Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33656

Architectural Evolution and Comparative Analysis of Vision-Centric BEV for End-to-End Urban Autonomous Driving

Yizhuo Wei

School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an, China
1229091669@stu.xjtu.edu.cn

Abstract. Complex urban scenarios remain the core bottleneck for the large-scale commercial deployment of high-level autonomous driving. Vision-centric Bird's Eye View (BEV)-driven end-to-end architectures have become the prevailing technical paradigms to address the inherent drawbacks of conventional modular autonomous driving solutions. Existing studies mainly focus on BEV-based 3D perception tasks, while lacking systematic comparative analysis of the end-to-end perception-decision closed loop. Based on representative works from top-tier conferences and journals (2020–2025), this paper sorts out the evolutionary path of vision-centric BEV technology, decomposes core technical modules of end-to-end architectures, conducts horizontal comparison of typical benchmark schemes in urban scenarios, and analyzes critical deployment bottlenecks. This work provides a systematic reference for academic research and mass production deployment of end-to-end urban autonomous driving.

Keywords: Vision-centric bird's eye view, end-to-end autonomous driving, spatiotemporal feature fusion, perception-planning closed-loop

1. Introduction

Urban autonomous driving functions have been widely deployed by mainstream automakers. However, the reliability of autonomous systems in complex urban scenarios (e.g., congested car-following, pedestrian crossing, intersection negotiation) remains the core bottleneck for large-scale commercialization of high-level autonomous driving. Unlike highway scenarios with stable traffic flow and standardized road structures, urban scenarios feature complex road topology, mixed traffic of vehicles and vulnerable road users, and diverse traffic rules. These features pose stringent requirements on the generalization, real-time performance and robustness of autonomous driving systems [1].

For decades, mainstream autonomous driving systems have adopted a modular architecture. It controls vehicles through cascaded modules including perception, localization, prediction, planning and control [2]. While widely applied in highway scenarios, this architecture suffers from inevitable error propagation from front-end perception to downstream planning and control modules in complex urban environments [3]. Also, such a modular architecture exhibits poor inter-module coordination. And it is difficult to upgrade and iterate as a whole. To address these drawbacks, end-to-end autonomous driving architectures, which directly output vehicle control commands from raw

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33656

sensor inputs and support joint optimization, have attracted extensive attention from academia and industry [2].

Vision-centric BEV representation has become the core foundation of end-to-end autonomous driving architectures. Compared with front-view perspective, BEV unifies the spatial coordinate system of multi-camera sensors. It is naturally aligned with the planning space of autonomous driving, effectively mitigating occlusion and scale variation issues in front-view perception [4]. Meanwhile, vision-based solutions have significant cost advantages over LiDAR-based counterparts. Cost is critical for large-scale commercial deployment. However, existing studies mainly focus on BEV-based 3D perception tasks, lacking systematic comparative analysis of the end-to-end perception-decision closed loop and targeted analysis for urban driving scenarios.

The core contributions of this paper are three-fold: First, it sorts out the three-stage evolutionary path of vision-centric BEV technology. Then, it decomposes the core technical modules of BEV-driven end-to-end architectures, and conducts horizontal comparative analysis of typical benchmark schemes in urban scenarios. Ultimately, it analyzes the critical bottlenecks for mass production deployment, and proposes actionable optimization directions for urban autonomous driving. The remainder of this paper is structured in the following sections: Section 2 introduces the core definition and evolutionary path of vision-centric BEV, Section 3 decomposes its core technical framework, Section 4 conducts comparative analysis of typical closed-loop architectures, Section 5 analyzes deployment challenges, and Section 6 concludes the paper.

2. Core definition and architectural evolution of vision-centric BEV

2.1. Definition of vision-centric BEV representation

Vision-centric BEV representation takes multi-view on-board camera images as the core input. Through feature encoding, implicit 3D spatial projection and view transformation, it maps 2D perspective image features into a unified BEV Cartesian coordinate system, and generates a global semantic feature representation that models the 3D surrounding environment of the ego-vehicle [4, 5]. Different from conventional single-view perception, this representation eliminates inherent spatial constraints of front-view detection, and unifies multi-camera coordinate systems to avoid occlusion and scale variation issues. Compared with LiDAR-based BEV, it achieves comparable scene modeling capability with significant cost advantages. Serving as the core feature backbone for end-to-end urban autonomous driving, it provides native support for a full spectrum of tasks including perception, prediction and planning. And it serves as the key paradigm connecting multi-sensor inputs to vehicle decision outputs.

