IEEE Open Journal of Vehicular Technology

Received 21 January 2025; accepted 12 February 2025. Date of publication 14 February 2025;
date of current version 11 March 2025. The review of this article was coordinated by Editor Nan Cheng.

Digital Object Identifier 10.1109/OJVT.2025.3542213

Infrastructure Assisted Autonomous Driving: Research, Challenges, and Opportunities

ROSHAN GEORGE [1,2], JOSEPH CLANCY[1,2], TIM BROPHY [1,2], GANESH SISTU[3], WILLIAM O'GRADY[3], SUNIL CHANDRA[3], FIACHRA COLLINS[3], DARRAGH MULLINS [1,2], EDWARD JONES [1,2] (Senior Member, IEEE), BRIAN DEEGAN [1,2] (Member, IEEE), AND MARTIN GLAVIN [1,2] (Member, IEEE)

[1]School of Engineering, University of Galway, H91 TK33 Galway, Ireland
[2]Ryan Institute, University of Galway, H91 TK33 Galway, Ireland
[3]Valeo, Tuam, Company, H91 TK33 Galway, Ireland

CORRESPONDING AUTHOR: ROSHAN GEORGE (e-mail: r.george5@universityofgalway.ie).

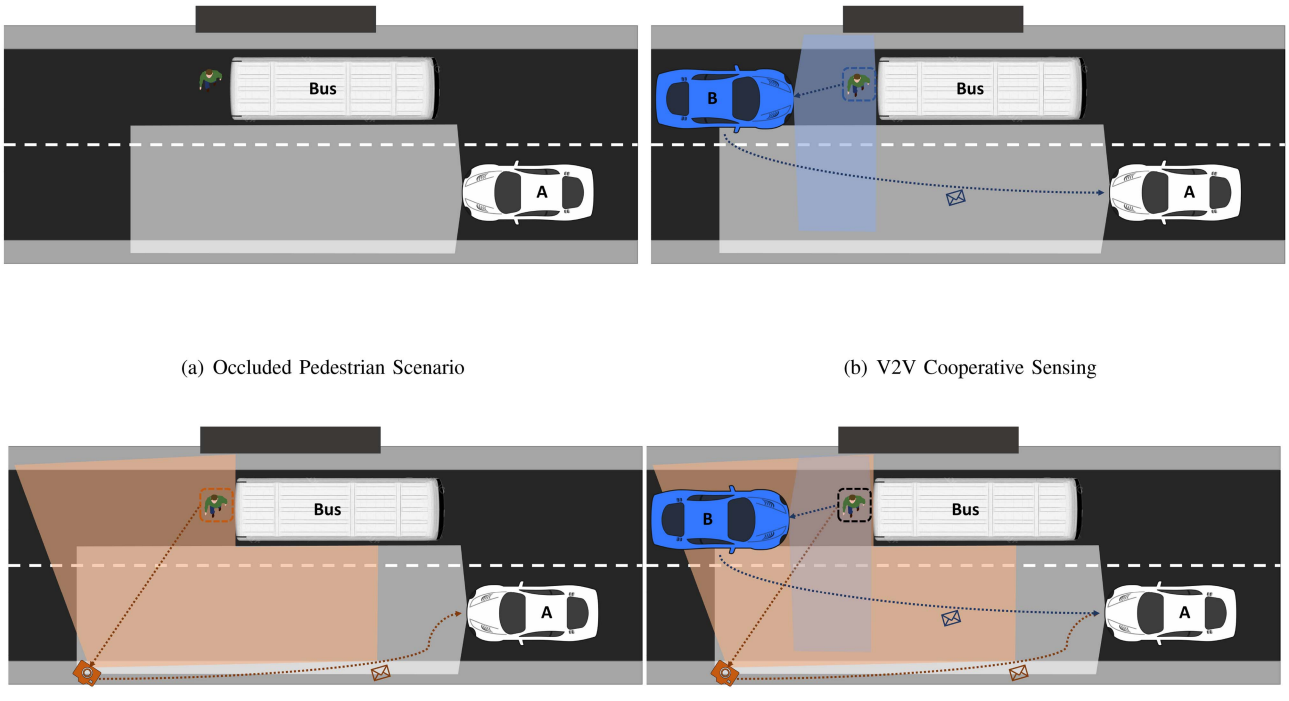
This work was produced by the Connaught Automotive Research (CAR) Group at the University of Galway and was supported in part by Taighde Éireann — Research Ireland under Grant 13/RC/2094 P2 and Grant 18/SP/5942, in part by the European Regional Development Fund through the Southern and Eastern Regional Operational Programme to Lero - the Research Ireland Center for Software (www.lero.ie), and in part by Valeo Vision Systems.

ABSTRACT Despite advancements in perception technology, achieving full autonomy in vehicles remains challenging partly due to limited situational awareness. Even with their sophisticated sensor arrays, autonomous vehicles often struggle to comprehend complex real-world environments due to the challenges associated with occlusion. A possible solution for addressing this limitation lies in the concept of vehicle-to-infrastructure cooperative driving, which enables vehicles to interact with various sensors implemented in the surrounding infrastructure. The infrastructure can share real-time data, such as traffic conditions, road hazards, and weather updates, facilitating safer and more efficient navigation. Within this framework, cooperative sensing is a crucial component, augmenting the onboard sensing capabilities of autonomous vehicles. Cooperative sensing surpasses traditional onboard sensors by leveraging a shared sensor network among vehicles and infrastructure. This approach mitigates challenges posed by occlusion, where objects are obscured from a vehicle's direct view. By pooling information from multiple sources, autonomous vehicles can gain a more comprehensive understanding of their surroundings, leading to enhanced safety and performance on the road. This study addresses a literature gap regarding information flow from real-world scenes to environmental models for cooperative V2I systems. It explores three core concepts essential for understanding the environment: sensing, perception, and mapping. This paper identifies the specific information required from infrastructure nodes, proposes an optimized sensor suite, discusses data processing algorithms, and investigates effective spatial model representations for cooperative sensing. This research informs the reader about the different challenges and opportunities associated with a V2I cooperative sensing system.

INDEX TERMS Cooperative intelligent transportation systems (C-ITS), infrastructure sensing, map fusion, roadside units, V2I, V2X.

I. INTRODUCTION

According to the World Health Organization, road traffic collisions claim the lives of approximately 1.3 million individuals each year, with over half of these fatalities affecting Vulnerable Road Users (VRU) such as pedestrians and cyclists [1]. With the rapid growth in the adoption of advanced driver assistance systems (ADAS) in vehicles [2], it is expected that ADAS and autonomous driving features will improve vehicle safety and reduce accidents. A study by the Insurance Institute for Highway Safety has reported a 7% reduction in traffic collisions with vehicles equipped with forward collision warning and up to 15% reduction in accidents with vehicles equipped with more advanced ADAS features, such as automatic emergency braking [3]. Companies such as Tesla [4], BMW [5],



(a) Occluded Pedestrian Scenario

(b) V2V Cooperative Sensing

(c) V2I Cooperative Sensing

(d) V2X Cooperative Sensing

FIGURE 1. **Examples of cooperative sensing for detecting occluded pedestrians. (a) The pedestrian is occluded by the bus and not visible to vehicle A. (b) V2V cooperative sensing where vehicle B detects the pedestrian occluded from vehicle A's view and shares this information with vehicle A. (c) V2I cooperative sensing where the FSN detects the occluded pedestrian and shares this information with vehicle A. (d) V2X cooperative sensing where vehicle A receives information about the occluded pedestrian from both vehicle B and a FSN.**

Toyota [6], and Audi [7] provide SAE L2 automated vehicles, and Mercedes-Benz provides L3 equipped vehicles [8]. Waymo [9], Cruise [10], Zoox [11], and Baidu [12] provide robotaxi services at L4 autonomy. These companies strive to use advanced autonomous driving technologies to reduce accidents, improve transportation efficiency, and increase mobility.

Despite the rapid pace of development of perception technologies, there are still several challenges that hinder the progress and safety of autonomous vehicles. Limited situational awareness is one of the key obstacles that challenge drivers and autonomous vehicles alike. Autonomous vehicles perceive the environment through a number of onboard multimodal sensors [13], [14], [15]. Vehicles generally employ several sensor fusion techniques [16], [17], [18] to overcome individual sensor limitations [19]. Nevertheless, the perception and understanding of the environment by the ego vehicle, which is the vehicle under study, is restricted as it depends solely on observing the environment from the ego vehicle's perspective. As a result, several common traffic scenarios such as darting pedestrians [20], VRU occlusion behind parked vehicles [21], and vehicles emerging from concealed entrances, all pose a significant challenge to autonomous vehicles. An occluded pedestrian emerging from behind a parked vehicle is illustrated in Fig. 1(a).

In [20], Palffy et al. proposed a solution to overcome the problem of darting pedestrians using a stereo camera and radar sensor fusion approach. A stereo camera pair is employed to identify the top half of the body. This information is coupled with the multi-path propagation property of radar to correctly estimate the likelihood of an occluded pedestrian.

Another approach to mitigate the issues associated with occlusion is through Vehicle-to-Everything (V2X) cooperative driving. Cooperative driving enables the ego vehicle to gather information about the surrounding environment by communicating with other vehicles, fixed infrastructure nodes, or other entities in the environment. A fixed infrastructure node, also known as a fixed sensor node (FSN) or a roadside unit (RSU), is a collection of sensors mounted at an elevated viewpoint, with built-in wireless connectivity and data processing units. For the remainder of this paper, the term FSN will be used to refer to this type of device. The primary advantage of a FSN is an elevated viewpoint that mitigates the issues of occlusion and is constantly available to monitor danger zones such as complex intersections, areas with dense pedestrian activity, and locations with limited visibility. With this

information, the cooperative driving participants jointly plan and execute driving maneuvers [22], increasing safety, and traffic efficiency [23].

Cooperative perception, also known as cooperative sensing [24], is a crucial component of cooperative driving. It involves sharing information about detected objects to enhance the environmental model of all participants [22], [24]. Cooperative sensing can be broken down into three key components, where the first two align with the 'sense' aspect of the sense-plan-act protocol [25] used in robotics:

1) Observing the scene through various sensors, that provide 'raw' sensor data.
2) Processing the raw data to generate meaningful perception task outputs, such as object, semantic, and trajectory information.
3) Representing the perception task outputs in a unified coordinate system within an environmental model, which is then shared using an appropriate communication method.

To clarify for the reader, in this paper we define cooperative sensing as the process of capturing raw data from sensors, processing these data to extract meaningful perception information, and representing this information in a unified environmental model [26] that can be shared among participants through appropriate communication methods. Fig. 1(b), (c), and (d) illustrates scenarios where Vehicle-to-Vehicle (V2V), Vehicle-to-Infrastructure (V2I), and V2X cooperation can aid the ego vehicle in detecting an occluded pedestrian. V2I cooperative sensing includes the vehicle and the FSN cooperatively communicating with each other and sharing complimentary information to enhance the environmental model.

In this paper, we define a map as an environmental model that is a unified representation of perception task outputs in a shared coordinate system. It is important to note that the definition of a map can vary depending on the specific application or context. For example, in [27], a map is defined as a spatial model of the environment with varying levels of abstraction, which supports localization, navigation, and exploration. Galceran et al. [28] discussed mapping in the context of path planning, where mapping refers to the process of a robot building a representation of the environment, that is used for coverage path planning to ensure efficient exploration of the entire space. In [29] active mapping is defined as the problem of a robot actively planning trajectories to explore the environment and incrementally construct a spatial model. Thrun et al. [30] define mapping as the act of acquiring spatial models of physical environments. These definitions indicate that mapping often consists of constructing a spatial model, which is used for various tasks such as localization, navigation, and exploration. In this paper, the focus is solely on the construction of a spatial model of the environment and does not introduce or explore ego vehicle localization, path planning, or exploration. With this definition, we focus the paper's discussion on the process of creating a unified representation of the environment based on the outputs of various perception tasks. The discussion on map fusion will consist of spatial and temporal alignment of these perception task outputs. Furthermore, we believe it is important to discuss the unification of perception data as a map because there are many mapping paradigms and numerous considerations required for map fusion. If we were to use a "black-box" approach, which performs all these alignments in the network and is then passed into a detection head, then these important considerations would be abstracted away.

The importance of cooperative driving is further highlighted with the release of the SAE J3216 [31] standards for cooperative driving applications. This standard defines the various components of cooperative driving along with the information that needs to be exchanged between the entities involved. There are four cooperative classes defined: Status-Sharing, Intent-Sharing, Agreement-Seeking and Prescriptive Cooperation. Table 1 provides a summary of the SAE J3216 standards.

Class A: Status-sharing cooperation involves the ego vehicle and any other participant in the environment, such as pedestrians, conventional vehicles, or FSNs. Participants share information about what they perceive along with their location. Status-sharing cooperation is intended to expand the spatial awareness of the receiver.

Class B: Intent-sharing cooperation also involves the ego vehicle and other participants within the environment, where the participants provide their future planned actions along with the future predicted actions of other participants in the scene to the ego vehicle. The future state of the sender is predicted from their kinematic state (i.e. position, velocity, and orientation) and projected path. Similarly, the other participants' future states can be predicted from their estimated projected path, based on their kinematic state, but also from visual indications of intention such as a turning signal or brake lights.

Class C: In Agreement-seeking cooperation, the ego vehicles and other automated driving systems on the road collaborate with each other. The ego vehicle proposes an action to jointly execute, and the accompanying participants can either accept or reject it to execute this dynamic driving task.

Class D: Prescriptive cooperation encapsulates all automated driving systems in the environment. These automated driving systems are required to adhere to a set of prescribed actions by an authoritative body.

In a scenario where a status-sharing cooperation scheme is utilized by a vehicle with support from a FSN, assuming the vehicle is making the decisions, then a FSN must perform two main tasks: the FSN must first accurately perceive and comprehend the environment, thus extending the perception range of the ego vehicle. The FSN must then share this information with all other connected entities. There are several potential benefits of using an infrastructure-based cooperative sensing system. The primary benefit includes improved safety of conventional and autonomous vehicles by diminishing the effects of occlusion and by perceiving beyond the line of sight. The mobility of vehicles on the road is enhanced by optimizing traffic flow. Finally, fuel consumption and efficiency of

TABLE 1. Cooperative Driving Levels as Outlined in SAE J3216 [31]

Cooperation Class	Architectures Involved	Information Exchanged
Class A: Status Sharing	V2V/V2I/V2X	Perceived object information (classification, pose, dimensions), environmental information (ambient weather, road surface conditions), and traffic participant information (density, speed, volume).
Class B: Intent Sharing	V2V/V2I/V2X	Ego vehicle intention (future path, future maneuvers), traffic participant intention, and visual cues of intention.
Class C: Agreement Seeking	V2V/V2I	Proposed action, and agreement/disagreement of action.
Class D: Prescriptive Cooperation	V2V/V2I	Motion control, and traffic/environmental rules.

on-road vehicles are improved by enabling vehicle platooning and optimized path planning.

A. RELATED WORKS

The importance of cooperation is further emphasized by numerous initiatives that many government and academic institutions pursue. This section will discuss the projects and literature that explore questions about real-world applications, the maturity of the technology required, and the policy issues associated with V2V and V2I cooperation.

1) VEHICLE-TO-VEHICLE COOPERATION

Across the globe, there are several ongoing efforts by many institutions to address the challenges associated with enabling cooperative driving.

The Grand Cooperative Driving Challenge (GCDC) provided a competitive platform to facilitate the advancement of cooperative driving. In 2009, the GCDC [32] primarily focused on V2V cooperation where the competition required participants to demonstrate vehicle platooning and platoon merging in scenarios such as highway and urban junctions. In 2016, the GCDC [33] introduced interoperable wireless communication and more complex traffic scenarios derived from the Cooperative Intelligent Transport System (C-ITS) standards. The SARTRE platooning program [34] is another initiative run by the European Commission to advance vehicle platooning capabilities on public motorways.

The US Department of Transportation (DoT) is advancing cooperative driving automation (CDA) through two key initiatives: CARMA [35] and IntelliDrive [36] program. These programs aim to integrate CDA technology into the existing transportation system. The CARMA Program [35] explores applications of CDA in three distinct tracks: traffic, reliability, and freight. The traffic track aims to identify how CDA can reduce congestion and improve traffic flows. The reliability track addresses issues around non-recurring traffic due to accidents or construction work, and how CDA can mitigate the effects of this. The freight research track investigates how CDA can improve efficiency in ports by improving freight movement reliability. Researchers from the IntelliDrive program [36] investigate how CDA can improve safety and mobility while reducing environmental impact.

While both European and US initiatives share the common goal of advancing cooperative driving, their approaches differ significantly. GCDC and SARTRE programs have emphasized competitive demonstrations and real-world implementation on public roads, particularly focusing on platooning technology. In contrast, the US programs take a more systematic, multi-track approach, addressing various aspects of transportation infrastructure integration. The CARMA program's three-track structure represents a more comprehensive strategy compared to the European focus on specific use cases. By integrating with existing transportation infrastructure, this approach will accelerate the adoption of the technology, as the associated industries could bear the initial costs.

2) VEHICLE-TO-INFRASTRUCTURE COOPERATION

Another core topic of interest for cooperation is V2I cooperative driving. As previously discussed, in scenarios with significant occlusions, leveraging infrastructure sensors can aid in improving safety and traffic efficiency. As a result, several projects aim to address the challenges and advantages of V2I cooperation. Both MEC-View [37] and Providentia++ [38] investigated how V2I cooperation can be beneficial for VRU safety and traffic flow. Particularly, they focus on how limitations of onboard sensing can be mitigated by a FSN sharing information with an autonomous vehicle. In MEC-view [37], the data from a multi-modal sensor system are fused to generate a local environmental model, this local model is shared with the ego vehicle to generate a unified environmental model. INFRAMIX [39] a Horizon 2020 funded project, studies how V2I cooperation can support the transition of fully autonomous vehicles onto roads while coexisting with conventional vehicles. As a result of this project, Infrastructure Support for Automated Driving (ISAD) [40] schemes were introduced. ISAD is an infrastructure classification scheme that defines the levels of infrastructure-based cooperation and the information required to be exchanged.

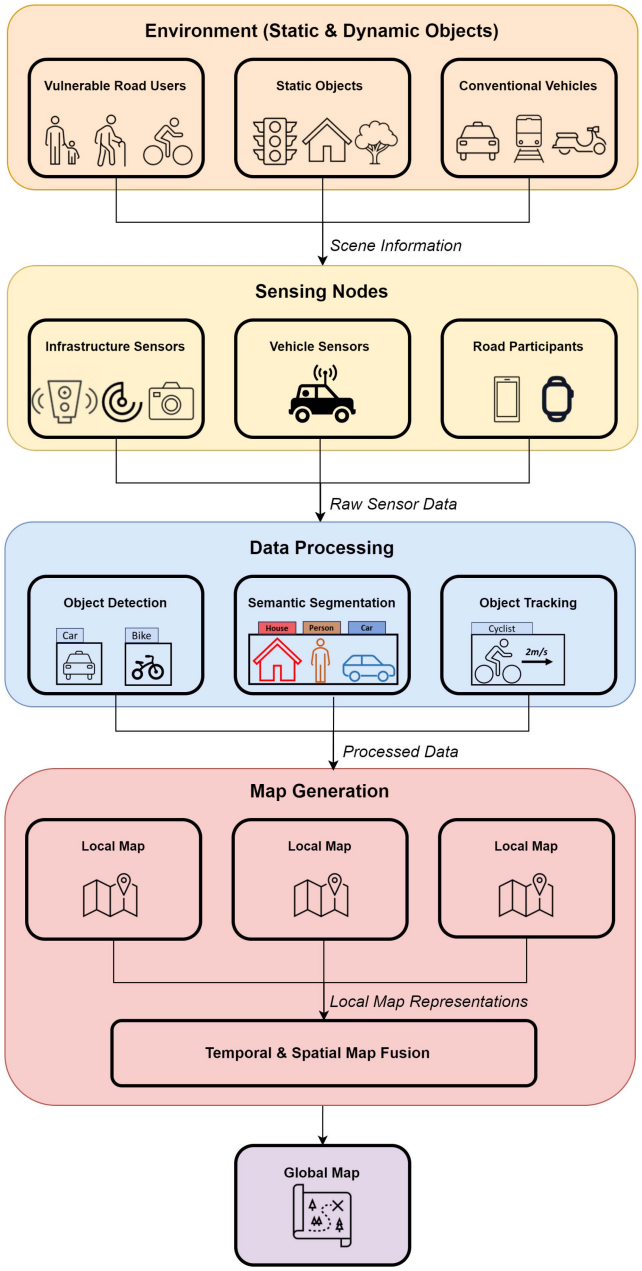


FIGURE 2. Cooperative V2X pipeline outlining the flow of information from a scene to a global map.

When comparing these initiatives, it is evident that distinct approaches are taken but they have a shared objective in addressing VRU safety and traffic flow. MEC-View emphasizes a multi-modal sensor fusion approach to create local environmental models, Providentia++ takes a broader perspective on infrastructure integration. INFRAMIX focuses on the transition period where autonomous and conventional vehicles must coexist, introducing the standardized ISAD classification scheme. This represents a more systematic approach to infrastructure integration compared to the more technically focused solutions of MEC-View and Providentia++.

Fig. 2 depicts a generalized pipeline depicting how environmental sensors in the infrastructure, vehicle sensors, and other road participants can share information and maintain a global map representation of the environment.

3) REVIEW PAPERS

In addition to the aforementioned projects, several review articles [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52] explore the various aspects of cooperative driving, such as perception, control, and communication. A summary of these papers comparing the different points of discussion is provided in Table 2.

Bai et al. [41] conducted a comprehensive survey proposing a hierarchical cooperative perception framework that integrates diverse tasks across various scenarios. Their review focuses on both infrastructure and vehicle perspectives for cooperative perception, examining sensing modalities and fusion schemes. The survey outlines three key directions for the future of cooperative perception in autonomous driving: heterogeneous cooperation, multi-modal cooperation, and scalable cooperation. The authors emphasize that V2I collaboration can effectively overcome physical occlusion limitations by leveraging the strengths of both mobile vehicles and FSN. They identify multi-modal sensor fusion across multiple perception nodes as a promising, yet underexplored area for improving overall system accuracy. To address scalability, the authors suggest moving beyond limited node interactions to more comprehensive cooperative systems. They propose a hierarchical structure encompassing vehicles, infrastructure, and cloud computing to tackle real-world implementation challenges, including varying computational resources and dynamic environments. This approach aims to optimize cost-effectiveness and performance by deploying lightweight modules in vehicles while utilizing high-performance infrastructure nodes for complex processing tasks.

Caillot et al. [46] conducted a review on V2V and V2I cooperation, which examined key aspects of cooperative systems, including system architecture, data sharing methods, fusion strategies, and approaches to mapping and localization. The author's analysis of cooperative systems revealed significant advantages, including improved localization precision even in GNSS-denied environments, drift-free operation with real-time updates, enhanced reliability, cost-efficiency, and an expanded field of view. Despite these benefits, the systems face challenges such as high computational demands, latency issues, and substantial data rate requirements. The study also identified promising opportunities, including raw sensor data fusion, V2I map generation, and enhanced trajectory planning. However, potential threats to the successful implementation of these systems include data management complexities and synchronization difficulties among participants.

Cui et al. [43] surveyed cooperative perception in autonomous driving within the Internet of Vehicles (IoV)

TABLE 2. Summary of Review Papers in Cooperative Perception. SS (Sensor Selection), T (Data Transmission), MF (Message Format), MS (Message Sharing), RA (Resource Allocation), SA-OD (Single Agent Object Detection), SA-T (Single Agent Tracking), SA-S (Single Agent Segmentation), C-OD (Collaborative Object Detection), C-T (Collaborative Tracking), C-S (Collaborative Segmentation), PA (Pose Alignment), TA (Temporal Alignment), BEV (Bird's Eye View), HD (HD Maps), SL (SLAM), RT (Real-Time Performance), PP (Privacy Preserving), A&p (Attacks & Defenses), EF (Economic Feasibility)

Survey	SS	DT	MF	MS	RA	SA-OD	SA-T	SA-S	C-OD	C-T	C-S	CS	PA	TA	BEV	HD	SL	RT	PP	A&P	EF
Bai et al. [41]	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Yu et al. [42]	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Cui et al. [43]	✗	✓	✓	✓	✓	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✓	✗
Bai et al. [44]	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Ghorai et al. [45]	✓	✗	✗	✗	✗	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗	✗
Caillot et al. [46]	✓	✓	✓	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✗	✓	✗	✗	✗	✗
Han et al. [47]	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗
Huang et al. [49]	✗	✓	✓	✓	✓	✗	✗	✗	✓	✓	✓	✗	✓	✓	✗	✗	✗	✓	✓	✗	✗
Liu et al. [50]	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✓	✗	✓	✓	✗	✗	✗	✓	✓	✓	✗
Bejarbaneh et al. [52]	✗	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗	✓	✓	✗	✗
Gao et al. [48]	✗	✓	✗	✗	✗	✓	✓	✓	✓	✗	✗	✗	✗	✗	✓	✗	✗	✓	✓	✓	✗
Wang et al. [51]	✗	✗	✗	✓	✗	✗	✗	✗	✓	✓	✓	✗	✓	✓	✗	✗	✗	✗	✗	✗	✗
Ours	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

environment, emphasizing sensor fusion, communication methodologies, and implementation challenges. The authors identified multimodal fusion of image and point cloud data as the most effective approach for balancing perception accuracy with communication load and computing efficiency. The study compared information fusion and sharing stages, noting that while processed data fusion reduces communication load, raw sensor data fusion offers better perception at the cost of increased communication burden. Broadcasting emerged as the common information-sharing strategy, but it introduces challenges like channel congestion and redundant data transmission, necessitating efficient resource allocation and task scheduling. Vehicle mobility complicates matters further by affecting communication stability and data integrity. The authors also highlighted significant security concerns due to the distributed nature of edge nodes, suggesting blockchain technology as a potential solution for ensuring data integrity and trust. Overall, the survey underscores the complex trade-offs between perception accuracy, communication efficiency, and system security in cooperative perception systems for autonomous vehicles.

Yu et al. [42] conducted a comprehensive review of cooperative perception and control strategies in V2V and V2I contexts. This study emphasizes that cooperative perception enhances perception capabilities, particularly in challenging scenarios like occluded intersections and dense traffic areas. The review explores various aspects of these systems, including data fusion techniques, V2X communication technologies, and cooperative control strategies, highlighting the critical role of robust, real-time communication protocols. The authors examine applications across diverse environments, from urban crossroads to mining sites, illustrating specific use cases such as blind spot perception and intersection management. While also acknowledging challenges like communication latency and data security. The author concludes that emerging technologies such as 5 G and cloud computing will likely propel Infrastructure-Vehicle Systems (IVS) to significantly enhance overall transportation efficiency and transform urban mobility.

Han et al. [47] conducted a comprehensive review of contemporary cooperative perception schemes, focusing on their deep learning architectures and providing a breakdown of intermediate collaborative perception methods. The study also outlined various cooperative perception datasets available for research. The authors identified several key challenges and future research directions for implementing collaborative perception in real-world autonomous driving. The authors noted that transmission efficiency remains a critical issue, suggesting future work should explore dynamic feature compression strategies based on data importance and address blind and weak perception areas. Another challenge is achieving robust perception in complex scenes, as existing datasets often fail to cover challenging conditions like bad weather, highways, and scenarios with small or distant objects. To enhance system robustness, the authors recommend gathering collaborative perception data in complex environments and developing tailored methods, such as multi-sensor fusion, virtual point cloud generation, and spatio-temporal data fusion. To address privacy and communication concerns, they propose federated learning-based collaborative perception as a solution, enabling decentralized model training without exchanging raw data. Additionally, the authors emphasize the need to reduce labeling dependencies in current methods, suggesting future research should focus on generalized weakly supervised learning and domain adaptation.

Bai et al. [44] reviewed techniques employed for infrastructure-based object detection and tracking systems for cooperative driving, focusing on addressing the challenges of infrastructure-based object detection methods. The authors conclude that the key challenges include developing efficient multi-sensor fusion schemes, acquiring and annotating roadside data to promote deep learning-based research, and addressing synchronization issues in large-scale implementations. Future trends point towards multi-sensor fusion

leveraging high-computational edge servers, cooperative perception to overcome physical occlusion limitations, and the development of lightweight on-board units for vehicles, which aims to balance cost-effectiveness with performance by relying more on infrastructure-based high-performance nodes for wider-range perception tasks.

Ghorai et al. [45] conducted a comprehensive study into motion prediction and state estimation of vehicles and VRUs for safe autonomous driving. Their study investigates some object detection and motion prediction methods, while also addressing the limitations of ego vehicle object detection. Some of the challenges outlined by the author include improving depth estimation for vision-based systems, enhancing sensor performance in poor weather conditions, optimizing sensor combinations for cost-effectiveness, and increasing detection accuracy for smaller objects. The survey emphasizes the need for better prediction of road user intentions, especially for VRUs like pedestrians and cyclists in urban environments. It also highlights the potential of Cooperative Perception and Navigation (CPN) systems, which face challenges related to data privacy, authenticity, and communication issues. The authors highlight the importance of further research in sensor fusion, pedestrian state estimation, and motion prediction. They also discuss the promise of V2X and V2I connectivity for enhancing VRU awareness and interaction with AVs.

Liu et al. [50] provided a comprehensive survey that reviews recent advancements in collaborative perception for V2X autonomous driving scenarios. The paper provides an overview of the V2X systematic architecture and proposes a taxonomy to categorize cooperative perception (CP) methods from various perspectives. It thoroughly summarizes existing datasets for CP and conducts extensive experimental analyses to compare state-of-the-art CP methods in terms of model efficiency, robustness, and generalization. The survey outlines key challenges and opportunities in CP for V2X, including optimizing the performance-bandwidth trade-off, developing more realistic simulators, addressing data management with fog/edge computing, improving model generalization across simulation-to-real and real-world scenarios, and integrating intelligent multi-agent cooperative control. The authors emphasize the importance of decentralized CP architectures, data management solutions using fog/edge computing, systematic cooperative control for mixed-traffic multi-intersections, optimization and security of infrastructure for CP, and translation of cooperative control theory into real-world applications. Overall, this survey provides a holistic view of the CP landscape for V2X autonomous driving.

Bejarbaneh et al. [52] provides a review of cooperative vehicle-intersection systems (CVIS), emphasizing shared perception and control. Bejarbaneh et al. proposed a three-tiered CVIS architecture, consisting of shared perception, intersection control, and vehicle control layers. The authors delve into cooperative perception methods, examining how connected and autonomous vehicles (CAVs) and infrastructure can share perceptual information to create a more complete understanding of the surrounding environment. This survey highlights key challenges and future research areas for CVIS, which include decentralized approaches to cooperative perception and control, data management strategies using fog/edge computing, control systems for mixed traffic flow across multiple intersections, optimizing and securing CVIS infrastructure, intelligent multi-agent cooperative control, and real-world testing and implementation.

Wang et al. [51] examined perception methods within vehicle-infrastructure cooperation systems, specifically focusing on adverse weather conditions. Their survey concentrates on three core areas: pre-processing LiDAR and image data in adverse weather; fusing multi-sensor data from images, point clouds, and image-point cloud combinations; and cooperative perception strategies between vehicles and infrastructure for information sharing and fusion. The authors emphasize the ongoing challenges in robust perception during dense fog, heavy rain, and snow. For future research, they suggest exploring multi-scale depth and semantic information for data preprocessing, adaptive calibration for heterogeneous data and detection results, hybrid information cooperation strategies, and lightweight deep learning networks.

Gao et al. [48] provides a detailed overview of collaborative perception for intelligent vehicles at intersections, focusing on specific tasks and goals within perception and communication for vehicle-road collaboration. The survey analyzes the timing and function of each task in the cooperative perception workflow. Perception tasks are categorized into detection, segmentation, and tracking, based on their goals. Communication tasks are examined, addressing challenges like delay robustness, bandwidth, positioning errors, security attacks, and privacy preservation.

Huang et al. [49] offers a comprehensive overview of recent advances and challenges in V2X cooperation for autonomous driving. The paper traces the evolution of CP technologies, from initial explorations to current state-of-the-art proposals, emphasizing advancements in V2X communication. A modern generic framework is presented to illustrate the V2X-based CP workflow, facilitating a systematic understanding of CP system components. This paper reviews V2X CP solutions, addressing practical driving scenario challenges like model and data heterogeneity, data rate requirements, lossy communication, data trustworthiness, perception uncertainties, task discrepancy, data privacy, and simulation-to-reality gaps. Open challenges and future research opportunities in V2X-based CP are discussed, considering both perception and V2X communication advancements. These challenges include collaboration triggers, realistic communication constraints, malicious and selfish behavior, real-world generalization, and handling challenging scenes and corner cases. Suggested future directions include hybrid collaboration schemes, multi-modal data sharing, integrated sensing and communication for CP, responsible AI for CP, privacy-preserving CP, and collaborative AI for enhanced CP and V2X communication.

B. MOTIVATION

The primary motivation behind this paper is to provide a comprehensive resource for designing a cooperative V2I system. As previously mentioned, cooperative sensing is defined as the process of capturing raw data from sensors, processing these data to extract meaningful perception information, and representing this information in a unified environmental map that can be shared among participants through appropriate communication methods. Eskandarian et al. [26] frames cooperative perception as a map merging problem, where the key steps are to estimate the relative pose between agents and merge the perception data into a global coordinate system. Thus, understanding single-agent paradigms is important for cooperative perception.

To achieve this, the paper will explore three core concepts essential for understanding the environment: sensing, perception, and mapping. While there are several well-regarded review papers on sensing and perception from the ego vehicle's perspective, this work aims to introduce these topics from the specific perspective of V2I.

This paper differs from the above-mentioned reviews by providing a guide for a 'ground-up' design approach for implementing a V2I cooperative sensing system. The paper provides a comprehensive analysis of the key sub-components: sensing, perception, and mapping from the perspective of FSNs, which is essential for tackling real-world challenges in deploying cooperative sensing systems. We explore the ideal sensor configuration for FSNs, considering factors such as cost, accuracy, reliability, and performance under various weather and lighting conditions while addressing sensor-specific challenges. Our paper also offers an in-depth examination of the critical perception tasks for extracting meaningful information from sensed data, including object detection, semantic segmentation, and tracking, discussing the strengths and limitations of different approaches. Additionally, we explore cooperative approaches for these perception tasks. Furthermore, we investigate the various mapping techniques used in cooperative perception, focusing on both Birds Eye View (BEV) and perspective view representations. We then focus on how to fuse these information spatially and temporally. Additionally, this paper provides an implementation section discussing aspects such as privacy, data security, trust, real-time performance, and economic feasibility of V2X systems. This exploration will contribute to the development of accurate and reliable environmental models that will enable effective V2I cooperation.

C. CONTRIBUTIONS

This paper aims to bridge the gap in the literature by providing a comprehensive analysis of the flow of information from scene capture to mapping in V2I cooperative sensing systems. Thus, the main contributions of this paper are as follows:

- Identifying the information that FSNs must share with connected vehicles to enhance their sensing capabilities.
- Defining an optimal sensor suite for FSNs in cooperative sensing systems. Analyze their advantages and disadvantages in specific scenarios and assess their suitability for deployment on infrastructure nodes.
- Examining various algorithms for perception tasks to process sensor data and generate object, semantic, and tracking information. Focusing on both single-agent and collaborative systems.
- Investigating different mapping approaches to represent processed data in a format that allows connected vehicles to extend their sensing range effectively. Particularly focusing on different methods for pose and temporal alignment.
- Discussing practical implementation considerations and presenting real-world use cases to illustrate the potential of cooperative perception systems in enhancing road safety and efficiency. Key emphasis on topics such as real-time performance, privacy preservation, trust, and cybersecurity.
- Exploring the economic feasibility of deploying a V2X system.

By addressing these key aspects, this work provides a comprehensive resource for researchers and practitioners looking to design and deploy cooperative systems, bridging the gap between theoretical concepts and practical implementation. To help readers effectively understand environmental mapping and the decisions involved in designing a V2I system, we will first explore some core concepts in environmental sensing and perception. This foundation will enable readers to better grasp the considerations and choices presented throughout the paper and follow along with the ground-up design approach.

