

How Emergent Architectures Solve the Bird's Eyes View Shortcomings: A Systematic Review

Abdelaziz Sahbani^{1*}, Rihem Sebai², Tarek Bejaoui³, Hela Mahersia⁴

Laboratory of Automatic (LARA), National Engineering School of Tunis (ENIT), University of Tunis El Manar, Tunisia¹

LR - IDEA, Faculty of Science of Bizerte, University of Carthage, Tunisia^{2,3,4}

Abstract—Autonomous driving requires robust, accurate and real-time environmental perception. Although cameras, LiDAR, and radar provide complementary sensing capabilities, individual modalities remain vulnerable to limitations such as depth ambiguity, point-cloud sparsity, adverse weather, and restricted angular resolution. Multi-sensor fusion therefore represents a key approach for improving 3D object detection and scene understanding. This systematic review analyzes research published from 2022 to 2026 using the PRISMA framework and studies retrieved from Scopus and Web of Science. The review examines sensor characteristics and major fusion strategies, including early, feature, late, and hybrid fusion. It also covers BEV representations, Transformer-based architectures, evaluation methods, and benchmark datasets. Particular attention is given to multimodal combinations involving cameras, LiDAR, and radar, as well as cooperative perception through V2X systems. The analysis shows a clear shift toward BEV-based feature fusion and attention-driven architectures, reflecting their ability to integrate heterogeneous spatial and semantic information. However, the literature remains fragmented in several critical areas. Real-world benchmarks covering adverse weather and challenging conditions are still limited, while computational requirements remain a major barrier to real-time onboard deployment. Robust online calibration, adversarial security, cross-domain generalization, and heterogeneous cross-manufacturer V2X perception are also insufficiently investigated. In addition, 3D occupancy representations remain less explored than conventional bounding-box detection, and emerging architectures such as Mamba and large language models require broader validation on real-world data. These findings reveal persistent gaps between research advances and practical deployment, highlighting the need for efficient, secure, weather-robust, and generalizable fusion approaches for heterogeneous sensors and cooperative autonomous driving scenarios in real-world applications.

Keywords—Autonomous driving; multi-sensor fusion; Bird's-Eye View (BEV); multimodal perception; deep learning; transformers; 3D object detection; adverse weather conditions

I. INTRODUCTION

A. Motivation

In recent years, the development and deployment of autonomous vehicles have increased significantly, driven by advances in artificial intelligence, sensing technologies, and intelligent transportation systems [1], [2], [3]. This growth was also accelerated by the COVID-19 pandemic, which highlighted the need for contactless delivery and transportation services. During periods of restriction of movement, autonomous vehicles showed potential to support the delivery of essential

goods, including food, medicines, and household products, while reducing human contact. The second important factor is accident. In fact, every day, approximately 3,200 to 3,500 people worldwide die in road accidents. According to the World Health Organization (WHO), this amounts to more than 1.16 million deaths per year [4]. The notable information is that often 90% of these fatalities are cited as driver errors. In autonomous driving systems, we discover perception. In fact, it is about localizing, classifying, and tracking the 3D object in real time with a high precision under any environmental conditions. Sensors are the first communicator with the environment, especially in object detection. Nevertheless, taking each modality cannot satisfy the perception functionalities for the advanced level of autonomous driving [5], [6]. In addition, sensors degradation can affect the driving systems performance. This degradation can be due to adverse weather conditions or physical attacks [7].

Studies [5], [8], [9] describe in detail the limitations of each sensor, the camera provides semantic information but cannot provide depth, and its performances decrease at night (low-luminosity). On the other hand, LiDAR is an important 3D precision, but is expensive, a high energy consumer, and its performance degrades in rain or dense fog [42]. The millimeter wave radar (mmWave) penetrates all precipitation and automatically measures Doppler velocity, but suffers from insufficient angular resolution and multi-trajectory noise in urban environments [8], [12], [13]. The fundamental physical complementarity allows multi-sensor fusion to be unavoidable [5], [6], [9].

Since 2022, the field presents experienced a depth architecture revolution: Bird's-Eye View representations (BEV) are known as the standard for heterogeneous data fusion [14], [15], the cross-attention mechanism [16], [17], [18] and the cooperative perception vehicle to everything (V2X) open up unprecedented possibilities [10], [11], [42], [19], [20], [103].

Survey articles in this research field are presented. Wang et al [5] is the first complete survey article focusing on multi-sensors with deep learning methods in autonomous driving that introduces a taxonomy based on fusion strategies(early, features, late) and quantitative depth analysis on KITTI [21] and nuScenes benchmarks. But it does not cover the architectures discovered after 2022. In the same context, Wang et al. [6] propose a more formalized taxonomy adding the BEV paradigm, and transformers for multi-modal fusion with extended comparisons in KITTI and nuScenes. This study is limited by the absence of recent development, too such as Mamba, and cooperative V2X and insufficient discussion of

*Corresponding author

TABLE I. CRITICAL ANALYSIS OF EXISTING SURVEYS

Survey	Ref.	Journal (Year)	BEV	4D Radar	V2X	Adverse Conditions	Adversarial Security	Emergent Architectures
Wang et al.	[5]	IJCV (2023)	Partial	No	No	No	No	No
Wang et al.	[6]	IEEE TIV (2023)	Yes	No	No	No	No	No
Fan et al.	[8]	IEEE TIV (2024)	Partial	Yes	No	Partial	No	No
Wang et al.	[9]	IEEE TITS (2024)	Yes	No	No	No	No	No
Valverde et al.	[23]	Sensors (2024)	Partial	Partial	No	No	No	No
Shi et al.	[24]	CMC (2025)	Yes	Partial	No	No	No	No
Li et al.	[15]	IEEE TITS (2026)	Yes	Partial	Partial	No	Partial	No
Umar et al.	[28]	Adv. Eng. Inf. (2026)	No	Partial	No	No	No	No
This SRL	–	2026	Yes	Yes	Yes	Yes	Yes	Yes

TABLE II. RESEARCH QUESTIONS BASED ON IDENTIFIED GAPS

QR	Question statement	Addressed gap	Section of response
QR1	What are the main fusion strategies for 3D object detection?	Lack of 4D radar coverage in fusion pipelines and heterogeneous Vehicle-to-Everything strategies.	V-A
QR2	How do these techniques behave under various conditions (night/rain/fog, etc.)?	Absence of a review covering adverse conditions and adversarial security.	V-B
QR3	What datasets and metrics are used for evaluation?	Absence of V2X datasets and occupancy metrics.	V-C
QR4	Which emerging architectures are fundamentally transforming fusion?	Absence of reviews covering Mamba/SSM and limited coverage of LLMs.	V-D
QR5	What are the current limitations and the most promising future research trends?	Lack of certain important information.	V-E

computational constraints for embedded systems. Zhang et al. [22] point out that learning algorithms applied to urban driving must specifically contend with this diversity of conditions and sensors to remain reliable in practice. Fan et al. [8] focus on 4D mmWave radar, providing a comprehensive architecture and analyzing the physical properties of the sensors. However, the integration of 4D radar into tri-modal fusion architectures combining LiDAR, cameras, and radar remains largely unexplored. Its application in V2X scenarios and standardized comparisons with LiDAR under adverse weather conditions also require further investigation. Wang et al. [9] propose a clear study introducing an approach based on the level of fusion integration and analyzing the environment applications in smart cities. Li et al. [15] have recently presented a complete BEV survey that introduces the new architectures and topics. However, adverse weather conditions and adverse security have developed insufficiently. Valverde et al. [23] demonstrated a transverse view of modalities based on clear taxonomy and the historic analysis of the evolution of different methods, but the width of the perimeter limits the depth of the analysis and decreases the precision of quantitative comparisons between single-modal and fused approaches. Shi et al. [24] address these dependencies and analyze architectural trends. However, their selection methodology lacks strength due to the absence of a PRISMA protocol and Bibliometrix [29] analysis, while multi-object tracking and adversarial robustness are still unaddressed. Adnan et al. [25] specified their survey on multi-object tracking using clustering techniques such as DBSCAN and HDBSCAN, and tracking approaches like Kalman, IMM et DNN. Dal'Col et al. [26] extend the perspective to joint perception and trajectory prediction for autonomous driving,

while Fawole and Rawat [27] survey recent advances in 3D detection for autonomous vehicles, confirming both the richness and the thematic dispersion of review efforts in this field.

Nevertheless, the study is limited to LiDAR only and does not cover the multi-sensor concepts and the transformers on tracking. The last survey was conducted by Umar and al [28]. It is a systematic survey using the Kitchenham methodology that includes 74 articles. A quality analysis and comprehensive coverage of multi-sensor depth estimation, including camera, LiDAR, radar, and fusion, were conducted. Nevertheless, this study remains focused only on depth estimation and does not address 3D detection, tracking, V2X cooperative perception, Mamba/LLM models, or adverse weather conditions.

In general, these surveys show significant progress in multi-modal autonomous perception, but still reveal important gaps in the integration of new AI architectures, robustness in real-world environments, the security of autonomous systems, and distributed cooperative fusion. Table I recaps the comparative study.

This article presents a systematic review covering studies between 2022 and 2026. The main subject of this study is the advancement in 3D object detection in autonomous driving using multi-data sensor fusion architecture.

B. Purpose

This systematic review introduces five main objectives:

- identification and categorization of multi-sensor fusion approaches for 3D object detection in autonomous

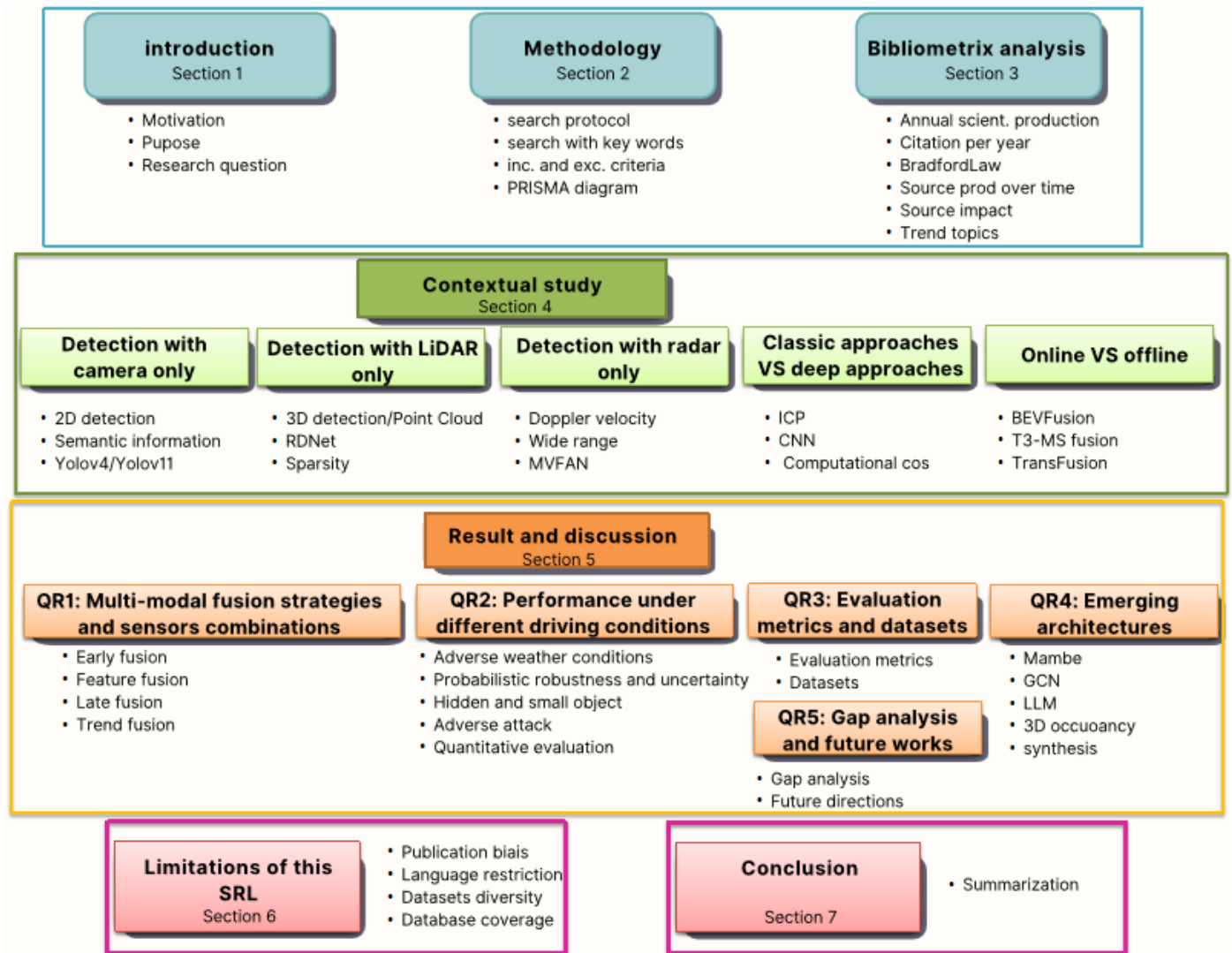


Fig. 1. The structure of review.

driving(2022-2026).

