A Survey and Framework of Cooperative Perception: From Heterogeneous Singleton to Hierarchical Cooperation

Zhengwei Bai, *Student Member, IEEE*, Guoyuan Wu, *Senior Member, IEEE*, Matthew J. Barth, *Fellow, IEEE*, Yongkang Liu, Emrah Akin Sisbot, *Member, IEEE*, Kentaro Oguchi, and Zhitong Huang

***Abstract*— Perceiving the environment is one of the most fundamental keys to enabling Cooperative Driving Automation, which is regarded as the revolutionary solution to addressing the safety, mobility, and sustainability issues of contemporary transportation systems. Although an unprecedented evolution is now happening in the area of computer vision for object perception, state-of-the-art perception methods are still struggling with sophisticated real-world traffic environments due to the inevitable physical occlusion and limited receptive field of single-vehicle systems. Based on multiple spatially separated perception nodes, Cooperative Perception (CP) is born to unlock the bottleneck of perception for driving automation. In this paper, we comprehensively review and analyze the research progress on CP, and we propose a unified CP framework. The architectures and taxonomy of CP systems based on different types of sensors are reviewed to show a high-level description of the workflow and different structures for CP systems. The node structure, sensing modality, and fusion schemes are reviewed and analyzed with detailed explanations for CP. A Hierarchical Cooperative Perception (HCP) framework is proposed, followed by a review of existing open-source tools that support CP development. The discussion highlights the current opportunities, open challenges, and anticipated future trends.**

***Index Terms*— Survey, cooperative perception, object detection and tracking, cooperative driving automation, sensor fusion.**

I. Introduction

THE rapid progress of the transportation system has improved the efficiency of our daily people and goods movement. Nevertheless, the rapidly increasing number of vehicles has resulted in several major issues in the transportation system in terms of safety [1], mobility [2], and environmental sustainability [3]. Taking advantage of recent strides in advanced sensing, wireless connectivity, and artificial intelligence, Cooperative Driving Automation (CDA) enables Connected and Automated Vehicles (CAVs) to communicate with each other, with roadway infrastructure, or with other road users such as pedestrians and cyclists equipped with mobile devices, to improve the system-wide performance. Hence, CDA has attracted increasingly more attention over the past few years and is regarded as a transformative solution to the aforementioned challenges [4].

Object Perception (OP), acting as the "vision" function of automated agents by analogy, plays a fundamental role in the basic structure of CDA applications [5]. Different kinds of onboard or roadside sensors have different capabilities of perceiving traffic conditions in the real-world environment. The perception data can act as the system input and support various kinds of downstream CDA applications, such as Collision Warning [6], Eco-Approach and Departure (EAD) [7], and Cooperative Adaptive Cruise Control (CACC) [8].

With the development of sensing technologies, transportation systems can retrieve high-fidelity traffic data from different sensors. For instance, cameras can provide detailed vision data to classify various kinds of traffic objects, such as vehicles, pedestrians, and cyclists [9]. LiDAR can provide high-fidelity 3D point cloud data to grasp the precise 3D location of the traffic objects [10]. Additionally, RADARs are more robust to visibility problems compared to cameras and LiDAR (e.g., atmospheric obscurants from precipitation, smoke, dust, etc.) and thus have been a critical component for safety-critical applications in the automotive industry [11].

However, during the last couple of decades, a large portion of the OP methods and high-fidelity perception data have come from onboard sensors while most of the roadside sensors are still used for traditional traffic data collection such as counting traffic volumes based on loop detectors, cameras, or RADARs [12]. Although empowered with advanced perception methods, onboard sensors are inevitably limited by sensing range and occlusion. Infrastructure-based perception systems have the potential to achieve better OP results with fewer occlusion effects and more flexibility in terms of mounting height and pose. However, due to the fixture of installation, infrastructure-based sensors will suffer from limited receptive ranges. Thus, neither onboard sensors nor infrastructure-based sensors alone can unlock these limitations based on a single *Perception Node* (PN) which is defined as a

Manuscript received 3 September 2022; revised 3 July 2023, 29 November 2023, and 6 July 2024; accepted 12 July 2024. Date of publication 30 September 2024; date of current version 1 November 2024. This work was supported by Toyota Motor North America, InfoTech Laboratories. The Associate Editor for this article was S. S. Nedevschi. *(Corresponding author: Zhengwei Bai.)*

Zhengwei Bai, Guoyuan Wu, and Matthew J. Barth are with the Department of Electrical and Computer Engineering, University of California at Riverside, Riverside, CA 92507 USA (e-mail: zbai012@ucr.edu).

Yongkang Liu, Emrah Akin Sisbot, and Kentaro Oguchi are with Toyota Motor North America, InfoTech Laboratories, Mountain View, CA 94043 USA.

Zhitong Huang is with Leidos Inc., McLean, VA 22101 USA.

Digital Object Identifier 10.1109/TITS.2024.3436012

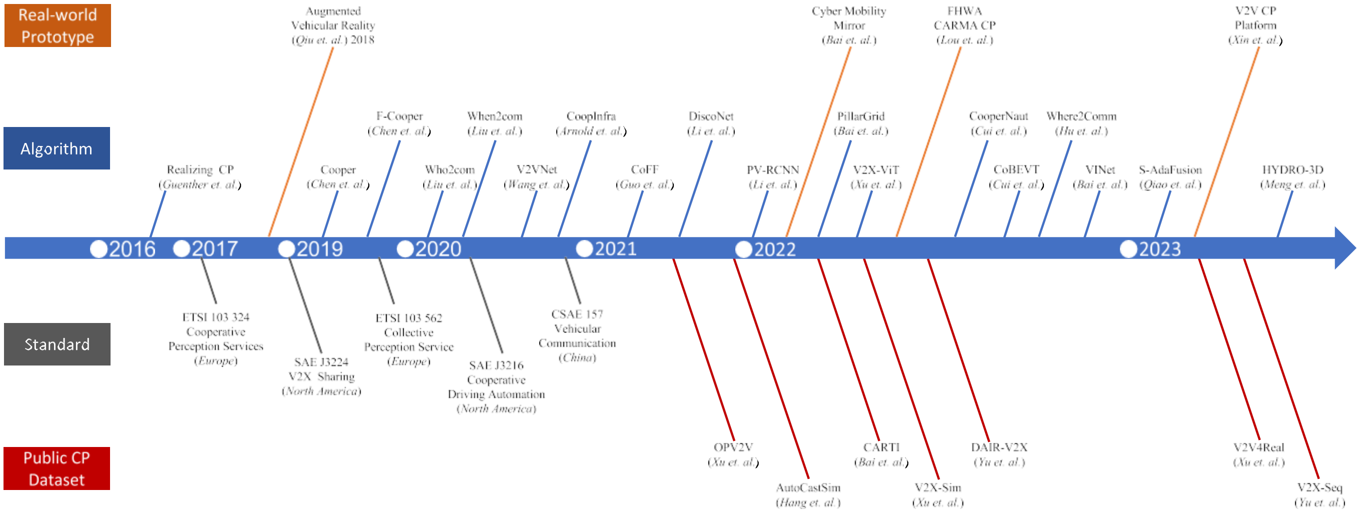


Fig. 1. The timeline diagram illustrating recent milestones in terms of different perspectives: 1) real-world prototype systems, 2) CP algorithms, 3) standards, and 4) public CP datasets.

singular entity equipped with perception and communication capabilities in this paper.

