome when the model's reasoning is poorly documented.

8.4.4. Safety and robustness

Deep learning models for AV perception must generalize well to diverse real-world conditions such as lighting changes, weather variations, and sensor noise. Failure under scenarios like fog or heavy rain can pose serious safety risks. Ensuring model reliability through extensive testing and validation is critical, as even small perception errors can lead to accidents. Addressing these challenges also involves tackling ethical and safety concerns, such as reducing training data bias, building transparent decision-making frameworks, and defining clear accountability in case of failures. Robust testing and clear ethical guidelines are essential for safe deployment. Balancing the need for high accuracy with safety, fairness, and accountability remains a major hurdle in advancing self-driving technology.

9. Open Research Issues and Future Directions

Although there has been substantial advancement in the field of deep learning-based perception for AVs, there are still numerous unresolved research challenges that necessitate further investigation, particularly in the context of the emerging developments of 2024.

9.1. End-to-end autonomous driving frameworks

Perception, planning, and control are divided into distinct stages in conventional modular pipelines. Recent advancements imply a transition toward end-to-end learning frameworks [134], wherein a singular neural architecture directly maps raw sensor data to driving actions or intermediate representations. Significant developments in 2024 encompass BEVFormer [135], TransFuser [136], and end-to-end BEV-based transformers [137], which exhibit enhanced performance in urban driving contexts by integrating spatial-temporal information from cameras, LiDAR, and high-definition maps. Improvements to the interpretability, generalization, and safety guarantees of these densely coupled architectures should be the primary focus of future research.

9.2. World models

The integration of world models, including DreamerV3 and PlaNet, into autonomous driving frameworks is another promising direction [138, 139]. These models acquire a latent dynamics representation of the environment, allowing agents to internally simulate future paths. Their application to AVs, although mainly investigated in robotics and gaming contexts, can markedly improve sampling efficiency, long-term prediction, and uncertainty-informed planning in intricate metropolitan settings. In order to align with real-world sensor dynamics, future research must investigate the training of such models using multimodal driving datasets.

9.3. Foundation models and multimodal pretraining

The emergence of foundation models [140], trained on extensive and diverse datasets (e.g., LiDAR, RGB pictures, semantic maps), has facilitated the development of generalizable perception systems. Pretraining on large autonomous driving datasets (such as nuScenes++ and Waymo Open Dataset 2024) has been investigated in 2024 to facilitate multitask learning, domain adaptation, and zero-shot transfer. Nonetheless, obstacles persist in scaling these models, aligning them with safety-critical goals, and minimizing computational demands for real-time implementation.

9.4. Emerging cross-domain directions

While traditional perception systems have focused on 3D object detection, sensor fusion, and real-time inference, emerging research has begun to address adjacent challenges related to data privacy, edge processing, and intelligent wireless sensing. Notably, works such as generative AI for privacy-preserving crowdsensing [141] and attention mechanisms in wireless networks [142] introduce generative models and attention-based methods to improve secure data handling and adaptive model behavior in distributed environments. These studies reflect a growing trend of integrating privacy-aware, intelligent edge computing techniques into perception systems—a direction that could significantly impact future sensor-driven AI frameworks.

10. Conclusion

The rapid advancement of deep learning techniques in AV perception has considerably improved self-driving cars' capabilities, allowing them to navigate complex and dynamic situations more accurately and safely. This review has examined recent advances from 2019 to 2024, emphasizing critical domains such as 3D object detection, sensor fusion, object localization, and environmental perception. Methods including CNNs, RNNs, GANs, and transformers, in conjunction with sensor fusion, have improved system precision and dependability. Nevertheless, ethical concerns, data scarcity, real-time processing, and computational complexity continue to be a challenge. Future research should focus on these difficulties in order to improve performance, ensure safety, and permit the widespread deployment of AVs in real-world circumstances.

Acknowledgment

The authors express their heartfelt gratitude to Mata Amritanandamayi Devi for her divine blessings and spiritual encouragement, which inspired and gave them the strength to carry out this research work. The views and opinions expressed in this paper are solely those of the authors and do not represent the official position or endorsement of Amazon or its affiliates.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Namitha Kalakunnath: Conceptualization, Validation, Writing - review & editing, Visualization, Supervision, Project administration. **Aneesh Varghese:** Supervision, Project administration. **Abekaesh Prakash Anuradha:** Methodology, Investigation, Resources, Data curation, Writing - original draft, Writing - review & editing. **Dhanush Kumar Girish:** Investigation, Data curation, Writing - original draft. **Renjith Sasidharan:** Writing - review & editing, Visualization.

References

[1] Yurtsever, E., Lambert, J., Carballo, A., & Takeda, K. (2020). A survey of autonomous driving: *Common practices and emerging technologies*. *IEEE Access*, *8*, 58443–58469. https://doi.org/10.1109/ACCESS.2020.2983149

[2] Megalingam, R. K., Rudravaram, G., Kumar, D. V., Deepika, A. S., & Smaran, K. S. (2022). Dynamic lane segmentation for autonomous vehicles using neural networks. In *2022 5th International Conference on Advances in Science and Technology*, 444–449. https://doi.org/10.1109/ICAST55766.2022.10039575

[3] Amara, D. K., Karthika, R., & Soman, K. (2020). Deeptracknet: Camera based end to end deep learning framework for real time detection, localization and tracking for autonomous vehicles. In *Intelligent Computing, Information and Control Systems*, 299–307. https://doi.org/10.1007/978-3-030-30465-2_34