- Comparative analysis of the fusion strategy based on the fusion level and the algorithmic architecture.
- Dataset and metrics analysis based on previous studies.
- Identification research gaps and concrete directions proposition for future work.
- New architecture emergence synthesizes.

C. Research Questions

Based on a critical analysis of existing surveys and the identified gaps, we define five research questions that structure this entire review. These questions specifically target areas that are not covered or are insufficiently addressed in the existing literature. Table II represents the research questions.

In the rest of this review, articles search and selection methodology is presented in Section II. The Bibliometrix

analysis is described in Section III. Contextual study is provided in Section IV. Discussion based on research questions is described in Section V. The limitations of this review is provided in Section VI. The conclusion is described in Section VII. The structure of this review is designed in Fig. 1.

II. METHODOLOGY

A. Search Protocol

This review is guided by the PRISMA protocol (Preferred Reporting Items for Systematic Reviews and Meta-Analyses). The research strategy is based on two complementary databases from Scopus and web of sciences(wos). In fact, PRISMA protocol provides a transparent, rigorous, and reproducible search strategy. It allows an evident description of all steps of the research process, from the identification of studies to their final selection specifying the databases consulted, the keywords used and the inclusion and exclusion criteria as demonstrating in the following sections.

B. Search Key Words and Filter on Databases

Detailed research on the field needs to carefully choose the key words in the search step. The key words used in the article search are described below:

- For Scopus: “multi-fusion” & “multi sensor” & “3D object detection” & “autonomous vehicles”
- For WoS: “multi sensor” & “3D object detection” & “autonomous vehicles”

This research finalized by finding 318 articles from Scopus and 155 articles from Web of Sciences(WoS). Then, in the Scopus website, We firstly filtered with “ year of publication (2022-2026)”, and the number of articles is decreased to 430.

C. Inclusion and Exclusion Criteria

Identifying inclusion and exclusion criteria is a crucial step in a systematic review. It defines which studies should be retained or excluded according to the research objectives. These criteria ensure a consistent and objective selection process, reduce potential bias, and improve the reliability of the results. Inclusion criteria are used to retain studies that are relevant to the research questions, whereas exclusion criteria eliminate studies that fall outside the scope of the review, such as incomplete or duplicate studies.

The inclusion and exclusion criteria used in this study are presented in Table III. A total of 26 duplicate articles were removed using RStudio, and 187 papers were then excluded that are not open access leaving 154 articles. Biblioshiny was then used to filter the articles by language, and 15 articles published in languages other than English were excluded. Finally, a comprehensive review of the remaining articles was conducted, resulting in 110 articles begin retained for the final analysis. The selected studies focus on 3D object detection or tracking in autonomous driving and provide quantitative results for multi-sensor detection or tracking.

D. PRISMA Diagram

The PRISMA diagram is very important. In fact, It outlines the various steps followed by the researcher, such as the identification of references, the removal of duplicates, the screening of titles and abstracts, the assessment of full-text article eligibility, and the final inclusion of the selected papers. Fig. 2 presents the PRISMA graph.

III. BIBLIOMETRIX ANALYSIS

Bibliometrix analysis is done thanks to the R-studio and biblioshiny command. The following subsections describe the analysis of the collected articles.

A. Annual Scientific Production

From 2023, the number of articles increase rapidly thanks to the BEVFusion paper, which introduces the fusion in BEV space as a dominant paradigm. The year 2025 presents the pic articles with more than 45 articles. Fig. 3 demonstrates the evolution of the number of articles per year.

B. Citation Per Year

The average number of citations decreases over time, which is expected given the different publication dates of the studies. Highly cited papers are mainly review articles and influential approaches, such as BEVFusion and TransFusion. In contrast, articles published in 2025 and 2026 have had less time to accumulate citations, which does not diminish their scientific contributions. Fig. 4 presents the average number of citations per year, with a notable increase in 2022 that coincides with the emergence of influential BEV-based fusion architectures such as BEVFusion.

C. Bradford Law

Based on Bradford law, the core of zone 1 presents mainly the dominant sources : Sensors (MDPI), IEEE Transactions on Intelligent transportation systems, IEEE Access, IEEE Robotics and Automation Letters, drones, world electric vehicle journal, and IEEE Transactions on Intelligent Vehicles. Fig. 5 demonstrates the classification zones based on BradfordLaw and the most relevant sources based on document number.

D. Source Production Over Time

Fig. 6 shows that Sensors is the dominant source producer from 2022 to 2026. On the other hand, the second dominant source is variable each year. Between 2022 and 2024, IEEE Robotics and Automation Letters wins to retain the second producer place, but in 2025 and 2026 it is back to the fourth producer. The attractive evolution is from IEEE Transactions on Intelligent transportation systems. In fact, it is translated from the fourth source with less than five articles in 2024 to the second one in 2026 with ten articles. In the same context, the world electric vehicle journal begins its way to produce articles on this field after 2023.

E. Source Impact

To improve the importance of the source in our search and based on Bibliometrix tools, we calculate the Local H-index of each source. The local impact H-index means that “x” articles from this source in the database(Scopus+WoS database) are cited at least “x” times on the studied articles(in the concatenate database).

Fig. 7 shows that the source “Sensors” is the source with the highest H-index impact that is equal to 6. As local impact H-index , 6 is an important value that improves the importance of these sources.

F. Trend Topics

Fig. 8 shows that the concepts 3D object detection and autonomous driving coincide in 2025 with its higher frequencies. In fact, it illustrates the novelty of the theme.

G. Summary

The previous subsections allow us to conclude the importance and tendance of our review. In fact, statics from the Bibliometrix analysis improve that the 3D object detection is a trendy subject and define the most important source in this field. As a conclusion, this Bibliometrix analysis helps us to

TABLE III. INCLUSION AND EXCLUSION CRITERIA

Criteria	Inclusion	Exclusion
Theme	3D object detection or tracking in autonomous driving	Map construction or trajectory planning
Language	Only English	Any other language
Date range	Published between 2022 and 2026	Before 2022
Source type	Peer-reviewed conference or journal articles	Non-peer-reviewed sources
Relevance	Quantitative results about multi-sensor detection or tracking	Articles using mono-modality (not critical)

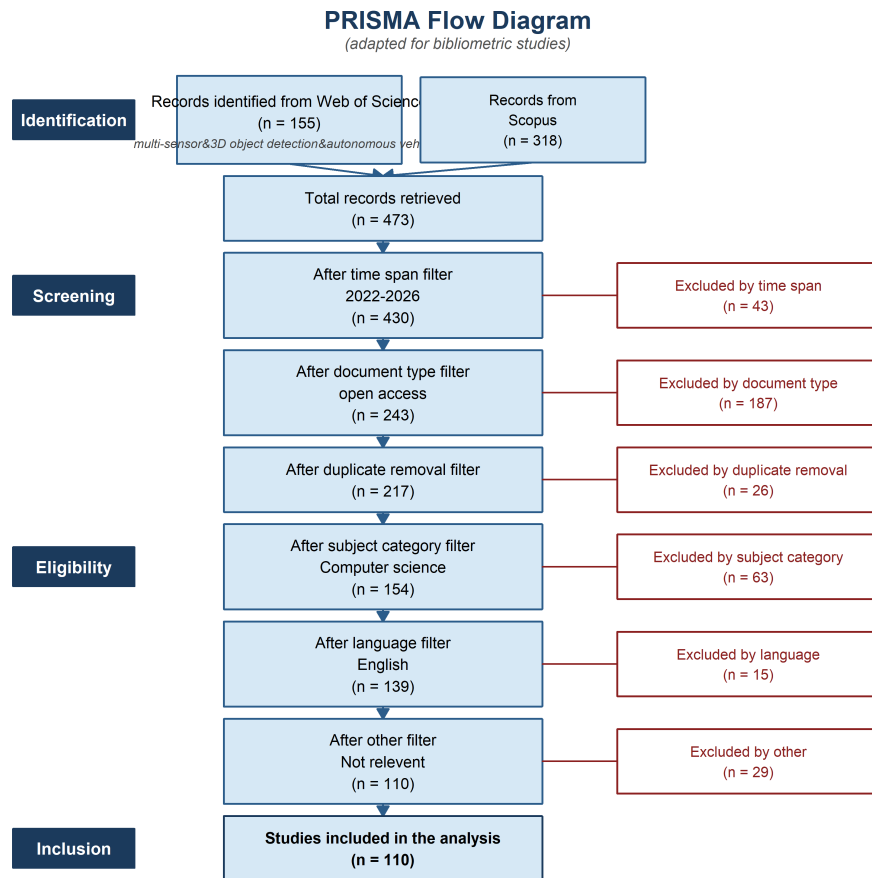


Fig. 2. PRISMA chart.

reflect on some notions related to 3D object detection theme which are defined in the research questions. In addition, it provides a look to the most important source of articles to enhance our study.

IV. CONTEXTUAL STUDY

The studies of the literature are interested in multi-sensor data fusion and depth estimation in autonomous driving. The comprehensive lecture of articles in the systematic review research allows us to conclude the the brilliance of the architectures such as transformer, Mamba, V2X and others. In addition, BEV is found in the literature as the basic of perception system and 3D detection. As a consequence, our research is

dedicated to improve how these emergent architectures solve the BEV shortcomings. To satisfy this search, we need to start from the basic of multi-sensor fusion. The new approaches and architectures start from the mono-modality with classic architectures. The following sections will present the evolution of perception tools and approaches.

A. Detection Using Camera Only

In production, monocular methods are the most useful because they are the cheap ones. A 2D detection with high precision in real time is satisfied thanks to YOLOv4 [30] and YOLOv11 [31] architectures. However, 3D detection from the camera presents only a fundamental challenge. In fact,

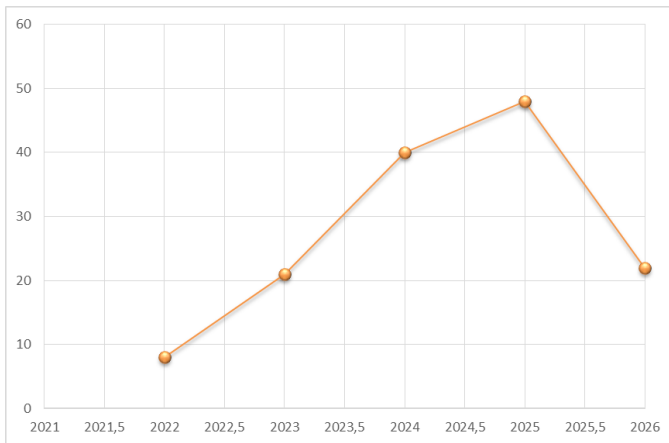


Fig. 3. Annual scientific production.

