frontiers | Frontiers in Physics

TYPE Review
PUBLISHED 18 October 2024
DOI 10.3389/fphy.2024.1485026

OPEN ACCESS

EDITED BY
Haoyu Chen,
University of Oulu, Finland

REVIEWED BY
Xianlu Tao,
Southeast University, China
Dai Jiangyan,
Weifang University, China

*CORRESPONDENCE
Wenhe Chen,
chenwh@jsut.edu.cn

RECEIVED 23 August 2024
ACCEPTED 02 October 2024
PUBLISHED 18 October 2024

CITATION
He W, Chen W, Tian S and Zhang L (2024) Towards full autonomous driving: challenges and frontiers.
Front. Phys. 12:1485026.
doi: 10.3389/fphy.2024.1485026

Towards full autonomous driving: challenges and frontiers

Wei He[1], Wenhe Chen[2,3]*, Siyi Tian[4] and Lunning Zhang[5]

[1]Shanghai Engineering Research Center of AI & Robotics, Academy for Engineering and Technology, Fudan University, Shanghai, China, [2]Artificial Intelligence Industry Academy School of Computer Engineering, Jiangsu University of Technology, Changzhou, China, [3]Shanghai Huace Navigation Technology Co., Ltd., Shanghai, China, [4]School of Sensing Science and Engineering, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China, [5]Shanghai Future Space-Time Technology Co., Ltd., Shanghai, China

With the rapid advancement of information technology and intelligent systems, autonomous driving has garnered significant attention and research in recent years. Key technologies, such as Simultaneous Localization and Mapping (SLAM), Perception and Localization, and Scene Segmentation, have proven to be essential in this field. These technologies not only evolve independently, each with its own research focus and application paths, but also complement and rely on one another in various complex autonomous driving scenarios. This paper provides a comprehensive review of the development and current state of these technologies, along with a forecast of their future trends.

KEYWORDS

autonomous driving, simultaneous localization and mapping, perception and localization, scene segmentation, deep learning

1 Introduction

Autonomous driving has developed rapidly in the past 2 decades and is now gradually evolving towards full automation. The premise for autonomous vehicles to achieve high-level tasks such as decision-making and planning is to obtain accurate self-state and environmental perception information in various complex scenarios, among which technologies such as Simultaneous Localization and Mapping (SLAM), Perception and Localization, Point Cloud Completion and Scene Segmentation are crucial, as shown in Figure 1. Specifically, SLAM is the basic framework for information association between agents and the environment, which provides agents with the ability to construct and locate real-time environment maps. Agents need to interact with the environment with a high degree of autonomy, and the Perception and Localization technology of autonomous driving systems is particularly critical. It covers a series of advanced functions from environmental perception to precise autonomous positioning. Scene Segmentation greatly enhances the agent's understanding and adaptability to complex scenes by performing detailed semantic analysis of the environment.

This paper will detail the development history, current implementation mechanisms and their practical roles in autonomous driving and broader computer vision and 3D data processing of these key technologies. Through in-depth analysis of the current situation and challenges of these technologies, this paper aims to explore their development trends and forecast how to improve the efficiency and intelligence level of the overall system through technology integration. In addition, it will also predict the future development direction of these technologies and their potential role in promoting the Frontier of automation and intelligent technologies.

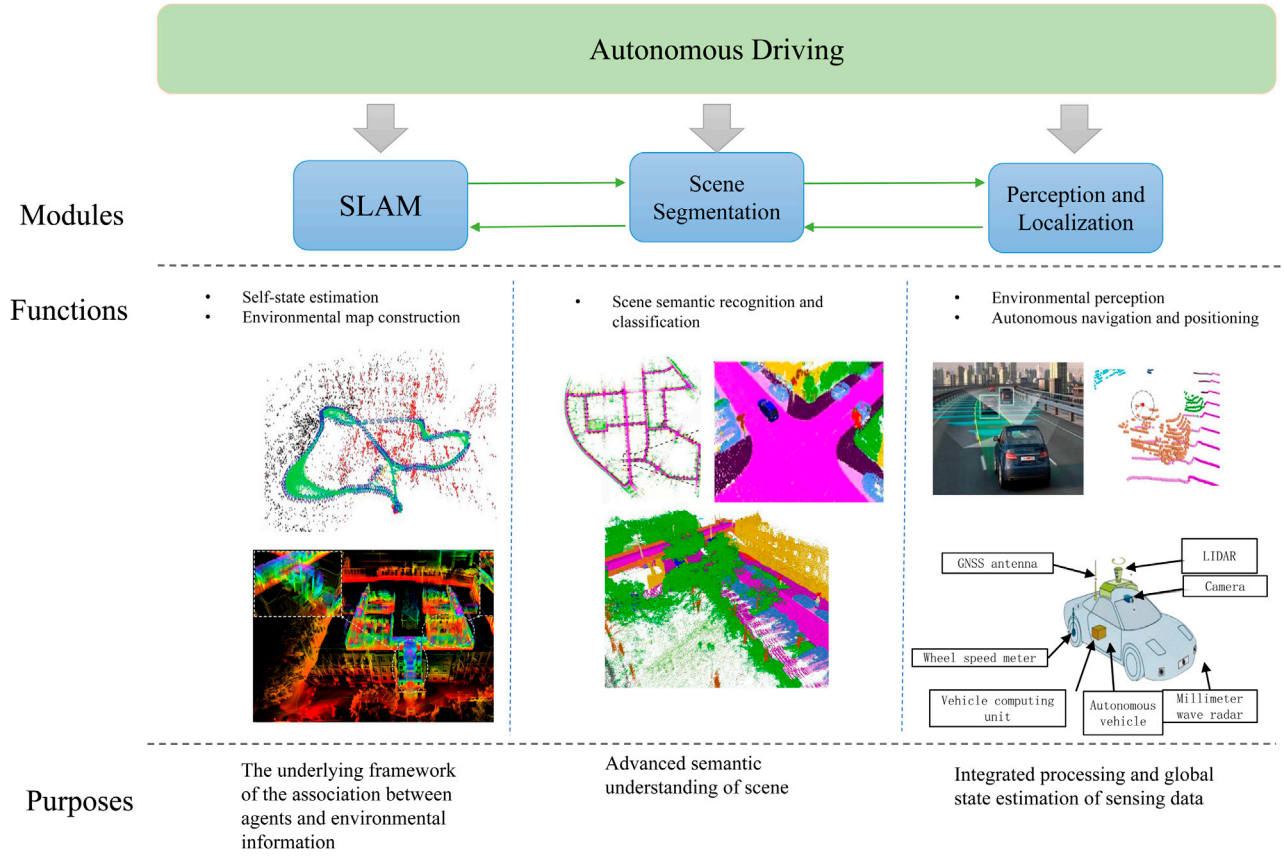

FIGURE 1
The overall framework of the paper.

2 Simultaneous Localization and Mapping (SLAM)

2.1 Definition, basic principles and development history of SLAM

Simultaneous Localization and Mapping (SLAM) is a technology in which a robot estimates its own state (position, speed, direction, sensor bias, etc.) in an unknown environment, and simultaneously constructs its motion environment based on sensor perception information. Over the past 30 years, there were many significant progress made in SLAM field which has been widely used in many industries. The basic principles and development of visual SLAM, laser SLAM and multi-sensor fusion SLAM in the order of different main sensors will be introduced in this section. The specific development route is shown in Figure 2. Due to the widespread application of the fusion of Inertial Measurement Unit (IMU) and SLAM, the development of such applications is also described in 2.1.1 and 2.1.2.

2.1.1 Visual SLAM

In the early stage of visual SLAM research, most of them belong to filtering-based methods, such as EKF-based MonoSLAM [1], tight coupling system composed of IMU and monocular camera [2] which realize real-time operation for the first time. Mourikis et al. [3] proposed the famous MSCKF based on the conventional EKF (Extended Kalman Filter). The state vector of MSCKF contains multiple camera states, and the measurement of the same feature point is used to define constraints between two or more camera poses. When some specific conditions are met, these constraints are used for filter updates. Compared with the conventional EKF method, the advantage of the MSCKF method is to maintain only one state variable with low dimension, and no longer store the coordinate information of map points, so as to reduce the amount of storage and calculation. MSCKF algorithm has become one of the classic algorithms of VIO, but it does not optimize the location of map points in the scene, therefore it is difficult to ensure the overall positioning accuracy for a long time. Optimization-based SLAM method is another mainstream solution, which optimizes the robot pose to be solved and the position of spatial waymark points through Bundle Adjustment (BA) technology. Compared with filtering-based methods, optimization-based methods usually achieve stronger robustness and higher accuracy, and their framework is more flexible. But it is more computational and time-consuming because its multiple iterative optimization process requires more computing resources. In 2007, Klein et al. [4] proposed the famous PTAM (Parallel Tracking and Mapping) algorithm, which applied graph optimization theory to solve SLAM problems for the first time, meanwhile, this algorithm pioneered the parallel implementation of

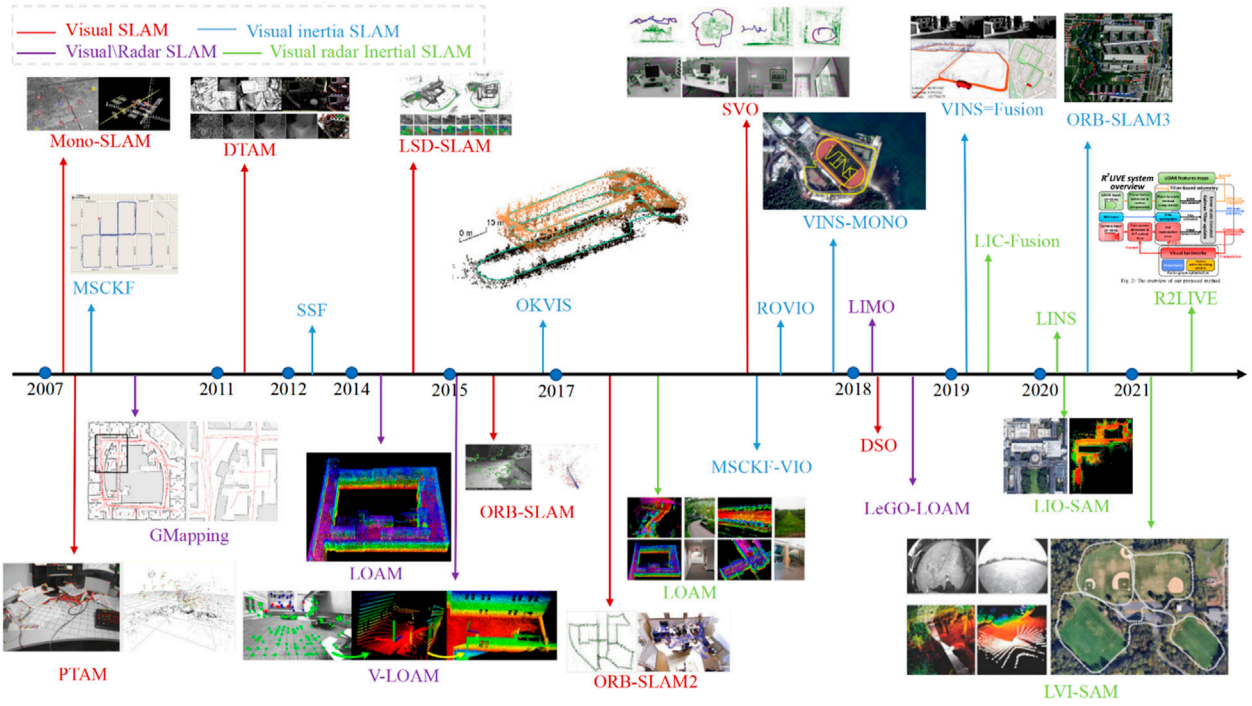

FIGURE 2
SLAM technology development roadmap.