[4] Kuutti, S., Bowden, R., Jin, Y., Barber, P., & Fallah, S. (2021). A survey of deep learning applications to autonomous vehicle control. *IEEE Transactions on Intelligent Transportation Systems*, *22*(2), 712–733. https://doi.org/10.1109/TITS.2019.2962338

[5] Tang, Y., Zhao, C., Wang, J., Zhang, C., Sun, Q., & Zheng, W. X. (2023). Perception and navigation in autonomous systems in the era of learning: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *34*(12), 9604–9624. https://doi.org/10.1109/TNNLS.2022.3167688

[6] Neethu, M., Remyakrishnan, P., Mukesh, S., Barath Kumar, J., & Royal, A. M. (2024). Perception system of autonomous vehicles: Sophisticated attacks, defense strategies, and future directions. In *2024 15th International Conference on Computing Communication and Networking Technologies*, 1–7. https://doi.org/10.1109/ICCCNT61001.2024.10725911

[7] Nahavandi, S. (2017). Trusted autonomy between humans and robots: Toward human-on-the-loop in robotics and autonomous systems. *IEEE Systems, Man, and Cybernetics Magazine*, *3*(1), 10–17. https://doi.org/10.1109/MSMC.2016.2623867

[8] Wang, K., Zhao, G., & Lu, J. (2024). A deep analysis of visual SLAM methods for highly automated and autonomous vehicles in complex urban environment. *IEEE Transactions on Intelligent Transportation Systems*, 25(9), 10524–10541. https://doi.org/10.1109/TITS.2024.3379993

[9] Rokonuzzaman, M., Mohajer, N., Nahavandi, S., & Mohamed, S. (2021). Review and performance evaluation of path tracking controllers of autonomous vehicles. *IET Intelligent Transport Systems*, *15*(5), 646–670. https://doi.org/10.1049/itr2.12051

[10] Alkendi, Y., Seneviratne, L., & Zweiri, Y. (2021). State of the art in vision-based localization techniques for autonomous navigation systems. *IEEE Access*, *9*, 76847–76874. https://doi.org/10.1109/ACCESS.2021.3082778

[11] Padmanabhan, A. A., Khan, J., Prakash, A., & Nair, A. R. (2023). Development and validation of lane keeping assist algorithm for advanced driver assistance system applications. In *2023 9th International Conference on Smart Computing and Communications*, 471–476. https://doi.org/10.1109/ICSCC59169.2023.10334980

[12] Nambisan, A., Chandran, A., Preetha, P., & Nair, M. G. (2021). Advanced body controlled safety system for connected and autonomous vehicle. In *2021 Second International Conference on Electronics and Sustainable Communication Systems*, 507–514. https://doi.org/10.1109/ICESC51422.2021.9532984

[13] Xiao, Y., Zhang, X., Xu, X., Liu, X., & Liu, J. (2023). Deep neural networks with koopman operators for modeling and control of autonomous vehicles. *IEEE Transactions on Intelligent Vehicles*, *8*(1), 135–146. https://doi.org/10.1109/TIV.2022.3180337

[14] Luo, T., Chen, Y., Luan, T., Cai, B., Chen, L., & Wang, H. (2024). IDS-MODEL: An efficient multitask model of road scene instance and drivable area segmentation for autonomous driving. *IEEE Transactions on Transportation Electrification, 10*(1), 1454–1464. https://doi.org/10.1109/TTE.2023.3293495

[15] Yang, K., Hu, X., & Stiefelhagen, R. (2021). Is context-aware CNN ready for the surroundings? Panoramic semantic segmentation in the wild. *IEEE Transactions on Image Processing*, *30,* 1866–1881. https://doi.org/10.1109/TIP.2020.3048682

[16] Liu, Q., & Zhou, S. (2024). Lightfusion: Lightweight CNN architecture for enabling efficient sensor fusion in free road segmentation of autonomous driving. *IEEE Transactions on Circuits and Systems II: Express Briefs*, *71*(9), 4296–4300. https://doi.org/10.1109/TCSII.2024.3384419

[17] Li, X., Li, J., Hu, X., & Yang, J. (2020). Line-CNN: End-to-end traffic line detection with line proposal unit. *IEEE Transactions on Intelligent Transportation Systems*, *21*(1), 248–258. https://doi.org/10.1109/TITS.2019.2890870

[18] Cao, Y., Li, C., Peng, Y., & Ru, H. (2023). MCS-YOLO: A multiscale object detection method for autonomous driving road environment recognition. *IEEE Access*, *11*, 22342–22354. https://doi.org/10.1109/ACCESS.2023.3252021

[19] Song, Y., Xie, Z., Wang, X., & Zou, Y. (2022). MS-YOLO: Object detection based on YOLOv5 optimized fusion millimeter-wave radar and machine vision. *IEEE Sensors Journal*, *22*(15), 15435–15447. https://doi.org/10.1109/JSEN.2022.3167251

[20] Wibowo, A., Trilaksono, B. R., Hidayat, E. M. I., & Munir, R. (2023). Object detection in dense and mixed traffic for autonomous vehicles with modified YOLO. *IEEE Access*, *11*, 134866–134877. https://doi.org/10.1109/ACCESS.2023.3335826

[21] Liu, J., Wang, H., Peng, L., Cao, Z., Yang, D., & Li, J. (2022). PNNUAD: Perception neural networks uncertainty aware decision-making for autonomous vehicle. *IEEE Transactions on Intelligent Transportation Systems*, *23*(12), 24355–24368. https://doi.org/10.1109/TITS.2022.3197602

