From Perception to End-To-End: A Comprehensive Exploration of Intelligent Driving Technology, Scenarios, and Industrialization Challenges

Zhekai Chen *

Department of Mathematics and Statistics, Skidmore College, Saratoga Springs, United States

* Corresponding Author Email: jackychen@skidmore.edu

Abstract. Intelligent driving technology is evolving from traditional modular architectures toward end-to-end learning and shifting from 3D detection-based perception solutions to occupancy prediction and world models. This paper surveys the current mainstream intelligent driving technology solutions, including pure visual BEV (Bird's-Eye-View) solutions, radar-vision fusion solutions, and LiDAR-dominated solutions, while analyzing the advantages and disadvantages of each solution in L2/L3-level intelligent driving on urban roads. Research indicates that under the trade-off between cost and stability, different scenarios require distinct technical routes: radar-vision fusion solutions are suitable for low-speed urban scenarios, pure visual solutions can be applied in highway scenarios, and multi-sensor redundancy is necessary under adverse weather conditions. Future intelligent driving will move toward a more integrated and end-to-end direction, with occupancy prediction and world models emerging as key technologies.

Keywords: Intelligent driving; BEV perception; sensor fusion; end-to-end learning; occupancy prediction.

1. Introduction

1.1. Development History and Technical Limitations of Intelligent Driving

The development of intelligent driving technology has undergone three critical paradigm shifts. The first is the transition from traditional modular architectures to end-to-end learning. In their landmark survey, Chen et al. pointed out that traditional autonomous driving systems adopt a serial architecture of perception-prediction-planning-control, where each module is optimized independently, leading to issues of information transmission loss and error accumulation [1]. The end-to-end approach, by contrast, leverages a unified neural network to directly output control commands from sensor data, enabling global optimization [2].

The second shift is from 3D object detection to occupancy prediction and world models. Traditional methods rely on bounding boxes for representation, which fail to capture complex geometric shapes and spatial relationships. Occupancy prediction provides voxel-level spatial understanding, making it more suitable for refined planning and decision-making in dense traffic environments.

The third shift is from perception-centric design to planning-oriented design. The planning-oriented autonomous driving framework proposed by Hu et al. emphasizes that perception and prediction modules should be specifically designed for planning tasks, rather than pursuing general perception accuracy metrics. This design philosophy is fully reflected in unified architectures such as UniAD [3].

In L2/L3-level intelligent driving on urban roads, systems encounter numerous technical challenges: the diverse behaviors of drivers, occlusion and lighting changes, irregular road signs, dynamic construction zones, and adverse weather conditions. These challenges require systems to balance perception accuracy and environmental robustness under cost constraints.

1.2. Research Questions and Requirement Scenarios

A fundamental contradiction exists in current intelligent driving systems between cost reduction and stability assurance. Cost-sensitive mass-produced vehicles tend to adopt pure visual solutions, but their performance is unstable in complex scenarios such as low-speed mixed traffic in urban areas, rainy/hazy weather, and nighttime conditions. Muhammad et al. analyzed the safety challenges of deep learning in autonomous driving, noting that improving system robustness under limited computing resources and cost constraints is a key technical issue [4].

This study focuses on enhancing the performance of intelligent driving on urban roads under limited costs, exploring the applicability of different sensor configurations and algorithm architectures in typical application scenarios, and providing a basis for technical selection in engineering practice.

1.3. Technical Routes and Main Contributions

This paper focuses on analyzing two mainstream technical routes: pure visual BEV perception solutions and multi-sensor fusion solutions. Pure visual solutions, represented by BEVFormer, adopt spatiotemporal transformers to realize feature learning from multi-camera inputs to BEV representations. Radar-vision fusion solutions, represented by BEVFusion and CenterFusion, achieve multi-modal information fusion in a unified BEV space. By comparing and analyzing the technical principles, performance metrics, and applicable scenarios of each solution, this paper provides targeted recommendations for different application requirements.

2. Case Analysis of Technical Solutions

2.1. Pure Visual BEV Solutions

2.1.1 BEVFormer & its performance, advantages, limitations

As a representative method for pure visual BEV perception, BEVFormer addresses key issues in multi-camera feature fusion and temporal information modeling [5]. This method uses learnable BEV queries to aggregate features from different cameras through a spatial cross-attention mechanism, and propagates historical BEV features via temporal self-attention.

