Multi-modal Data-driven Positioning and Navigation Technology for Autonomous Driving

Xingrui Dong [1,*], Shaozu Han [2], Mingyang Zhang [3]

{* 22721045@bjtu.edu.cn}

[1] School of Telecommunication and Information Engineering, Beijing Jiaotong University, Weihai, 264400, China
[2] School of Electronics Engineering, Dalian Maritime University, Dalian, 116000, China
[3] School of Automation, Nanjing University of Science and Technology, Nanjing, 210094, China

Abstract. Against the backdrop of the rapid development of artificial intelligence and sensor technology, autonomous driving technology is gradually moving from the laboratory to reality. However, due to the complex and uncertain changes of the real environment (extreme weather, signal blockage, dynamic interference, etc.), it brings serious challenges to the location of autonomous driving and autonomous navigation technology. The single-modal perception and decision-making schemes are typically not capable of handling these challenges. Consequently, it has become an inevitable choice to make driving more robust and reliable by utilizing multi-modal information. To this end, this report focuses on reviewing and analyzing the current key technologies for multimodal data-driven unmanned vehicle positioning and navigation. First, by combining the development history of unmanned driving, the report makes clear the background and research significance of this research. Second, some basic fundamental theories, including Kalman filtering, dynamic Bayesian networks, and convolutional neural networks, are introduced. After that, some frontier methods such as MixedFusion, tightly coupled SLAM, MDSTF, DeepInteraction++, RoboTron-Drive, and CCTP-Net were analyzed and their innovation in ideas, technical routes, and advantages and disadvantages in performance were compared from the perspective of positioning and navigation. Finally, it summarizes the problems and progress of this topic research, to provide a practical theoretical basis and analysis framework for future detailed research.

Keywords: Unmanned driving; Multimodal data, Positioning technology, Navigation technology, Sensor fusion.

1 Introduction

The development history of unmanned driving technology has experienced twists and turns. From the early radio-controlled car in the 1920s, the initial exploration based on computer vision in the 1980s, and the technology boost of DARPA in the early 21st century, the technology of unmanned driving gradually became a reality from conception. Waymo 2009 launched by Google, and the wave of commercialization was born. Since 2016, the industry has begun to

ICMEEA 2026, April 24-26, Singapore, Singapore

DOI 10.4108/eai.24-4-2026.2364849

step into a new stage of diversified development and coexistence of development difficulties, with Tesla's FSD as the leader.

However, there must always be problems while technology is developing. For example, in tunnels, urban canyons, and other scenarios, the GPS signal is likely to be lost; in extreme rain, snow, fog weather conditions, the normal operation of sensors such as cameras and lidars may be seriously affected; the moving pedestrians and vehicles on the road will also cause great interference to environmental perception and decision-making. The essence of such practical challenges is: how can the autonomous driving system be able to perform stable, accurate positioning and reliable navigation in any environment. It has been proved that relying on a single sensor modality will be defective, for example, in rainy weather, when only relying on cameras, the recognition accuracy of lane lines may decrease to 41.9% [1], and when only relying on lidar, although relatively less weather interference, the sparseness of point clouds may also cause missing of nearly half of objects. Multi - modal data - driven technology route shows great potential. Integrating the information of multiple sensors, such as camera, lidar, radar, etc., to achieve complementary advantages, will greatly improve its robustness in extreme environments. The multi-modal fusion solution can also still recognize 72.2% of lane lines when there is heavy rain, which also fully reflects its value [1].

The application and development of various multimodal technologies in the positioning and navigation of unmanned vehicles are summarized and summarized in this paper to provide new ideas and a clear technical overview for more reliable and more human-like automatic driving of unmanned vehicles.

2 Basic Theory

Short-term state estimation and prediction are mainly carried out by the Kalman Filter. By integrating the dynamic model of the system with continuous observation information, the Kalman Filter can give the optimal estimation result of the vehicle position, speed, etc. It has the advantages of good real-time and high accuracy, and can smooth the trajectory and compensate for the positioning error caused by the loss of GPS signal for a short time.

Dynamic Bayesian Networks (DBNs) are used for long-term inference and decision-making, suitable for situations where data is incomplete, and sample sizes are small. Dynamic Bayesian Networks are typically used to describe causal relationships between variables in the form of nodes and directed edges, and to describe dependency strength using conditional probabilities. It can reason about uncertainty and is often used to predict long-term trajectory behavior.

