Pure Vision and Multi-modal Perception in Autonomous Driving: Performance, Challenges and Architectural Insights

Yaode Han[1] and Kaiyang Li[2,*]

[1] Xi'an Jiaotong-Liverpool University, School of AI and Advanced Computing, Taicang, China
[2] College of Arts and Sciences, University of Washington, Seattle, USA
*kaiyangli096@gmail.com

Abstract. Environmental perception is a core technical aspect of autonomous driving systems, and its architecture design directly determines the vehicle's understanding ability of the surrounding environment and driving safety. With the development of artificial intelligence and sensor technology, pure visual perception and multimodal fusion perception have become two mainstream technical routes. The trade-off in performance, cost, and reliability between the two is of great significance for the practical deployment of autonomous driving systems. This article focuses on the design of a new perception architecture and explores how to enhance the overall performance and safety of autonomous driving systems. The article systematically analyzes and compares the performance differences between pure visual perception systems and multimodal perception systems (fusing cameras, millimeter-wave radars, laser radars, etc.) in terms of perception accuracy, generalization ability, and robustness. It summarizes the challenges faced by each path (such as generalization, stability, and reliability) and their impacts on downstream tasks, and reviews several representative research works. In addition, this article also analyzes the advantages of pure visual perception in system construction from multiple dimensions, while also pointing out its shortcomings and deficiencies in dealing with complex environments, in order to provide theoretical references and practical inspirations for the design of future perception architectures.

Keywords: Autonomous driving; Pure visual perception; Multimodal fusion; Generalization ability; Robustness

1 Introduction

Environmental perception is the cornerstone for an autonomous driving system to achieve environmental understanding and decision-making planning. The choice of its technical route directly determines the safety, reliability, and implementation cost of the system. Currently, the autonomous driving perception architecture is mainly divided into two technical routes: pure visual perception and multimodal perception. The pure visual solution relies on camera data and achieves end-to-end scene understanding

K. Subramanian (ed.), *Proceedings of the International Workshop on Advances in Deep Learning for Image Analysis and Computer Vision (IWADIC 2025)*, Advances in Computer Science Research 128,
https://doi.org/10.2991/978-94-6239-648-7_79

through deep neural networks (such as BEVFormer, BEVDepth, etc.), featuring low cost, easy deployment, and strong generalization ability; while the multimodal solution integrates multiple sensors such as cameras, lidars, and millimeter-wave radars, enhancing perception accuracy and robustness through information complementation, especially performing more reliably in extreme weather and complex scenarios. Both paths have their advantages and disadvantages, and have become the focus of long-term debate in the academic and industrial communities.

The Wang Yongtao team proposed the RCBEVDet architecture, which realizes 3D target detection through BEV feature fusion of millimeter-wave radar and cameras. The designed deformable cross-attention mechanism effectively solves the problem of modality differences, and outperforms the pure visual benchmark model by 3.4 NDS on the nuScenes dataset, while maintaining real-time inference speed and modality absence robustness [1]; The team from Tsinghua University proposed a modality-invariant causal representation method at CVPR 2024, which can eliminate sensor deviations caused by rain and fog weather, improving the domain adaptation performance of multimodal perception by 28.3% [2]. The research by the Mercedes-Benz team in 2023 based on the CLIP-derived framework constructed a cross-modal embedding model, achieving an mAP score of 0.481 on the nuScenes dataset, improving the accuracy by 34% compared to the single-modal solution, providing an efficient paradigm for multimodal feature alignment [3].

Although these methods have promoted the development of the perception system at different levels, both the pure visual and multi-modal architectures still face key challenges: the former has insufficient reliability in adverse weather conditions, while the latter is limited by the difficulty of cross-modal alignment and the complexity of system integration. Additionally, the output quality of the perception module directly affects downstream tasks such as prediction and planning, and current research in the collaborative optimization of perception and decision-making still falls short in this regard. Especially in practical deployment, how to balance system cost, energy consumption and performance, and construct a hybrid architecture that is both efficient and robust has become a core issue restricting the implementation of high-level autonomous driving.

In view of this, this paper systematically compares the performance differences and generalization capabilities of pure visual and multi-modal perception models, analyzes the synchronization, alignment and robustness challenges in the fusion process, and deeply explores the impact mechanism of the perception architecture on high-level tasks. Through a review of a number of cutting-edge literature and related experimental verification, this paper aims to provide theoretical basis and architectural inspiration for the design of the next-generation autonomous driving perception system, and promote the evolution of perception technology towards an efficient hybrid direction of "visual dominance and multi-modal assistance".