Specifically, spatial cross-attention associates BEV queries with multi-view image features through pre-defined reference points, enabling cross-camera information aggregation. Temporal self-attention propagates historical information in the BEV space, enhancing the ability to track moving objects. This design cleverly avoids the inaccurate depth prediction problem inherent in traditional LSS (Lift-Splat-Shoot) methods.

On the nuScenes dataset, BEVFormer achieved the state-of-the-art nuScenes Detection Score (NDS) metric for 3D detection tasks. In BEV map segmentation tasks, its mean Intersection over Union (mIoU) metric also significantly outperformed baseline methods. Particularly in terms of object velocity estimation, BEVFormer showed a noticeable improvement over traditional methods—a critical advantage for subsequent motion planning.

In terms of computational efficiency, BEVFormer achieves real-time inference through a parallelized design, reaching a processing speed of approximately 15 frames per second (FPS) on a single V100 GPU, which meets the requirements of practical deployment.

The main advantages of pure visual BEV solutions lie in their low cost and ease of mass production deployment. Cameras are far less expensive than LiDAR and possess strong capabilities for acquiring rich semantic information. However, they are highly sensitive to lighting and weather conditions, with performance declining significantly under nighttime, backlighting, rainy, or foggy conditions.

In terms of occupancy prediction, the SurroundOcc method proposed by Wei et al. demonstrates the ability of multi-camera-based 3D occupancy prediction, providing more refined spatial representation for closed-loop planning and risk assessment. Compared with traditional 3D detection, occupancy prediction better handles objects with irregular shapes and models the free space [6].

2.2. Radar-Vision Fusion Solutions

2.2.1 Multi-modal fusion in CenterFusion

The CenterFusion method exhibits significant stability advantages under rainy, foggy, and low-light conditions [7]. This method adopts a center-based detection framework, projecting radar point clouds onto the image plane and integrating complementary information through feature-level fusion. The radial velocity observation capability of radar effectively compensates for the shortcomings of visual sensors in depth and velocity estimation.

In adverse weather tests, CenterFusion showed significant improvements over pure visual methods in both objects recall rate and velocity estimation accuracy. Particularly in rainy and nighttime scenarios, the all-weather perception capability of radar ensures stable system operation.

2.2.2 Unified spatial fusion in BEVFusion

BEVFusion further advances the multi-sensor fusion architecture based on a unified BEV space [8]. This method transforms both camera and LiDAR features into a shared BEV representation space, preserving the semantic density of cameras and the geometric accuracy of LiDAR.

Its key technical innovation is an efficient camera-to-BEV transformation core. Through precomputation and interval reduction optimization, the computational complexity of traditional view transformation is reduced by approximately 40 times, making unified BEV fusion feasible with low latency. On the nuScenes dataset, BEVFusion achieved state-of-the-art performance in both 3D detection and BEV segmentation tasks.

2.2.3 Performance metrics and robustness evaluation

Across multiple evaluation metrics, radar-vision fusion solutions perform excellently in terms of velocity estimation error and robustness under adverse weather. The radial velocity observation capability makes this solution more secure in high-speed scenarios. However, the radar false alarm issue and additional sensor costs need to be weighed in practical deployment.

2.3. LiDAR-Dominated Solutions

LiDAR has inherent advantages in geometric accuracy and 3D structure understanding. Point cloud data is efficiently processed through methods such as voxelization and sparse convolution to achieve accurate 3D object detection and semantic segmentation. Compared with visual solutions, LiDAR is more reliable in-depth measurement and geometric modeling.

LiDAR demonstrates unique value in localization and odometry. LiDAR-inertial odometry methods such as FAST-LIO2 have been widely applied in mobile platforms such as vehicles, UAVs (Unmanned Aerial Vehicles), and AGVs (Automated Guided Vehicles), providing high-precision pose estimation and map construction capabilities [9].

Although the DROID-SLAM method proposed by Teed and Deng is primarily designed for visual SLAM (Simultaneous Localization and Mapping), its excellent performance on monocular, stereo, and RGB-D cameras has inspired the development of multi-sensor SLAM [10].

The main limitations of LiDAR include high sensor costs, complex installation and calibration requirements, and sensitivity to environmental pollution. In rainy or snowy weather, surface contamination of LiDAR can significantly affect ranging accuracy. Additionally, LiDAR also has limitations in long-distance detection capability and occlusion handling.