By providing the reader with a comprehensive resource that encompasses the full design stack for V2I systems, we hope to facilitate and accelerate the deployment and research of these. Table 2 compares our work to previous survey papers, highlighting the unique contributions and broader scope of this paper in relation to the existing literature.

The paper is structured as follows: Section II focuses on the environment and the information that needs to be shared from an FSN to an ego vehicle to extend its perception capabilities. This section provides a comprehensive discussion of sensing modalities, highlighting their relative strengths, weaknesses, and performance. Additionally, a review of V2X data transmission and considerations is included.

Section III delves into the processing of sensor data to generate a greater understanding of the environment. It provides a detailed examination of 3D object detection, semantic segmentation, and object tracking approaches for a wide range of sensing modalities. Special emphasis is placed on cooperative sensing algorithms and their advantages compared to traditional approaches, providing valuable insights into the benefits of collaborative perception.

Section IV explores the different mapping strategies and map fusion approaches. It investigates various techniques for creating and combining maps from multiple data sources, addressing the challenges of integrating heterogeneous sensor

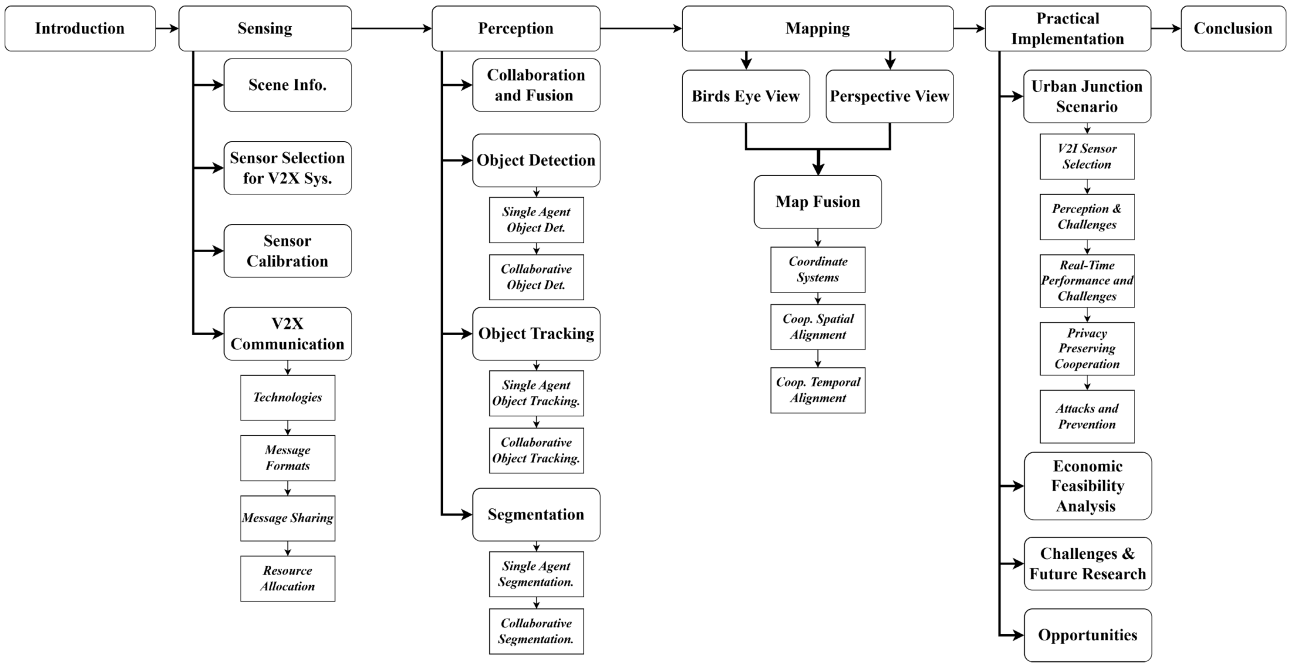


FIGURE 3. Overview of this review paper. The flow of information is outlined to show the relation from scene capture to mapping.

data. The section also discusses methods for aligning these maps in a unified coordinate system, thus maintaining map coherence and accuracy in dynamic environments, and ensuring the reliability of the mapping system.

Section V presents a practical implementation section, which builds on the enabling technologies introduced in the previous sections. We focus on a unique scenario where a cooperative V2I system can be deployed: an urban junction pedestrian interaction scenario. For this use case, a bottom-up approach is provided, covering the selection of suitable sensors, the identification of appropriate sensor information processing techniques, and the relevant mapping strategies for creating a unified environmental model. This section also highlights and discusses real-world challenges such as data privacy, security, and real-time performance. Furthermore, this section explores the economic feasibility of a V2X system. This section concludes with some existing challenges and opportunities associated with implementing cooperative sensing systems, while also outlining potential future research directions for this topic.

Finally, Section VI concludes the findings of this review, summarizing the key takeaways and offering a comprehensive understanding of the current state and future potential of cooperative sensing in intelligent transportation systems. As a support to the reader, Fig. 3 presents the structure of the remainder of this paper.

II. ENVIRONMENTAL SENSING

The goal of this section is to provide the reader with an overview of the components required to accurately perceive the environment from a FSN perspective. As previously discussed, accurate perception and comprehension of the environment is a crucial task for any autonomous driving application as it provides information to make reliable and robust decisions for vehicle control [53]. Firstly, the information that needs to be captured by a FSN will be identified, along with the various sensing modalities that are required to obtain this information. Following this, an overview of different calibration methodologies is provided. Finally, a discussion on V2X communication technologies, message formats, message-sharing strategies, and resource allocation will be provided.

It is important to discuss these enabling technologies since they form the basis for cooperative sensing. The sensor selection section will provide an overview of various sensor modalities and the factors to consider when deploying them. By exploring these topics, readers will gain insights into the key considerations involved in implementing cooperative sensing systems. This is essential for designing and deploying effective solutions that leverage the strengths of different sensor modalities.

A. SCENE INFORMATION

It is crucial to identify the information that needs to be captured and processed from a given scene. This can help determine the sensors that are required on a FSN, while also outlining the functional capabilities of the FSN. The discussion and analysis provided are necessary to understand what sort of information is contained in a scene, and whether this information can be sensed by a FSN. The five main information types that a FSN can share include physical, semantic, localization, environmental, and metadata information.

Expanding on the work of Malik et al. [54], Eising et al. [55] introduced the concept of the 4Rs of automotive computer vision: Reconstruction, Recognition, Relocalization, and Reorganization. Reconstruction refers to inferring 3D scene geometry in a point cloud or voxelized representation along with object pose, which directly corresponds to the physical information that a FSN must share. Recognition refers to adding semantic labels to entities that are detected in the scene, highlighting the importance of semantic information. Relocalization refers to localizing the vehicle relative to its surroundings. Reorganization is a fusion of object, semantic, and localization information into a unified representation. Thus, the 4Rs framework reinforces the idea that a FSN must be capable of sharing physical, semantic, and localization information as a unified map, demonstrating the interdependency of these information types.

The SAE J3216 standards [31] specify that for a status-sharing cooperation system, the minimum information that must be shared includes physical scene information, visual indicator cues, and environmental information. Physical information, consisting of object class, attributes, and pose, is a core requirement, aligning with the reconstruction aspect of the 4Rs. Visual indicator cues, such as brake lights and turning signals, can be considered a subset of metadata, providing crucial context for understanding vehicle intentions. Lastly, environmental information contains details about road surface conditions, ambient weather, traffic density, recurring events such as buses pulling into stops, and non-recurring events such as accidents. In contrast to the 4Rs, which focus on the core perception tasks, the SAE J3216 standard emphasizes the information needed for varying levels of cooperation.

Similarly, according to the ISAD levels [40], a FSN-based cooperative sensing system must perceive and transmit detailed traffic information to the ego vehicle. This is achieved by sharing a High-Definition (HD) map containing precise geo-referencing points for localization along with dynamic updates of the lane topology. The HD map encapsulates both physical and semantic information, similar to the unified map concept derived from the 4Rs. However, the ISAD levels place a greater emphasis on the precision and detail of the information, particularly for localization and lane topology, compared to the more general requirements of the SAE J3216 standard. This highlights a difference in focus, the SAE standard focuses on minimum requirements for basic cooperation, while the ISAD levels emphasize the need for detailed, high-precision information for advanced autonomous driving.

The importance of physical object information and semantic scene labeling is clear, as it is vital for scene mapping, obstacle avoidance, and vehicle control. Studies conducted by Wen et al. [56], Deng et al. [57], and Gupta et al. [58] further support the idea that physical and semantic information needs to be shared for a complete understanding of the scene. In [56], Wen et al. thoroughly reviewed LiDAR, camera, and LiDAR-camera fusion approaches for both 3D object detection and drivable area detection, providing a comprehensive overview of different approaches and their effectiveness in capturing physical and semantic information. In addition to this, [57] also identifies the critical role that localization plays in accurately traversing an environment, highlighting the interdependency of localization with physical and semantic information. A study by Gupta et al. [58] identified that 3D physical information and semantic information are needed for accurate scene understanding, reinforcing the findings of [56], [57].

The above-mentioned studies reinforce the importance of physical, semantic, and localization information shared as a unified map for accurate scene understanding. Based on [31], environmental conditions can also be shared to further augment an ego vehicle's situational awareness.

Some useful metadata can include information timestamp, calibration information and accuracy, global location of the FSN, unique identifiers of the FSN, and relative pose. Timestamps of captured information can be useful for temporally syncing multiple perspectives, which can help mitigate errors in map fusion tasks and multi-perspective localization tasks. Due to aging and environmental conditions, deviations can occur in camera parameters over time, thus the information captured and shared will contain errors that reduce the accuracy of downstream tasks. By sharing calibration information, the ego vehicle can account for the deviation and make more accurate decisions. In a system with multiple FSNs jointly managing a global map, unique node ID and accurate GPS location can be useful metadata to share. This allows the ego vehicle to accurately geo-locate information shared by the FSN and also compute the relative pose. Additionally, if the relative pose is known by the FSN, then sharing this information can simplify the spatial alignment process. While all metadata types are useful, they serve different purposes, with some being more critical for accuracy and others for system management. As a reference for the reader, Table 3 categorizes the types of information that a FSN can share and describes why it can be useful.

B. SENSOR SELECTION FOR V2X SYSTEMS

A FSN can employ a number of exteroceptive sensing modalities for the perception and mapping of the environment. These sensors can be categorized as either active or passive sensors. Active sensors emit energy into the environment and react to the returned signals, whereas passive sensors react to the energy received from the environment. LiDARs and radars are examples of active sensors, and cameras are examples of passive sensors.

When selecting a sensor configuration for a FSN, several key characteristics must be considered. The field of view (FOV) determines the angular extent of scene capture on a given axis, whereas spatial resolution, measured in pixels per degree (px/deg), defines the information density within that FOV. Temporal resolution, inversely related to frame rate, indicates the discrete time intervals at which information is captured. These factors, combined with other sensor-specific components, determine the sensor's data rate that dictates

TABLE 3. Overview of the Information That a FSN Can Share

Information Type	*Description*	*Justification*
Physical	Object information: 3D geometry (centroid, dimensions, orientation), classification (vehicle, person, cyclist), pose (position, velocity).	Crucial for scene mapping, obstacle avoidance, object tracking, and future state prediction.
Semantic	Semantic information: drivable area (road boundaries), lane topology (lane markings, curvatures), and traffic sign recognition.	Allows for a thorough understanding of the scene by adding contextual information.
Localization	Localization information: known landmarks in the scene with accurate global/shared coordinates.	Helps to accurately localize itself/ego-vehicle in the scene, which improves the overall accuracy of the global map.
Environmental	Environmental information: road surface conditions (friction), ambient weather (humidity, air temperature), recurring events (traffic stops), and non-recurring events (collisions).	Provides more information to the vehicle, improving path planning.
Metadata	Metadata information: visual indicators (turning signals), calibration accuracy (reprojection error), timestamp, and global/shared location of FSN (longitude, latitude, altitude, heading).	Metadata information can help the vehicle verify and improve the accuracy of its own detections.

the volume of information transmission. The sensor's ability to provide absolute distance measurements and semantic information significantly influences its utility in scene interpretation. Distance measurement capability is particularly crucial for accurate spatial mapping, while semantic information facilitates higher-level scene understanding. Practical considerations such as sensor cost, which is influenced by manufacturing maturity and material requirements, play a vital role in large-scale deployment decisions. Additionally, the sensor's resilience to adverse conditions, including low light and inclement weather, is critical for ensuring consistent performance across various environmental scenarios.

These characteristics: FOV, spatial and temporal resolution, data rate, distance and semantic measurement capabilities, cost, and resilience to environmental conditions, are a set of parameters that must be balanced to define a sensor configuration suitable for a specific deployment of the FSN. The chosen configuration must provide comprehensive scene coverage, accurate measurements, and robust performance while remaining cost-effective and suitable for the specific deployment environment.

1) PASSIVE SENSORS

Cameras are exteroceptive sensors that operate passively by capturing different wavelengths of the electromagnetic (EM) spectrum and storing that information as an image. While there are various camera technologies employed in automotive applications, the discussion in this section will revolve around three camera types: RGB, infrared, and neuromorphic. A comparison of images from these three camera types is presented in Fig. 5.

a) RGB cameras: Conventional RGB cameras operate in the visible light spectrum, ranging from 0.4 μm to 0.7 μm, converting photons to an electric charge via an image sensor and subsequently converting it to a digital representation with further processing. A Color Filter Array (CFA) is used to limit specific wavelengths of visible light incident upon the image sensor. A Bayer filter, acting as an RGB CFA, will only allow wavelengths that correspond to red, green, and blue light spectra to be incident on the image sensor. The Bayer image will be passed to an image signal processor (ISP), that will interpolate the full-color information and optimize the image quality [59]. Fig. 4 illustrates the operating principle of an RGB camera and shows how an image is produced.

The FOV of cameras is dependent on the focal length of the lens [60] that's attached and the image sensor size [61]. A larger FOV allows for the capture of more information in the scene. Wide-angle lenses, like those in fish-eye cameras, offer FOV of approximately 180°, ideal for capturing more information closer to the sensor but at the cost of detection range. Conversely, narrow FOV cameras, 60° or less, provide better long-range detection [62]. This ties into spatial resolution, determining the information density within the FOV. Higher-resolution sensors capture more detail, crucial for perception tasks like object detection and localization. Current automotive standards typically use 2-4 MP sensors for fish-eye cameras, and up to 8MP sensors for narrow FOV cameras. Temporal resolution, which is the inverse of frame rate, is equally important for safety-critical applications, with automotive-grade cameras usually operating at 20-30 frames per second (FPS).

RGB camera performance degrades in low-light conditions and adverse weather. Poor lighting can cause motion blur due to increased exposure time, while bright sunlight can lead to over-saturation from specular reflections [14]. RGB camera performance can also degrade in adverse weather. For example in rain, droplets can build up on the lens causing distortions and occluding the image plane [62], [63]. Photons can

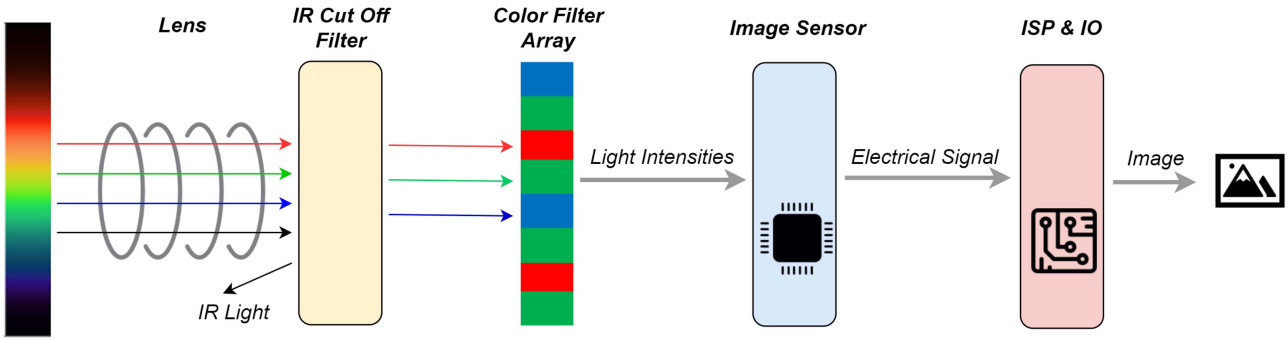


FIGURE 4. Operating principles of an RGB camera, adapted from [14].

(a) RGB Camera Image (b) Thermal Camera Image (c) Neuromorphic Camera Image

FIGURE 5. Comparison of images from different passive sensing modalities: (a) RGB camera, (b) Thermal camera, (c) Neuromorphic camera.

interact with the fog particles causing mie scattering that leads to a reduction in visibility [19]. Despite these challenges, RGB cameras utilizing CMOS sensors are relatively inexpensive to manufacture, leveraging existing semiconductor production equipment. This cost-effectiveness, combined with their rich visual data capture, makes them a vital component in automotive sensing systems, despite their limitations in certain environmental conditions. Their data rates are influenced by several factors such as the spatial and temporal resolution, whether ISP tuning is applied, color depth, and compression standard.

b) Infrared cameras: Near-infrared (NIR), short-wavelength infrared (SWIR), and long-wavelength infrared (LWIR) cameras operate in different parts of the infrared spectrum. NIR cameras work from 0.7 μm to 1.4 μm, and SWIR from 1.4 μm to 3 μm. NIR and SWIR cameras function by capturing reflected infrared wavelengths originating from the sun or other active illumination sources [14]. Similar to RGB cameras, NIR image sensors are also silicon-based, whereas SWIR imagers use Indium Gallium Arsenide (InGaAs) because of its higher sensitivity above 1.1 μm. Unlike NIR and SWIR, long wavelength IR (LWIR), also known as thermal IR, operates in the range of 8 μm to 14 μm and can function through thermal emissions alone. LWIR uses a bolometer, which is a device that measures thermal radiation, and supporting circuitry to convert thermal radiation into electrical signals. The principle of thermal sensing is described in [64].

The FOV of thermal cameras is constrained by manufacturing challenges and cost considerations. Unlike RGB cameras that use standard glass for optics, thermal cameras require specialized lens materials, such as germanium, that is transparent to infrared radiation. These materials are significantly more expensive and difficult to shape into wide-angle lenses. Consequently, thermal cameras are typically designed with narrower FOVs, balancing performance with cost feasibility. The spatial resolution of an IR optical system is dependent on the specifications of its IR detector [65]. Wilson et al. [64] provided a survey on recent advancements in thermal imaging. This study shows that recent technological advancements have enabled the development of thermal cameras capable of capturing images with megapixel-level spatial resolution. Thermal cameras are widely used in military applications, as a result, thermal cameras operating at 60/30 FPS are export-controlled by the U.S. government; however, export control laws do not apply to thermal cameras less than 9 FPS [66].

Compared to RGB cameras, thermal cameras are more expensive due to the imager architecture and lens materials. However, the cost of detectors has decreased with the reduction in pixel pitch size of microbolometers, making the lenses account for a significant portion of the overall cost [67]. IR sensors typically do not suffer from performance degradation in low light conditions as they do not require an external light source. Additionally, the effects of glare are also mitigated in LWIR cameras [14]. This makes it a

valuable contender for low-light imaging. Similarly, thermal cameras have a higher resilience to adverse weather than RGB cameras. Rivera Velázquez et al. [68] identified the resilience of thermal imaging in extremely foggy conditions. YOLOv5 was used to generate a detection rate ranging from 10 m-15 m meteorological optical range (MOR). It was concluded that thermal cameras have a higher detection rate of pedestrians than cars since the detection rate is directly linked to the temperature of the target. The data rate is determined similarly to an RGB camera, with the factors of influence being temporal resolution, spatial resolution, bit depth, and compression standards.

c) Neuromorphic cameras: Neuromorphic cameras are a bio-inspired sensor that mimics the retina of the eye [69]. Unlike traditional frame-based cameras that send information at a fixed frame rate, neuromorphic cameras, also known as event cameras, utilize an asynchronous pixel design that triggers a pixel when there is a brightness level change in the scene. Similar to RGB image sensors, event camera sensors are made of silicon and operate on the visible light and NIR spectrum [14]. The lack of an IR filter on the neuromorphic camera allows more light to be captured by the sensor. Neuromorphic cameras provide a stream of events rather than an image frame, however, these events can then be accumulated over a window of time to generate frames.

Modern neuromorphic cameras offer FOV and spatial resolution comparable to RGB cameras, while their asynchronous design enables microsecond-level temporal resolution [69]. This high temporal precision is particularly beneficial for autonomous driving, allowing the capture of objects with high relative velocity with very little motion blur [69].

Event sensors, being silicon-based, are cost-effective. They also offer low data rates when the scene is static, as they only transmit data upon detecting changes. These cameras perform well in low-light conditions due to their logarithmic-scale photoreceptors detecting brightness changes even in near-darkness [69]. Neuromorphic cameras offer a unique advantage in terms of data rate, particularly in static scenes where they transmit very little information. This event-based data stream contrasts with the frame-based output of traditional cameras, which generate a constant stream of data regardless of scene changes. While the low data rate is beneficial for processing efficiency, it also requires specialized algorithms to interpret the event data.

d) Semantic and distance measurements: Passive sensors excel at capturing semantic information, however, these sensors cannot measure distances directly. RGB cameras are excellent at capturing rich semantic information from different color bands, making them ideal for complete scene comprehension. Thermal images can also provide sufficient information for segmentation tasks, as seen in [70]. Similarly, for neuromorphic cameras, Jia et al. [71] showed that events captured from a neuromorphic camera can also be extended to infer semantic information from the scene. In general, distance measurement is a difficult task for 2D sensors since they don't provide absolute 3D measurements. Some deep learning methods utilize prior knowledge such as camera intrinsic and extrinsic parameters, or object size to allow the network to learn pixel and distance relationships in the scene to estimate 3D information from 2D sensors [72], [73].

2) ACTIVE SENSORS

Active sensors employ a time-of-flight (TOF) principle to measure depth, thus generating a 3D representation of the environment. The three different TOF principles are: pulsed TOF, amplitude-modulated continuous-wave (AMCW) TOF, and frequency-modulated continuous-wave (FMCW) TOF, which is described in detail in [74]. This allows these sensors to get accurate measurements of distances from objects in the scene. This section will discuss the two commonly used active sensors in automotive applications: LiDARs and radars.

a) LiDAR: Automotive LiDARs are an active sensing modality that uses the TOF principle with either 905 nm or 1550 nm lasers to image the environment [75]. LiDARs provide a 3D representation of the scene as a point cloud, which is a collection of points that contain (x, y, z) distance relative to the sensor in metric space along with the intensity of each return. The two most common LiDAR imaging strategies are either scanner-based or detector array-based [76]. Scanner-based LiDAR operates by repositioning the laser spot on the target by varying the angular direction. Detector array LiDARs utilize an array of detectors along with an active illumination scheme to generate a point cloud of the environment. Royo et al. [76] provided an extensive discussion on the fundamental operating principles of various LiDAR technologies, including scanner-based systems such as; rotor-based mechanical scanners, microelectromechanical (MEMS) scanners, and optical phased arrays (OPA), as well as detector array LiDARs like flash imagers and AMCW cameras.

LiDAR's FOV varies depending on their imaging strategy. Rotor-based mechanical scanner LiDARs offer a 360° horizontal FOV, but their vertical FOV is typically limited to 20°–40°, depending on the number of parallel detectors and emitters [77]. MEMS, OPA, and detector array LiDARs generally have a more restricted horizontal FOV compared to rotor-based systems [16]. The spatial resolution of LiDARs is inversely related to the angular resolution of their lasers. MEMS, OPA, and detector array LiDARs achieve higher spatial resolution due to their lower angular resolution, outperforming rotor-based mechanical scanner LiDARs in this aspect. Temporal resolution in LiDARs also depends on the imaging strategy. Scanner-based systems typically operate at 10-25 Hz [16], limited by their mechanical components. Detector-based solutions can achieve higher temporal resolutions, varying with the scanning angle, as they lack moving parts.

LiDAR systems perform well in low-light conditions since they actively emit lasers, however, adverse weather, particularly fog and rain, can significantly impact their performance. In a study conducted by Abdo et al. [78] the detrimental effects of fog on LiDAR performance were identified. Similarly,

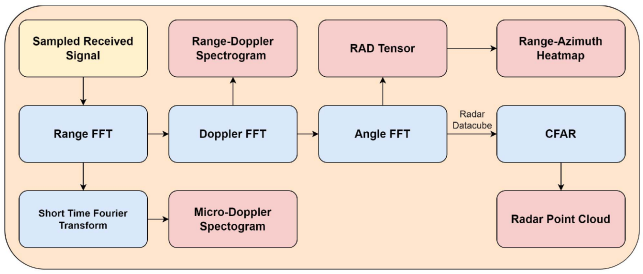


FIGURE 6. Example of how radar ADC signal can be processed (blue) to output various radar data formats (red), adapted from [82].

in [19], the negative impact of rain on LiDAR performance is explored. This degradation is primarily due to water absorption phenomena [79] and Mie scattering effects. The cost of LiDAR systems varies depending on materials and manufacturing processes. For instance, 1550 nm LiDARs use expensive materials like Indium Phosphide (InP), Gallium Arsenide (GaAs), and InGaAs [80]. Additionally, the manufacturing process requires extremely tight tolerances, further increasing production costs.

b) Radar: Radar is another active sensing modality that operates using a TOF principle. Automotive radars typically use a linearly chirped FMCW TOF configuration in the 77-81 GHz frequency bands [81]. This configuration transmits an electromagnetic wave at a frequency that is linearly increased over time. Zhou et al. [81] explained how the range, Doppler, and angle information can be extracted from the reflected waveforms. Fig. 6 illustrates the processing pipeline for transforming the raw Analog-to-Digital Converter (ADC) signal into various radar representations, including Range-Doppler (RD) maps, Range-Azimuth (RA) maps, Range-Angle-Doppler (RAD) tensors and 3D point clouds.

Radar systems vary in their FOV and resolution capabilities. While conventional radar, or 3D radar, operates only on the horizontal axis, 4D radar, or imaging radars, provides both horizontal and vertical information. 4D radar offers a 120° horizontal FOV and a 30° vertical FOV. Close-range 3D radars maintain the 120° horizontal FOV, while far-range 3D radars have a narrower 20° horizontal FOV [82]. Spatial resolution in radars, like in LiDARs, depends on the angular resolution. 4D radars generally have better angular resolution compared to 3D radars. Specifically, 3D near-range and far-range radars have horizontal angular resolutions of 4° and 1.5° respectively, while 4D radars achieve 1° resolution in both horizontal and vertical axes [82].

Radars, being active sensors, perform well in low-light conditions. They also demonstrate greater resilience in adverse weather compared to LiDARs or cameras [19]. 3D radar datacubes typically contain about 32 million cells. These cells are organized into 1024 range bins, 1024 velocity bins, and 32 angle bins. Each cell stores a 16-bit value. The data volume is even larger for 4D radar systems. This substantial amount of information results in high data transfer rates. For 3D radar systems, the data rate can reach up to 2.5 Gbps. 4D radar systems, with their additional dimensional data, require even higher transfer rates, potentially up to 10 Gbps [82].

The maturity of FMCW radar technology has led to lower manufacturing costs. This cost-effectiveness, combined with its performance in challenging conditions and ability to measure absolute distance and velocity, makes radar an attractive option for automotive sensing applications.

c) Semantic and Distance Measurements: LiDAR systems provide absolute distance measurements with low angular resolution. These point clouds can also be used to infer semantic information, particularly for ground segmentation tasks [80]. Radar, another active sensor, offers absolute distance and velocity measurements, however, radar systems typically have lower angular resolution compared to LiDAR, resulting in less dense spatial information [82]. While radar datacubes inherently lack semantic information, making it challenging for direct use in perception tasks, methods have been developed to infer semantic data from radar measurements [83]. This process requires additional computational steps specific to the radar datacube. The key difference lies in the data richness: LiDAR provides denser spatial information that more readily lends itself to semantic interpretation, while radar excels in direct velocity measurements but requires more processing for semantic understanding.

To better assist readers in understanding the relative strengths and weaknesses of the sensing modalities, Table 4 offers a comparison of the different modalities, highlighting the key differences.

3) DISCUSSION ON SENSING MODALITIES

To get a complete representation of the scene, an FSN must reliably perceive the environment through a wide FOV. This can be done through a heterogeneous sensor suite with varying sensor FOV or a single sensor with a wide FOV. Depending on the specific deployment of a cooperative V2I system, the sensor suite must be able to provide semantic information with accurate distance estimates to downstream perception tasks. In large-scale deployments, the sensor cost is also a crucial consideration.

The choice of sensors for FSNs depends heavily on the deployment scenario. For urban junctions or blind intersections [84], where the focus is on pedestrians and slow-moving vehicles, a combination of RGB cameras and LiDAR proves effective. RGB cameras offer FOV coverage and rich semantic information, while LiDAR provides accurate distance measurements. This setup balances performance and cost-effectiveness, with manageable data transmission rates due to potential compression and resolution adjustments. In scenarios prioritizing low-light detection, thermal cameras can replace RGB cameras, however, this substitution involves a trade-off with FOV and is less suitable for scenes with spatially distributed dynamic objects. Similarly, for high-speed applications in low-light conditions, such as highways, neuromorphic cameras excel. LiDAR systems are strong contenders

TABLE 4. Comparison of Different Sensing Modalities

Characteristic	*RGB*	*Infrared*	*Neuromorphic*	*LIDAR*	*Radar*
Principle of Op.	Passive. Operates on the visible light spectrum	Passive. Operates on the IR spectrum (reflected/thermal)	Passive. Operates by detecting brightness changes	Active. Operates using TOF principal with lasers	Active. Operates using TOF principal with radio waves.
Field of View (FOV)	Varies. Dependent on lens, ranges from wide to narrow.	Narrower due to lens limitations	Comparable to RGB	Varies with LiDAR type	Varies with Radar type
Spatial Resolution	High. Up to 8MP for narrow FOV	Megapixel-level advancements	Comparable to RGB	High. Depends on angular resolution	Lower than LIDAR, varies with radar type
Temporal Resolution	20-30 FPS (automotive grade)	Up to 60 FPS. Export controlled	Microsecond level	Varies with type.	Varies with type.
Low Light	Degrades due to motion blur	Robust, especially LWIR	Robust	Robust	Robust
Adverse Weather	Degrades.	Resilient compared to RGB	Robust	Slight degradation	Most resilient
Data Rate	Influenced by resolution, frame rate	Similar factors to RGB	Low when static, high with dynamic scenes	High	High (especially 4D)
Distance Meas.	Indirect, requires additional computation	Indirect, requires additional computation	Indirect, requires additional computation	Direct and accurate	Direct and accurate
Semantic Info.	Rich, excellent for scene understanding	Good	Can be inferred from event streams, not as good as RGB	Can be inferred, especially for ground textures	Requires preproc. to infer
Cost	Relatively inexpensive	More expensive due to lens materials	Cost-effective since sensor is silicon-based	Varies, can be expensive	Cost-efficient

overall, offering both horizontal and vertical resolution for comprehensive distance measurements. They complement cameras well by addressing their inability to measure distances directly. While radars are cost-effective and perform well in adverse conditions, their main drawback in FSN applications is the limited vertical information when mounted from an elevated perspective. 4D radars can overcome this limitation but at a higher cost, potentially exceeding that of LiDAR systems. Each sensor type has its strengths and limitations, and the optimal configuration depends on the specific requirements of the deployment scenario, balancing factors such as FOV, resolution, low-light performance, speed of objects, and cost-effectiveness.

To maximize the benefits of a cooperative system, the sensor placement must be optimized to maximize the FOV. Cai et al. [85] investigate optimal infrastructure LiDAR placement for autonomous driving perception using a Realistic LiDAR Simulation (RLS) library within CARLA. Their pipeline simulates various LiDAR placements, evaluates perception accuracy via detection models, and correlates point cloud metrics (density, uniformity) with performance. A greedy algorithm iteratively selects placements maximizing perceptual gain. While computationally intensive and reliant on predefined scenarios, this approach benefits from the RLS library's realism and the efficiency of simulation-based evaluation. Similarly, Jiang et al. [86] introduced a method to optimize roadside LiDAR placement for multi-agent cooperative perception. The authors use a greedy algorithm to maximize scene coverage and 3D detection performance by sequentially selecting positions based on a learned perception predictor model.

Deploying LiDAR extensively for V2X systems is expensive due to the high cost of individual sensors. Liu et al. [87] addressed this challenge by investigating cost-effective V2I systems using roadside LiDAR sensors with varying densities. They introduced V2X-DSI, a benchmark dataset designed to analyze the impact of infrastructure LiDAR density on cooperative perception performance. Initially, the models were trained on high-density LiDAR data and then tested on low-density data to quantify the performance difference. Subsequently, a fine-tuning approach was employed to adapt the models to the low-density LiDAR scenarios. This methodology reveals the trade-off between LiDAR deployment cost and perception accuracy.

C. SENSOR CALIBRATION

Accurate calibration of sensors is a crucial step in cooperative V2I systems. It is a necessary precursor for sensor fusion,

localization, tracking, and mapping. Furthermore, accurate calibration ensures that the estimated relative pose is reliable improving the map fusion process.

Geometric camera calibration estimates the intrinsic and extrinsic parameters of a camera by analyzing features in a scene with known positions in a fixed world coordinate system [88]. The intrinsic parameters are the optical and digital properties of the camera, such as focal length, principal point, skew, and geometric lens distortion [89]. The extrinsic parameters define the transformation from world coordinates, which is a fixed 3D coordinate system in the scene, to camera coordinates, which is a 2D coordinate system in pixel units. This transformation is described by the translation values along the X, Y, and Z axes, and the rotation values (pitch, yaw, and roll) around these axes [90].

A target-based camera calibration method observes a calibration object with known dimensions or geometry in a 3D space. Zhang et al. [90] proposed a method that utilized a 2D planar target with a geometric pattern. In another study, Liu et al. [91] proposed a grid spherical target approach for camera calibration. Other approaches consist of goniometric calibration [92] and diffractive optical element (DOE) [93] calibration. Goniometric and DOE approaches are suitable in factory environments and provide high pixel accuracy, ideal for intrinsic calibration. Conversely, online calibration, refers to a calibration algorithm running on the ego vehicle or the FSN, using scene features or optical flow to constantly estimate the extrinsic calibration parameters. Motion-based online calibration approaches are more commonly used by the ego vehicle. Several deep learning approaches for standard camera model calibration, high distortion camera model calibration, multi-view models, and multi-sensor models are provided by Liao et al. [94].

In a multi-modal perception system, sensor-pair calibration finds the position of sensors relative to each other. This allows for view transformations and projections from different sensors onto the world plane. Domhof et al. [95] proposed a novel method for LiDAR-radar-camera extrinsic calibration, that utilizes a unique calibration target. Chai et al. [96] introduced a novel method that uses a cube with ArUcO markers and a plane fitting method to estimate the transformation between LiDAR and an RGB camera frame. Similarly, in [97], Rashd et al. introduced a unique target board that leverages plane feature extraction from sparse LiDAR data to generate 3D-to-2D correspondences for LiDAR and RGB frames.

For real-time applications in a multi-modal perception system, temporal calibration plays a significant role. This process ensures that the data streams from several sensors are synchronized. Internal temporal calibration exploits encoded sensor information (e.g., timestamps) and external temporal calibration uses an external source of time (e.g., GPS) to synchronize the sensor streams.

Lee et al. [98] proposed a target-based geometric and temporal calibration approach for a radar-LiDAR pair by utilizing pre-selected objects in the scene. Similarly, Wang et al. [99] proposed an online geometric and temporal calibration method for a LiDAR-Camera pair that applies a pose estimation model and line feature extractor to temporally and geometrically calibrate objects.

From a FSN perspective, an online target-less calibration approach is a more viable solution for extrinsic calibration. This approach allows for constant monitoring of any deviation in the extrinsic parameters between sensors. Online calibration approaches used by the ego vehicle which relies on vehicle motion are not suitable for FSN online calibration. Fan et al. [100] proposed a new calibration-free BEV representation framework, by leveraging scene-based features to estimate the camera parameters and thus generate a BEV representation of the scene. In a multi-sensor system with shared views, deep learning-based cross-view or cross-sensor calibration models can be used. These models are extensively reviewed in [94].