2.2. Evolutionary trajectory of vision-centric BEV technology

The evolution of BEV technology falls into three well-defined eras: the foundational period, the explosive growth phase, and the mature stage. The core of BEV technology lies in spatial transformation from 2D to 3D BEV space. Two mainstream technical paradigms are adopted for this task. Geometry-based schemes achieve low accuracy yet high inference speed, whereas learning-based approaches possess superior generalization capability with higher computational overhead. Such trade-offs constitute the central dilemma in contemporary BEV view transformation research.

The first is the foundational stage (pre-2022), with the core breakthrough of geometry-based view transformation. The representative work LSS first realizes end-to-end trainable pure vision BEV transformation through implicit depth estimation [5]. It lays the core foundation for subsequent

DOI: 10.54254/2755-2721/2026.GU33656

frameworks. The second is the explosive growth stage (2022-2023). Learning-based spatiotemporal fusion is the main progress. The representative work BEVFormer adopts spatial and temporal Transformers to realize multi-camera view transformation and multi-frame feature fusion [4]. It significantly improves modeling capability for occluded targets and dynamic scenes. The third is the end-to-end closed-loop maturity stage (2023-2025). Its core breakthrough is deep integration of BEV into the full end-to-end pipeline. In this stage, BEV is no longer limited to perception output, but serves as the global feature carrier for joint optimization of full-stack modules.

The three-stage evolution of vision-centric BEV technology essentially reflects the deepening understanding of autonomous driving in academia and industry. Early foundational work solves the feasibility of pure vision BEV construction. The growth stage then breaks through the perception performance bottleneck. Finally, the closed-loop maturity stage realizes the transformation from perception technology to full-stack decision-making carrier. This evolutionary context provides a clear developmental basis for the decomposition of the core technical modules in the following sections.

3. Core technical modules of BEV-driven end-to-end urban autonomous driving

Vision-centric BEV representation learning forms a complete technical pipeline for end-to-end urban autonomous driving. It is centered on three core modules: multi-view spatiotemporal fusion, unified static-dynamic scene perception, and perception-planning collaborative adaptation. They enable efficient feature flow from image input to decision output. These three modules lay a technical foundation for the BEV-driven end-to-end architecture.

3.1. Multi-camera temporal BEV feature encoding module

This module is the fundamental link for BEV representation generation. It converts multi-view 2D image features into unified BEV spatial features and integrates temporal information to enhance scene modeling stability. It includes three mainstream technical routes with distinct performance trade-offs.

The first one is implicit depth estimation. LSS is the representative work. It is the foundational scheme of vision-only BEV [5]. This method first lifts image features into 3D space through implicit depth estimation and then projects them onto the BEV plane. It achieves fast inference. Nevertheless, LSS is susceptible to environmental interference in complex urban scenarios.

Second, Transformer-based spatiotemporal fusion is the current mainstream scheme. BEVFormer [4] and UniFusion [6] are its representative works. They realize collaborative optimization of BEV features in spatial and temporal dimensions. Compared with implicit depth estimation, this module significantly improves perception accuracy and generalization. For instance, UniFusion doubles the inference speed of serial fusion schemes and improves BEV map segmentation performance by 12.0%.

Third, temporal optimization strategies, represented by BEVDet4D [7] and BEVerse [8], are designed for dynamic urban scenes. They alleviate feature blur caused by fast-moving objects through multi-frame feature alignment.

3.2. Static and dynamic scene occupancy network and semantic modeling module

This module is the core for complex urban environment understanding. It realizes collaborative perception of static road structures and dynamic traffic participants through semantic segmentation

DOI: 10.54254/2755-2721/2026.GU33656

and occupancy networks. And it provides accurate scene priors for downstream planning and control. Three technical routes are adopted for different urban scene requirements.

First, static scene modeling focuses on the semantic representation of roads and stationary obstacles. BEVDet achieves joint optimization of 3D detection and BEV semantic segmentation by extending output branches [9]. Therefore, it enables accurate modeling of road structures. Experimental results reported in its arXiv preprint show that this method achieves a semantic segmentation mIoU of 62.1% on the nuScenes dataset.

Second, dynamic scene modeling relies on occupancy networks to encode the spatial states of objects. Tesla's Occupancy Network and DynamicBEV encode the spatial states of objects through BEV grids and predicts motion trajectories of dynamic targets [10]. Compared with static modeling schemes, it achieves 23% lower trajectory prediction error for dynamic objects through dynamic attention mechanism and motion consistency loss. As a result, it significantly improves robustness in urban mixed traffic scenarios.

