Integrating Multi-Modal Sensors: A Review of Fusion Techniques for Intelligent Vehicles

Chuheng Wei*, *Student Member, IEEE*, Ziye Qin, *Student Member, IEEE*, Ziyan Zhang, Guoyuan Wu, *Senior Member, IEEE*, Matthew J. Barth, *Fellow, IEEE*,

***Abstract*— Multi-sensor fusion plays a critical role in enhancing perception for autonomous driving, overcoming individual sensor limitations, and enabling comprehensive environmental understanding. This paper first formalizes multi-sensor fusion strategies into data-level, feature-level, and decision-level categories and then provides a systematic review of deep learning-based methods corresponding to each strategy. We present key multi-modal datasets and discuss their applicability in addressing real-world challenges, particularly in adverse weather conditions and complex urban environments. Additionally, we explore emerging trends, including the integration of Vision-Language Models (VLMs), Large Language Models (LLMs), and the role of sensor fusion in end-to-end autonomous driving, highlighting its potential to enhance system adaptability and robustness. Our work offers valuable insights into current methods and future directions for multi-sensor fusion in autonomous driving.**

I. Introduction

A. Background

Perception serves as the cornerstone of autonomous driving systems, enabling vehicles to understand and interact with their surroundings safely and effectively [1]. Sensors play a pivotal role in this perception process, as they collect critical environmental data necessary for tasks such as adaptive cruise control (ACC), lane keeping, and collision avoidance [2]. However, the limitations of individual sensors, including restricted range and susceptibility to environmental factors, necessitate cooperative sensing [3].

Relying solely on one type of sensor often results in performance bottlenecks due to inherent weaknesses. For example, cameras struggle under low-light conditions [4], LiDAR can be hampered by adverse weather [5], and Radar provides limited resolution for fine-grained object recognition [6]. To overcome these challenges, multi-sensor fusion has emerged as a vital approach in autonomous driving. By integrating data from multiple sensor modalities, multi-sensor fusion provides a more comprehensive and robust understanding of the environment, enhancing the reliability and safety of autonomous systems [3].

As the complexity of driving scenarios increases, the role of multi-sensor fusion becomes even more significant. Modern autonomous driving systems demand a seamless integration of data from heterogeneous sensors to address challenges such as adverse weather conditions, dynamic road environments, and real-time decision-making [7]. This has fueled extensive research and innovation in multi-sensor fusion techniques, making it a critical area of study in the advancement of autonomous driving.

Chuheng Wei, Ziyan Zhang, Guoyuan Wu, and Matthew J. Barth are with the College of Engineering, Center for Environmental Research and Technology, University of California at Riverside, Riverside, CA, 92507.

Ziye Qin is with the School of Transportation and Logistics, Southwest Jiaotong University, Chengdu, China.

*Corresponding author. e-mail: chuheng.wei@email.ucr.edu

B. Sensors in Autonomous Driving

a) Camera: Cameras are among the most widely used sensors in autonomous driving due to their ability to capture high-resolution images and detailed visual information [8]. They excel in recognizing objects [9], traffic signs [10], and lane markings [11], providing rich semantic data crucial for decision-making. However, cameras are sensitive to lighting conditions and can perform poorly in scenarios such as nighttime driving or glare from direct sunlight.

b) LiDAR: LiDAR sensors utilize laser beams to measure distances and generate precise 3D maps of the environment [8]. Their high spatial resolution and ability to operate independently of ambient light make them invaluable for object detection and localization tasks [12]. Despite these advantages, LiDAR systems are costly, generate large volumes of data, and can be adversely affected by heavy rain or fog [5].

c) Radar: Radar sensors are known for their robustness in adverse weather conditions, such as rain, fog, and snow [13]. They excel in measuring object velocity and detecting moving targets at long ranges, making them particularly suitable for highway scenarios [14]. However, Radar's resolution is lower compared to cameras and LiDAR, limiting its ability to provide detailed object classification and environmental mapping [3].

TABLE I: Comparison of Camera, LiDAR, and Radar capabilities.

Characteristic	Camera	LiDAR	Radar
Cost	Low	High	Medium
Data Volume	Medium	High	Low
Field of View	Wide	Ultra-Wide†, Narrow‡	Narrow
Color Resolution	High	None	None
Rain Performance	Poor	Medium	High
Fog Performance	Poor	Poor	High
Night Performance	Medium	High	High

† Refers to mechanical LiDAR.
‡ Refers to solid-state LiDAR.

The complementary strengths and weaknesses of these sensors underscore the necessity of sensor fusion in autonomous driving. By combining data from cameras, LiDAR, and Radar, a more holistic and resilient perception system

can be achieved. Table I provides a comparative evaluation of these sensors based on various criteria.

C. Contributions

Previous works such as [2], [3], [7], [15], [16] have summarized various autonomous driving algorithms, but rarely provide a formalized mathematical analysis of sensor fusion techniques. Additionally, few studies comprehensively discuss emerging hot topics in sensor fusion applications.

This paper makes the following contributions:

- **Formalized Sensor Fusion:** Provided a formalized mathematical summary of multi-sensor fusion in autonomous driving.
- **Comprehensive Algorithm Review:** Reviewed state-of-the-art multi-sensor fusion algorithms, categorizing them into data, feature, and decision-level fusion while analyzing their methodologies, strengths, and limitations in autonomous driving.
- **Datasets:** Conducted a thorough review of autonomous driving datasets, highlighting their suitability for multi-sensor fusion studies.
- **Emerging Trends:** Discussed multi-sensor fusion in the context of emerging trends such as end-to-end autonomous driving and the integration of Vision-Language Models (VLM) and Large Language Models (LLM), providing insights into future directions for research.