Empowered by mobile connectivity, Connected Vehicles (CVs) and CAVs can grasp perception information from others who are equipped with perception systems and connectivity, such as smart infrastructures or other CAVs. It is conceivable that combining perceptual information from spatially separated nodes is a natural way to overcome the aforementioned limitations, which is named Cooperative Perception (CP) or Collaborative Perception. As an emerging topic, CP attracts fast-increasing attention (as shown in Fig. 1). Research has been conducted from various aspects, including perception nodes (PNs), sensor modalities, and fusion schemes. Specifically, in terms of PNs, CP research includes vehicle-to-vehicle (V2V) CP [13] or vehicle-to-infrastructure (V2I) CP [14], [15], and vehicle-to-everything (V2X) CP [16]. For sensor modalities, CP research considers cameras [17], LiDAR [18], RADARs [19], etc. Additionally, different fusion schemes are investigated in CP which include early fusion [20], late fusion [21], or intermediate fusion [16]. Although a recent overview conducted by Caillot et al. [22] reviewed the cooperative perception in an automotive context, their focus is mainly on the sensing of the ego-vehicle using multiple sensors (e.g., vehicle localization, map generation, etc). Thus, a comprehensive survey of CP from the perspective of CDA is still missing. Meanwhile, different cooperative perception methods are typically associated with some specific transportation scenarios, which makes the implementation and integration of system-level design for cooperative perception in real-world conditions more challenging.

In this paper, CP technology is reviewed comprehensively, which aims to establish an overall landscape for this emerging area. Recent CP milestones are summarized in Fig. 1 to briefly illustrate the development of CP in terms of real- world prototype systems, algorithms, standards, and public datasets. CP methods are overviewed based on three primary aspects including 1) node structures, 2) sensor modalities, and 3) fusion schemes. Furthermore, a hierarchical CP framework is proposed to unify different scenarios in terms of the different perspectives mentioned above and to provide inspiration for future studies in this field to expedite the implementation of cooperative perception.

The rest of this paper is organized as follows: Architectures and taxonomy for CP systems are reviewed in Section II to lay the foundation. Major pillars for CP including node structure, sensing modality, and fusion scheme are reviewed in Section III to V, respectively. The hierarchical cooperative perception framework is proposed and discussed in Section VI, followed by the summarizing of open-source CP Datasets and Platforms in Section VII. Section VIII highlights the current states, open challenges and future trends, followed by Section IX that concludes the paper.

II. Architecture and Taxonomy

A. Standards

Due to the revolutionary impact that cooperative perception would have on the transportation industry, various standards related to CP technologies had been initiated by different automotive societies around the globe such as European [23], North America [5], and China [24]. As shown in Fig. 1, early-stage studies (e.g., Guenther et. al. [25], Thandavarayan et. al. [26], etc.) demonstrated the significant potential of CP systems and inspired the drafting of the European CP standards such as TS 103 324 on Cooperative Perception Services and TR 103 562 on Collective Perception Service [23].

For the development of driving automation, the Society of Automotive Engineers (SAE) initiated the SAE J3016 Standard, commonly known as the *SAE Levels of Driving Automation* [27], which has been the fundamental source guiding the development of driving automation. Six levels of driving automation are classified from Level 0 (No driving automation) to Level 5 (Full driving automation) in terms of motor vehicles. Defined by the SAE J3216 Standard [5], CDA enables communication and cooperation between equipped vehicles, infrastructure, and other road users, which will, in turn, improve the safety, mobility, and sustainability of

TABLE I

Relationship Between Classes of CDA Cooperation and Levels of Automation Based on SAE Standards [5]

		SAE Driving Automation (DA) Levels					
		Level 0: **No DA**	**Level 1:** **Driver Assistance**	**Level 2:** **Partial DA**	**Level 3:** **Conditional DA**	**Level 4:** **High DA**	**Level 5:** **Full DA**
CDA Cooperation Classes	**No Cooperation**	e.g., Signage	Relies on driver to supervise performance in real-time		Relies on ADS under defined conditions		
	Class A: status-sharing	e.g., Traffic Signal	Limited Cooperation: Human is driving and supervise CDA features		Improved C-ADS situational awareness by on-board sensing and surrounding road users and operators		
	Class B: intent-sharing	e.g., Turn Signal	Limited Cooperation (only longitudinal OR lateral)	Limited Cooperation (both longitudinal AND lateral)	Improved C-ADS situational awareness through prediction reliability		
	Class C: agreement-seeking	e.g., Hand Signals	N/A	N/A	Improved Ability of C-ADS by coordination with surrounding road users and operators		
	Class D: prescriptive	e.g., Lane Assignment	N/A	N/A	C-ADS has full authority to decide actions except for very specific cases		

transportation systems. By further extending the SAE levels of Driving Automation, SAE J3216 defines the CDA levels into five classes including 1) No cooperative automation, 2) Class A: Status-sharing, 3) Class B: Intent-sharing, 4) Class C: Agreement-seeking, and 5) Class D: Prescriptive. Table I summarized the details and relationship between classes of CDA cooperation and levels of driving automation. According to Table I, cooperative perception plays a significant and fundamental role in supporting both CDA and Automated Driving systems. Based on the analysis of those standards, the architecture and taxonomy of CP are introduced and explained in the following sections.

B. Architecture

In CDA, the fidelity and range of perception information have a significant impact on the system performance for subsequent cooperative maneuvers. Fig. 2 demonstrates a system architecture of the cooperative perception system for enabling CDA. Specifically, four typical phases can be identified in the CP process: 1) *Information Collection*; 2) *Local Computing*; 3) *Perceptual Cooperation*; and 4) *Message Distribution*.