Third, spatio-temporal joint modeling, represented by BEVerse [8], integrates static semantics and dynamic prediction into a unified framework. It simultaneously outputs semantic maps and trajectory prediction results.

3.3. BEV feature alignment and optimization module for the planning and control pipeline

This module is the critical link connecting perception to decision-making. It realizes the adaptation of BEV features to planning and control space, end-to-end joint learning, and automotive-grade lightweight deployment. And it solves the problem of disconnection between perception and planning, as well as computing power mismatch. Three core optimization directions are adopted with distinct focuses.

Dimension alignment is a prerequisite for the joint optimization of perception and planning and control. It maps high-dimensional BEV features to low-dimensional feature vectors compatible with planning modules through adaptive downsampling. Also, it accelerates planning inference speed by 3× with performance loss below 3%, laying the foundation for perception-planning joint optimization [11].

The second module is end-to-end joint learning. It directly binds BEV features to trajectory planning. Representative schemes include UniAD and ChatBEV [11, 12]. UniAD realizes collaborative optimization across perception, prediction and planning modules. ChatBEV enhances semantic features through large language models to improve decision rationality in complex long-tail urban scenarios.

Furthermore, lightweight design is a necessary condition for mass production deployment. It reduces the computational cost and parameter count of BEV encoders through depthwise separable convolution, model distillation and network pruning. This design enables the model to meet automotive-grade real-time requirements for mass production.

These three technical modules enable efficient generation of BEV representations, providing core support for constructing a closed-loop perception-decision architecture.

4. Comparative analysis of BEV-driven perception-planning closed-loop architectures

4.1. Core design principles and underlying logic of closed-loop architecture

The BEV-driven perception-planning closed-loop architecture takes unified BEV representation as the global feature carrier. It breaks down the technical barriers between perception and planning in

DOI: 10.54254/2755-2721/2026.GU33656

traditional modular systems, and realizes end-to-end joint optimization from multi-camera inputs to vehicle control commands. Its core design logic includes two dimensions: feature-level closed loop and data-level closed loop. The feature-level closed loop directly binds BEV perception features with planning decisions, eliminating error propagation between modules. The data-level closed loop realizes model iterative optimization based on real-world vehicle operation data, which is the key support for mass production deployment.

4.2. Comparative analysis of typical benchmark schemes and urban scenario adaptability

In academia, three representative benchmark schemes have been proposed for BEV-driven closed-loop architectures, with distinct technical focuses and urban scenario adaptability, as shown in Table 1.

Table 1. Comparative analysis of typical BEV-driven closed-loop benchmark schemes [8, 11, 12]

Benchmark Scheme	Core Innovation	Advantages in Urban Scenarios	Core Limitations
UniAD	Full-stack joint optimization of perception, prediction, planning and control	Holistic error suppression, optimal comprehensive performance, suitable for mass production deployment	High computational overhead, limited performance in long-tail scenarios
BEVerse	Unified perception and prediction with integrated spatio-temporal modeling	Superior dynamic target processing capability, excellent performance in congested urban scenarios	Weak semantic reasoning ability, insufficient adaptability to complex intersection scenarios
ChatBEV	Large language model-enhanced BEV semantic-visual cross-modal fusion	Strong semantic reasoning ability, excellent performance in long-tail urban scenarios	High computational cost, difficult to meet automotive-grade real-time requirements

Overall, the three schemes follow the core design logic of feature-level coupling and data-level iteration. UniAD is the preferred architecture for mass production deployment due to its full-stack optimization capability. BEVerse is more suitable for urban congested scenarios with dense dynamic targets, while ChatBEV represents the future development trend of integrating large models with BEV for long-tail scenario adaptation.

The differences between the three benchmark architectures essentially reflect three differentiated development paths of end-to-end autonomous driving: UniAD follows the mass production-oriented path, balancing full-stack performance and deployability; BEVerse follows the task-oriented path, pursuing optimal performance of unified perception and prediction in academic benchmarks; ChatBEV represents the future scenario-oriented path, focusing on long-tail scenario adaptation through large model enhancement. This comparison also reveals a core academic paradox: schemes with the best benchmark performance are often not the most suitable for mass production, which is rooted in the mismatch between static academic evaluation metrics and dynamic safety requirements of open urban driving.

5. Core challenges for mass production deployment

Despite the academic validation of BEV-driven end-to-end architectures, vision-centric BEV solutions still face three core bottlenecks in large-scale mass production for urban autonomous driving.