locating and mapping on two independently running threads. Mur-Artalet [5] continued and improved the basic idea of PTAM, and the famous ORB-SLAM algorithm was proposed. ORB-SLAM is a more complete monocular SLAM system, which includes three threads: tracking, partial mapping and loopback detection. In each thread, ORB (Oriented FAST and Rotated BRIEF) operator [6] is used to extract and describe image features. In 2017, Mur-Artal [7] further proposed ORB-SLAM2, which supports both monocular camera input and binocular cameras and RGB-D cameras input. Similar to the filtering-based method, the combination of visual inertial sensors is often used to build visual inertial SLAM system in the optimization-based framework. Leutenegger [8] proposed keyframe-based OKVIS (Open Keyframe-based Visual-Inertial Syetem). In 2021, Campos et al. [9] proposed ORB-SLAM3, which can support monocular, binocular, and RGB-D image input of pinhole or fisheye lens models, and can perform visual, visual inertial, and multi-map SLAM processes. VINS-Mono (Monocular Visual-Inertial Navigation System) proposed by Qin et al. [10] is another representative work of optimization methods. VINS-Mono uses a sparse direct method similar to SVO [11] as its front end, simple corner points are extracted on the image, and the corner points are tracked by KLT (Kanade-Lucas-Tomasi) optical flow method. In constructing the BA, a quaternion-based IMU pre-integration model [12], a sliding window, and a two-step marginalization technique are used. Depending on the type of visual front end, all of the above visual SLAM methods can be referred to as feature point methods. Feature point method has long been regarded as the mainstream method of SLAM, but the disadvantage of this method is that it easily leads to poor feature extraction or feature tracking effect when encountering weak texture environment, fast robot movement speed or blurred visual imaging, which affects the performance of the algorithm, and the key point extraction and descriptor calculation are also time-consuming. Therefore, some researchers have also studied another kind of direct SLAM method, which directly estimates camera motion according to pixel gray information. Concha [13] proposed a monocular visual inertial odometer VIDS (Visual-Inertial Direct SLAM) based on the direct method. Forster proposed a SVO (Semi-Direct Monocular Visual Odometry), which combines feature point method with direct method. SVO executes motion estimation thread and map construction thread in parallel, and can obtain fast and accurate positioning effect when the observation scene is approximately plane. Engel et al. [14] proposed a monocular SLAM method LSD-SLAM (Large-Scale Direct Monocular SLAM) based on direct method, which can obtain more accurate motion estimation in large-scale scenarios and construct large-scale environmental maps. On this basis, Engel et al. [15] incorporated the photometric calibration strategy and further exploited sparsity, and proposed the DSO (Direct Sparse Odometry) algorithm. Stumberg [16] proposed the VI-DSO method with further fusion of IMU measurement information.

In addition, thanks to the development of semantic segmentation technology and object detection technology based on deep learning, the integration of higher-level semantic information into the design and implementation of SLAM algorithm has become a new direction for researchers. Bowman et al. [17] used probabilistic representation to theoretically analyze the solution of semantic SLAM problems, and proposed a theoretical framework for semantic data association and iterative solution in SLAM by using Expectation-Maximum algorithm (EM). On this

basis, Lianos et al. [18] proposed a semantic SLAM algorithm VSO (Visual Semantic Odometry) that uses semantic information to assist visual feature tracking. Yang et al. [19] proposed a monocular SLAM algorithm that fuses indoor plane features (walls, floors, etc.) with object-level road signs. Frost et al. [20] solved the problem of missing scale in monocular SLAM by constructing 2D projection constraints of vehicle targets with known scales in BA. Nicholson et al. [21] proposed a three-dimensional modeling method of object-level road marks, i.e., an ellipsoid is used to represent three-dimensional object road marks, and a semantic constraint residual term with geometric significance is added to the optimization function of BA to improve positioning accuracy. Li et al. [22] proposed that the closed-loop detection function in complex situations such as large viewing angle changes and occlusion can be enhanced by constructing object-level semantic mapping.

2.1.2 Radar SLAM

The measurement data of LiDAR is a point cloud, and each point cloud contains the spatial coordinates of many spatial points in the Ontology coordinate system at the time of LiDAR measurement. LiDAR SLAM uses point cloud registration, i.e., pose estimation is realized by finding the matching item between the source frame and the target frame and inferring the pose transformation from the source frame to the target frame.

Early LiDAR SLAM studies have mainly focused on 2D LiDAR, and several 2D laser SLAM based on filtering and optimization frameworks have been proposed, including EKF-based frameworks, Unscented Kalman Filter (UKF)-based frameworks [23], and classic framework GMapping [24] based on Particle Filter (PF) [25]. A representative work of graph optimization-based methods is GraphSLAM [26].

With the development of technology, SLAM based on 3D LiDAR has gradually become a research hotspot. The research focus of SLAM method based on 3D LiDAR is mainly on point cloud registration because the basic theory of SLAM has gradually matured when 3D LiDAR began to be studied. Iterative Closest Points (ICP) [27] is the most classic point cloud registration method, which correlates points in a source frame with points in a target frame according to the nearest neighbor criterion, and then solves the optimal transformation between two point clouds. Based on ICP, Mendes et al. [28] proposed to achieve positioning by ICP registration between the current frame and key frames, and then detect loopbacks by ICP registration between different key frames. In order to overcome the defects that the original ICP is sensitive to initial values and measurement noise, many variants of ICP were proposed and applied to LiDAR SLAM. According to the curvature, LOAM [29] extracts surface feature points and corner feature points from the point cloud, and these feature points are registered with adjacent frames and world maps through point-surface and point-line ICP to realize low drift pose estimation. Based on LOAM, LeGO-LOAM [30] introduces ground point constraints in inter-frame registration to suppress height drift, and the pitch angle, roll angle and vertical axis coordinates related to height are first optimized, and then other pose components are optimized, which improves the solution efficiency of inter-frame registration. Also based on point-surface ICP, IMLS-SLAM [31] and SuMa [32] represent planes in maps in the form of hidden planes and patches, respectively. ICP based on normal distribution describes the local geometry of the point cloud through the local covariance matrix of the point cloud, so that the registration takes into account the local orientation of the point cloud. Among them, the representative methods are Normal Distribution Transformation (NDT) [33] and Generalized ICP (GICP) [34].

In addition to ICP, researchers are also actively exploring the application of other point cloud registration schemes in LiDAR SLAM. S4-SLAM [35] uses Super4PCSI [36], a method for point cloud registration based on affine invariance of line segment crossover ratio. GP-SLAM+ [37] uses Gaussian process regression to predict "test points" evenly distributed in space on the current point cloud, and then registers them with the results predicted from the map. SegMap [38] uses machine learning to extract feature points and calculate descriptors from the point cloud, adds semantic information to the point cloud, which can achieve more robust registration, and can reach a pose output frequency of 1 Hz, so as to lay a foundation for the introduction of subsequent machine learning methods.

2.1.3 Multi-sensor fusion SLAM

Generally, multi-sensor fusion positioning methods can be divided into loose coupling method and tight coupling method. The former fuses the independent positioning results of single sensors, while the latter fuses the original measurement information of various sensors.

As the cost of sensors decreases, SLAM methods that integrate three or more sensors have attracted more and more attention from academia and industry in order to obtain higher precision and robust performance and further extend the applicable scenarios of SLAM systems. In 2018, Zhang et al. [39] proposed a sequential multi-sensor fusion SLAM-VLOAM. In this method, IMU firstly provides pose prediction for a loosely coupled VIO, and then the localization results of the VIO are further loosely coupled with LiDAR data to realize a pose estimation from coarse to fine. LVI-SAM [40] combines the VIO system and the LIO system to construct a tightly coupled LVIO. Among them, VIO provides the initial value for the point cloud registration of LIO, and the output of LIO system helps the VIO system to initialize and obtain the depth of visual feature points. Moreover, LVI-SAM also detects the working conditions of these two subsystems respectively. When one subsystem fails, the other system can run independently to ensure the robustness of the system. At the back end, LVI-SAM uses a factor map to receive the inter-frame pose constraints provided by the two subsystems to smooth the trajectory and improve the overall estimation accuracy. Based on FAST-LIO and VINS-Mono, R2LIVE [41] uses ESIKF to tightly couple IMU data with camera data and LiDAR data respectively, and uses a local factor map to adjust key frame pose and visual feature point position. LIC-Fusion [42] is based on the architecture of tightly coupled VIO method MSCKF, LiDAR frames are introduced on the basis of visual frames, and the constraints of LiDAR common view features are added between LiDAR frames. Meanwhile, the external parameters and time differences between sensors are estimated as filtering parameters, which achieves tight coupling well. LIC-Fusion2.0 [43] proposes a more robust plane tracking method between LiDAR frames on the basis of LIC-Fusion, which further improves the system performance.

2.2 Application cases of SLAM in different fields

Since SLAM is essentially autonomous positioning and environmental information correlation in unknown environments, and involves a variety of sensors, direct needs exist in many fields. Therefore, the application of SLAM technology in various industries has been fully studied after decades of development, covering robotics, industrial automation, autonomous driving, augmented reality, medical care, aerospace, geology and environmental science, military security, etc.

Robotics is the hottest field of SLAM technology application. In indoor environments, service robots use SLAM for localization mapping to autonomously navigate and perform tasks in hotel, hospital and home environments [44]; SLAM is used for autonomous navigation and intelligent obstacle avoidance of material delivery trolleys on the factory floor to improve logistics efficiency and automation levels [45]; Unmanned aerial vehicles can use SLAM to carry out autonomous flight [46], and realize surveying and mapping, express delivery and other tasks. Autonomous driving vehicles rely on SLAM to build high-precision maps and assist vehicles in path planning, obstacle avoidance and positioning to ensure driving safety [47]. On AR and VR, SLAM enables such devices to build and update virtual environment maps in real time [48], and can be further used for highly immersive gaming experiences, create dynamic and interactive learning environments or help designers create virtual prototypes and simulations in the field of industrial design; In the medical field, SLAM can also be used for surgical navigation, assisting the safe movement of instruments by building a high-precision model of the surgical area. In the military field, SLAM helps reconnaissance drones navigate and position under the denial condition of no external available signals, and realize tasks such as reconnaissance, surveillance, and target tracking [49].

2.3 Key issues and challenges of SLAM technology

2.3.1 Front-end data association

The SLAM front-end module is responsible for feature extraction, description and tracking on the raw measurement data of the sensor, so as to establish data association on continuous time frames. The state of the carrier can be preliminarily estimated and optimized based on the correctly associated image or point cloud frame. The results of front-end estimation are crucial in the accuracy of the whole SLAM system, but modern SLAM systems generally require the front-end to have high real-time and robustness, which puts forward high requirements for the selection and matching of correlation features. Meanwhile, it is also challenging to correctly correlate the sensor data of different modes in time and space because the front end directly manipulates the sensor data. In addition, various degradation scenarios for vision and LiDAR (lack of features, low feature discrimination, and tracking loss caused by fast motion) require the front end to have accurate, reliable, and stable data processing performance.