1) Information Collection: Collecting raw data of traffic information serves as a fundamental step in facilitating subsequent perception tasks. As transportation systems evolve, a diverse range of sensors has been deployed to address specific objectives and scenarios. Traditional sensors such as Loop Detectors and Microwave RADAR have found widespread application in traffic surveillance, primarily focus- ing on providing mesoscopic traffic information, including traffic volume and queue length [28]. However, the capabilities of these conventional sensors are limited when it comes to offering comprehensive 3D object-level information necessary for supporting CDA. To address this requirement, high-resolution sensors such as cameras and LiDAR have emerged as indispensable tools capable of generating the desired object-level information.

Several decades ago, the development of intelligent transportation systems faced challenges in leveraging high-resolution sensor-based object perception due to computational limitations and the nascent stage of the computer vision field [29]. Although some vision-based methods were proposed during that period, their performance remained considerably constrained [30]. However, with the rapid advancement in high-performance computation and the proliferation of artificial intelligence (AI) techniques [31], high-resolution sensors now possess the ability to provide precise object-level perception outcomes. These sensors can be deployed on vehicles or integrated into roadside infrastructures to capture the surrounding environment. Subsequently, the collected data is transmitted to a processing server through a communication hub for further analysis and interpretation.

2) Local Computing: Traditional traffic surveillance systems typically do not require high-frequency and low-latency processing. However, in the context of CDA, perception data with a minimal frequency of $1-10Hz$ and a time delay of less than $100ms$ are essential [32]. Transmitting a large volume of raw data, such as point cloud data, over limited bandwidth can lead to unacceptable time delays, particularly in safety- critical scenarios. To address this challenge, it is advantageous to process the information collected from sensors on local servers located on vehicles or infrastructures. Processing the raw sensing data at a single PN typically can be mainly divided into the following blocks [18], outlined as follows (it is noted that the exact order of these blocks may vary according to the exact system design):

- *Preprocessing*: Manipulations of raw data to provide a ready-to-use format for perception modules with respect to specific sensors, such as coordinate transformation, geo-fencing, and noise reduction.
- *Feature Extraction*: Feature extraction for subsequent perception tasks by applying deep neural networks (DNNs) or traditional statistical methods.
- *Multi-Sensor Fusion*: Multi-sensor fusion algorithms may be applied if there is more than one sensor used for a single PN.
- *Detection & Tracking*: Generation of object detection and tracking results for demonstrating position, pose,

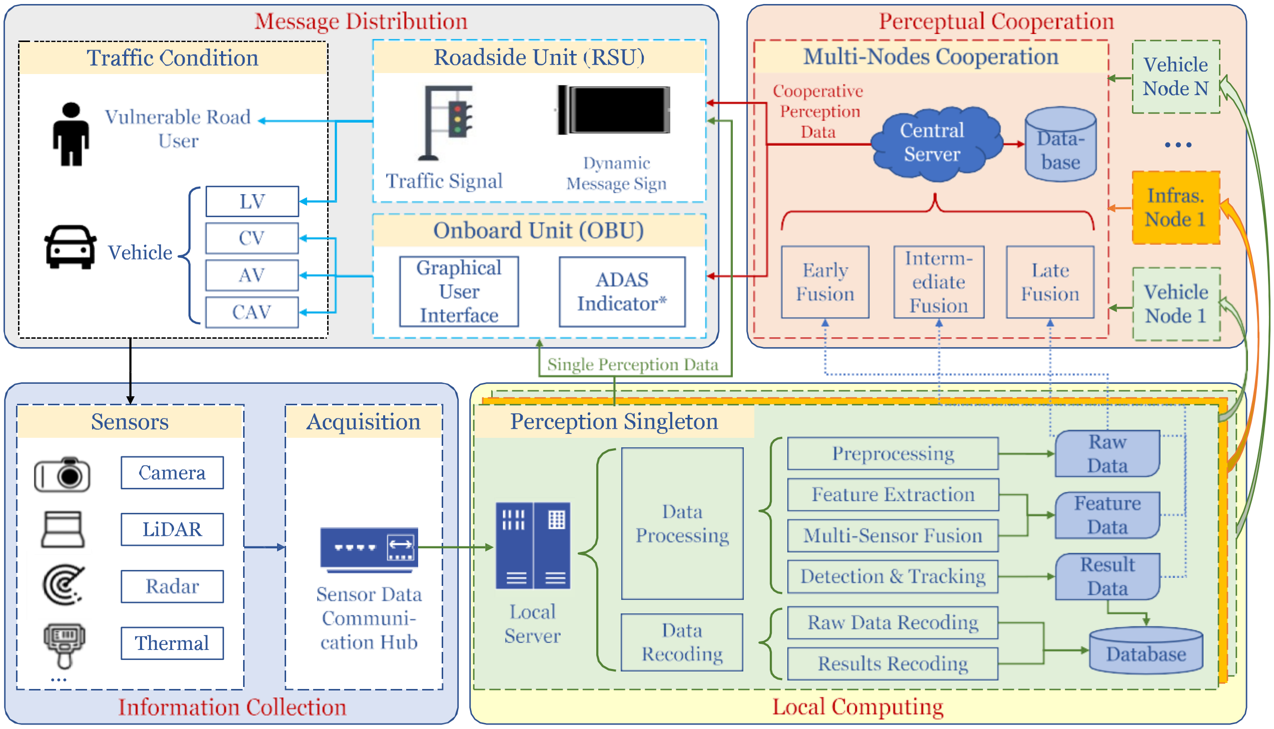


Fig. 2. Systematic architecture for cooperative perception system (*: Other non-visual driving advisory signals for Advanced driver-assistance systems (ADAS), such as auditory, haptic, or even control commands).

and identification of certain road users, such as rotated bounding boxes with unique IDs and classification tags.

- *Raw Data Logging*: Recording of raw sensing data with timestamps for post-analysis.
- *Results Logging*: Recording of semantic perception data with timestamps for post-analysis.

Different types of PNs play different roles in a CP system. For a Vehicle PN (V-PN), local computing mainly serves itself, i.e., perceiving the environment to support the downstream driving tasks such as decision-making or control. For an Infrastructure PN (I-PN), its main purpose is to improve the situation awareness at a fixed location by advanced ranging sensing (e.g., camera, LiDAR) and communications. Generally, three types of perception data are generated from PNs:

- Raw data which contains the original information from sensors, e.g., RGB images from the camera, point cloud data (PCD) from LiDAR, etc.
- Feature data which contains the hidden feature extracted by neural network or statistical methods for representing the raw data in higher dimensional spaces.
- Result data which contains the semantic perception information such as 2D/3D location, size, rotation, etc.