2 Performance differences and generalization capabilities between pure visual perception and multimodal perception

End-to-end pure vision is achieved through cameras, using models such as Transformer to directly generate BEV perception results. There is no multimodal data alignment, making it easy to deploy and highly generalizable. Anupkumar Bochare et al. have demonstrated that a pure vision system based on multi-view camera data and deep neural networks can achieve 85% road segmentation accuracy and 85-90% vehicle detection rate, with an average position error controlled within 1.2 meters. Its core advantage lies in the joint optimization of depth estimation and target detection, enabling reliable spatial perception without the need for lidar [4]; As shown in Table 1, comparative experiments indicate that the pure vision GPU computation is less and the inference speed is fast, with an algorithm mAP of 0.456 and NDS of 0.555. It has strong robustness and better generalization ability in complex scenes under illumination variations. It can learn a large amount of environmental knowledge through extensive data augmentation and be applied to solve challenging unknown tasks. The BEVMap architecture proposed by Mincheol Chang et al. further proves that by integrating map prior information, the IoU of the pure vision system in vehicle segmentation tasks can be improved by 1.5%, and the IoU of drivable area segmentation can be improved by 13.4%, significantly improving the accuracy of depth estimation and scene semantic understanding [5].

Table1: Performance Comparison of Mainstream Pure Vision BEV Models (NuScenes Dataset)

Model	Gflops	FPS	mAP	NDS
BEVENet	161.42	47.6	0.456	0.555
BEVFormer	1303.5	1.7	0.416	0.476
SOLOFusion	223.2	11.1	0.427	0.534
BEVDepth	233.7	14.7	0.356	0.477

BEVENet is a pure convolutional model proposed by Yuxin Li et al. Its computing speed during testing is 47.6 FPS, surpassing the multimodal methods based on ViT [6]. However, pure vision is very sensitive to data volume and is difficult to ensure generalization performance in complex environments; at the same time, quantization shows that the three-modal scheme has better anti-obstruction accuracy, but due to its more complex scheme, its generalization ability has declined; in restricted scenarios, using pure vision is more convenient and direct, while multimodal methods are more vulnerable; in clear structural scenarios such as roads, pure vision combined with structural information can achieve precise recognition.

Liu Fei et al. mentioned from the perspective of visual perception [7], pure vision has many advantages compared to other methods, but in harsh environments such as foggy days, pure vision cannot fully exert its own functions; combining radar and vision can achieve these in the same time. Overall, the performance gap between pure vision and multimodal methods mainly depends on the degree of redundancy at the input end.

Pure vision pursues obtaining a larger degree of generalization ability with the smallest computational cost, but multi-sensor fusion pursues the optimal solution without considering the cost issue.

3 Challenges of Synchronization, Alignment and Robustness in Modal Fusion

The coordination of multi-modal sensors requires ensuring the synchronization of the input sources of perception and completing the registration of sensor outputs. For complex environments with large modality variations, how to achieve better perception robustness is also a problem. For example, Dang Xiangwei et al. proposed [8] that an effective two-step registration algorithm can be established between two sensors to improve the robustness and reliability of the perception module, and by using features to effectively register the spatial alignment from millimeter-wave radar (2D point cloud) and laser radar (3D point cloud), noise and multipath interference can be reduced. Experimental results show that the positioning error after fusion is reduced by 50% compared to before, especially in smoke conditions, it compensates for the defect of a single modality. Zhiqing Wei et al. classified the fusion process into data level, decision level and feature level [9], and pointed out that sensor calibration (such as coordinate calibration and radar point filtering) is a key challenge in the fusion process.

Keli Huang et al. summarized 50+ articles classified by fusion stage (data level/feature level) [10], and found that synchronization (time-space imbalance under high-speed conditions) and late fusion (dimensionality reduction simplification, ignoring cross-modal connections, poor robustness). Di Feng et al. proposed corresponding optimization measures for different types of obstacles ahead and complex and diverse detection objects in various weather conditions [11], to solve the decrease in target detection indicators due to interference, and raised new questions about how to achieve better detection performance in open sets under such conditions. The redundancy provided by multi-modalities will introduce interference, thereby misleading the final result, which is more difficult to distinguish than single-modal vision, that is, single-modal vision is affected by different conditions (such as weather conditions), but to some extent, multi-modalities can counteract this influence. Therefore, further research on multi-modalities in the future still needs to propose more applicable to various environments and achieve effective adaptive fusion algorithms.