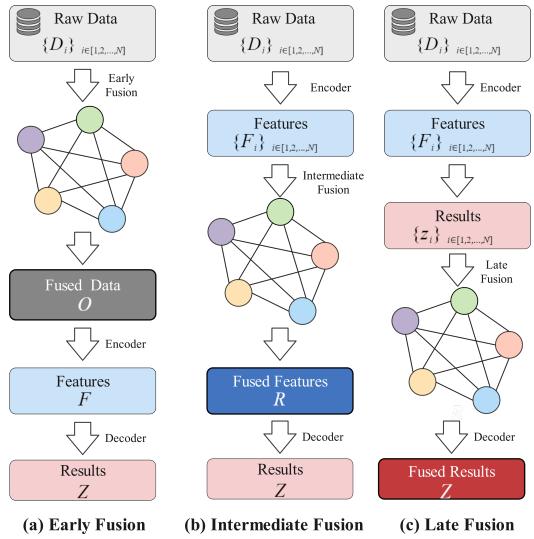

Fig. 1: Architecture for Different Sensor Fusion Strategies

II. Formalization

Multi-sensor fusion in autonomous driving can be framed as a process that transforms raw data from multiple sensors into a unified output. Let $D = \{D_1, D_2, \ldots, D_n\}$ denote the set of raw data collected from n sensors, and Z represent the final fused output. The fusion process is formulated as:

$$Z = \Omega(D; \theta), \tag{1}$$

where $\Omega(\cdot; \theta)$ is the overall fusion function parameterized by θ, responsible for integrating sensor data for tasks such as object detection, tracking, or decision-making.

To maintain consistency across fusion strategies, we define α, ψ, and ϕ as parameter vectors associated with the fusion, encoding, and decoding functions, respectively. These parameters may vary in form depending on the fusion level. While not all parameters are explicitly instantiated in every equation, they represent tunable components within the respective functions.

The following subsections illustrate three common fusion strategies: Data-Level (Early), Feature-Level (Intermediate), and Decision-Level (Late). Figure 1 provides a visual representation of these strategies.

A. Data-Level Fusion (Early)

In data-level fusion, all raw sensor data are fused first, then transformed into features, and finally decoded:

a) Fuse Raw Data:

$$O = G(D_1, D_2, \ldots, D_m; \alpha), \tag{2}$$

where $G(\cdot; \alpha)$ merges the raw inputs into an intermediate representation O.

b) Extract Features:

$$F = E(O; \psi), \tag{3}$$

where $E(\cdot; \psi)$ encodes O into feature space.

c) Produce Output:

$$Z = H(F; \phi), \tag{4}$$

where $H(\cdot; \phi)$ decodes the features F into the final output Z.

B. Feature-Level Fusion (Intermediate)

Feature-level fusion separately encodes each sensor's data, then fuses the resulting features before producing the final output:

a) Encode Each Sensor:

$$F_i = E(D_i; \psi), \tag{5}$$

where each F_i is a feature vector extracted from sensor D_i.

b) Fuse Features:

$$R = G(F_1, F_2, \ldots, F_m; \alpha), \tag{6}$$

where $G(\cdot; \alpha)$ aggregates feature vectors $F_1 \ldots F_m$ into a fused representation R.

c) Produce Output:

$$Z = H(R; \phi), \tag{7}$$

where $H(\cdot; \phi)$ converts the fused features R into the final output Z.

C. Decision-Level Fusion (Late)

Decision-level fusion first decodes each sensor's feature into an intermediate prediction, then fuses these predictions to obtain the final result:

a) Encode Sensor Data:

$$F_i = E(D_i; \psi), \tag{8}$$

where $E(\cdot; \psi)$ extracts features from the raw data D_i.

b) Sensor-Specific Output:

$$z_i = H(F_i; \phi), \tag{9}$$

where $H(\cdot; \phi)$ decodes each feature vector F_i into a sensor-specific output z_i.

c) Final Fusion:

$$Z = G(z_1, z_2, \ldots, z_m; \alpha), \tag{10}$$

where $G(\cdot; \alpha)$ merges these sensor-specific outputs into the final fused result Z.

III. Fusion Methods for Multi-Sensor Integration

Multi-sensor fusion is categorized by the integration stage within the perception pipeline into data-level, feature-level, and decision-level fusion, corresponding to early, intermediate, and late fusion techniques with different trade-offs in information preservation, computation, and robustness. Mix fusion methods further enhance adaptability by dynamically selecting fusion levels based on context. Table II summarizes fusion strategies, sensor combinations, and key applications.

A. Data-Level Fusion (Early Fusion)

Data-level fusion, or early fusion, combines raw sensor data at the beginning of the processing pipeline, enabling the model to learn from aligned sensor inputs directly, as illustrated in Figure 1-(a) This approach is common in radar and camera fusion, where radar point clouds are projected onto the camera frame using a calibration matrix, creating a unified front-view perspective. Since radar points often lie on the ground plane, further data augmentation and correction are applied to enhance accuracy. For instance, YOdar [17] and CRF-Net [18] introduce a learnable vertical axis for each radar point, allowing the network to adaptively find the effective vertical depth coverage of radar points. Nabati et al. proposed RRPN [19], a radar-based real-time region proposal network, which generates object proposals using radar data and predefined bounding boxes, replacing the RPN module in two-stage image detectors to accelerate processing.

As LiDAR sensors become more prevalent in autonomous vehicles, early fusion methods for camera-LiDAR data have advanced significantly. Qi et al. proposed Frustum PointNet [20], which first detects objects in 2D images, converts them into 3D frustums via camera-LiDAR projections, and applies deep learning for 3D object detection. Wang et al. proposed PointAugmenting [21], which enriches LiDAR point clouds with semantic features extracted from 2D image models and applies cross-modal data augmentation to improve detection performance under challenging conditions. A related method, PointPainting [22], enhances detection by projecting pixel-level semantic scores from an image segmentation network onto LiDAR points. Although PointPainting is often classified as a data-level or early fusion method due to the integration of semantic information before feature encoding, it is also considered by some to lie between data-level and feature-level fusion, as the semantic scores represent features extracted from multi-view images.