2.3.2 Back-end state estimation

With the idea of minimizing errors, the back-end state estimation optimizes and modifies the initial estimation provided by the front-end globally or locally, so as to obtain more accurate and robust trajectory and three-dimensional environment map. In addition, when the system detects a loop, the back-end module will cooperate with the loop detection module to introduce new constraints to correct the accumulated error, so as to improve the accuracy and robustness of the whole SLAM system. It is necessary to develop more efficient optimization algorithms and data structures to cope with it because the complexity of back-end optimization may increase with the expansion of state and map scales. Meanwhile, nonlinear optimization is easy to fall into local minimum, so it is necessary to set appropriate initial values, optimization strategies and constraints to solve it. In real-time applications, the back-end module needs to complete the optimization process in a limited time, and it is also a challenge how to achieve better optimization results in the shortest time.

2.3.3 Loopback detection

Loopback detection is a key component of SLAM, especially in navigation and mapping tasks over long distances or large ranges. However, there is the possibility of misjudgment: one is to identify different scenes as the same scene, and the other is to detect the same scene as different scenes. The main reasons for misjudgment are as follows: (1) The scale inconsistency caused by the change of distance ratio between camera and scene at different time points in visual SLAM. (2) The judgment error caused by the change of viewing angle when observing the same scene at different time points. (3) Dynamic objects may be incorrectly identified as cyclic features, and may also cause changes in the location and appearance of the visited scene. The front-end module of the system may also generate erroneous guidance when tracking dynamic targets. (4) Weather, time, season and other factors may change the characteristics of the same scene. All the above items are all challenges in SLAM loopback detection.

2.4 Future development direction of SLAM technology

2.4.1 Deep learning-based SLAM

At present, deep learning has shown its potential in the field of SLAM, and there are studies on the introduction and replacement of deep learning methods in each module, including image matching [50, 51], point cloud registration [52], semantic segmentation [53], closed-loop detection [54] and pose estimation [55], etc. In addition, SLAM systems directly based on end-to-end networks [56] also appeared. All the above studies have injected new vitality into the field of SLAM, but so far SLAM methods based on deep learning have not been able to reach the accuracy and reliability of conventional methods. The future development trends of learning-based SLAM systems include: (1) Deep learning networks are needed for online learning on long-term SALM systems in open environments to cope with new scenes and objects independent of training data. (2) Deep learning networks are inseparable from training data. Learning-based SLAM is highly dependent on the richness of training data, requires a lot of labeling

work, and needs to explore low-sample learning techniques. (3) At present, many large models have emerged in the field of deep learning. They have the advantages of powerful data processing capabilities, complex problem solving capabilities, high precision and high performance. Large models are expected to be deployed in SLAM systems to achieve all-round improvement in the future.

2.4.2 Multi-agent collaborative SLAM

Multi-agent refers to the overall system in which various forms of intelligent robots cooperate to complete complex tasks according to task division in a certain time and space [57]. Due to the limitation of the endurance time of a single robot, the efficiency of obtaining 3D information is low with small range; Moreover, it is difficult to comprehensively analyze the complex structure and scene information in real time due to the limitation of working mode. Meanwhile, SLAM has error accumulation characteristics, which makes it difficult to ensure the accuracy of long-term and large-scale mapping. These problems can be solved through the collaborative SLAM of multiple agents. The realization of multi-agent SLAM requires multiple agents to cooperate in a single-machine or cross-machine collaboration manner. Meanwhile, multiple agents share scene maps and perform information interaction and fusion, so as to significantly improve the efficiency, accuracy and robustness of single SLAM.

2.4.3 New type sensors

A variety of new sensors are expected to be introduced into SLAM system with the development of sensor technology. For example, the Event Camera, which is designed to imitate the animal vision system to record the time and location of the event stream. Compared with conventional cameras, it has the advantages of no motion blur, sub-millisecond time delay and ultra-high dynamic range, which has been applied to feature tracking [58], optical flow [59], 3D reconstruction [60], and SLAM [61]. However, due to the uniqueness of event cameras, the processing of noise and spatiotemporal information is different from that of traditional vision, and all task-level algorithms need to be redesigned [62].

3 Perception and positioning technology for autonomous driving

3.1 The importance of perception and positioning technology in autonomous driving system

In the autonomous driving system, the main task of perception and positioning is to obtain the environmental information around the vehicle through relevant sensors, and determine the position and attitude of the vehicle in the environment, so that the vehicle can achieve safe driving under complex traffic road conditions. Perception technology identifies road conditions, obstacles, traffic signs, and other vehicles based on vehicle sensor data. This kind of understanding of the environment is crucial for the vehicle, because it must be able to dynamically respond to rapid changes on the road, such as avoiding sudden obstacles and adapting to different environmental conditions such as weather and light. Positioning technology can estimate the motion state quantities of the vehicle, including position, pose and speed, in real time and accurately based on the vehicle sensor information, so as to meet the demand of other functional modules of the autonomous driving system for motion state information. Perception and positioning technology provides key underlying information and support for the autonomous driving system, and provides the foundation for the advanced functions of the system such as decision-making and planning, which directly affects the safety, efficiency and reliability of autonomous vehicles. Autonomous driving perception and localization technologies are explained from two aspects: perception and localization. Perception technologies include visual perception, LiDAR perception, and millimeter wave radar sensing, while localization technologies include inertial odometer, satellite navigation and positioning, wheel speed odometer, and map matching. The specific technology is shown in Figure 3.

3.2 Autonomous driving perception technology

Real-time, accurate and robust perception of road traffic environment is the basic but most challenging task in autonomous driving. By equipped with multi-modal sensors, the autonomous driving system needs to accurately identify information such as the type, location, trajectory and motion status of targets in road traffic. Autonomous driving perception technology can be mainly divided into visual perception, LiDAR perception and millimeter wave radar perception according to sensor principles.

3.2.1 Visual perception

Vision sensors can obtain images with rich color, texture and semantic information with low cost, so they are widely used in perception tasks of autonomous driving, including traffic target detection, drivable area segmentation and lane line recognition [63]. Object detection ensures the safety of autonomous driving by identifying and locating traffic targets such as vehicles, pedestrians, cyclists, and traffic signs. Object detection methods can be divided into two categories: two-stage networks and single-stage networks. Two-stage networks (such as the R-CNN series, including Fast R-CNN [64] and Faster R-CNN [65]) achieve high accuracy through regional proposal method, but with slower inference. On the other hand, single-stage networks (such as SSD [66] and YOLO [67]) sacrifice partial accuracy in exchange for faster inference speed by simultaneously handling bounding box regression and target classification. Such networks divide input images into meshes or use anchor boxes of various sizes to extract multi-scale features. For autonomous driving scenarios, D. Gragnaniello [68] proposed a 2D multi-object detection and tracking algorithm to solve the problem of multi-class object detection and tracking. OVTrack proposed by Li et al. [69] handles the detection and tracking of arbitrary object classes through visual language models. Huang et al. [70] proposed a multi-object tracking algorithm based on self-supervised appearance model.

Drivable area segmentation enables autonomous vehicles to effectively plan safe trajectories by identifying drivable areas on the road. CNN-based deep learning models perform well in semantic segmentation, which are widely used for pixel-level

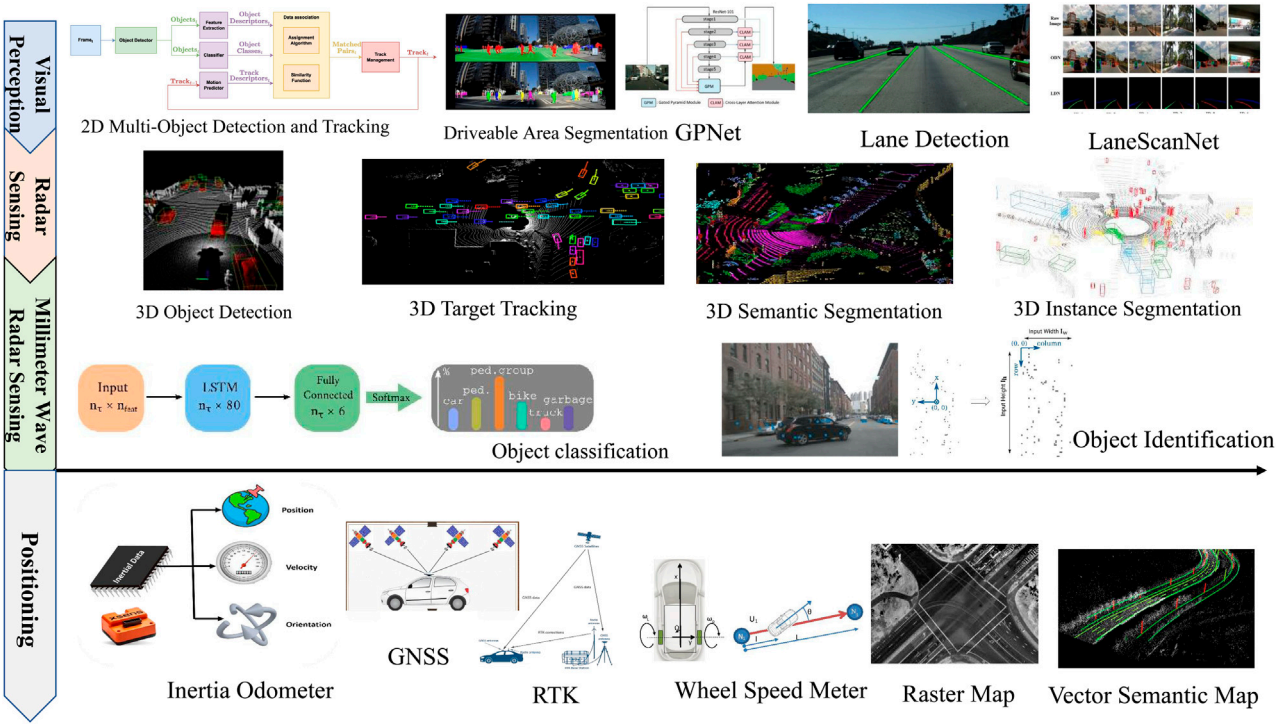

FIGURE 3
Schematic diagram of perception and positioning technology of autonomous driving.

segmentation of drivable regions. Xia and Kim [71] proposed a semantic segmentation architecture that combines multiscale contextual features and low-level features, using hybrid spatial pyramid pooling and global attention fusion. Zhang et al. [72] proposed a GPNet for traffic scene segmentation, combining multi-scale features of gating and pairwise techniques. SegFormer [73] provides sufficient segmentation efficiency and performance through a position-independent hierarchical Transformer encoder and lightweight decoder network. Real-time semantic segmentation is a prerequisite for autonomous driving which needs to achieve competitive segmentation accuracy at low computational costs. DSANet [74] is a computationally efficient network consisting of channel segmentation and shuffling modules and dual attention modules using expanded spatial attention and channel attention to achieve higher segmentation accuracy and lower computational cost.

Accurate lane marker detection and segmentation enable autonomous vehicles to remain within the appropriate lane for precise trajectory control. Conventionally, lane lines are detected using a Canny edge detector [75] and then located in the scene using either a Hough transform [76] or RANSAC [77]. However, these methods are susceptible to illumination and occlusion [78]. CNN-based deep learning models overcome these limitations by annotating lane segments at the pixel level [79]. Zou et al. [80] adopted a segmentation method based on multimodal fusion network for lane detection. Qin et al. [81] proposed an anchor frame-driven sequential classification method for lane detection, which can significantly reduce the computational cost. LaneScanNET [82] assists autonomous driving systems in lane change or lane keeping decisions by combining obstacle detection networks (ODN) and lane detection networks (LDN). The proposed architecture combines the results of obstacle detection and lane line segmentation to predict the obstacle lane state in the field of view of autonomous vehicles. DSUNet [83] is a UNet-based architecture designed for lane detection and path prediction in autonomous driving, using deep separable convolution for faster inference in real-time autonomous driving.

