ITM Web of Conferences **84**, 04024 (2026) https://doi.org/10.1051/itmconf/20268404024
ATCIDS 2026

Analysis of Target Detection Technology in the Field of Autonomous Driving by Integrating Lidar and Camera

Yutong Wu[1] *
[1]College of Information Science and Technology, Beijing University of Chemical Technology, 102200, Beijing, China

Abstract. Environmental perception is the core support for the implementation of autonomous driving technology. The heterogeneous fusion of LiDAR and cameras, through the complementary advantages of geometric and semantic information, effectively breaks through the performance bottleneck of a single sensor. This paper systematically reviews the technological evolution in this field, constructs a dual-dimensional classification system of the fusion stage and data form, and deeply analyzes the innovative mechanisms and dataset performance of representative technologies such as BEVFusion and BEVFormer. Research shows that the core technology achieves an NDS level of 69.2% to 74.1% in 3D object detection tasks, while revealing core bottlenecks such as insufficient cross-modal spatiotemporal homogeneity robustness, difficulty in balancing real-time performance and accuracy, and limited generalization ability. Finally, we look forward to the development directions such as dynamic elastic integration, software and hardware collaborative optimization, generalization and expansion, and standardized implementation, providing important references for the research and development and engineering application of high-level autonomous driving perception systems.

1 Introduction

Automatic driving is evolving from L2 level to L4 level, with strict requirements on the accuracy, robustness and real-time performance of the environmental awareness system. As a core task, target detection needs to accurately identify the three-dimensional information of the target to support path planning and control. Camera detection has gone through the stages of manual features and deep learning. The latter generates rich semantic images at low cost, but is sensitive to light and weather and lacks direct depth information [1]. LiDAR is based on the principle of time of flight, through the geometric features, depth learning phase. PointNet pioneered point cloud learning [2]. In recent years, BEV technology has become a hotspot, which is light resistant and has high 3D accuracy, but remote point clouds are sparse, lack of semantics, and cost is high [3]. The two perception dimensions are complementary, and integration can improve system robustness, which is an inevitable trend of technology development. This paper focuses on multi-stage fusion classification, introduces related

* Corresponding author's email: 2023040041@buct.edu.cn

technologies and performance, and analyzes challenges and trends.

2. Development of LiDAR and Camera Fusion Technology

LiDAR and camera fusion technology can be classified from two complementary perspectives of fusion stage and data form. The former is divided into early, middle, late and cross stage multi-level fusion according to the depth of information fusion, and is a mature classification framework of the field [4], which is described below. The latter is divided into dense level and sparse level fusion according to the data processing form [5], which makes up for the lack of attention to sparse scenes in traditional classification. The two are related and complementary, providing reference for the selection of fusion schemes.

2.1 Early Fusion (Data-Level Fusion)

Early fusion is to realize sensor information interaction at the level of original data or low dimensional features, complete accurate data alignment through strict spatial calibration, and maximize the integrity of original information. The geometric projection class, represented by PointFusion, projects the LiDAR point cloud to the pixel position of the camera image, splices geometric information and image features to form joint features, effectively making up for the lack of direct depth information of the camera. The semantic enhancement class is represented by PointPainting [6]. First, the image is subject to pixel level semantic segmentation, and then the semantic information is "sprayed" to the corresponding point cloud. This stage is suitable for short distance, unobstructed simple traffic scenes, but it is difficult to deal with the complex situation of sparse point clouds or poor image quality.

2.2 Mid-term Fusion (Feature-level Fusion)

Mid-term fusion is the current mainstream strong fusion paradigm for automated driving 3D detection. The core is to complete cross modal alignment and deep fusion in uniform space such as BEV after extracting high-order features from dual sensors, which balances accuracy and efficiency.

BEV is the core technology direction and also belongs to dense level fusion. The key is to solve the problem of the difference between the camera and LiDAR. The core process is to transform the multi view camera features into BEV features through the view transformation module, and then fuse them with the BEV features after voxelization of LiDAR point clouds. LSS provides a unified spatial carrier for cross modal fusion by lifting image pixels along the optical axis to 3D voxel features and then projecting them into BEV space to form dense features [7]. In BEVFusion, the camera and LiDAR can generate BEV features independently. In case of LiDAR failure, the performance can still be maintained by camera stream, and 40 times speed increase can be achieved through kernel function optimization. NuScenes dataset mAP reaches 69.2% [8, 9]. With the help of spatial cross attention and temporal self attention mechanism [3], BEVFormer has improved the feature alignment accuracy and the robustness of moving objects and occluded scenes, and the NDS of the test set has reached 74.1%. TransFusion reduced sensor calibration dependency through Transformer soft correlation mechanism [10], and NDS reached 71.7%.