More complex early fusion methods incorporate multiple sensor modalities, such as camera, LiDAR, and radar. MVXNet [23] integrates RGB images and LiDAR data using simple yet effective fusion techniques like PointFusion and VoxelFusion within the VoxelNet architecture. Similarly, Lin et al. [24] developed a dedicated stage to preprocess radar depth data before fusing it with image data, achieving improved depth estimation by leveraging complementary information from both radar and camera.

B. Feature-Level Fusion (Intermediate Fusion)

Feature-level fusion, as shown in Figure 1-(b), combines sensor data at the feature extraction stage, effectively integrating complementary information from different modalities. This approach allows for richer feature representations and is currently the most widely used method in multi-sensor fusion, especially for applications in autonomous driving.

Initial works primarily focused on LiDAR-camera fusion for object detection. Chen et al. introduced MV3D [26], which generates 3D region proposals from LiDAR Bird's-eye view (BEV) data and fuses them with RGB image features to improve detection accuracy. Ku et al. proposed AVOD [28], a two-stage fusion approach that enhances bounding box regression and classification by integrating LiDAR and image features at the region proposal stage.

Subsequent approaches further enriched feature representation and adaptability. Liang et al. developed ContFuse [27], which integrates multi-scale LiDAR reflections while preserving spatial details. Huang et al. introduced EPNet [33], aligning point features with semantic information via point-guided image fusion and consistency-enforcing losses. Yang et al. proposed Graph-RCNN [36], leveraging neighborhood graphs within region proposals for iterative message passing between image and point cloud features. Zhu et al. presented VPFNet [35], introducing "virtual" intermediate points to bridge the resolution gap between sparse point clouds and dense images. Chen et al. introduced AutoAlign [39], a learnable strategy that fuses image and voxel features through cross-attention for improved multi-modal consistency.

BEV-based fusion has become essential for 3D object detection. Liang et al. proposed MMF [30], integrating image features into BEV for high-accuracy detection and depth completion. Yoo et al. developed 3D-CVF [32], which projects 2D features into a smooth BEV space to refine proposals in LiDAR-camera fusion. Liu et al. introduced BEVFusion [40], unifying multi-modal features in BEV to preserve geometric and semantic details efficiently. Zhao et al. extended this with CRKD [42], transferring feature knowledge from a LiDAR-camera model to a camera-radar model via knowledge distillation. Zhou et al. proposed RCBEV [41], addressing radar's sparse, noisy points through a spatio-temporal encoder. Multi-modal fusion has also been applied in adverse conditions like fog. Zhang et al. introduced Transfusion [63], a two-stage model that processes LiDAR and radar data independently before fusion, improving perception in low-visibility scenarios.

Beyond object detection, feature-level fusion supports semantic segmentation and 3D semantic occupancy prediction. John et al. proposed SO-Net [29], a radar-vision framework facilitating object detection and segmentation. Meyer et al. integrated camera and LiDAR features within a localized view for joint feature extraction. Zhuang et al. introduced PMF [34], dynamically weighting LiDAR depth and camera

TABLE II: Overview of multi-sensor fusion methods and their applications

Method	Proposed Year	Fusion Level			Fused Sensors			Task
		Data	Feature	Decision	Camera	LiDAR	Radar	
Frustum PointNet [25]	2018	✓			✓	✓		Object Detection
RRPN [19]	2019	✓			✓		✓	Object Detection
CRF-Net [18]	2019	✓			✓		✓	Object Detection
MVX-Net [23]	2019	✓			✓	✓		Object Detection
[24]	2020	✓			✓		✓	Depth Estimation
PointPainting [22]	2020	✓*			✓	✓		Object Detection
YOdar [17]	2021	✓			✓		✓	Object Detection
PointAugmenting [21]	2021	✓			✓	✓		Object Detection
MV3D [26]	2017		✓		✓	✓		Object Detection
ContFuse [27]	2018		✓		✓	✓		Object Detection
AVOD [28]	2018		✓		✓	✓		Object Detection
SO-Net [29]	2019		✓		✓		✓	Semantic Segmentation & Object Detection
MMF [30]	2019		✓		✓	✓		Object Detection
[31]	2019		✓		✓	✓		Semantic Segmentation & Object Detection
3D-CVF [32]	2020		✓		✓	✓		Object Detection
EPNet [33]	2020		✓		✓	✓		Object Detection
PMF [34]	2021		✓		✓	✓		Semantic Segmentation & Object Detection
VPFNet [35]	2022		✓		✓	✓		Object Detection
Graph RCNN [36]	2022		✓		✓	✓		Object Detection
SFD [37]	2022		✓		✓	✓		Object Detection
DeepFusion [38]	2022		✓		✓	✓		Object Detection
AutoAlign [39]	2022		✓		✓	✓		Object Detection
BEVfusion [40]	2023		✓		✓	✓		Object Detection
RCBEV [41]	2023		✓		✓		✓	Object Detection
CRKD [42]	2024		✓		✓		✓	Object Detection
GATR [43]	2024		✓		✓	✓		Object Detection
GAFusion [44]	2024		✓		✓	✓		Object Detection
HyDRa [45]	2024		✓		✓		✓	3D Semantic Occupancy Prediction
CO-Occ [46]	2024		✓		✓	✓		3D Semantic Occupancy Prediction
FOP-MOC [47]	2015			✓	✓	✓	✓	Object Detection
[48]	2015			✓	✓	✓	✓	Object Detection
[49]	2018			✓	✓	✓		Object Detection
[50]	2018			✓	✓	✓		Object Detection
RoarNet [51]	2019			✓	✓	✓		Object Detection
PI-RCNN [52]	2020			✓	✓	✓		Object Detection
[53]	2020			✓		✓	✓	Multi Object Tracking
[54]	2020			✓	✓	✓		Object Detection
CLOCs [55]	2020			✓	✓	✓		Object Detection
Paigwar [56]	2021			✓	✓	✓		Object Detection
SCF [57]	2024			✓	✓	✓		Object Detection
[58]	2016	✓	✓	✓	✓	✓		Object Detection
[59]	2019	✓	✓	✓	✓	✓		Object Detection
CenterFusion [60]	2021	✓		✓	✓		✓	Object Detection
CRAFT [61]	2023	✓		✓	✓		✓	Object Detection
RCM-Fusion [62]	2024		✓	✓	✓	✓		Object Detection