3) Perceptual Cooperation: Considering the large-scale implementation of cooperation, central computing is involved to act as the fusion center for multiple PNs. Information from heterogeneous PNs will be transmitted to the *Central Server* via different kinds of communications. For mobile road users (e.g., vehicles, cyclists, pedestrians), wireless communication, such as *Cellular Network*, *Wireless Local Area Network (WLAN)*, etc. is used to exchange information with the *Central Server*. Additionally, infrastructure can take advantage of both wireless and wired communications (e.g., *Optical Fiber*, *Local Area Network (LAN)*, etc) by well-balancing the cost and system performance such as delay [33].

One of the key components for CP is data fusion and different fusion schemes will be applied, depending on the types of data to be shared between PNs and the *Cloud*. For instance, early fusion, intermediate fusion, and late fusion are based on raw data, feature data, and result data, respectively. Due to the limited bandwidth of wireless communication, result data are most widely implemented for real-world CP systems, such as sharing the object lists from camera- based object detection systems [34] or LiDAR-based object detection systems [18]. A few systems that have high-speed communication capability, which allow high-volume low-latency data transmission, can also transmit raw data to the Cloud for processing, and some work has been conducted to enhance driving automation [20], [35].

In terms of multi-node perception systems, i.e., simultaneously perceiving the environment from different locations, time alignment (with the necessity of delay compensation) and object association need to be considered for spatiotemporal information assimilation and synchronization. Recently, intermediate fusion attracts increasingly popular attention due to its superiority in CP performance [13], [14], [16]. Detailed review and discussion of fusion schemes are provided in Section V.

4) Message Distribution: Perception information (along with advisory or actuation signals) can be distributed to road users in two main ways, depending on connectivity status. For conventional road users without wireless connectivity, this information can be delivered to end devices on the roadside, such as *Dynamic Message Sign* or signal head display of traffic lights through the Traffic Management Center. For

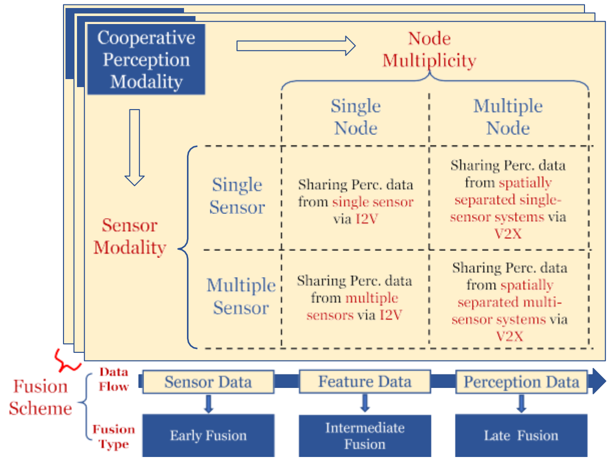

Cooperative Perception Modality
Node Multiplicity
Single Node
Multiple Node
Sensor Modality
Single Sensor
Multiple Sensor
Sharing Perc. data from single sensor via I2V
Sharing Perc. data from spatially separated single-sensor systems via V2X
Sharing Perc. data from multiple sensors via I2V
Sharing Perc. data from spatially separated multi-sensor systems via V2X
Fusion Scheme
Data Flow
Sensor Data
Feature Data
Perception Data
Fusion Type
Early Fusion
Intermediate Fusion
Late Fusion

Fig. 3. Taxonomy of CP in terms of node multiplicity, sensor modality, and fusion scheme.

road users with connectivity, customized information, e.g., surrounding objects and *Signal Phase and Timing* (SPaT) of upcoming signals, and various visual/non-visual ADAS indicators can be accessed to enable various connected driving automation applications, such as Connected Eco-Driving [7], [36]. CP messages can support more sophisticated cooperative maneuvers in mixed-traffic environments. For example, vulnerable road users and legacy vehicles can react to the message shown in DMS [6]. CVs can use CP information to get better situational awareness and pass through intersections in a safer manner [37]. Autonomous vehicles (AVs) and CAVs can improve their driving performance via better coordination algorithms [38]. By leveraging CP messages, road users across different categories can benefit from enhanced safety, efficiency, and coordination in the traffic ecosystem.

C. Taxonomy

Based on the architecture of CP illustrated above, three key aspects are identified for a CP system, namely 1) Node Multiplicity, 2) Sensor Modality, and 3) Fusion Scheme, and Fig. 3 illustrates these aspects in detail. In terms of node multiplicity and sensing modality, four types of CP systems can be identified as follows:

- *Single-Node Single-Mode CP (SS-CP)*: Cooperation between a PN equipped with the single-modal sensor(s) and other users with connectivity only [18], [34].
- *Multi-Node Single-Mode CP (MS-CP)*: Cooperation between multiple PNs equipped with the single-modal sensor(s) and connectivity [13], [20], [21].
- *Single-Node Multi-Mode CP (SM-CP)*: Cooperation between a PNs equipped with the multi-modal sensor(s) and other users with connectivity only.
- *Multi-Node Multi-Mode CP (MM-CP)*: Cooperation between multiple PNs equipped with the multi-modal sensor(s) and connectivity [14], [16].

In the following, a comprehensive literature review is conducted with detailed analyses of the aspects of node multiplicity, sensing modality, and fusion scheme, respectively.

III. Node Structure for CP

This section aims to review the CP system from the perspective of node structure as mentioned in Fig. 3. For comprehensiveness and conciseness, we discuss CP methods with different types of PNs including the vehicle PN (V-PN), the infrastructure PN (I-PN), and the heterogeneous PN (H-PN).

A. I-PN-Based CP

Object perception based on roadside sensors has a great potential to break the current bottleneck for automated driving, especially in a mixed traffic environment via cooperative perception [39]. This section reviews the infrastructure-based object detection and tracking approaches in the literature. Specifically, a single I-PN equipped with communication devices can be used for enhancing the perception capacity of vulnerable road users or vehicles with connectivity within certain scenarios, such as intersection areas. Thus, in this section, both single-I-PN perception and multi- I-PN perception models are regarded as I-PN-based CP methods.