[22] Petrovai, A., & Nedevschi, S. (2022). Semantic cameras for 360-degree environment perception in automated urban driving. *IEEE Transactions on Intelligent Transportation Systems*, *23*(10), 17271–17283. https://doi.org/10.1109/TITS.2022.3156794

[23] Ma, X., Liu, S., Xia, Z., Zhang, H., Zeng, X., & Ouyang, W. (2020). Rethinking pseudo-LiDAR representation. In *Computer Vision–ECCV 2020: 16th European Conference,* 311–327. https://doi.org/10.1007/978-3-030-58601-0_19

[24] Li, B., Ouyang, W., Sheng, L., Zeng, X., & Wang, X. (2019). GS3D: An efficient 3D object detection framework for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1019–1028. https://doi.org/10.1109/CVPR.2019.00111

[25] Li, L., Zhao, W., Xu, C., Wang, C., Chen, Q., & Dai, S. (2021). Lane-change intention inference based on RNN for autonomous driving on highways. *IEEE Transactions on Vehicular Technology*, *70*(6), 5499–5510. https://doi.org/10.1109/TVT.2021.3079263

[26] Jeong, Y., Kim, S., & Yi, K. (2020). Surround vehicle motion prediction using LSTM-RNN for motion planning of autonomous vehicles at multi-lane turn intersections. *IEEE Open Journal of Intelligent Transportation Systems*, *1*, 2–14. https://doi.org/10.1109/OJITS.2020.2965969

[27] Xu, W., Souly, N., & Brahma, P. P. (2021). Reliability of GAN generated data to train and validate perception systems for autonomous vehicles. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 171–180. https://doi.org/10.1109/WACVW52041.2021.00023

[28] Qazi, T., Kumar, M. R., Mukherjee, P., & Lall, B. (2024). EgoFormer: Ego-gesture classification in context of autonomous driving. *IEEE Sensors Journal, 24*(11), 18133–18140. https://doi.org/10.1109/JSEN.2024.3390794

[29] Hu, Y., Li, S., Weng, W., Xu, K., & Wang, G. (2023). NSAW: An efficient and accurate transformer for vehicle LiDAR object detection. *IEEE Transactions on Instrumentation and Measurement, 72,* 5028310. https://doi.org/10.1109/TIM.2023.3318680

[30] Zhuang, L., Jiang, T., Qiu, M., Wang, A., & Huang, Z. (2024). Transformer generates conditional convolution kernels for end-to-end lane detection. *IEEE Sensors Journal, 24*(17), 28383–28396.https;//doi.org/10.1109/JSEN.2024.3430234

[31] Wang, G., Wu, J., Tian, B., Teng, S., Chen, L., & Cao, D. (2022). CenterNet3D: An anchor free object detector for point cloud. *IEEE Transactions on Intelligent Transportation Systems*, *23*(8), 12953–12965. https://doi.org/10.1109/TITS.2021.3118698

[32] Yang, Z., Sun, Y., Liu, S., & Jia, J. (2020). 3DSSD: Point-based 3D single stage object detector. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11040–11048. https://doi.org/10.1109/CVPR42600.2020.01105

[33] Wang, T., Zhu, X., Pang, J., & Lin, D. (2021). FCOS3D: Fully convolutional one-stage monocular 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 913–922. https://doi.org/10.1109/ICCVW54120.2021.00107

[34] Zhang, Y., Hu, Q., Xu, G., Ma, Y., Wan, J., & Guo, Y. (2022). 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*, 18953–18962. https://doi.org/10.1109/CVPR52688.2022.01838

[35] Liu, Z., Wu, Z., & Tóth, R. (2020). SMOKE: Single-stage monocular 3D object detection via keypoint estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 996–997. https://doi.org/10.1109/CVPRW50498.2020.00506

[36] He, C., Zeng, H., Huang, J., Hua, X.-S., & Zhang, L. (2020). Structure aware single-stage 3D object detection from point cloud. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11870–11879.* https://doi.org/10.1109/CVPR42600.2020.01189

[37] Kuang, H., Wang, B., An, J., Zhang, M., & Zhang, Z. (2020). Voxel-FPN: Multi-scale voxel feature aggregation for 3D object detection from LIDAR point clouds. *Sensors*, *20*(3), 704. https://doi.org/10.3390/s20030704

[38] Ye, M., Xu, S., & Cao, T. (2020). HVNet: Hybrid voxel network for LiDAR based 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1631–1640. https://doi.org/10.1109/CVPR42600.2020.00170

[39] Yang, B., Liang, M., & Urtasun, R. (2018). HDNET: Exploiting HD maps for 3D object detection. In *2nd Conference on Robot Learning, 87,* 146-155.