D. V2X COMMUNICATION

1) COMMUNICATION TECHNOLOGIES

Data transmission is a crucial aspect of any cooperative system. There are several survey papers on this topic, where the different wireless access technologies are explored for V2X applications. This section aims to provide a brief introduction to these technologies without delving too deeply into the subject as it falls outside the scope of this paper. Additionally, automotive sensor data often has a large footprint, making it common practice to compress the data before transmission and decompress it at the receiving end. However, a detailed discussion of data compression and decompression techniques falls outside the scope of this review.

Clancy et al. [101] outlines the two primary candidate technologies for V2X communication; Dedicated Short Range Communications (DSRC) based on IEEE 802.11p and Cellular V2X (C-V2X) based on 4G LTE and 5G NR standards. Both DSRC and C-V2X aim to facilitate communication for vehicular applications, but they differ significantly in their underlying technologies and capabilities. DSRC is designed for direct V2V safety communication in the 5.9 GHz band, utilizing orthogonal frequency division multiplexing (OFDM) and Carrier Sense Medium Access with Collision Avoidance (CSMA/CA) protocols, similar to Wi-Fi. This makes DSRC a more localized and direct communication method. In contrast, C-V2X supports a wider range of communications including V2I, vehicle-to-pedestrians (V2P), and vehicle-to-network (V2N). C-V2X leverages the existing cellular networks and spectrum, using cellular uplink/downlink and direct sidelink between vehicles. This reliance on cellular infrastructure provides C-V2X with potentially wider coverage compared to DSRC's direct, short-range focus. Additionally, the 5G NR technology enhances 4G LTE with a new radio interface and core network architecture, offering potential improvements in speed and latency for C-V2X. It is demonstrated that C-V2X offers several advantages over DSRC in terms of application coverage, scalability, and evolution, however, neither technology can meet the strict

requirements revolving around reliability and latency [101], [102]. To address these challenges and improve capacity, researchers are exploring various enabling technologies, such as mmWave spectrum [103], massive MIMO antennas [104], new waveforms like Orthogonal Time Frequency Space (OTFS) [105], non-orthogonal multiple access schemes [106], and mobile edge computing [84]. The adoption of these key enabling technologies will be critical in advancing V2X communications.

Despite the advantages of C-V2X over DSRC in terms of broader application and scalability, there are several challenges that need to be addressed, which are explored by Gyawali et al. [107]. At the physical layer, the existing LTE design cannot support high carrier frequencies and vehicle velocities, leading to Doppler effects and inaccurate channel estimation. Proposed solutions to this issue include enhancing the reference signal structure, employing scalable subcarrier spacing, and using extended cyclic prefixes. Another challenge specific to C-V2X is synchronization due to frequent topology changes and timing offsets among neighboring vehicles. Using the global navigation satellite system (GNSS) and synchronization signals among out-of-coverage users on the sidelink can help mitigate these issues. Resource allocation in scenarios with dense vehicular interactions is also a concern for C-V2X. While DSRC's contention-based access mechanism handles resource allocation differently, it too can suffer from collisions in dense scenarios. Potential solutions for C-V2X include semi-persistent allocation, sensing-based allocation, and non-orthogonal multiple access (NOMA) based resource sharing. Resource allocation in scenarios with dense vehicular interactions is also a concern, and potential solutions such as semi-persistent allocation, sensing-based allocation, and non-orthogonal multiple access (NOMA) based resource sharing have been proposed.

The authors of [108] provide a review on V2X, infrastructure-to-everything (I2X), and pedestrian-to-everything (P2X) communications. These communication technologies are used to enhance road safety, traffic efficiency, and enable new smart city applications. The main enablers for these technologies include spectrum allocation, IEEE 802.11p for DSRC, 4G/5G for C-V2X, and national programs, while the lack of international spectrum and standards remains a significant barrier. A key difference lies in the standardization and global spectrum allocation, where DSRC has faced fragmentation, while C-V2X benefits from the more globally harmonized cellular spectrum. Future enabling technologies, such as 5G, mmWave, massive MIMO, NOMA, mobile edge computing, and blockchain, are being investigated to further enhance the capabilities of these communication technologies. The paper also outlines that despite the significant progress in this field the key challenges are related to scalability, robustness, security, privacy, and infrastructure investments, challenges that are relevant to both DSRC and C-V2X, albeit with varying degrees of impact.

2) V2X MESSAGE FORMATS

V2X communication also relies on several message formats to enable cooperative awareness, safety, and efficiency applications. These formats include Cooperative Awareness Messages (CAMs), Decentralized Environmental Notification Messages (DENMs), MAP messages, and Collective Perception Messages (CPMs). These message types are introduced and reviewed in [101], [109], [110], [111], [112].

CAMs are periodically broadcasted by Intelligent Transport Systems (ITS) stations at a frequency of 1-10 Hz, adapting based on changes in vehicle dynamics. They contain basic vehicle status information, such as type, position, speed, acceleration, and heading, enabling ITS stations to be aware of each other's presence and state. CAMs support a wide range of cooperative awareness and safety applications, such as collision warning, lane change assistance, vehicle platooning, and adaptive cruise control. In essence, CAMs provide a continuous stream of basic awareness data, similar to a regular heartbeat signal for each vehicle.

DENMs, in contrast to the periodic nature of CAMs, are event-triggered messages generated when an ITS station detects a noteworthy event, such as an accident, road hazard, or adverse weather condition. They are disseminated and relayed by receiving ITS stations until a cancellation DENM is received or the event validity duration expires. DENMs provide critical safety information about events and can be used for road hazard warnings, emergency vehicle warnings, and traffic condition warnings. While CAMs establish a baseline awareness, DENMs serve as urgent alerts for specific, potentially hazardous situations.

MAP messages, often used in conjunction with Signal Phase and Timing (SPaT) messages, provide detailed road topology information for intersections or road segments. They describe lane-level geometry using node, link, and lane elements to represent the road network. MAP messages are typically broadcasted by FSNs periodically or on-demand, enabling vehicles to locate themselves within the road network and understand the layout. These messages support intersection safety applications, eco-driving, green light optimal speed advisory (GLOSA) [113], and localization enhancement. Unlike CAMs and DENMs which focus on dynamic information, MAP messages offer static infrastructural context.

CPMs share information about perceived objects and road topology, enhancing perception range and accuracy by fusing data from multiple sources. They are broadcasted by ITS stations periodically at 1-10 Hz with adaptive triggering based on detected object dynamics. CPMs include mandatory containers for sensor information and perceived objects, as well as optional containers for originating vehicle or RSU information. These messages enable advanced cooperative applications, such as collective perception, cooperative sensing, and automated driving. CPMs can be seen as an extension of CAMs, going beyond basic vehicle status to include environmental perception data, enabling a shared understanding of the surroundings.

Each message format has its strengths and limitations. CAMs and CPMs are single-hop broadcasts, limiting their direct reach but reducing network congestion compared to multi-hop messages. DENMs, being multi-hop, can disseminate critical information over a wider area. MAP messages are typically broadcasted by the FSN, making their availability dependent on infrastructure deployment. CAMs enable basic safety applications, DENMs support hazard warnings, MAP messages enhance intersection safety and efficiency, and CPMs enable advanced cooperative sensing and driving. Therefore, the message types cater to different levels of application complexity and information needs. However, they also face challenges, such as potential channel congestion, particularly for frequently broadcasted CAMs and CPMs, reliance on accurate event detection or map data, which is crucial for the effectiveness of DENMs and MAP messages respectively, and the need for high transmission rates and data fusion, especially for CPMs which carry detailed sensor data.

3) V2X MESSAGE SHARING

Information sharing mechanisms are another important consideration for a V2I system to ensure time-sensitive and accurate data transfer while minimizing the network load. As a result, the frequency and content of shared data must be carefully considered to avoid network congestion and redundancy. According to the ETSI standards [112], CPMs can be shared at a frequency of 1-10 Hz. A new CPM should be generated if a vehicle detects a new object or if any previously detected object meets one of these criteria since last being included in a CPM: its absolute position changed by over 4 m, its absolute speed changed by over 0.5 m/s, or it was last included in a CPM 1 or more seconds ago. This approach ensures that excessive CPM generation is avoided while generating CPMs when meaningful updates happen. This rule-based approach to CPM generation aims for a balance between providing timely updates and minimizing redundant transmissions, acting as a baseline strategy for efficient information dissemination.

Several different information-sharing strategies based on CPM value and CPM generation rules have been reviewed and summarized by Cui et al. [43]. CPM generation rule-based methods, similar to the ETSI standard, aim to optimize and dynamically generate CPMs to reduce redundant information. In contrast, CPM value-based methods focus on reducing communication load by requesting confirmation of critical information or selecting data to share based on predicted CPM importance. While rule-based methods rely on predefined thresholds for triggering updates, value-based methods introduce a level of intelligence by prioritizing information based on its perceived significance. Wang et al. [51] surveyed and summarized different information-sharing methods. Some of these approaches include: graph techniques to calculate optimal information distribution weights; selecting specific spatial region data and distributing it based on optimal CPM update frequencies to reduce channel congestion and packet loss; and knowledge distillation models with attention mechanisms to reduce shared data features.

4) V2X RESOURCE ALLOCATION

Effectively sharing network resources among several agents with diverse communication needs presents a significant challenge [49]. Safety-critical messages demand high reliability and low latency, whereas infotainment applications can tolerate delays but require high throughput. This inherent conflict in requirements necessitates careful resource allocation strategies. Although dedicating spectrum exclusively for safety ensures reliability, it often leads to under-utilization. This dedicated approach prioritizes safety above all else but sacrifices overall spectrum efficiency. Resource sharing between agents can optimize utilization but introduces issues like cross-link interference, unfair allocation, added overhead costs, and increased complexity. While resource sharing aims for better spectrum utilization, it introduces challenges related to interference management and fairness, requiring more sophisticated allocation mechanisms. To address these issues, Cui et al. [114] proposed a b-matching-based spectrum-sharing scheme to ensure low latency. This method allocates dedicated resource blocks to agents for safety messages while allowing them to share resource blocks with cellular users for infotainment purposes. This hybrid approach attempts to balance the need for guaranteed resources for safety with the desire for efficient utilization for other applications. Similarly, Saad et al. [115] developed a collaborative multi-agent deep reinforcement learning approach for distributed resource allocation. In this approach, each resource pool is managed by a shared deep Q-network, and vehicles within the same pool collaborate by sharing a common reward function, avoiding competition for resources. This learning-based approach offers a more dynamic and adaptive solution compared to static allocation schemes, allowing the system to learn optimal resource allocation policies based on network conditions and traffic patterns. This strategy significantly outperforms random allocation, particularly in dense network scenarios, by improving the packet delivery ratio. Thus, highlighting the potential of intelligent resource allocation techniques to overcome the limitations of simpler methods, especially in challenging network environments. In another study, Ji et al. [116] proposed a method combining graph neural networks (GNNs) and deep reinforcement learning (DRL) for resource allocation in C-V2X communications. They construct a dynamic graph with communication links as nodes and use the GraphSAGE model to adapt to graph structure changes, extracting low-dimensional features containing structural information from local vehicle observations. Integrating the GNN with a Double Deep Q-Network allows vehicles to make independent, distributed resource allocation decisions based on the graph features. Simulations show that GNNs enhance DRL agent decision-making compared to other methods, with only a modest computational load increase.

III. ENVIRONMENTAL PERCEPTION TASKS

Perception involves processing raw data collected by sensors and transforming it into a concise and understandable format. It is fundamental for multi-agent systems to perceive, understand, and interact with their environment, enabling them to navigate safely and efficiently without human intervention. Even in cooperative systems, single-agent perception tasks remain crucial. Each agent can independently process its own sensor data to extract object-level information using single-agent perception tasks. This information can then be combined through various alignment or fusion methods, effectively enabling a form of late collaboration. Furthermore, in [26], cooperative sensing is described as the unification of perception information from individual agents onto a global coordinate frame.

As a result, this section will delve into the extraction of 3D object data from diverse sensors, obtaining semantic comprehension of the environment, conducting object motion prediction and tracking, and different sensor fusion methodologies. For each of the perception tasks, the discussion will cover single-agent perspectives to provide the reader with a foundational understanding of how these approaches work, following their cooperative counterparts. A specific focus will be given to spatial and temporal alignment in Section IV.

A. SENSOR FUSION AND COLLABORATION APPROACHES

Sensor fusion is another important task that needs to be considered for cooperative V2I systems. When sending data from the multi-modal FSN to the ego vehicle, sensor fusion plays a crucial role. By fusing data from different modalities at the FSN before transmission, the vehicle's cooperative perception network receives robust and reliable information. This approach leverages the complementary strengths of various sensor types, to provide accurate object detection and tracking even in adverse conditions. Furthermore, sensor fusion at the FSN level reduces data transmission overhead, as sending individual raw sensor data from each modality is not feasible due to bandwidth and processing constraints. Typically, sensor fusion approaches can be categorized into low-level fusion (LLF), mid-level fusion (MLF), and high-level fusion (HLF) [16].

In low-level fusion, data streams at the lowest abstraction level are spatially aligned and projected using their geometric correspondences before being fed into a deep learning network for perception tasks. The advantage of this approach is that all raw data are retained and fused, thus, reducing latency [118] and improving performance [119]. The authors of [120] proposed a low-level LiDAR-Camera fusion methodology. This method addresses the issues of fusion on sparse LiDAR point clouds. Kim et al. [121] proposed a low and mid-level fusion approach for radar-camera fusion. Abbasi et al. [122] introduced an LLF approach for human tracking and surveillance using a greyscale and neuromorphic sensor.

MLF fuses information on the feature level. Yoo et al. [123] proposed a 3D cross-view fusion network, where LiDAR and RGB feature maps are fused and aligned for 3D object detection. The authors of [124] proposed a LiDAR-camera feature fusion system for 3D bounding box estimation for monocular images.

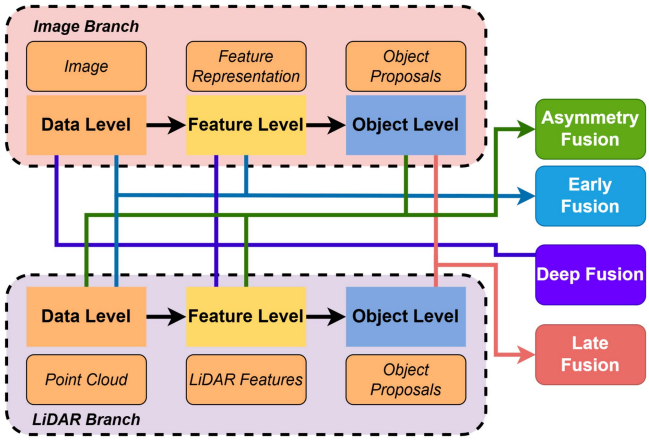


FIGURE 7. Sensor fusion overview adapted from [117].

In late fusion, the detection outputs are spatially aligned and fused on an object level. Typically, these architectures have the lowest data rate requirements, as they operate at the object level [41]. MV3D [125] and AVOD [126] both proposed a late fusion approach for LiDAR-camera fusion. MV3D [125] proposed a late fusion framework that leverages multiple LiDAR views and camera input for 3D object detection. AVOD [126] proposed a late LiDAR-camera fusion methodology, where the architecture leverages feature maps in the LiDAR BEV map with an image for region proposals that can be used for 3D object detection.

These fusion paradigms extend to collaboration schemes. Early collaboration, or LLF, involves transmitting raw sensor data from individual agents and aligning before performing downstream tasks. Intermediate collaboration entails sharing feature-level information for downstream processing, while late collaboration focuses on sharing only object-level information for fusion.

Although HLF, MLF, and LLF are common, the authors of [117] introduced the concept of asymmetry fusion that treats different data branches with different privileges. For example, fusing object proposals from one branch and feature proposals from another branch is an example of asymmetry fusion. This approach allows for the retention of the strongest features from each data branch, resulting in improved robustness and performance. Fig. 7 depicts the various sensor fusion approaches outlined in [117].

One of the challenges of multi-modal sensor fusion is the data representation of each sensing modality. For example, consider images that are 2D representations, and 3D point cloud representations. The challenge arises when inferring 3D information from a 2D image, due to the lack of depth information. Thus, depth information needs to be estimated for cameras to project from 2D to 3D, allowing the fusion of

camera information and 3D LiDAR information into a unified space. The inverse approach is projecting 3D LiDAR onto a 2D camera plane map (CPM) using the geometric correspondence information. The authors of [117], [127] discussed these challenges for sensor fusion along with the various data representations for LiDAR and camera.

In recent years, end-to-end (E2E) sensor fusion approaches have gained in popularity. FUTR3D [128] introduced a unified E2E sensor fusion network for 3D detection. This network incorporates a model-agnostic feature sampler that utilizes queries and a transformer decoder to generate 3D detections. Chittia et al. [18] proposed a multi-modal fusion transformer network for LiDAR BEV and monocular camera representation, that utilizes several transformer modules for intermediate feature map fusion.

Late fusion is a practical approach given considerations for data transfer limitations, processing constraints, and implementation complexity.

B. SINGLE AGENT OBJECT DETECTION

Object detection can be considered one of the most important perception tasks for single-agent or multi-agent systems. The goal of object detection, irrespective of the sensing modality, is to provide the spatial location of objects in a given image, point cloud, or data cube. 3D object detection can be carried out on both active and passive sensors. These algorithms use features, shape, and orientation to accurately locate and classify objects in 3D space.

Based on these studies [129], [130], [131], [132], [133] LiDAR object detection can be categorized into projection-based, discretization-based, point-based, and graph-based methods. Projection-based methods operate by projecting the 3D point cloud onto 2D views such as front-view (FV) or BEV. For FV projection [129] the 3D point cloud is projected onto a spherical or cylindrical surface with the sensor at the origin. This is then represented as a 2D dense range image, where each pixel is encoded with features like range, height, and reflectance. Compared to other methods, FV projection is computationally efficient due to its reliance on 2D CNNs. However, a significant drawback is the potential loss of 3D spatial information during the projection process. Furthermore, object scale variations and occlusions can pose challenges in FV projections. Similar to FV, BEV projection benefits from the efficiency of 2D CNNs. 2D CNNs are then used on these projected views for object detection. While both FV and BEV projections offer computational advantages, they inherently involve a loss of 3D information, a limitation not shared by point-based methods that operate directly on the 3D point cloud.

Discretization methods include voxelization and sparse discretization approaches for object detection. Voxelization methods discretize the point cloud into a 3D grid of voxels, where points are grouped into voxels. 3D CNNs are then applied to the voxelized representation for object detection. This method provides a structured 3D representation, allowing 3D CNNs to effectively learn spatial features, a capability that projection-based methods lack. However, a key disadvantage of voxelization, especially dense voxelization, is its high computational cost and memory requirements. In contrast to projection methods, voxelization retains more 3D information but at the cost of increased computational burden. Sparse discretization methods exploit the sparsity of LiDAR data by using efficient sparse convolution [134] operations for object detection [131]. Sparse discretization offers an improvement over dense voxelization by reducing computational overhead while still maintaining a 3D representation. This makes it more efficient than standard voxelization but potentially more complex to implement than projection-based methods. Fixed voxelization approaches can also suffer information loss in terms of point dropouts. This information loss, although different in nature, is a shared concern with projection-based methods.

Point-based approaches operate directly on the raw point clouds and use architectures like PointNet [135] to learn point-wise features in the point cloud. A significant advantage of point-based methods is their ability to process the raw point cloud directly, avoiding the information loss inherent in projection and some discretization methods. Segmentation is performed on the point cloud to generate foreground and background masks which are then coupled with the point-wise features to generate object proposals [133]. While preserving the original point cloud structure, point-based methods can be computationally intensive compared to projection-based methods, especially when dealing with large point clouds.

Graph-based approaches represent the point cloud as a graph, with points as nodes and edges connecting neighboring points. These methods excel at capturing the local structure and connectivity of the point cloud, offering a different perspective compared to voxelization which focuses on grid-based structures. These approaches use graph neural networks (GNNs) to learn node features by aggregating features along edges. Max pooling operations or Multilayer Perceptrons (MLPs) can be used for feature extraction and then subsequently used for object detection [130]. However, a major challenge with graph-based methods is their high computational complexity and the difficulty in defining effective graph convolution operations, making them potentially less suitable for real-time applications compared to projection-based methods. Fig. 8 illustrates some of the different methods for LiDAR object detection.

Radar object detection can either be classed as point-cloud-based methods or pre-constant false alarm rate (CFAR) data-based methods, that use range-angle (RA) maps, range-Doppler (RD) maps, or range-angle-Doppler (RAD) tensors. Fig. 9 illustrates a number of approaches for radar-based object detection as outlined. Point-cloud-based methods use PointNet variants to extract features and perform detection on the radar point clouds. These methods are directly compatible with LiDAR-based 3D detection methods, making it easier to develop multi-modal sensor fusion techniques. 3D radars provide sparse point clouds that can limit detection performance. This sparsity is a key differentiator from LiDAR, where point

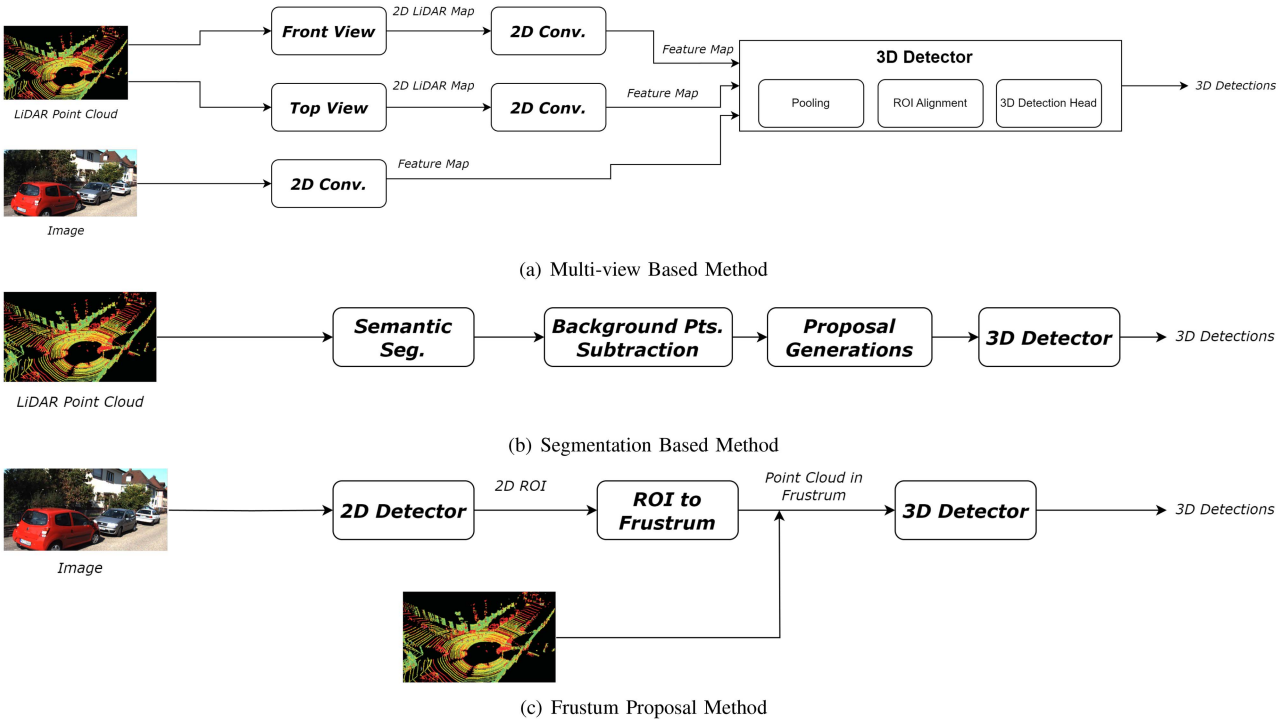


FIGURE 8. Overview of some LiDAR 3D object detection approaches, adapted from [132]. (a) Multi-view-based method that uses various views in the LiDAR point cloud and an RGB frame to generate 3D detections in a LiDAR point cloud. (b) Segmentation-based approach for generating 3D detections in a LiDAR point cloud. (c) Frustum proposal approach projects a frustum from an image onto a LiDAR point cloud to generate 3D detections.

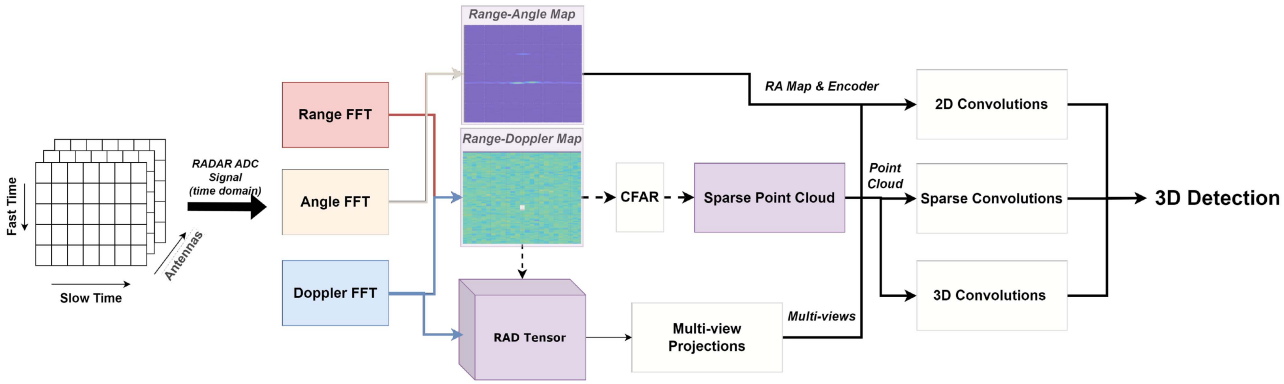


FIGURE 9. Overview of object detection methodologies for various radar data types, adapted from [81], [82]. The purple blocks denote different radar data formats, while the yellow blocks represent the corresponding data processing methods.

clouds are typically denser, potentially leading to lower accuracy for radar point-cloud methods compared to their LiDAR counterparts.

The use of 4D radars aims to bridge the performance gap with LiDAR by increasing point cloud density, but it also increases the complexity and cost of the sensor. Pan et al. [136] introduced RaTrack, focusing on a radar object detection and tracking approach using 4D radar point clouds. This paper overcomes the need for 3D bounding box estimation which can be challenging for sparse radar point clouds, instead, it uses motion segmentation and clustering to detect objects and uses estimated scene flow vectors to help with tracking. This approach offers a way to mitigate the limitations of sparse radar data without relying solely on denser point clouds, presenting an alternative to direct 3D bounding box estimation. Kohler et al. [137] use a grid rendering approach, called KP-BEV, to convert the radar point cloud to a BEV image for CNNs. This method shares similarities with the projection-based methods in LiDAR, leveraging the efficiency of 2D CNNs after transforming the point cloud. Kernel point convolutions are used to fuse features from the BEV rendering, resulting in 3D detections. Compared to directly processing

the point cloud, this BEV approach might lose some fine-grained 3D information but can be more computationally efficient.

Pre-CFAR methods use 2D CNNS on the RA map, RD map, or RAD tensor data representations for object detection. The primary advantage to these approaches is that they can accurately capture the spatial distribution of Doppler velocities as well as the position and they can work with lower-resolution radar hardware, unlike point cloud methods. This ability to work with lower-resolution data and capture velocity information is a significant advantage over point-cloud methods, especially when considering cost and the unique information provided by radar. Kim et al. [138] demonstrated a deep learning-based approach for object detection and tracking using RA maps. Here, the RA map is preprocessed and passed into a YOLOv3 network, that extracts features at different scales. Predictions are made on features using a feature pyramid network (FPN) at different resolutions, which is fused to obtain information about the detected objects. This approach, using RA maps and YOLOv3, is computationally efficient and leverages well-established 2D object detection architectures, similar to the projection-based methods in LiDAR. DAROD [139] introduced an object detection model for RD maps. Here the authors use 2D CNNs for spatial feature extraction, and a region proposal network (RPN) is used on the feature map for object proposals. The radial velocity of objects is used as an additional feature to handle translation variance. By focusing on RD maps, this method explicitly incorporates velocity information, a key advantage of radar over passive sensors like cameras. This approach proves to be lightweight and computationally efficient. The computational efficiency of RD map-based methods makes them suitable for real-time applications, potentially more so than some of the more complex point-cloud-based methods. Zhang et al. [140] introduced RADNet, that performs 3D object detection directly in the RAD space. A coordinate transform layer can be used for 2D detections in the Cartesian coordinate system. Using RAD tensors as the input ensures that no data, such as azimuth, is discarded leading to better detection performance. Utilizing the full RAD tensor aims to maximize the information used for detection, potentially leading to higher accuracy compared to methods using only RA or RD maps. However, the RAD inputs have high dimensionality, resulting in computationally demanding algorithms. This high dimensionality and computational cost are a trade-off for the increased information and potential accuracy, making it less suitable for resource-constrained applications compared to RA or RD map methods. Paek et al. [141] uses RAD tensors from a 4D radar for object detection. The tensor is converted to a Cartesian coordinate system to extract an ROI, from which features are extracted using a 3D sparse convolution network or a 2D dense convolution network. The detection head outputs 3D bounding boxes from the fused feature maps. This approach combines the benefits of using the full RAD tensor with techniques like sparse convolutions to manage the computational complexity, aiming for a balance between accuracy and efficiency.

While active sensors can provide absolute measurements of the scene, passive sensors cannot. As a result, the task of 3D object detection becomes a challenge when working with images. Nevertheless, monocular 3D object detection is a topic of interest for vision-centric autonomous driving. Cameras provide rich visual information, crucial for holistic scene understanding, and are scalable in terms of data collection and model training.

Kim et al. [142] provided an extensive survey on monocular 3D object detection methodologies, categorizing them into multi-stage methods and end-to-end learning approaches. Multi-stage methods include 2D detection-driven approaches, 3D shape information methods, depth estimation methods, and representation transform methods. End-to-end (E2E) learning approaches encompass direct regression and 2D-3D correspondence methods. Some of these methods are depicted in Fig. 10.

2D detection-driven methods, such as GS3D [73], employ a 2D detector to generate region proposals, which is coupled with prior knowledge such as object size, location, or shape. With this information and certain assumptions, such as all objects exist on the ground plane, the task of 3D detection becomes a loss minimization task. Since these methods use pre-existing 2D detectors, they are easy to implement, however, their performance is heavily reliant on the accuracy of the 2D detectors. Depending on 2D detection accuracy is a significant limitation, as errors in the initial 2D detection will propagate to the 3D detection stage, a problem not shared by methods that directly estimate 3D properties. 3D shape information methods, such as Mono3D++ [143], uses shape template matching for 3D object detection. While this approach provides more accurate 3D shape representation, offering an advantage over methods that rely on simpler geometric assumptions, it requires a large database of 3D models and can be computationally expensive, making real-time deployment challenging. The computational cost and reliance on a comprehensive 3D model database are significant drawbacks compared to methods that learn shape implicitly from data.

Depth estimation approaches, such as MonoGRNet [144], fuse depth features with 2D proposals to generate 3D detections. These methods can leverage existing depth estimation techniques to improve 3D localization, potentially leading to more accurate 3D positioning compared to 2D detection-driven methods that rely on assumptions. However, the multi-stage pipeline can be complex and computationally expensive, affecting real-time processing speeds.

Representation transformation methods convert from image domain to point cloud or BEV representations for 3D detection. Pseudo-LiDAR approaches convert from an image representation to a point cloud format using a depth map, on which a point-based approach can be used for 3D detection. The advantage of this approach is that it is computationally

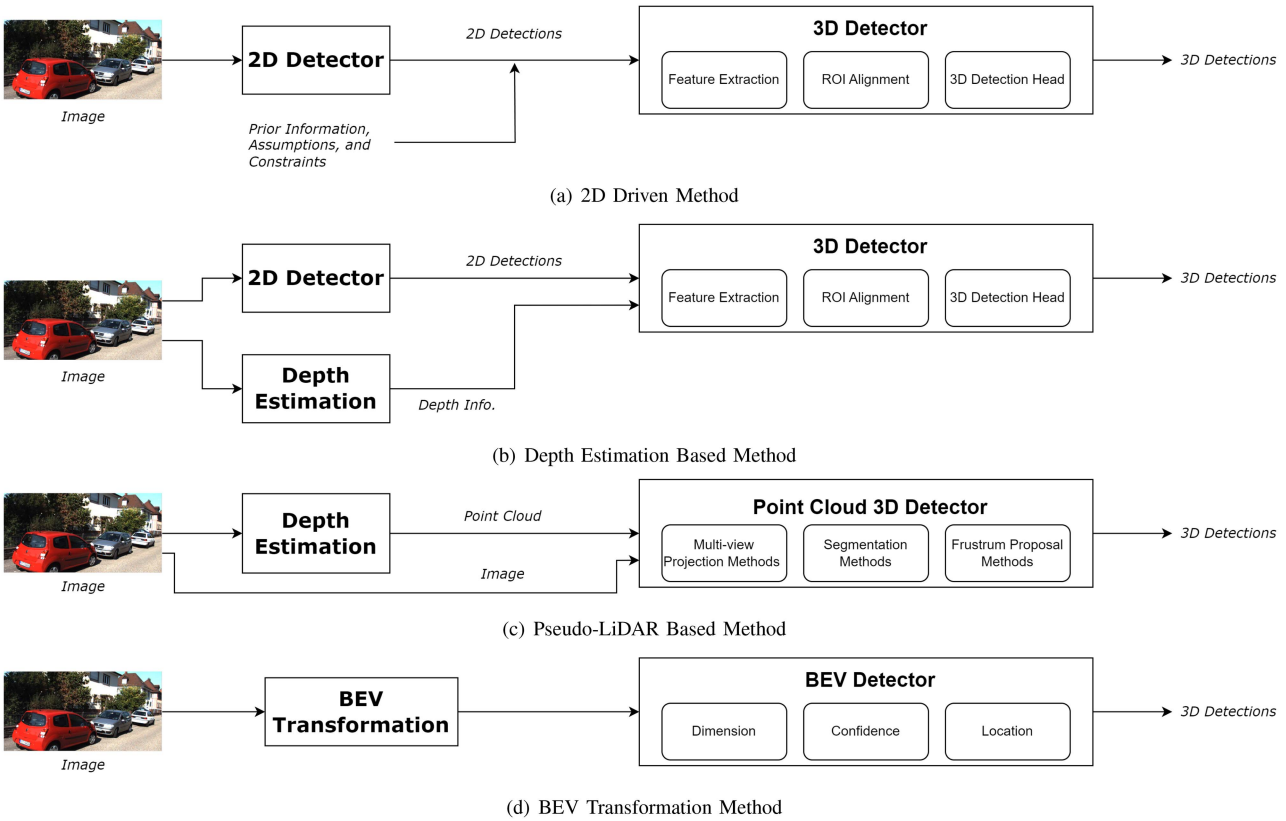


FIGURE 10. **Overview of some approaches used for monocular RGB 3D Object Detection, adapted from [145] (a) A 2D-driven method that couples prior assumptions or constraints with 2D detections to generate 3D detections. (b) Depth estimation-based method that utilizes depth information with 2D detections to generate 3D detections. (c) Pseudo-LiDAR-based method that uses a depth estimator to generate a pseudo point cloud and is coupled with an image to generate 3D detections. (d) BEV transformation method that performs 3D detections in the BEV space.**

cheaper to infer a 3D location in a point cloud than an image. The BEV space can also be utilized for 3D detections. The input image can be transformed into a BEV space, with or without the help of camera intrinsic and extrinsic parameters. The 3D detections are then performed in the BEV space which allows for a top-down perspective, simplifying certain 3D detection tasks and facilitating fusion with other sensor data that is naturally represented in BEV, such as LiDAR.