Convolutional Neural Network (CNN) is the core of deep learning in the visual perception field. CNN is able to extract hierarchical features (from edges to complex objects) from the original image automatically through the structure, such as convolutional layers and pooling layers. It is also widely used in tasks such as lane line recognition and vehicle detection.

In addition, multimodal SLAM as a core idea, the data level、feature level, decision level, and other levels of fusion strategy to match the data from different sensors in time and space, which is also the most basic paradigm of stable localization and mapping with high precision. Under such a framework, there is a typical research example: Han Team designed and proposed a high-precision positioning and mapping spacecraft feature-assisted pose estimation method based on the variational autoencoder structure in a complex space environment. This method achieves the deep fusion at the feature level by fusing the data from multi-sensor sources like

vision and inertial sensors, effectively solving the robustness problem under illumination changes, feature loss and other problems of SLAM systems, improving the autonomy and reliability of in-orb service spacecraft. In addition, from the perspective of high-quality scene reconstruction in visual SLAM, the real-time and high-realism SLAM system (RP-SLAM) based on efficient 3D Gaussian spraying was proposed by Bai L's team. This system combined the feature-level SLAM and 3D Gaussian mapping, and solved the real-time and fine-quality mapping and positioning problem in scenarios of single camera and RGB-D camera. Under this system, based on feature-level camera pose estimation using ORB-SLAM3, 3D Gaussian elements are adaptively sampled and keyframe-optimized dynamically in the mapping layer to obtain high-quality, low-redundancy, real-time, realistic reconstruction in a complex environment [2][3].

3 In-depth Analysis of Typical Methods

3.1 Multimodal Localization Methods

MixedFusion: 3D Object Data Fusion Framework. To address the limitations of single-sensor perception in adverse weather conditions, MixedFusion proposes a layered fusion strategy. First, in the detection phase, a Transformer is used to fuse camera images and LiDAR point cloud features to generate accurate 3D detection results. Then, in the tracking phase, these results are matched with the radar detection output using multiple priorities to achieve stable trajectory tracking [4]. Its HLSG (High-Level Semantic Guidance) fusion detection and MPM (Multi-Priority Matching) tracking mechanism avoids the noise interference of the radar in the detection stage, and uses the radar's strong penetration in adverse weather conditions. On the K-Radar dataset, its mAP can reach as high as 55.08%, which greatly outperforms the advantages of, especially for extreme weather such as heavy fog and heavy snow.

A multimodal data fusion 3D multi-object detection framework was proposed by Zheng’s team in order to enhance comprehensive and reliable perception ability in uncertain scenarios[5]. Experiments conducted in the Argoverse 2 (AV2) and RADIATE datasets also confirmed the generalisation ability and effectiveness of the proposed method.

Tightly-Coupled Multimodal Localization. Modal tight-coupled positioning mainly solves the problem of accurate positioning in scenarios where GNSS signals fail. By using a tight-coupled optimization method, the global pose constraints of GNSS/INS are fused with the local relative observation depth of LiDAR SLAM to correct the accumulated error of the IMU [6]. In urban areas where GNSS is completely ineffective, this method can effectively suppress IMU drift and reconstruct a smooth trajectory and high-precision point cloud that closely matches the real road shape. Quantitative data show that it can achieve decimeter-level positioning accuracy (such as a horizontal error of 0.329 meters) in complex urban environments, which is significantly better than the traditional loosely coupled scheme.

Wang's team proposed a factor graph optimization-based framework that tightly integrates GNSS double-difference pseudorange and carrier phase observations with inertial, visual, and lidar information, thereby achieving high-precision simultaneous positioning and mapping (SLAM) performance in large-scale environments [7]. They achieved continuous positioning accuracy from centimeter to decimeter levels under harsh GNSS signal conditions.

Multimodal Spatio-Temporal Fusion Trajectory Prediction Model. Multimodal spatio-temporal fusion trajectory prediction model. This model models spatio-temporal interaction

from the full perspective of the agent, and obtains more reasonable trajectory prediction by modeling the spatial relationship of the agent at different locations, the spatial relationship of the agent at the same location in the past, and the spatial relationship of the agent at different locations at the beginning and end. Among them, the full-process spatial modeling is based on an LSTM network that models the change of historical spatial relationships with a Transformer that models long-distance dependence [8]. The gated fusion mechanism is used to dynamically weight the spatial information of each dimension (local L, global G, full-process F) to obtain an integrated spatial expression GLF. We observe that on the ApolloScape dataset, the model performs significantly better on WSADE and WSFDE compared with baseline models, especially in terms of long-term prediction.