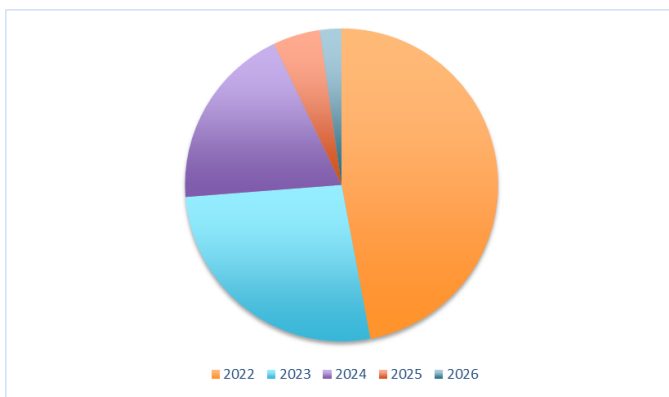


Fig. 4. Average citations per year.

numerous 3D scenes can coincide with the same 2D projection. To enhance the performance of detection with the camera only, studies propose exploring polar representation [33] and depth guidance [32]. Stable diffusion is used to outperform nocturnal images [34], the 3D object annotation via pseudo-LiDAR and the VLM validation [35] guide to an efficient annotation without LiDAR.

Even with all improvements, the monocular camera introduces several problems: depth, light and weather conditions, and insufficient cross-domain generalization for advanced levels of autonomous driving [10], [23]. The presented limits make the fusion with other modals crucial.

B. Detection Using LiDAR

Due to its capacity to directly measure the time-of-flight laser generating point clouds with high resolution and unchangeable under light conditions, LiDAR is presented as a reference to 3D detection [5], [6]. The approach RDNet [36] introduces a Rotate-ground-truth augmentation that allows to correct the unbalance of classes annotation (pedestrians are underrepresented) combined with a decoupled attention head to enhance the performance on KITTI and nuScenes datasets. Transformers in Multi-Feature Fusion Hierarchical-GCN (MUFFIN-HGCN) [37] allow pixel-point mapping without extrinsic parameters, combined with GCN hierarchic to

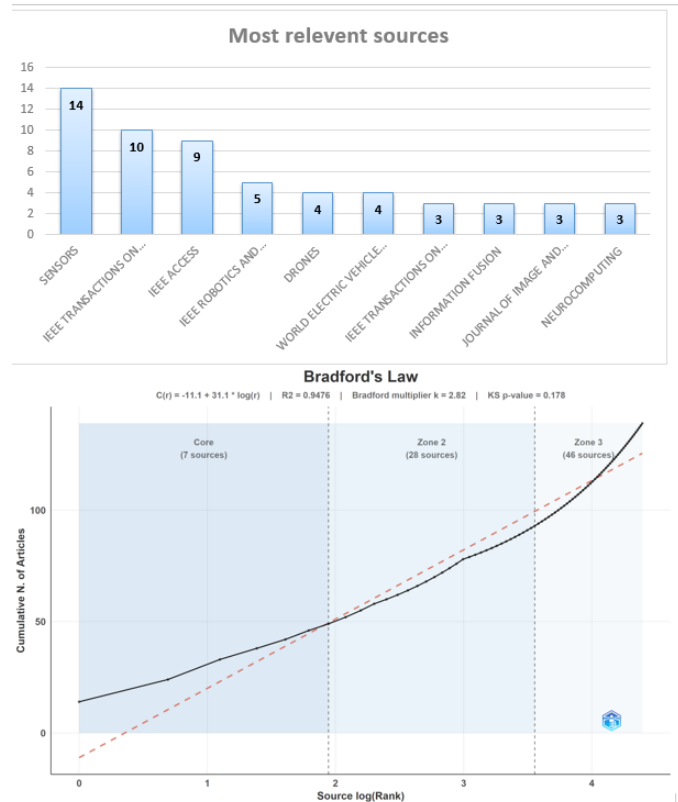


Fig. 5. BradfordLaw and most relevant sources.

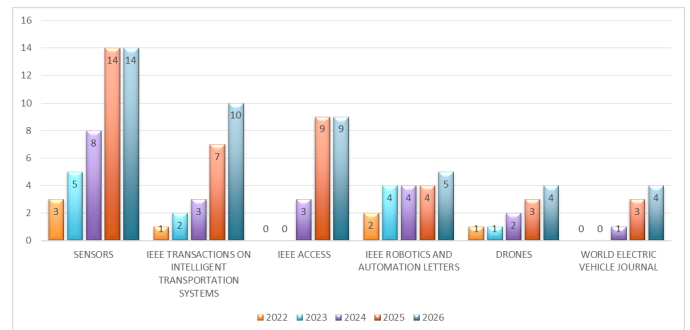


Fig. 6. Source dynamics.

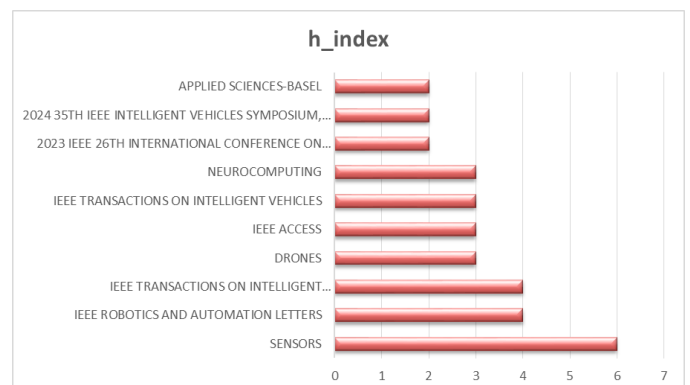


Fig. 7. Source impact.

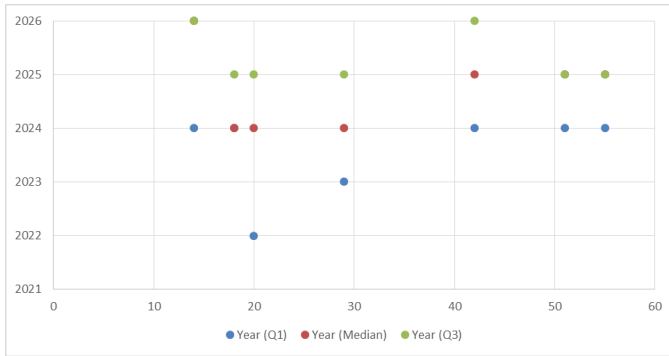


Fig. 8. Trend topics.

capture the non-Euclidean structure of point clouds, surpassing standard voxelizations. In real world deployment, LiDAR presents several deficiencies, including degradation under heavy precipitation (about 80% returns lost in dense fog) [42], sparsity and lack of semantic information for fine classification [5], [6].

C. Detection Using Radar Only

The unique physic complementarity is represented by the Millimeter radar. In fact, it does not degrade under intense precipitation, Doppler velocity measurement and provides a wide range of about 200-250m [8]. The 4D mmWave radar introduces the vertical dimension to provide a real spatial dimension. The SMFormer [38] approach uses multi-view transformers for 3D detection with 4D radar. The Range-Angle-Doppler Net(RADNet) [39] method treats the 3D Range-Angle-Doppler tensor, avoiding information loss due to conversion to point clouds. Multi-View Feature Assisted Network for 4D Radar Object Detection (MVFAN) [13] introduces a multi-view architecture to enhance 3D detection using only 4D radar. Radar alone is still significantly inferior to LiDAR in spatial resolution, suffers from important speckle noise, and to false positives due to multi-path detection in dense urban environments. Precise object classification remains difficult without additional visual information [8], [12].

D. Classic Approaches VS Deep Learning Approaches

Classic approaches such as stereo triangulation, ICP for LiDAR, etc., are interpretable and do not need labeled data, but they are noise-sensing, not very generalizable, and assume quasi-static conditions [8], [9]. Exploring large annotated corpora allows deep learning to change the 3D detection. CNNs, Transformers, and State-Space Models (Mamba) are the best for autonomous driving under various conditions, but a higher computational cost is the price of this enhancement [14], [15], [37]. Table IV represents the comparative table of sensor performance for 3D object detection tested on KITTI and evaluated with the metric Average precision, except the combination radar+camera due to the absence of radar in KITTI dataset.

E. Real-Time and Offline

The real-time and offline distinctions are crucial for deployment. Real-time models should treat each frame in less than

50 ms in embedded calculators with energy vulnerabilities. They use light backbones such as MobileNet, PointPillars, and light ResNet-18, with quantification and pruning. T-3MS fusion [40] improves that an economical VLP-32C LiDAR and a 360° camera achieves state of the art without increased test-time. 360° DAMOT [41] marks energy efficiency with fast encoders and a portable Docker/ROS architecture. Offline models such as BEVFusion, TransFusion, and CoMCM [42] prioritize maximum accuracy for algorithmic validation but are computationally prohibitive in production. Tran et al. [43] propose T360Fusion, a multimodal 360° temporal fusion method based on Transformers that aims for a balance between accuracy and latency, while Hossain [44] applies lightweight multi-sensor fusion to short- and long-range obstacle detection for a delivery robot, illustrating the transferability of these architectures beyond the realm of autonomous road vehicles.

V. RESULTS AND DISCUSSION BASED ON RESEARCH QUESTIONS

This section answers the five research questions defined from the gaps identified in Section I-C. Each subsection provides a structured, quantified, and critical analysis of the 110 articles.

A. QRI: Multi-Modal Fusion Strategies and Sensors Combinations

Fusion approaches categorize in four classes based on the integration level. Table V synthesizes the characteristics, advantages, and limitations of each method.

1) *Early fusion*: Early fusion combines raw data directly. PixelFusion [45] introduces a pixel fusion by projecting the point clouds of LiDAR on the image plan, allowing integration of fine information(texture/color) before the voxelization step. ExistenceMap-PointPillars [46] integrates pseudo-cards to PointPillars, improving robustness when data degradation. The Segmentation Guided Fusion Network(SGFNet) [65] uses a foreground segmentation head to guide the projection of the pixel-point, avoiding LiDAR sparsity artifacts. Zhao et al. [47] introduce a multi-scale voxel fusion module specifically optimized for small objects, achieving 73.81% AP_{3D} hard cars and 48.03% AP_{3D} hard pedestrians on KITTI. Other works refine the pixel-point fusion process itself: Tian et al. [48] employ singular value decomposition (SVD) to geometrically align LiDAR point clouds and camera images; Zuo et al. [49] propose a supervised cross-fusion method that reduces the need for dense annotations; and Yildiz et al. [50] combine real-time detection with semantic distance measurement using camera-LiDAR fusion. Chen et al. [51] propose CFFCNet, an object-center-guided feature completion network designed for more accurate localization and dimension estimation from LiDAR point clouds. Fig. 9 represents the process of early fusion methods.