This stage adapts to more complex scenes, and can fully mine high-order complementary information between modes, but it is difficult to align features, and has high requirements on backbone network expression ability and hardware computing power.

2.3 Late-Stage Fusion (Decision-level Fusion)

Late-fusion belongs to weak fusion attribute. The two sensors independently perform detection tasks. Only the detection frame, confidence level and other results are fused at the decision-making level. The module has strong independence and high fault tolerance.The rule-driven approach establishes the correspondence between detection boxes through the Hungarian algorithm, and then optimizes the results through Kalman filtering and weighted averaging. It has a simple logic, high computational efficiency, and is suitable for engineering scenarios with high real-time requirements. The learning-driven approach is an advanced direction. CLOCS achieved the state-of-the-art (SOTA) performance on the KITTI dataset at that time [11]. Methods such as ValID verify LiDAR detection results through camera features, reducing the false positive rate by more than 60% [12]. Both belong to sparse-level fusion. This stage adapts to the deployment scenario of heterogeneous sensors. In case of a single sensor failure, the system can still maintain basic functions. It is difficult to land the project, but there is a lack of intermediate feature interaction, insufficient information utilization, and detection accuracy is generally lower than that of mid term and multi-level fusion.

2.4 Multi-level Integration (Cross-Stage Advanced Integration)

The core of multi-level fusion is to integrate the fusion logic of two or more levels in "data feature decision". The representative technology SuperFusion builds a three-level architecture of "data level - feature level - BEV level" [13]. The data level uses LiDAR depth to monitor camera depth estimation. Feature level complements the semantics of remote point clouds by cross attention mechanism. Weighted fusion after BEV level correction mode dislocation. The 90 meter long distance high-definition map is generated in the nuScenes dataset, and the detection accuracy of each interval is better than that of the single-stage method. This kind of method has the advantages of integration at all stages, but the model structure is complex and the calculation cost is high.

2.5 Fusion of dense level and sparse level

As an important supplement to the classification of "fusion stage", this classification focuses on whether to introduce dense BEV feature map [5]. Dense level fusion is based on dense features such as BEV/voxels, such as BEVFusion and TransFusion. It has high information density, can mine fine-grained complementary information, and has significant precision advantages. However, most methods are computationally complex, and only some query driven methods reduce computational power consumption through sparse optimization. The sparse level fusion directly operates sparse data, and the point level such as PointPainting adapts to near distance semantic enhancement, and the candidate box level such as CLOCS adapts to distant, point cloud sparse or high real-time scenes, with efficient computing, simple engineering, and low-cost hardware adaptation, but the precision of complex scenes is lower than that of the dense level. The two classifications complement each other. The classification in the fusion stage reflects the evolution of technology. The "dense/sparse" classification provides a direct reference for hardware selection and scene adaptation, forming a complete technology classification system.

3. Core technical challenges

3.1 Weak robustness in cross-modal spatiotemporal alignment

Depending on offline calibration and fixed timing synchronization, sensor vibration and temperature drift will destroy feature alignment in practice; There is no dynamic weight mechanism in extreme environment, and the recognition accuracy of remote and small targets drops sharply.

3.2 Balance between real-time performance and detection accuracy

High-precision methods such as dense BEV fusion need to handle high-resolution feature maps and complex attention calculations, and the inference delay exceeds the real-time requirements of on-board hardware. Although lightweight solutions reduce computational overhead, they often sacrifice fine-grained features and are difficult to meet the precision requirements of L3 and above autonomous driving.

3.3 Dual constraints of generalization and technical implementation

The mainstream methods rely on large annotation sets, and the lack of long tail target annotations leads to weak cross dataset migration. The cost of LiDAR for medium and high wire bundles is high, the sensor interface/format has no unified standard, and cross platform adaptation is difficult.