The Shi et al. team introduced a multimodal trajectory prediction model that combines the spatio-temporal features extracted by GCN(Graph Convolutional Network) and the temporal attention mechanism. The average Root Mean Square Error values of the model using GCN, GAT (Graph Attention Network), and LSTM decreased by 15.6%, 16.3%, and 23.8% respectively [9].

3.2 Multimodal Localization Methods

DeepInteraction++:The Evolution from Data Fusion to Interactive Modeling. DeepInteraction++ does not directly fuse the multimodal features together, but retains two kinds of dual-stream feature representations of LiDAR and images at all times. So it performs hierarchical deep interaction through a bidirectional cross-attention mechanism [10]. In the interaction, the LiDAR features are semantically enhanced, while the image features can obtain more geometric structure information, which realizes the real 1+1>2 complementarity. It reaches the state-of-the-art results on the nuScenes benchmark (72.0%mAP, 74.4%NDS), which shows the superiority of the interaction paradigm.

RoboTron-Drive. Unified Large Multimodal Model. To overcome the problem of fragmentation of “expert model”, RoboTron-Drive establishes a unified architecture for solving the complete task of the whole chain from perception and prediction to planning, extracts features from the SigLIP visual encoder, and projects them to the semantic space of the language model (Llama-3.1), and builds a unified cognitive architecture through four-stage course learning [11]. This end-to-end integrated model represents the change of paradigm from a tool for one specific field to a general-purpose "cognitive brain". The improved question-answering mechanism and data standardization ensure the high quality and diversity of training data. It also achieves better results than expert models and general large models on multiple benchmark tests, particularly in complex tasks with deep reasoning (Nullstruct), which demonstrates the high potential of the combined model.

The Sun team proposed a sparse scene representation-based end -to-end autonomous driving framework. It uses a symmetric sparse perception module to integrate object detection, tracking, and online mapping, and avoids the computational bottleneck caused by traditional dense BEV features [12]. The team also proposed a parallel motion planner, which can perform motion prediction and path planning simultaneously. By using semantics- and geometry-aware ego-instance initialization and collision-aware rescoring, it can greatly improve interactive understanding and decision-making safety, and the planning collision rate is lowered by 71.4% on the nuScenes benchmark, providing a high-performance and high-efficiency solution for end-to-end autonomous driving.

CCTP-Net. Trajectory Prediction by Integrating Causal Reasoning with Driving Cognition. Aiming at the problem that traditional prediction models do not have causal understanding and lack human-like decision-making characteristics, CCTP-Net performs causal intervention to remove spurious correlations and simulates the speed-field-of-view adaptive feature of human decision-making for key node selection [13], realizing a theoretical do(Traj) operation through deformable cross-attention, and adding cognitive mechanisms to realize decision-making operations in a faster and more reasonable way. In the nuScenes dataset, the proposed CCTP-Net also leads the performance of minADE_5 and MR_5, which proves that causal reasoning effectively enhances the generalization ability of the model.

Cheng et al. proposed a Multimodal Causal Analysis Model that used a feature decoupling approach to construct direct causal connections, indirect causal connections, and confounding causal connections between scene features and behavioral features [14]. The model built a latent causal model between visual and linguistic modalities and suppressed false information.

4 Challenges and Future Directions

However, there are still some challenges to be solved for autonomous vehicle localization and navigation multi modal data - driven technologies. For the localization, the sensor calibration are still sensitive to the changes of the environment and the cumulative drift of Inertial Measurement Units (IMUs) is not fully eliminated. In addition, the complex algorithms cannot achieve the real - time demand in practical deployment. For navigation, deep learning models usually require high computational and memory resource to train and operate. Data-driven approaches may learn spurious correlations from data rather than learning the underlying principles . Integrated end-to-end models are costly to deploy, while causal models are often built based on strong prior assumptions.

Future research is probably going to concentrate on the following main directions. (i) Improving system trustworthiness. There is a growing trend from correlation-based learning to modeling with causal relationships. This is to say, researchers want to design frameworks that are aware of and can intervene on key causal variables, to make decisions more transparent and verifiable. Second, to bypass the shortcomings of the current modularity, a natively embedded end-to-end architecture is being explored. For example, general driving models are built on world models, and holistic optimization and decision-making are pursued. Third, a trade-off between performance and cost. We are shifting away from expensive sensors to a ``light sensing, heavy reasoning'' approach, using prior knowledge, commonsense reasoning, and high-fidelity simulation, to enable lightweight yet highly cognitive algorithms. Making progress in these directions should systematically enhance the robustness, generalization, and utility of intelligent systems in complex real-world settings.