* Although often seen as early fusion, PointPainting [22] also shares traits with feature-level fusion due to the use of extracted semantic features.

semantics for robust 3D segmentation. Extending feature fusion to 3D occupancy, Wolters et al. proposed HyDRa [45], refining radar-camera fusion with radar-weighted depth consistency and a Height Association Transformer. Pan et al. introduced Co-Occ [46], enhancing LiDAR-camera fusion via a geometry-aware module with KNN-based feature merging and volume rendering for improved 3D occupancy prediction.

C. Decision-Level Fusion (Late Fusion)

Decision-level fusion integrates sensor outputs at a higher abstraction level, where individual sensors provide their independent predictions, and the system combines them to reduce uncertainty and enhance robustness. This approach is particularly beneficial in autonomous driving, where fusing multiple sensor decisions improves overall perception and adaptability in dynamic environments.

Decision-level fusion remains widely explored in sensor-based perception, with most studies focusing on LiDAR-camera fusion. Asvadi et al. [50] processed LiDAR point clouds into depth, reflectivity, and color maps, which were input into a YOLO-based network, followed by a scoring function and non-maximum suppression for refinement. Similarly, Du et al. [49] introduced a model-based approach, refining 3D structures using predefined car models and a 2D CNN for enhanced accuracy. Shin et al. [51] developed RoarNet, a two-stage pipeline estimating object poses from monocular images before 3D refinement. To improve pillar-based detection, Paigwar et al. [56] combined visual and point cloud features, enhancing localization in sparse data. Meanwhile, Melotti et al. [54] leveraged distance-based weighting to account for sensor performance variations. Addressing fusion consistency, Pang et al. [55] proposed CLOCs, aligning 2D and 3D detections through geomet-

ric and semantic constraints to improve efficiency. Further refining detections, Yu et al. [57] introduced the Semantic Consistency Filter (SCF), which filters false positives based on 2D segmentation masks.

Beyond camera-LiDAR fusion, several studies have incorporated radar to enhance perception. Kunz et al. [48] introduced a hierarchical modular framework that probabilistically integrates radar, camera, and LiDAR data. Chavez-Garcia and Aycard's FOP-MOC model [47] applied an evidential fusion framework to combine classification outputs, improving object detection and tracking accuracy. Kurapati et al. [53] fused radar and camera outputs for multi-target tracking, using the Hungarian algorithm to refine associations and improve tracking stability.

By integrating sensor outputs at the decision level, these methods effectively enhance detection accuracy, mitigate individual sensor limitations, and improve robustness in real-world driving scenarios.

D. Mix Fusion

Mix fusion methods integrate data, features, and decisions from multiple sensors at various stages of the processing pipeline, offering a flexible and adaptive approach to multi-sensor perception in autonomous systems. Unlike strictly defined fusion strategies, mix fusion dynamically selects fusion levels based on task requirements and environmental conditions, allowing models to leverage the complementary strengths of different sensor modalities more effectively.

Schlosser et al. [58] were among the first to explore fusion of LiDAR and camera data within convolutional neural networks, demonstrating that different fusion levels impact performance uniquely. Their findings highlighted that combining multiple strategies enhances adaptability in complex scenarios. Expanding on this, Nabati et al. introduced CenterFusion [60], which integrates radar-based spatial features into an object detection pipeline by dynamically incorporating radar information at various processing stages rather than committing to a fixed fusion strategy. Similarly, Kim et al. [61] proposed CRAFT, which employs a soft polar association and a spatio-contextual fusion transformer, enriching spatial and contextual information while making fusion decisions based on confidence levels.

Caltagirone et al. [59] applied mix fusion in road detection, using an FCN-based approach that projects LiDAR point clouds onto the image plane to produce dense 2D road images while incorporating semantic segmentation outputs to refine road boundaries. More recently, Kim et al. introduced RCM-Fusion [62], a radar-camera 3D object detection framework utilizing both feature-level and instance-level fusion. This method employs a Radar Guided BEV Encoder to convert camera data into a BEV format while refining object positions through radar-adaptive sampling, adjusting the fusion process based on sensor reliability.

In summary, mix fusion enables robust perception by adapting sensor integration across multiple levels, from raw data processing to final decision-making. By leveraging the strengths of various fusion techniques dynamically, mix fusion enhances detection accuracy and robustness in complex, dynamic environments, making it a promising direction for future multi-sensor fusion research.