4 Impact Mechanism of Perception Architecture on High-Level Tasks

The output results of the perception system will affect the working effect of subsequent navigation, path planning and decision-making and other important parts.

Yan Gong et al. from the perspective of "progressive BEV perception" summarized various methods for BEV perception over the past few years [12], and they believe that improving the accuracy of obstacle avoidance through multi-modal fusion is a very

good means, but it increases the input of multiple models from different perspectives, thereby increasing the computational pressure, and the additional benefits brought by this approach will not increase indefinitely, and eventually reach a "benefit boundary", when this boundary is exceeded, further increasing additional modalities will not bring too much benefit return.

In terms of the perception paradigms of the two, pure vision relying on BEV representation can directly analyze temporal and spatial relationships and reason, while multi-modalities due to stronger redundancy (can provide more backup information) will lead to slower decision-making.

In addition, in 2015, Brody Huval et al. proved that a CNN model relying only on vision can affect the planning stage [13], which is also a single-modal application. Although adding multi-modalities can make up for the deficiency in accuracy, it is difficult to directly fuse the data of different modalities.

The input structure of the perception architecture determines the high-level task mechanism: Pure vision is achieved through end-to-end optimization using Transformers, reducing the propagation of intermediate errors; for multimodal perception, attention should be paid to the selection of different modalities' weights, and errors should not be allowed to interfere with the decision-making process redundantly. The higher the complexity of the environmental scene, the better the generalization ability of multiple-angle fusion is, but beyond three modalities, there is little room for improvement. The reasons for choosing pure vision are mentioned in the literature because such tasks have strong timeliness and do not require high safety and reliability requirements, while choosing multimodal perception is done when planning and decision-making are affected and safety requirements are higher. Anupkumar Bochare et al. have shown that the pure vision system, through the collaborative perception of six or seven panoramic camera viewpoints, can achieve 360-degree environment coverage without blind spots, effectively supplementing the blind spots of a single viewpoint, and the unified BEV representation it generates can provide continuous and complete spatial information support for global path planning, playing a key role in the rationality of planning decisions [4]. Additionally, high-level tasks such as path planning are completed based on the consistency judgment of the spatiotemporal continuity of perception data. Therefore, an architecture based on pure vision is convenient for directly constructing the overall gradient flow of perception to optimize the model effect, while multimodal perception requires the use of attention to concentrate attention between modalities, which may lead to potential time delays. In future simulation environments, when task goals can be quickly switched, the model based on pure vision is more reliable under dynamic decision-making, while in the face of dynamic uncertainties, such as night driving, the multimodal model shows stronger advantages.

5 Conclusion

This paper systematically analyzed the performance differences and architectural characteristics of the pure visual and multi-modal perception technologies in

autonomous driving. The study shows that the pure visual solution demonstrates significant deployment advantages due to its low hardware cost and high inference speed. Models such as BEVENet can achieve a real-time inference speed of 47.6 FPS on the NuScenes dataset, demonstrating good engineering feasibility. However, its perception reliability in adverse weather conditions still faces bottlenecks, and the average accuracy indicators are generally lower than those of the multi-modal solution. In contrast, multi-modal perception shows greater potential in terms of accuracy and robustness through sensor redundancy fusion. For example, in extreme scenarios such as smoke, the fusion method can reduce the positioning error by 50%. However, it also faces system complexity and high computational overhead due to data alignment, synchronization, and complex fusion algorithms.

Current technology development still faces three core challenges. The complexity of deep integration is the top priority. How to design advanced fusion algorithms that can adaptively handle asynchronous, missing, or conflicting information between modalities remains an open issue. The bottleneck of data and simulation constrains system evolution. The massive, precise, and synchronized labeled data required by multi-modal systems is extremely costly, and the construction of realistic simulation environments for extreme scenarios still needs improvement. The real-time performance and cost constraints of the system are also not to be ignored. Complex fusion models pose severe challenges to the vehicle-mounted computing platform, and meeting the requirements of low latency and high throughput real-time inference under limited power and cost is an obstacle that must be overcome in engineering implementation.