1) Single-I-PN Perception: Infrastructure-based camera systems have been widely used for object detection and a survey conducted by Zou et al. [12] shows various camera-based applications in traffic scenes, such as traffic surveillance, safety warning, traffic management, etc. Monovision camera plays a significant role in object detection. Ojala et al. proposed a *Convolutional Neural Network* (CNN) based pedestrian detection and localization approach using roadside cameras [40]. The perception system consists of a monovision camera streaming video and a computing unit that performs object detection and positioning. Besides, Guo et al. proposed a 3D vehicle detection method based on a monocular camera [41], which consists of three steps: 1) clustering arbitrary object contours into linear equations; 2) estimating positions, orientations, and dimensions of vehicles by applying the K-means method; and 3) refining 3D detection results by maximizing a posterior probability.

Instead of using a fixed roadside camera, some researchers try to take advantage of Unmanned Aerial Vehicle (UAV) based cameras. MultEYE [42] is a monitoring system for real-time vehicle detection, tracking, and speed estimation proposed by Balamuralidhar et al. Different from general roadside sensors equipped on signal poles or light poles, the data source of MultEYE comes from a UAV equipped with an embedded computer and a video camera. Inspired by the multi-task learning methodology, a segmentation head [43] is added to the object detector backbone [44]. Dedicated object tracking [45] and speed estimation algorithms have been optimized to track objects reliably from a UAV with limited computational efforts. Cicek and Gören proposed a deep-learning-based automated curbside parking spot detection approach through a roadside camera [46]. To identify the road boundaries, object detection and road segmentation methods are employed by utilizing *the FCN-VGG16* model [47] on the *KITTI* dataset [48] and *Faster R-CNN* [49] on *MS-COCO* dataset [50], respectively. Then, a method is designed to differentiate parked vehicles from moving ones and then give guidance on the nearest spot information to drivers.

In recent years, roadside LiDAR sensors have attracted increasing attention from researchers about object perception in transportation. Using roadside LiDAR, Zhao et al. proposed a detection and tracking approach for pedestrians and vehicles [51]. As one of the early studies utilizing roadside LiDAR for perception, a classical detection and tracking pipeline for PCD was designed. It mainly consists of 1) *Background Filtering*: To remove the laser points reflected from road surfaces or buildings by applying a statistics-based background filtering method [52]; 2) *Clustering*: To generate clusters for the laser points by implementing a DBSCAN method [53]; 3) *Classification*: To generate different labels for different traffic objects, such as vehicles and pedestrians, based on neural networks [54]; and 4) *Tracking*: To identify the same object in continuous data frames by applying a discrete Kalman filter [55]. Based on the aforementioned work, Cui et al. designed an automatic vehicle tracking system by considering vehicle detection and lane identification [56]. A real-world operational system is developed, which consists of a roadside LiDAR, an edge computer, a *Dedicated Short- Range Communication* (*DSRC*) *Roadside Unit* (*RSU*), a Wi-Fi router, and a DSRC *On-board Unit* (*OBU*), and a GUI. Following a similar workflow, Zhang et al. proposed a vehicle tracking and speed estimation approach based on a roadside LiDAR [57]. Vehicle detection results are generated by the "*Background Filtering-Clustering-Classification*" process. Then, a centroid-based tracking flow is implemented to obtain initial vehicle transformations, and the unscented Kalman Filter [58] and joint probabilistic data association filter [59] are adopted in the tracking flow. Finally, vehicle tracking is refined through a Bird's-Eye-View (BEV) LiDAR-image matching process to improve the accuracy of estimated vehicle speeds. Following the bottom-up pipeline mentioned above, numerous roadside LiDAR-based methods are proposed from various points of view [60], [61], [62], [63], [64].

On the other hand, using learning-based models to cope with LiDAR data is another main methodology. Bai et al. [37] proposed a deep-learning-based real-time vehicle detection and reconstruction system from roadside LiDAR data. Specifically, CARLA simulator [65] is implemented for collecting the training dataset, and ComplexYOLO model [66] is applied and retrained for the object detection on the CARLA dataset. Finally, a co-simulation platform is designed and developed to provide vehicle detection and object-level reconstruction, which aims to empower subsequent CDA applications with readily retrieved authentic detection data. In their following work for real-world implementation, Bai et al. [18] proposed a deep-learning-based 3D object detection, tracking, and reconstruction system for real-world implementation. The field operational system consists of three main parts: 1) 3D object detection by adopting PointPillar [67] for inference from roadside PCD; 2) 3D multi-object tracking by improving DeepSORT [68] to support 3D tracking, and 3) 3D reconstruction by geodetic transformation and real-time onboard *Graphic User Interface* (*GUI*) display.

2) Multi-I-PN Perception: Leveraging multiple I-PNs can significantly improve the perception range. For I-PN-based CP systems using cameras, Arnold et al. proposed a cooperative 3D object detection model by utilizing multiple depth cameras to mitigate the limitation of field-of-view (FOV) of a single-sensor system [21]. For each camera, a depth image is projected to pseudo-point-cloud data [69]. Two sensor-fusion schemes are designed: early fusion and late fusion which are adapted from Voxelnet [70]. The evaluation in a T-junction and a roundabout scenario in the CARLA simulator [65] demonstrates that the proposed method can enlarge the detection coverage without compromising accuracy. To take advantage of LiDAR for I-PN-based CP systems, VINet [71] was proposed to consider a scalable number of I-PN LiDAR inputs. Specifically, CNNs are designed for feature extraction and a specific two-stream fusion method was proposed to fuse features from scalable numbers of I-PNs.

B. V-PN-Based CP

Cooperative perception between vehicles mainly emerged from the research for Unmanned Aerial Vehicles (UAVs) to provide estimated localization in the region of interest. Back in 2006, Merino et al. [72] proposed a multi-UAV CP system based on a distributed-centralized CP framework (similar to the current "local-central" framework). The sensor data (such as images) collected from UAVs will be processed on the UAV side including image segmentation, stabilization of sequences of images, and geo-referencing. The location of objects in the region of interest will be estimated by UAVs and then sent to a central server for further fusion by utilizing a probabilistic model.

For on-road vehicles, Rockl et al. [73] propose a *Multi-Sensor Multi-Target Tracking* method by associating the received sensor data via V2V communication. A more notable CP system for on-road vehicles was proposed by Rauch et al. [74] in 2012. A Car2X-based module was proposed to fuse perception results for both spatial and temporal dimensions via the Unscented Kalman filter (UKF). Specifically, the object data shared from other vehicles need to be aligned to the coordinate of the host vehicle and synchronized in time. Rawashdeh and Wang [75] proposed a machine learning-based method to fuse proposals generated by different connected agents. A specific center-point estimation method was proposed for generating the object location into the coordinate system of the host vehicle. Xiao et al. [76] proposed a CP method by sharing semantic segmentation information generated by a DNN and vision-feature matching data from the BEV-projected image data. GPS data was required for spatial alignment.