IV. Datasets

In this section, we introduce several available datasets collected from multi-sensor setups deployed on vehicles. The discussion focuses primarily on the use of cameras, radars, and LiDARs in each dataset and their potential applications to autonomous driving tasks. To obtain a valid dataset, several key steps are required: sensor deployment, sensor calibration, data collection, time synchronization, and data annotation. As summarized in Table III, these datasets are sourced from vehicles operating across North America, Europe, and Asia, facilitating research into autonomous driving within diverse traffic scenarios. All of the vehicles used for data collection are equipped with at least cameras and LiDARs; however, radars are less commonly included, which limits the capability of some datasets to accurately sense distance and speed. Consequently, many datasets are less suitable for motion forecasting tasks. Each dataset has unique characteristics that make it valuable for specific research applications. For example, KITTI [64], the first large-scale open-source dataset, remains widely used and serves as a benchmark for the format of related datasets. KITTI covers a wide range of scenarios, including highways and urban roads. Datasets like Argoverse [66] and Lyft Level 5 [67], equipped with multiple cameras, offer 360-degree image data, enabling comprehensive environmental perception. Notably, the Argoverse dataset is the first large-scale dataset to include a semantic vector map, enhancing its utility for 3D tracking and motion forecasting. Meanwhile, the Lyft Level 5 dataset provides 1,118 hours of recorded data, making it the largest dataset for motion prediction and a rich resource for exploring diverse traffic scenarios. The A2D2 dataset [72] offers additional CAN bus information, which can be leveraged for end-to-end autonomous driving development. Furthermore, recent cooperative perception datasets, such as DAIR-V2X-V [76], TUMTraf [77] , and V2XReal-VCC [78], integrate data from vehicle-mounted sensors, enabling advanced research into multi-sensor fusion.

V. Discussion and Future Directions

A. Algorithmic Perspectives

Feature-level fusion, the most widely used approach in multi-sensor fusion, primarily follows "Shallow Feature Intermediate Fusion", where sensor data is aligned based on spatial correspondences, merely merging positional information without fundamentally transforming features. While effective, this approach limits cross-modal interactions, as features largely retain their original characteristics with minimal modifications beyond alignment and concatenation. Future research should explore "Deep Feature Intermediate Fusion", where feature representations undergo learned transformations, such as convolutional operations, cross-modal attention, or adaptive re-weighting. This deeper integration

TABLE III: A detailed comparison of autonomous driving datasets for multi-sensor fusion.

Dataset	Year†	Camera	LiDAR	Radar	Region/Platform	Tasks
KITTI [64]	2013	4	1	0	Germany	Object Detection, Object Tracking, Semantic Segmentation
ApolloScape [65]	2018	6	2	0	China	Object Detection, Object Tracking, Semantic Segmentation
Argoverse [66]	2019	7	2	0	USA	Object Tracking, Motion Forecasting
Lyft Level 5 [67]	2019	7	3	5	USA	Object Detection, Semantic Segmentation, Motion Prediction
Waymo Open Dataset [68]	2019	5	5	0	USA	Object Detection, Object Tracking
H3D [69]	2019	3	1	0	USA	Object Detection, Object Tracking
nuScenes [70]	2020	6	1	5	USA, Singapore	Object Detection, Object Tracking
A*3D [71]	2020	2	1	0	Singapore	Object Detection
A2D2 [72]	2020	6	5	0	Germany	Object Detection, Semantic Segmentation
PandaSet [73]	2020	6	2	0	USA	Object Detection
PixSet [74]	2021	3	1	1	Canada	Object Detection
View-of-Delft [75]	2022	1	1	1	The Netherlands	Object Detection, Trajectory Prediction
DAIR-V2X-V [76]	2022	3	1	1	China	Object Detection
TUMTraf V2X-V [77]	2024	1	1	0	Germany	Object Detection, Semantic Segmentation
V2XReal-VC [78]	2024	4	1	0	USA	Object Detection

† The publication year refers to the year the dataset was introduced in the respective paper.

can enhance feature synergy across modalities, improving robustness and adaptability in complex driving environments.

Advancements in hardware and algorithms enable the exploration of other fusion strategies. Data-level fusion is becoming more feasible with efficient preprocessing and high-throughput data handling. Similarly, decision-level fusion benefits from improved sensor-specific algorithms and contextual integration across modalities, enhancing perception robustness. Future research should move beyond structural integration, leveraging feature transformations for richer cross-modal interactions and deeper semantic fusion.

B. Dataset Considerations

The availability and diversity of datasets are critical for advancing multi-sensor fusion research. Current datasets often focus on specific sensor combinations, such as camera-LiDAR or camera-Radar, and may lack comprehensive multi-modal coverage for real-world scenarios. There is an increasing need for datasets that capture challenging conditions, including adverse weather, nighttime environments, and complex urban settings. Furthermore, datasets that provide synchronized, high-quality data from multiple modalities can enable more effective training and evaluation of fusion algorithms, accelerating progress in the field.

C. Sensor Fusion in End-to-End Autonomous Driving

In the context of end-to-end autonomous driving systems, multi-sensor fusion algorithms are expected to evolve towards seamless integration within neural architectures. The goal is to enable direct mapping from sensor inputs to driving actions while preserving interpretability and reliability. End-to-end fusion frameworks must address challenges such as handling diverse sensor data representations, ensuring real-time performance, and maintaining robustness under dynamic and unpredictable conditions. As these systems mature, they are likely to incorporate hybrid fusion strategies, leveraging the strengths of each fusion level to optimize decision-making.

D. Sensor Fusion in VLM and LLM Integration

The emergence of Vision-Language Models (VLMs) and Large Language Models (LLMs) offers exciting opportunities for advancing multi-sensor fusion. These models can facilitate cross-modal understanding by incorporating semantic context into perception tasks, enabling autonomous systems to make more informed decisions. Future developments may focus on integrating VLMs and LLMs with multi-sensor fusion frameworks to enhance their ability to process unstructured, multi-modal data. Additionally, the potential for large-scale pretraining on diverse datasets could improve generalization across various driving scenarios, paving the way for more intelligent and adaptable autonomous systems.

VI. Conclusion

This paper provides a comprehensive review of multi-sensor fusion techniques for autonomous driving, categorizing methods into data-level, feature-level, object-level, and decision-level fusion. By analyzing state-of-the-art algorithms, we highlight the strengths and limitations of each fusion strategy, emphasizing the importance of feature-level fusion in current applications. The formulation of multi-sensor fusion provides a structured framework for understanding and implementing these methods effectively.