2.4. End-to-End and Planning-Oriented Solutions

End-to-end autonomous driving learns to directly output waypoints or control signals from sensor data through behavioral cloning, enabling joint optimization of perception-prediction-planning [2]. This approach avoids information loss and error accumulation in traditional modular architectures but imposes extremely high requirements on the quality and diversity of training data.

World models and occupancy-driven planning represent the latest development direction of end-to-end methods. By learning the dynamic evolution rules of the environment, they provide a more reliable predictive basis for planning and decision-making.

Evaluating end-to-end systems faces challenges in open-loop versus closed-loop assessment. Traditional open-loop metrics such as mean Average Precision (mAP) and NDS fail to reflect closed-loop driving performance. Simulation platforms such as CARLA (Car Learning to Act) and nuPlan provide closed-loop evaluation environments, comprehensively assessing system performance through metrics such as Driving Score, collision rate, and comfort.

End-to-end methods encounter significant challenges in interpretability, safety redundancy, and out-of-distribution generalization. The black-box nature of the decision-making process makes it difficult to meet the safety standards of the automotive industry. Causal confusion issues may lead to unexpected behaviors of the system in new scenarios, which is a core technical challenge for current end-to-end methods.

3. Technical Analysis and Comparison

3.1. In-Depth Analysis of Perception-Layer Technologies

3.1.1 Technical Characteristics of Visible-Light Cameras

Camera sensors acquire rich texture and semantic information based on optical imaging principles. Their information dimensions include color, texture, and edges, providing a natural advantage for semantic understanding. However, their depth observability is limited— especially in monocular configurations, where depth estimation relies on motion parallax or learned priors, resulting in limited accuracy.

Camera performance fluctuates significantly under varying lighting conditions. Low illumination at night, glare from backlighting, and reflections in rainy weather can all severely degrade image quality. In contrast, cameras have a distinct cost advantage: the cost of a single camera sensor is typically several hundred RMB, far lower than that of other sensors [5].

3.1.2 Advantages and limitations of LiDAR technology

By actively emitting laser pulses to measure distances, LiDAR achieves extremely high geometric accuracy, typically at the centimeter level. In SLAM and odometry applications, the precise geometric information provided by LiDAR is irreplaceable by other sensors. Methods such as FAST-LIO2 demonstrate LiDAR's excellent performance in real-time localization and mapping [9].

However, LiDAR is susceptible to contamination from rain and snow, requiring regular cleaning and calibration maintenance. Sensor costs remain high: mechanical LiDAR typically costs tens of thousands of RMB, and while solid-state LiDAR are cheaper, they are still far more expensive than cameras. The complexity of installation and calibration also increases the difficulty of system integration.

3.1.3 Complementary advantages of millimeter-wave radar

The core advantages of millimeter-wave radar lie in its direct radial velocity observation capability and all-weather operation. Velocity observation is critical for tracking moving objects and assessing collision risks. The fusion of radar and cameras can significantly improve the observability and stability of systems under adverse weather, as demonstrated by the CenterFusion method [7].

3.2. Comparison of Underlying Representation Frameworks

3.2.1 Detection metrics vs. occupancy representation

Traditional 3D detection metrics such as mAP and NDS focus on object-level detection accuracy but fail to fully reflect spatial occupancy relationships. Metrics such as BEV mIoU and Occupancy IoU are more suitable for evaluating planning-related spatial understanding capabilities. Occupancy prediction provides more refined collision detection and free space modeling capabilities, but it also significantly increases computational overhead and annotation costs.

3.2.2 Trade-off between end-to-end and modular architectures

End-to-end architectures avoid error propagation between modules through joint optimization, theoretically achieving better overall performance. However, modular architectures still have obvious advantages in interpretability and debugging convenience. In safety-critical automotive applications, the interpretability and fault localization capabilities of systems are crucial.

Closed-loop stability is another challenge faced by end-to-end systems. Chen et al. noted that end-to-end systems may exhibit unstable behaviors in complex interactive scenarios, which need to be mitigated through large-scale data training and careful safety verification [2].

3.3. Computing Platforms and Acceleration Optimization

Modern intelligent driving systems primarily rely on GPUs and dedicated AI accelerators for real-time inference. View transformation, sparse computing, and tensor parallelism are the main computational bottlenecks and optimization priorities. Through the design of an efficient pooling core, BEVFusion reduces the computational complexity of view transformation by 40 times, demonstrating the great potential of algorithm optimization [8].