5 Conclusion

In this paper, some key technologies of multi-modal data-driven localization and navigation in autonomous driving have been investigated. By studying the technology development, application status and key issues, multi-modal fusion is a key direction to improve the ability of the vehicle to adapt to the environment. In the future, it may focus on the transition from shallow to deep fusion, from only data fitting to uncovering the causes, and from modular to integrated,

which will ultimately serve the long-term development of more intelligent, reliable, and practical autonomous driving systems.

Authors Contribution

All the authors contributed equally and their names were listed in alphabetical order.

References

[1] K. Park, Y. Kim, D. Kim, and J. W. Choi, "Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2025, pp. 6720–6729.

[2] H. Shen-Tu, Z. Bai, Y. Chen, Y. Wei, J. Zhao, and Z. Cao, "A GM-PHD-SLAM algorithm based on pose and map alternating update," *Chin. J. Aeronaut.*, vol. 38, no. 11, Art. no. 103739, 2025.

[3] L. Bai, C. Tian, J. Yang, S. Zhang, M. Suganuma, and T. Okatani, "RP-SLAM: Real-time photorealistic SLAM with efficient 3D Gaussian splatting," *IEEE Trans. Vis. Comput. Graph.*, vol. 32, no. 2, pp. 1452–1466, Feb. 2026, doi: 10.1109/TVCG.2025.3616173.

[4] C. Zhang, H. Wang, L. Chen, Y. Li, and Y. Cai, "MixedFusion: An efficient multimodal data fusion framework for 3-D object detection and tracking," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 36, no. 1, pp. 1842–1856, Jan. 2025.

[5] Z. Shao, H. Wang, Y. Cai, L. Chen, and Y. Li, "UA-Fusion: Uncertainty-aware multimodal data fusion framework for 3-D object detection of autonomous vehicles," *IEEE Trans. Instrum. Meas.*, vol. 74, pp. 1–16, 2025, doi: 10.1109/TIM.2025.3548184.

[6] S. Lu, S. Bao, W. Shi, et al., "Multimodal sensor dataset from vehicle-mounted mobile mapping system for comprehensive urban scenes," *Sci. Data*, vol. 12, Art. no. 1411, 2025, doi: 10.1038/s41597-025-05471-1.

[7] X. Wang, X. Li, H. Yu, H. Chang, Y. Zhou, and S. Li, "GIVL-SLAM: A robust and high-precision SLAM system by tightly coupled GNSS RTK, inertial, vision, and LiDAR," *IEEE/ASME Trans. Mechatronics*, vol. 30, no. 2, pp. 1212–1223, 2025.

[8] X. Wang, Z. Wu, B. Jin, et al., "MDSTF: A multi-dimensional spatio-temporal feature fusion trajectory prediction model for autonomous driving," *Complex Intell. Syst.*, vol. 10, no. 5, pp. 6647–6665, Oct. 2024, doi: 10.1007/s40747-024-01490-4.

[9] X. Shi, H. Wang, Y. Ji, et al., "Multimodal vehicle trajectory prediction method fusing spatio-temporal features," *Comput. Eng. Appl.*, vol. 61, no. 7, pp. 325–333, 2025.

[10] Z. Yang, N. Song, W. Li, X. Zhu, L. Zhang, and P. H. S. Torr, "DeepInteraction++: Multi-modality interaction for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 8, pp. 6749–6763, Aug. 2025, doi: 10.1109/TPAMI.2025.3565194.

[11] Z. Huang et al., "RoboTron-Drive: All-in-one large multimodal model for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2025, pp. 8011–8021.

[12] W. Sun, X. Lin, Y. Shi, C. Zhang, H. Wu, and S. Zheng, "SparseDrive: End-to-end autonomous driving via sparse scene representation," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2025, pp. 8795–8801, doi: 10.1109/ICRA55743.2025.11128800.

[13] Z. Yang, J. Yang, Y. Zhou, et al., "Multimodal trajectory prediction for intelligent connected vehicles in complex road scenarios based on causal reasoning and driving cognition characteristics," *Sci. Rep.*, vol. 15, Art. no. 7259, 2025, doi: 10.1038/s41598-025-91818-y.

[14] T. Cheng, R. Li, Y. Xiong, T. Zhang, J. Wang, and K. Liu, "MCAM: Multimodal causal analysis model for ego-vehicle-level driving video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2025, pp. 5479–5489.