E2E learning approaches aim to simplify the 3D detection pipeline. In contrast to the multi-stage methods, E2E approaches aim for a more direct mapping from input to output, potentially leading to greater efficiency. Direct regression methods estimate the 3D location and pose parameters from monocular images. These methods can achieve real-time performance due to their simplified architecture. This real-time capability is a significant advantage over many multi-stage methods. However, they may struggle with the nonlinearity of the rotation space and require post-processing for refinement. The challenges with rotation representation and the need for post-processing are limitations that need to be addressed for robust performance. 2D-3D correspondence methods focus on predicting 2D projections of 3D key points and solve the Perspective-n-Point (PnP) problem for 3D pose estimation. These methods are more efficient compared to multi-stage methods but are sensitive to occlusion and truncation of key points, affecting their accuracy in complex scenes. While offering efficiency gains, the sensitivity to occlusion and truncation highlights a vulnerability in challenging real-world scenarios.

Chen et al. [145] discussed stereo and multi-view approaches for 3D detection. These approaches leverage multiple views to overcome the inherent depth ambiguity of monocular vision, offering a significant advantage in terms of depth perception. Stereo-based approaches leverage the disparity between two views to obtain stronger depth perception, which can be coupled with 2D detectors on both views for accurate 3D localization. These methods can provide more reliable depth information compared to monocular methods, making them inherently more accurate in estimating 3D locations. The need for precise calibration and the increased computational load are practical considerations when choosing between monocular and stereo approaches. Multi-view approaches tend to utilize multiple different camera views and transform them into the BEV space. This allows for a complete scene representation and stronger 3D proposals. By integrating information from multiple viewpoints, multi-view

approaches can achieve a more comprehensive understanding of the scene and improve the robustness of 3D detection. Lift-Splat-Shoot (LSS) [146] is a multi-view approach for generating semantic information in the BEV space. While the architecture is not explicitly designed to provide 3D detection, but instead focuses on the segmentation task, the outputs from the "Lift" and "Splat" stages can be used to generate 3D detections in the BEV space. The LSS approach enables the model to perform various tasks such as object detection, semantic segmentation, and motion planning directly in the BEV space from multi-camera inputs. The multi-task capability and the direct generation of BEV representations are key strengths of LSS. Moreover, the LSS architecture is designed to be robust to camera calibration errors and missing cameras, making it adaptable to real-world scenarios. Robustness to calibration errors and missing data is a significant advantage in practical deployments compared to traditional stereo vision systems.

C. COOPERATIVE OBJECT DETECTION

Object detection advancements in recent years have led to the development of unified architectures that perform downstream tasks collaboratively. This integrated approach offers enhanced performance and optimization. The key distinction between collaborative and single-agent approaches lies in the centralized nature of collaborative systems. These systems handle the entire task within a single architecture, explicitly addressing factors like compression and communication. While these unified approaches offer numerous benefits, they often tightly couple performance to specific sensor types and model choices. This inherent dependence can create challenges when scaling to heterogeneous sensor systems, both within individual vehicles and across multiple agents.

1) EARLY COLLABORATION

Arnold et al. [147] introduced a cooperative V2I object detection framework, where the authors examined early and late fusion schemes. With the early fusion scheme, raw point clouds from spatially diverse sensors are combined and transformed into a global coordinate system. This fused point cloud is then fed into Voxelnet 3D object detection model for detection. Early fusion leverages the full resolution of the raw data, potentially preserving more information compared to late fusion. Similarly, in the late fusion scheme, point clouds from each sensor are passed through a detection model, and the generated objects are fused. A Non-Maximum Suppression (NMS) algorithm is used to remove multiple detections of a single object. Late fusion is generally simpler to implement and less demanding in terms of communication bandwidth, as only the detection results need to be transmitted. The framework was tested on a simulated dataset, and the authors concluded that the early fusion scheme outperforms late fusion in terms of detection accuracy, especially for occluded or low-visibility objects. This improved accuracy is a key advantage of early fusion, resulting from the richer information being available before detection. However, this improvement in performance is accompanied by an increased communication and computation cost. The higher communication and computation costs are significant drawbacks of early fusion, especially in resource-constrained environments.

Cooper [148] is another cooperative perception framework that performs early fusion on 3D point clouds. Cooper fuses the sensor data collected from different positions and angles of agents in the scene and applies a point cloud-based 3D object detection method called Sparse Point-cloud Object Detection (SPOD) to detect objects in the fused point clouds. The SPOD method is designed to work with both high-density and low-density point clouds, making it suitable for heterogeneous LiDAR inputs. Cooper's adaptability to varying point cloud densities is a key advantage in a cooperative setting where vehicles might have different sensor configurations. Firstly, point clouds from different agents are transformed into a common coordinate system using GPS and IMU data and fused to create a more comprehensive representation of the environment. This representation is passed into the SPOD network which detects objects in the fused point cloud. The authors also analyze the communication requirements of the framework and show that it is feasible to transmit ROI point cloud data using existing vehicular network technologies like DSRC. The main advantage of Cooper is its ability to enhance object detection performance by leveraging fused point cloud data from multiple connected vehicles, providing a more complete view of the environment and mitigating occlusions compared to single-vehicle perception. However, the framework's effectiveness may be limited by the quality of the point cloud fusion and the communication bandwidth available in real-world scenarios. The reliance on accurate point cloud registration and the potential strain on communication bandwidth are practical challenges.

Guo et al. [149] introduced a method for monocular 3D vehicle detection in a cooperative V2I system. Firstly a K-means approach is used to cluster the contour points of the vehicle's bottom edge into different sets, with each corresponding to a side of the vehicle. The camera intrinsic parameters and the ground plane equation are used to calculate the 3D position in the 2D frame. Using the vehicle-ground contact points, the 3D bounding box vertices are derived, and this allows for the orientation and dimension to be estimated. A maximum a posteriori estimation refines the 3D bounding box by optimizing the vehicle's pose and dimensions to maximize their posterior probability conditioned on the 2D bounding box. The main advantage of this approach is its ability to perform 3D object detection without requiring 3D labels, making it easier to deploy in new environments; this reduces the annotation burden compared to methods requiring full 3D supervision. However, the accuracy of the 3D bounding boxes may be affected by the quality of the 2D segmentation masks and the presence of occlusions. The dependence on accurate 2D segmentation and susceptibility to occlusions are limitations shared with other monocular vision-based methods.

2) INTERMEDIATE COLLABORATION

Yu et al. [150] introduced a V2I cooperative 3D object detection system using feature flow prediction. Firstly, the feature flow net (FFNet) generates a feature flow from sequential FSN frames, which is compressed and transmitted to the vehicle. The feature flow serves as a prediction function to describe the changes in infrastructure features over time, allowing the vehicle to predict the infrastructure feature at any future timestamp aligned with the vehicle data. The vehicle then decompresses the feature flow and reconstructs the FSN perspective features, aligning them with the vehicle's features for object detection. A self-supervised approach is used to train the feature flow generator by constructing ground truth features from nearby FSN frames. The main advantage of FFNet is its ability to overcome temporal asynchrony through the feature flow prediction and reduce communication costs by transmitting compressed feature flows instead of raw data, offering a balance between the high communication cost of early fusion and the potential information loss of late fusion. The effectiveness of feature flow prediction may depend on the complexity and dynamics of the scene, which could be a potential limitation as rapidly changing or highly complex scenes might challenge the prediction accuracy.

Carrying on from [148], F-Cooper [151] is a feature-based cooperative perception framework designed to improve object detection performance in connected vehicles using 3D point clouds. F-Cooper represents a move towards feature sharing, aiming to reduce the communication burden of raw data fusion while retaining much of its performance benefits. F-Cooper consists of a Voxel Feature Fusion (VFF) network and a Spatial Feature Fusion (SFF) network. In the VFF approach, features generated through voxel feature encoding are fused and passed to sparse convolutional layers and a RPN to generate 3D detections. In the SFF approach, the voxel features are first processed by sparse convolutional layers to generate spatial feature maps, which are then fused and passed to the RPN. F-Cooper also includes a compression step to reduce the amount of transmitted data. The main advantage of F-Cooper is its ability to improve object detection performance by fusing features from multiple vehicles while reducing the amount of data transmitted compared to raw data fusion methods. SFF achieves similar performance to raw data fusion, while SFF allows for dynamic adjustment of the feature map size for better compression. The drawback is that the architecture relies on accurate alignment of the point clouds, which can be challenging in real-world scenarios with sensor drift and calibration errors. Similar to Cooper, accurate point cloud alignment remains a critical requirement for F-Cooper.

Similarly, CooPercept introduced by Zhang et al. [152] fuses LiDAR point clouds and camera images from multiple CAVs to enhance 3D object detection. CooPercept extends the concept of feature sharing to multi-modal data, leveraging the complementary strengths of LiDAR and cameras. Firstly, a self-data processing module transforms the raw point clouds and images into compressible voxel features, where semantic segmentation scores from camera images are appended to each LiDAR point, and voxel feature encoding. Compressed voxel feature encodings are broadcasted through V2X messages. Then a cross-CAV fusion aligns the received feature maps from various CAVs spatially and temporally before fusing them. The fused feature maps are then passed through a sparse convolutional layer and an RPN for object detection. The main advantage of this approach is the multi-modal fusion of LiDAR and camera data allows for the point cloud to be augmented with semantic information from the camera images. The architecture also demonstrates robustness against sensor drift and adverse weather conditions. The demonstrated robustness to sensor drift and adverse weather is a significant advantage in real-world deployments compared to methods relying solely on a single sensor modality.

Marvasti et al. [153] introduced a feature-sharing cooperative object detection (FS-COD) framework. FS-COD provides another example of feature sharing, focusing on projecting point clouds onto a common image plane for feature fusion. Firstly, the point clouds from the vehicles are aligned into a global coordinate system and then projected into a 2D image plane using a BEV projector. Feature maps generated from the BEV images are transmitted to cooperative vehicles along with GPS information. The received feature maps are aligned and fused to the ego vehicle feature maps. The fused feature map is passed into a CNN for the detections. By sharing partially processed features instead of raw data, FS-COD achieves a balance between performance and communication cost. This balance between performance and communication cost is a key strength of feature-sharing approaches compared to raw data fusion.

3) LATE COLLABORATION

Shi et al. [154] proposed VIPS, a vehicle-infrastructure perception fusion system that enhances vehicle perception by integrating LiDAR data from roadside infrastructure and vehicles. VIPS first independently detects objects of interest from each agent's point clouds. It then tracks the detected objects using a Kalman filter and Hungarian method to temporally align the infrastructure and vehicle frames, handling inconsistent frame rates and missing data. Next, it constructs multi-affinity graphs encoding object locations, classes, sizes, and spatial relationships for both the vehicle and infrastructure detections. An efficient graph-matching algorithm identifies co-visible objects between the two perspectives. Finally, based on the matched object pairs, VIPS optimally transforms and aligns all objects from the infrastructure's view into the vehicle's coordinate frame. This results in a fused perception of the surrounding environment with enhanced range and accuracy. VIPS offers several advantages. It efficiently utilizes limited infrastructure resources by transmitting only compact detection data, scales well to multiple vehicles, and maintains robustness to communication delays and frame misalignment by tracking object motion. However,

VIPS relies on the presence of co-visible objects between the infrastructure and vehicle views for accurate alignment, and the 3D object detection method is computationally expensive.

Zheng et al. [155] introduced a real-time cooperative perception framework designed to improve object detection and tracking for CAVs in challenging urban environments like intersections. Motivated by the limitations of existing cooperative perception methods, particularly concerning data sharing, computational resources, and real-time validation, CooperFuse employs a late fusion approach that considers object detection confidence, kinematic and dynamic consistency, and scale consistency. First, the framework ensures precise temporal synchronization and localization of CAVs using LiDAR, IMU, and high-definition maps. The object data are then fused from agents based on the confidence scores, kinematic and dynamic consistency, and scale consistency. A multi-object tracking module uses motion models to track detected objects, even when occluded. Finally, a feature-based fusion module refines bounding box fusion by incorporating various feature scores, rather than relying solely on confidence scores. Results demonstrate that CooperFuse achieves real-time performance with minimal latency and surpasses baselines in key metrics like average precision, translation error, orientation error, and scale error, particularly in V2X scenarios with heterogeneous detection models.

The studies mentioned above demonstrate that the most suitable method for sharing information while meeting communication bandwidth constraints is to share feature maps or object data. The conclusion highlights the practical limitations of raw data sharing in cooperative perception and the effectiveness of more efficient information representations. This approach ensures that the maximum amount of relevant information is transmitted efficiently. Furthermore, by fusing perspectives from multiple agents, the impact of occlusions and other factors is mitigated, leading to a more complete and accurate representation of objects in the scene. This results in improved performance for 3D object detection. Cooperative object detection architectures like F-Cooper [151], CooPercept [152], and FS-COD [153] demonstrate significant improvements in detection performance compared to the baseline methods. The cooperative sensing capabilities of the network provide a more robust and reliable representation of the environment, reducing false positives and false negatives. Also, relying on multi-agent data ensures that the ego vehicle can maintain object detection and tracking capabilities even if there is sensor failure on the ego vehicle. The redundancy provides a crucial safety mechanism, ensuring continued operation even in the event of individual sensor failures. Thus, improving the overall reliability and fault tolerance of the perception system, which is crucial for ensuring the safety of autonomous driving. Furthermore, FFNet [150] and Cooper [148], proposed techniques to compress sensor data and to transmit only relevant regions of interest, reducing the communication bandwidth requirements. These techniques address the practical challenge of limited communication bandwidth in real-world cooperative systems. These approaches show that efficient sharing of sensor data even in the presence of limited communication resources, enables effective object detection.

D. SINGLE AGENT OBJECT TRACKING

Object trajectory prediction has a crucial role in scene understanding, as it allows the ego vehicle to estimate the future states of objects in the surrounding environment. In a survey conducted by Gulzar et al. [156], the authors examined various models for motion prediction, categorizing them into physics-based and learning-based approaches. Physics-based methods utilize kinematic models, such as the constant velocity (CV) and constant acceleration (CA) models, to predict the trajectory of objects. These models are useful for short-term predictions since they assume objects maintain their current velocity or acceleration over a short timeframe. A key advantage of physics-based models is their simplicity and computational efficiency, making them suitable for real-time applications with limited computational resources. However, this simplicity is also their limitation, as this approach is limited in capturing complex environmental factors or long-term dependencies. Kalman filtering techniques can be used to address uncertainty in object states and physics models, improving their robustness to noise but not fundamentally addressing their inability to model complex behaviors.

Learning-based models rely on data such as scene context, object attributes, and motion history to predict future outcomes. These models can be further classified into sequential models, that consider the temporal evolution of an object's motion, and non-sequential models, that learn patterns from the data distribution. In contrast to physics-based models, learning-based approaches excel at capturing complex behaviors and long-term dependencies by learning from vast amounts of data. Learning-based approaches have the advantage of incorporating environmental influences, such as road structure, traffic rules, and interactions between objects, leading to more accurate and context-aware predictions.

In [157], the authors provided a comparative analysis of the different categories of motion prediction models, while highlighting the trade-off between model complexity and interpretability. The paper states that physics-based models are more explainable but fail to capture complex behaviors, while learning-based models offer better performance at the cost of reduced interpretability. This trade-off is a crucial consideration when choosing a prediction model. While physics-based models offer transparency and ease of understanding, their predictive power is limited. Conversely, learning-based models, while achieving higher accuracy, often operate as "black boxes," making it challenging to understand the reasoning behind their predictions.

Huang et al. [158] provided a review on deep learning approaches for motion prediction in autonomous driving. Various aspects like input representations, scene context refinement, and prediction rationality improvement were discussed. Another topic of discussion was the importance of effectively

representing the scene information, considering the interactions between objects, and integrating map information to enhance the context understanding.

1) COOPERATIVE OBJECT TRACKING

Cooperative object tracking allows for tracking algorithms to leverage spatially diverse sensors and provide more robust and reliable tracking information.

In HYDRO-3D [159] a LiDAR-based hybrid cooperative object detection and tracking framework is introduced. HYDRO-3D leverages historical object tracking information to enhance object detection performance, particularly in areas with occlusion, sparsity, and out-of-range issues. Firstly, the LiDAR point cloud is passed into the object detection feature extraction module using a PointPillar [160] backbone and a transformer to extract and fuse features from multiple agents. Then the object detection and tracking feature fusion module concatenates the object detection features with historical object tracking features generated by the object tracking trajectory network to predict 3D bounding boxes. The object tracking module uses a Kalman filter and Hungarian algorithm [161] to associate detected objects and update their trajectories. The main advantage of HYDRO-3D is its ability to utilize historical object tracking information to improve detection performance, especially in challenging scenarios, however, the framework's complexity may increase computational requirements and latency.

Su et al. [162] introduced a cooperative 3D multi-object tracking (MOT) framework that leverages both deep learning-based object detections and spatial-temporal trajectory information from multiple connected vehicles. The authors proposed a novel data association strategy that exploits complementary information from neighboring vehicles to compensate for detection failures and occlusions, enabling more robust and persistent tracking. The architecture consists of a cooperative detection stage, followed by a cooperative tracking stage. In the detection stage, a CNN backbone is used to extract features from the individual point clouds. These feature maps are compressed and shared with nearby vehicles over a communication link, providing a good balance between information density and communication efficiency. The received feature maps are fused with the ego vehicle's feature map using a middle fusion strategy, and the fused feature map is then processed by a detection header network to output 3D bounding boxes of detected objects. In the tracking stage, the ego vehicle aims to associate these cooperative detections with previously tracked object trajectories to obtain temporally consistent identities and state estimates. A CV Kalman Filter is used to predict the current states of previously tracked objects based on their past trajectories. These predicted states are then associated with current detections using a distance-based cost metric and Hungarian matching. Dropped detections due to occlusions, limited sensor FOV, or algorithmic error, can lead to tracked objects being missed in the current frame's detections. To overcome this a complementary data association (CDA) strategy is used, where the key idea is if an object is missed in the ego vehicle's detections, it might still be successfully detected and tracked by a neighboring vehicle with a different vantage point. The key advantage of this approach is that it effectively exploits spatial diversity and redundancy across multiple vehicles to compensate for individual vehicles' sensing limitations. The CDA module enables robust tracking through occlusions and detection gaps by using complementary tracking information from neighbors.

In another study [163], the authors proposed a novel method for controlling the geometry of a multi-robot formation with the primary objective of cooperatively tracking a target with minimum uncertainty. The system architecture integrates a cooperative target estimator (CTE), which is a cooperative estimation framework, based on a particle filter, where each agent fuses target observations received from all formation members to generate a more accurate and robust target state estimate. For each timestep, each agent receives a target observation along with a confidence score. These observations are fused by each agent based on the confidence score, and this approach allows for uncertain observations to be weighted lower during the fusion stage. Furthermore, each agent transmits its own localization confidence, and the CTE uses this confidence to discount observations from agents with high localization uncertainty. This paper shows that cooperative fusion provides more accurate and robust target estimates by leveraging spatially diverse observations, however, with this approach the computation and communication costs increase with more agents in the scene.

Cooperative tracking has also demonstrated significant improvements in tracking accuracy, continuity, and robustness. By using historical tracking information from multiple agents, HYDRO-3D [159] demonstrated resilience to temporary occlusions and out-of-range issues, resulting in robust object tracking performance. Lima et al. [163] illustrated that the cooperative tracking approach minimized the target localization error and improved tracking accuracy while being able to withstand sensor or communication-link failures. Similarly, Su et al. [162] outlined that cooperative tracking provided increased tracking continuity by compensating for missed objects from other agents. The proposed method achieved a higher MOT accuracy, while also reducing the instances of false negatives, indicating better tracking performance.

E. SEMANTIC SEGMENTATION

The task of semantic segmentation is essentially pixel-wise classification. This is an important task for autonomous driving applications as it allows for better scene understanding by categorizing the scene into different classes. Segmentation algorithms allow the processing of semantic scene information, such as lane topology, traffic signs, and drivable areas.

Classical monocular semantic segmentation approaches consisted of thresholding and clustering approaches that use

features in the image to generate location boundaries. Surveys by Ulku et al. [164] and Minaee et al. [165] provide a comprehensive review of the various deep learning architectures used for image-based semantic segmentation. These architectures include fully convolutional networks (FCNs), convolutional neural networks (CNNs) with graphical models, encoder-decoder models, multiscale and pyramid networks, R-CNN based models, dilated convolutional models, recurrent neural network (RNN) based models, attention-based models, generative models, and transformer-based models.

Additionally, Thisanke et al. [166] provided a survey on segmentation approaches using vision transformers. These transformer-based models leverage self-attention mechanisms to capture global context and long-range dependencies. They differ in terms of the specific architectures used, such as the type of attention mechanisms, pyramid structures, and integration with convolutional layers. Advantages of transformer-based models include better global context modeling and handling of long-range dependencies, however, they generally require more data and compute resources compared to convolutional models. Some transformer models like Swin Transformer [167] are designed to be more efficient in terms of computation and memory usage, addressing a key limitation of standard transformers. Thus, making them more practical for resource-constrained applications, although they might still not reach the efficiency of some lightweight convolutional models.

Multi-view segmentation approaches have gained significant popularity in recent years. BEVSegFormer [168] proposes a transformer-based architecture to flexibly perform BEV segmentation from any camera inputs. A shared backbone encodes multi-view images, which are integrated into BEV space using deformable attention between BEV queries and image features. This enables handling arbitrary camera rigs without camera parameters, a significant advantage over methods requiring precise calibration, however, the multi-stage transformer architecture is fairly complex and computationally costly. This complexity can be a drawback compared to simpler, potentially faster, single-view methods, although the improved spatial understanding offered by BEV representations can justify the added computational cost for certain applications. Similarly, M ^{2} BEV [169] proposes an efficient multi-task framework for 3D object detection and BEV map segmentation from multi-view images. 2D image features are projected into a 3D voxel representation assuming uniform depth, which is encoded to BEV features efficiently using a novel spatial-to-channel operator. Task-specific heads then operate on the shared BEV representation to provide 3D detection or BEV map segmentation. A benefit of M ^{2} BEV is its ability to perform both 3D detection and BEV segmentation in a unified framework, increasing efficiency by sharing computations. Furthermore, the 2D-3D projection overcomes the need for expensive depth estimation, making it a more cost-effective approach compared to methods relying on explicit depth sensors.

Several 3D semantic segmentation approaches are commonly used with LiDAR point clouds. In [131] Guo et al. reviewed projection-based, discretization-based, and point-based methods for 3D semantic segmentation and discussed these principles of operations. The PointSeg architecture [170] projects the 3D point cloud into a spherical image representation, on which a CNN is used for point-wise segmentation. The segmentation mask is then projected back onto the 3D point cloud. This approach provides a good balance between efficiency and prediction performance, however, information loss occurs during the representation transformation stage. Milioto et al. [171] introduced RangeNet++, an architecture that provides fast and accurate semantic segmentation on 3D LiDAR point clouds. RangeNet++ takes a range image representation of the point cloud as input and applies a convolutional encoder-decoder architecture to predict per-pixel semantic labels. An efficient post-processing algorithm is used to address some issues stemming from the range image representation. The semantic labels are transferred from pixels back to their corresponding 3D points and refined using a K-NN search in the 3D point cloud space, which improves the handling of discretization artifacts and blurry CNN outputs. RangeNet++ achieves good performance while running in real-time on a single GPU, however, it still operates on a 2D projection of the point cloud losing some 3D structure.

Radar-oriented semantic segmentation has also been explored in the literature. Ouaknine et al. [172] proposed a multi-view segmentation approach for a RAD tensor. The RAD tensor is decomposed into three different 2D views: RA, RD, and an angle-Doppler view. These views are processed by dedicated encoders and mapped to a common latent space, and then two decoder heads produce semantic segmentation of the RA and RD views. This architecture exploits the entire radar data while addressing challenges of volume and noise, at the cost of increased complexity due to multi-view processing. Prophet et al. [83] introduced a semantic segmentation approach that generates a semantic grid map. Accumulated radar point clouds are transformed into 2D or 3D occupancy grids and then passed into a CNN. Following this, an encoder-decoder is used to generate an output of semantic grids.

1) COOPERATIVE SEMANTIC SEGMENTATION

Cooperative semantic segmentation has been shown to offer several benefits including increased spatial coverage, complementary information, and improved reliability and robustness. Studies like [173], [174] have introduced different approaches that perform cooperative semantic segmentation.

Xu et al. [173] introduced CoBEVT, a cooperative BEV segmentation framework that fuses information from multiple agents and camera views. Firstly, the SinBEVT module takes multi-view camera images from a single agent as input. A shared backbone is used to encode the images into features, and the encoded features are integrated into BEV space using a cross-attention mechanism called fused axial

attention (FAX) which captures sparse local and global interactions. The BEV embedding is progressively downsampled and refined with image features at each scale, and finally, BEV features for a single agent are generated. Following this, the BEV features from each agent are compressed and broadcasted to other agents along with the agent's pose. Upon receiving features, the receiving agents use the pose to warp the features to their own coordinate system via a spatial transformation. The FuseBEVT module then stacks the received and transformed BEV features from multiple agents and uses a transformer encoder with FAX self-attention blocks to attentively fuse information across agents. A decoder then applies convolutional and upsampling layers to the fused BEV features from all the agents to generate the final segmentation output. CoBEVT achieves state-of-the-art performance on multi-agent BEV segmentation by effectively fusing multi-view, multi-agent information. Additionally, the compressed BEV feature-sharing approach is communication-efficient and scalable. The strength of CoBEVT lies in its ability to effectively fuse information from multiple agents using attention mechanisms in the BEV space, leading to high accuracy. Furthermore, its compressed feature sharing makes it more practical for real-world deployment compared to methods of sharing raw sensor data. However, the reliance on relative pose information between agents for feature transformation can lead to errors if localization is inaccurate, and the multi-scale transformer architecture can be computationally complex and memory-intensive, especially with many agents and high-resolution feature maps.

Liu et al. [174] introduced the vehicle-infrastructure cooperative semantic segmentation (VICSS) framework to enhance vehicle perception by fusing features from an infrastructure-side LiDAR and the vehicle's onboard LiDAR. The infrastructure feature extraction (IFE) block takes the infrastructure point cloud, transforms it into the vehicle's coordinate system, and then extracts features from the overlapping FOV. Then a local feature extraction is applied to the overlapping and non-overlapping points of the vehicle, and a K-NN algorithm is applied to find neighboring points and encode their relative positions to capture the local structure. The feature aggregation stage fuses the overlapping FOV infrastructure features with the overlapping FOV vehicle features using cross-attention. Vehicle features are used to compute queries and the infrastructure features provide keys and values, allowing the vehicle to selectively fuse relevant information from infrastructure. Finally, the overlapping FOV, non-overlapping FOV, and global features are concatenated and passed through an MLP to obtain semantic labels for each 3D point in the vehicle point cloud. Selectively sharing compressed features instead of raw point clouds reduces the communication requirements, however, communication delays can lead to temporal misalignment between the sensor data.

Cooperative segmentation approaches have been shown to outperform single-agent segmentation, and still provide segmentation results in the case of full sensor suite failure on the ego vehicle. In [173], CoBEVT provides better performance on static and dynamic classes. Furthermore, this cooperative segmentation approach is more robust against occlusion, and even if all cameras on the vehicle are in failure, CoBEVT still provides some results for the segmentation task. Similarly, the VICSS framework [174], outperforms all non-cooperative baselines on the vehicle, drivable area, and lane segmentation tasks on the OPV2V dataset.

IV. ENVIRONMENTAL MAPPING

Maps are dense digital representations of the environment, incorporating essential features such as roads, lanes, traffic signs, signals, obstacles, pedestrians, and other relevant elements, facilitating a thorough comprehension of the surroundings. These maps vary in information density and are tailored to specific requirements. For V2I cooperative sensing, the maps can contain information varying from coarse object detection to dense object tracking and semantic information, with the information density adjusted according to distance from the FSN.

In [26], Eskandarian et al. classify cooperative perception as a map merging problem, where the goal is to unify the perception information gathered by individual vehicles/agents and map it onto a global coordinate frame. To achieve this, the first step is to estimate the relative poses between the vehicles. Once the relative poses are determined, the perception information from each vehicle can be merged. As a result, it is important to discuss single-agent mapping paradigms within the cooperative perception context. Furthermore, the ego-vehicle uses these methods to enhance its environmental understanding and can integrate shared information from other agents, like object data from an FSN, into its existing maps (e.g., a semantic map). This decoupled approach avoids dependencies on specific collaborative methods and their sensor/model constraints, promoting flexibility in heterogeneous systems. The discussion and understanding of these mapping techniques will act as an enabling technology.

Consequently, this section explores two main mapping paradigms: BEV and perspective view (PV). Within the BEV framework, we discuss the creation of object and semantic maps, while for PV, we examine the use of SLAM and HD maps. Object maps and semantic maps in the BEV space are illustrated in Fig. 12. Once these mapping approaches are discussed, an analysis of different map fusion techniques will be provided.

A. BIRDS-EYE-VIEW REPRESENTATIONS

Front-view or PV is a representation of the scene captured from the 2D viewpoint of the sensor, whereas BEV consists of the same scene transformed into a top-down view. PV representation is widely accepted for various perception tasks. However, the BEV space has several inherent advantages that are particularly useful from a V2I cooperative sensing standpoint.

Several perception tasks such as object detection [169], [177], [178], [179], semantic segmentation [146], [169], [180], [181], tracking [182], and fusion [183], [184] can be

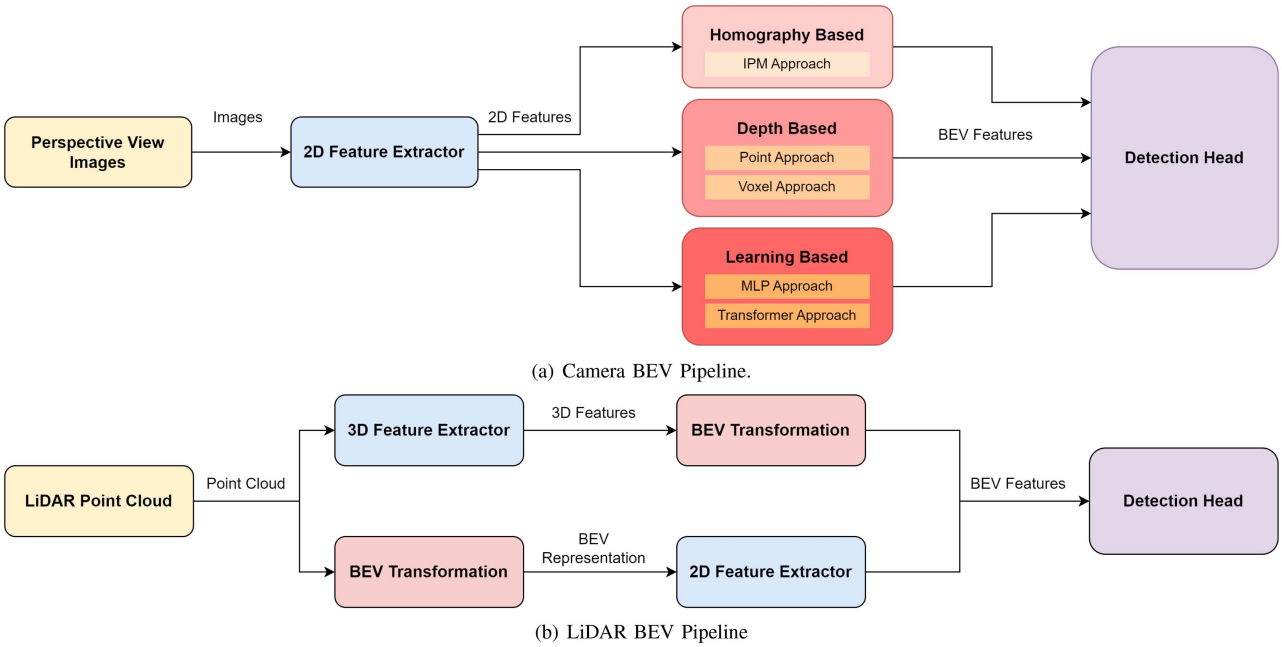


FIGURE 11. Comparison of BEV pipeline architectures: (a) Camera-based and (b) LiDAR-based pipelines, highlighting the components in each pipeline, adapted from [175], [176].

performed in the BEV space. The main advantage of the BEV space is that it provides a complete representation of the scene; this allows it to overcome the scale ambiguity [185] and occlusion problems present in PV perception tasks [145], [175]. BEV space also represents objects in a more intuitive way, a feature that is beneficial for planning and control modules. Finally, the unified representation of the BEV space benefits downstream perception tasks such as motion forecasting and tracking.

Li et al. [175] outlined the key components of a camera and LiDAR BEV perception pipeline as shown in Fig. 11. The camera-only BEV pipeline, Fig. 11(a) consists of 3 stages: feature extraction, view transformation, and 3D decoder. The 2D feature extraction stage uses backbone networks to process the PV input images and extract their 2D features. The following view transformation module, considered a crucial element for a camera-only BEV pipeline, converts these 2D features into BEV representations. Finally, the 3D decoder takes the transformed BEV features and generates the final 3D perception results, such as 3D bounding boxes or BEV map segmentation.

A LiDAR BEV pipeline Fig. 11(b) also consists of 3 stages: feature extraction, feature processing, and detection. The feature processing stage has 2 approaches pre-BEV and post-BEV feature processing. In a pre-BEV approach, 3D convolutions are used to extract 3D features and are transformed into a BEV representation. The post-BEV techniques directly convert the point cloud to a BEV representation using statistics or pillar-based methods, following which the BEV features are processed using 2D convolutions. While pre-BEV processing retains richer 3D information leading to potentially better accuracy, the increased computational burden might make it less suitable for real-time applications compared to post-BEV methods. The post-BEV methods offer efficiency at the potential expense of loss of spatial details due to the BEV space conversion.

The view transformation module is concerned with converting from PV to BEV. Ma et al. [176] discussed the different view transformation approaches for a camera-based BEV pipeline. As a reference to the reader, Table 5 provides the advantages and limitations of each view transformation approach.

Classical view transformation methodologies rely on geometry to transform from PV to BEV. These homography-based methods use Inverse Perspective Mapping (IPM) techniques that employ the intrinsic and extrinsic properties of the camera to warp the PV to a BEV. IPM can also be conducted as a pre-processing or post-processing methodology for downstream tasks such as detection and tracking. The simplicity and computational efficiency of homography-based methods like IPM make them attractive for resource-constrained applications, but their reliance on the flat world assumption limits their applicability in complex urban environments.

Depth-based transformation approaches, also known as 2D-3D methods, were introduced to overcome the flat world assumption by IPM methodologies. These methods convert 2D pixels or features to a 3D representation by learning and estimating the depth in the 2D plane. These approaches can be categorized into two types: point-based and voxel-based methods. Point-based methods convert depth maps into pseudo-LiDAR points and are then passed into LiDAR-based 3D detection models, whereas voxel-based methods discretize

TABLE 5. Comparison of View Transformation Methods in a Camera-Only BEV pipeline [175], [176]

View Transformation Approach	*Advantages*	*Limitations*
Homography Based	1. Simple and computationally efficient. 2. No learning required.	1. Flat world assumption. 2. Struggles with occlusions.
Depth Based	1. Overcomes flat world assumption. 2. Provides explicit 3D representations.	1. Computationally intensive. 2. Requires accurate depth estimation.
MLP Based	1. Can learn complex mappings. 2. No need for geometric priors.	1. Large datasets may be required to learn effectively. 2. Slower convergence due to lack of geometric priors.
Transformer Based	1. Can handle multi-view inputs. 2. Strong relation modeling ability	1. High computational cost. 2. Memory complexity leads to resolution trade-offs.

the 3D space to construct a regular structure for feature transformation, which is then shared with subsequent BEV-based modules.

Deep learning approaches for view transformations attempt to overcome the constraints set by the geometric transformation. MLP-based approaches use neural networks as complex mapping functions to transform PV to BEV. These methods learn the implicit representations of camera calibrations without relying on the geometry for mapping. MLP-based methods offer flexibility by learning complex mappings without explicit geometric priors, but this often requires large datasets and can lead to slower convergence compared to geometrically informed approaches. Similarly, transformer-based approaches use the cross-attention mechanisms to map from PV to BEV. This approach uses the cross-attention mechanisms to construct queries and then find corresponding image features. Transformer-based methods excel at handling multi-view inputs and modeling complex relationships, making them robust to occlusions, but their high computational and memory costs can be a limiting factor.