DOI: 10.54254/2755-2721/2026.GU33656

First, in terms of perception robustness, the depth estimation of pure-vision BEV is susceptible to complex urban environments. Low-light conditions and rainy weather lead to a significant degradation in 3D modeling accuracy [2].

Engineering implementation is the second challenge. The high computational overhead of high-performance BEV models and large model semantic enhancement conflicts with the limited computing resources of automotive-grade hardware. This situation makes it difficult to balance real-time performance and accuracy.

Ultimately, safety and compliance are critical issues. The industry has not yet established unified safety alignment standards for end-to-end architectures. Meanwhile, the imperfect fail-safe degradation mechanism and insufficient coverage of extreme scenarios in simulation verification make it difficult to meet mass production compliance requirements [13].

6. Conclusion

Based on representative works from top-tier conferences and journals (2020–2025), this paper sorts out the three-stage evolutionary path of vision-centric BEV technology, decomposes its three core technical modules for end-to-end urban autonomous driving, and conducts a systematic comparative analysis of typical closed-loop benchmark architectures and their adaptability to urban scenarios. This study clarifies the technical context of vision-centric BEV from perception technology to end-to-end full-link carrier, and provides a systematic reference for academic research and mass production deployment of end-to-end urban autonomous driving.

In the long run, vision-centric BEV-driven end-to-end autonomous driving will inevitably move towards a human-like driving paradigm. The current BEV architecture still follows the logic of "geometric modeling first, decision-making second", while human driving relies on the deep integration of geometric perception, semantic understanding and traffic causal reasoning. Therefore, the next-generation end-to-end BEV architecture must integrate traffic causal inference, large model semantic understanding and edge-cloud collaborative iteration, to build a safe, interpretable and robust system that can adapt to complex urban scenarios, and ultimately realize the large-scale commercial deployment of high-level autonomous driving. Future research will focus on improving the perception robustness of pure-vision BEV in complex urban scenarios, realizing automotive-grade lightweight deployment, and establishing a unified safety compliance system for end-to-end architectures.

References

[1] Chib, P. S., & Singh, P. (2023). Recent advancements in end-to-end autonomous driving using deep learning: A survey. IEEE Transactions on Intelligent Vehicles, 9(1), 103-118.

[2] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger and H. Li, "End-to-End Autonomous Driving: Challenges and Frontiers, " in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 12, pp. 10164-10183, Dec. 2024.

[3] Chen, W.-H.; Wu, J.-C.; Davydov, Y.; Yeh, W.-C.; Lin, Y.-C. Impact of Perception Errors in Vision-Based Detection and Tracking Pipelines on Pedestrian Trajectory Prediction in Autonomous Driving Systems. Sensors 2024, 24, 5066.

[4] Z. Li et al., "BEVFormer: Learning Bird's-Eye-View Representation From LiDAR-Camera via Spatiotemporal Transformers, " in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 47, no. 3, pp. 2020-2036, March 2025, doi: 10.1109/TPAMI.2024.3515454.

[5] Philion J, Fidler S. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d [C]//European conference on computer vision. Cham: Springer International Publishing, 2020: 194-210.

DOI: 10.54254/2755-2721/2026.GU33656

[6] Qin Z, Chen J, Chen C, et al. Unifusion: Unified multi-view fusion transformer for spatial-temporal representation in bird's-eye-view [C]//Proceedings of the IEEE/CVF international conference on computer vision. 2023: 8690-8699.
[7] Huang, J., & Huang, G. (2022). Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. arXiv preprint arXiv: 2203.17054.
[8] Zhang W, Zhu H, Zhou H, et al. BEVerse: Unified Perception and Prediction in Bird's-Eye-View for Autonomous Driving [C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023: 1612-1622.
[9] Huang J, Huang G, Zhu Z, et al. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view [J]. arXiv preprint arXiv: 2112.11790, 2021.
[10] Yao, J., & Lai, Y. (2023). Dynamicbev: Leveraging dynamic queries and temporal context for 3d object detection.
[11] Li H, Yang Z, Wang Y, Hu J, Li Y, Chen L, et al. Planning-oriented Autonomous Driving [C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023: 2219-2229.
[12] Xu Q, Chen S, Chen G, et al. ChatBEV: A visual language model that understands BEV maps [J]. arXiv preprint arXiv: 2503.13938, 2025.
[13] Zhu Z, Zhang Y, Chen H, et al. Understanding the robustness of 3D object detection with bird's-eye-view representations in autonomous driving [C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 21600-21610.