3.4. Datasets and Evaluation Systems

Open-source datasets such as nuScenes, Waymo, and Occ3D provide standardized evaluation benchmarks for algorithm development. These datasets cover different geographic regions, weather conditions, and traffic scenarios, but limitations remain in terms of data distribution and annotation quality.

Simulation platforms such as CARLA and nuPlan provide closed-loop evaluation environments, but the gap between simulation and reality remains significant. Metrics such as scenario coverage, interaction conflict handling, comfort, and traffic rule compliance require further verification in real-vehicle tests.

3.5. Risk Assessment and Safety Standards

Typical risks encountered in real-world driving include glare interference, water reflection, nighttime pedestrian detection, intersection visibility occlusion, and temporary construction signs. These scenarios pose severe challenges to sensor performance and algorithm robustness.

The functional safety standard for the automotive industry (ISO 26262) imposes strict requirements on the Safety Integrity Level (SIL) of intelligent driving systems. Systems need to have fault detection, degradation modes, and safety redundancy capabilities, which pose significant challenges to the engineering deployment of end-to-end methods.

4. Technical Recommendations for Application Scenarios

4.1. Low-Speed Complex Urban Scenarios

For low-speed, congested, and traffic-dense urban road environments, the radar + vision unified BEV + occupancy-driven planning technical solution is practical. Urban environments feature diverse and complex behaviors of traffic participants, with frequent occlusion. The velocity observation capability of radar provides critical motion information, compensating for the shortcomings of visual sensors in depth estimation. Unified BEV representation enables effective fusion of multi-modal information, while occupancy-driven planning provides refined spatial understanding.

To address radar false alarms, it is necessary to develop false alarm filtering algorithms based on multi-frame fusion. For the nighttime performance issues of cameras, large-aperture lenses, HDR imaging, and active noise reduction technologies are able to solve these performance issues.

Additionally, a redundancy mechanism for cross-validation among multiple sensors should be established.

4.2. Structured Highway Scenarios

For highly structured scenarios such as highways, it is practical to adopt the pure visual BEV as the main component + long-range radar as auxiliary + lightweight HD map solution.

Highway scenarios are relatively simple, with clear lane markings and regular traffic flow. Pure visual solutions can effectively control costs, meeting the needs of large-scale mass production. Long-range radar is mainly used for accurate distance measurement and velocity observation of vehicles ahead, while lightweight HD maps provide lane-level navigation information. To address challenges such as backlight glare and rainy/snowy weather, it is necessary to improve depth estimation accuracy through self-supervised deep learning methods, adopt active windshield wiper and defogging strategies to enhance visibility, and establish a degraded control mode based on vehicle dynamics models.

4.3. Adverse Weather and Nighttime Scenarios

Under adverse weather conditions (rain, snow, fog) and low-light nighttime environments, the optimal solution is to adopt a multi-redundancy configuration that uses radar as the primary component, vision as an auxiliary component, and short-range LiDAR as a supplementary component.

The all-weather advantage of radar makes it the primary perception tool, while visual sensors provide supplementary semantic information, and short-range LiDAR is used for precise short-distance obstacle avoidance. Although this configuration has higher costs, it ensures driving safety under harsh conditions. Using this solution requires focusing on monitoring object loss rate, sensor degradation detection latency, and the smoothness of degraded mode switching. Establish a dynamic weight assignment mechanism based on sensor confidence.

4.4. Large-Scale Low-Cost Deployment

For cost-sensitive large-scale mass production applications, the lightweight solution of pure visual occupancy prediction and deep learning model compression is applicable.

To implement this solution, it is necessary to achieve algorithm lightweight through techniques such as model distillation, quantization and pruning; reduce deployment costs through automatic annotation and online extrinsic parameter calibration; and establish a comprehensive algorithm and model monitoring system to ensure long-term stable operation of the system. This solution has to focus on sensor calibration drift, model performance degradation, and edge case detection, and further establish a continuous learning and model update mechanism based on cloud big data.

5. Evolution Trends and Opportunities

5.1. Prediction of Technical Evolution Trends

Future intelligent driving technology will move toward greater integration and intelligence. Occupancy prediction and world models will serve as key bridges connecting perception and planning, providing more refined environmental understanding and more reliable behavior prediction. Multi-modal fusion will achieve more efficient information integration in a unified BEV space, and the complementary advantages among sensors will be fully leveraged.