2) *BEV feature fusion*: since 2022, the fusion of BEV features has been the dominant paradigm [14], [15]. In fact, BEVFusion combines camera and LiDAR characteristics in the same BEV space without depth problems. AttBEV [18] outperforms BEVFusion with the channel-spatial attention module CBAM, obtaining NDS = 67.95% (+2.63%) on nuScenes. MV2DFusion [16] generates independent semantic queries

TABLE IV. COMPARATIVE TABLE OF SENSOR PERFORMANCE FOR 3D OBJECT DETECTION

Modal	Advantages	Limits	Performance (tested on KITTI)
Camera	High resolution, rich semantic information and texture, and low cost	Depth errors, absence of direct 3D detection, and sensitivity to lighting and weather conditions	~70–80% AP
LiDAR	3D precision and robust under lighting conditions	Degradation under precipitation, sparsity, and high cost	~85–92% AP
4D Radar	Wide range, direct measurement of Doppler velocity, and robustness under adverse weather conditions	Low angular resolution and multipath noise	~40–60% AP
LiDAR + Camera (BEV fusion)	Geometric and semantic complementarity	Critical calibration, high computational cost, and need for temporal synchronization	~88–95% AP
Camera + Radar	Robustness under adverse conditions, semantic information, and Doppler velocity measurement	Modest 3D precision	~55–70% NDS
LiDAR + Radar	Geometric complementarity and robustness under lighting and weather conditions	Pipeline is not yet standardized and datasets are scarce	~87–92% AP

TABLE V. COMPARISON OF MULTIMODAL FUSION APPROACHES FOR 3D OBJECT DETECTION

Approach	Description	Advantages	Limitations	Examples (Refs.)
Early fusion	Fusion of raw sensor data before feature extraction	Rich multimodal information with low-level details preserved	Requires precise calibration and involves high computational cost	[45], [46], [47], [83], [82]
Feature fusion	Fusion of extracted features in a shared BEV space	Good performance–efficiency trade-off and improved robustness to miscalibration	Heterogeneous feature alignment and increased training complexity	[16], [17], [18], [59], [60], [66], [53], [67], [84], [85], [86], [87]
Late fusion	Combination of the outputs of independent models	Modularity, fault tolerance, and flexibility across different domains	Limited modeling of fine-grained cross-modal correlations	[61], [62], [63], [64], [88], [89], [90], [91]
Deep hybrid	Multi-level feature integration with adaptive fusion	High detection accuracy, context awareness, and adaptive multi-modal integration	Requires substantial annotated training data and high computational resources	[10], [42], [12], [16], [20], [92], [93], [103]

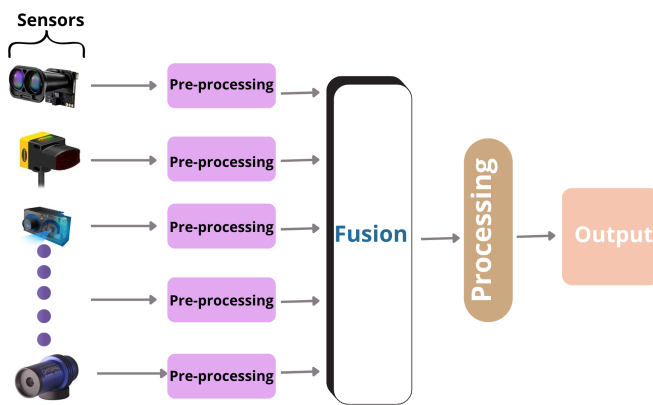


Fig. 9. Early fusion approach.

from images and point clouds, then combines them sparingly via cross-attention, achieving state-of-the-art performance on nuScenes and Argoverse2 [52]. Xu et al. [17] propose a Frustum-Aware Fusion Network with cross-attention, and a multi-modal alignment module providing 90,88% /84,47% /80,99% mAP_{3D} respectively for cars/pedestrians/cyclists on KITTI. Li et al. [66] provide a Deformable Cross-Modal Spatial Attention with learnable offsets, robust for calibration

errors via the Height-Adaptive Aggregation strategy. BEV is enhanced by ECL3D [53] through the combination between Depth-Semantic Feature Enhancement (DSE) and Edge-guided Channel-Spatial Enhancement (ECSE) as part of Lift-Splat-Shoot without additional inference cost. RM²Occ[54] extends the BEV paradigm towards semantic 3D voxel occupancy prediction with bidirectional (from 2D to 3D and vice versa) reprojection and integration of the Inertial Measurement Unit(IMU), more suitable for trajectory planning. In the same context, Li et al.[55] propose Panoptic-VSNet, which leverages prior semantic knowledge to guide multimodal 3D panoptic segmentation, combining instance detection and dense scene understanding into a single BEV pipeline. Based on a similar principle of selectivity, Jacobson et al. [56] propose Center Feature Fusion, which first identifies relevant object locations in camera and LiDAR streams before fusing only a limited subset of associated features in BEV, thereby reducing computational cost compared to approaches that project the entire image. Liu et al.[57] enhance detection robustness in complex traffic through unified BEV feature alignment, while Wang et al. [58] propose TF-CMFA, which combines temporal fusion and cross-modal alignment for robust 3D detection in dynamic environments. The feature level fusion is described in Fig. 10.

3) *Late and cooperative fusion*: LCF3D [61] introduces a late cascade fusion (late + cascade) that provides two main functionalities: *Filtering the false positives of LiDAR*

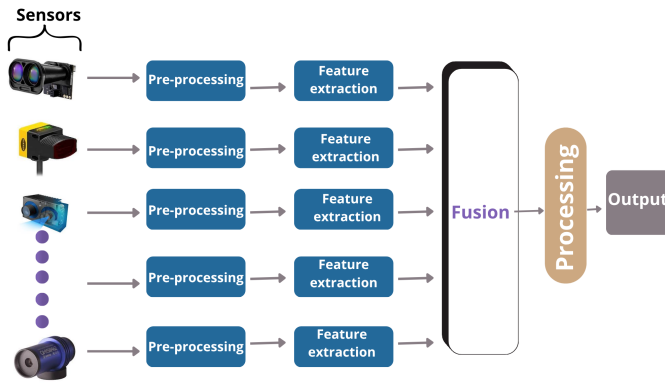


Fig. 10. Feature level fusion approach.

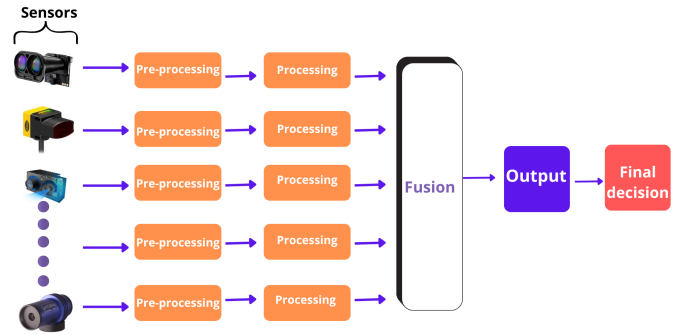


Fig. 11. Late fusion approach.

by the association of the 2D image camera detection and from unmatched 2D detections, the 3D frustums retrieve of missed objects. Its main objective is cross-domain. In fact, it allows for different configurations for sensors in training and testing. Without access to internal architectures, Fadili et al. [66] provide multi-source black-box cooperation that enables collaboration between agents from different manufacturers (crucial conditions for real deployment). CLF-BEV SORT [62] uses a Kalman filter for 3D multi-object tracking by combining camera and LiDAR in BEV. It offers a lightweight alternative to intensive deep learning methods. Naveen et al. [97] combine multiple sensors using deep learning for the onboard perception of autonomous electric vehicles. Dheepika and Umadevi [98] propose a hybrid approach combining 3D detection and tracking within a single deep learning pipeline. Regarding multi-object tracking, Cho and Kim [107] combine multi-positive contrastive learning with deep reinforcement learning for 3D LiDAR tracking, while Sochaniwsky et al. [108] employ a camera-LiDAR fusion framework designed for road traffic monitoring. These architectures also find applications in related fields: Liu et al. [109] and Bultmann et al. [110] apply similar multi-sensor fusion principles to drone landing and semantic segmentation; Rho et al. [105] to underwater 3D reconstruction; and Shang et al. [111] to obstacle avoidance for autonomous vehicles, demonstrating the broader transferability of these fusion techniques beyond the road traffic context alone. Yan et al. [100] enhance cooperative vehicle-infrastructure perception using enriched point cloud features and dedicated spatial attention, while Zimmer et al. [101] propose InfraDet3D, a multimodal detection architecture leveraging cameras and LiDAR mounted on road infrastructure. Yuan et al. [104] propose a keypoint-based deep feature fusion method for cooperative vehicle detection, serving as a foundational reference for this type of approach. The late fusion process is demonstrated in Fig. 11.

4) *Trend fusion: Camera-radar:* Since 2023, the combination of camera and radar has experienced significant growth. In fact, it is an economical and robust alternative [8], [64], [68], [69], [70], [71]. With spatial information and radar speed, CR3DT [64] extends BEV detection and camera only methods. It provides +5,3% mAP and +14,9% average multi object tracking accuracy (AMOTA) in nuScenes. To improve detection and tracking in bad weather, NeXtFusion [68] uses

attention mechanisms. The Dual-branch network allows to design interactions between camera and radar with point cloud offset compensation. Xiao et al. [70] combine perspective view fusion for depth estimation and BEV fusion for final detection. MLF-4DRCNet [71] introduces a 4D radar + camera multi-level fusion with cross-attention on nuScenes. The LiDAR and radar fusion is possible and offers a robust system under adverse conditions and a better range as demonstrated in [72] but it lacks semantic information. Qi et al. [112] proposes FusionBev, an architecture that also combines LiDAR and 4D radar in BEV space, confirming the growing relevance of this sensor combination for robust 3D detection. Li et al. [99], for their part, guide radar vision fusion using multi-scale region proposals, whereas Ruddat et al. [102] specifically address the detection of vulnerable road users through fusion that is robust to 4D radar sparsity. In the same context, Khatab et al. [106] specifically evaluate the detection of vulnerable objects using a dedicated multi-sensor system.

B. QR2: Performance Under Different Driving Conditions

1) *Adverse weather conditions:* To face adverse weather conditions, ContextualFusion [73] proposes a fusion with convolutional gate. As a result, ContextualFusion performs better than the standard BEVFusion providing +11,7% mAP at night and +8,3% in rain in nuScenes. Depending on detected weather conditions, SAMFusion [74] proposes to dynamically weight the modalities. The adaptive sensor fusion demonstrates significant gains in rain, fog, and night. Tian et al. [75] demonstrate that the LiDAR-radar combination with voxel augmentation and BEV synchronization significantly improves 3D detection in dense fog in Oxford Radar RobotCar. HRFuser [76] is evaluated on nuScenes DENSE (snow, fog, rain), demonstrating the robustness of multi-resolution radar fusion in nighttime conditions. Fast All-day [77] fuses LiDAR and a visible camera for detection in all weather conditions. More recently, Wang et al. [113] proposed SWG-Fusion, which leverages a vision-language model (VLM) to dynamically guide multimodal fusion in BEV based on the severity of detected weather conditions, illustrating the growing integration of language models into mechanisms for robustness against adverse conditions.

2) *Probabilistic robustness and uncertainty*: An adaptive uncertainty quantification for the Kalman filter is proposed by Ulreich et al. [78]. In fact, the proposed method allows to dynamically estimating the noise measure based on the condition, making the tracking more robust face to sensor degradation. Yang et al. [79] introduce Evidential Fusion based on Dempster-Shafer theory, separately treating epistemic and random uncertainty for a dynamic weighting of modalities according to their instantaneous reliability.

3) *Hidden and small objects*: The detection of hidden and small objects is a major and crucial challenge. In fact, the extraction of LiDAR features in probable occlusion regions from 3D frustums defined by 2D detections is guided by a Spatial Attention Frustum [80]. Adaptive Multiscale Attention Feature Aggregation is developed with contextual priors in occlusion regions [81]. Zhao et al. [47] target small objects thanks to a multi-scale voxel fusion module with three granularities.

4) *Adverse attack*: El-Fatyany [94] identifies a fundamental structural vulnerability: multi-sensor fusion pipelines can be manipulated by physical patch attacks inserted into the camera modality, deceiving the 3D detector despite an intact LiDAR. The proposed solution based on Unet segmentation is a first step to face this type of attack.

This gap constitutes an unquantified operational risk for the public deployment of autonomous vehicles.

5) *Quantitative evaluation*: Table VI demonstrates the quantitative evaluation of the methods of the articles in this state-of-the-art.

Entries are grouped by dataset and that cross-row comparisons across different metrics/datasets are illustrative only, not statistically equivalent.