[40] Shi, S., Guo, C., Jiang, L., Wang, Z., Shi, J., Wang, X., & Li, H. (2020). PV-RCNN: Point-voxel feature set abstraction for 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10529–10538. https://doi.org/10.1109/CVPR42600.2020.01054

[41] Wang, H., Chen, Z., Cai, Y., Chen, L., Li, Y., Sotelo, M. A., & Li, Z. (2022). Voxel-RCNN-complex: An effective 3-D point cloud object detector for complex traffic conditions. *IEEE Transactions on Instrumentation and Measurement*, *71*, 2507112. https://doi.org/10.1109/TIM.2022.3165251

[42] Chen, Y., Liu, S., Shen, X., & Jia, J. (2019). Fast point R-CNN. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9775–9784. https://doi.org/10.1109/ICCV.2019.00987

[43] Shi, S., Wang, Z., Shi, J., Wang, X., & Li, H. (2021). 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), 2647–2664. https://doi.org/10.1109/TPAMI.2020.2977026

[44] Feng, C., Xiang, C., Xie, X., Zhang, Y., Yang, M., & Li, X. (2023). HPV-RCNN: Hybrid point–voxel two-stage network for LiDAR-Based 3-D object detection. *IEEE Transactions on Computational Social Systems, 10*(6), 3066–3076. https://doi.org/10.1109/TCSS.2023.3286543

[45] Sheng, H., Cai, S., Liu, Y., Deng, B., Huang, J., Hua, X.-S., & Zhao, M.-J. (2021). Improving 3D object detection with channel-wise transformer. In *Proceedings of the IEEE/CVF*

International Conference on Computer Vision, 2743–2752. https://doi.org/10.1109/ICCV48922.2021.00274

[46] Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,* 3354–3361. https://doi.org/10.1109/CVPR.2012.6248074

[47] Yang, L., Zhang, X., Li, J., Wang, L., Zhu, M., Zhang, C., & Liu, H. (2023). Mix-teaching: A simple, unified and effective semi-supervised learning framework for monocular 3D object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, *33*(11), 6832–6844. https://doi.org/10.1109/TCSVT.2023.3270728

[48] Wang, H., Cong, Y., Litany, O., Gao, Y., & Guibas, L. J. (2021). 3DIoUMatch: Leveraging IoU prediction for semi-supervised 3D object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 14615–14624. https://doi.org/10.1109/CVPR46437.2021.01438

[49] Zhang, J., Liu, H., & Lu, J. (2022). A semi-supervised 3D object detection method for autonomous driving. *Displays*, *71*, 102117. https://doi.org/10.1016/j.displa.2021.102117

[50] Park, J., Xu, C., Zhou, Y., Tomizuka, M., & Zhan, W. (2022). DetMatch: Two teachers are better than one for joint 2D and 3D semi-supervised object detection. In *Computer Vision – ECCV 2022: 17th European Conference*, 370–389. https://doi.org/10.1007/978-3-031-20080-9_22

[51] Lee, J.-H., Lee, J.-K., Kim, J.-S., & Kwon, S. (2024). 3DSimDet: Simple yet effective semi-supervised 3D object detector for autonomous driving. In *2024 IEEE Intelligent Vehicles Symposium*, 2834–2840. https://doi.org/10.1109/IV55156.2024.10588763

[52] Hwang, S., Kim, Y., Kim, S., Bahk, S., & Kim, H.-S. (2023). UpCycling: Semi-supervised 3D object detection without sharing raw-level unlabeled scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 23351–23361. https://doi.org/10.1109/ICCV51070.2023.02134

[53] Peng, L., Yan, S., Wu, B., Yang, Z., He, X., & Cai, D. (2022). WeakM3D: Towards weakly supervised monocular 3D object detection. In *International Conference on Learning Representations*, 1–20.

[54] Qin, Z., Wang, J., & Lu, Y. (2020). Weakly supervised 3D object detection from point clouds. In *Proceedings of the 28th ACM International Conference on Multimedia*, 4144–4152. https://doi.org/10.1145/3394171.3413805

[55] Qin, Z., Wang, J., & Lu, Y. (2022). MonoGRNet: A general framework for monocular 3D object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(9), 5170–5184. https://doi.org/10.1109/TPAMI.2021.3074363

[56] Shi, W., & Rajkumar, R. R. (2022). Self-supervised pretraining for point cloud object detection in autonomous driving. In *2022 IEEE 25th International Conference on Intelligent Transportation Systems*, 4341–4348. https://doi.org/10.1109/ITSC55140.2022.9922494

[57] Chen, H., Huang, Y., Tian, W., Gao, Z., & Xiong, L. (2021). MonoRUn: Monocular 3D object detection by reconstruction and uncertainty propagation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10379–10388. https://doi.org/10.1109/CVPR46437.2021.01024

[58] Chen, Y., Li, N., Zhu, D., Zhou, C. C., Hu, Z., Bai, Y., & Yan, J. (2024). BEVSOC: Self-supervised contrastive learning for calibration-free BEV 3-D object detection. *IEEE Internet of Things Journal, 11*(12), 22167–22182. https://doi.org/10.1109/JIOT.2024.3379471

[59] Yang, H., He, T., Liu, J., Chen, H., Wu, B., Lin, B., & Ouyang, W. (2023). GD-MAE: Generative decoder for MAE pre-training on LiDAR point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9403–9414. https://doi.org/10.1109/CVPR52729.2023.00907

[60] Min, C., Xiao, L., Zhao, D., Nie, Y., & Dai, B. (2024). Occupancy-MAE: Self-supervised pre-training large-scale LiDAR point clouds with masked occupancy autoencoders. *IEEE Transactions on Intelligent Vehicles, 9(7),*5150–5162. https://doi.org/10.1109/TIV.2023.3322409

[61] Hess, G., Jaxing, J., Svensson, E., Hagerman, D., Petersson, C., & Svensson, L. (2023). Masked autoencoder for self-supervised pre-training on LiDAR point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 350–359. https://doi.org/10.1109/WACVW58289.2023.00039

[62] Tian, X., Ran, H., Wang, Y., & Zhao, H. (2023). GeoMAE: Masked geometric target prediction for self-supervised point cloud pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13570–13580. https://doi.org/10.1109/CVPR52729.2023.01304