1) OBJECT MAPS

Object maps are defined as representations of the environment that can be generated in either PV or BEV. For an accurate representation, object maps include object information, such as location in metric space relative to the sensor, classification, size, and orientation. Fig. 12(b) illustrates an example of a BEV object map derived from a monocular camera frame.

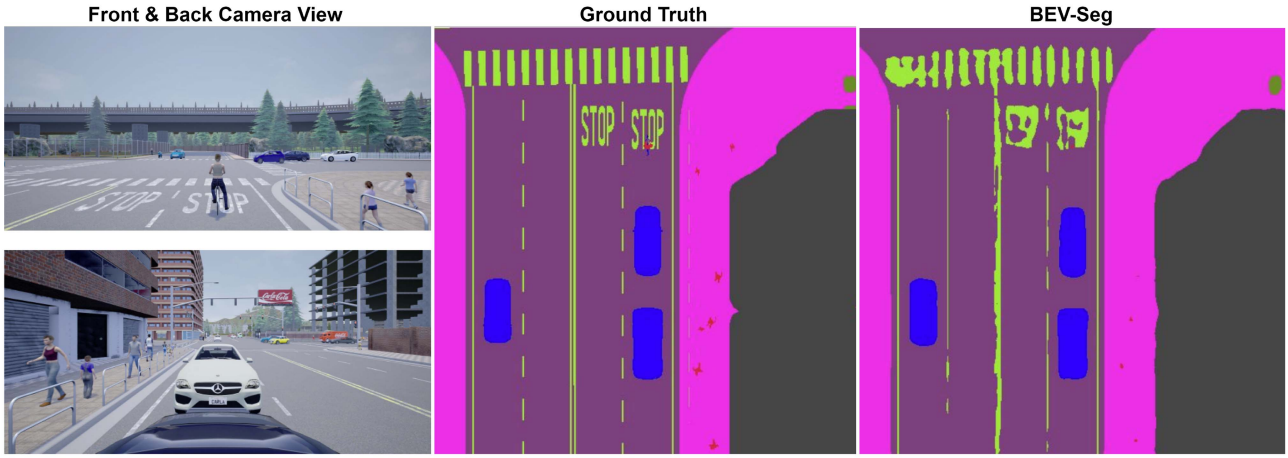
An occupancy grid map (OGM) is an example of an object map. OGMs are evenly spaced grid representations of the environment in a BEV space, and each grid contains a probability that an object occupies that space [187], [188].

The OGM framework uses sensor noise models and sensor measurements to compute the probability of occupancy in each grid. Objects from the environment can span across multiple cells, but an OGM does not estimate the class or connection of objects between cells [189]. Furthermore, they lack object-level information such as class, position, and orientation. This makes OGM unsuitable for path-planning tasks.

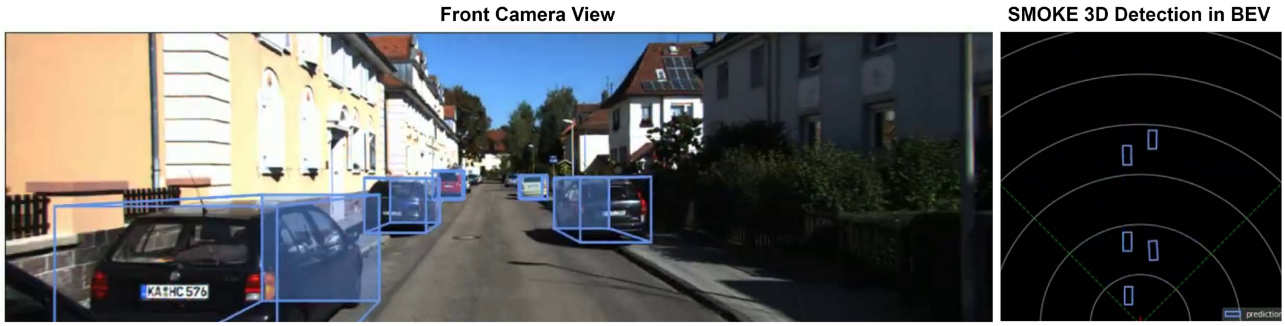
Collins et al. [187] reviewed a number of established OGM frameworks and described the limitations of each approach, while also defining a benchmark for evaluating OGMs. Defauw et al. [189] proposed different models for vehicle detection that utilize OGMs as input in both a camera and LiDAR point cloud framework. Lu et al. [190] introduced a framework that extends classical OGM representations to include semantic and metric information. By incorporating semantic understanding and making them more informative, a key limitation of traditional OGMs is addressed. OGMs can also be extended to V2V and V2I cooperative applications as outlined by Li et al. [191] and Caillot et al. [192].

Another method of generating an object map that includes class, location, and orientation information is by converting a 3D detection in PV to a BEV space. This method involves projecting a 3D bounding box onto a 2D plane. The 3D bounding box is in the format of (centerX, centerY, centerZ, length, width, height, rotation). The outermost corners of the box can be obtained by finding the maximum and minimum values along the x and y axes and generating a rectangle that encompasses the object. The BEV bounding box can then be rotated to adjust for orientation. One limitation of this method is that the centerX, centerY, and centerZ points of the 3D bounding box are subject to spatial uncertainty, meaning that there is potential inaccuracy in the estimated location of the object's center. This uncertainty arises from various factors such as sensor noise, calibration errors, and algorithmic limitations. As a consequence, representing objects solely as centroids in the PV to BEV conversion process can introduce errors that propagate through the subsequent steps, potentially affecting the accuracy of the resulting object map.

An alternative technique for generating an object map is by performing 3D object detection directly in the BEV space. Huang et al. [177] introduced BEVDet which performs multi-view 3D object detection in the BEV space. The architecture consists of an image encoder that extracts features from the images and generates a depth map. The depth map coupled with the camera model generates a point cloud that can be used to transform the features from PV to BEV. Once in



(a) BEV-Seg Semantic Map.



(b) SMOKE Object Map.

FIGURE 12. Comparison of BEV mapping techniques from input camera images: (a) BEV-Seg [180] generating a semantic map, and (b) SMOKE [186] producing an object map, both in BEV space.

the BEV space, the detection head outputs 3D detections. BEVDet4D [178] enhances the performance of detection and velocity prediction tasks by utilizing temporal information from previous frames.

The architecture in M ^{2} BEV [169] employs multi-view images to perform 3D detection and segmentation in the BEV space. Zhou et al. [193] introduced PersDet which performs object detection in a perspective BEV space without the need for feature sampling. Fig. 13 depicts the difference between camera PV, regular BEV, and perspective BEV.

Zhu et al. [194] evaluated the robustness of these vision-dependent BEV detection models by comparing the effects of natural and adversarial attacks.

The BEV space can also be extended to include single object tracking (SOT) information. BEVTrack [182] performs BEV SOT on LiDAR point clouds. Features from temporally preceding point clouds are extracted, and a motion model in the BEV space is generated. This motion model is then used to predict the object's location in the following point cloud. Similarly, Li et al. [195] proposed PowerBEV, a method that performs instance prediction in a BEV representation. PowerBEV approach efficiently combines pixel-level and instance-level association, enabling better object tracking over time. The authors demonstrate that PowerBEV performs well in various driving scenarios, such as urban traffic, parking lots, and adverse weather, as well as outperforming SOTA baselines on the NuScenes dataset for Intersection-over-Union (IoU) and Video Panoptic Quality (VPQ) metrics.

2) SEMANTIC MAPS

Semantic maps extend object maps by providing semantic labels of the environment. This allows for a better understanding and representation of the environment. Fig. 12(a) shows an example of semantic mapping. Traditional BEV segmentation methods first performed segmentation in PV and then applied IPM to transform the results into BEV space [168].

As previously mentioned, Philion et al. [146] introduced Lift, Splat, Shoot, a method for inferring vehicle drivable area, and lane segmentation in the BEV space using multi-view images. In the "Lift" stage, the model estimates a depth distribution and generates context features for each pixel in the input images, constructing a 3D frustum of features for each pixel. Then, during the "Splat" phase, the 3D features from all cameras are projected onto a shared BEV grid using the camera calibration parameters, allowing for a unified representation of the scene from a top-down perspective. A

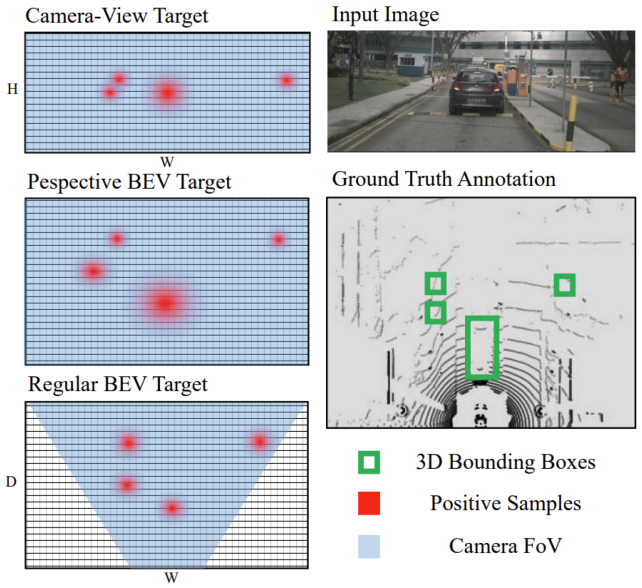


FIGURE 13. Comparison of camera PV, regular BEV, and perspective BEV [193].

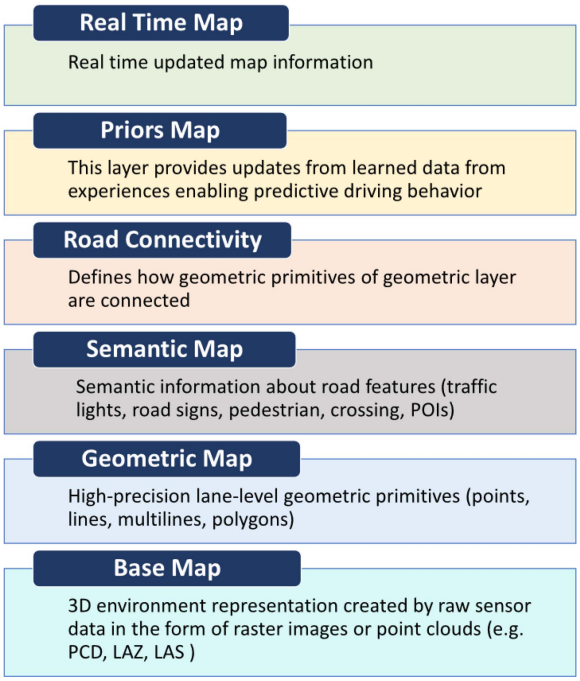


FIGURE 14. Six layers for a HD Map [202].

BEV CNN is responsible for the segmentation task in the BEV space. Lastly, in the "Shoot" step, which is specifically relevant for motion planning tasks, the model evaluates and scores a set of predefined trajectory templates by integrating costs along each trajectory path within the learned BEV representation, enabling the selection of the most suitable trajectory for the given scenario.

BEV-MODNet [196] performs object detection on moving objects in the BEV space using a single RGB frame as input. The dual-stage architecture fuses both RGB and optical flow streams for motion segmentation in the BEV space. BEV-Seg [180] utilizes a two-stage pipeline for semantic segmentation in the BEV space. PV images are passed into a depth estimation and semantic segmentation network. The combined results of these networks are then used to make a semantic point cloud representation. This point cloud is then transformed into an incomplete BEV representation that is passed into a parser network that provides the final BEV segmentation representation.

BEVSegFormer [168] uses a transformer module to generate a BEV semantic segmentation map from input images, using a shared backbone to produce a multi-scale feature map. These features are passed into a transformer encoder that enhances the feature representations that are subsequently passed into a BEV transformer decoder yielding a BEV segmentation result.

Yang et al. [197] introduced a novel framework that estimates the road scene layout and vehicle occupancies in a BEV representation. FIERY [181] predicts future instance models by predicting future instance segmentation and motion in a BEV representation. Xu et al. [173] introduced CoBEVT utilizing sparse transformers for cooperative camera-based BEV segmentation.

B. PERSPECTIVE VIEW MAPPING

While the aforementioned methods discussed how a map can be generated in the BEV space, there are two common mapping representations from an ego vehicle perspective which are: HD maps and SLAM.

1) HD-MAPS

HD maps are a digital representation of the environment that contains semantic [198], [199], and geometric information [198] about both static and dynamic objects [200]. These models of the environment are pre-built enabling the ego vehicles to precisely localize themselves in the environment relying on the features and information in the pre-defined map. HD maps are discussed in this section because they enable accurate self-localization for multiple agents [155], facilitating both accurate inter-agent pose determination, which is crucial for effective collaboration.

Typically, HD maps can be decomposed into 3 models [201]: the road model, the lane model, and the localization model. The road model contains topological information about the road, such as the direction of travel, slopes, curbs, and boundaries. The lane model is a perceptual layer that encodes road lines, road width, and speed limits. The localization layer contains static features in the environment such as lane markings, buildings, and traffic signs; these static features help the vehicle localize itself in the environment.

Elghazaly et al. [202] further decompose these models into six layers for HD maps, as seen in Fig. 14. Elghazaly et al. [202] uses a highly accurate 3D base representation of the scene. The geometric layer represents the road model,

encoding the location, shape, and features. The semantic layer is a fusion of the lane model and the localization layer. Information such as traffic signs, speed limits, and lane markings are encoded into this layer. The road connectivity layer defines how the various lane boundaries and geometric representations of the road are connected. The priors map utilizes pre-learned information about the environment to predict the future states of road users. The real-time layer provides non-recurrent environmental information such as collisions and road closures for better motion planning. The conditions and frequency for updating an HD map are described by Boubakri et al. [201].

Once data collection is completed, the first step in creating an HD map is to generate a point cloud base map of the scene. Once the base map is constructed, scene features such as lane markings, and road boundaries are appended to it. Bao et al. [203] describe many point cloud map generation and feature extraction methodologies. Li et al. [199] introduced an online HD Map semantic layer framework that utilizes images and LiDAR point clouds to generate a vectorized HD Map. A potential use case where FSNs and HD maps co-exist is by utilizing the FSN to update the dynamic layer of the HD map.

The main downside of HD-Maps is the time and monetary cost associated with data collection. Elghazaly et al. [202] also outlined the various challenges associated with HD maps, particularly the lack of standard representations, that can pose a challenge if HD maps are used to enable L5 autonomy.

2) SIMULTANEOUS LOCALIZATION & MAPPING

Simultaneous Localization and Mapping (SLAM) involves computing the location of the autonomous vehicle while concurrently mapping its environment [204]. SLAM can be performed on a wide range of sensor sets such as LiDAR, monocular camera, stereo camera, and RGB-D cameras [205]. Kazerouni et al. [205] reported on a generic SLAM framework that consists of feature extraction and matching, pose estimation, loop closure, and map generation. The pose estimation stage identifies the vehicle's relative pose given the information from the feature extraction and matching stage. The loop closure stage is used to identify whether the agent has returned to a previously visited location. Finally, the map-building stage generates a representation of the environment. SLAM algorithms are another important consideration in cooperative multi-agent systems because they enable precise localization within a local coordinate frame. If the FSN is within the SLAM loop, then a more accurate pose estimation is possible, ultimately improving collaboration.

Classical approaches for SLAM consisted of probabilistic approaches such as Kalman Filters, Particle Filters, and Expectation Maximization methods. These probabilistic methods were popularised due to the algorithm's ability to model the effects of uncertainty and noise on the SLAM outcome [204].

Placed et al. [206] and Mokssit et al. [207] reviewed deep learning-based SLAM techniques. Mokssit et al. [207] identified the various modules for SLAM such as depth estimation, visual odometry, and optical flow, while comparing classical approaches with deep learning methods that address traditional challenges of these modules. This study also investigated joint learning approaches for SLAM, which leverage relationships between SLAM modules for optimization and more accurate models, and also examined uncertainty estimation and reduction techniques to enhance SLAM reliability.

C. MAP FUSION

The previous discussions have focused on the technical details of generating local maps from individual agents. These individual approaches can be further enhanced by fusing the maps to create a global representation. Map fusion refers to taking maps from different perspectives and aligning them temporally and spatially in a global frame of reference. For V2I cooperative sensing, map fusion consists of temporal and spatial alignment, which merges the object information from different local maps.

1) COORDINATE SYSTEMS

Before discussing spatial alignment, it is important to introduce the various coordinate systems and their transformations. Typically, the vehicle can have its own coordinate system, as defined in [208]. The vehicle is the origin and every detected object is represented relative to the vehicle. A similar approach can be taken for the FSN coordinate system. The data shared by the FSN must be in the vehicle coordinate system so that the vehicle can append it to its environmental map. To achieve this, an accurate relative pose must be established, allowing the sender to transform the data into the receiver's coordinate system. If the relative pose has high errors, the agent will receive unreliable data.

When the vehicle and the FSN are viewing overlapping regions, then an easily distinguishable landmark can be used to align both the vehicle and the FSN to a local coordinate system. This landmark can be the origin, and every detected object can be relative to it. The FSN can always append to this landmark since it will always be in the FOV of the FSN, removing the need to continuously determine its location. However, one drawback to this approach is that the vehicle must switch to this coordinate system when it is in range of the landmark to receive shared information, and then switch out of the coordinate system when it is out of range. This can potentially result in higher processing overheads and delays.

Alternatively, a global reference system can be used, where the FSN and the ego vehicle always work with a global coordinate system. This approach heavily relies on accurate GNSS positioning. WGS84 is a geodetic coordinate system that is used for the global positioning system (GPS). It consists of longitude (λ), latitude (ϕ), and altitude (h). Earth-Centered Earth-Fixed (ECEF) is a Cartesian coordinate system that uses the center of the earth as the fixed origin. Similarly, East-North-Up (ENU) is also another cartesian coordinate system that uses the geodetic coordinate of the reference as a

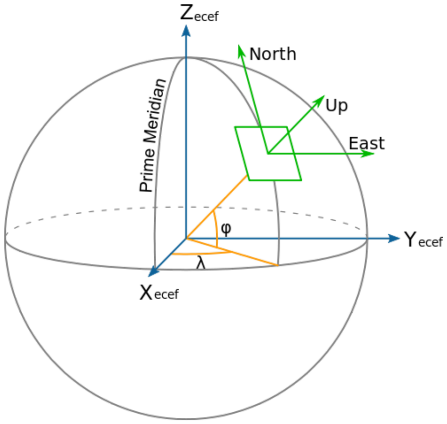


FIGURE 15. Global coordinate systems and their relations.

fixed origin. Fig. 15 illustrates an example of how the above-mentioned coordinate systems are related to each other. Allig et al. [209] discuss the different coordinate systems in detail while also discussing the relative transformation between them. One significant disadvantage of using this global coordinate system to represent object information is that factors such as temporal jitter or inaccuracies in GNSS can result in poor mapping outcomes. However, if these errors aren't significant, then the global coordinate systems can be used for spatial alignment, which we will discuss in the next section.

2) COOPERATIVE SPATIAL ALIGNMENT

Spatial alignment ensures that the data shared from cooperative agents are aligned in the same coordinate system. If the spatial alignment is poor, it can lead to problems like duplicated objects, where a single object might appear as two separate objects in the fused data. The discussion in this section will introduce different approaches for spatial alignment. These methods can be broadly categorized as either relative-pose-based or learning-based approaches.

Pose-based approaches use the shared pose information or calculate the relative pose based on shared data, to compute the transformation for alignment. Thrun et al. [210] introduced a method for spatial alignment where the initial relative pose is known. As the agent moves through the scene, an incremental maximum likelihood estimator is used to determine the current pose that aims to maximize the consistency of current sensor measurements. This produces a map by predicting the most likely continuation of the previous maps, considering the latest sensor measurements. To handle revisits of mapped areas, a full probabilistic posterior is maintained. The posterior identifies the accumulated pose error which is used to correct the map. This approach enables efficient real-time mapping of closed or looped environments.

Similarly, in a study conducted by Seeliger et al. [211], the authors introduced a cooperative perception and warning system to provide early advisory warnings, especially in critical situations with occluded VRUs. In this system, the perception data from other vehicles, FSNs, and the ego vehicle is fused via an inter-vehicle information fusion (IV-IF) module. The IV-IF module transforms the received object states to the ego vehicle's coordinate system using the sender's relative pose and uses an unscented transform for uncertainty propagation. Objects are associated using Mahalanobis-Distance on the x and y components of object position and covariance intersection is used to fuse the associated objects and obtain a global object map.

In [212], the spatial alignment process transforms object tracks from the sender's coordinate frame to the ego vehicle's frame for fusion. Firstly, objects from the sender's coordinate frame are transformed into an intermediate ECEF coordinate system. The sender's orientation in ECEF is obtained from the received data. Following this, using the ego vehicle's pose in ECEF, the objects are transformed from the ECEF coordinate system to the ego vehicle coordinate system. One key aspect is the intermediate frame used, should be used consistently for all transformations to avoid any further errors.

In [213] the authors introduced a method for inter-vehicle object association for cooperative perception systems, where the relative pose is computed using the shared data. Detected objects from each agent are represented as a point cloud, where different point matching algorithms like Iterative Closest Point (ICP), Expectation-Maximization ICP (EM-ICP), and Gaussian Mixture Model Registration (GMM-Registration) are used to estimate position and orientation offsets between object lists from different vehicles. Singular value decomposition (SVD) is used to compute the optimal rotation and translation to align point clouds from different agents. Limitations of this approach are that it assumes initial vehicle pose to enable coarse point cloud alignment, and large initial pose errors can degrade performance.

Similarly, Ballesta et al. [214] focuses on aligning multiple landmark-based maps offline to find the relative rigid transformations between them. Each map is in its own local frame based on the robot's initial pose. The alignment problem is formulated as estimating the 2D translation (tx, ty) and rotation (θ) between two maps. This approach uses 3D features in the scene extracted based on their Euclidean distances, and a list of corresponding feature pairs (m, m') is constructed. The translation and rotation can then be computed using RANSAC, SVD, ICP, and ImpICP algorithms.

In BB-Align [215], a two-stage approach is used to estimate the relative pose transformation between agents, that enables accurate spatial alignment of perception data for fusion. Firstly, the coarse alignment stage performs a BEV transformation from the agent's LiDAR point cloud. Then using Log-Gabor filters keypoint matching in the sparse BEV image is enhanced, and a coarse transformation is estimated using RANSAC on the matched keypoint positions. The fine alignment stage aligns the corners of the corresponding 3D bounding boxes in the BEV space and estimates the transformation to correct for misalignments. Since this isn't a

learning-based approach, the BB-Align framework can integrate with existing V2V systems, however, this approach still depends on sufficient landmarks and objects for matching.

Jiang et al. [216] introduced the map container framework using HD maps as a spatial reference for alignment. Firstly, the ego vehicle localizes itself in the map frame by aligning visual features to map elements. In this step, the ego vehicle pose is estimated by minimizing the reprojection error of matching the monocular segmentation output to the HD map features. This pose is then further refined using the Levenberg-Marquardt algorithm [228]. Using the ego vehicle pose as a reference, received poses and detected objects from connected vehicles are projected into the map frame of reference. Finally, all projected poses and objects in the map coordinate system are jointly optimized, to minimize the weighted residual error between the predicted states and the map-aligned observations.

Qu et al. [217] introduced V2I-Calib to perform calibration for a LiDAR-based V2I system. Firstly, an affinity matrix based on 3D IoU is used to capture the similarity between the infrastructure and the vehicle. Then a Hungarian matching algorithm [161] is used to find an optimal correspondence between the boxes. A second affinity matrix is constructed with additional cues like category, size, and angle to discard incorrect matches. Extrinsic calibration parameters are computed by minimizing the distance between the matched boxes. This approach doesn't require an initial extrinsic estimate, however, the bounding box matching approach will only work if the same object is detected by both agents.

Zhao et al. [218] introduced a framework that uses geolocation information from calibrated LiDARs or radars to compute the relative pose and align multiple uncalibrated cameras. Firstly, the cameras use learning-based camera calibration approaches, discussed in Section II, to estimate vehicle locations in the scene. The LiDAR or radar also estimates the vehicle location, acting as a ground truth. The relative camera parameters are optimized using the ground truth data to tolerate global parameter errors. Then the spatial alignment module transforms the multi-view cameras and aligns them in a global coordinate system, using the relative poses estimated.

In [219], shared pose information is used to align the feature maps to the ego vehicle's coordinate system. A pose regression module is used to predict relative pose correction between pairs of vehicles, and a consistency module formulated by a Markov network is used to refine the global poses. An attention mechanism is then used to filter out any outliers or pose noise based on learned weights, contributing to improved robustness. The drawback to the attention mechanism is that it requires supervised labels for noisy and clean examples to learn from.

Learning-based methods use graph neural networks (GNNs) or transformer networks to compute the relative pose and fuse the data for perception tasks. In V2V-Net [220] each agent broadcasts an intermediate feature map based on their own coordinate system. A GNN is used by the receiver to spatially transform and fuse the received feature maps based on the relative vehicle poses. The transformed features are fused and passed through CNN layers to 3D object detections and future trajectories. This approach aggregates information from multiple agents to handle occlusions and also compresses intermediate representations to meet bandwidth requirements. However, this approach requires relative vehicle pose estimates for spatial transformation, and an error in the initial pose can lead to errors in the spatial alignment.

Similarly, Lei et al. [221] introduced FreeAlign for spatial-temporal alignment. FreeAlign constructs a salient-object graph from each agent representing spatial relations between the detected objects. Within this graph, the nodes are detected objects and the edges encode relative distances and orientations. A GNN is used to learn compact edge features, which capture the local geometric structure invariant to global coordinate frames. Following this, graph matching is used to find the maximum common subgraph between two agents' salient-object graphs, signifying distinct and similar geometric structures. Correspondence between the nodes detected by different agents is established, and the relative transformation is estimated using RANSAC algorithms. The main advantage is that the salient-object graphs provide a concise representation for spatial alignment compared to raw point clouds, however, the salient-object graphs may be sensitive to detection errors or missed detections.

Additionally, HM-ViT [222] introduced a spatial-aware heterogeneous 3D graph transformer to fuse the information from camera and LiDAR agents in different coordinate frames. In this framework, each agent is represented as a node in the graph, and a spatial transformation matrix converts the intermediate features of each node to the ego vehicle's coordinate system. The transformation is composed of the relative poses that are shared by the agents, along with the compressed features. Non-overlapping areas are masked when computing attention scores within each transformer block. Finally, the refined feature map is passed into a detection head to predict 3D bounding boxes.

The authors of [223] proposed the SCOPE framework for multi-agent collaborative perception that captures spatio-temporal semantics. Compressed feature maps are shared with the ego vehicle and a deformable cross-attention module is used to aggregate multi-scale spatial features from other agents while being robust to localization errors. The primary advantage of this approach is that the cross-attention module enables localization error-tolerant spatial feature aggregation, however, the cross-attention module also introduces additional compute and memory requirements.

The MASH network [224] performs patch-based spatial alignment of visual features across agents. With this approach, the agents extract spatial features on their input data using a segmentation backbone, which is then compressed into a query and keymaps via learned encoders. The query map is broadcasted, and the receiving agents compare it to their key map to produce a dense correspondence volume. A context-aware autoencoder is used to smooth the correspondence volume, which is then used to transform the ego vehicle's segmentation output.

Zhou et al. [229] introduced a new framework that uses a single-step communication process and employs an intermediate fusion technique. The system is built on a unified transformer-based architecture, which effectively models the complex spatiotemporal relationships across various domains, including temporal per-frame data, spatial per-agent information, and high-definition maps. To support their research, the authors also created the V2XPnP Sequential Dataset, which addresses the limitations of existing real-world datasets by supporting all V2X cooperation modes and providing data beyond single-frame or single-mode cooperation.

Finally, Shi et al. [225] introduced MCoT, a transformer-based model for multi-modal V2V cooperative perception. For each agent, camera and LiDAR feature maps are converted to the BEV space and aligned using a rigid transformation. The authors implemented an average sampling mechanism that overcome the discrepancies when aligning LiDAR and image features. A cross-attention mechanism is used to fuse the aligned multi-modal feature maps, used for 3D detections. Since this approach leverages both camera and LiDAR data, detection performance is improved, however, since the pose estimation is required for BEV feature alignment, the model performance may degrade with pose errors.

The above-mentioned methods utilize a wide range of techniques for spatial alignment. Most of these methods are used as a precursor for cooperative sensing tasks. Further research in this area can focus on developing more efficient and robust spatial alignment methods that can handle challenging scenarios like large pose errors, occlusions, and dynamic environments. Table 6 presents a comparison of various spatial alignment methods along with their respective advantages and disadvantages.

3) COOPERATIVE TEMPORAL ALIGNMENT

Temporal alignment is crucial for accurate spatial alignment and map fusion in cooperative systems. Factors such as communication latency, processing delays, and object motion introduce discrepancies between the time when information is captured and when the vehicle receives it. Thus, the received information may be outdated, leading to poor accuracy in spatial alignment and map fusion.

In V2V-Net [220] the time delay is measured from the received message and current time. A small CNN is applied that uses the estimated time delay as input into the received feature map. This CNN learns to compensate for time delay that fuses time-compensated features from past messages to refine the current feature map, thus producing a temporally-aligned feature map. In another study [221], each cooperative agent maintains a temporal sequence of its salient-object graphs, and graph matching between the received graph and the historical graph sequence is performed, from which the timestamp that yields the maximum common subgraph is selected. This subgraph is then used for pose estimation for spatial alignment.

How2comm [226] introduced a flow-guided delay compensation strategy to temporally align the shared features. A flow generator uses the past sequence of feature maps and predicts a feature flow map and scale matrix for a fixed time interval. The ego vehicle will use the received flow to warp the received feature map to the current timestamp via bilinear interpolation and scaling, before feeding it into the spatio-temporal fusion module. A temporal cross-attention module is used in the spatiotemporal fusion module, fusing historical context across all agents to reinforce the current ego feature map.

Another approach is introduced in [212], where the authors first temporally align the data before spatial alignment. Firstly, using a constant turn rate and acceleration (CTRA) motion model, the sender's pose and object states are predicted to the current fusion time. Additionally, the sender's motion between the received message timestamp and the fusion timestamp is calculated and accounted for in the predicted object states. Similarly, in [227], the authors use a CTRA motion model to predict the sender's motion. Based on this motion model, an unscented Kalman filter is used to predict the received object data at the current timestep. The above-mentioned approaches predict the object state, account for motion, and then transform to the receiver frame. In [209], the authors proposed predicting the object state and then transforming it to the receiver frame using the last known sender pose, thus skipping the motion compensation. While this approach is computationally simpler, this method can introduce significant errors if the sender has moved significantly since the last transmission.

Table 7 presents a comparison of some of the temporal alignment methods along with their respective advantages and disadvantages.

V. PRACTICAL IMPLEMENTATION

A. URBAN JUNCTION SCENARIO

One common yet challenging scenario for autonomous vehicles is navigating urban junctions. These intersections often pose complex situations consisting of low-speed vehicles, dense pedestrian traffic, and unpredictable pedestrian behavior, making it a complex situation for autonomous agents. Pedestrian occlusions are a common occurrence due to the presence of multiple parked and slow-moving vehicles, signposts, and other obstacles that can obscure the vehicle's line of sight. VRUs may suddenly emerge from behind these obstructions, leaving little time for the vehicle to react and avoid collision. Additionally, darting pedestrians are another frequent challenge in these scenarios, where pedestrians may unexpectedly change their direction or speed quickly moving into the path of the ego-vehicle, due to distraction or obstacles on the footpath. This unpredictable behavior requires the ego-vehicle to have quick reaction capabilities, and the ability to anticipate and respond to sudden changes in the pedestrian movement.

The use of FSNs, viewing the scene from an elevated perspective, can significantly mitigate some of the risks associated with these urban junction scenarios. These sensor nodes can provide a complementary viewpoint that enhances the situational awareness of autonomous vehicles. The elevated viewpoint, with their higher vantage point, can detect

TABLE 6. Analysis of Spatial Alignment Methods

Ref.	Category	Technique	Advantages	Disadvantages
Thrun et al. [210]	Pose-Based	Incremental maximum likelihood estimator, probabilistic posterior	Efficient real-time mapping of closed/looped environments	Requires initial relative pose
Seeliger et al. [211]	Pose-Based	Relative pose transformation, unscented transform, Mahalanobis distance	Fuses data from multiple sources	Relies on accurate relative pose
Seeliger et al. [212]	Pose-Based	ECEF intermediate frame transformation	Consistent intermediate frame	Relies on accurate ECEF pose
Rauch et al. [213]	Pose-Based	Point cloud matching (ICP, EM-ICP, GMM), SVD	Computes relative pose from data	Assumes initial pose, sensitive to errors
Ballesta et al. [214]	Pose-Based	3D feature extraction, RANSAC, SVD, ICP	Aligns landmark-based maps offline	Offline, requires landmarks
Song et al. [215]	Pose-Based	Two-stage BEV, keypoint matching, RANSAC	Integrates with V2V systems	Depends on landmarks and objects
Jiang et al. [216]	Pose-Based	HD map alignment, Levenberg-Marquardt	Uses HD maps	Requires HD maps, accurate matching
Qu et al. [217]	Pose-Based	3D IoU, Hungarian matching	No initial extrinsic estimate	Matching needs same object detection
Zhao et al. [218]	Pose-Based	LiDAR/radar geolocation, camera calibration	Aligns uncalibrated cameras	Relies on accurate calibration
Vadivelu et al. [219]	Pose-Based	Pose regression, attention mechanism	Refines poses, filters outliers	Supervised labels for attention
Wang et al. [220]	Learning-Based	GNN, relative pose, CNN	Handles occlusions, compresses data	Requires relative pose, error sensitive
Lei et al. [221]	Learning-Based	Salient-object graph, GNN, RANSAC	Concise representation	Sensitive to detection errors
Xiang et al. [222]	Learning-Based	3D graph transformer, relative pose	Fuses camera and LiDAR data	Requires relative pose
Yang et al. [223]	Learning-Based	Deformable cross-attention	Robust to localization errors	High compute and memory
Glaser et al. [224]	Learning-Based	Patch-based alignment, autoencoder	Handles bandwidth constraints	High compute.
Shi et al. [225]	Learning-Based	BEV, cross-attention	Leverages camera and LiDAR, improves detection	Pose needed for BEV, error sensitive

these occluded pedestrians and share this critical information with the ego vehicle. Additionally, the FSNs allow for the accumulation of historical data on pedestrian behaviors at specific junctions. This accumulated historical data can reveal the probability distribution of pedestrians crossing at certain locations influenced by the presence of obstacles or the desire to take shortcuts. By accounting for these historical patterns, the vehicle can anticipate and proactively respond to unexpected pedestrian movements, even if they are not immediately visible from the vehicle's perspective.

We will explore two sub-scenarios that require different levels of involvement from the FSN. The first sub-scenario is a safety-critical application, where the FSN detects an emerging occluded pedestrian and alerts the ego-vehicle, as seen in Fig. 16. The primary goal in this case is to increase the time-to-collision, allowing the vehicle to react appropriately or change its trajectory to ideally avoid a potential collision.