C. QR3: Evaluation Metrics and Datasets

Evaluation metrics and dataset selection are a crucial step in algorithm evaluation and comparison study between different methods

1) *Evaluation metrics*: Metrics allows a comparative study between different systems. The most useful metrics are:

- mean Average Precision (mAP): the main metric for 3D detection, it represents the average precision for all classes and thresholds IoU (0.5 for pedestrians and 0.7 for cars in KITTI dataset)
- (nuScenes Detection Score (NDS): the standard of leaderboard nuScenes, it is a composite score combining the mAP with five kinematic attributes.
- Higher Order Tracking Accuracy (HOTA): detection, association and localization are the three fundamentals components of object tracking, the HOTA score is designed to balance them.
- Average Multiple Object Tracking Accuracy (AMOTA): the standard metric of nuScenes for the multiple object tracking dataset.
- AP_BEV/AP_3D: precision in Bird's Eye View and in 3D on KITTI with IoU threshold 0.5/0.7.

- Intersection over Union (IoU): the main metric used in computer vision to measure the overlap between two geometric regions. The IoU is used to evaluate the accuracy of object detection, image segmentation, and object tracking models.
- Identity Switches (IDS): represents the number of identity switches indicating the quality of temporal association in tracking.

2) *Datasets*: Datasets are crucial for autonomous driving systems. Its variety allows to test the systems in different conditions, to evaluate its performance and to compare between different modals. Each dataset has its own characteristics such as the equipments of the vehicle (sensors), scenes, classes (pedestrians, cars, roads,...). Table VII represents an overview of the datasets used in multi-sensor fusion.

D. QR4: Emerging Architectures

Each modality is encoded separately using a CNN or Transformer, and the resulting features are projected into a shared BEV space and fused through concatenation or cross-attention. This represents the dominant paradigm for multi-sensor fusion in 3D detection. Nevertheless, this paradigm has several limitations when moving beyond standard conditions, particularly in non-Euclidean point clouds, multi-vehicle co-operation, complex semantic contexts, and direct trajectory planning.

To address these limitations, emerging architectures, including Mamba, GCNs, LLMs, and 3D occupancy-based models, have been developed. This section identifies the limitations of conventional architectures and discusses how emerging architectures address these challenges. Table VIII summarizes the main limitations and the corresponding architectures proposed to address them.

1) *From local fusion to sequential cooperative fusion: Mamba/SSM*:

a) *The addressed problems*: Cooperative multi-vehicle perception requires combining BEV representations from multiple moving agents. The representations are captured at marginally different timestamps, with point cloud densities that vary based on each vehicle's relative position. Transformers address this fusion through cross-attention between the agents' BEV representations. However, two structural issues render this approach suboptimal in this precise context.

- first problem: The quadratic complexity $O(n^2)$ of self-attention: In multi-vehicle, the sequence of BEV tokens for N vehicles grows linearly with N, while memory and computation time grow quadratically.
- Second problem: The non-causal aspect of attention: Each vehicle transmits its BEV state at regular intervals. Standard attention treats all tokens as temporally equivalent, ignoring the causal structure of this successive communication. The context transmitted by the vehicle at time $t-1$ should influence the interpretation of the transmission at time t , yet attention does not model this type of sequential dependency.

To resume, the perception of V2V is based on temporal state of BEV of N agents. Transformers ignored sequential

TABLE VI. QUANTITATIVE EVALUATION OF MULTIMODAL 3D OBJECT DETECTION METHODS

Ref.	Modalities	Datasets	Results	Test conditions
[17]	LiDAR + Camera	KITTI	90.88%, 84.47%, and 80.99% AP _{3D} for the Car category	Easy, Moderate, and Hard settings
[17]	LiDAR + Camera	nuScenes	89.2% and 64.9% AP _{3D} for Pedestrian and Bicycle detection	Moderate setting
[18]	LiDAR + Camera	nuScenes	67.95% NDS	Standard evaluation
[16]	LiDAR + Camera	nuScenes	76.7% NDS and 74.5% mAP	Standard evaluation
[75]	LiDAR + Radar	Oxford RobotCar	Accuracy rates of 93.2%, 92.5%, and 88.7% at IoU levels of 0.5, 0.65, and 0.8, respectively	Foggy conditions
[74]	Camera + LiDAR + Radar	Seeing Through Fog	Improvement of 17.2 AP points for pedestrians in dense fog and 15.62 AP points in heavy snow at long range	Adverse weather conditions
[73]	LiDAR + Camera	AdverseOp3D	6.2% mAP improvement over state-of-the-art methods	Adverse weather conditions (rain and night)
[64]	Camera + Radar	nuScenes	5.3% mAP improvement and 14.9% AMOTA improvement	Standard evaluation
[61]	LiDAR + Camera	KITTI / nuScenes	Best or second-best performance among the evaluated methods	Different evaluation parameters
[40]	LiDAR + Camera	nuScenes	74.9% NDS and 72.8% mAP	Class imbalance
[47]	LiDAR + Camera	KITTI	73.81% 3D AP for cars and 48.03% 3D AP for persons	Hard level for cars and persons (small objects)
[42]	LiDAR V2V	OPV2V	0.704 AP@0.5 and 0.495 AP@0.7	Cooperative perception
[10]	LiDAR + Camera	OPV2V and DAIR-V2X	State-of-the-art performance	Heterogeneous cooperative perception tasks
[20]	LiDAR-based single vehicle + infrastructure camera	V2XSet-V2I	0.292 improvement in AP@0.5 and 0.289 improvement in AP@0.7	Infrastructure and collaborating vehicles

TABLE VII. COMPARISON OF DATASETS USED FOR MULTIMODAL 3D OBJECT DETECTION

Dataset	Year	Scenes	Annotated frames	Classes	Sensors
nuScenes	2019	1000	40K	23	Cameras + LiDAR + Radar
KITTI	2012	50	15K	8	Cameras + LiDAR
AdverseOp3D	2024	100	4K	23	Cameras + LiDAR + Radar
VoD	2022	~8700	~8700	13	Camera + Radar + LiDAR
K-Radar	2022	62	35K	5	Camera + LiDAR + Radar
Oxford Radar RobotCar	2019	32	240K	Focus on zone geometry	Camera + LiDAR + Radar + GPS/INS
Seeing Through Fog (nuScenes DENSE)	2020	>10K km recorded in real test scenarios	1600	4	Camera + LiDAR + Radar
Waymo Open Dataset	2019	1150	12.5M	4	Camera + LiDAR
OPV2V	2021	73	12K	Vehicles	Camera + LiDAR
DAIR-V2X	2022	Not reported	71,254	15	Camera + LiDAR

TABLE VIII. LIMITATIONS OF BEV TRANSFORMERS AND CORRESPONDING ARCHITECTURES FOR MULTI-SENSOR FUSION

BEV Transformer Limitations	Effects on Multi-Sensor Fusion	Corresponding Architecture
Quadratic complexity ($O(n^2)$)	High computational cost when processing long BEV sequences and multi-vehicle scenarios	Mamba [42]
Non-causal BEV representation	Agent cooperation and temporal dependencies are not explicitly captured by the attention mechanism	Mamba [42]
Processing on a Euclidean grid	Voxelization can lose fine-grained spatial relationships in non-Euclidean point clouds	GCN [37]
Limited semantic context fusion	Limited ability to reason about scene-level semantic context and assess interactions between heterogeneous agents	LLM [12]
Discrete representation	Limited representation of the continuous space between objects	3D Occupancy [54]

causal structure and suffer from quadratic complexity, making them very expensive for real-time multi-agent pipelines.

b) How Mamba face to these problems?: Mamba is a selective State-Space Model that models sequences through a segmented continuous-time state equation. Its operation in the context of cooperative fusion relies on two main aspects.

- First, linear complexity ($O(n)$): Due to a parallel scan algorithm, Mamba processes sequences in linear time. For a sequence of BEV tokens of length n , the memory cost is ($O(n)$), rather than ($O(n^2)$). This property makes it possible to perform fusion over long multi-agent BEV sequences for on-board deployment.
- Second, contextual selection: The parameters of the state equation are input-dependent and change according to the content of each BEV token. This property allows Mamba to dynamically select relevant information from previous agents to interpret the current state.

In conclusion, CoMCM [42] proposes two Mamba modules for cooperative fusion. The Contextual Clustering Mamba module scans each agent's BEV tokens in two details: global, the agent's full sequence, and local, semantic clusters within each BEV map. The Collaborative Mamba in the BEV module then merges the hidden states of N agents into a unified cooperative BEV representation via a multi-agent scan. The result is state-of-the-art performance on OPV2V, with inference latency suitable for real-time on-board deployment.

2) From voxelization to non-Euclidean structure: GCN:

a) The addressed problem: A Point cloud is a set of points irregularly distributed of 3D space without regular grid. Before treating it with CNN or Transformer, most of architectures such as PointPillars, VoxelNet, and BEVFusion convert the point clouds in a regular representation. This voxelization resolves the problem of irregularity, but loses fundamental structural information. Standard voxelization erases non-Euclidean spatial relationships between LiDAR points, such as connectivity, local surface properties, and curvature. CNNs and Transformers applied to voxel grids cannot rebuild these lost relationships. Consequently, detection accuracy for small objects and fine contours is critical.

b) How GCNs respond to this problem?: A hierarchical Graph Neural Network clearly builds a graph from the LiDAR point cloud, where each point is a node, and edges encode spatial relationships between neighboring points. Messages passing between nodes enable the aggregation of local information and the computation of feature representations that reflect the actual structure of points cloud. The key element is hierarchy: the hierarchical GCN creates multiple scales of graph representation, ranging from the finest (relationships between individual points) to the coarsest (relationships between object regions). At each level, a graph pooling operator reduces the number of nodes while preserving important structural relationships. This multi-level approach is conceptually similar to multi-scale CNNs, but operates on graph structures rather than grids.

In fact, It outperforms standard voxelization architectures on KITTI and nuScenes, especially for small objects.

3) From geometric fusion to contextual semantic fusion: LLM:

a) The addressed problems: All multi-sensor fusion architectures operate on the fundamental geometric and statistic level. They detect objects by associating visual and geometric features with predefined categories, using supervised learning on a closed dataset. The major limit of this approach is designed in three situations.

- First, contextual reasoning between agents is impossible. In a V2X system where a vehicle shares its detections with a pedestrian on the street, a geometric fusion system cannot reason about the meaning of this configuration: is it a pedestrian about to cross? A cyclist stopping? CNNs and geometric Transformers do not possess semantic understanding of human behavior that can reasoning these scenes.
- Second, open-vocabulary classification is beyond the reach of current systems. In fact, any object not present in the training categories is ignored or misclassified.
- Third, semantic evaluation of risk. It requires a reasoning beyond the simple detection. This needs a causal model of the world that geometric architectures cannot build.

b) How LLM deal with these problems?: The UAV-vehicle is a cooperative framework driven by LLM(GPT-4) proposed by de Curtò & de Zarzà [12]. In fact, the language model merges and interprets multi-modal data from heterogeneous points of view, providing a semantic risk assessment. Although based on simulated data, this study opens up a fundamentally new path for contextual semantic fusion that goes beyond purely geometric approaches.

4) From object detection to spatial understanding: 3D occupancy:

a) The addressed limitations: 3D bounding box detection has been the standard representation for multi-sensor fusion systems since PointPillars and VoxelNet. This representation is standardized, but it presents three major limitations for trajectory planning.

- First, Bounding boxes do not represent free space. For safe trajectory planning, the module needs to know not only the location of the object but also where the navigable space lies.
- Second, Bounding boxes are inappropriate for objects with irregular forms.
- Third, Bounding boxes need a precise temporal association for tracking. The data-to-track association problem must be solved at all times.