A comprehensive automated driving system (ADS) was implemented by Kim et al. [77], whose core innovation is a CP system that provides ego-vehicle information beyond occlusion by a leading vehicle. A real-world system was deployed to validate the effectiveness of CP by various tasks (e.g., collision warning, overtaking/lane-changing assistance, etc.), which demonstrated the potential of improving driving automation through CP technology.

TABLE II
SUMMARY OF DIFFERENT NODE STRUCTURES FOR COOPERATIVE PERCEPTION

Structure	Modality	Pros. and Cons.	Highlighted Features	Literature
Single-Node	Infrastructure	Pros: Higher location with flexible pose leads to less occlusion and system-level cost-effective.	Infrastructure assisted high-fidelity traffic surveillance	Bai et al. [18]
		Cons: Need infrastructure support.		
	Vehicle	Pros: Low latency perception for ego-vehicle.	Everything on the vehicle side: sensing, processing, analysis.	Arnold et al. [10]
		Cons: Easily occluded by the surrounding vehicles or buildings.		
Multi-Node	Vehi. + Vehi.	Pros: Extend perception range from vehicle side.	Sharing features generated from convolutional neural networks.	Chen et al. [13]
		Cons: Occlusion by other vehicles.		
	Infra. + Infra.	Pros: Extend perception range from the infrastructure side.	Sharing preprocessed RGB data among all roadside sensors.	Arnold et al. [21]
		Cons: Have blind zone under the sensor.		
	Infra. + Vehi.	Pros: Achieve a comprehensive range and field of view (FOV) for perception.	Considering asynchronous information sharing, pose errors, and heterogeneity of V2X components.	Xu, et al. [16]
		Cons: Require heterogeneity of the model.		

For CP systems based on LiDAR data, Chen et al. proposed an early fusion method (*Cooper* [20]) by aligning raw point cloud data (PCD) from multiple vehicles. To fulfill the limited bandwidth of V2V communication, raw PCD was preprocessed to reduce its size. Additionally, GPS and *Inertial Measurement Unit* (IMU) data were required for PCD alignment. Then a PCD detector was designed based on VoxelNet [70], Sparse Convolution [78], and Region Proposal Network (RPN) [79]. The experiments demonstrated that Cooper was capable of improving perception performance by expanding sensing data. Following the *Cooper*, Chen et al. proposed *F-Cooper* [13], a feature-based CP system using PCD. The core idea of F-Cooper is a two-step process: 1) to extract the hidden feature from sensor data via a DNN at each vehicle side, i.e., V-PN; 2) to generate perception results based on cross-vehicle feature data sharing.

CNN-based feature sharing was also applied in the work proposed by Marvasti et al. [80] for the V2V CP task, named *Feature Sharing Cooperative Object Detection* (FS-COD). Both FS-COD and F-Cooper complete spatial alignment at the feature level. However, different from F-Cooper which uses *maxout* operation [81] (i.e., output maximum value for corresponding multi-source data points) to fuse the multi-source data, FS-COD uses summation for multi-source feature fusion.

Considering compressing the feature data for transmission, Wang et al. proposed V2VNet [82], which leverages the power of both deep neural networks and data compression. Specifically, a pipeline of "*feature extraction-compression-decompression-object detector*" is created to further consider the limitation of communication. Additionally, a novel simulator, *Lidarsim* [83], is involved for cooperative perception to generate a PCD-based V2V dataset in a more realistic manner.

Zhang et al. [84] proposed a vehicle-edge-cloud framework for dynamic map fusion. Federated learning is applied for generating object detection results from multiple V-PNs and a three-stage fusion scheme is proposed to generate the final objects based on overlapping results from multiple PNs.

Xu et al. [85] propose a feature-sharing-based CP model by V2V communication. Vehicles' relative pose information with respect to ego-vehicle is required for spatial alignment and feature generation. Specifically, the attention operation [86] is applied for multi-node feature fusion and an open-source simulation-based dataset is developed and implemented for model training and validation.

C. H-PN-Based CP

Although many researchers have dug into cooperative perception from the perspectives of infrastructure perception and V2V cooperation, so far, only a few pieces of research are conducted for CP between heterogeneous PNs, i.e., cooperation between vehicles and infrastructure.

For cooperation between vehicles and infrastructure, Bai et al. [14] proposed a CP method, named *PillarGrid*, to generate 3D object detection results based on PCD from onboard-roadside LiDAR sensors. Specifically, decoupled multi-stream CNNs are applied for feature extraction. The vehicle pose information is required for spatial alignment and the feature data are shared via V2X communication. A Grid-wise Feature Fusion (GFF) method is proposed for multi-PN feature fusion, which endows the PillarGrid with better scalability and capacity to handle heterogeneity.

Using Vision Transformer (ViT) [87], Xu et al. [16] proposed a CP method named *V2X-ViT*, which applied a share-weights CNNs for feature extraction. Ego-vehicle pose information is transmitted to surrounding vehicles and infrastructures for raw data alignment. Heterogeneous Graph Transformer (HGT) [88] is designed to deal with different feature fusion types, e.g., V2V, V2I, etc. A window attention module is designed to capture hidden features from the fused feature map, which is then used to generate the object detection results.

Table II summarizes the advantages and disadvantages of different node structures for cooperative perception. In a nutshell, V-PN is more ego-efficient (i.e., improving the perception capability from the standpoint of ego-vehicle.) while I-PN is more suitable for scalable cooperation. CP between homogeneous PNs, such as V2V or I2I, can mainly extend the perceptive range while CP between heterogeneous PNs, such as V2X, can achieve better FOV by complementing different sensor configurations.

IV. Sensor Modality for CP

For the CP system, sensors are the most fundamental modules due to their roles in raw data collection. This section aims to overview the CP system from the perspective of sensing modality by 1) introducing the specification and performance of different types of sensors that are mainly utilized in transportation systems, 2) reviewing the development of single-sensor perception systems based on cameras, LiDAR and RADARs, and 3) summarizing the studies of multi-sensor perception systems in terms of homogeneous sensor fusion and heterogeneous sensor fusion.

A. Sensor Specification and Performance

For sensors equipped with current ADS, the most popular ones are cameras, LiDAR, and RADAR [89]. Onboard RADAR has been deployed on vehicles to mainly achieve ADAS functionalities for many years [90], such as Adaptive Cruise Control (ACC), Collision Avoidance, etc [91].