The second sub-scenario focuses on darting pedestrians, which is certainly a safety-critical scenario when it happens unexpectedly near the ego vehicle. But, in this discussion, we will examine how sharing a heatmap of historical pedestrian

TABLE 7. Analysis of Temporal Alignment Methods

Reference	Technique	Advantages	Disadvantages
Wang et al. [220]	Small CNN using time delay as input to compensate for latency	Fuses time-compensated features, refines current feature map	Limited to small time delays, effectiveness depends on CNN training
Lei et al. [221]	Temporal sequence of salient-object graphs, graph matching for timestamp selection	Uses graph matching for accurate timestamp selection	Sensitive to graph construction and matching accuracy
Yang et al. [226]	Flow-guided delay compensation, flow generator, bilinear interpolation, temporal cross-attention	Handles varying time delays, reinforces features with historical context	Computational overhead of flow generation and warping
Seeliger et al. [212]	CTRA motion model to predict sender pose and object states, motion compensation	Accounts for sender motion, improves spatial alignment accuracy	Accuracy depends on CTRA model assumptions
Rauch et al. [227]	CTRA motion model, unscented Kalman filter for object data prediction	Predicts object data at current timestep, handles uncertainty	Accuracy depends on CTRA model and filter parameters
Allig et al. [209]	Object state prediction, transformation using last known sender pose	Computationally simpler	Significant errors if sender moves considerably

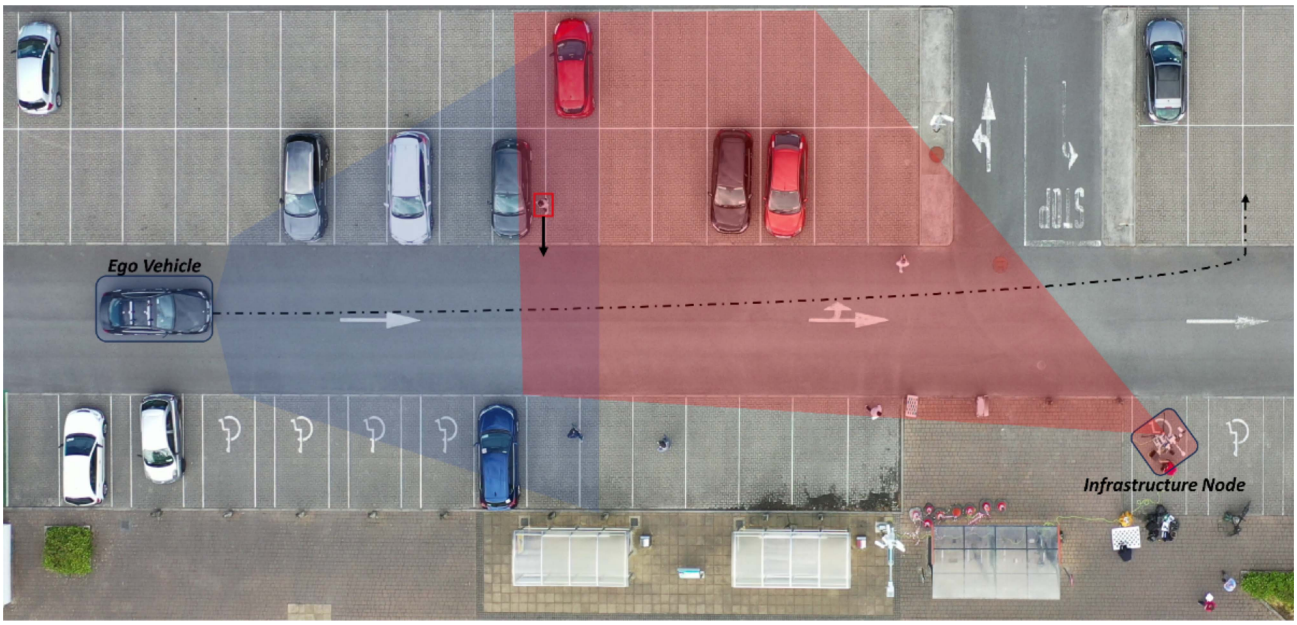


FIGURE 16. Example of sub-scenario 1, where a pedestrian, depicted in red, is occluded behind a parked vehicle and is about to walk in front of the ego-vehicle. The vehicle's trajectory is indicated by a black dotted arrow, and the pedestrian's trajectory is shown as a black solid arrow. The FSN detects the pedestrian and shares this information with the ego-vehicle.

movements can help the vehicle plan its path accordingly and account for potential darting pedestrians due to environmental factors. By providing the vehicle with information about areas where pedestrians are more likely to dart out unexpectedly, based on past behavior, the FSN can assist the vehicle in making more informed decisions and adjusting its speed or route preemptively. Therefore, we will consider sub-scenario two to be a non-critical scenario.

To simplify the scenario, we make the following assumptions. Firstly, data transmission between the car and the FSN is instantaneous and reliable. Secondly, The car is the arbiter of all final decisions, and the role of the FSN is to send information to the car to aid in its decision-making process. Finally, the vehicle operates within its own coordinate system.

When considering each sub-scenario, it's essential to identify the most crucial information for the vehicle to receive. As discussed in Section II-A, a wide range of information can be shared with the vehicle to extend its sensor cocoon, however, for sub-scenario one, the primary goal is to ensure that the ego-vehicle is aware of an object that is about to appear in its

path. To achieve this, the minimum viable information that needs to be shared is the object information in the vehicle coordinate system. This can be done in two ways, either the global location of the FSN is shared along with the object information, allowing the ego-vehicle to find the relative pose using its own location and align the object to its own map. Or, the FSN knows the vehicle's global location, finds the relative pose using its own location, and shares the transformed object information with the ego vehicle. Another critical consideration is that this safety-critical information needs to be sent at a relatively high frequency, with a minimum of 10 Hz [101].

In sub-scenario two, the aim is to provide enhanced route guidance to the ego-vehicle. To achieve this, the information shared with the vehicle can include a broader range of details about the surrounding environment compared to sub-scenario one. The most crucial piece of information to share is the heatmap of pedestrian behaviors, helping the vehicle anticipate and respond to potential pedestrian movements. In addition to this, other valuable recurring events, such as bus stops, can be included in the shared data. This enables the vehicle to anticipate areas where pedestrians are likely to gather or cross the road, even if no actual pedestrian crossing is present. By combining the heatmap data with these additional details, the vehicle can apply the same level of caution when approaching these areas as it would when approaching a designated pedestrian crossing, such as slowing down and preparing to stop if needed. It's important to note that the shared information is meant to act as supplementary data that can be passed to the vehicle well before it reaches its destination, therefore a minimum frequency of 1 Hz needs to be met [101].

1) FSN SENSOR SELECTION

To reliably provide accurate object and semantic data for sub-scenarios one and two, an FSN must have a wide FOV for comprehensive scene capture, the ability to measure absolute distances, and perform consistently in various lighting and weather conditions. As previously discussed in Section II-B, a heterogeneous system integrating wide FOV RGB cameras and LiDARs offers a good balance of capabilities and cost-effectiveness, meeting the above-mentioned requirements. RGB cameras provide excellent FOV, high spatial and temporal resolution, and affordability, while LiDARs complement them by offering robust performance in adverse conditions and absolute distance measurements, addressing the cameras' limitations. However, current manufacturing constraints result in a limited vertical FOV for LiDAR technology, ranging between 10° and 45° [77], [80]. This restricted vertical FOV, coupled with the raised perspective of an FSN, reduces the effective range of LiDARs. Fig. 17 illustrates this by comparing the effective range of a LiDAR sensor with a narrow vertical FOV to a sensor with a wider vertical FOV, both placed at the same height on an infrastructure node. The sensor with the narrower vertical FOV is unable to detect objects at further distances, conversely, the sensor with the wider FOV is unable to detect objects closer to the infrastructure node. Additionally, natural sensor deviations can cause camera geometric parameters to drift, potentially compromising the accuracy of information shared by the camera. To overcome these sensor deviations, one approach could involve continuously monitoring calibration discrepancies using known scene landmarks. The detected deviations could then be used to either directly correct the drift or be shared with ego vehicles, enabling them to compensate for sensor inaccuracies in their calculations. Achieving temporal synchronization between the ego-vehicle and FSN also presents a significant challenge. Direct hardware synchronization using a shared clock is unfeasible due to the lack of a physical connection. Therefore, a software-based approach that utilizes an external reference, such as GNSS timing, becomes necessary for system synchronization. The temporal discrepancies due to variations in communication link latency must then be accounted for in the perception pipeline.

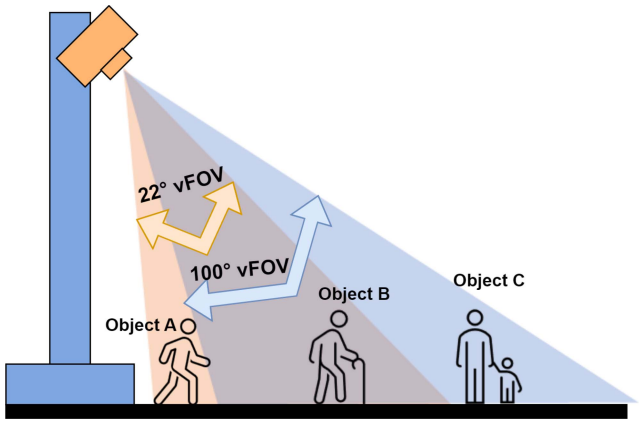


FIGURE 17. Effects of FOV at an FSN perspective.

2) FSN PERCEPTION AND CHALLENGES

When deploying models for a V2I system, several factors are needed to ensure accurate and reliable performance. For sub-scenario one, which is an occluded pedestrian scenario, getting LiDAR object information quickly and accurately is the main consideration. A projection-based model is suitable for this task. As previously discussed in Section III-B, projection-based models are computationally efficient and easy to implement. Furthermore, by projecting the LiDAR data into 2D views, the fusion task of heterogeneous data becomes easier. Methods for generating semantic scene information and object trajectories were discussed in Sections III-B and III-C. These methods can be applied to the camera or LiDAR data to get semantic scene information and pedestrian heatmaps. The model efficiency and performance is reliant on the input data from the camera and LiDAR. If the LiDAR outputs sparse point clouds, then additional computations like temporal point accumulation might be needed to increase cloud density for accurate detection. This comes at the cost of

additional processing time. Furthermore, monocular detection models often rely on either scene-specific background properties or camera parameters, leading to reduced generalization, context-dependent models, and inefficient utilization of model resources. Additionally, models tied to specific sensor characteristics and parameters can result in inaccurate detections if sensor parameters drift over time.

One of the ways to overcome reduced model generalization is through learning collaboration. Learning collaboration through federated learning (FL) is a scalable approach that enables long-term deployment. In FL, multiple agents (ego vehicles, FSNs, etc) work together to train a global model supervised by a central server. Each agent trains the model locally on their own data and shares only model updates, such as gradients or weights, with the server. The server then aggregates these updates to improve the global model [230]. By learning from diverse data while maintaining privacy, FL allows the global model to become more robust, particularly in handling long-tail scenarios [231]. FL can be categorized as centralized and decentralized [230], [232]. In centralized FL, a central server acts as an aggregator where agents train the model locally on their perception data and send updates to the server, which then averages these updates and applies them to the global model. However, the main drawback of this approach is the reliance on a central server. Alternatively, decentralized FL removes the need for a central server by having agents communicate directly with each other to share updates. Although decentralized FL offers increased robustness, it comes with challenges related to synchronization and scalability.

Posner et al. [233] introduced a concept of a Federated Vehicular Network (FVN) framework that utilizes vehicles as workers in a distributed machine learning system, where advanced communication and computing technology on vehicles is used along with existing venue infrastructure. The proposed hybrid architecture combines centralized components for resource management and model aggregation, decentralized vehicle-to-vehicle collaboration for model training, a federated vehicular cloud for efficient data and model sharing, and blockchain for incentivization, reputation management, and secure transactions. Similarly, Barbieri et al. [232] introduced a decentralized FL approach for cooperative sensing and object classification. In this approach, each agent is used to collaboratively train a PointNet [135] model, where each vehicle learns a local model and shares weights of the last few layers with the neighboring vehicles through 6G networks. Then each agent refines their model iteratively through a weighted average of neighboring models.

While for simplification purposes we assumed that data is transferred instantaneously, in realistic deployments, data transfer limitation is a key challenge in a cooperative system. The volume of data generated by automotive systems exceeds the practical transmission capabilities of real-time vehicular communication networks. For example, raw sensor data from an automotive camera can be up to 800 Mbps [84]. Similarly, for 3D radars, this can be up to 2.5 Gbps, and 10 Gbps for 4D radars [82]. High-density LiDARs used in automotive applications can produce up to 9.6 million points per second [234]. Handling these large volumes of data during transmission, processing, and storage can be particularly challenging for embedded perception systems with limited memory and bandwidth resources. In comparison to the raw data, the object data rates for an automotive camera in a busy scene can be approximately 50 Kbps [84]. As a result, for a reliable real-time V2I cooperative sensing system with hardware and cost considerations, the practical approach is to transmit processed data such as object information or feature maps.

Edge computing [84] refers to computation on the FSN, where only relevant information such as object information or feature maps is passed to the vehicle, thus lowering the data rate requirement. However, the processing and transmission delay incurred means that temporal synchronization is necessary to ensure the data received by the vehicle is relevant and up-to-date, further emphasizing the need for an efficient model to be deployed. Alternatively, FSNs can have low-latency and high-bandwidth links capable of sending raw information to a cloud server for processing and dissemination to connected vehicles. This cloud computing approach may drive down the cost of both vehicles and FSNs but has a higher processing and communication delay associated with it, making it more suitable for sub-scenario 2, where the information shared is not time-critical.

Given we have the processed information from the perception tasks, the FSN must represent this in a map so it can be shared with the ego vehicle. The ego vehicle is using different mapping approaches to generate a local map of its surrounding environment, and actively localizing itself in it. This map contains information about obstacles relative to the vehicle, from which the vehicle path and trajectory are determined. The goal of the FSN sharing its local map containing information about the occluded pedestrian is to append this information to the ego vehicle's local map so that it can adapt its trajectory and avoid collisions. There are several approaches to sharing and appending this information. The first approach for sharing object information from the infrastructure to the vehicle involves transforming the detected object's local coordinates to a global system (e.g., GPS) before transmitting the data to the vehicle. The ego-vehicle will then transform this object information to its own local coordinate system, thus appending this information to its own local map. Alternatively, the FSN can share the detected object's local coordinates along with its own global location, allowing the vehicle to estimate a relative pose and transform the object accordingly. In the case that the shared data doesn't contain sufficient information to resolve the object to a unified coordinate system, then the FSN and vehicle must align themselves to a shared coordinate system based on common landmarks. With this approach, the FSN can directly share the object's coordinates without requiring further transformation by the vehicle.

Multiple challenges need to be overcome before sharing this information with the vehicle. Discrepancies in shared

GPS locations, relative pose calculations, or coordinate system alignment can lead to object duplication. Additionally, time delays, localization errors, lossy communication, uncertainties introduced by sensor noise, and calibration errors, can result in degraded performance. Therefore, robust data association methods are necessary to mitigate these errors and prevent the vehicle from receiving inaccurate information. Furthermore, sharing these error bounds with the ego vehicle ensures that the vehicle operates with an awareness of the potential inaccuracies in the provided information. Temporal alignment is another consideration for data sources that are not synchronized. CV or CTRA models must be employed to predict object information at specific timestamps, compensating for communication delays and ensuring synchronization between the FSN and the vehicle.

Liu et al. [50] investigated the robustness of collaborative perception algorithms to time delays, localization errors, and lossy communication, which are unique challenges that impede the performance of cooperative systems. Experiments involved fine-tuning pre-trained models with latency ranging from 0 to 500 ms, using both fixed and random delay scenarios. On the V2XSet dataset [235], V2X-ViT [235] exhibited strong robustness to time delays, while early and late fusion methods showed poorer robustness. DiscoNet [236] performed best on DAIR-V2X-C [237]. Overall, time delay significantly impacts V2X perception, underscoring the need for robust fusion methods. To simulate localization errors during testing, Gaussian noise with varying standard deviations was used, while training occurred in perfect settings. Early fusion proved most sensitive to localization errors on DAIR-V2X-C [237]. V2XSet [235] results indicated heading errors have less impact than positional errors, with intermediate fusion demonstrating better robustness compared to early and late fusion methods. To examine lossy communication, models were trained in ideal settings and fine-tuned in lossy environments simulated by replacing elements in collaborative features with random noise at varying probabilities. Lossy communication negatively impacted most models, especially Where2comm [238] and V2X-ViT [235] on V2XSet [235]. V2VNet [220] and CoBEVT [173] showed better robustness with smaller performance drops. The methods highlight the importance of developing methods that are robust to time delays, localization errors, and lossy communication, which are common real-world challenges in V2X perception.

Several methods have been explored to overcome localization errors in cooperative systems. Gao et al. [239] proposed a multisensor fusion framework for cooperative localization in V2V networks, combining object detection with point cloud matching. Real-world experiments demonstrated that the relative pose refinement algorithm improved relative position accuracy by at least 20% compared to using only object detection, and the cooperative localization framework achieved decimeter-level absolute position accuracy, outperforming range-based and pure INS methods. The method was resilient to localization errors in neighboring vehicles and benefited from undirected communication topologies with multiple vehicles sharing information.

Similarly, Gao et al. [240] proposed an end-to-end learning framework to estimate and correct relative pose errors between an ego vehicle and its neighbors by leveraging shared intermediate neural features. In this framework, neighbors transform their LiDAR scans into the ego vehicle's frame, extract features using a backbone network, and share them with the ego vehicle, which concatenates them with its own features. A localization head network then estimates the relative pose error from the combined feature map. Experiments showed that voxel size significantly affected accuracy, with smaller voxels yielding better results but increasing transmission data size. The learned approach outperformed point cloud matching (PCM) under large initial pose errors but was inferior for small noise levels.

Xia et al. [241] introduced a secure cooperative localization method using consensus estimation while considering potential cyber attacks on shared sensory data. The method consists of a Consensus Kalman Information Filter (CKIF) for fusing sensory information from the ego vehicle and its neighbors, a Generalized Likelihood Ratio Test (GLRT) with a novel delay-prediction framework for real-time attack detection, and a rule-based method for isolating compromised measurements. Numerical simulations validated the effectiveness of the proposed method, demonstrating its ability to achieve higher accuracy and better security compared to a standard Kalman filter and the state-of-the-art MSMV method. The localization accuracy improved as the number of connected vehicles increased, and the CKIF was adaptable to dynamic communication topologies.

3) REAL TIME PERFORMANCE AND CHALLENGES FOR COLLABORATIVE SYSTEMS

The importance of real-time performance in cooperative V2I systems cannot be overstated, as it directly impacts both safety and reliability [242]. Real-time capabilities enable near-instantaneous decision-making in dynamic environments, however, the inherent complexity of processing and transmitting vast amounts of data within stringent time constraints raises concerns over latency. Latency can arise at various stages, such as perception algorithms, communication links, planning and control methods, and hardware constraints [49], [52], [243]. High latency can significantly impact performance, causing the perceived position of an object to lag behind its actual position. Resulting in a positional offset that leaves a trailing discrepancy along the object's path. As a result, optimization of these processes is essential to ensure the safety and reliability of the cooperative V2I system.

Within the perception task, the computational demands of tasks like object detection, semantic segmentation, sensor fusion, and localization make it difficult to achieve low-latency results at the high sensor frame rates required for dynamic scenes. Table 8 summarizes the real-time performance capabilities of 3D object detection tasks. Additionally, the computational costs of multi-agent sensor fusion hinder real-time perception [52]. Communication challenges arise

TABLE 8. Overview of Commonly Used 3D Object Detection Methods and Their Real-Time Performance Considerations

Sensor Modality	Method	Real-Time Performance Considerations
LiDAR	Projection Based	Lightweight 2D CNNs enable real-time inference but lower accuracy for small or occluded objects.
	Discretization Based	Moderate performance due to voxelization overhead and slower 3D convolutions.
	Point Based	Poor performance due to high computational overhead and point-level processing.
	Graph Based	Not suitable for real-time inference due to high computational costs and graph traversal overhead.
Radar	Point Cloud Based	Sparse point clouds reduce accuracy, and additional computations make real-time inference challenging.
	Pre-CFAR Based	Real-time performance is difficult due to the complexity of datacubes used in pre-processing.
Camera	2D Detection Driven	Moderate performance, heavily reliant on the speed and accuracy of the 2D detector.
	3D Shape Information	Poor real-time performance due to high computational requirements for reconstructing objects.
	Depth Estimation	Depth estimation introduces significant latency, making real-time inference difficult.
	Representation Transformation	Moderate performance, as detection is performed in the 3D domain, which scales computational complexity.
	End-to-End	Good performance due to an optimized, streamlined pipeline.

from the strict real-time data exchange requirements in safety-critical scenarios. For certain safety-critical applications, the acceptable latency typically ranges from 20 to 100 milliseconds [101]. Latency, limited bandwidth, packet loss, and scalability issues are significant barriers [244]. High data volumes [84] can exceed network capacities, making real-time transmission infeasible without specialized compression techniques or compact feature sharing. Planning and control face the complex task of generating safe, optimal trajectories in dynamic environments while considering vehicle dynamics and multi-agent interactions [52], [243]. Long planning horizons, non-linear vehicle models, non-convex constraints, and the complexity of multi-agent collision avoidance make real-time optimization computationally expensive. Hardware resource constraints on embedded platforms optimized for weight, size, power, and thermal efficiency result in limited computational resources, memory, and energy budgets. Balancing the demands of perception, planning, and control modules leads to resource contention, exacerbating latency issues and reducing throughput. High-bandwidth sensor data processing and high-frequency control loops strain these limited resources, especially as scene complexity or operating speeds increase

a) Edge Computing for V2X Latency Minimization: Edge computing [84], as discussed earlier, can be a key enabler for optimizing real-time performance in cooperative systems by bringing computation and data storage closer to the sources of data generation, such as vehicles and FSNs [52]. By processing data at the edge, latency is significantly reduced, responsiveness is improved, and real-time decision-making becomes possible. In a cooperative system powered by edge computing, agents are equipped with edge nodes capable of processing sensor data locally. Instead of transmitting vast amounts of raw data to central cloud servers, which can increase latency, edge nodes filter and aggregate only the most relevant information. Edge computing also enhances the reliability and scalability of V2I systems. In areas with limited or unreliable network connectivity, edge nodes ensure that critical functions can continue operating without interruption. As the number of connected vehicles and devices grows, the distributed nature of edge computing allows the system to scale by leveraging multiple edge nodes working in parallel.

4) FEDERATED LEARNING FOR PRIVACY-PRESERVING COOPERATIVE PERCEPTION

Some additional concerns for a cooperative sensing system include data protection, cybersecurity, and trust. Mulder et al. [245] examined the issues surrounding data protection for autonomous vehicles and driver monitoring systems (DMS). The authors discuss various aspects of GDPR, particularly focusing on the collection, protection, and sharing of sensitive data. The paper suggests that data protection could become part of approval requirements for automated vehicles. One of the challenges outlined was distinguishing between collecting data needed for safety and unnecessary collection or processing of personal data. It was also noted that managing cyber risks is crucial for autonomous agents. In some

countries, there are regulations that require the anonymization of sensitive and personal data in recorded information. This includes data related to government entities as well as individuals' faces, to ensure they are not recognizable or identifiable in the recorded data [246].

FL is one of the ways to tackle the privacy protection issue [247], [248], [249]. Vast amounts of real-time data are processed for perception and downstream tasks. The decentralized nature of FL ensures that the sensitive data remains on individual agents, however, this alone does not eliminate privacy risks. Several different techniques such as differential privacy, secure multiparty computation (SMPC), trusted execution environments (TEEs), and anonymization enable privacy-preserving FL approaches for connected vehicles. These techniques have been reviewed by Chellapandi et al. [230] and Yin et al. [250]. Each of these techniques introduces unique benefits, challenges, and trade-offs in terms of accuracy, scalability, latency, and computational overhead.

The goal of differential privacy [251] methods is to ensure individual data points can not be reverse engineering from FL updates. This is achieved by adding controlled noise into model updates, which will protect sensitive information while allowing meaningful learning. Local differential privacy techniques apply noise on the agent side, whereas global differential privacy methods add noise on the server side. The challenge with this method is balancing the privacy-accuracy trade-off. Excessive noise and overly stringent privacy measures can hinder the model's learning ability in real time. Truex et al. [252] introduced the LDP-Fed framework which proposed a local differential privacy (LDP) module and k-Client selection module to ensure data security during model training. The LDP module applies noise to the model update before transmitting to the server, while the k-Client selection module regulates which client can participate in the communication process, outperforming simpler mechanisms in balancing privacy and performance.

SMPC methods [253] aims to ensure that the data is never exposed during the FL process by applying homomorphic encryption. This approach is particularly useful in environments where secure aggregation of model updates from multiple agents ensures privacy even if one agent is compromised. Due to its high computational and communication overheads, this approach does not scale well and is not suitable for latency-sensitive applications. Li et al. [247] introduced the chain-PPFL framework where the agents are organized in a chain, and a single-masking mechanism is used. The single-masking mechanism ensures that each agent's local update is masked by the cumulative token. Each agent adds their local model update to a masked aggregate received from the previous participant and the masked value is passed randomly through the chain until the final participant sends the fully masked update to the central server. Following this, the server then removes the initial random mask to reveal only the aggregated update, without accessing individual contributions. This approach reduces communication overhead and outperforms simple homomorphic encryption methods, however, the performance is reliant on good connectivity between agents.

Anonymization techniques aim to remove or obscure identifiable attributes from model updates. Data masking methods can mask and transform sensitive information such as GPS coordinates or object location and pseudonymization [254] methods can assign pseudonyms to vehicle updates so that the identities are not linked to the sender. Ensuring that neither the central server or adversarial parties can't link it back to individual agents. Anonymization techniques alone cannot fully safeguard against certain attacks, such as inference attacks. Therefore, these techniques need to be implemented alongside differential privacy or SMPC methods.

Hardware-based approaches such as TEEs [255] ensure privacy by isolating sensitive computations within a secure, tamper-proof enclosure. TEEs onboard the agents can perform model updates securely, while server-side TEEs protect results during global aggregation. This approach holds an advantage over software-based approaches since the data can be safeguarded, even if the OS is compromised. However, this is not a standalone solution since TEEs are vulnerable to side-channel attacks, and furthermore, implementation in resource-constrained agent hardware also remains a challenge. In [256] the authors introduced a novel framework that performs model training in secure TEEs. This approach trains the model layers sequentially, where each layer is loaded into an agent's secure TEE, linked to the server's frozen layers, and trained locally on the agent's data. After each epoch, updated layer weights are aggregated across clients to refine the model for the next iteration. This approach mitigates privacy risks and remains computationally efficient by optimizing communication costs and utilizing memory resources during its layer-wise training.

5) ADVERSARIAL ATTACKS AND PREVENTION

While FL is a promising approach to address privacy concerns, significant data security challenges [257] still exist, as highlighted by Mothukuri et al. [258] in their comprehensive review of the security landscape in FL. The FL architecture exposes several points of vulnerability that can be exploited by adversaries. These include the communication link between participating agents and the central server, agent data manipulation, weaknesses in aggregation algorithms that may fail to detect abnormal updates, and the risk of a compromised central server. Malicious agents can leverage these vulnerabilities to launch various attacks, such as poisoning attacks [259] that manipulate the training data or model, inference attacks [260] that leak sensitive information about private training data, and backdoor attacks [261] that inject malicious functionality while preserving the model's performance on the main task. Additionally, advanced attacks using generative adversarial networks [262] pose a significant threat to the security of FL systems.

Various defensive techniques to address these challenges have been developed. These include Sniper [263], which uses Euclidean distance metrics to identify and exclude malicious client updates. Anomaly detection methods using clustering or auto-encoders to detect abnormal client behavior. Moving target defense strategies [264] dynamically change the attack surface, making it harder for attackers to exploit vulnerabilities. Model pruning techniques defend against backdoor attacks while reducing communication costs. Data sanitization methods filter out suspicious training data points to mitigate poisoning risks. TEEs provide hardware-based protections for code and data integrity and confidentiality. Decentralized FL architectures based on blockchain technology can enhance the robustness and transparency of FL systems by leveraging inherent security features such as immutability and distributed consensus.

Additionally, Kim et al. [265] provided an in-depth review of the cybersecurity landscape surrounding autonomous vehicles. The paper highlights the increasing susceptibility of autonomous vehicles to complex and interconnected attacks targeting various components, including in-vehicle systems, sensors, V2X communications, and infotainment systems. They discuss a wide range of vulnerabilities and the corresponding attacks that exploit them, such as ECU attacks, in-vehicle network attacks, sensor attacks, mobile app attacks, VANET attacks, and infotainment system attacks. Furthermore, the authors explore defensive techniques proposed by researchers to mitigate these threats, which include strengthening ECU and in-vehicle network security, improving sensor and perception security, securing external interfaces, protecting V2X communications, and leveraging AI/ML techniques for attack detection. The paper emphasizes the need for a comprehensive approach to address the complex cybersecurity challenges posed by autonomous vehicles, particularly considering the potential vulnerabilities introduced by 'black-box' approaches in autonomous driving systems that can be difficult to foresee.

6) TRUST IN COLLABORATIVE SYSTEMS

The concept of trust is a crucial consideration for self-driving applications. This concept allows the ego vehicle to validate the data collected by its own sensors, and ensure that the data has not been corrupted by adversarial attacks [266], [267], [268]. Furthermore, trust plays a vital role in cooperative driving applications, where the ego vehicle must be able to verify the reliability of the information it receives from other agents. Rosenstatter et al. [269] introduced the concept of a trust system that allows cooperative vehicles to assess trust in data from their own sensors and received data. To do this a trust index was generated to represent the reliability of the vehicle data and the received data. The results showed that the trust index can identify various traffic situations and adapt the level of trust during a merge scenario. One of the challenges outlined was that the ego-vehicle trust models must be capable of handling different levels of uncertainty in the received data and vehicle data. Furthermore, constantly estimating a trust metric requires consistent modeling and data fusion, which can be computationally expensive. Allig et al. [270] proposed a method to estimate the trust of vehicles in a Vehicular ad hoc Network (VANET) based on the consistency of their collective perception data. The authors stated that false perception data can dangerously mislead vehicles if trusted, and developing robust trust models that capture all the intricate factors affecting perception data is challenging.

B. ECONOMIC FEASIBILITY OF V2X DEPLOYMENT

The deployment of a cooperative V2I system presents several opportunities to enhance safety, mobility, and sustainability within the transportation ecosystem, however, enabling the full potential of these technologies requires navigating a complex landscape of economic, technical, and institutional challenges [271]. The deployment of cooperative V2I systems will require significant investments from various stakeholders, especially road operators and transportation agencies, who will play a key role in tasks such as installing FSNs, upgrading current infrastructure, and deploying the required support system [272]. Federal, state, and local bodies will likely provide the majority of the funding, however, there may be opportunities to share costs through public-private partnerships with auto manufacturers, mobile network operators, and other private sector stakeholders who stand to benefit from V2I services. For example, auto OEMs may be willing to subsidize some infrastructure investments in order to enable V2I features in their vehicles and gain a competitive advantage in the market. However, one significant barrier to mass adoption is the high upfront capital costs [272]. These expenses, coupled with market and regulatory uncertainty surrounding the enabling technologies, result in difficulty for agencies and manufacturers to build strong business cases for investment.

Regardless of the challenges, there exist several potential benefits of V2I technologies, which enable a range of safety, mobility, and environmental applications. Key safety use cases, such as red light violation warnings, curve speed warnings, and reduced speed zone warnings can reduce injuries and fatalities on the roads [273]. Several pilot development programs have quantified the benefits of these V2I systems. Queue warning systems for trucks were deployed that alerted motorists about traffic queues ahead of interstate work zones, resulting in an 80% decrease in hard braking events and thus lowering the likelihood of crashes [274]. Similarly, a cloud-based hazard warning system demonstrated a 90% reduction in the risk of collision with roadside workers [275].

While the numerous benefits are compelling, the deployment costs of these systems cannot be overlooked. The key cost components of V2I systems include FSNs, upgrades to existing roadside infrastructure, back-haul communications, software integration, and ongoing operations and maintenance [272], [276]. FSN costs can range from USD $900 to $5,250 per unit, with an additional $1,000 to $8,000 for installation and integration. Traffic signal controller upgrades can cost between $2,200 and $13,000 [276]. Back-haul and other

supporting infrastructure costs can vary widely depending on site conditions and deployment scale, Furthermore, ongoing operations and maintenance expenses are estimated to fall between $2,000 and $3,000 per deployment annually [276]. The cost of large-scale V2I deployment will vary, but it undoubtedly requires a multi-million dollar investment, particularly for larger states [272]. A considerable portion of the cost is associated with the number of devices needed to ensure sufficient coverage across a region. As the scale of deployment increases, the unit costs for equipment and installation decrease due to economies of scale, however, despite the declining technology costs over time, the overall investment necessary for V2I implementation will remain significant.

It is challenging to do a rigorous cost-benefit analysis (CBA) due to the uncertainty around the deployment costs. However, some of the key factors that will determine the cost-effectiveness of V2I include the number and severity of crashes prevented, the value of travel time saved through mobility applications, the lifespan and market penetration of the V2I infrastructure, the pace of deployment, the effectiveness of V2I applications in modifying driver behavior, and the ability to scale and replicate benefits from initial pilot sites. While the upfront costs of V2I deployment can be substantial, the long-term economic and societal benefits, such as reduced crashes, decreased fuel consumption, and travel time savings can be significant if the technology is effectively implemented and adopted by drivers.

Dong et al. [277] conducted a simulated CBA to compare the economic feasibility of manual, V2V-equipped, and V2I-equipped vehicles under various market penetration scenarios. The study considered four cost categories: road wear, fuel consumption, travel time, and equipment costs. The findings revealed that at low market penetration rates, less than 30%, manual vehicles had the lowest total cost per vehicle, closely followed by V2V and V2I-equipped vehicles. This was due to the limited benefits and high equipment costs of the technologies at this stage. However, as market penetration increased, the total cost per vehicle decreased for all scenarios. V2V-equipped vehicles outperformed manual vehicles at around 50% market penetration, primarily due to significant reductions in road wear, fuel consumption, and travel time costs. On the other hand, V2I-equipped vehicles had the highest total cost per vehicle at all market penetration rates, largely because of the substantial infrastructure costs involved. The sensitivity analysis, which considered a 50% reduction in V2V and V2I equipment costs, showed that V2V-equipped vehicles became the most cost-effective option at all market penetration rates, followed by V2I-equipped vehicles. This suggests that the economic feasibility of these technologies is highly sensitive to equipment costs.

To optimize the benefits of V2I deployments, agencies must carefully evaluate the costs and benefits on a case-by-case basis, considering factors such as crash history, traffic volumes, intersection geometry, and local conditions. By prioritizing locations with the highest potential for safety and mobility improvements, leveraging existing infrastructure, exploring cost-sharing partnerships, and phasing deployments over time, agencies can ensure that their V2I investments deliver the greatest possible benefits to road users and society as a whole.

C. CHALLENGES AND FUTURE RESEARCH

1) REQUIRED INFORMATION

One of the challenges is identifying what information is required to be shared in a V2I system. We explored one scenario, where object information has to be shared, however, this requirement is heavily dependent on the use case and the level of involvement by the FSN. One avenue for future research is investigating safety-critical scenarios where an FSN can be deployed, and then identifying what exact information is required by the vehicle. Identifying the specific information required for a variety of scenarios presents opportunities to tackle one of the key challenges of data protection highlighted by Mulder et al. [245], which stresses the importance of collecting only the data that is strictly necessary. Furthermore, identifying useful information for safety-critical and non-safety-critical applications will allow the design of FSN sensor systems to be efficient and more cost-effective.

2) MODEL COMPLEXITY AND DOMAIN GAP

Deep learning approaches have shown great promise in tackling complex perception tasks, however, their success heavily relies on the availability of substantial amounts of diverse and representative annotated training data. Recent advancements in multi-agent multi-modal research include the publication of several real-world datasets [235], [237], [278], [279]. Liu et al. [50] summarizes and analyzes some of these dataset. However, a significant challenge still remains due to the lack of datasets that are tailored to specific scenarios. Without such datasets, it becomes challenging to identify the specific scenarios where a FSN would be most beneficial and cost-effective to deploy. Additionally, the lack of datasets also limits our understanding of how V2I cooperative sensing would perform in intricate and dynamic scenes, where multiple agents with varying sensor capabilities interact.

To evaluate and enhance the robustness of LiDAR-based V2X perception, Xiang et al. [280] proposed V2XP-ASG, a framework for generating adversarial driving scenarios. It first identifies collaborator combinations whose viewpoints create perception vulnerabilities, prioritizing weaker collaborators using learned attention weights. Then, it strategically perturbs agent poses using occlusion heuristics and black-box optimization to minimize perception accuracy. Additionally, studies into domain adaptation approaches are needed, where annotated simulated data can be used alongside real-world data to overcome the need for extensive labeling [281], [282], [283], [284].