In addition, the visual perception system can obtain different types of information through multiple configurations such as monocular, binocular, and multi-ocular cameras. Such multi-modal data helps to improve the robustness and accuracy of perception, such as obtaining depth information through binocular vision, and achieve panoramic perception through multi-eye vision. However, visual perception technology relies heavily on ambient lighting conditions. The effect of visual perception will be greatly reduced under low light, strong light, backlight and night conditions, requiring additional processing and compensation technology; The visual perception system is easily affected by obstructions, resulting in part of the visual field being blocked. In open environment, visual perception system can provide comprehensive environmental information, but it needs to be used in combination with other sensors in complex environment to make up for the deficiency of visual perception [84].

3.2.2 LiDAR perception

LiDAR directly measures the distance of traffic scenes by transmitting and receiving laser beams to obtain high-precision point cloud data. There are different processing methods for point

cloud data. The projection method tries to project the point cloud data into a two-dimensional plane, and then uses a two-dimensional method to process it. Another part of the research voxelizes 3D point cloud data (Voxelization), i.e., the space is divided into small cubes (called voxels). However, a large amount of original information is lost in the process of preprocessing data whether it is the projection method or the voxelization method, and the full performance of high-precision LiDAR cannot be exerted. In order to make full use of the collected information, the mainstream method perception tasks in recent years directly use point cloud data [85]. LiDAR is also used for a variety of perception tasks on autonomous vehicles, such as 3D target detection, 3D target tracking, 3D semantic segmentation, and instance segmentation [86].

LiDAR operates independently of natural light, providing reliable environmental awareness day and night and in various weather conditions. Direct ranging is more accurate and reliable than visual inference of depth information, especially in long-distance and complex scenes. However, the relatively high cost of LiDAR limits its large-scale commercial application, but the cost is expected to decrease with technological progress and expansion of mass production. Due to the large amount of three-dimensional point cloud data generated by LiDAR, it requires powerful data processing capabilities and efficient algorithms for real-time processing and analysis. Therefore, higher requirements are put forward for the computing platform of autonomous driving systems, and data processing pipelines and algorithms need to be optimized to meet real-time needs. And although it can work in a variety of weather conditions, the attenuation and scattering of laser signals may affect the measurement accuracy in extreme environments such as dense fog and heavy rain and snow.

3.2.3 Millimeter wave radar sensing

Millimeter-wave radar was mainly used in automotive assisted driving systems in the past. In recent years, with the improvement of semiconductor radio frequency technology, millimeter-wave radar has shown huge advantages in bandwidth, size and cost, and has also shown great application potential in advanced perception tasks of autonomous driving. Scholars have studied the problem of target recognition based on millimeter wave radar point cloud. These methods are divided into two categories: one is to extract information through hand-designed feature extractors, such as Schumann et al. [87] obtain the target area through clustering, and classify pedestrians, vehicles and other targets based on hand-designed multi-dimensional features; The other is to directly extract features through deep neural networks. Danzer et al. used PointNet [88] and PointNet++ [89] methods for pedestrian and vehicle target recognition respectively, and Lombacher et al. [90, 91] converted radar point cloud into rasterized data, and then proposed a series of CNN methods for feature extraction and target recognition.

3.3 Autonomous driving positioning technology

Autonomous driving positioning technology can accurately estimate the motion state quantities of the vehicle in real time based on the vehicle sensor information to meet the functional requirements of other autonomous driving modules. The following is an introduction based on the main sensing equipment [92].

3.3.1 Inertial odometer

Inertial odometers use the measurement values of inertial devices such as gyroscopes and accelerometers to estimate the carrier's running trajectory. The calculation accuracy depends on the measurement accuracy and stability of inertial devices. For cost considerations, autonomous vehicle platforms usually deploy consumer-grade inertial devices based on Micro Electro Mechanical System (MEMS) structure. MEMS inertial devices often have large measurement noise and complex error characteristics. In order to improve navigation and positioning accuracy, they need to be compensated for their errors in use. The errors of MEMS inertial devices can be roughly divided into static errors, dynamic errors and random errors. Among them, static error and dynamic error are generally considered to be deterministic error related to the motion state of the carrier, and static error can be compensated by offline calibration method [93], or online estimation through other sensor information, while dynamic error is difficult to calibrate or estimate. For random error, it cannot be eliminated by calibration or estimation method, but only an identification model can be established to estimate the parameters of random error. In addition, researchers have explored the application of Gauss-Markov processes [94], wavelet transform methods [95], generalized wavelet moment methods [96] in random error identification and denoising. Due to the complexity of dynamic error and random error models, researchers have begun to try to model inertial odometer errors in a data-driven way in recent years. Martin Brossard et al. [97] employed CNN to predict the bias error of gyroscopes online based on time series window data. Another solution is to directly use neural network to model the calculation process of inertial odometer in an end-to-end way. Joao Paulo et al. [98] encoded the original angular velocity measurement and acceleration measurement input into a discrete CNN channel, and then used a bidirectional Long Short-Term Memory (LSTM) network to encode the time series inertial information to predict the pose increment in an end-to-end manner.

3.3.2 Satellite navigation and positioning

The Global Navigation Satellite System (GNSS) uses navigation satellite wireless signals to perform pseudo-range or carrier ranging, calculates the geometric intersection of spatial straight lines based on the ranging information, and estimates the position of the signal receiver in the global coordinate system. GNSS is widely used in location services for autonomous driving due to its simplicity, speed and wide coverage. Standard Point Positioning (SPP), also known as pseudo-range single point positioning, is the most common GNSS positioning method. Influenced by clock error, ionospheric interference, tropospheric interference and other factors, the positioning accuracy of SPP is low. Differential GNSS technology eliminates the temporal and spatial correlation factors such as satellite orbit error, clock error, ionospheric error and tropospheric error by differentiating satellite signals with similar geographical locations, and improves the stability of satellite positioning. In order to improve satellite positioning accuracy, carrier phase ranging technology was born, which can provide centimeter-level ranging accuracy. Combining carrier ranging and

difference principles gave birth to real-time dynamic carrier phase difference technology (Real time Kinematic, RTK) [99], RTK can complete the solution of the ambiguity of the whole circle in a short time and provide position measurement up to centimeter level. At this stage, the RTK technology of reference station network with wide coverage is formed mainly by establishing multiple RTK reference stations for networking and using wireless networks to transmit differential signals.

Satellite signals are easily affected by clock error, clock drift, clock jump, etc. during transmission, which will result in data failure. Therefore, it is necessary to enhance the reliability of self-localization through fault detection. Traffic accidents caused by positioning deviation can be effectively avoided by analyzing the validity of observation data, identifying and eliminating fault data. At present, localization fault detection is usually divided into three categories: snapshot detection, sequence detection, and density anomaly detection [100]. Snapshot detection mode focuses on the consistency test of current observations, which can identify step faults more accurately. The sequence detection method comprehensively uses historical data and current data for consistency test, which can effectively improve the detection effect of slope faults. Furthermore, the distribution uncertainty of observed data and the dependence on prior knowledge can be overcome through identifying anomaly localization data based on the density difference between current data and neighboring data. In the actual operation scenarios of autonomous vehicles, GNSS positioning faces the risk of signal interference; In scenes such as tree-lined road sections, high-rise streets, and under viaducts, blocked by environmental obstacles, the multi-path effect caused by multiple reflections and propagation of satellite signals will greatly interfere with the signal calculation ability of the receiver, resulting in deviations in position and speed measurements. In scenarios such as tunnels and underground garages, satellite signals are completely blocked, and GNSS will completely lose its positioning capabilities [101].

3.3.3 Wheel speed odometer

The wheel speed odometer recovers the motion state of the vehicle from the wheel speed information measured by the wheel speed meter. Wheel speed information is essentially the observation information of the vehicle moving speed. Compared with inertial navigation, the number of integrations involved in recovering the vehicle position state through wheel speed is fewer, so the wheel speed odometer is generally more accurate than the inertial odometer. Wheel speed odometers also face the problem of error accumulation, and researchers are also trying to use data-driven methods to improve the accuracy of wheel speed odometers. Uche Onyekpe et al. [102] used the position error between the speed difference model and the GNSS measurement as a neural network supervised signal to train the LSTM network, and the output of the network was used to compensate the position output of the speed difference model; After that, the team further proposed a structurally optimized wheel speed odometry network WhONet [103], using Recurrent Neural Network (RNN) to improve the real-time performance of prediction. Experiments show that the accuracy of this method exceeds the conventional speed differential motion model. Martin Brossard et al. [104] used Gaussian Processes (GP) to model the wheel speed model and its uncertainty, and combined variational inference to train the neural network as the kernel function of GP to reduce the computational complexity of GP.

3.3.4 Map matching

Map matching technology matches the positioning features provided by high-precision maps with sensor signals to estimate the position and pose of the vehicle in the map. Different from the SLAM system, high-precision map features are collected through professional mapping equipment, and converted into the global coordinate system through offline optimization and other steps, with excellent position accuracy. Therefore, map matching based on high-precision maps can achieve high-precision global positioning. According to the feature form of map positioning and the type of vehicle sensor, map matching technology has different implementation ideas. In the early autonomous driving, map matching technologies with LiDAR as the main body have been widely studied, such as ICP [29] and NDT [35]. In the grid positioning method proposed by Jesse Levinson et al. [105], the high-precision map records the environmental reflection intensity and elevation information in a plane two-dimensional raster, and the map matching process uses histogram filtering to calculate the likelihood probability corresponding to pose sampling points. Compared with dense point cloud map scheme, vector semantic map models road objects in the environment with parametric geometric vector shapes, and records its geometric attributes and semantic category attributes. Its lightweight characteristics are beneficial to real-time transmission applications of autonomous driving. In addition, compared with conventional visual descriptors, semantic tags, as higher-level abstract information, are less affected by changes in light conditions, seasonal weather changes, and dynamic obstacle occlusion [106]. Therefore, high-precision vector semantic maps have the potential for large-scale application deployment.

3.4 Challenges and future development directions of autonomous driving perception and positioning technology

Although the perception and positioning technology of autonomous driving system has made great progress, with the continuous improvement of the intelligence of autonomous driving vehicles, the requirements for corresponding technologies are also constantly increasing, and the current technology still faces some challenges. For example, how does the system maintain the accuracy and robustness of perception in complex scenes and environments such as severe weather like rain and fog, low-recognition scenes with insufficient lighting conditions, and urban congested road sections; Ensure the accuracy and reliability of positioning in GNSS occlusion or denial environments such as tunnels and urban canyons; Strike a balance between real-time performance and computing resource cost when a large number of sensors and computing tasks are involved. In view of these challenging problems, there are the following future development directions.

3.4.1 Multi-source fusion sensing and localization

The data of a single sensor will fail in some environments. The current practical solution is to combine multiple complementary sensors to compensate for their respective shortcomings. Different environments rely on different sensor combinations for effective sensing [107]. In the future, multi-sensor fusion will further develop in the direction of multi-modality, scalability and low computing requirements, thereby achieving robust and reliable real-time perception and positioning.

3.4.2 Collaborative perception

There are blind spots and limited perception range in the sensor perception of a vehicle. With the continuous development of intelligent network connection technology composed of wireless communication V2X [108] (Vehicle to Everything, including V2V: Vehicle to Infrastructure and V2P: Vehicle to Persons), a new generation of autonomous driving perception technology will further develop to the level of high-dimensional network connection collaborative perception. The information of vehicles, roads, traffic facilities and pedestrians can be shared and interacted through V2X to achieve integrated, global and high-performance traffic status collaborative perception.