Different ADS may take different sensor configurations. For instance, the ADS developed by *Comma.ai* [92], only deploys a single-camera-based perception system at the middle-top of the windshield. *Waymo ADS* [93], utilizes multi-modality sensors including 1) multiple cameras installed around the top-surrounding positions of the vehicle, 2) LiDAR sensors equipped on the top and two front sides of the vehicle, and 3) RADARs integrated at lower-surrounding positions around the vehicle.

Regarding the installation of roadside sensors, typical locations may include signal arms and street lamp posts, with some minimum height requirements to avoid tampering. As a result, roadside sensors can have a much higher position (compared to onboard sensors) to minimize the occlusion effect due to dense traffic. The specific installation position may vary based on different roadside sensors. For example, the roadside LiDAR sensors are mainly installed at the height of $3 - 6m$ (but no more than $10m$), while fisheye cameras prefer a higher installation [18], [37], [51].

To form a comprehensive view of the general performance of different sensors used for perception in transportation systems, Table III provides a summary of those that are widely utilized in ADS, traffic surveillance, and other transportation systems. Each of these sensors has its own capabilities and strengths in different use cases.

- Camera: High-resolution. Not great for 3D position and speed measurements, especially in dense traffic.
- LiDAR: High-accuracy 3D perception with resilience to environmental changes. Not great with its relatively high price and data sparsity.
- RADAR: Measuring speed, unlocking applications like stop bar & dilemma zone detection. Not great for distinguishing objects.
- Thermal Camera: Getting thermal information, which provides resilience to lighting changes.
- Fisheye Camera: 360-degree full field-of-view (FOV) for detection. Requires a high-accuracy calibration matrix to account for distortion.
- Loops: Measuring traffic counts and speed. Costly to install and maintain due to intrusiveness.

B. Multi-Sensor Perception

Owing to the complementary of different sensors, multi-sensor-based perception systems have the potential to achieve better object detection and tracking performance via sensor fusion. In this section, homogeneous multi-sensor perception, which fuses sensor data from the same type of sensors, is reviewed first followed by the discussion of heterogeneous multi-sensor perception methods, such as *Camera+LiDAR*, *Camera+RADAR*, etc.

1) Homogeneous Multi-Sensor Perception: The multi-camera system has been developed for decades and lots of applications have been designed and implemented in our current transportation systems [96], such as object detection and object tracking.

For object detection, before the surge of CNN, the extraction and fusion of object-level features is a major challenge for traditional methods due to the high-dimensional complexity of RGB data. Merino et al. [72] proposed a multi-UAV CP system based on heterogeneous sensor systems including infrared and visual cameras, fire sensors, and others. A set of functions were designed for object detection including image segmentation, and stabilization of image sequences. By coordinating the processed results from spatially separated sensors, the targeting object can be detected and localized based on a geo-referencing process.

With the tremendous power of CNN to extract hidden features, object detection based on multi-camera systems quickly attracts lots of attention from researchers. For spatial alignment for the multi-node cameras, Arnold et al. [21] chose to project camera data from RGB images to pseudo-PCD. Owing to the 3D attribute of PCD, this pseudo-PCD could be easily aligned and merged into a unified coordinate system. Then a deep learning-based object detector was applied to generate perception results.

Object tracking has been widely developed in multi-camera systems for several decades to enable traffic surveillance and thus to analyze the traffic scenarios for further traffic optimization [97]. The most typical way of multi-camera tracking is to calibrate the multi-camera systems to make all views stitched together in a unified coordinate system [98]. Meanwhile, consecutively tracking multi-objects under occluded conditions is one of the main strengths of a multi-camera tracking system which can provide sequences of images from different viewpoints. Specifically, based on the unified coordinate system gained from calibration, the Kalman Filter [99], the particle filter [100], etc., have been widely applied in multi-video object tracking systems.

The tracking schemes mentioned above generally require joint FOV for computing association across cameras. For the

TABLE III

SENSOR PERFORMANCE MATRIX FROM DIFFERENT ASSESSING PERSPECTIVES (RATING RANGE FROM 1 TO 3 STARS BASED ON THEIR RELATIVE PERFORMANCE)

Capabilities	Camera	LiDAR	RADAR	Thermal	Fisheye	Loop
Privacy-safe data	☆	☆☆☆	☆☆☆	☆☆	☆	☆☆☆
Accurately detects and classifies objects	☆☆	☆☆☆	☆	☆☆	☆☆	☆
Accurately measures object speed and position	☆	☆☆☆	☆☆☆	☆☆	☆	☆☆
Extensive FOV	☆	☆☆☆	☆☆	☆	☆☆☆	☆
Robustness across different environmental conditions	☆	☆☆	☆☆☆	☆☆	☆	☆☆☆
Ability to read signs and differentiate color	☆☆☆	☆☆	☆	☆	☆☆☆	☆
Cost for deployment and maintenance	☆☆☆	☆	☆☆	☆☆	☆☆	☆

TABLE IV

SUMMARY OF DIFFERENT SENSOR MODALITIES FOR COOPERATIVE PERCEPTION

Structure	Modality	Pros. and Cons.	Highlighted Features	Literature
Single-Sensor	Camera	Pros: abundant vision data with cost-effective system.	Using shifted window multi-head attention	Liu et al. [94]
		Cons: difficult to provide high-fidelity 3D information and significantly impacted by the vulnerability of weather conditions.		
	Lidar	Pros: capable to provide high-fidelity 3D information with panoramic FOV.	Encoding point cloud into voxelized pillars	Lang et al. [67]
		Cons: sparse data without vision information.		
	Radar	Pros: tolerance to adverse weather conditions and low visibility situations, and direct measurement of speed of motion.	Roadside millimeter RADAR for detecting vulnerable road users.	Liu et al. [19]
		Cons: lower spatial resolution and may struggle with detailed object recognition tasks.		
Multi-Sensor	Homogeneous modality	Pros: a straightforward way to expand the FOV and perception area by fusing the sensor data with similar information distributions.	Projecting RGB camera data into pseudo-LiDAR point cloud.	Arnold et al. [21]
		Cons: difficult to provide comprehensive information such as high-fidelity 3D information from cameras, and vulnerability to the same adverse environmental conditions.		
	Heterogeneous modalities	Pros: taking advantage of different sensor modalities to achieve more comprehensive sensing information, such as camera, LiDAR, RADAR.	Capturing BEV features from different types of sensors via CNN.	Liu et al. [95]
		Cons: different data distribution retrieved from different sensor modalities, which are difficult to be fused effectively.		

disjoint camera system, appearance cues are designed for capturing the common features between multiple views by integrating spatial-temporal information [101]. To overcome the dynamically changed spatial-temporal information in vision information, e.g., lighting condition and traffic speed, the tracking model should also be able to update its model adaptively. Thus, Expectation-Maximization (EM) framework [102], unsupervised learning network [103], etc., have been implemented to dynamically update the model.