Domain gaps issues resulting from private data sources and multi-agent system discrepancies remain a challenge. Xu et al. [285] proposed a solution to address the domain gap issue that arises from agents using diverse neural network architectures by proposing a Multi-agent Perception Domain Adaptation (MPDA) framework. In [286], Li et al. address the distribution gap arising from when agents trained on

different private datasets exhibit feature distribution mismatches despite using identical network architectures.

3) MAP FUSION AND SYNCHRONIZATION

Map fusion and data alignment is another key challenge of V2I cooperative sensing. Sensor extrinsic parameter drift, due to deviations over time, can lead to geo-localization errors of objects. Additionally, data transmission delays can add temporal discrepancies, resulting in misaligned objects reducing the effectiveness of cooperation. Furthermore, errors compounding from the sensing systems to the mapping task can result in misaligned and duplicated objects. Identifying what coordinate system to represent the unified map also has challenges. The FSN sharing information in the vehicle coordinate system is only feasible if the relative pose between the vehicle and the FSN is accurate. Inaccuracies in this relative pose may result in duplicated objects. Future studies should explore how these object location errors affect the performance of ego vehicles for safety-critical scenarios. This will define acceptable object location error tolerances, which can then be considered for spatial and temporal alignment tasks. Investigations into appropriate coordinate systems for V2I cooperative sensing must be conducted. Although the impact of relative pose errors on object alignment is discussed, adopting a local coordinate system based on landmarks or adhering to a global coordinate system may involve tradeoffs between performance and reliability, and these tradeoffs need to be thoroughly investigated.

D. OPPORTUNITIES

The deployment of cooperative V2I systems presents several significant opportunities. As previously mentioned one of the main benefits is the enhanced road safety by improved object detection and tracking performance, as well as providing wider spatial coverage. The cooperative perspective from the FSN enables the connected vehicles to quickly detect and react to potential hidden dangers on the road, resulting in safer conditions for vehicles and VRUs. Additionally, V2I systems offer improved traffic efficiency through applications such as enhanced road guidance, GLOSA systems, and vehicle platooning [271]. The improved traffic efficiency leads to a substantial reduction in the time passenger cars and buses spend waiting on roads in urban areas, as reported in [287]. Furthermore, this enhanced traffic efficiency contributes to reduced environmental impacts [272]. Additionally, Li et al. [23] proposed a cooperative perception framework to enhance real-time traffic management by estimating and predicting individual vehicle speeds and positions. This approach proves to be effective in situations with limited connected and automated vehicle adoption, achieving 80-90% accuracy with a 50% penetration rate. The framework also leverages roadside detector data to further improve accuracy, especially when fewer connected vehicles are present, and offers robust short-term prediction capabilities.

Enhanced emergency response [271] through cooperation is another opportunity that can greatly enhance emergency response efforts by enabling collaboration between emergency vehicles, conventional vehicles, roadside infrastructure, and traffic management systems. These systems work together to ensure emergency responders reach incident sites quickly, safely, and efficiently. By preemptively clearing paths for emergency vehicles and dynamically adjusting traffic flow, they reduce response times while minimizing disruptions to regular traffic.

Several methods have been proposed to improve emergency vehicle preemption. Figueiredo et al. [288] proposed a method for detecting emergency vehicles using CAMs and radar data. The system predicts the emergency vehicle's future location in real-time and disseminates warning messages to nearby agents ahead of its arrival. By utilizing both I2V and V2V communication, the approach minimizes lost warning messages and achieves 80% message delivery within 100 ms. Similarly, Ding et al. [289] proposed a method to improve emergency vehicle preemption efficiency using a deep reinforcement learning-based mobility-aware lane change algorithm (DRL-MLC). The emergency vehicle employs a policy-based deep reinforcement learning algorithm to determine the shortest trajectory for lane changes, while connected agents execute mobility-aware lane changes to clear the path for the emergency vehicle, guided by the emergency messages they receive. However, disseminating emergency messages faces several challenges, as highlighted by the authors in [290], [291]. These challenges include the broadcast storm problem [292], hidden node problem [293], packet collision, scalability, and accurate positioning.

Cost is a critical consideration that automotive OEMs must carefully assess when developing and deploying ADAS and autonomous driving technologies, particularly for systems classified as SAE level 2 or above. As the levels of automation required by vehicles progress, the complexity of the tasks they need to perform increases significantly. This results in the use of more sophisticated sensors and the deployment of more advanced algorithms. Furthermore, the hardware requirements for running these complex algorithms while meeting stringent safety and performance criteria will also increase. All of this will result in a substantial increase in the overall system cost, which may be a burden to the OEMs.

V2I cooperative sensing can potentially reduce some of these costs in the future [277], [287]. In Germany, Mercedes-Benz offers vehicles with L3 capabilities that can only be activated on motorways, however, the high cost of these vehicles hinders mass adoption. If large-scale adoption of V2I technology becomes feasible, FSN deployed along roads or motorways can bear some responsibility for some of the sensing and processing tasks. By offloading certain tasks to the FSN, vehicles can potentially reduce their reliance on expensive onboard sensors and processing capabilities, leading to cost savings for OEMs [277]. Furthermore, government agencies can invest in the installation and maintenance of these FSNs, creating a shared resource that benefits all road users. This collaborative approach can help distribute the costs associated with advanced sensing and processing across

multiple stakeholders, making it more affordable for individual OEMs to develop and deploy ADAS and autonomous driving technologies.

VI. CONCLUSION

This comprehensive review provides a systematic overview of the technologies required for an effective V2I cooperative sensing system. This paper explores the existing technologies and challenges surrounding each of the critical sub-components of a cooperative system. Following this, a scenario where a V2I cooperative sensing framework can be deployed is explored while discussing the challenges the entire architecture faces, and highlighting future research topics and opportunities in this field.

Our ground-up design approach for this cooperative sensing system begins by identifying the scene information that must be captured and the appropriate sensors required to acquire this data. Sensor calibration methods are then delved into and an overview of data transmission technologies is provided. With the collected data, various methods for perception tasks are explored, first focusing on individual agent perspectives and then on cooperative approaches. Different mapping techniques such as PV and BEV-based mapping approaches are then investigated, followed by an extensive review of map fusion methods. The technologies and methods mentioned above are then used to implement a practical V2I cooperative sensing system from the ground up. Here the difficulties that the architecture faces such as data protection, cybersecurity, real-time performance, map fusion and synchronization, task-specific model selection and deployment, and identifying the most beneficial use cases for cooperative sensing are discussed. Following this, the economic feasibility of this system is explored.

The paper highlights some of the potential benefits of a cooperative V2I system, however, significant challenges exist at each stage of the pipeline, requiring careful consideration of the specific requirements and design choices for each use case. It is not possible to deploy a single solution for this cooperative system as the details are heavily dependent on the use-case requirements. Large-scale deployment of V2I cooperative sensing systems is only going to be possible through collaboration between research institutes, automotive OEMs, and government organizations. These entities must work together to identify the specific requirements for safety-critical use cases, thus defining the involvement required by the FSN to enable the design and deployment of cost-effective and efficient FSNs.

REFERENCES

[1] World Health Organization, "Road traffic injuries," Accessed: Jul. 26, 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

[2] Economic Times India, "Rising adoption of adas to drive auto tech market to USD 134.06 bn by 2025," Accessed: Jul. 26, 2024. [Online]. Available: https://auto.economictimes.indiatimes.com/news/auto-technology/rising-adoption-of-adas-to-drive-auto-tech-market-to-usd-134-06-bn-by-2025-report/87960929

[3] Consumer Reports, "Avoiding crashes with self driving cars," Accessed: Jul. 26, 2024. [Online]. Available: https://www.consumerreports.org/cro/magazine/2014/04/the-road-to-self-driving-cars/index.htm

[4] Tesla, "Autopilot and full self driving capability," Accessed: Jul. 26, 2024. [Online]. Available: https://www.tesla.com/en_ie/support/autopilot

[5] BMW, "Overview of the main driver assistance systems," Accessed: Jul. 26, 2024. [Online]. Available: https://www.bmw.com/en/innovation/the-main-driver-assistance-systems.html

[6] Toyota, "Toyota safety sense," Accessed: Jul. 26, 2024. [Online]. Available: https://www.toyota.ie/discover-toyota/safety/toyota-safety-sense

[7] Audi, "Driver assistance systems," Accessed: Jul. 26, 2024. [Online]. Available: https://www.audi.ie/ie/web/en/customer-area/audi-owners/driver-assistance-systems.html

[8] Mercedes-Benz Group, "Mercedes-Benz World's first automotive company to certify SAE level 3 system for U.S. market," Accessed: Jul. 26, 2024. [Online]. Available: https://group.mercedes-benz.com/innovation/product-innovation/autonomous-driving/drive-pilot-nevada.html

[9] Waymo, "Waymo one," Accessed: Jul. 26, 2024. [Online]. Available: https://waymo.com/waymo-one/

[10] Cruise, "Cruise self driving cars," Accessed: Jul. 26, 2024. [Online]. Available: https://getcruise.com/rides/

[11] Zoox, "The 'full-stack' - behind autonomous driving," Accessed: Jul. 26, 2024. [Online]. Available: https://zoox.com/autonomy/

[12] Baidu, "Baidu starts offering nighttime driverless taxis in China," Accessed: Jul. 26, 2024. [Online]. Available: https://techcrunch.com/2022/12/26/baidu-night-time-driverless-robotaxis-china/

[13] E. Yurtsever, J. Lambert, A. Carballo, and K. Takeda, "A survey of autonomous driving: Common practices and emerging technologies," *IEEE Access*, vol. 8, pp. 58443–58469, 2020.

[14] Y. Li, J. Moreau, and J. Ibanez-Guzman, "Emergent visual sensors for autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 5, pp. 4716–4737, May 2023.

[15] F. de Ponte Müller, "Survey on ranging sensors and cooperative techniques for relative positioning of vehicles," *Sensors*, vol. 17, no. 2, 2017, Art. no. 271.

[16] 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, 2021, Art. no. 2140.

[17] Y. Peng, Y. Qin, X. Tang, Z. Zhang, and L. Deng, "Survey on image and point-cloud fusion-based object detection in autonomous vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 22772–22789, Dec. 2022.

[18] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger, "Transfuser: Imitation with transformer-based sensor fusion for autonomous driving," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 12878–12895, Nov. 2023.

[19] J. Vargas, S. Alsweiss, O. Toker, R. Razdan, and J. Santos, "An overview of autonomous vehicles sensors and their vulnerability to weather conditions," *Sensors*, vol. 21, no. 16, 2021, Art. no. 5397.

[20] A. Palffy, J. F. Kooij, and D. M. Gavrila, "Detecting darting out pedestrians with occlusion aware sensor fusion of radar and stereo camera," *IEEE Trans. Intell. Veh.*, vol. 8, no. 2, pp. 1459–1472, Feb. 2023.

[21] European New Car Assessment Programme, "Test protocol-aeb VRU systems," Accessed: Jul. 26, 2024. [Online]. Available: https://cdn.euroncap.com/media/58226/euro-ncap-aeb-vru-test-protocol-v303.pdf

[22] BOSCH, "Cooperative driving of automated vehicles with V2X technology," Accessed: Sep. 03, 2024. [Online]. Available: https://www.bosch.com/stories/cooperative-driving/

[23] T. Li, X. Han, and J. Ma, "Cooperative perception for estimating and predicting microscopic traffic states to manage connected and automated traffic," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 8, pp. 13694–13707, Aug. 2022.

[24] G. Thandavarayan, M. Sepulcre, J. Gozalvez, and B. Coll-Perales, "Scalable cooperative perception for connected and automated driving," *J. Netw. Comput. Appl.*, vol. 216, 2023, Art. no. 103655.

[25] R. N. Albustanji, S. Elmanaseer, and A. A. Alkhatib, "Robotics: Five senses plus one—An overview," *Robotics*, vol. 12, no. 3, 2023, Art. no. 68.

[26] A. Eskandarian, C. Wu, and C. Sun, "Research advances and challenges of autonomous and connected ground vehicles," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 2, pp. 683–711, Feb. 2021.

[27] I. Kostavelis and A. Gasteratos, "Semantic mapping for mobile robotics tasks: A survey," *Robot. Auton. Syst.*, vol. 66, pp. 86–103, 2015.

[28] E. Galceran and M. Carreras, "A survey on coverage path planning for robotics," *Robot. Auton. Syst.*, vol. 61, no. 12, pp. 1258–1276, 2013.

[29] I. Lluvia, E. Lazkano, and A. Ansuategi, "Active mapping and robot exploration: A survey," *Sensors*, vol. 21, no. 7, 2021, Art. no. 2445.

[30] S. Thrun et al., "Robotic mapping: A survey," in *Exploring Artificial Intelligence in the New Millennium*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2003, pp. 1–35.

[31] *Taxonomy and Definitions for Terms Related to Cooperative Driving Automation for On-Road Motor Vehicles*, SAE International Standard J3216, SAE International, Warrendale, PA, USA, Jul. 2021.

[32] M. Lauer and A. Gerrits, "Next steps for the grand cooperative driving challenge [its events]," *IEEE Intell. Transp. Syst. Mag.*, vol. 1, no. 4, pp. 24–32, Winter 2009.

[33] C. Englund et al., "The grand cooperative driving challenge 2016: Boosting the introduction of cooperative automated vehicles," *IEEE Wireless Commun.*, vol. 23, no. 4, pp. 146–152, Aug. 2016.

[34] T. Robinson, E. Chan, and E. Coelingh, "Operating platoons on public motorways: An introduction to the sartre platooning programme," in *Proc. 17th World Congr. Intell. Transport Syst.*, Busan, South Korea, 2010, vol. 1, Art. no. 12.

[35] CARMA, "CARMA research tracks FHWA," Accessed: Jul. 26, 2024. [Online]. Available: https://highways.dot.gov/research/operations/carma-research-tracks

[36] ITS Inteternational, "IntelliDrive, connectivity, safety, mobility and the environment?" Accessed: Jul. 26, 2024. [Online]. Available: https://www.itsinternational.com/its8/feature/intellidrive-connectivity-safety-mobility-and-environment

[37] M. Gabb, H. Digel, T. Müller, and R.-W. Henn, "Infrastructure-supported perception and track-level fusion using edge computing," in *Proc. IEEE Intell. Veh. Symp.*, 2019, pp. 1739–1745.

[38] Providentia, "Providentia: The testfield for autonomous and automated driving," Accessed: Jul. 26, 2024. [Online]. Available: https://innovation-mobility.com/en/project-providentia/

[39] CORDIS, "Road Infrastructure ready for mixed vehicle traffic flows," Accessed: Jul. 26, 2024. [Online]. Available: https://cordis.europa.eu/project/id/723016/reporting

[40] A. Carreras, X. Daura, J. Erhart, and S. Ruehrup, "Road infrastructure support levels for automated driving," in *Proc. 25th ITS World Congr.*, Copenhagen, Denmark, 2018, pp. 17–21.

[41] Z. Bai et al., "A survey and framework of cooperative perception: From heterogeneous singleton to hierarchical cooperation," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 11, pp. 15191–15209, Nov. 2024.

[42] G. Yu, H. Li, Y. Wang, P. Chen, and B. Zhou, "A review on cooperative perception and control supported infrastructure-vehicle system," *Green Energy Intell. Transp.*, vol. 1, no. 3, 2022, Art. no. 100023.

[43] G. Cui, W. Zhang, Y. Xiao, L. Yao, and Z. Fang, "Cooperative perception technology of autonomous driving in the internet of vehicles environment: A review," *Sensors*, vol. 22, no. 15, 2022, Art. no. 5535.

[44] Z. Bai, G. Wu, X. Qi, Y. Liu, K. Oguchi, and M. J. Barth, "Infrastructure-based object detection and tracking for cooperative driving automation: A survey," in *Proc. IEEE Intell. Veh. Symp.*, 2022, pp. 1366–1373.

[45] P. Ghorai, A. Eskandarian, Y.-K. Kim, and G. Mehr, "State estimation and motion prediction of vehicles and vulnerable road users for cooperative autonomous driving: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 16983–17002, Oct. 2022.

[46] A. Caillot, S. Ouerghi, P. Vasseur, R. Boutteau, and Y. Dupuis, "Survey on cooperative perception in an automotive context," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 14204–14223, Sep. 2022.

[47] Y. Han, H. Zhang, H. Li, Y. Jin, C. Lang, and Y. Li, "Collaborative perception in autonomous driving: Methods, datasets, and challenges," *IEEE Intell. Transp. Syst. Mag.*, vol. 15, no. 6, pp. 131–151, Nov./Dec. 2023.

[48] X. Gao et al., "A survey of collaborative perception in intelligent vehicles at intersections," *IEEE Trans. Intell. Veh.*, early access, May 01, 2024, doi: 10.1109/TIV.2024.3395783.

[49] T. Huang et al., "V2X cooperative perception for autonomous driving: Recent advances and challenges," 2023, *arXiv:2310.03525*.

[50] S. Liu et al., "Towards vehicle-to-everything autonomous driving: A survey on collaborative perception," 2023, *arXiv:2308.16714*.

[51] J. Wang, Z. Wu, Y. Liang, J. Tang, and H. Chen, "Perception methods for adverse weather based on vehicle infrastructure cooperation system: A review," *Sensors*, vol. 24, no. 2, 2024, Art. no. 374.

[52] E. Y. Bejarbaneh, H. Du, and F. Naghdy, "Exploring shared perception and control in cooperative vehicle-intersection systems: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 11, pp. 15247–15272, Nov. 2024.

[53] J. Van Brummelen, M. O'brien, D. Gruyer, and H. Najjaran, "Autonomous vehicle perception: The technology of today and tomorrow," *Transp. Res. Part C, Emerg. Technol.*, vol. 89, pp. 384–406, 2018.

[54] J. Malik et al., "The three R's of computer vision: Recognition, reconstruction and reorganization," *Pattern Recognit. Lett.*, vol. 72, pp. 4–14, 2016.

[55] C. Eising, J. Horgan, and S. Yogamani, "Near-field perception for low-speed vehicle automation using surround-view fisheye cameras," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 13976–13993, Sep. 2022.

[56] L.-H. Wen and K.-H. Jo, "Deep learning-based perception systems for autonomous driving: A comprehensive survey," *Neurocomputing*, vol. 489, pp. 255–270, 2022.

[57] Y. Deng, T. Zhang, G. Lou, X. Zheng, J. Jin, and Q.-L. Han, "Deep learning-based autonomous driving systems: A survey of attacks and defenses," *IEEE Trans. Ind. Informat.*, vol. 17, no. 12, pp. 7897–7912, Dec. 2021.

[58] A. Gupta, A. Anpalagan, L. Guan, and A. S. Khwaja, "Deep learning for object detection and scene perception in self-driving cars: Survey, challenges, and open issues," *Array*, vol. 10, 2021, Art. no. 100057.

[59] D. Molloy et al., "Impact of ISP tuning on object detection," *J. Imag.*, vol. 9, no. 12, 2023, Art. no. 260.

[60] X. Yang and S. Fang, "Effect of field of view on the accuracy of camera calibration," *Optik*, vol. 125, no. 2, pp. 844–849, 2014.

[61] Z. Long et al., "Bio-inspired visual systems based on curved image sensors and synaptic devices," *Mater. Today Electron.*, vol. 6, 2023, Art. no. 100071.

[62] T. Brophy et al., "A review of the impact of rain on camera-based perception in automated driving systems," *IEEE Access*, vol. 11, pp. 67040–67057, 2023.

[63] Y. Zhang, A. Carballo, H. Yang, and K. Takeda, "Perception and sensing for autonomous vehicles under adverse weather conditions: A survey," *ISPRS J. Photogrammetry Remote Sens.*, vol. 196, pp. 146–177, 2023.

[64] A. Wilson, K. A. Gupta, B. H. Koduru, A. Kumar, A. Jha, and L. R. Cenkeramaddi, "Recent advances in thermal imaging and its applications using machine learning: A review," *IEEE Sensors J.*, vol. 23, no. 4, pp. 3395–3407, Feb. 2023.

[65] K. Mowafy, T. EL-Dessouky, and M. Medhat, "Design of IR zoom lens system for long-range detection in uncooled LWIR camera," *J. Opt.*, vol. 52, no. 1, pp. 281–289, 2023.

[66] FLIR, "As far as 30 FPs vs. 9 FPs video rates are concerned, why use one over the other?" Accessed: Jul. 26, 2024. [Online]. Available: https://www.flir.com/support-center/oem/as-far-as-30-fps-vs.-9-fps-video-rates-are-concerned-why-use-one-over-the-other/

[67] G. Druart et al., "Evaluation of the potential of high index chalcogenide lenses for automotive applications," in *Proc. Conf. SPIE Adv. Opt. Imag. Appl., UV Through LWIR*, 2022, vol. 12103, pp. 113–124.

[68] J. M. Rivera Velázquez et al., "Analysis of thermal imaging performance under extreme foggy conditions: Applications to autonomous driving," *J. Imag.*, vol. 8, no. 11, 2022, Art. no. 306.

[69] G. Gallego et al., "Event-based vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 1, pp. 154–180, Jan. 2022.

[70] Z. Kütük and G. Algan, "Semantic segmentation for thermal images: A comparative survey," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 286–295.

[71] S. Jia, "Event camera survey and extension application to semantic segmentation," in *Proc. 4th Int. Conf. Image Process. Mach. Vis.*, 2022, pp. 115–121.

[72] X. Chen, K. Kundu, Z. Zhang, H. Ma, S. Fidler, and R. Urtasun, "Monocular 3D object detection for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2147–2156.

[73] B. Li, W. Ouyang, L. Sheng, X. Zeng, and X. Wang, "GS3D: An efficient 3D object detection framework for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1019–1028.

[74] N. Li et al., "A progress review on solid-state lidar and nanophotonics-based lidar sensors," *Laser Photon. Rev.*, vol. 16, no. 11, 2022, Art. no. 2100511.

[75] Z. Dai, A. Wolf, P.-P. Ley, T. Glück, M. C. Sundermeier, and R. Lachmayer, "Requirements for automotive lidar systems," *Sensors*, vol. 22, no. 19, 2022, Art. no. 7532.

[76] S. Royo and M. Ballesta-Garcia, "An overview of lidar imaging systems for autonomous vehicles," *Appl. Sci.*, vol. 9, no. 19, 2019, Art. no. 4093.

[77] J. Lambert et al., "Performance analysis of 10 models of 3D lidars for automated driving," *IEEE Access*, vol. 8, pp. 131699–131722, 2020.

[78] J. Abdo, S. Hamblin, and G. Chen, "Effective range assessment of lidar imaging systems for autonomous vehicles under adverse weather conditions with stationary vehicles," *ASCE-ASME J. Risk Uncertainty Eng. Syst., Part B, Mech. Eng.*, vol. 8, no. 3, 2022, Art. no. 031103.

[79] E. Browell, S. Ismail, and W. Grant, "Differential absorption lidar (dial) measurements from air and space," *Appl. Phys. B*, vol. 67, pp. 399–410, 1998.

[80] R. Roriz, J. Cabral, and T. Gomes, "Automotive lidar technology: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6282–6297, Jul. 2022.

[81] Y. Zhou, L. Liu, H. Zhao, M. López-Benítez, L. Yu, and Y. Yue, "Towards deep radar perception for autonomous driving: Datasets, methods, and challenges," *Sensors*, vol. 22, no. 11, 2022, Art. no. 4208.

[82] A. Srivastav and S Mandal, "Radars for autonomous driving: A review of deep learning methods and challenges," *IEEE Access*, vol. 11, pp. 97147–97168, 2023.

[83] R. Prophet, A. Deligiannis, J.-C. Fuentes-Michel, I. Weber, and M. Vossiek, "Semantic segmentation on 3D occupancy grids for automotive radar," *IEEE Access*, vol. 8, pp. 197917–197930, 2020.

[84] J. Clancy et al., "Evaluating the feasibility of intelligent blind road junction V2I deployments," *Smart Cities*, vol. 7, no. 3, pp. 973–990, 2024.

[85] X. Cai et al., "Analyzing infrastructure lidar placement with realistic lidar simulation library," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 5581–5587.

[86] W. Jiang et al., "Optimizing the placement of roadside lidars for autonomous driving," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 18381–18390.

[87] X. Liu et al., "V2X-DSI: A density-sensitive infrastructure lidar benchmark for economic vehicle-to-everything cooperative perception," in *Proc. IEEE Intell. Veh. Symp.*, 2024, pp. 490–495.

[88] D. A. Forsyth and J. Ponce, *Computer Vision: A Modern Approach*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2002.

[89] O. Kainz, F. Jakab, P. Fecil'ak, R. Vápeník, A. Deák, and D. Cymbalák, "Estimation of camera intrinsic matrix parameters and its utilization in the extraction of dimensional units," in *Proc. Int. Conf. Emerg. eLearn. Technol. Appl.*, 2016, pp. 153–156.

[90] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 11, pp. 1330–1334, Nov. 2000.

[91] Z. Liu, Q. Wu, S. Wu, and X. Pan, "Flexible and accurate camera calibration using grid spherical images," *Opt. Exp.*, vol. 25, no. 13, pp. 15269–15285, 2017.

[92] J. Hieronymus, "Comparison of methods for geometric camera calibration," *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 39, pp. 595–599, 2012.

[93] S. Krüger, M. Scheele, and R. Schuster, "New calibration scheme for panoramic line scanners," in *Proc. 2nd Panoramic Photogrammetry Workshop. Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, 2005, vol. 36, no. 5, Art. no. W8.

[94] K. Liao et al., "Deep learning for camera calibration and beyond: A survey," 2023, *arXiv:2303.10559*.

[95] J. Domhof, J. F. Kooij, and D. M. Gavrila, "An extrinsic calibration tool for radar, camera and lidar," in *Proc. Int. Conf. Robot. Automat.*, 2019, pp. 8107–8113.

[96] Z. Chai, Y. Sun, and Z. Xiong, "A novel method for lidar camera calibration by plane fitting," in *Proc. IEEE/ASME Int. Conf. Adv. Intell. Mechatron.*, 2018, pp. 286–291.

[97] A. Rashd, W. Hardt, A. Kolker, M. Bdiwi, and M. Putz, "Open-box target for extrinsic calibration of lidar, camera and industrial robot," in *Proc. 3rd Int. Conf. Mechatron., Robot. Automat.*, 2020, pp. 121–125.

[98] C.-L. Lee, Y.-H. Hsueh, C.-C. Wang, and W.-C. Lin, "Extrinsic and temporal calibration of automotive radar and 3D lidar," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2020, pp. 9976–9983.

[99] S. Wang et al., "Temporal and spatial online integrated calibration for camera and lidar," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst.*, 2022, pp. 3016–3022.

[100] S. Fan, Z. Wang, X. Huo, Y. Wang, and J. Liu, "Calibration-free BEV representation for infrastructure perception," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2023, pp. 9008–9013.

[101] J. Clancy et al., "Wireless access for V2X communications: Research, challenges and opportunities," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 3, pp. 2082–2119, Thirdquarter 2024.

[102] Z. MacHardy, A. Khan, K. Obana, and S. Iwashina, "V2X access technologies: Regulation, research, and remaining challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1858–1877, Thirdquarter 2018.

[103] Ericsson, "Leveraging the potential of 5G millimeter wave," Accessed: Sep. 20, 2024. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/further-insights/leveraging-the-potential-of-5g-millimeter-wave

[104] IEEE Future Networks, "Massive MIMO," Accessed: Sep. 20, 2024. [Online]. Available: https://futurenetworks.ieee.org/topics/massive-mimo

[105] R. Hadani et al., "Orthogonal time frequency space modulation," in *Proc. IEEE Wireless Commun. Netw. Conf.*, 2017, pp. 1–6.

[106] E. Dahlman, S. Parkvall, and J. Sköld, "Chapter 20 - Beyond the first release of 5G," in *5G NR: The Next Generation Wireless Access Technology*, E. Dahlman, S. Parkvall, and J. Sköld, Eds. New York, NY, USA: Academic, 2018, pp. 407–414.

[107] S. Gyawali, S. Xu, Y. Qian, and R. Q. Hu, "Challenges and solutions for cellular based V2X communications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 222–255, Firstquarter 2021.

[108] J. M. Lozano Dominguez and T. J. Mateo Sanguino, "Review on V2X, I2X, and P2X communications and their applications: A comprehensive analysis over time," *Sensors*, vol. 19, no. 12, 2019, Art. no. 2756.

[109] F. A. Schiegg, I. Llatser, D. Bischoff, and G. Volk, "Collective perception: A safety perspective," *Sensors*, vol. 21, no. 1, 2020, Art. no. 159.

[110] G. Thandavarayan, M. Sepulcre, and J. Gozalvez, "Generation of cooperative perception messages for connected and automated vehicles," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 16336–16341, Dec. 2020.

[111] M. Correia, J. Almeida, P. C. Bartolomeu, J. A. Fonseca, and J. Ferreira, "Performance assessment of collective perception service supported by the roadside infrastructure," *Electronics*, vol. 11, no. 3, 2022, Art. no. 347.

[112] ETSI, "Intelligent transport system (ITS); vehicular communications," ITS ETSI, Sophia Antipolis, France, DraftTech. Rep. 103 562 V0. 0.15, 2019.

[113] N. Mellegård and F. Reichenberg, "The day 1 C-ITS application green light optimal speed advisory—A mapping study," *Transp. Res. Procedia*, vol. 49, pp. 170–182, 2020.

[114] H. Cui, Q. Wei, and L. Wang, "Matching theory based spectrum sharing for V2X communications in platoon scenarios," in *Proc. IEEE Int. Conf. Commun. Syst.*, 2018, pp. 326–331.

[115] M. M. Saad, M. M. Islam, M. A. Tariq, M. T. R. Khan, and D. Kim, "Collaborative multi-agent resource allocation in C-V2X mode 4," in *Proc. 12th Int. Conf. Ubiquitous Future Netw.*, 2021, pp. 7–10.

[116] M. Ji et al., "Graph neural networks and deep reinforcement learning based resource allocation for V2X communications," *IEEE Internet Things J.*, vol. 12, no. 4, pp. 3613–3628, Feb. 2025.

[117] K. Huang, B. Shi, X. Li, X. Li, S. Huang, and Y. Li, "Multi-modal sensor fusion for auto driving perception: A survey," 2022, *arXiv:2202.02703*.

[118] M. Pollach, F. Schiegg, and A. Knoll, "Low latency and low-level sensor fusion for automotive use-cases," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 6780–6786.

[119] Aptiv, "What is sensor fusion?" Accessed: Jul. 26, 2024. [Online]. Available: https://www.aptiv.com/en/insights/article/what-is-sensor-fusion

[120] A. Rövid, V. Remeli, and Z. Szalay, "Raw fusion of camera and sparse lidar for detecting distant objects," *At-Automatisierungstechnik*, vol. 68, no. 5, pp. 337–346, 2020.

[121] J. Kim, Y. Kim, and D. Kum, "Low-level sensor fusion network for 3D vehicle detection using radar range-azimuth heatmap and monocular image," in *Proc. Asian Conf. Comput. Vis.*, 2020, pp. 388–402.

[122] A. Abbasi, S. Queirós, N. M. da Costa, J. C. Fonseca, and J. Borges, "Sensor fusion approach for multiple human motion detection for indoor surveillance use-case," *Sensors*, vol. 23, no. 8, 2023, Art. no. 3993.

[123] 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 *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug., 2020, pp. 720–736.

[124] D. Xu, D. Anguelov, and A. Jain, "Pointfusion: Deep sensor fusion for 3D bounding box estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 244–253.

[125] X. Chen, H. Ma, J. Wan, B. Li, and T. Xia, "Multi-view 3D object detection network for autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1907–1915.

[126] J. Ku, M. Mozifian, J. Lee, A. Harakeh, and S. L. Waslander, "Joint 3D proposal generation and object detection from view aggregation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2018, pp. 1–8.

[127] Y. Cui et al., "Deep learning for image and point cloud fusion in autonomous driving: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 2, pp. 722–739, Feb. 2022.

[128] X. Chen, T. Zhang, Y. Wang, Y. Wang, and H. Zhao, "FUTR3D: A unified sensor fusion framework for 3D detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 172–181.

[129] Y. Wu, Y. Wang, S. Zhang, and H. Ogai, "Deep 3D object detection networks using lidar data: A review," *IEEE Sensors J.*, vol. 21, no. 2, pp. 1152–1171, Jan. 2021.

[130] G. Zamanakos, L. Tsochatzidis, A. Amanatiadis, and I. Pratikakis, "A comprehensive survey of lidar-based 3D object detection methods with deep learning for autonomous driving," *Comput. Graph.*, vol. 99, pp. 153–181, 2021.

[131] Y. Guo, H. Wang, Q. Hu, H. Liu, L. Liu, and M. Bennamoun, "Deep learning for 3D point clouds: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 12, pp. 4338–4364, Dec. 2021.

[132] Z. Liang and Y. Huang, "Survey on deep learning-based 3D object detection in autonomous driving," *Trans. Inst. Meas. Control*, vol. 45, no. 4, pp. 761–776, 2023.

[133] S. Y. Alaba and J. E. Ball, "A survey on deep-learning-based lidar 3D object detection for autonomous driving," *Sensors*, vol. 22, no. 24, 2022, Art. no. 9577.

[134] Z. Zhou, "How does sparse convolution work?" Accessed: Sep. 09, 2024. [Online]. Available: https://towardsdatascience.com/how-does-sparse-convolution-work-3257a0a8fd1

[135] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 652–660.

[136] Z. Pan, F. Ding, H. Zhong, and C. X. Lu, "RaTrack: Moving object detection and tracking with 4D radar point cloud," in *Proc. IEEE Int. Conf. Robot. Autom.*, May 2024, pp. 4480–4487.

[137] D. Köhler, M. Quach, M. Ulrich, F. Meinl, B. Bischoff, and H. Blume, "Improved multi-scale grid rendering of point clouds for radar object detection networks," in *Proc. 26th Int. Conf. Inf. Fusion*, 2023, pp. 1–8.

[138] J.-H. Kim, M.-C. Lee, and T.-S. Lee, "Deep-learning based multi-object detection and tracking using range-angle map in automotive radar systems," in *Proc. IEEE 95th Veh. Technol. Conf.*, 2022, pp. 1–6.

[139] C. Decourt, R. VanRullen, D. Salle, and T. Oberlin, "DAROD: A deep automotive radar object detector on range-doppler maps," in *Proc. IEEE Intell. Veh. Symp.*, 2022, pp. 112–118.

[140] A. Zhang, F. E. Nowruzi, and R. Laganiere, "RADDet: Range-Azimuth-Doppler based radar object detection for dynamic road users," in *Proc. 18th Conf. Robots Vis.*, 2021, pp. 95–102.

[141] D.-H. Paek, S.-H. Kong, and K. T. Wijaya, "K-Radar: 4D radar object detection for autonomous driving in various weather conditions," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 3819–3829.

[142] S.-h. Kim and Y. Hwang, "A survey on deep learning based methods and datasets for monocular 3D object detection," *Electronics*, vol. 10, no. 4, 2021, Art. no. 517.

[143] T. He and S. Soatto, "Mono3D++: Monocular 3D vehicle detection with two-scale 3D hypotheses and task priors," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 01, pp. 8409–8416.

[144] Z. Qin, J. Wang, and Y. Lu, "MonoGRNet: A geometric reasoning network for monocular 3D object localization," in *Proc. AAAI Conf. Artif. Intell.*, 2019, vol. 33, no. 01, pp. 8851–8858.

[145] W. Chen, Y. Li, Z. Tian, and F. Zhang, "2D and 3D object detection algorithms from images: A survey," *Array*, vol. 19, 2023, Art. no. 100305.

[146] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3D," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 194–210.

[147] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3D object detection in driving scenarios using infrastructure sensors," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 1852–1864, Mar. 2022.

[148] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3D point clouds," in *Proc. IEEE 39th Int. Conf. Distrib. Comput. Syst.*, 2019, pp. 514–524.

[149] E. Guo, Z. Chen, S. Rahardja, and J. Yang, "3D detection and pose estimation of vehicle in cooperative vehicle infrastructure system," *IEEE Sensors J.*, vol. 21, no. 19, pp. 21759–21771, Oct. 2021.

[150] H. Yu et al., "Vehicle-infrastructure cooperative 3D object detection via feature flow prediction," 2023, *arXiv:2303.10552*.