3.4.3 Unified perspective perception

In recent years, the Bird's Eye View (BEV) [109] unified perception large model based on surround-view camera has attracted a lot of attention from academia and industry, and has become a hot spot in autonomous driving perception research. The BEV perception paradigm converts the information of the vehicle-mounted surround-view sensor into the BEV space through a series of operations, and represents it in the vehicle body coordinate system in the form of a two-dimensional spatial grid. Accordingly, a series of perceptual tasks share the same BEV spatial features, and perform neural network decoding for their respective task objectives. The BEV awareness model is expected to be constructed as a large-parameter neural network model that supports multi-modal, long-time series data input and is oriented to multi-task applications.

4 Scene segmentation technology

4.1 Application of the definition of scene segmentation technology in 3D data processing

Scene segmentation aims to divide the whole three-dimensional scene into several regions with different semantics, which refers to the category information of real objects observed by scene data. Scene segmentation is the foundation of scene understanding and plays an important role in various fields involving 3D data processing. In autonomous driving, scene segmentation is used to identify roads, vehicles, pedestrians and other obstacles, and generate semantic maps of the surrounding environment of vehicles in real time, providing a basis for decision-making of autonomous driving system; In robotics, scene segmentation helps robots understand their working environment, correctly identify work areas and paths, and enable them to navigate autonomously and interact with the environment; In medical image processing, for three-dimensional CT or MRI data, scene segmentation technology can be used to identify and label different organs and diseased areas, thereby improving the accuracy of diagnosis; In the field of remote sensing mapping, scene segmentation can be used for environmental monitoring, urban modeling and so on.

4.2 Classification of scene segmentation techniques

In various applications, most of the objects processed by scene segmentation are represented in the form of point clouds, i.e., the three-dimensional data obtained by scanning and reconstructing the real scene with depth sensors. Since point cloud data is usually disordered, unorganized, and unstructured, and point clouds are huge in open scenarios, it is extremely challenging to segment it and semantically label each point. From the method point of view, semantic segmentation can be divided into: (1) semantic segmentation based on 2D-3D mapping; (2) voxel-based segmentation method; (3) semantic segmentation based on graph convolution; (4) semantic segmentation based on sparse convolution; (5) semantic segmentation based on point convolution. The development route of scene segmentation technology is shown in Figure 4.

4.2.1 2D-3D mapping-based method

Compared with three-dimensional computer vision, two-dimensional vision has a longer development history, so in some methods, the semantic segmentation problem of three-dimensional point clouds is tried to be solved by using technologies in the field of two-dimensional vision. V-MVFusion [110] proposes a two-dimensional projection [111] from multiple perspectives to represent a three-dimensional point cloud, and then uses a two-dimensional semantic segmentation network framework [112–114] to process the two-dimensional projection. Based on one-way feature mapping, a bidirectional fusion between two-dimensional features and three-dimensional features is proposed, i.e., two-dimensional image segmentation and three-dimensional point cloud segmentation are performed simultaneously on the scene, and two-way feature mapping is performed in the decoder network, and the experimental results show that bidirectional mapping can improve the performance of semantic segmentation of 3D point clouds better than unidirectional mapping. Because the mapping between point cloud and image often involves preprocessing operations of depth map and occlusion information estimation, the early semantic segmentation methods of point cloud are difficult to be applied in practice. In order to solve this problem, DeepViewAgg [115] proposes a mapping method without preprocessing operations, which can estimate the pixel depth in real time to obtain the correspondence between points and pixels. For the point cloud semantic segmentation method based on 2D-3D mapping, its advantages are that on the one hand, it can make full use of the mature segmentation technology in the field of image, and on the other hand, the 2D image features from multiple perspectives can provide rich context information for 3D semantic segmentation. However, such methods require additional 2D image data and

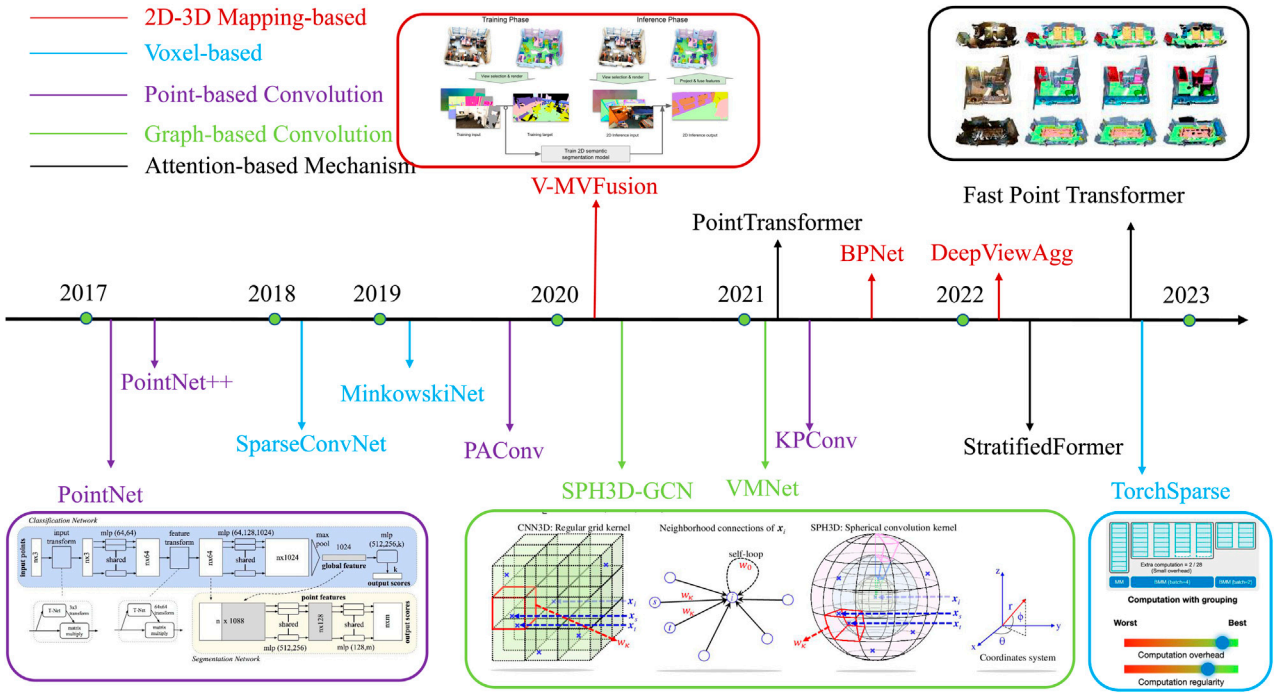

FIGURE 4
Development roadmap of scene segmentation technology.

involve complex multi-view projections, so they rely too much on the choice of camera viewing angle.

4.2.2 Voxel-based method

In order to reduce the reliance on redundant image and perspective information, some methods [116, 117] choose to convert point clouds into three-dimensional voxels, i.e., spatially small-volume elements, and then use sparse convolution for semantic segmentation. Sparse convolution concentrates the computation on a non-empty voxel grid, which can effectively reduce the computational overhead. However, since the convolution operation may pass the features of one non-empty voxel to multiple voxels, the number of non-empty voxels will always be high as the multi-layer network is convolved. To solve this problem, SparseConvNet [118] proposes a submanifold sparse convolution operation. This operation requires that for a certain voxel grid point, its non-empty condition is that the central grid point of the receptive field is also non-empty. This method can effectively reduce the number of non-empty voxel grid points, improve the segmentation performance and reduce the computational overhead. After that, more work has been done to try to improve the efficiency and performance of sparse convolution. MinkowskiNet [119] proposes a 4D sparse convolution network, which can process the time series data of three-dimensional point clouds through sparse convolution, and has achieved good results on both indoor and outdoor scene data sets. Haotian et al. [120] further proposed that TorchSparse should be used to improve problems such as computational irregularity and high video memory occupancy in sparse convolution processes. The main advantage of voxel-based method is that it has high efficiency in processing point clouds and is easy to be applied to large-scale point cloud scene data; The disadvantage is that the point cloud needs to be voxelized first. Some key details may be lost when the voxel resolution is low.

4.2.3 Point convolution-based method

Also due to the success of convolutional neural networks (CNNs) in the field of images, a lot of work has been done to try to migrate convolutional operations to point clouds. Point cloud segmentation is a technique in computer vision and 3D graphics used to divide point cloud data into different regions or categories. A point cloud is a set of discrete points representing objects or scenes in three-dimensional space, obtained through scanning devices such as LiDAR or 3D cameras. These points typically contain coordinate information (X, Y, Z), and sometimes include additional attributes such as color or intensity. The purpose of point cloud segmentation is to divide these points into meaningful subsets, such as separating buildings, roads, vehicles, and pedestrians from a complex point cloud. PointNet [121] and PointNet++ [122] are representative works in this regard, which aggregate global or local features through max-pooling operations to avoid the negative effects of point cloud disorder. PointNet++ proposes hierarchical local feature aggregation based on PointNet, which is used to improve the network's ability to recognize local features, and lays the foundation for more semantic segmentation methods based on point convolution in the future. KPConv [123] used kernel points to replace the convolution kernel of conventional convolution operations. The features of input points in the convolution process are obtained by the weighted sum of features of adjacent kernel points. Because the kernel points are continuously distributed in the geometric space, their positions can be learned

through the network, and this variable convolution operation can effectively adapt to the problem of uneven local point distribution in the point cloud. On this basis, PAConv [124] used the ScoreNet network to estimate the weighting coefficients of kernel points in the convolution process, and further improved the performance of point convolution through the learnability of the network. The advantage of this method is that it directly processes the point cloud without additional image data or data conversion operations, so it can retain the detailed information of the point cloud to the maximum extent.

4.2.4 Graph convolution-based method

Graph convolutional neural network (GNN) is a kind of neural network that specializes in dealing with graph structure. The spatial interaction between three-dimensional points in a point cloud can be represented by a graph, and each point is used as a node of the graph. Therefore, it tries to apply graph convolutional network to point cloud semantic segmentation in some work. L Jiang et al. [125] proposed to enhance point cloud semantic segmentation by an edge feature branch that uses graph convolution techniques to explicitly establish the semantic relationship of each point with its neighborhood points and extract contextual information within the local neighborhood. Similarly, SPH3D-GCN [126] proposes a spherical kernel-based graph convolution operation for point cloud processing, which also directly establishes local graph relations through point coordinates. Another way to use graph convolution to point cloud semantic segmentation is to additionally use the grid model corresponding to the point cloud. Since the grid model has its own coordinate and edge information, graph convolution network can be well applied. DCM-Net [127] proposes to extract geodesic information on the grid model through graph convolution operation, and uses two convolution operations to extract Euclidean distance and geodesic distance respectively, and fuses the two kinds of information through feature stitching. VMNet [128] uses a dual-branch network structure to process point clouds and grid models separately, and proposes a feature fusion module based on attention mechanism to selectively perform fusion, thereby improving the performance of this method in semantic segmentation.