[63] Wu, D., Liao, M.-W., Zhang, W.-T., Wang, X.-G., Bai, X., Cheng, W.-Q., & Liu, W.-Y. (2023). YOLOP: You only look once for panoptic driving perception. *Machine Intelligence Research*, *20*(6), 952–952. https://doi.org/10.1007/s11633-023-1452-6

[64] Zhou, D., Fang, J., Song, X., Liu, L., Yin, J., & Dai, Y. (2020). Joint 3D instance segmentation and object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1839–1849. https://doi.org/10.1109/CVPR42600.2020.00191

[65] Xie, E., Yu, Z., Zhou, D., Philion, J., Anandkumar, A., Fidler, S., . . . , & Alvarez, J. M. (2022). M^2 BEV: Multi-camera joint 3D detection and segmentation with unified birds-eye view representation. *arXiv Preprint:2204.05088*. https://doi.org/10.48550/arXiv.2204.05088

[66] Meyer, G. P., Charland, J., Hegde, D., Laddha, A., & Vallespi-Gonzalez, C. (2019). Sensor fusion for joint 3D object detection and semantic segmentation. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 1230–1237.

[67] Li, P., & Jin, J. (2022). Time3D: End-to-end joint monocular 3D object detection and tracking for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3885–3894. https://doi.org/10.1109/CVPR52688.2022.00386

[68] Huang, K., & Hao, Q. (2021). Joint multi-object detection and tracking with camera-LiDAR fusion for autonomous driving. In *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 6983–6989. https://doi.org/10.1109/IROS51168.2021.9636311

[69] Koh, J., Kim, J., Yoo, J. H., Kim, Y., Kum, D., & Choi, J. W. (2022). Joint 3D object detection and tracking using spatio-temporal representation of camera image and LiDAR point clouds. *Proceedings of the AAAI Conference on*

Artificial Intelligence, *36*(1), 1210–1218. https://doi.org/10.1609/aaai.v36i1.20007

[70] Hu, H.-N., Cai, Q.-Z., Wang, D., Lin, J., Sun, M., & Krahenbuhl, P. (2019). Joint monocular 3D vehicle detection and tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5390–5399. https://doi.org/10.1109/ICCV.2019.00549

[71] Meng, Z., Xia, X., Xu, R., Liu, W., & Ma, J. (2023). HYDRO-3D: Hybrid object detection and tracking for cooperative perception using 3D LiDAR. *IEEE Transactions on Intelligent Vehicles*, *8*(8), 4069–4080. https://doi.org/10.1109/TIV.2023.3282567

[72] Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., ..., & Beijbom, O. (2020). nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference On Computer Vision and Pattern Recognition*, 11621–11631. https://doi.org/10.1109/CVPR42600.2020.01164

[73] Li, Z. (2022). LiDAR-based 3d object detection for autonomous driving. In *2022 International Conference on Image Processing, Computer Vision and Machine Learning*, 507–512. https://doi.org/10.1109/ICICML57342.2022.10009752

[74] Houston, J., Zuidhof, G., Bergamini, L., Ye, Y., Chen, L., Jain, A., & Omari, S. (2021). One thousand and one hours: Self-driving motion prediction dataset. In *Proceedings of the 2020 Conference on Robot Learning*, *155*, 409–418.

[75] Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., & Benenson, R. (2016). The cityscapes dataset for semantic urban scene understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3213–3223. https://doi.org/10.1109/CVPR.2016.350

[76] Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., & Liu, F. (2020). BDD100K: A diverse driving dataset for heterogeneous multitask learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2636–2645. https://doi.org/10.1109/CVPR42600.2020.00271

[77] Wang, Y., Wang, G., Hsu, H., Liu, H., & Hwang, J. (2021). Rethinking of radar's role: A camera-radar dataset and systematic annotator via coordinate alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2815–2824. https://doi.org/10.1109/CVPRW53098.2021.00316

[78] Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning, 78,* 1–16.

[79] Grigorescu, S., Trasnea, B., Cocias, T., & Macesanu, G. (2020). A survey of deep learning techniques for autonomous driving. *Journal of Field Robotics*, *37*(3), 362–386. https://doi.org/10.1002/rob.21918

[80] Wang, L., Chen, T., Anklam, C., & Goldluecke, B. (2020). High dimensional frustum PointNet for 3D object detection from camera, LiDAR, and radar. In *2020 IEEE Intelligent Vehicles Symposium*, 1621–1628. https://doi.org/10.1109/IV47402.2020.9304655

[81] Wang, L., & Goldluecke, B. (2021). Sparse-PointNet: See further in autonomous vehicles. *IEEE Robotics and Automation Letters*, *6*(4), 7049–7056. https://doi.org/10.1109/LRA.2021.3096253

[82] Wen, L.-H., & Jo, K.-H. (2021). Fast and accurate 3D object detection for LiDAR-camera-based autonomous vehicles using one shared voxel-based backbone. *IEEE Access*, *9*, 22080–22089. https://doi.org/10.1109/ACCESS.2021.3055491

[83] Chen, Y., Liu, G., Xu, Y., Pan, P., & Xing, Y. (2021). PointNet++ network architecture with individual point level and global features on centroid for ALS point cloud classification. *Remote Sensing*, *13*(3), 472. https://doi.org/10.3390/rs13030472