[151] Q. Chen, X. Ma, S. Tang, J. Guo, Q. Yang, and S. Fu, "F-cooper: Feature based cooperative perception for autonomous vehicle edge computing system using 3D point clouds," in *Proc. 4th ACM/IEEE Symp. Edge Comput.*, 2019, pp. 88–100.

[152] Y. Zhang, B. Chen, J. Qin, F. Hu, and J. Hao, "Coopercept: Cooperative perception for 3D object detection of autonomous vehicles," *Drones*, vol. 8, no. 6, 2024, Art. no. 228.

[153] E. E. Marvasti, A. Raftari, A. E. Marvasti, Y. P. Fallah, R. Guo, and H. Lu, "Cooperative lidar object detection via feature sharing in deep networks," in *Proc. IEEE 92nd Veh. Technol. Conf.*, 2020, pp. 1–7.

[154] S. Shi et al., "Vips: Real-time perception fusion for infrastructure-assisted autonomous driving," in *Proc. 28th Annu. Int. Conf. Mobile Comput. Netw.*, 2022, pp. 133–146.

[155] Z. Zheng, X. Xia, L. Gao, H. Xiang, and J. Ma, "Cooperfuse: A real-time cooperative perception fusion framework," in *Proc. IEEE Intell. Veh. Symp.*, 2024, pp. 533–538.

[156] M. Gulzar, Y. Muhammad, and N. Muhammad, "A survey on motion prediction of pedestrians and vehicles for autonomous driving," *IEEE Access*, vol. 9, pp. 137957–137969, 2021.

[157] P. Karle, M. Geisslinger, J. Betz, and M. Lienkamp, "Scenario understanding and motion prediction for autonomous vehicles—Review and comparison," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 10, pp. 16962–16982, Oct. 2022.

[158] R. Huang, G. Zhuo, L. Xiong, S. Lu, and W. Tian, "A review of deep learning-based vehicle motion prediction for autonomous driving," *Sustainability*, vol. 15, no. 20, 2023, Art. no. 14716.

[159] Z. Meng, X. Xia, R. Xu, W. Liu, and J. Ma, "HYDRO-3D: Hybrid object detection and tracking for cooperative perception using 3D lidar," *IEEE Trans. Intell. Veh.*, vol. 8, no. 8, pp. 4069–4080, Aug. 2023.

[160] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "PointPillars: Fast encoders for object detection from point clouds," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 12697–12705.

[161] H. W. Kuhn, "The Hungarian method for the assignment problem," *Nav. Res. Logistics Quart.*, vol. 2, no. 1/2, pp. 83–97, 1955.

[162] H. Su, S. Arakawa, and M. Murata, "Cooperative 3D multi-object tracking for connected and automated vehicles with complementary data association," in *Proc. IEEE Intell. Veh. Symp.*, 2024, pp. 285–291.

[163] P. U. Lima et al., "Formation control driven by cooperative object tracking," *Robot. Auton. Syst.*, vol. 63, pp. 68–79, 2015.

[164] I. Ulku and E. Akagündüz, "A survey on deep learning-based architectures for semantic segmentation on 2D images," *Appl. Artif. Intell.*, vol. 36, no. 1, 2022, Art. no. 2032924.

[165] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image segmentation using deep learning: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3523–3542, Jul. 2022.

[166] H. Thisanke, C. Deshan, K. Chamith, S. Seneviratne, R. Vidanaarachchi, and D. Herath, "Semantic segmentation using vision transformers: A survey," *Eng. Appl. Artif. Intell.*, vol. 126, 2023, Art. no. 106669.

[167] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 10012–10022.

[168] L. Peng, Z. Chen, Z. Fu, P. Liang, and E. Cheng, "BEVSegFormer: Bird's eye view semantic segmentation from arbitrary camera rigs," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2023, pp. 5935–5943.

[169] E. Xie et al., "M^2BEV: Multi-camera joint 3D detection and segmentation with unified birds-eye view representation," 2022, *arXiv:2204.05088*.

[170] Y. Wang, T. Shi, P. Yun, L. Tai, and M. Liu, "PointSeg: Real-time semantic segmentation based on 3D LiDAR point cloud," 2018, *arXiv:1807.06288*.

[171] A. Milioto, I. Vizzo, J. Behley, and C. Stachniss, "RangeNet++: Fast and accurate lidar semantic segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2019, pp. 4213–4220.

[172] A. Ouaknine, A. Newson, P. Pérez, F. Tupin, and J. Rebut, "Multi-view radar semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15671–15680.

[173] R. Xu, Z. Tu, H. Xiang, W. Shao, B. Zhou, and J. Ma, "CoBEVT: Cooperative bird's eye view semantic segmentation with sparse transformers," 2022, *arXiv:2207.02202*.

[174] H. Liu, Z. Gu, C. Wang, P. Wang, and D. Vukobratovic, "A lidar semantic segmentation framework for the cooperative vehicle-infrastructure system," in *Proc. IEEE 98th Veh. Technol. Conf.*, 2023, pp. 1–5.

[175] H. Li et al., "Delving into the devils of bird's-eye-view perception: A review, evaluation and recipe," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 4, pp. 2151–2170, Apr. 2024.

[176] Y. Ma et al., "Vision-centric BEV perception: A survey," 2022, *arXiv:2208.02797*.

[177] J. Huang, G. Huang, Z. Zhu, Y. Ye, and D. Du, "BEVDet: High-performance multi-camera 3D object detection in bird-eye-view," 2021, *arXiv:2112.11790*.

[178] J. Huang and G. Huang, "BEVDet4D: Exploit temporal cues in multi-camera 3D object detection," 2022, *arXiv:2203.17054*.

[179] Y. Li et al., "Towards efficient 3D object detection in bird's-eye-space for autonomous driving: A convolutional-only approach," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst.*, 2023, pp. 2170–2177.

[180] M. H. Ng, K. Radia, J. Chen, D. Wang, I. Gog, and J. E. Gonzalez, "BEV-Seg: Bird's eye view semantic segmentation using geometry and semantic point cloud," 2020, *arXiv:2006.11436*.

[181] A. Hu et al., "Fiery: Future instance prediction in bird's-eye view from surround monocular cameras," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2021, pp. 15273–15282.

[182] Y. Yang, Y. Deng, J. Zhang, J. Nie, and Z.-J. Zha, "BEVTrack: A simple and strong baseline for 3D single object tracking in bird's-eye view," 2023, *arXiv:2309.02185*.

[183] Z. Liu et al., "BEVFusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 2774–2781.

[184] A. Laddha, S. Gautam, S. Palombo, S. Pandey, and C. Vallespi-Gonzalez, "MVFuseNet: Improving end-to-end object detection and motion forecasting through multi-view fusion of lidar data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2865–2874.

[185] A. El Amin et al., "Monocular VO scale ambiguity resolution using an ultra low-cost spike rangefinder," *Positioning*, vol. 11, no. 04, 2020, Art. no. 45.

[186] Z. Liu, Z. Wu, and R. Tóth, "Smoke: Single-stage monocular 3D object detection via keypoint estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 996–997.

[187] T. Collins and J. Collins, "Occupancy grid mapping: An empirical evaluation," in *Proc. Mediterranean Conf. Control Automat.*, 2007, pp. 1–6.

[188] S. Mentasti and M. Matteucci, "Multi-layer occupancy grid mapping for autonomous vehicles navigation," in *Proc. AEIT Int. Conf. Elect. Electron. Technol. Automot.*, 2019, pp. 1–6.

[189] N. Defauw, M. Malfante, O. Antoni, T. Rakotovao, and S. Lesecq, "Vehicle detection on occupancy grid maps: Comparison of five detectors regarding real-time performance," *Sensors*, vol. 23, no. 3, 2023, Art. no. 1613.

[190] C. Lu, M. J. G. Van De Molengraft, and G. Dubbelman, "Monocular semantic occupancy grid mapping with convolutional variational encoder–decoder networks," *IEEE Robot. Automat. Lett.*, vol. 4, no. 2, pp. 445–452, Apr. 2019.

[191] H. Li, M. Tsukada, F. Nashashibi, and M. Parent, "Multivehicle cooperative local mapping: A methodology based on occupancy grid map merging," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2089–2100, Oct. 2014.

[192] A. Caillot, S. Ouerghi, P. Vasseur, Y. Dupuis, and R. Boutteau, "Multi-agent cooperative camera-based evidential occupancy grid generation," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst.*, 2022, pp. 203–209.

[193] H. Zhou, Z. Ge, W. Mao, and Z. Li, "PersDet: Monocular 3D detection in perspective bird's-eye-view," 2022, *arXiv:2208.09394*.

[194] Z. Zhu et al., "Understanding the robustness of 3D object detection with bird's-eye-view representations in autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 21600–21610.

[195] P. Li, S. Ding, X. Chen, N. Hanselmann, M. Cordts, and J. Gall, "PowerBEV: A powerful yet lightweight framework for instance prediction in bird's-eye view," 2023, *arXiv:2306.10761*.

[196] H. Rashed, M. Essam, M. Mohamed, A. E. Sallab, and S. Yogamani, "BEV-ModNet: Monocular camera based bird's eye view moving object detection for autonomous driving," in *Proc. IEEE Int. Intell. Transp. Syst. Conf.*, 2021, pp. 1503–1508.

[197] W. Yang et al., "Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15536–15545.

[198] S. Klaudt, A. Zlocki, and L. Eckstein, "A-priori map information and path planning for automated valet-parking," in *Proc. IEEE Intell. Veh. Symp.*, 2017, pp. 1770–1775.

[199] Q. Li, Y. Wang, Y. Wang, and H. Zhao, "HDMapNet: An online HD map construction and evaluation framework," in *Proc. Int. Conf. Robot. Automat.*, 2022, pp. 4628–4634.

[200] J. Ziegler et al., "Making bertha drive—An autonomous journey on a historic route," *IEEE Intell. Transp. Syst. Mag.*, vol. 6, no. 2, pp. 8–20, Summer 2014.

[201] A. Boubakri, S. M. Gammar, M. B. Brahim, and F. Filali, "High definition map update for autonomous and connected vehicles: A survey," in *Proc. Int. Wireless Commun. Mobile Comput.*, 2022, pp. 1148–1153.

[202] G. Elghazaly, R. Frank, S. Harvey, and S. Safko, "High-definition maps: Comprehensive survey, challenges and future perspectives," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 527–550, 2023.

[203] Z. Bao, S. Hossain, H. Lang, and X. Lin, "A review of high-definition map creation methods for autonomous driving," *Eng. Appl. Artif. Intell.*, vol. 122, 2023, Art. no. 106125.

[204] J. Aulinas, Y. Petillot, J. Salvi, and X. Lladó, "The SLAM problem: A survey," in *Proc. Int. Conf. Catalan Assoc. Artif. Intell.*, 2008, pp. 363–371.

[205] I. A. Kazerouni, L. Fitzgerald, G. Dooly, and D. Toal, "A survey of state-of-the-art on visual slam," *Expert Syst. Appl.*, vol. 205, 2022, Art. no. 117734.

[206] J. A. Placed et al., "A survey on active simultaneous localization and mapping: State of the art and new frontiers," *IEEE Trans. Robot.*, vol. 39, no. 3, pp. 1686–1705, Jun. 2023.

[207] S. Mokssit, D. B. Licea, B. Guermah, and M. Ghogho, "Deep learning techniques for visual SLAM: A survey," *IEEE Access*, vol. 11, pp. 20026–20050, 2023.

[208] *Road Vehicles–Vehicle Dynamics and Road-Holding Ability – Vocabulary*, Standard ISO 8855:2011, International Organization for Standardization, Geneva, Switzerland, 2011.

[209] C. Allig and G. Wanielik, "Alignment of perception information for cooperative perception," in *Proc. IEEE Intell. Veh. Symp.*, 2019, pp. 1849–1854.

[210] S. Thrun, "A probabilistic on-line mapping algorithm for teams of mobile robots," *Int. J. Robot. Res.*, vol. 20, no. 5, pp. 335–363, 2001.

[211] F. Seeliger et al., "Advisory warnings based on cooperative perception," in *Proc. IEEE Intell. Veh. Symp.*, 2014, pp. 246–252.

[212] F. Seeliger and K. Dietmayer, "Inter-vehicle information-fusion with shared perception information," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst.*, 2014, pp. 2087–2093.

[213] A. Rauch, S. Maier, F. Klanner, and K. Dietmayer, "Inter-vehicle object association for cooperative perception systems," in *Proc. 16th Int. IEEE Conf. Intell. Transp. Syst.*, 2013, pp. 893–898.

[214] M. Ballesta, A. Gil, O. Reinoso, M. Juliá, and L. M. Jiménez, "Multi-robot map alignment in visual SLAM," *WSEAS Trans. Syst.*, vol. 9, no. 2, pp. 213–222, 2010.

[215] L. Song, W. Valentine, Q. Yang, H. Wang, H. Fang, and Y. Liu, "BB-Align: A lightweight pose recovery framework for vehicle-to-vehicle cooperative perception," in *Proc. IEEE 44th Int. Conf. Distrib. Comput. Syst.*, 2024, pp. 1016–1026.

[216] K. Jiang et al., "Map container: A map-based framework for cooperative perception," 2022, *arXiv:2208.13226*.

[217] Q. Qu, Y. Xiong, X. Wu, H. Li, and S. Guo, "V2I-Calib: A novel calibration approach for collaborative vehicle and infrastructure lidar systems," 2024, *arXiv:2407.10195*.

[218] Z. Zhao, Y. Li, Y. Chen, X. Zhang, and R. Tian, "A spatial alignment framework using geolocation cues for roadside multi-view multi-sensor fusion," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst.*, 2023, pp. 3633–3640.

[219] N. Vadivelu, M. Ren, J. Tu, J. Wang, and R. Urtasun, "Learning to communicate and correct pose errors," in *Proc. Conf. Robot Learn.*, 2021, pp. 1195–1210.

[220] T.-H. Wang, S. Manivasagam, M. Liang, B. Yang, W. Zeng, and R. Urtasun, "V2VNet: Vehicle-to-vehicle communication for joint perception and prediction," in *Proc. 16th Eur. Conf. Comput. Vis.*, Glasgow, U.K., Aug. 2020, pp. 605–621.

[221] Z. Lei et al., "Robust collaborative perception without external localization and clock devices," in *Proc. 2024 IEEE Int. Conf. Robot. Autom.*, Yokohama, Japan, 2024, pp. 7280–7286.

[222] H. Xiang, R. Xu, and J. Ma, "HM-ViT: Hetero-modal vehicle-to-vehicle cooperative perception with vision transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 284–295.

[223] K. Yang et al., "Spatio-temporal domain awareness for multi-agent collaborative perception," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 23383–23392.

[224] N. Glaser, Y.-C. Liu, J. Tian, and Z. Kira, "Overcoming obstructions via bandwidth-limited multi-agent spatial handshaking," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2021, pp. 2406–2413.

[225] S. Shi, C. Zhang, A. Lv, and S. He, "MCOT: Multi-modal vehicle-to-vehicle cooperative perception with transformers," in *Proc. IEEE 29th Int. Conf. Parallel Distrib. Syst.*, 2023, pp. 1612–1619.

[226] D. Yang et al., "How2comm: Communication-efficient and collaboration-pragmatic multi-agent perception," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2024, vol. 36, pp. 25151–25164.

[227] A. Rauch, F. Klanner, R. Rasshofer, and K. Dietmayer, "Car2x-based perception in a high-level fusion architecture for cooperative perception systems," in *Proc. IEEE Intell. Veh. Symp.*, 2012, pp. 270–275.

[228] A. Fischer, A. F. Izmailov, and M. V. Solodov, "The Levenberg–Marquardt method: An overview of modern convergence theories and more," *Comput. Optim. Appl.*, vol. 89, pp. 33–67, 2024.

[229] Z. Zhou et al., "V2XPNP: Vehicle-to-everything spatio-temporal fusion for multi-agent perception and prediction," 2024, *arXiv:2412.01812*.

[230] V. P. Chellapandi, L. Yuan, C. G. Brinton, S. H. Żak, and Z. Wang, "Federated learning for connected and automated vehicles: A survey of existing approaches and challenges," *IEEE Trans. Intell. Veh.*, vol. 9, no. 1, pp. 119–137, Jan. 2024.

[231] W. Zhou et al., "Long-tail prediction uncertainty aware trajectory planning for self-driving vehicles," in *Proc. IEEE 25th Int. Conf. Intell. Transp. Syst.*, 2022, pp. 1275–1282.

[232] L. Barbieri, S. Savazzi, M. Brambilla, and M. Nicoli, "Decentralized federated learning for extended sensing in 6G connected vehicles," *Veh. Commun.*, vol. 33, 2022, Art. no. 100396.

[233] J. Posner, L. Tseng, M. Aloqaily, and Y. Jararweh, "Federated learning in vehicular networks: Opportunities and solutions," *IEEE Netw.*, vol. 35, no. 2, pp. 152–159, Mar./Apr. 2021.

[234] R. Roriz, H. Silva, F. Dias, and T. Gomes, "A survey on data compression techniques for automotive lidar point clouds," *Sensors*, vol. 24, no. 10, 2024, Art. no. 3185.

[235] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer," in *Proc. Eur. Conf. Comput. Vis.*, 2022, pp. 107–124.

[236] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, vol. 34, pp. 29541–29552.

[237] H. Yu et al., "Dair-V2X: A large-scale dataset for vehicle-infrastructure cooperative 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21361–21370.

[238] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, vol. 35, pp. 4874–4886.

[239] L. Gao, H. Xiang, X. Xia, and J. Ma, "Multi-sensor fusion for vehicle-to-vehicle cooperative localization with object detection and point cloud matching," *IEEE Sensors J.*, vol. 24, no. 7, pp. 10865–10877, Apr. 2024.

[240] L. Gao, H. Xiang, X. Xia, and J. Ma, "End-to-end cooperative localization via neural feature sharing," in *Proc. IEEE Intell. Veh. Symp.*, 2024, pp. 553–558.

[241] X. Xia, R. Xu, and J. Ma, "Secure cooperative localization for connected automated vehicles based on consensus," *IEEE Sensors J.*, vol. 23, no. 20, pp. 25061–25074, Oct. 2023.

[242] W. Choi, M. Shin, H. Lee, J. Cho, J. Park, and S. Im, "Multi-task learning for real-time autonomous driving leveraging task-adaptive attention generator," in *Proc. 2024 IEEE Int. Conf. Robot. Autom.*, Yokohama, Japan, 2024, pp. 14732–14739.

[243] J. Betz et al., "Autonomous vehicles on the edge: A survey on autonomous vehicle racing," *IEEE Open J. Intell. Transp. Syst.*, vol. 3, pp. 458–488, 2022.

[244] J. Clancy et al., "Investigating the effect of handover on latency in early 5G NR deployments for C-V2X network planning," *IEEE Access*, vol. 11, pp. 129124–129143, 2023.

[245] T. Mulder and N. E. Vellinga, "Exploring data protection challenges of automated driving," *Comput. Law Secur. Rev.*, vol. 40, 2021, Art. no. 105530.

[246] China Business Law Journal, "Driving data management," Accessed: Dec. 30, 2024. [Online]. Available: https://law.asia/china-automotive-industry-data-compliance/

[247] Y. Li, X. Tao, X. Zhang, J. Liu, and J. Xu, "Privacy-preserved federated learning for autonomous driving," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 8423–8434, Jul. 2021.

[248] P. A. Tonga, Z. S. Ameen, A. S. Mubarak, and F. Al-Turjman, "A review on on device privacy and machine learning training," in *Proc. Int. Conf. Artif. Intell. Everything*, 2022, pp. 679–684.

[249] Z. Zhang et al., "On the federated learning framework for cooperative perception," *IEEE Robot. Autom. Lett.*, vol. 9, no. 11, pp. 9423–9430, Nov. 2024.

[250] X. Yin, Y. Zhu, and J. Hu, "A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions," *ACM Comput. Surv.*, vol. 54, no. 6, pp. 1–36, 2021.

[251] A. El Ouadrhiri and A. Abdelhadi, "Differential privacy for deep and federated learning: A survey," *IEEE Access*, vol. 10, pp. 22359–22380, 2022.

[252] S. Truex, L. Liu, K.-H. Chow, M. E. Gursoy, and W. Wei, "LDP-fed: Federated learning with local differential privacy," in *Proc. 3rd ACM Int. Workshop Edge Syst., Anal. Netw.*, 2020, pp. 61–66.

[253] I. Zhou, F. Tofigh, M. Piccardi, M. Abolhasan, D. Franklin, and J. Lipman, "Secure multi-party computation for machine learning: A survey," *IEEE Access*, vol. 12, pp. 53881–53899, 2024.

[254] B. Khalfoun, S. Ben Mokhtar, S. Bouchenak, and V. Nitu, "Eden: Enforcing location privacy through re-identification risk assessment: A federated learning approach," *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.*, vol. 5, no. 2, pp. 1–25, 2021.

[255] Y. Chen, F. Luo, T. Li, T. Xiang, Z. Liu, and J. Li, "A training-integrity privacy-preserving federated learning scheme with trusted execution environment," *Inf. Sci.*, vol. 522, pp. 69–79, 2020.

[256] F. Mo, H. Haddadi, K. Katevas, E. Marin, D. Perino, and N. Kourtellis, "PPFL: Privacy-preserving federated learning with trusted execution environments," in *Proc. 19th Annu. Int. Conf. Mobile Syst., Appl., Serv.*, 2021, pp. 94–108.

[257] L. Lyu, H. Yu, and Q. Yang, "Threats to federated learning: A survey," 2020, *arXiv:2003.02133*.

[258] V. Mothukuri, R. M. Parizi, S. Pouriyeh, Y. Huang, A. Dehghantanha, and G. Srivastava, "A survey on security and privacy of federated learning," *Future Gener. Comput. Syst.*, vol. 115, pp. 619–640, 2021.

[259] V. Tolpegin, S. Truex, M. E. Gursoy, and L. Liu, "Data poisoning attacks against federated learning systems," in *Proc. 25th Eur. Symp. Res. Comput. Secur.*, Guildford, U.K., Sep. 2020, pp. 480–501.

[260] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in *Proc. IEEE Symp. Secur. Privacy*, 2019, pp. 739–753.
[261] X. Gong, Y. Chen, Q. Wang, and W. Kong, "Backdoor attacks and defenses in federated learning: State-of-the-art, taxonomy, and future directions," *IEEE Wireless Commun.*, vol. 30, no. 2, pp. 114–121, Apr. 2023.
[262] J. Zhang, B. Chen, X. Cheng, H. T. T. Binh, and S. Yu, "PoisonGAN: Generative poisoning attacks against federated learning in edge computing systems," *IEEE Internet Things J.*, vol. 8, no. 5, pp. 3310–3322, Mar. 2021.
[263] D. Cao, S. Chang, Z. Lin, G. Liu, and D. Sun, "Understanding distributed poisoning attack in federated learning," in *Proc. IEEE 25th Int. Conf. Parallel Distrib. Syst.*, 2019, pp. 233–239.
[264] E. T. Martínez Beltrán et al., "Mitigating communications threats in decentralized federated learning through moving target defense," *Wireless Netw.*, vol. 30, pp. 7407–7421, 2024.
[265] K. Kim, J. S. Kim, S. Jeong, J.-H. Park, and H. K. Kim, "Cybersecurity for autonomous vehicles: Review of attacks and defense," *Comput. Secur.*, vol. 103, 2021, Art. no. 102150.
[266] M. Girdhar, J. Hong, and J. Moore, "Cybersecurity of autonomous vehicles: A systematic literature review of adversarial attacks and defense models," *IEEE Open J. Veh. Technol.*, vol. 4, pp. 417–437, 2023.
[267] Y. Cao, S. H. Bhupathiraju, P. Naghavi, T. Sugawara, Z. M. Mao, and S. Rampazzi, "You can't see me: Physical removal attacks on {lidar-based} autonomous vehicles driving frameworks," in *Proc. 32nd USENIX Secur. Symp.*, 2023, pp. 2993–3010.
[268] A. Chahe, C. Wang, A. Jeyapratap, K. Xu, and L. Zhou, "Dynamic adversarial attacks on autonomous driving systems," 2023, *arXiv:2312.06701*.
[269] T. Rosenstatter and C. Englund, "Modelling the level of trust in a cooperative automated vehicle control system," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 4, pp. 1237–1247, Apr. 2018.
[270] C. Allig, T. Leinmüller, P. Mittal, and G. Wanielik, "Trustworthiness estimation of entities within collective perception," in *Proc. IEEE Veh. Netw. Conf.*, 2019, pp. 1–8.
[271] Transportation Research Board and National Academies of Sciences, Engineering, and Medicine, *Business Models to Facilitate Deployment of Connected Vehicle Infrastructure to Support Automated Vehicle Operations*, WSP USA, Washington, DC, USA: The National Academies Press, 2020. [Online]. Available: https://nap.nationalacademies.org/catalog/25946/business-models-to-facilitate-deployment-of-connected-vehicle-infrastructure-to-support-automated-vehicle-operations
[272] 5G Automotive Association, "Cost analysis of V2I deployment," Accessed: Sep. 20, 2024. [Online]. Available: https://5gaa.org/cost-analysis-of-v2i-deployment/
[273] T. Le and K. Eccles, "Safety-based deployment assistance for location of V2I applications pilot: Red-light-violation warning application," Federal Highway Admin., Office Safety Res. Develop., Washington, DC, USA, HSIS Task Rep., Jun. 2016. [Online]. Available: https://www.transportation.gov/research-and-technology/safety-based-deployment-assistance-location-v2i-applications-pilot-red
[274] ITS Deployment Evaluation, "Indiana dot's 26 month pilot study shows that deployment of queue warning trucks ahead of interstate work zones to alert motorists of queued traffic reduced hard braking events by 80 percent," Accessed: Dec. 30, 2024. [Online]. Available: https://www.itskrs.its.dot.gov/2023-b01745
[275] ITS Deployment Evaluation, "Implementation of connected vehicle technologies in connecticut revealed benefits with a 90 percent reduction in collision risk due to a cloud-based digital alert system protecting roadside workers," Accessed: Dec. 30, 2024. [Online]. Available: https://www.itskrs.its.dot.gov/2023-b01745
[276] ITS Deployment Evaluation, "V2x technology," Accessed: Dec. 30, 2024. [Online]. Available: https://www.itskrs.its.dot.gov/briefings/executive-briefing/vehicle-everything-v2x-technology
[277] C. Dong, H. Wang, Y. Li, Y. Liu, and Q. Chen, "Economic comparison between vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) at freeway on-ramps based on microscopic simulations," *IET Intell. Transport Syst.*, vol. 13, no. 11, pp. 1726–1735, 2019.
[278] H. Xiang et al., "V2X-real: A largs-scale dataset for vehicle-to-everything cooperative perception," in *Proc. Eur. Conf. Comput. Vis.*, 2025, pp. 455–470.
[279] W. Zimmer, G. A. Wardana, S. Sritharan, X. Zhou, R. Song, and A. C. Knoll, "Tumtraf V2X cooperative perception dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 22668–22677.
[280] H. Xiang, R. Xu, X. Xia, Z. Zheng, B. Zhou, and J. Ma, "V2XP-ASG: Generating adversarial scenes for vehicle-to-everything perception," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 3584–3591.
[281] M. Sahu, A. Mukhopadhyay, and S. Zachow, "Simulation-to-real domain adaptation with teacher–student learning for endoscopic instrument segmentation," *Int. J. Comput. Assist. Radiol. Surg.*, vol. 16, no. 5, pp. 849–859, 2021.
[282] W. M. Kouw and M. Loog, "A review of domain adaptation without target labels," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 3, pp. 766–785, Mar. 2021.
[283] J. Li et al., "Domain adaptation based object detection for autonomous driving in foggy and rainy weather," *IEEE Trans. Intell. Veh.*, early access, Jun. 27, 2024, doi: 10.1109/TIV.2024.3419689.
[284] J. Li et al., "S2R-ViT for multi-agent cooperative perception: Bridging the gap from simulation to reality," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2024, pp. 16374–16380.
[285] R. Xu, J. Li, X. Dong, H. Yu, and J. Ma, "Bridging the domain gap for multi-agent perception," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2023, pp. 6035–6042.
[286] J. Li, B. Li, X. Liu, R. Xu, J. Ma, and H. Yu, "Breaking data silos: Cross-domain learning for multi-agent perception from independent private sources," in *Proc. 2024 IEEE Int. Conf. Robot. Autom.*, Yokohama, Japan, 2024, pp. 18414–18420.
[287] N. Asselin-Miller et al., "Study on the deployment of C-its in Europe: Final report," *Rep. DG MOVE MOVE/C*, vol. 3, pp. 2014–794, 2016.
[288] A. Figueiredo, P. Rito, M. Luís, and S. Sargento, "Mobility sensing and V2X communication for emergency services," *Mobile Netw. Appl.*, vol. 28, no. 3, pp. 1126–1141, 2023.
[289] C. Ding, I. W.-H. Ho, E. Chung, and T. Fan, "V2X and deep reinforcement learning-aided mobility-aware lane changing for emergency vehicle preemption in connected autonomous transport systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 25, no. 7, pp. 7281–7293, Jul. 2024.
[290] A. Ahmed, M. M. Iqbal, S. Jabbar, M. Ibrar, A. Erbad, and H. Song, "Position-based emergency message dissemination schemes in the internet of vehicles: A review," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 13548–13572, Dec. 2023.
[291] M. U. Ghazi, M. A. K. Khattak, B. Shabir, A. W. Malik, and M. S. Ramzan, "Emergency message dissemination in vehicular networks: A review," *IEEE Access*, vol. 8, pp. 38606–38621, 2020.
[292] S.-Y. Ni, Y.-C. Tseng, Y.-S. Chen, and J.-P. Sheu, "The broadcast storm problem in a mobile ad hoc network," in *Proc. 5th Annu. ACM/IEEE Int. Conf. Mobile Comput. Netw.*, 1999, pp. 151–162.
[293] L. B. Jiang and S. C. Liew, "Hidden-Node removal and its application in cellular WiFi networks," *IEEE Trans. Veh. Technol.*, vol. 56, no. 5, pp. 2641–2654, Sep. 2007.

ROSHAN GEORGE received the B.Eng. (Hons.) degree in 2021 from the University of Galway, Galway, Ireland, where he is currently working toward the Ph.D. degree. He is also a Member of the Connaught Automotive Research (CAR) Group under the supervision of Prof. Martin Glavin and Prof. Edward Jones. His research interests include V2I cooperative perception and sensor fusion for intelligent transportation systems (ITS).

JOSEPH CLANCY received the B.Eng. (Hons.) degree and the Ph.D. degree in electronic engineering from the University of Galway, Galway, Ireland, in 2019 and 2024, respectively. His Ph.D. research topic involved the investigation of the feasibility of vehicular communications with modern wireless access technologies. He is currently a Research Associate with the School of Engineering, University of Galway.

TIM BROPHY received the B.Eng. (Hons.) degree in 2018 from the University of Galway, Galway, Ireland, where he is currently working toward the Ph.D. degree. He is a Member of the Connaught Automotive Research (CAR) Group under the supervision of Prof. Edward Jones and Prof. Martin Glavin. His research interests include computer vision and sensor availability within an autonomous vehicle context.

GANESH SISTU is currently a Principal AI Architect with Valeo Ireland and Adjunct Assistant Professor with the University of Limerick, Limerick, Ireland, leads innovations in automated driving and parking. With more than 13 years in computer vision and machine learning, he has contributed to more than 30 publications in top-tier conferences like ICCV and ICRA. He also plays a pivotal role in guiding the future of AI education as an industry board member for Science Foundation Ireland's Data Science PhD and the National MSc in AI programs with the University of Limerick, blending academic insight with industry expertise.

WILLIAM O'GRADY received the B.Eng. (Hons.) degree from the University of Galway, Galway, Ireland, in 2012. He is currently a Lead Computer Vision Product Owner and Senior Expert with Valeo Ireland for automated parking and driving solutions and validation tooling and methodology. He has more than 11 years of experience in computer vision and machine learning applications, validation, and camera extrinsic calibration in series production automotive systems.

SUNIL CHANDRA received the MPhil degree in mathematics from the Indian Institute of Technology, Roorkee, India, in 1997, and the Ph.D. degree in computer vision from the University of Surrey, Guildford, U.K., in 2006. From 2007 to 2012, he was an Imaging Researcher in Face and Gesture Recognition technologies. Since 2013, he has been with Valeo Vision Systems, where he is currently a Valeo Expert and Lead Algorithm Developer. He has more than 16 years of experience in computer vision and machine learning applications, including 3D reconstruction, object detection/tracking, and camera calibration.

FIACHRA COLLINS received the B.E. degree in mechanical engineering from University College Dublin, Dublin, Ireland, in 2007, and the Ph.D. degree from Dublin City University, Dublin, in 2011. From 2014 to 2018, he was Chief Technology Officer for a start-up company specializing in IoT and sensors, which was a spin-out of his postdoctoral research from 2011 to 2014. Since 2019, he has been the Geometric Perception Team Manager with Valeo Vision Systems, overseeing the software development for camera calibration and perception algorithms (3D reconstruction, localization, mapping). His research interests include computer vision and data analytics.

DARRAGH MULLINS received the B.E. degree in energy systems engineering and the Ph.D. degree in electronic engineering from the University of Galway, Galway, Ireland, in 2013 and 2018, respectively. His Ph.D. research topic involved the application of imaging sensors and signal processing to wastewater treatment plant performance sensing. He was a Postdoctoral Research Fellow with the University of Galway, from 2018 to 2022, where he managed a research program and co-supervised six Ph.D. student projects involving sensors and V2X communication systems for pedestrian and vehicle monitoring from both vehicle and fixed infrastructure point-of-view. He is currently a Senior Technical Officer and Adjunct Lecturer with the School of Engineering, University of Galway. He was awarded Lero Director's Prize for Education and Public Engagement, in 2020.

EDWARD JONES (Senior Member, IEEE) received the B.E. and Ph.D. degrees in electronic engineering from the University of Galway, Galway, Ireland. His Ph.D. research topic was on the development of computational auditory models for speech processing. From 2009 to 2010, he was a Visiting Researcher with the Department of Electrical Engineering, Columbia University, New York, NY, USA, and has also been appointed a Visiting Fellow with the School of Electrical Engineering and Telecommunications, University of New South Wales, Sydney, NSW, Australia. He is currently a Professor of electrical & electronic engineering with the School of Engineering, University of Galway. He also has a number of years of industrial experience in senior positions, in both start-ups and multinational companies, including Toucan Technology Ltd., PMC-Sierra Inc., Innovada Ltd., and Duolog Technologies Ltd. He also represented Toucan Technology and PMC-Sierra on international standardization groups ANSI T1E1.4 and ETSI TM6. His research interests include DSP algorithm development and embedded implementation for applications in biomedical engineering, speech and audio processing, and image processing. He is a Chartered Engineer and Fellow of the Institution of Engineers of Ireland.

BRIAN DEEGAN (Member, IEEE) received the bachelor's degree in computer engineering and the M.Sc. degree in biomedical engineering from the University of Limerick, Limerick, Ireland, in 2004 and 2005, respectively, and the Ph.D. degree in biomedical engineering from the University of Galway, Galway, Ireland, in 2011. The focus of his research was the relationship between blood pressure and cerebral blood flow in humans. From 2011 to 2022, he was with Valeo Vision Systems as a Vision Research Engineer focusing on Image Quality. In 2022, he joined the Department of Electrical & Electronic Engineering, University of Galway, as a Lecturer and Researcher. His research focuses on high dynamic range imaging, LED flicker, Topview harmonization algorithms, and the relationship between image quality and machine vision.

MARTIN GLAVIN (Member, IEEE) received the B.E. degree in electronic engineering and the Ph.D. degree in the area of algorithms and architectures for high-speed data communications systems from the University of Galway, Galway, Ireland, in 1997 and 2004, respectively, and the Higher Diploma in Third Level Education in 2007. He was a Lecturer from 1999 to 2003 and became a permanent Member of the academic staff in 2004. He is currently the Joint Director of the Connaught Automotive Research (CAR) Group, University of Galway. He is a Lero Funded Investigator, supervising researchers with industry on signal processing and embedded systems for automotive and agricultural applications.