4.2.5 Attention mechanism-based method

As attention mechanism shows powerful feature representation capabilities in the fields of natural language processing and computer vision, and it also tried to apply it to semantic segmentation of 3D point clouds in many works. PointTransformer [129] is one of the representative works. Different from previous scalar attention mechanisms, this method proposes a vector attention mechanism for point clouds and uses learnable position coding to improve the network's ability to capture spatial geometric information. However, this method uses a local attention mechanism to reduce the computational overhead. When dealing with complex scenes, it is necessary to superimpose multiple layers of attention modules to expand the receptive field of features. To solve this problem, StratifiedFormer [130] proposes a hierarchical attention mechanism to establish long-distance relationships between features. For each point, this method will simultaneously sample adjacent points in its nearer and farther distances to calculate attention. The sampling is denser in the nearer distance and sparser in the far distance, which can directly expand the receptive field. In addition, some efforts have been made to improve the attention mechanism in point cloud semantic segmentation in terms of efficiency and performance. Fast Point Transformer [131] utilizes a voxel hash architecture to speed up attention modules. Point Transformer V2 [132] groups vector attention on the basis of Point Transformer, further strengthens position coding information, and improves the robustness of network processing point clouds.

4.3 Advantages, disadvantages and development trends of existing scene segmentation technologies

Existing scene segmentation technologies are outstanding in high precision and detail capture, which can achieve high-precision segmentation in three-dimensional space, capture subtle geometric details, and provide richer information for the understanding and processing of complex scenes. However, the technology also has some shortcomings. First of all, processing 3D point cloud and voxel data requires a lot of computing resources and high-performance hardware, especially in high-resolution and large-scale scenarios, where computing costs and storage requirements are high. Secondly, it is expensive to obtain high-quality 3D data and perform accurate annotation, and the complexity of data annotation increases the difficulty of preparing training data. Meanwhile, 3D scene segmentation algorithms are usually complex with weak real-time processing capabilities, and are difficult to run efficiently on resource-constrained devices, which is a significant bottleneck in applications that require rapid response. In addition, the existing 3D scene segmentation models lack robustness and generalization ability when they meet complex environments and different scenes, and may require additional tuning and training for specific scenes.

Future development trends mainly focus on the following aspects. First, with the continuous advancement of deep learning technology, especially the application of Transformer and GNN, the accuracy and efficiency of 3D scene segmentation will be further improved. These advanced models are better able to handle large-scale and complex 3D data. Secondly, future research will focus more on multi-task learning and self-supervised learning to reduce the dependence on large-scale labeled data, thereby reducing the cost of data labeling and improving the generalization ability and robustness of the model. Third, with the improvement of hardware performance and the optimization of algorithms, it will be possible to achieve efficient real-time 3D scene segmentation on mobile devices and edge devices, which will promote the practical application of 3D scene segmentation technology in autonomous driving, intelligent robots and other fields. Fourthly, the accuracy and reliability of scene segmentation can be improved by fusing different types of 3D sensor data. Multi-sensor fusion technology will become an important direction of future 3D scene segmentation research. In addition, combining 3D scene information with other modalities (such as text, audio, etc.) can enhance the performance of scene segmentation, and cross-modal fusion technology will provide more comprehensive and accurate information support for 3D scene segmentation.

5 Summary and outlook

In this paper, the research status of SLAM, perception and positioning of autonomous driving, scene segmentation and other technologies are introduced respectively. These technologies are interdependent and work together, which constitute the core of modern autonomous driving system. SLAM provides basic positioning and map construction capabilities, scene segmentation provides advanced semantic understanding of the environment, and autonomous driving perception and positioning technology integrates this information for autonomous navigation and decision-making. They are actually closely related and interacted with each other although these technologies belong to different fields on the surface. For example, in the field of autonomous driving, the acquisition of high-precision maps relies on high-precision mapping of SLAM, while higher-level environmental awareness requires scene segmentation and target detection, and real-time positioning also requires SLAM.The future development direction and trend of each technology are prospected when summarizing it. On this basis, development directions applicable to all mentioned technologies will be summarized in the paper, aiming at the common characteristics of all reviewed technologies.

(1) Continuous Application of Deep Learning

Since all the above technologies involve feature extraction and calculation, deep learning has unparalleled advantages in this respect. In the future, deep learning will continue to play a role in various technical fields, integrating more deep learning technologies such as 3D scene reconstruction, 3D target detection, and point cloud completion, which are even expected to completely replace conventional methods in some fields.

(2) Fusion of multi-source and multi-modal information

When the above calculations for three-dimensional data processing or scene perception are applied, multi-source and multi-modal data can provide a more comprehensive and integrated description and understanding of real scenes and objects than single type of data.

(3) High Real-time Performance and Low Computing Load

All data processing technologies will further pursue real-time performance and low computing load to improve processing efficiency and reduce processing costs, so as to promote the real implementation of these technologies in various fields.

The future development of autonomous driving will not only transform transportation technology but will also have profound impacts on law, ethics, and society. In terms of law, the division of responsibility for autonomous vehicle accidents will become a core issue. On the ethical front, the decision-making challenges brought by autonomous driving technology are also a major concern. For example, how should a vehicle make moral judgments when faced with unavoidable accidents (such as the “trolley problem”)? The social impacts are equally important. Autonomous driving could significantly reduce traffic accidents and improve road safety, but it will also disrupt the job market, particularly in the transportation industry. Moreover, as private vehicle ownership declines and shared autonomous vehicle fleets rise, urban planning could be reshaped, changing the way people travel. However, the widespread application of this technology will also raise privacy issues, and how to protect user data will spark ongoing debates in social and policy realms. Overall, the future of autonomous driving is not just about technological breakthroughs but also about comprehensive transformations in law, ethics, and social structures, with the key challenge being how to find a balance in these areas.

Author contributions

WH: Data curation, Formal Analysis, Investigation, Methodology, Supervision, Writing–review and editing, Conceptualization, Project administration, Writing–original draft. WC: Conceptualization, Data curation, Formal Analysis, Investigation, Methodology, Project administration, Supervision, Writing–original draft, Writing–review and editing. ST: Investigation, Validation, Writing–review and editing. LZ: Formal Analysis, Investigation, Writing–review and editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This research was funded in part by the National Natural Science Foundation of China (No. 62306128), the Basic Science Research Project of Jiangsu Provincial Department of Education (No. 23KJD520003), the Leading Innovation Project of Changzhou Science and Technology Bureau (No. CQ20230072), and the Qingpu District Industry University Research Cooperation Development Foundation of Shanghai (No. 202314).

Conflict of interest

Author WC was employed by Shanghai Huace Navigation Technology Co., Ltd. Author LZ was employed by Shanghai Future Space-Time Technology Co., Ltd.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher’s note

References

1. Davison AJ, Reid ID, Molton ND, Stasse O. MonoSLAM: real-time single camera SLAM. *IEEE Trans pattern Anal machine intelligence* (2007) 29(6):1052–67. doi:10.1109/tpami.2007.1049

2. Jones ES, Soatto S. Visual-inertial navigation, mapping and localization: a scalable real-time causal approach. *The Int J Robotics Res* (2011) 30(4):407–30. doi:10.1177/0278364910388963

3. Mourikis AI, Roumeliotis SI. A multi-state constraint Kalman filter for vision-aided inertial navigation. In: *Proceedings 2007 IEEE international conference on robotics and automation*. IEEE (2007). 3565–72.

4. Klein G, Murray D. Parallel tracking and mapping for small AR workspaces. In: *6th IEEE and ACM international symposium on mixed and augmented reality*. IEEE (2007). 225–34.

5. Mur-Artal R, Montiel JMM, Tardos JD. ORB-SLAM: a versatile and accurate monocular SLAM system. *IEEE Trans robotics* (2015) 31(5):1147–63. doi:10.1109/tro.2015.2463671

6. Rublee E, Rabaud V, Konolige K, Bradski G ORB: an efficient alternative to SIFT or SURF. In: *2011 International conference on computer vision*. Barcelona, Spain: IEEE (2011). 2564–71.

7. Mur-Artal R, Tardós JD. Orb-slam2: an open-source slam system for monocular, stereo, and rgb-d cameras. *IEEE Trans robotics* (2017) 33(5):1255–62. doi:10.1109/tro.2017.2705103

8. Leutenegger S, Lynen S, Bosse M, Siegwart R, Furgale P. Keyframe-based visual–inertial odometry using nonlinear optimization. *The Int J Robotics Res* (2015) 34(3):314–34. doi:10.1177/0278364914554813

9. Campos C, Elvira R, Rodríguez JJG, M. Montiel JM, D. Tardos J. Orb-slam3: an accurate open-source library for visual, visual–inertial, and multimap slam. *IEEE Trans Robotics* (2021) 37(6):1874–90. doi:10.1109/tro.2021.3075644

10. Qin T, Li P, Shen S. Vins-mono: a robust and versatile monocular visual-inertial state estimator. *IEEE Trans robotics* (2018) 34(4):1004–20. doi:10.1109/tro.2018.2853729

11. Forster C, Pizzoli M, Scaramuzza D. SVO: fast semi-direct monocular visual odometry. In: *2014 IEEE international conference on robotics and automation (ICRA)*. IEEE (2014). 15–22.

12. Shen S, Mulgaonkar Y, Michael N, Kumar V Initialization-free monocular visual-inertial state estimation with application to autonomous MAVs. In: *Experimental robotics: the 14th international symposium on experimental robotics*. Cham: Springer International Publishing (2015). p. 211–27.

13. Concha A, Loianno G, Kumar V, Civera J *Visual-inertial direct SLAM 2016 IEEE international conference on robotics and automation (ICRA)*. IEEE (2016). 1331–8.

14. Engel J, Schöps T, Cremers D. *LSD-SLAM: large-scale direct monocular SLAMEuropean conference on computer vision*. Cham: Springer International Publishing (2014). 834–49.

15. Engel J, Koltun V, Cremers D. Direct sparse odometry. *IEEE Trans pattern Anal machine intelligence* (2017) 40(3):611–25. doi:10.1109/tpami.2017.2658577

16. Von Stumberg L, Usenko V, Cremers D. Direct sparse visual-inertial odometry using dynamic marginalization. In: *IEEE international conference on robotics and automation (ICRA)*. IEEE (2018). 2510–7.

17. Bowman SL, Atanasov N, Daniilidis K, Pappas GJ *Probabilistic data association for semantic slam 2017 IEEE international conference on robotics and automation (ICRA)*. IEEE (2017). 1722–9.

18. Lianos KN, Schonberger JL, Pollefeys M, Sattler T Vso: visual semantic odometry. In: *Proceedings of the European conference on computer vision*. Munich, Germany: ECCV (2018). 234–50.

19. Yang S, Scherer S. Monocular object and plane slam in structured environments. *IEEE Robotics Automation Lett* (2019) 4(4):3145–52. doi:10.1109/lra.2019.2924848

20. Frost D, Prisacariu V, Murray D. Recovering stable scale in monocular SLAM using object-supplemented bundle adjustment. *IEEE Trans Robotics* (2018) 34(3):736–47. doi:10.1109/tro.2018.2820722

21. Nicholson L, Milford M, Sünderhauf N. Quadricslam: dual quadrics from object detections as landmarks in object-oriented slam. *IEEE Robotics Automation Lett* (2018) 4(1):1–8. doi:10.1109/lra.2018.2866205

22. Lin S, Wang J, Xu M, Zhao H, Chen Z. Topology aware object-level semantic mapping towards more robust loop closure. *IEEE Robotics Automation Lett* (2021) 6(4):7041–8. doi:10.1109/lra.2021.3097242

23. Julier SJ, Uhlmann JK. New extension of the Kalman filter to nonlinear systemsSignal processing, sensor fusion, and target recognition VI. *Spie* (1997) 3068:182–93. doi:10.1117/12.280797

24. Grisetti G, Stachniss C, Burgard W. Improved techniques for grid mapping with rao-blackwellized particle filters. *IEEE Trans Robotics* (2007) 23(1):34–46. doi:10.1109/tro.2006.889486

25. Godsill S. *Particle filtering: the first 25 years and beyondICASSP 2019-2019 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE (2019). 7760–4.