Looking ahead, advancements in hardware, datasets, and algorithmic innovations will continue to drive progress in multi-sensor fusion research. The integration of end-to-end frameworks, combined with the capabilities of VLMs and LLMs, holds significant promise for enhancing the robustness and scalability of autonomous systems. Future work will need to address challenges such as real-time performance, interpretability, and adaptability to diverse driving environments, ensuring the safe and reliable deployment of autonomous vehicles in real-world scenarios.

References

[1] Z. Wang, Y. Wu, and Q. Niu, "Multi-sensor fusion in automated driving: A survey," *IEEE Access*, vol. 8, pp. 2847–2868, 2019.

[2] M. L. Fung, M. Z. Chen, and Y. H. Chen, "Sensor fusion: A review of methods and applications," in *2017 29th Chinese Control And Decision Conference (CCDC)*. IEEE, 2017, pp. 3853–3860.

[3] J. Kocić, N. Jovičić, and V. Drndarević, "Sensors and sensor fusion in autonomous vehicles," in *2018 26th Telecommunications Forum (TELFOR)*. IEEE, 2018, pp. 420–425.

[4] C. Wei, G. Wu, and M. J. Barth, "Feature corrective transfer learning: End-to-end solutions to object detection in non-ideal visual conditions," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 23–32.

[5] M. Kutila, P. Pyykönen, H. Holzhüter, M. Colomb, and P. Duthon, "Automotive lidar performance verification in fog and rain," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 1695–1701.

[6] Z. Lu, D. Dzurisin, H.-S. Jung, J. Zhang, and Y. Zhang, "Radar image and data fusion for natural hazards characterisation," *International Journal of Image and Data Fusion*, vol. 1, no. 3, pp. 217–242, 2010.

[7] D. J. Yeong, G. Velasco-Hernandez, J. Barry, and J. Walsh, "Sensor and sensor fusion technology in autonomous vehicles: A review," *Sensors*, vol. 21, no. 6, p. 2140, 2021.

[8] P. Wang, "Research on comparison of lidar and camera in autonomous driving," in *Journal of Physics: Conference Series*, vol. 2093, no. 1. IOP Publishing, 2021, p. 012032.

[9] C. Wei, G. Wu, M. Barth, P. H. Chan, V. Donzella, and A. Huggett, "Enhanced object detection by integrating camera parameters into raw image-based faster r-cnn," in *2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2023, pp. 4473–4478.

[10] A. De La Escalera, L. E. Moreno, M. A. Salichs, and J. M. Armingol, "Road traffic sign detection and classification," *IEEE transactions on industrial electronics*, vol. 44, no. 6, pp. 848–859, 1997.

[11] D. C. Andrade, F. Bueno, F. R. Franco, R. A. Silva, J. H. Z. Neme, E. Margraf, W. T. Omoto, F. A. Farinelli, A. M. Tusset, S. Okida *et al.*, "A novel strategy for road lane detection and tracking based on a vehicle's forward monocular camera," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 4, pp. 1497–1507, 2018.

[12] K. Żywanowski, A. Banaszczyk, and M. R. Nowicki, "Comparison of camera-based and 3d lidar-based place recognition across weather conditions," in *2020 16th International Conference on Control, Automation, Robotics and Vision (ICARCV)*. IEEE, 2020, pp. 886–891.

[13] Z. Liu, Y. Cai, H. Wang, L. Chen, H. Gao, Y. Jia, and Y. Li, "Robust target recognition and tracking of self-driving cars with radar and camera information fusion under severe weather conditions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 7, pp. 6640–6653, 2021.

[14] G. Chen, W. Meckle, and J. Wilson, "Speed and safety effect of photo radar enforcement on a highway corridor in british columbia," *Accident Analysis & Prevention*, vol. 34, no. 2, pp. 129–138, 2002.

[15] H. A. Ignatious, M. Khan *et al.*, "An overview of sensors in autonomous vehicles," *Procedia Computer Science*, vol. 198, pp. 736–741, 2022.

[16] Q. Tang, J. Liang, and F. Zhu, "A comparative review on multi-modal sensors fusion based on deep learning," *Signal Processing*, p. 109165, 2023.

[17] K. Kowol, M. Rottmann, S. Bracke, and H. Gottschalk, "Yodar: Uncertainty-based sensor fusion for vehicle detection with camera and radar sensors," in *Proceedings of the 13th International Conference on Agents and Artificial Intelligence*. SCITEPRESS-Science and Technology Publications, 2021.

[18] F. Nobis, M. Geisslinger, M. Weber, J. Betz, and M. Lienkamp, "A deep learning-based radar and camera sensor fusion architecture for object detection," in *2019 Sensor Data Fusion: Trends, Solutions, Applications (SDF)*. IEEE, 2019, pp. 1–7.

[19] R. Nabati and H. Qi, "Rrpn: Radar region proposal network for object detection in autonomous vehicles," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 3093–3097.

[20] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[21] C. Wang, C. Ma, M. Zhu, and X. Yang, "Pointaugmenting: Cross-modal augmentation for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 11 794–11 803.

[22] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.

[23] V. A. Sindagi, Y. Zhou, and O. Tuzel, "Mvx-net: Multimodal voxelnet for 3d object detection," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 7276–7282.

[24] J.-T. Lin, D. Dai, and L. Van Gool, "Depth estimation from monocular images and sparse radar data," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 233–10 240.

[25] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum pointnets for 3d object detection from rgb-d data," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 918–927.

[26] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3d object detection network for autonomous driving," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2017, pp. 1907–1915.

[27] M. Liang, B. Yang, S. Wang, and R. Urtasun, "Deep continuous fusion for multi-sensor 3d object detection," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 641–656.