[84] Deng, Y., Shen, J., Wen, H., Chi, C., Zhou, Y., & Xu, G. (2024). Dense frustum-aware fusion for 3D object detection in perception systems. *Expert Systems with Applications*, *238*, 122061. https://doi.org/10.1016/j.eswa.2023.122061

[85] Gong, Z., Lin, H., Zhang, D., Luo, Z., Zelek, J., & Chen, Y. (2020). A frustum-based probabilistic framework for 3D object detection by fusion of LiDAR and camera data. *ISPRS Journal of Photogrammetry and Remote Sensing*, *159*, 90–100. https://doi.org/10.1016/j.isprsjprs.2019.10.015

[86] He, X., Zhang, X., Wang, Y., Ji, H., Duan, X., & Guo, F. (2022). Spatial attention frustum: A 3D object detection method focusing on occluded objects. *Sensors*, *22*(6), 2366. https://doi.org/10.3390/s22062366

[87] Alaba, S. Y., Gurbuz, A. C., & Ball, J. E. (2024). Emerging trends in autonomous vehicle perception: Multimodal fusion for 3D object detection. *World Electric Vehicle Journal*, *15*(1), 20. https://doi.org/10.3390/wevj15010020

[88] Wen, L.-H., & Jo, K.-H. (2022). Deep learning-based perception systems for autonomous driving: A comprehensive survey. *Neurocomputing*, *489*, 255–270. https://doi.org/10.1016/j.neucom.2021.08.155

[89] Park, D., Ambrus, R., Guizilini, V., Li, J., & Gaidon, A. (2021). Is pseudo-LiDAR needed for monocular 3D object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3142–3152. https://doi.org/10.1109/ICCV48922.2021.00313

[90] Simonelli, A., Rota Bulò, S., Porzi, L., Kontschieder, P., & Ricci, E. (2021). Are we missing confidence in Pseudo-LiDAR methods for monocular 3D object detection? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3225–3233. https://doi.org/10.1109/ICCV48922.2021.00321

[91] Peng, W., Pan, H., Liu, H., & Sun, Y. (2020). IDA-3D: Instance-depth-aware 3D object detection from stereo vision for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13015–13024. https://doi.org/10.1109/CVPR42600.2020.01303

[92] Li, P., Chen, X., & Shen, S. (2019). Stereo R-CNN based 3D object detection for autonomous driving. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7636–7644. https://doi.org/10.1109/CVPR.2019.00783

[93] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., & Lu, T. (2022). BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *Computer Vision – ECCV 2022: 17th European Conference*, 1–18. https://doi.org/10.1007/978-3-031-20077-9_1

[94] Wu, Y., Li, R., Qin, Z., Zhao, X., & Li, X. (2024). HeightFormer: Explicit height modeling without extra data for camera-only 3D object detection in bird's eye view. *IEEE Transactions on Image Processing, 34*, 689–700. https://doi.org/10.1109/TIP.2024.3427701

[95] Yang, C., Lin, T., Huang, L., & Crowley, E. J. (2024). WidthFormer: Toward efficient transformer-based BEV view transformation. In *Proceedings of the IEEE/RSJ International*

Conference on Intelligent Robots and Systems, 8457–8464. https://doi.org/10.1109/IROS58592.2024.10801452

[96] Lee, Y., & Park, S. (2020). A deep learning-based perception algorithm using 3D LiDAR for autonomous driving: Simultaneous segmentation and detection network (SSAD-Net). *Applied Sciences*, *10*(13), 4486. https://doi.org/10.3390/app10134486

[97] Liang, H., Jiang, C., Feng, D., Chen, X., Xu, H., & Liang, X. (2021). Exploring geometry-aware contrast and clustering harmonization for self-supervised 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3293–3302. https://doi.org/10.1109/ICCV48922.2021.00328

[98] Mao, J., Xue, Y., Niu, M., Bai, H., Feng, J., Liang, X., ..., & Xu, C. (2021). Voxel Transformer for 3D object detection. In *2021 IEEE/CVF International Conference on Computer Vision*, 3164–3173. https://doi.org/10.1109/ICCV48922.2021.00315

[99] Sun, P., Tan, M., Wang, W., Liu, C., Xia, F., Leng, Z., & Anguelov, D. (2022). SWFormer: Sparse window transformer for 3D object detection in point clouds. In *Computer Vision – ECCV 2022: 17th European Conference*, 426–442. https://doi.org/10.1007/978-3-031-20080-9_25

[100] Meyer, G. P., Laddha, A., Kee, E., Vallespi-Gonzalez, C., & Wellington, C. K. (2019). LaserNet: An efficient probabilistic 3D object detector for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12677–12686. https://doi.org/10.1109/CVPR.2019.01296

[101] Lang, A. H., Vora, S., Caesar, H., Zhou, L., Yang, J., & Beijbom, O. (2019). Pointpillars: Fast encoders for object detection from point clouds. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12697–12705. https://doi.org/10.1109/CVPR.2019.01298

[102] Fan, L., Xiong, X., Wang, F., Wang, N., & Zhang, Z. (2021). RangeDet: In defense of range view for LiDAR-based 3D object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2918–2927. https://doi.org/10.1109/ICCV48922.2021.00291

[103] He, C., Li, R., Li, S., & Zhang, L. (2022). 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*, 8417–8427. https://doi.org/10.1109/CVPR52688.2022.00823

[104] Cheng, J.-H., Kuan, S.-Y., Liu, H.-I., Latapie, H., Liu, G., & Hwang, J.-N. (2024). CenterRadarNet: Joint 3D object detection and tracking framework using 4D FMCW radar. In *Proceedings of the 2024 IEEE International Conference on Image Processing*, 998–1004. https://doi.org/10.1109/ICIP51287.2024.10648077