26. Thrun S, Montemerlo M. The graph SLAM algorithm with applications to large-scale mapping of urban structures. *The Int J Robotics Res* (2006) 25(5-6):403–29. doi:10.1177/0278364906065387

27. Besl PJ, McKay ND. Method for registration of 3-D shapesSensor fusion IV: control paradigms and data structures. *Spie* (1992) 1611:586–606. doi:10.1117/12.57955

28. Mendes E, Koch P, Lacroix S. ICP-based pose-graph SLAM. In: *2016 IEEE international symposium on safety, security, and rescue robotics (SSRR)*. IEEE (2016). 195–200. doi:10.15607/RSS.2014.X.007

29. Zhang J, Singh S. LOAM: lidar odometry and mapping in real-timeRobotics. *Sci Syst* (2014) 2(9):1–9.

30. Shan T, Englot B. *Lego-loam: lightweight and ground-optimized lidar odometry and mapping on variable terrain*. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (2018). 4758–65.

31. Deschaud JE. IMLS-SLAM: scan-to-model matching based on 3D data2018. *IEEE Int Conf Robotics Automation (Icra) IEEE* (2018) 2480–5. doi:10.48550/arXiv.1802.08633

32. Behley J, Stachniss C *Efficient surfel-based SLAM using 3D laser range data in urban environmentsRobotics: science and systems*, 2018 (2018). 59.

33. Biber P, Straßer W. The normal distributions transform: a new approach to laser scan matching. In: Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems IROS 2003, 3. IEEE (2003). 2743–8. doi:10.1109/iros.2003.1249285

34. Segal A, Haehnel D, Thrun S. *Generalized-icpRobotics: Sci Syst* (2009) 2(4):435. doi:10.7551/mitpress/8727.003.0022

35. Zhou B, He Y, Qian K, Ma X, Li X. S4-SLAM: a real-time 3D LIDAR SLAM system for ground/watersurface multi-scene outdoor applications. *Autonomous Robots* (2021) 45:77–98. doi:10.1007/s10514-020-09948-3

36. Cohen-OrD ADMNJ. 4-points congruent sets for robust pairwise surface registration. *ACM SIGGRAPH 2008 Pap on-SIGGRAPH* (2008) 8:11–5. doi:10.1145/1399504.1360684

37. Ruan J, Li B, Wang Y, Fang Z *GP-SLAM+: real-time 3D lidar SLAM based on improved regionalized Gaussian process map reconstruction*. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (2020). 5171–8.

38. Dube R, Cramariuc A, Dugas D, Sommer H, Dymczyk M, Nieto J, et al. SegMap: segment-based mapping and localization using data-driven descriptors. *The Int J Robotics Res* (2020) 39(2-3):339–55. doi:10.1177/0278364919863090

39. Zhang J, Singh S. Laser–visual–inertial odometry and mapping with high robustness and low drift. *J field robotics* (2018) 35(8):1242–64. doi:10.1002/rob.21809

40. Shan T, Englot B, Ratti C, Rus D Lvi-sam: tightly-coupled lidar-visual-inertial odometry via smoothing and mapping. In: *IEEE international conference on robotics and automation (ICRA)*. IEEE (2021). 5692–8.

41. Lin J, Zheng C, Xu W, Zhang F. R LIVE: a robust, real-time, LiDAR-inertial-visual tightly-coupled state estimator and mapping. *IEEE Robotics Automation Lett* (2021) 6(4):7469–76. doi:10.1109/lra.2021.3095515

42. Zuo X, Geneva P, Lee W, Liu Y, Huang G Lic-fusion: lidar-inertial-camera odometry. In: *IEEE/RSJ international conference on intelligent robots and systems IROS*. IEEE (2019). 5848–54.

43. Zuo X, Yang Y, Geneva P, Lv J, Liu Y, Huang G, et al. *Lic-fusion 2.0: lidar-inertial-camera odometry with sliding-window plane-feature tracking*. In: IEEE/RSJ International Conference on Intelligent Robots and Systems IROS. IEEE (2020). 5112–9.

44. Cadena C, Carlone L, Carrillo H, Latif Y, Scaramuzza D, Neira J, et al. Past, present, and future of simultaneous localization and mapping: toward the robust-perception age. *IEEE Trans robotics* (2016) 32(6):1309–32. doi:10.1109/tro.2016.2624754

45. Wang W, Wu Y, Jiang Z, Qi J. A clutter-resistant SLAM algorithm for autonomous guided vehicles in dynamic industrial environment. *IEEE Access* (2020) 8:109770–82. doi:10.1109/access.2020.3001756

46. Faessler M, Fontana F, Forster C, Mueggler E, Pizzoli M, Scaramuzza D. Autonomous, vision-based flight and live dense 3D mapping with a quadrotor micro aerial vehicle. *J Field Robotics* (2016) 33(4):431–50. doi:10.1002/rob.21581

47. Cheng J, Zhang L, Chen Q, Hu X, Cai J. A review of visual SLAM methods for autonomous driving vehicles. *Eng Appl Artif Intelligence* (2022) 114:104992. doi:10.1016/j.engappai.2022.104992

48. Zou Z. Application of SLAM technology in VR and AR. *AIP Conf Proc AIP Publishing* (2024) 3144(1):030007. doi:10.1063/5.0215525

49. Wang K, Kooistra L, Pan R, Wang W, Valente J. UAV-based simultaneous localization and mapping in outdoor environments: a systematic scoping review. *J Field Robotics* (2024) 41:1617–42. doi:10.1002/rob.22325

50. DeTone D, Malisiewicz T, Rabinovich A. Superpoint: self-supervised interest point detection and description. In: *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (2018). 224–36.

51. SAR-optical feature matching: A large-scale patch dataset and a deep local descriptor

52. Wang Y, Solomon JM. Deep closest point: learning representations for point cloud registration. In: *Proceedings of the IEEE/CVF international conference on computer vision* (2019). p. 3523–32.

53. Milioto A, Vizzo I, Behley J, Stachniss C Rangenet++: fast and accurate lidar semantic segmentation. In: *IEEE/RSJ international conference on intelligent robots and systems (IROS)*. IEEE (2019). p. 4213–20.

54. Cattaneo D, Vaghi M, Valada A. Lcdnet: deep loop closure detection and point cloud registration for lidar slam. *IEEE Trans Robotics* (2022) 38(4):2074–93. doi:10.1109/tro.2022.3150683

55. Pvn3d: a deep point-wise 3d keypoints voting network for 6dof pose estimation

56. Droid-slam: deep visual slam for monocular, stereo, and rgb-d cameras

57. Lajoie PY, Beltrame G. Swarm-slam: sparse decentralized collaborative simultaneous localization and mapping framework for multi-robot systems. *IEEE Robotics Automation Lett* (2023) 9(1):475–82. doi:10.1109/lra.2023.3333742

58. Kueng B, Mueggler E, Gallego G, Scaramuzza D *Low-latency visual odometry using event-based feature tracks*. IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). IEEE (2016). 16–23.

59. Benosman R, Clercq C, Lagorce X, Sio-Hoi Ieng Bartolozzi C. Event-based visual flow. *IEEE Trans Neural networks Learn Syst* (2013) 25(2):407–17. doi:10.1109/tnnls.2013.2273537

60. Matsuda N, Cossairt O, Gupta M. Mc3d: motion contrast 3d scanning. In: *2015 IEEE international conference on computational photography (ICCP)*. IEEE (2015). p. 1–10.

61. Zhou Y, Gallego G, Shen S. Event-based stereo visual odometry. *IEEE Trans Robotics* (2021) 37(5):1433–50. doi:10.1109/tro.2021.3062252

62. Gallego G, Delbrück T, Orchard G, Bartolozzi C, Taba B, Censi A, et al. Event-based vision: a survey. *IEEE Trans pattern Anal machine intelligence* (2020) 44(1):154–80. doi:10.1109/tpami.2020.3008413

63. Kerbl B, Kopanas G, Leimkühler T, Drettakis G. 3d Gaussian splatting for real-time radiance field rendering. *ACM Trans Graphics* (2023) 42(4):1–14. doi:10.1145/3592433

64. Girshick R. Fast r-cnn. *Proc IEEE Int Conf Comput Vis* (2015) 1440–8. doi:10.48550/arXiv.1504.08083

65. Ren S, He K, Girshick R, Sun J Faster r-cnn: towards real-time object detection with region proposal networks. *Adv Neural Inf Process Syst* (2015) 28. doi:10.48550/arXiv.1506.01497

66. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu C, et al. *Ssd: single shot multibox detector Computer vision–ECCV 2016: 14th European conference, Amsterdam, The Netherlands, october 11–14, 2016, Proceedings, Part I 14*. Springer International Publishing (2016). p. 21–37.

67. Redmon J, Divvala S, Girshick R, Farhadi A You only look once: unified, real-time object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016). p. 779–88.

68. Gragnaniello D, Greco A, Saggese A, Vento M, Vicinanza A. Benchmarking 2D multi-object detection and tracking algorithms in autonomous vehicle driving scenarios. *Sensors* (2023) 23(8):4024. doi:10.3390/s23084024

69. Li S, Fischer T, Ke L, Ding H, Danelljan M, Yu F Ovtrack: open-vocabulary multiple object tracking. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2023). 5567–77.

70. Huang K, Lertniphonphan K, Chen F, Li J, Wang Z Multi-object tracking by self-supervised learning appearance model. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2023). 3163–9.

71. Xia Z, Kim J. Mixed spatial pyramid pooling for semantic segmentation. *Appl Soft Comput* (2020) 91:106209. doi:10.1016/j.asoc.2020.106209

72. Zhang Y, Sun X, Dong J, Chen C, Lv Q. GPNet: gated pyramid network for semantic segmentation. *Pattern Recognition* (2021) 115:107940. doi:10.1016/j.patcog.2021.107940

73. Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P SegFormer: simple and efficient design for semantic segmentation with transformers. *Adv Neural Inf Process Syst* (2021) 34:12077–90. doi:10.48550/arXiv.2105.15203

74. Elhassan MAM, Huang C, Yang C, Munea TL. DSANet: Dilated spatial attention for real-time semantic segmentation in urban street scenes. *Expert Syst Appl* (2021) 183:115090. doi:10.1016/j.eswa.2021.115090

75. Ding L, Goshtasby A. On the Canny edge detector. *Pattern recognition* (2001) 34(3):721–5. doi:10.1016/s0031-3203(00)00023-6

76. Illingworth J, Kittler J. A survey of the Hough transform. *Computer Vis graphics, image Process* (1988) 44(1):87–116. doi:10.1016/s0734-189x(88)80033-1

77. Choi S, Kim T, Yu W. Performance evaluation of RANSAC family. *J Computer Vis* (1997) 24(3):271–300. doi:10.5244/C.23.81

78. Shyam P, Yoon KJ, Kim KS. Weakly supervised approach for joint object and lane marking detection. *Proc IEEE/CVF Int Conf Comput Vis* (2021) 2885–95. doi:10.1109/ICCVW54120.2021.00323

79. Pan X, Shi J, Luo P, Wang X, Tang X. Spatial as deep: spatial cnn for traffic scene understanding. *Proc AAAI Conf Artif intelligence* (2018) 32(1). doi:10.1609/aaai.v32i1.12301