Looking forward to the future, autonomous driving perception technology will develop in four main directions. Lightweight fusion and "software-defined" perception will become the research focus, shifting from relying on expensive hardware for "strong perception" to intelligent fusion based on advanced algorithms, using more efficient neural network architectures to achieve performance breakthroughs with low-cost sensor combinations. Dynamic adaptation and trustworthy perception are inevitable requirements. The system needs to have the ability to adjust fusion strategies in real time based on scene complexity and weather conditions and be able to output uncertainty estimates, building a transparent and trustworthy autonomous driving system. Active perception of interaction with the environment represents a new direction of technological evolution. The perception system will be deeply integrated with prediction and planning modules, actively controlling vehicle movement to reduce blind areas, verify uncertainties, and achieve a leap from "seeing" to "understanding" and "exploring". Cross-modal self-supervised learning will break through the bottleneck of labeled data, enabling the model to mine correlations between modalities from a large amount of natural driving data, reducing reliance on manual labeling.

In conclusion, autonomous driving perception technology is moving from independent perception of a single modality to deep and intelligent multi-modal fusion. Future breakthroughs will rely on collaborative innovation in algorithms, data, and system architecture. By constructing a "vision-led, multi-modal-assisted" hybrid architecture, a balance point between performance, cost, and reliability can be found, ultimately achieving a new generation of perception system with high cost-

effectiveness, high reliability, and high adaptability to the environment. This is not only the inevitable trend of technological development but also the key path for the large-scale implementation of high-level autonomous driving.

Authors Contribution

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

References

1. Wang, Y. T. et al.: RCBEVDet: Radar-Camera BEV Fusion for Robust 3D Object Detection in Autonomous Driving, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 12456-12465. (2024)
2. Qinghua University Team: Modal-Invariant Causal Representation for Multimodal Perception Under Extreme Weather, in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 8972-8981. (2024)
3. Mercedes-Benz R&D Team.: Cross-Modal Embedding Model for Autonomous Driving Perception, Nature Machine Intelligence, vol. 5, no. 3, pp. 218-227, (2023) doi: 10.1038/s42256-023-00654-x.
4. Anupkumar, B.: Camera-Only Bird's Eye View Perception: A Neural Approach to LiDAR-Free Environmental Mapping for Autonomous Vehicles. arXiv preprint arXiv:2505.06113v1 [cs.CV]. (2025)
5. Mincheol, C., Seokha, M., Reza, M., Jinkyu, K.: BEVMap: Map-Aware BEV Modeling for 3D Perception. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 7419-7428. (2024)
6. Li, Y. et al.: Towards Efficient 3D Object Detection in Bird's-Eye-Space for Autonomous Driving: A Convolutional-Only Approach, 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, pp. 2170-2177, (2023) doi: 10.1109/ITSC57777.2023.10422223.
7. Liu, F., Lu, Z. H. et al.: Vision-based environmental perception for autonomous driving. arXiv:2212.11453. (2023)
8. Dang, X. W., Qin, F., Bu, X. X. et al.: A Robust Perception Algorithm for Millimeter-Wave Radar and LiDAR Fusion for Intelligent Driving, Journal of Radars, vol. 10, no. 4, pp. 622–631 (2021) doi: 10.12000/JR21036.
9. Wei, Z. Q., Zhang, F. K., Chang, S. et al.: MmWave Radar and Vision Fusion for Object Detection in Autonomous Driving: A Review. arXiv:2108.03004. (2021)
10. Huang, K. L., Shi, B. T. et al.: Multi-modal Sensor Fusion for Auto Driving Perception: A Survey. IEEE Intelligent Transportation Systems Magazine, vol. 15, no. 5, pp. 36-58, (2023)
11. Di, F., Christian, H. et al.: Deep Multi-modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges. IEEE Transactions on Intelligent Transportation Systems. (2020)
12. Gong, Y., Wang, N. B. et al.: Progressive Bird's Eye View Perception for Safety-Critical Autonomous Driving: A Comprehensive Survey. arXiv:2508.07560 (2024)
13. Brody, H., Tao, W. et al.: An Empirical Evaluation of Deep Learning on Highway Driving. arXiv:1504.01716 (2015)