[28] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3d proposal generation and object detection from view aggregation," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1–8.

[29] V. John, M. Nithilan, S. Mita, H. Tehrani, R. Sudheesh, and P. Lalu, "So-net: Joint semantic segmentation and obstacle detection using deep fusion of monocular camera and radar," in *Image and Video Technology: PSIVT 2019 International Workshops, Sydney, NSW, Australia, November 18–22, 2019, Revised Selected Papers 9*. Springer, 2019, pp. 138–148.

[30] M. Liang, B. Yang, Y. Chen, R. Hu, and R. Urtasun, "Multi-task multi-sensor fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 7345–7353.

[31] G. P. Meyer, J. Charland, D. Hegde, A. Laddha, and C. Vallespi-Gonzalez, "Sensor fusion for joint 3d object detection and semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, 2019, pp. 0–0.

[32] J. H. Yoo, Y. Kim, J. Kim, and J. W. Choi, "3d-cvf: Generating joint camera and lidar features using cross-view spatial feature fusion for 3d object detection," in *Computer vision–ECCV 2020: 16th European conference, Glasgow, UK, August 23–28, 2020, proceedings, part XXVII 16*. Springer, 2020, pp. 720–736.

[33] T. Huang, Z. Liu, X. Chen, and X. Bai, "Epnet: Enhancing point features with image semantics for 3d object detection," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XV 16*. Springer, 2020, pp. 35–52.

[34] Z. Zhuang, R. Li, K. Jia, Q. Wang, Y. Li, and M. Tan, "Perception-aware multi-sensor fusion for 3d lidar semantic segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 16 280–16 290.

[35] H. Zhu, J. Deng, Y. Zhang, J. Ji, Q. Mao, H. Li, and Y. Zhang, "Vpfnet: Improving 3d object detection with virtual point based lidar and stereo data fusion," *IEEE Transactions on Multimedia*, vol. 25, pp. 5291–5304, 2022.

[36] H. Yang, Z. Liu, X. Wu, W. Wang, W. Qian, X. He, and D. Cai, "Graph r-cnn: Towards accurate 3d object detection with semantic-decorated local graph," in *European Conference on Computer Vision*. Springer, 2022, pp. 662–679.

[37] X. Wu, L. Peng, H. Yang, L. Xie, C. Huang, C. Deng, H. Liu, and D. Cai, "Sparse fuse dense: Towards high quality 3d detection with depth completion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5418–5427.

[38] Y. Li, A. W. Yu, T. Meng, B. Caine, J. Ngiam, D. Peng, J. Shen, Y. Lu, D. Zhou, Q. V. Le *et al.*, "Deepfusion: Lidar-camera deep fusion for multi-modal 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 17 182–17 191.

[39] Z. Chen, Z. Li, S. Zhang, L. Fang, Q. Jiang, F. Zhao, B. Zhou, and H. Zhao, "Autoalign: Pixel-instance feature aggregation for multi-modal 3d object detection," *arXiv preprint arXiv:2201.06493*, 2022.

[40] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. L. Rus, and S. Han, "Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *2023 IEEE international conference on robotics and automation (ICRA)*. IEEE, 2023, pp. 2774–2781.

[41] T. Zhou, J. Chen, Y. Shi, K. Jiang, M. Yang, and D. Yang, "Bridging the view disparity between radar and camera features for multi-modal fusion 3d object detection," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 2, pp. 1523–1535, 2023.

[42] L. Zhao, J. Song, and K. A. Skinner, "Crkd: Enhanced camera-radar object detection with cross-modality knowledge distillation," in

Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15 470–15 480.

[43] Y. Luo, L. He, S. Ma, Z. Qi, and Z. Fan, "Gatr: Transformer based on guided aggregation decoder for 3d multi-modal detection," *IEEE Robotics and Automation Letters*, 2024.

[44] X. Li, B. Fan, J. Tian, and H. Fan, "Gafusion: Adaptive fusing lidar and camera with multiple guidance for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 21 209–21 218.

[45] P. Wolters, J. Gilg, T. Teepe, F. Herzog, A. Laouichi, M. Hofmann, and G. Rigoll, "Unleashing hydra: Hybrid fusion, depth consistency and radar for unified 3d perception," *arXiv preprint arXiv:2403.07746*, 2024.

[46] J. Pan, Z. Wang, and L. Wang, "Co-occ: Coupling explicit feature fusion with volume rendering regularization for multi-modal 3d semantic occupancy prediction," *IEEE Robotics and Automation Letters*, 2024.

[47] R. O. Chavez-Garcia and O. Aycard, "Multiple sensor fusion and classification for moving object detection and tracking," *IEEE Transactions on Intelligent Transportation Systems*, vol. 17, no. 2, pp. 525–534, 2015.

[48] F. Kunz, D. Nuss, J. Wiest, H. Deusch, S. Reuter, F. Gritschneder, A. Scheel, M. Stübler, M. Bach, P. Hatzelmann *et al.*, "Autonomous driving at ulm university: A modular, robust, and sensor-independent fusion approach," in *2015 IEEE intelligent vehicles symposium (IV)*. IEEE, 2015, pp. 666–673.

[49] X. Du, M. H. Ang, S. Karaman, and D. Rus, "A general pipeline for 3d detection of vehicles," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 3194–3200.

[50] A. Asvadi, L. Garrote, C. Premebida, P. Peixoto, and U. J. Nunes, "Multimodal vehicle detection: fusing 3d-lidar and color camera data," *Pattern Recognition Letters*, vol. 115, pp. 20–29, 2018.

[51] K. Shin, Y. P. Kwon, and M. Tomizuka, "Roarnet: A robust 3d object detection based on region approximation refinement," in *2019 IEEE intelligent vehicles symposium (IV)*. IEEE, 2019, pp. 2510–2515.