80. Zou Z, Zhang X, Liu H, Li Z, Hussain A, Li J. A novel multimodal fusion network based on a joint-coding model for lane line segmentation. *Inf Fusion* (2022) 80:167–78. doi:10.1016/j.inffus.2021.10.008

81. Qin Z, Zhang P, Li X. Ultra fast deep lane detection with hybrid anchor driven ordinal classification. *IEEE Trans pattern Anal machine intelligence* (2022) 46(5):2555–68. doi:10.1109/tpami.2022.3182097

82. LaneScanNET: A deep-learning approach for simultaneous detection of obstacle-lane states for autonomous driving systems

83. Lee DH, Liu JL. End-to-end deep learning of lane detection and path prediction for real-time autonomous driving. *Signal Image Video Process.* (2023) 17(1):199–205. doi:10.1007/s11760-022-02222-2

84. Wang Y, Du S, Xin Q, He Y, Qian W. Autonomous driving system driven by Artificial intelligence perception fusion. *Acad J Sci Technology* (2024) 9(2):193–8. doi:10.54097/e0b9ak47

85. Zha Y, Shangguan W, Chai L, Chen J. Hierarchical perception Enhancement for different levels of autonomous driving: a review. *IEEE Sensors J* (2024) 24:17366–86. doi:10.1109/jsen.2024.3388503

86. Aung NHH, Sangwongngam P, Jintamethasawat R, Shah S, Wuttisittikulkij L. A review of LiDAR-based 3D object detection via deep learning Approaches towards robust connected and autonomous vehicles. *IEEE Trans Intell Vehicles* (2024) 1–23. doi:10.1109/tiv.2024.3415771

87. Schumann O, Wöhler C, Hahn M, Dickmann J. Comparison of random forest and long short-term memory network performances in classification tasks using radar. In: *2017 sensor data fusion: trends, solutions, applications (SDF)*. IEEE (2017). p. 1–6.

88. Prophet R, Li G, Sturm C, Vossiek M Semantic segmentation on automotive radar maps 2019 IEEE Intelligent Vehicles Symposium (IV). *IEEE* (2019) 756–63. doi:10.1109/IVS.2019.8813808

89. Qi CR, Yi L, Su H, Guibas LJ Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *Adv Neural Inf Process Syst* (2017) 30. doi:10.48550/arXiv.1706.02413

90. Lombacher J, Hahn M, Dickmann J, Wöhler C Object classification in radar using ensemble methods. In: *2017 IEEE MTT-S international conference on Microwaves for intelligent Mobility (ICMIM)*. IEEE (2017). 87–90.

91. Dreher M, Erçelik E, Bänziger T, Knoll A Radar-based 2D car detection using deep neural networks. In: *2020 IEEE 23rd international conference on intelligent transportation systems (ITSC)*. IEEE (2020). p. 1–8.

92. Zhao C, Song A, Zhu Y, Jiang S, Liao F, Du Y. Data-driven indoor positioning correction for infrastructure-enabled autonomous driving systems: a lifelong framework. *IEEE Trans Intell Transportation Syst* (2023) 24(4):3908–21. doi:10.1109/tits.2022.3233563

93. Liu A, Tucker R, Jampani V, Makadia A, Snavely N, Kanazawa A. Infinite nature: Perpetual view generation of natural scenes from a single image. In: *Proceedings of the IEEE-CVF International Conference on computer vision*. Montreal, Canada: ICCV (2021).

94. Unsal D, Demirbas K. Estimation of deterministic and stochastic IMU error parameters. In: *Proceedings of the 2012 IEEE/ION position, location and navigation symposium*. IEEE (2012). 862–8.

95. Dong P, Cheng J, Liu L, Zhang W *Application of improved wavelet de-noising method in MEMS-IMU signals 2019 Chinese Control Conference (CCC)*. IEEE (2019). 3881–4.

96. Radi A, Sheta B, Nassar S, Arafa I, Youssef A, El-Sheimy N Accurate identification and implementation of complicated stochastic error models for low-cost MEMS inertial sensors. In: *2020 12th international conference on Electrical Engineering (ICEENG)*. IEEE (2020). 471–5.

97. Brossard M, Bonnabel S, Barrau A. Denoising imu gyroscopes with deep learning for open-loop attitude estimation. *IEEE Robotics Automation Lett* (2020) 5(3):4796–803. doi:10.48550/arXiv.2002.10718

98. Silva do Monte Lima JP, Uchiyama H, Taniguchi R. End-to-end learning framework for imu-based 6-dof odometry. *Sensors* (2019) 19(17):3777. doi:10.3390/s19173777

99. Ku J, Mozifian M, Lee J, Harakeh A, Waslander SL *Joint 3d proposal generation and object detection from view aggregation 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE (2018). p. 1–8.

100. Wang W, Shangguan W, Liu J, Chen J. Enhanced fault detection for GNSS/INS integration using maximum correntropy filter and local outlier factor. *IEEE Trans Intell Vehicles* (2023) 9:2077–93. doi:10.1109/tiv.2023.3312654

101. Hou P, Zha J, Liu T, Zhang B. Recent advances and perspectives in GNSS PPP-RTK. *Meas Sci Technology* (2023) 34(5):051002. doi:10.1088/1361-6501/acb78c

102. Onyekpe U, Palade V, Kanarachos S, Christopoulos SRG *Learning uncertainties in wheel odometry for vehicular localisation in GNSS deprived environments*. In: 19th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE (2020). 741–6.

103. Onyekpe U, Palade V, Herath A, Kanarachos S, Fitzpatrick ME. WhONet: wheel Odometry neural Network for vehicular localisation in GNSS-deprived environments. *Eng Appl Artif Intelligence* (2021) 105:104421. doi:10.1016/j.engappai.2021.104421

104. Brossard M, Bonnabel S. *Learning wheel odometry and IMU errors for localization2019 international conference on robotics and automation (ICRA)*. IEEE (2019). 291–7.

105. Levinson J, Thrun S. *Robust vehicle localization in urban environments using probabilistic maps2010 IEEE international conference on robotics and automation*. IEEE (2010). p. 4372–8.

106. Xiao Z, Yang D, Wen T, Jiang K, Yan R. Monocular localization with vector HD map (MLVHM): a low-cost method for commercial IVs. *Sensors* (2020) 20(7):1870. doi:10.3390/s20071870

107. Ye X, Song F, Zhang Z, Zeng Q. A review of small UAV navigation system based on multi-source sensor fusion. *IEEE Sensors J* (2023) 23:18926–48. reinforcement learning. doi:10.1109/jsen.2023.3292427

108. Chen S, Hu J, Shi Y, Peng Y, Fang J, Zhao R, et al. Vehicle-to-everything (V2X) services supported by LTE-based systems and 5G. *IEEE Commun Stand Mag* (2017) 1(2):70–6. doi:10.1109/mcomstd.2017.1700015

109. Ma Y, Wang T, Bai X, Yang H, Hou Y, Wang Y, et al. Vision-centric bev perception: a survey. *IEEE Trans Pattern Anal Mach Intell* (2024). doi:10.1109/TPAMI.2024.3449912

110. Kundu A, Yin X, Fathi A, Ross D, Brewington B, Funkhouser T, et al. Virtual multi-view fusion for 3d semantic segmentation. In: *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXIV 16*. Springer International Publishing (2020). 518–35.

111. Lawin FJ, Danelljan M, Tosteberg P, Bhat G, Khan FS, Felsberg M Deep projective 3D semantic segmentation/computer analysis of images and Patterns. In: *17th international conference, CAIP 2017, Ystad, Sweden, August 22-24, 2017, Proceedings, Part I 17*. Springer International Publishing (2017). 95–107.

112. Huang J, Zhang H, Yi L, Funkhouser T, Nießner M, Guibas L Texturenet: Consistent local parametrizations for learning from high-resolution signals on meshes. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019). 4440–9.

113. Tatarchenko M, Park J, Koltun V, Zhou QY Tangent convolutions for dense prediction in 3d. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018). 3887–96.

114. Hu W, Zhao H, Jiang L, Jia J, Wong TT Bidirectional projection network for cross dimension scene understanding. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021). 14373–82.

115. Robert D, Vallet B, Landrieu L. Learning multi-view aggregation in the wild for large-scale 3d semantic segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). p. 5575–84.

116. Graham B. Sparse 3D convolutional neural networks. (2015) 150.1–9. doi:10.5244/c.29.150

117. Engelcke M, Rao D, Wang DZ, Tong CH, Posner I Vote3deep: fast object detection in 3d point clouds using efficient convolutional neural networks. In: *IEEE international conference on robotics and automation (ICRA)*. IEEE (2017). 1355–61.

118. Graham B, Engelcke M, Van Der Maaten L. 3d semantic segmentation with submanifold sparse convolutional networks. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018). 9224–32.

119. Choy C, Gwak JY, Savarese S. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2019). 3075–84.

120. Tang H, Liu Z, Li X, Lin Y, Han S Torchsparse: efficient point cloud inference engine. *Proc Machine Learn Syst* (2022) 4:302–15. doi:10.48550/arXiv.2204.10319

121. Qi CR, Su H, Mo K, Guibas LJ Pointnet: deep learning on point sets for 3d classification and segmentation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017). 652–60.

122. Qi CR, Yi L, Su H, Guibas LJ Pointnet+: deep hierarchical feature learning on point sets in a metric space. *Adv Neural Inf Process Syst* (2017) 30.

123. Thomas H, Qi CR, Deschaud JE, Marcotegui B, Goulette F, Guibas LJ Kpconv: Flexible and deformable convolution for point clouds. In: *Proceedings of the IEEE/CVF international conference on computer vision* (2019). 6411–20.

124. Xu M, Ding R, Zhao H, Qi X Paconv: position adaptive convolution with dynamic kernel assembling on point clouds. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2021). 3173–82.

125. Jiang L, Zhao H, Liu S, Shen X, Fu CW, Jia J Hierarchical point-edge interaction network for point cloud semantic segmentation. *Proc IEEE/CVF Int Conf Computer Vis* (2019) 10433–41. doi:10.1109/ICCV.2019.01053

126. Lei H, Akhtar N, Mian A. Spherical kernel for efficient graph convolution on 3d point clouds. *IEEE Trans pattern Anal machine intelligence* (2020) 43(10):3664–80. doi:10.1109/tpami.2020.2983410

127. Schult J, Engelmann F, Kontogianni T, Leibe B *Dualconvmesh-net: Joint geodesic and euclidean convolutions on 3d meshes. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2020). 8612–22.

128. Hu Z, Bai X, Shang J, Zhang R, Dong J, Wang X, et al. Vmnet: voxel-mesh network for geodesic-aware 3d semantic segmentation. *Proc IEEE/CVF Int Conf Computer Vis* (2021) 15488–98. doi:10.48550/arXiv.2107.13824

129. Zhao H, Jiang L, Jia J, Torr P, Koltun V Point transformer. In: *Proceedings of the IEEE/CVF international conference on computer vision* (2021). 16259–68.

130. Lai X, Liu J, Jiang L, Wang L, Zhao H, Liu S, et al. Stratified transformer for 3d point cloud segmentation. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). 8500–9.

131. Park C, Jeong Y, Cho M, Park J Fast point transformer. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (2022). 16949–58.

132. Wu X, Lao Y, Jiang L, Liu X, Zhao H Point transformer v2: Grouped vector attention and partition-based pooling. *Adv Neural Inf Process Syst* (2022) 35:33330–42. doi:10.48550/arXiv.2210.05666