[105] Saini, L., Su, Y., Tercan, H., & Meisen, T. (2024). Centerpoint transformer for BEV object detection with automotive radar. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 4451–4460. https://doi.org/10.1109/CVPRW63382.2024.00448

[106] Nabati, R., & Qi, H. (2019). RRPN: Radar region proposal network for object detection in autonomous vehicles. In *2019 IEEE International Conference on Image Processing*, 3093–3097. https://doi.org/10.1109/ICIP.2019.8803392

[107] Dalbah, Y., Lahoud, J., & Cholakkal, H. (2023). RadarFormer: Lightweight and accurate real-time radar object detection model. In *Image Analysis: 22nd Scandinavian Conference*, 341–358. https://doi.org/10.1007/978-3-031-31435-3_23

[108] Wang, Y., Jiang, Z., Gao, X., Hwang, J.-N., Xing, G., & Liu, H. (2021). RODNet: Radar object detection using cross-modal supervision. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 504–513. https://doi.org/10.1109/WACV48630.2021.00055

[109] Choi, J. D., & Kim, M. Y. (2023). A sensor fusion system with thermal infrared camera and LiDAR for autonomous vehicles and deep learning based object detection. *ICT Express*, *9*(2), 222–227. https://doi.org/10.1016/j.icte.2021.12.016

[110] Rashed, H., Ramzy, M., Vaquero, V., El Sallab, A., Sistu, G., & Yogamani, S. (2019). FuseMODNet: Real-time camera and LiDAR based moving object detection for robust low-light autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops,* 2393–2402. https://doi.org/10.1109/ICCVW.2019.00293

[111] Zhao, X., Sun, P., Xu, Z., Min, H., & Yu, H. (2020). Fusion of 3D LiDAR and camera data for object detection in autonomous vehicle applications. *IEEE Sensors Journal*, *20*(9), 4901–4913. https://doi.org/10.1109/JSEN.2020.2966034

[112] Zhuang, Z., Li, R., Jia, K., Wang, Q., Li, Y., & Tan, M. (2021). Perception-aware multi-sensor fusion for 3D LiDAR semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 16280–16290. https://doi.org/10.1109/ICCV48922.2021.01597

[113] Kumar, G. A., Lee, J. H., Hwang, J., Park, J., Youn, S. H., & Kwon, S. (2020). LiDAR and camera fusion approach for object distance estimation in self-driving vehicles. *Symmetry*, *12*(2), 324. https://doi.org/10.3390/sym12020324

[114] Wu, Q., Li, X., Wang, K., & Bilal, H. (2023). Regional feature fusion for on-road detection of objects using camera and 3D-LiDAR in high-speed autonomous vehicles. *Soft Computing*, *27*(23), 18195–18213. https://doi.org/10.1007/s00500-023-09278-3

[115] Nobis, F., Geisslinger, M., Weber, M., Betz, J., & Lienkamp, M. (2019). A deep learning-based radar and camera sensor fusion architecture for object detection. In *2019 sensor data fusion: Trends, solutions, applications*, 1–7. https://doi.org/10.1109/SDF.2019.8916629

[116] Yadav, R., Vierling, A., & Berns, K. (2020). Radar+ RGB fusion for robust object detection in autonomous vehicle. In *2020 IEEE International Conference on Image Processing*, 1986–1990. https://doi.org/10.1109/ICIP40778.2020.9191046

[117] Liu, Z., Cai, Y., Wang, H., Chen, L., Gao, H., Jia, Y., & Li, Y. (2022). Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions. *IEEE Transactions on Intelligent Transportation Systems*, *23*(7), 6640–6653. https://doi.org/10.1109/TITS.2021.3059674

[118] Kang, D., & Kum, D. (2020). Camera and radar sensor fusion for robust vehicle localization via vehicle part localization. *IEEE Access*, *8*, 75223–75236. https://doi.org/10.1109/ACCESS.2020.2985075

[119] Ogunrinde, I., & Bernadin, S. (2023). Deep camera–radar fusion with an attention framework for autonomous vehicle vision in foggy weather conditions. *Sensors*, *23*(14), 6255. https://doi.org/10.3390/s23146255

[120] Roche, J., De-Silva, V., & Kondoz, A. (2022). A multimodal perception-driven self evolving autonomous ground vehicle.

IEEE Transactions on Cybernetics, *52*(9), 9279–9289. https://doi.org/10.1109/TCYB.2021.3113804

[121] Roy, D., Li, Y., Jian, T., Tian, P., Chowdhury, K., & Ioannidis, S. (2023). Multi-modality sensing and data fusion for multi-vehicle detection. In *IEEE Transactions on Multimedia, 25,* 2280–2295.https://doi.org/10.1109/TMM.2022.3145663

[122] Xiao, Y., Codevilla, F., Gurram, A., Urfalioglu, O., & López, A. M. (2022). Multimodal end-to-end autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, *23*(1), 537–547. https://doi.org/10.1109/TITS.2020.3013234

[123] Man, Y., Gui, L.-Y., & Wang, Y.-X. (2023). BEV-guided multi-modality fusion for driving perception. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21960–21969. https://doi.org/10.1109/CVPR52729.2023.02103

[124] Huang, J., Qing, L., Han, L., Liao, J., Guo, L., & Peng, Y. (2023). A collaborative perception method of human-urban environment based on machine learning and its application to the case area. *Engineering Applications of Artificial Intelligence*, *119*, 105746. https://doi.org/10.1016/j.engappai.2022.105746