[52] L. Xie, C. Xiang, Z. Yu, G. Xu, Z. Yang, D. Cai, and X. He, "Pi-rcnn: An efficient multi-sensor 3d object detector with point-based attentive cont-conv fusion module," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 12 460–12 467.

[53] K. R. Kurapati, M. Suma, and A. Chavan, "Multiple object tracking using radar and vision sensor fusion for autonomous vehicle," in *2020 IEEE International Conference for Innovation in Technology (INOCON)*. IEEE, 2020, pp. 1–6.

[54] G. Melotti, C. Premebida, and N. Gonçalves, "Multimodal deep-learning for object recognition combining camera and lidar data," in *2020 IEEE International Conference on Autonomous Robot Systems and Competitions (ICARSC)*. IEEE, 2020, pp. 177–182.

[55] S. Pang, D. Morris, and H. Radha, "Clocs: Camera-lidar object candidates fusion for 3d object detection," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020, pp. 10 386–10 393.

[56] A. Paigwar, D. Sierra-Gonzalez, Ö. Erkent, and C. Laugier, "Frustum-pointpillars: A multi-stage approach for 3d object detection using rgb camera and lidar," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2926–2933.

[57] L. Yu, J. Zhang, Z. Liu, H. Yue, and W. Chen, "Rethinking the late fusion of lidar-camera based 3d object detection," in *2024 IEEE 19th Conference on Industrial Electronics and Applications (ICIEA)*. IEEE, 2024, pp. 1–6.

[58] J. Schlosser, C. K. Chow, and Z. Kira, "Fusing lidar and images for pedestrian detection using convolutional neural networks," in *2016 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2016, pp. 2198–2205.

[59] L. Caltagirone, M. Bellone, L. Svensson, and M. Wahde, "Lidar–camera fusion for road detection using fully convolutional neural networks," *Robotics and Autonomous Systems*, vol. 111, pp. 125–131, 2019.

[60] R. Nabati and H. Qi, "Centerfusion: Center-based radar and camera fusion for 3d object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 1527–1536.

[61] Y. Kim, S. Kim, J. W. Choi, and D. Kum, "Craft: Camera-radar 3d object detection with spatio-contextual fusion transformer," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 1160–1168.

[62] J. Kim, M. Seong, G. Bang, D. Kum, and J. W. Choi, "Rcm-fusion: Radar-camera multi-level fusion for 3d object detection," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 18 236–18 242.

[63] C. Zhang, H. Wang, Y. Cai, L. Chen, and Y. Li, "Transfusion: Multi-modal robust fusion for 3d object detection in foggy weather based on spatial vision transformer," *IEEE Transactions on Intelligent Transportation Systems*, 2024.

[64] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[65] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang, "The apolloscape dataset for autonomous driving," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2018, pp. 954–960.

[66] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan *et al.*, "Argoverse: 3d tracking and forecasting with rich maps," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 8748–8757.

[67] R. Kesten, M. Usman, J. Houston, T. Pandya, K. Nadhamuni, A. Ferreira, M. Yuan, B. Low, A. Jain, P. Ondruska, S. Omari, S. Shah, A. Kulkarni, A. Kazakova, C. Tao, L. Platinsky, W. Jiang, and V. Shet, "Lyft level 5 av dataset 2019," https://level5.lyft.com/dataset/, 2019.

[68] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[69] A. Patil, S. Malla, H. Gang, and Y.-T. Chen, "The h3d dataset for full-surround 3d multi-object detection and tracking in crowded urban scenes," in *2019 International Conference on Robotics and Automation (ICRA)*. IEEE, 2019, pp. 9552–9557.

[70] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[71] Q.-H. Pham, P. Sevestre, R. S. Pahwa, H. Zhan, C. H. Pang, Y. Chen, A. Mustafa, V. Chandrasekhar, and J. Lin, "A* 3d dataset: Towards autonomous driving in challenging environments," in *2020 IEEE International conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 2267–2273.

[72] J. Geyer, Y. Kassahun, M. Mahmudi, X. Ricou, R. Durgesh, A. S. Chung, L. Hauswald, V. H. Pham, M. Mühlegg, S. Dorn *et al.*, "A2d2: Audi autonomous driving dataset," *arXiv preprint arXiv:2004.06320*, 2020.

[73] P. Xiao, Z. Shao, S. Hao, Z. Zhang, X. Chai, J. Jiao, Z. Li, J. Wu, K. Sun, K. Jiang *et al.*, "Pandaset: Advanced sensor suite dataset for autonomous driving," in *2021 IEEE International Intelligent Transportation Systems Conference (ITSC)*. IEEE, 2021, pp. 3095–3101.

[74] J.-L. Déziel, P. Merriaux, F. Tremblay, D. Lessard, D. Plourde, J. Stanguennec, P. Goulet, and P. Olivier, "Pixset: An opportunity for 3d computer vision to go beyond point clouds with a full-waveform lidar dataset," in *2021 ieee international intelligent transportation systems conference (itsc)*. IEEE, 2021, pp. 2987–2993.

[75] A. Palffy, E. Pool, S. Baratam, J. F. P. Kooij, and D. M. Gavrila, "Multi-class road user detection with 3+1d radar in the view-of-delft dataset," *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 4961–4968, 2022.

[76] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan, and Z. Nie, "Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.

[77] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, and A. C. Knoll, "Tumtraf v2x cooperative perception dataset," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*. IEEE, 2024.

[78] H. Xiang, Z. Zheng, X. Xia, R. Xu, L. Gao, Z. Zhou, X. Han, X. Ji, M. Li, Z. Meng *et al.*, "V2x-real: a largs-scale dataset for vehicle-to-everything cooperative perception," *arXiv preprint arXiv:2403.16034*, 2024.