This association problem is several under the conditions of occlusion, object crossings, or missing detections.

b) How 3D occupancy face these limitations?: Re-projection Multi-Task Multi-Sensor [54] is the first 3D occupancy network to integrate LiDAR-camera fusion within a multi-task framework. Its architecture comprises three modules:

TABLE IX. SYNTHESIS OF THE IDENTIFIED RESEARCH GAPS

Gap	Description	Impact
Adverse-weather benchmarks	Lack of standardized real-world datasets covering adverse weather conditions (night, rain, fog, etc.)	Critical
Adversarial security of multi-sensor fusion	Only one article investigates adversarial attacks in multi-sensor fusion [25]	Critical
Computational efficiency	Most state-of-the-art methods are too computationally demanding for on-board systems	High
Heterogeneous cross-manufacturer V2X	Multi-agent systems equipped with heterogeneous sensor configurations remain insufficiently studied	High
Multi-sensor 3D occupancy	Few alternatives to bounding-box representations are available for downstream planning	Medium
Cross-domain generalization	Models often exhibit limited transferability to new environments and geographical regions [96]	High
Robust online calibration	Few approaches remain robust under dynamic sensor miscalibration [95]	High
Mamba/LLM	Promising approaches, but their performance has not yet been extensively validated on large-scale real-world datasets	Medium

- A dual-branch encoder is represented in the LiDAR branch via voxelization and in the camera branch via Lift-Splat-Shoot
- A BEV fusion module featuring bidirectional 2D/3D reprojection and IMU data integration for temporal consistency
- A multi-task head that predicts the semantic voxel occupancy grid and standard 3D bounding boxes at the same time

Compared to previous monocular camera occupancy methods, the results demonstrate an improvement in voxel IoU on nuScenes, while maintaining standard 3D detection performance.

5) *Synthesis*: Emerging architectures are fundamentally transforming multi-sensor fusion not through incremental improvements, but through a paradigm shift over four dimensions: the temporal scalability of multi-agent cooperation [42], the topological fidelity of LiDAR encoding [37], contextual semantic scene reasoning [12], and the very nature of the fusion output representation [54]. Each addresses a structural limitation that dominant BEV Transformers cannot resolve without fundamentally adjusting their paradigm.

E. QR5: Gap analysis and future works

1) *Gap analysis*: This study allows us to introduce eight gaps. These gaps define the most promising opportunities for future work and establish the field's priority areas for development. Table IX synthesizes the identified gaps.

2) *Future direction*:

- Build self-calibrating models that adapt to sensor drift in real time.
- real-world datasets with standardized weather conditions suitable for comparison.
- Light onboard architectures.
- Developing robust systems against adversarial attacks.
- Lightweight 3D occupancy.
- Mamba and LLMs for contextual fusion.

VI. LIMITATIONS OF THIS STUDY

This systematic literature review was conducted according to rigorous academic protocols. However, several limitations must be acknowledged for the spirit of transparency and the credibility of the results.

A. Publication Bias

The articles studied are peer-reviewed publications. This excludes unpublished work, industrial technical reports, and preprints from major companies. The excluded works could offer valuable insights into fusion systems under real-world operating conditions.

B. Language Restriction

Only articles in English were considered. Relevant studies published in Chinese, Japanese, or Korean may have been excluded.

C. Datasets Diversity

The predominance of KITTI and nuScenes, which can bias evaluations toward specific geographic and climatic conditions. This can affect driving in rural environments and developing countries.

D. Database Coverage

Although Scopus and WoS are the most comprehensive reference databases, some relevant articles may have been missed due to indexing variations or alternative terminology.

VII. CONCLUSION

This systematic literature review, conducted in accordance with the PRISMA protocol, analyzed 110 articles published between 2022 and 2026, rigorously selected from Scopus and Web of Science.

The results confirm that multi-sensor fusion provides significantly superior accuracy, robustness, and reliability compared to single-sensor approaches. Among the various strategies, feature fusion in BEV space with cross-attention mechanisms or deformable transformers has become the dominant

paradigm. Late fusion introduces a cross-domain generalization advantage, and cooperative V2X hybrid deep fusion opens up significant prospects.

Architecture has been enhanced since 2024. The major architectural contributions are: Mamba state-space models for Bird's Eye View Vehicle to vehicle cooperation, fusion based on probabilistic uncertainty quantification, LLMs for multi-agent semantic fusion, multi-sensor, multi-task 3D occupancy, and heterogeneous cross-modality V2X cooperation. 4D mmWave radar has established itself as an essential third complementary modality.

The identified gaps define a clear research docket:

- A unified adverse weather benchmark based on real world data.
- Adversarial security for multi-sensor fusion.
- Lightweight architecture for embedded deployment.
- Cross-domain generalization.
- Robust online calibration.

These challenges conduct the transition from academic results to safe, certifiable, and large-scale deployed autonomous driving systems.

ACKNOWLEDGMENT

The authors acknowledge that a large language model was used only to correct grammatical and syntactic errors and to improve the fluency and readability of the manuscript. The scientific content and ideas remain those of the authors.

REFERENCES

- [1] A. Sahbani, NARMA-L2 Neuro controller for speed regulation of an intelligent vehicle based on image processing techniques, 21st Saudi Computer Society National Computer Conference, NCC 2018, doi.org/10.1109/NCG.2018.8593173.
- [2] J. Cao, Y. Han, D. Wang, W. Zhao, Y. Zhao, and Q. Hou, "3D object detection method based on BEV feature fusion", *Qiche Gongcheng/Automotive Engineering*, vol.48, no.1, pp.80-90, 2026.
- [3] X. Li, "Transformer-based fusion of LiDAR and polarimetric imaging for autonomous driving", in *Proc. SPIE*, vol.14064, 2026.
- [4] Road traffic injuries, Accessed: September 13, 2026, Available at: <https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries>.
- [5] Y. Wang et al., "Multi-Modal 3D Object Detection in Autonomous Driving: A Survey", in *International Journal of Computer Vision* 131, 2122-2152, 2023, doi: 10.1007/s11263-023-01784-z.
- [6] L. Wang et al., "Multi-Modal 3D Object Detection in Autonomous Driving: A Survey and Taxonomy", in *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 7, pp. 3781-3798, July 2023, doi: 10.1109/TIV.2023.3264658.
- [7] R. Sebai, A. Sahbani and T. Bejaoui, "Robust Multi-Sensor Fusion Architecture Tested on CARLA Sim", 2026 International Conference on Smart Applications, Communications and Networking (SmartNets), Rome, Italy, 2026, pp. 1-6, doi: 10.1109/SmartNets69662.2026.11604736.
- [8] L. Fan et al., "4D mmWave Radar for Autonomous Driving Perception: A Comprehensive Survey", in *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 4, pp. 4606-4620, April 2024, doi: 10.1109/TIV.2024.3380244.
- [9] X. Wang, K. Li and A. Chehri, "Multi-Sensor Fusion Technology for 3D Object Detection in Autonomous Driving: A Review", in *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 2, pp. 1148-1165, Feb. 2024, doi: 10.1109/TITS.2023.3317372.
- [10] Z. Li et al., "CD-HCP: A Heterogeneous Cooperative Perception Framework Utilizing Cross-Modality Dual-Attention", in *IEEE Transactions on Intelligent Transportation Systems*, vol. 27, no. 6, pp. 6787-6799, June 2026, doi: 10.1109/TITS.2026.3670252.
- [11] Y. Zhang et al., "CooPercept: Cooperative Perception for 3D Object Detection of Autonomous Vehicles", in *Drones Journal*, 2024, Volume 8, Issue 6, doi: 10.3390/drones8060228.
- [12] J. De Curtò and I. De Zarzà, "LLM-Powered Cooperative Perception Framework for Mixed UAV-Vehicle Platoons", 2025 10th International Conference on Fog and Mobile Edge Computing (FMEC), Tampa, FL, USA, 2025, pp. 282-289, doi: 10.1109/FMEC65595.2025.11119241.
- [13] Q. Yan, & Y. Wang, "MVFAN: Multi-view Feature Assisted Network for 4D Radar Object Detection", In *Neural Information Processing: 30th International Conference, ICONIP 2023, Changsha, China, November 20-23, 2023, Proceedings, Part IV*, Pages 493 - 511, doi: /10.1007/978-981-99-8070-3_38.
- [14] H. Li et al., "Delving into the Devils of Bird's-eye-view Perception: A Review, Evaluation and Recipe", in *IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE*, VOL. 46, NO. 4, APRIL 2024.
- [15] H. Li, Y. Zhao, J. Zhong, B. Wang, C. Sun and F. Sun, "Delving Into the Secrets of BEV 3D Object Detection in Autonomous Driving: A Comprehensive Survey", in *IEEE Transactions on Intelligent Transportation Systems*, vol. 27, no. 1, pp. 119-144, Jan. 2026, doi: 10.1109/TITS.2025.3624830.
- [16] Z. Wang, Z. Huang, Y. Gao, N. Wang and S. Liu, "MV2DFusion: Leveraging Modality-Specific Object Semantics for Multi-Modal 3D Detection", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 48, no. 1, pp. 609-623, Jan. 2026, doi: 10.1109/T-PAMI.2025.3609348.
- [17] J. Xu, K. Xiao, W. Ji and C. Li, "A Frustum-Aware Fusion Network With Cross-Attention for Multi-Modal 3D Detection," in *IEEE Access*, vol. 13, pp. 200777-200790, 2025, doi: 10.1109/ACCESS.2025.3637156.
- [18] N. Zhang et al., "AttBEV: enhancing multi-modal 3D object detection with CBAM attention in BEVFusion for autonomous driving", in *IEEE robotics and automation letters Journal*, Feb 2025, vol. 11, num. 2, DOI:10.1109/LRA.2025.3641130.
- [19] J. Zhong et al., "Leveraging Temporal Contexts to Enhance Vehicle-Infrastructure Cooperative Perception", 2024 IEEE 27th International Conference on Intelligent Transportation Systems (ITSC), Edmonton, AB, Canada, 2024, pp. 915-922, doi: 10.1109/ITSC58415.2024.10920140.
- [20] X. Zhou, C. Wang, Q. Xie and T. Qiu, "V2I-Coop: Accurate Object Detection for Connected Automated Vehicles at Accident Black Spots With V2I Cross-Modality Cooperation", in *IEEE Transactions on Mobile Computing*, vol. 24, no. 3, pp. 2043-2055, March 2025, doi: 10.1109/TMC.2024.3486758.
- [21] Y. Liao et al., "KITTI-360: A Novel Dataset and Benchmarks for Urban Scene Understanding in 2D and 3D", in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3292-3310, 1 March 2023, doi: 10.1109/TPAMI.2022.3179507.
- [22] Sh. Zhang et al., "Learning-based algorithms with application to urban scene autonomous driving", in *Artificial Life and Robotics Journal*, Volume 28, pages 244-252, 2022, doi: 10.1007/s10015-022-00813-3.
- [23] M. Valverde et al., "A Survey of Deep Learning-Based 3D Object Detection Methods for Autonomous Driving Across Different Sensor Modalities". *Sensors*. 2025; 25(17):5264. <https://doi.org/10.3390/s25175264>.
- [24] P. Shi et al., "Research Progress on Multi-Modal Fusion Object Detection Algorithms for Autonomous Driving: A Review", in *Computers, Materials and Continua Journal*, Volume 83, Issue 3, 19 May 2025, Pages 3877-3917, doi: 10.32604/cmc.2025.063205.
- [25] M. Adnan et al., "Systematic and Comprehensive Review of Clustering and Multi-Target Tracking Techniques for LiDAR Point Clouds in Autonomous Driving Applications", in *Sensors Journal* 23, no. 13: 6119, 2023, doi: 10.3390/s23136119.
- [26] L. Dal'Col et al., "Joint Perception and Prediction for Autonomous Driving: A Survey", in *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 12, pp. 21427-21452, Dec. 2025, doi: 10.1109/TITS.2025.3607743.
- [27] O.A. Fawole et al., "Recent Advances in 3D Object Detection for Self-