4. Future development trends

4.1 Dynamic Spatio-temporal Alignment and Elastic Fusion Paradigm

The visual SLAM online calibration module and the sequential self attention network are developed to realize the real-time correction structure of spatiotemporal parameters. Modeling state confidence adaptive strategy, switching fusion logic in case of single mode failure.

4.2 Software and Hardware Collaborative Optimization and Multi-Task Unified Framework.

The algorithm layer uses sparse BEV and model compression to reduce reasoning delay. The hardware layer is designed with special acceleration unit adapted to the car end chip; Multi task integration into shared feature space improves system efficiency.

4.3 Generalization Expansion and Standardization Implementation Support.

Reduce label dependency through unsupervised domain adaptation and other technologies, and expand to special environments. Establish a unified sensor interface/synchronization standard to reduce adaptation costs and promote industrialization.

5. Conclusion

The fusion technology of LiDAR and camera has evolved from single-mode complementarity to BEV space unified representation paradigm, and the detection accuracy and robustness of conventional scenes have been significantly improved. However, bottlenecks such as cross modal dynamic alignment, extreme environment adaptation, and engineering balance still restrict its large-scale implementation. In the future, it is necessary to promote technological innovation such as adaptive fusion and software hardware collaboration, and improve sensor interfaces and data standards based on scenario requirements. After the technology iteration, the field will be more efficient and reliable, providing a solid perceptual support for the safety upgrade and industrialization popularization of automatic driving.

References

1. Z. He, W. Ran, S. Liu, et al. Low-light Image Enhancement with Multi-scale Attention and Frequency-domain Optimization. IEEE Transactions on Circuits and Systems for Video Technology, 34(4): 2861-2875. (2024)
2. C. R. Qi, H. Su, K. Mo, et al. PointNet++: Deep hierarchical feature learning on point sets in a metric space// Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), 5099-5108. (2017)
3. 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, doi: 10.1109/TPAMI.2024.351545. (2025)
4. Y. Pang, J. Zhang, and Y. Ning Research Progress on LiDAR and Camera Fusion Target Detection Technology in Autonomous Driving. Automobile technology, 1-14 (2026) https://doi.org/10.19620/j.cnki.1000-3703.20250225.
5. W. Ji, Y. Li, and J. Wang, et al. A Review of 3D Object Detection Based on the Fusion of Lidar and Camera. Computer Science,1-24. (2026) https://link.cnki.net/urlid/50.1075.tp.20250717.1650.024.
6. S. Vora, A. H. Lang, B. Helou, et al. PointPainting: Sequential Fusion for 3D Object Detection// Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA: IEEE, 4604-4612. (2020)
7. J. Philion, S. Fidler. Lift, Splat, Shoot: Encoding Images from Arbitrary Camera Rigs by Implicitly Unprojecting to 3D// Proceedings of the European Conference on Computer Vision (ECCV). Glasgow, UK: Springer, 194-210. (2020)
8. Z. Liu, H. Tang, A. Amini, et al. BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation//2023 IEEE International Conference on Robotics and Automation. IEEE, 2774-2781. (2023)
9. T. T. Liang, H. W. Xie, K. C. Yu, et al. BEVFusion: A Simple and Robust LiDAR-Camera Fusion Framework// Proceedings of the 36th Conference on Neural Information Processing Systems (NIPS). New Orleans, LA, USA: NIPS, (2022)
10. X. Bai, Z. Hu, X. Zhu, et al. TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1090-1099. (2022)
11. S. Pang, D. Morris, H. Radha. CLOCs: Camera-LiDAR object candidates fusion for 3D object detection//2020 IEEE/RSJ International Conference on Intelligent Robots and Systems. IEEE, 10386-10393. (2020)
12. V. Vats, M. B. Nizam, J. Davis. VaLID: Verification as Late Integration of Detections for LiDAR-Camera Fusion. https://arxiv. org/abs/2409.15529. (2025)
13. H. Dong, W. H. Gu, X. J. Zhang, et al. SuperFusion: Multilevel LiDAR-Camera Fusion for Long-Range HD Map Generation// Proceedings of the 2024 IEEE International Conference on Robotics and Automation (ICRA). Yokohama, Japan: IEEE, 9056-9062. (2024)