In terms of algorithm architecture, end-to-end learning will gradually address interpretability and safety issues, enhancing system transparency through technologies such as causal reasoning and attention visualization. The development of model compression and edge computing technologies will enable complex algorithms to be deployed on resource-constrained in-vehicle platforms.

5.2. Industrialization Challenges and Opportunities

The balance between cost control and performance assurance remains a core challenge for industrialization. With advances in sensor technology and the expansion of mass production scales, hardware costs will continue to decline. Algorithm optimization and chip specialization will further reduce computational costs.

Safety regulations and standard formulation will drive technical convergence in the industry. Functional safety, information security, and expected functional safety will become important constraints for technological development. Data privacy protection and algorithm fairness will also influence the selection of technical routes.

6. Conclusions

The development of intelligent driving technology shows a diversified trend, with different technical solutions having unique characteristics. Pure visual solutions dominate mass-produced vehicles due to their cost advantages and perform well in structured scenarios such as highways, but they have obvious limitations in complex urban environments and adverse weather. Radar-vision fusion solutions significantly improve system robustness through multi-modal information complementarity, especially demonstrating unique advantages in velocity observation and all-weather perception— though additional sensor costs and system complexity need to be weighed.

While LiDAR solutions are irreplaceable in geometric accuracy and SLAM applications, their high costs and maintenance complexity limit their popularity in consumer-grade products. End-to-end methods represent the future development direction but still face significant challenges in interpretability, safety redundancy, and engineering deployment.

Future research will focus on several key areas: first, interpretable AI technology based on neuro-symbolic reasoning to achieve transparent decision-making in end-to-end systems; second, multi-agent collaborative perception and planning to achieve beyond-line-of-sight perception capabilities through Vehicle-to-Everything (V2X) communication; third, continuous learning and domain adaptation technologies to improve the generalization ability of systems in new scenarios; and fourth, simulation testing technology based on digital twins to accelerate algorithm verification and safety assessment.

As technology maturity continues to improve and costs keep declining, intelligent driving will rapidly penetrate from high-end vehicles to ordinary consumer vehicles, ultimately realizing a fundamental transformation in the way people travel across society. This process not only requires the drive of technological innovation but also demands collaborative efforts across all links of the industrial chain and supporting policies and regulations.

References

[1] Chen, L., Li, Y., Huang, C., Li, B., Zhang, B., Nie, L., Zhang, A., Xue, Z., & Wang, F. Y. (2022). Milestones in autonomous driving and intelligent vehicles: Survey of surveys. IEEE Transactions on Intelligent Vehicles, 8(2), 1046-1056.

[2] Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., & Li, H. (2024). End-to-End autonomous driving: Challenges and frontiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 10164-10183.

[3] Hu, Y., Yang, J., Chen, L., Li, K., Sima, C., Zhu, X., Chai, S., Du, S., Lin, T., Wang, W., Lu, L., Jia, X., Liu, P., Wang, D., Chen, Y., Wang, H., & Li, H. (2023). Planning-oriented autonomous driving. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 17853-17862).

[4] Muhammad, K., Ullah, A., Lloret, J., Del Ser, J., & de Albuquerque, V. H. C. (2020). Deep learning for safe autonomous driving: Current challenges and future directions. IEEE Transactions on Intelligent Transportation Systems, 22(7), 4316-4336.

[5] Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., & Dai, J. (2022). BEVFormer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 47(3), 2020-2036.

[6] Wei, Y., Zhao, L., Zheng, W., Zhu, Z., Zhou, J., & Lu, J. (2023). SurroundOcc: Multi-camera 3D occupancy prediction for autonomous driving. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 21729-21740).

[7] Nabati, R., & Qi, H. (2021). CenterFusion: Center-based radar and camera fusion for 3D object detection. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 1527-1536).

[8] Liu, Z., Tang, H., Amini, A., Yang, X., Mao, H., Rus, D., & Han, S. (2022). BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. arXiv preprint arXiv:2205.13542. https://arxiv.org/abs/2205.13542

[9] Xu, W., Cai, Y., He, D., Lin, J., & Zhang, F. (2022). FAST-LIO2: Fast direct LiDAR-inertial odometry. IEEE Transactions on Robotics, 38(4), 2053-2073.

[10] Teed, Z., & Deng, J. (2021). DROID-SLAM: Deep visual SLAM for monocular, stereo, and RGB-D cameras. Advances in Neural Information Processing Systems, 34, 16558-16569.