- Driving Vehicles: A Survey”, in AI Journal 5, no. 3: 1255-1285, 2024, doi: 10.3390/ai5030061.
- [28] M. Umar et al., “Advances in multi-sensor fusion for depth estimation in autonomous vehicles: A comprehensive survey”, in Advanced Engineering Informatics, September 2026, Article: 104657, Volume 74, Part A, doi: 10.1016/j.aei.2026.104657.
- [29] Aria, M., & Cuccurullo, C. (2017). bibliometrix: An R-tool for comprehensive science mapping analysis. Journal of Informetrics , 11(4), 959-975.
- [30] R. Murendeni et al., “Using a YOLO Deep Learning Algorithm to Improve the Accuracy of 3D Object Detection by Autonomous Vehicles”, in World Electric Vehicle Journal. 2025; 16(1):9,doi:10.3390/wevj16010009.
- [31] M. Simeonov, P. Kamencay and M. Dado, “Evaluating YOLOv11’s Role in Robust Real-Time Object Detection for Autonomous Driving”, 2025 35th International Conference Radioelektronika (RADIOELEKTRONIKA), Czech Republic, 2025, pp. 1-5, doi: 10.1109/RADIOELEKTRONIKA65656.2025.11008422.
- [32] Y. He, “3D Object Detection via Multi-Sensor Fusion and Deep Guidance”, in Applied and Computational Engineering Journal,DOI: 10.54254/2755-2721/2026.GU33954.
- [33] J. Tang, Y. Hou, J. Liu and B. Yu, “PRTF: Polar Space Represented Multi-View 3D Object Detection With Temporal Fusion Enhancement”, in IEEE Transactions on Intelligent Transportation Systems, vol. 27, no. 1, pp. 1538-1549, Jan. 2026, doi: 10.1109/TITS.2025.3633448.
- [34] Y. Zhang et al., “A stable diffusion-based method for low-light image enhancement in autonomous driving”, in Intelligent Journal Machine Learning & Cyber. 17, 238 2026,doi:10.1007/s13042-026-03053-0.
- [35] Ma. Yalong, et al, “Efficient 3D object annotation via vision-derived pseudo-LiDAR and Vision Language Model (VLM) validation”, in Transportation Research Part C: Emerging Technologies, Volume 182, 2026, 105429, ISSN 0968-090X, doi: 10.1016/j.trc.2025.105429.
- [36] Z. Xia et al., “RDNet: Rotate-Groundtruth Augmentation and Decoupled Attention HEAD for 3D Object Detection”, in IEEE Transactions on Big Data, vol. 12, no. 2, pp. 389-402, April 2026, doi: 10.1109/TB-DATA.2025.3624947.
- [37] C. Liu, et al, “MUFFIN-HGCN: Multi-Feature Fusion Hierarchical-GCN for 3D Object Detection”, in IEEE Transactions on Vehicular Technology, vol. 74, no. 9, pp. 13618-13630, Sept. 2025, doi: 10.1109/TVT.2025.3566696.
- [38] W. Shi et al., “SMIFormer: Learning Spatial Feature Representation for 3D Object Detection from 4D Imaging Radar via Multi-View Interactive Transformers”, in Sensors Journal. 2023; 23(23):9429. <https://doi.org/10.3390/s23239429>.
- [39] K. Hu et al., “RADNet: A Radar Detection Network for Target Detection Using 3D Range-Angle-Doppler Tensor”, 2023 IEEE 19th International Conference on Automation Science and Engineering (CASE), Auckland, New Zealand, 2023, pp. 1-6, doi: 10.1109/CASE56687.2023.10260345.
- [40] B. R. Lamichhane et al., “T-3MS Fusion: Gradient-aware Transformer-guided Middle-Fusion”, in Autonomous Intelligent Systems Journal, Volume 6, article number 5, 2026, DOI: 10.1007/s43684-026-00128-1.
- [41] C. Gómez-Huélamo et al., “360° real-time and power-efficient 3D DAMOT for autonomous driving applications”. in Multimed Tools Appl 81, 26915–26940 (2022). doi: 10.1007/s11042-021-11624-2.
- [42] T. Wang et al., “CoMCM: Collaborative 3D Detection With Multiscale Clustering Mamba”, in IEEE Journal of Selected Topics in Signal Processing, vol. 19, no. 8, pp. 1996-2009, Dec. 2025, doi: 10.1109/JSTSP.2025.3650028.
- [43] K.B. Tran et al., “T360Fusion: Temporal 360 Multimodal Fusion for 3D Object Detection via Transformers”, in Sensors Journal 25, no. 16: 4902. doi: 10.3390/s25164902.
- [44] S. Hossain & X. Lin, “Short-range and Long-range Obstacle Detection Method for a Delivery Robot Based on Multi-sensor Fusion”, in Conference: 2023 IEEE Transportation Electrification Conference & Expo (ITEC), 2023, DOI: 10.1109/ITEC55900.2023.10186935.
- [45] Q. Ning et al., “PixelFusion: A Pixel-level Multi-Sensor Fusion Method for 3D Object Detection”, 2024 8th CAA International Conference on Vehicular Control and Intelligence (CVCI), Chongqing, China, 2024, pp. 1-5, doi: 10.1109/CVCI63518.2024.10830053.
- [46] K. Hariya et al., “ExistenceMap-PointPillars: A Multi-Fusion Network for Stable 3D Object Detection with Pseudo 2D Maps”, 2023 IEEE Intelligent Vehicles Symposium (IV), Anchorage, AK, USA, 2023, pp. 1-7, doi: 10.1109/IV55152.2023.10186740.
- [47] Y. Zhao et al., “A Multi-Sensor 3D Detection Method for Small Objects”, in World Electric Vehicle Journal. 2024; 15(5):210,doi:10.3390/wevj15050210.
- [48] K. Tian et al., “Singular Value Decomposition (SVD) Method for LiDAR and Camera Sensor Fusion and Pattern Matching Algorithm”, Sensors 25, no. 13: 3876. <https://doi.org/10.3390/s25133876>.
- [49] C. J. Zuo, C. Y. Gu, Y. K. Guo and X. D. Miao, “Cross-Supervised LiDAR-Camera Fusion for 3D Object Detection,” in IEEE Access, vol. 13, pp. 10447-10458, 2025, doi: 10.1109/ACCESS.2024.3518564.
- [50] A.S Yildiz et al., “Real-Time Object Detection and Distance Measurement Enhanced with Semantic 3D Depth Sensing Using Camera-LiDAR Fusion”, Applied Sciences 15, no. 10: 5543, 2025, doi: 10.3390/app15105543.
- [51] X. Chen et al., “CFFCNet: Center-Guided Feature Fusion Completion for Accurate Vehicle Localization and Dimension Estimation from Lidar Point Clouds”, Remote Sensing 18, no. 1: 39, 2025, doi: 10.3390/rs18010039.
- [52] B. Wilson et al., “Argoverse 2: Next Generation Datasets for Self-Driving”, 2023, DOI: 10.48550/arXiv.2301.00493.
- [53] M. Chen et al., “Multi-Modal BEV Enhancement Fusion for 3D Object Detection in Autonomous Driving”, in IEEE Transactions on Intelligent Transportation Systems, vol. 26, no. 11, pp. 18447-18459, Nov. 2025, doi: 10.1109/TITS.2025.3589581.
- [54] Y. Ren et al., “RM2Occ: Re-Projection Multi-Task Multi-Sensor Fusion for Autonomous Driving 3D Object Detection and Occupancy Perception”, in IEEE Transactions on Intelligent Transportation Systems, vol. 26, no. 11, pp. 20864-20881, Nov. 2025, doi: 10.1109/TITS.2025.3606554.
- [55] X. Li et al., “Panoptic-VSNet: Visual-semantic prior knowledge-driven multimodal 3D panoptic segmentation”, in Pattern Recognition Journal, 2026, Article: 113239, Volume 176, doi: 10.1016/j.patcog.2026.113239.
- [56] P. Jacobson, Y. Zhou, W. Zhan, M. Tomizuka and M. C. Wu, “Center Feature Fusion: Selective Multi-Sensor Fusion of Center-based Objects”, 2023 IEEE International Conference on Robotics and Automation (ICRA), London, United Kingdom, 2023, pp. 8312-8318, doi: 10.1109/ICRA48891.2023.10160616.
- [57] A. Liu et al., “Robust 3D Object Detection in Complex Traffic via Unified Feature Alignment in Bird’s Eye View”, World Electric Vehicle Journal 16, no. 10: 567, 2025, doi: 10.3390/wevj16100567.
- [58] Y. Wang et al., “TF-CMFA: Robust Multimodal 3D Object Detection for Dynamic Environments Using Temporal Fusion and Cross-Modal Alignment”, in IEEE Access, vol. 13, pp. 74821-74832, 2025, doi: 10.1109/ACCESS.2025.3563483.
- [59] S. Sharma et al., “AEPF: Attention-Enabled Point Fusion for 3D Object Detection”, in Sensors Journal, 24, no. 17: 5841, 2024, doi:10.3390/s24175841.
- [60] W. Zhang et al., “MMAF-Net: Multi-view multi-stage adaptive fusion for 3D detection”, in Expert Systems with Applications Journal, Article: 122716, Volume 242, 2024, doi: 10.1016/j.eswa.2023.122716.
- [61] C. Sgaravatti et al., “LCF3D: A Robust Late-Cascade Fusion Framework for 3D Detection”, in Pattern Recognition Journal, volume 175, Article: 113046, doi: 10.1016/j.patcog.2026.113046.
- [62] L. Alfeqy et al., “Kalman Filter-Based Fusion of LiDAR and Camera Data in Bird’s Eye View for Multi-Object Tracking in Autonomous Vehicles”, in Sensors Journal 24, no. 23: 7718, 2024, doi:10.3390/s24237718.
- [63] M. Fadili et al., “A late collaborative perception framework for 3d multi-object and multi-source association and fusion”, 2024, In 2025 9th International Conference on Robotics and Automation Sciences (ICRAS) (pp. 259-266).
- [64] N. Baumann et al., “CR3DT: Camera-RADAR Fusion for 3D Detection and Tracking”, in Conference: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), DOI: 10.1109/IROS58592.2024.10801848.
- [65] Y. Wang et al., “SGFNet: Segmentation Guided Fusion Network for 3D Object Detection”, in IEEE Robotics and Automation Letters, vol. 8, no. 12, pp. 8239-8246, Dec. 2023, doi: 10.1109/LRA.2023.3326697.