[125] Farahani, M., Razavi-Termeh, S. V., Sadeghi-Niaraki, A., & Choi, S.-M. (2023). A hybridization of spatial modeling and deep learning for people's visual perception of urban landscapes. *Sustainability*, *15*(13), 10403. https://doi.org/10.3390/su151310403

[126] Wang, D., Li, W., Liu, X., Li, N., & Zhang, C. (2020). UAV environmental perception and autonomous obstacle avoidance: A deep learning and depth camera combined solution. *Computers and Electronics in Agriculture*, *175*, 105523. https://doi.org/10.1016/j.compag.2020.105523

[127] Zakaria, N. J., Shapiai, M. I., Ghani, R. A., Mohd Yassin, M. N., Ibrahim, M. Z., & Wahid, N. (2023). Lane detection in autonomous vehicles: A systematic review. *IEEE Access*, *11*, 3729–3765. https://doi.org/10.1109/ACCESS.2023.3234442

[128] Khan, M. A.-M., Haque, M. F., Rakib Hasan, K., Alajmani, S. H., Baz, M., Masud, M., & Nahid, A.-A. (2022). LLDNet: A lightweight lane detection approach for autonomous cars using deep learning. *Sensors*, *22*(15), 5595. https://doi.org/10.3390/s22155595

[129] Huang, J., Choudhury, P. K., Yin, S., & Zhu, L. (2021). Real-time road curb and lane detection for autonomous driving using LiDAR point clouds. *IEEE Access*, *9*, 144940–144951. https://doi.org/10.1109/ACCESS.2021.3120741

[130] Jebamikyous, H.-H., & Kashef, R. (2022). Autonomous vehicles perception (AVP) using deep learning: Modeling, assessment, and challenges. *IEEE Access*, *10*, 10523–10535. https://doi.org/10.1109/ACCESS.2022.3144407

[131] Fayyad, J., Jaradat, M. A., Gruyer, D., & Najjaran, H. (2020). Deep learning sensor fusion for autonomous vehicle perception and localization: A review. *Sensors*, *20*(15), 4220. https://doi.org/10.3390/s20154220

[132] Benamer, I., Yahiouche, A., & Ghenai, A. (2021). Deep learning environment perception and self-tracking for autonomous and connected vehicles. In *Machine Learning for Networking Third International Conference*, 305–319. https://doi.org/10.1007/978-3-030-70866-5_20

[133] Yalamanchili, T. C., Sreedhar, N. S., Nouduri, S., Amritesh, P. J., & Mr, N. (2024). Hybrid optimized A* algorithm for dynamic rerouting based on real time traffic congestion. In *2024 International Conference on Computer, Communications and Allied Sciences*, 1–7. https://doi.org/10.1109/ICCCNT61001.2024.10724594

[134] Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., & Li, H. (2024). End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 46*(12), 10164–10183. https://doi.org/10.1109/TPAMI.2024.3435937

[135] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., & Lu, T. (2025). BEVFormer: Learning bird's-eye-view representation from LiDAR-camera via spatiotemporal transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *47*(3), 2020–2036. https://doi.org/10.1109/TPAMI.2024.3515454

[136] Abdulmaksoud, A., & Ahmed, R. (2025). Transformer-based sensor fusion for autonomous vehicles: A comprehensive review. *IEEE Access, 13,* 41822–41838. https://doi.org/10.1109/ACCESS.2025.3545032

[137] Lang, B., Li, X., & Chuah, M. C. (2024). BEV-TP: End-to-end visual perception and trajectory prediction for autonomous driving. *IEEE Transactions on Intelligent Transportation Systems, 25*(11), 18537–18546 https://doi.org/10.1109/TITS.2024.3433591

[138] Wang, X., Zhu, Z., Huang, G., Chen, X., Zhu, J., & Lu, J. (2024). DriveDreamer: Towards real-world-drive world models for autonomous driving. In *Computer Vision – ECCV 2024: 18th European Conference,* 55–72. https://doi.org/10.1007/978-3-031-73195-2_4

[139] Guan, Y., Liao, H., Li, Z., Hu, J., Yuan, R., Zhang, G., & Xu, C. (2024). World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*. Advance online publication. https://doi.org/10.1109/TIV.2024.3398357

[140] Wang, T.-H., Maalouf, A., Xiao, W., Ban, Y., Amini, A., Rosman, G., & Karaman, S. (2024). Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models. In *2024 IEEE International Conference on Robotics and Automation*, 6687–6694. https://doi.org/10.1109/ICRA57147.2024.10611590

[141] Yang, Y., Zhang, B., Guo, D., Du, H., Xiong, Z., Niyato, D., & Han, Z. (2024). Generative AI for secure and privacy-preserving mobile crowdsensing. *IEEE Wireless Communications, 31*(6), 29–38. https://doi.org/10.1109/MWC.004.2400017

[142] Yang, Y., Du, H., Xiong, Z., Niyato, D., Jamalipour, A., & Han, Z. (2025). Enhancing wireless networks with attention mechanisms: Insights from mobile crowdsensing. *Advance online publication*. https://doi.org/10.1109/MWC.003.2400231

How to Cite: Kalakunnath, N., Varghese, A., Anuradha, A. P., Girish, D. K., & Sasidharan, R. (2025). Advances in Deep Learning for Autonomous Vehicle Perception: A Comprehensive Review. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE52025836