- [66] J. Li et al., "Height-Adaptive Deformable Multi-Modal Fusion for 3D Object Detection", in IEEE Access, vol. 13, pp. 52385-52396, 2025, doi: 10.1109/ACCESS.2025.3553372.
- [67] P. Jia et al., "ORSU: An Online Road Scene Understanding Framework with Multi-level Information Fusion and Bird's-Eye View Representation", Volume 9, pages 296-308, 2026, doi: 10.1007/s42154-025-00387-3.
- [68] P. Kalgaonkar et al., "NeXtFusion: Attention-Based Camera-Radar Fusion Network for Improved Three-Dimensional Object Detection and Tracking", in Future Internet 16, no. 4: 114,doi:10.3390/fi16040114.
- [69] Z. Yu, W. Wan, M. Ren, X. Zheng and Z. Fang, "SparseFusion3D: Sparse Sensor Fusion for 3D Object Detection by Radar and Camera in Environmental Perception", in IEEE Transactions on Intelligent Vehicles, vol. 9, no. 1, pp. 1524-1536, Jan. 2024, doi: 10.1109/TIV.2023.3331972.
- [70] Y. Xiao et al., "Radar-Camera Fusion in Perspective View and Bird's Eye View for 3D Object Detection", in Sensors Journal 25, no. 19: 6106, doi: 10.3390/s25196106.
- [71] Y. Wu et al., "MLF-4DRCNet: Multi-Level Fusion with 4D Radar and Camera", in Information Fusion, October 2026, Article: 104377, Volume 134, doi: 10.1016/j.inffus.2026.104377.
- [72] L. Wang et al., "Multi-Modal and Multi-Scale Fusion 3D Object Detection of 4D Radar and LiDAR for Autonomous Driving", in IEEE Transactions on Vehicular Technology, vol. 72, no. 5, pp. 5628-5641, May 2023, doi: 10.1109/TVT.2022.3230265.
- [73] S. Sural et al., "ContextualFusion: Context-Based Multi-Sensor Fusion for 3D Object Detection in Adverse Operating Conditions", in 2024 IEEE Intelligent Vehicles Symposium (IV) June 2-5, 2024, Jeju Island, Korea ,DOI: 10.1109/IV55156.2024.10588584.
- [74] E. Palladin et al., "SAMFusion: Sensor-Adaptive Multimodal Fusion for Adverse Weather", in Computer Vision – 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXI, doi: 10.1007/978-3-031-73030-6_27.
- [75] J. Tian et al., "3D Detection Under Foggy Weather: The Effective Multi-Sensor Fusion of LiDAR and Radar", 2024 21st Annual IEEE International Conference on Sensing, Communication, and Networking (SECON), Phoenix, AZ, USA, 2024, pp. 1-2, doi: 10.1109/SECON64284.2024.10934948.
- [76] T. Brödermann, C. Sakaridis, D. Dai and L. Van Gool, "HRFuser: A Multi-Resolution Sensor Fusion Architecture for 2D Object Detection", 2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 2023, pp. 4159-4166, doi: 10.1109/ITSC57777.2023.10422432.
- [77] L. Xiao et al., "Fast All-day 3D Object Detection Based on Multi-sensor Fusion", 2023 IEEE Conference on Artificial Intelligence (CAI), Santa Clara, CA, USA, 2023, pp. 71-73, doi: 10.1109/CAI54212.2023.00038.
- [78] F. Ulreich et al., "Uncertainty Quantification for Adaptive Measurement Noise in Kalman Filter Based Object Tracking", 2024 Sensor Data Fusion: Trends, Solutions, Applications (SDF), Bonn, Germany, 2024, pp. 1-7, doi: 10.1109/SDF63218.2024.10773809.
- [79] Q. Yang et al. "Uncertainty-Aware Evidential Fusion for Multi-Modal Object Detection in Autonomous Driving", in Drones Journal, 2026, 10, no. 2:130, doi: 10.3390/drones10020130.
- [80] X. He et al., "Spatial Attention Frustum: A 3D Object Detection Method Focusing on Occluded Objects", in Sensors Journal 22, no. 6: 2366, doi:10.3390/s22062366.
- [81] Y. Han et al., "Adaptive Multiscale Attention Feature Aggregation for Multi-Modal 3D Occluded Object Detection", in IET Computer Vision Journal, Volume19, Issue1, 2025, doi: 10.1049/cvi2.70035.
- [82] A. Mahmoud et al., "Dense Voxel Fusion for 3D Object Detection", in 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), DOI: 10.1109/WACV56688.2023.00073.
- [83] M. Abdelfattah et al., "Multi-Modal Streaming 3D Object Detection", in IEEE Robotics and Automation Letters, vol. 8, no. 10, pp. 6163-6170, Oct. 2023, doi: 10.1109/LRA.2023.3303696.
- [84] E. S. Honorato and D. F. Wolf, "Enhancing 3D Object Detection in Autonomous Vehicles: Multi-Sensor Fusion with Attention Mechanisms", 2024 Latin American Robotics Symposium (LARS), Arequipa, Peru, 2024, pp. 1-6, doi: 10.1109/LARS64411.2024.10786434.
- [85] H. Liu et al., "ProEqBEV: Product Group Equivariant BEV Network for 3D Object Detection in Road Scenes of Autonomous Driving", 2024 IEEE International Conference on Robotics and Automation (ICRA), Yokohama, Japan, 2024, pp. 16178-16184, doi: 10.1109/ICRA57147.2024.10610492.
- [86] C. Zhang et al., "MotionTrack: End-to-End Transformer-based Multi-Object Tracing with LiDAR-Camera Fusion", in Computer Vision and Pattern Recognition, 2023, DOI: 10.48550/arXiv.2306.17000.
- [87] M. Wang et al., "Adaptive Fusion of LiDAR Features for 3D Object Detection in Autonomous Driving", in Sensors Journal 25, no. 13: 3865, 2025, doi: 10.3390/s25133865.
- [88] S. Usmankhujaev, S. Baydadaev and J. W. Kwon, "Perception, Distance Estimation, and Tracking Integration From BEV Representations", in IEEE Access, vol. 12, pp. 169136-169148, 2024, doi: 10.1109/ACCESS.2024.3488510.
- [89] . Wang et al., "Vision meets mmWave Radar: 3D Object Perception Benchmark for Autonomous Driving", 2024 IEEE Intelligent Vehicles Symposium (IV), Jeju Island, Korea, Republic of, 2024, pp. 2769-2775, doi: 10.1109/IV55156.2024.10588620.
- [90] H. Florea et al., "Enhanced Perception for Autonomous Driving Using Semantic and Geometric Data Fusion", in Sensors Journal 22, no. 13: 5061, 2022, doi: 10.3390/s22135061.
- [91] H. Yin et al., "Multi-Object Tracking with Object Candidate Fusion for Camera and LiDAR Data", in IEEE 26th International Conference on Intelligent Transportation Systems (ITSC), Bilbao, Spain, 2023, pp. 2965-2970, doi: 10.1109/ITSC57777.2023.10422106.
- [92] Y. Zhou, C. Yang, P. Wang, C. Wang, X. Wang and N. Ngoc Van, "ViT-FuseNet: Multimodal Fusion of Vision Transformer for Vehicle-Infrastructure Cooperative Perception", in IEEE Access, vol. 12, pp. 31640-31651, 2024, doi: 10.1109/ACCESS.2024.3368404.
- [93] G. Wang, C. Peng, Y. Gu, J. Zhang and H. Wang, "Interactive Multi-Scale Fusion of 2D and 3D Features for Multi-Object Vehicle Tracking", in IEEE Transactions on Intelligent Transportation Systems, vol. 24, no. 10, pp. 10618-10627, Oct. 2023, doi: 10.1109/TITS.2023.3275954.
- [94] A. El-Fatyany, "A robust multi-sensor fusion model against adversarial patch attack", in Wireless Networks Journal, Volume 32, pages 719-735, 2026, doi: 10.1007/s11276-025-04065-4. 10.1109/Smart-Nets69662.2026.11604736.
- [95] X. Li et al., "An Online Calibration Method for Robust Multi-Modality 3D Object Detection", 2024 IEEE 11th International Conference on Data Science and Advanced Analytics (DSAA), San Diego, CA, USA, 2024, pp. 1-10, doi: 10.1109/DSAA61799.2024.10722831.
- [96] Z. Chen et al., "SynthDrive: Scalable Real2Sim2Real Sensor Simulation Pipeline for High-Fidelity Asset Generation and Driving Data Synthesis", 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Hangzhou, China, 2025, pp. 912-919, doi: 10.1109/IROS60139.2025.11247690.
- [97] B. Naveen et al., "Advanced artificial intelligence-based multi-sensor fusion for environmental perception in autonomous electric vehicles", in Bulletin of Electrical Engineering and Informatics Journal, Vol 15, No 1, 2026, DOI: 10.11591/eei.v15i1.10739.
- [98] P.S. Dheepika & V. Umadevi, "A NOVEL HYBRID DEEP LEARNING APPROACH FOR 3D OBJECT DETECTION AND TRACKING IN AUTONOMOUS DRIVING", in Computer Science Journal, 2024, DOI: 10.7494/csci.2024.25.3.5597.
- [99] H. Li et al., "Proposal-Guided Multi-Scale Radar and Vision Fusion for 3D Object Detection", in IEEE Transactions on Intelligent Transportation Systems, vol. 26, no. 12, pp. 22232-22245, Dec. 2025, doi: 10.1109/TITS.2025.3616963.
- [100] S. Yan et al., "Research on Cooperative Vehicle-Infrastructure Perception Integrating Enhanced Point-Cloud Features and Spatial Attention", World Electric Vehicle Journal 17, no. 4: 164, 2026, doi: 10.3390/w-evj17040164
- [101] W. Zimmer et al., "InfraDet3D: Multi-Modal 3D Object Detection based on Roadside Infrastructure Camera and LiDAR Sensors", 2023 IEEE Intelligent Vehicles Symposium (IV), Anchorage, AK, USA, 2023, pp. 1-8, doi: 10.1109/IV55152.2023.10186723.
- [102] L. Ruddat et al., "Sparsity-Robust Feature Fusion for Vulnerable Road-User Detection with 4D Radar", Applied Sciences 14, no. 7: 2781, 2024, doi: 10.3390/app14072781.
- [103] L. Yan et al., "A Vehicle-Infrastructure Cooperative 3D Object Detection Scheme Based on Adaptive Feature Selection", in Journal of

- Electronics & Information Technology, 2025, 47(12): 5214-5225. doi: 10.11999/JEIT250601.
- [104] Y. Yuan et al., "Keypoints-Based Deep Feature Fusion for Cooperative Vehicle Detection of Autonomous Driving", in *IEEE Robotics and Automation Letters*, vol. 7, no. 2, pp. 3054-3061, April 2022, doi: 10.1109/LRA.2022.3143299.
- [105] S. Rho et al., "Multi-Sonar Fusion-Based Precision Underwater 3D Reconstruction for Optimal Scan Path Planning of AUV", in *IEEE Access Journal*, PP(99):1-1, 2025, DOI: 10.1109/ACCESS.2025.3542084.
- [106] E. Khatab et al., "Evaluation of 3D Vulnerable Objects' Detection Using a Multi-Sensors System for Autonomous Vehicles", in *Sensors Journal* 22, no. 4: 1663, 2022, doi: 10.3390/s22041663.
- [107] Ch. Minhø & K. Euntai, "3D LiDAR Multi-Object Tracking Using Multi Positive Contrastive Learning and Deep Reinforcement Learning", in *IEEE Access Journal* PP(99):1-1, DOI: 10.1109/ACCESS.2024.3521334.
- [108] A. Sochaniwsky et al., "A Camera-LiDAR Fusion Framework for Traffic Monitoring", in *2024 IEEE Transportation Electrification Conference and Expo (ITEC)* (2024): 1-6, doi: 10.1109/ITEC60657.2024.10598852.
- [109] F. Liu et al., "A Real-Time and Multi-Sensor-Based Landing Area Recognition System for UAVs", in *Drones Journal* 6, no. 5: 118, 2022, doi: 10.3390/drones6050118.
- [110] S. Bultmann et al., "Real-time multi-modal semantic fusion on unmanned aerial vehicles with label propagation for cross-domain adaptation", in *Robotics and Autonomous Systems Journal*, 2023, Article: 104286, Volume 159, doi: 10.1016/j.robot.2022.104286.
- [111] Y. Shang et al., "3D obstacle avoidance system for autonomous vehicles based on multi-sensor fusion", In: *Intelligent Manufacturing and Cloud Computing: Proceedings of the 2nd International Conference (ICIMCC 2025)*, Wuhan, China, 12-14 December 2025. SAGE Publications; 2026:194-204. doi:10.3233/ATDE260211.
- [112] Y. Qi et al., "FusionBev: LiDAR and 4D radar fusion for 3D object detection", in *Information Fusion Journal*, 2026, Article: 104240, Volume 132, doi: 10.1016/j.inffus.2026.104240.
- [113] W. Wang et al., "SWG-Fusion: Soft weather-guided multimodal fusion with VLM-assistance for BEV object detection under harsh weather", in *Pattern Recognition Journal*, 2026, Article: 113398, Volume 178, doi: 10.1016/j.patcog.2026.113398.