Although one single LiDAR can provide panoramic FOV around the ego-vehicle, physical occlusion may easily block the perceptive range and cause the ego-vehicle to lose some crucial perception information which significantly affects its decision-making or control process. On the other hand, a spatially separated LiDAR perception system can expand the perceptive range for intelligent vehicles or smart infrastructure.

One of the straightforward inspirations of the multi-LiDAR perception system is sharing the raw PCD via V2V communication [20]. However, limited wireless communication bandwidth may significantly limit real-time performance. Feature data generated from CNN requires much less bandwidth and is more robust to sensor noises, thus becoming a popular solution to multi-LiDAR fusion [13], [82]. Marvasti et al. [80] used two sharing-parameter CNNs to extract the feature map for PCD retrieved from two vehicle nodes. Feature maps were then aligned based on the relative position and fused by element-wise summation. By applying an attention mechanism, Xu et al. [85] proposed a V2V- based cooperative object detection method. A similar CNN process [67] was designed for extracting feature maps for V2V sharing. Furthermore, self-attention was involved in data aggregation based on spatial location in the feature map.

Recently, researchers started focusing on cooperation between V-PN and I-PN based on the multi-LiDAR system. For handling the data heterogeneity from the roadside and onboard PCD, Bai et al. [14] proposed a decoupled multi-stream CNN framework for generating feature maps accordingly. Relative position information was applied to PCD alignment and the shared feature maps were then fused based on grid-wise *maxout* operation. Additionally, Xu et al. [16] proposed a ViT-based CP method for heterogeneous PNs. Feature maps were extracted using sharing-parameter CNNs and V2X communications. For dealing with heterogeneity, specific graph transformer structures were designed for data extraction.

2) Heterogeneous Multi-Sensor Perception: As different sensor modalities, the camera and LiDAR seem to be a naturally complementary couple for perception. For instance, the camera is good at perceiving vision information but lacks 3D distance data, while the LiDAR excels at collecting 3D information but lacks vision data.

TABLE V
SUMMARY OF DIFFERENT FUSION SCHEMES FOR COOPERATIVE PERCEPTION

Fusion Scheme	Methodology	Pros. and Cons.	Highlighted Features	Literature
Early Fusion	Model-based Fusion	Pros: Raw data is shared and gathered to form a holistic view. Cons: Low tolerance to the noise and delay of the transmitted data; potentially constrained by the communication bandwidth.	Raw point cloud data is compressed to fit the limited bandwidth.	Chen et al. [20]
Intermediate Fusion	Model-free Fusion	Pros: High tolerance to the noise, delay, and difference between different nodes and sensor models. Cons: Require training data and hard to find a systematic way for model design.	Deep neural features are extracted and fused based on spatial correspondence.	Bai et al. [14]
Late Fusion	Model-based Fusion	Pros: Easy to design and deploy in the real-world system. Cons: Significantly limited by the wrong perception results or the difference between sources.	A late-fusion is proposed based on joint re-scoring and non-maximum suppression.	Zhang et al. [84]

One typical way for the fusion of multi-modal sensor data is using CNN to extract hidden features in parallel and then combine them on the corresponding scale level. Zhu et al. proposed *Multi-Sensor Multi-Level Enhanced YOLO* (*MME-YOLO*) for vehicle detection in traffic surveillance [17]. MME-YOLO consists of two tightly coupled structures: 1) The enhanced inference head is empowered by attention- guided feature selection blocks and anchor-based/anchor-free ensemble head in terms of better generalization abilities in real-world scenarios; 2) The LiDAR-Image composite module is based on CBNet [104] to cascade the multi-level feature maps from the LiDAR subnet to the image subnet, which strengthens the generalization of the detector in complex scenarios. MME-YOLO can achieve better performance for vehicle detection compared with YOLOv3 [105] for roadside sensor data.

Since the camera and LiDAR have different poses and FOV, creating an intermediate feature level to unify LiDAR and image data before sending it to the feature-extraction backbone becomes a promising way for multi-modal sensor fusion [106]. A popular way is to project camera information into LiDAR data to endow PCD with vision information. *PointPainting* [107], a point-level feature fusion method, decorates the PCD with semantic segmentation results from vision data. The point cloud data decorated with vision information are then fed into detectors, e.g., *PointPillar* [67] for generating object detection results. Recently, Liu et al. [95] proposed a novel framework, named *BEVFusion*, to project both RGB and PCD information into a BEV feature map for fusion. Specifically, two dedicated encoders were designed to extract RGB and PCD inputs into the BEV feature map. Then, multi-modal feature fusion was conducted based on the spatial correspondence of BEV feature maps.

Additionally, empowered by remarkable depth-sensing capability, RADARs are innately complementary to cameras to improve the overall perception ability [108]. In the early stage of RADAR-camera fusion studies, RADAR data was usually extracted to enhance the depth information for visual data [109], which was straightforward but not very reliable and high-performance. Conversely, perception pipelines can be designed separately with respect to camera data and RADAR data respectively. Then traditional multi-sensor fusion methods can be applied to fuse these multi-source perception results, such as *Probabilistic Reasoning*-based fusion studies [110], and *Kalman Filter*-based fusion methods [111].

Recently, DNN-based methods became a dominant solution to fuse camera and RADAR data with higher performance. For instance, CNNs were applied to extract the hidden feature for both camera data and RADAR data and then these features were fused together to enhance the feature representation [112], [113]. Meanwhile, Transformer models [86] also attracted increasing attention to fuse features from different sensor modalities using their self-attention or cross-attention mechanism [114].

V. FUSION SCHEME FOR CP

In terms of the stage of sensor fusion, a multi-sensor perception system can be divided into three classes: 1) *Early Fusion* – to fuse raw sensor data with basic preprocessing steps; 2) *Intermediate Fusion* – to fuse intermediate feature data within the perception models (typically the intermediate feature map within a neural network); and 3)*Late Fusion* – to fuse perception results from individual perception pipelines for different perception nodes. It is noted that, in the context of CP, raw data typically means the output data after the proprietary decoding process of the sensors, such as the pixel matrix from cameras or point cloud data from LiDAR, which have a common format.

Different fusion schemes have their specific advantages and disadvantages in terms of distinct perspectives. Early Fusion and Intermediate Fusion have higher accuracy but need more computational power and complex model design. Conversely, Late Fusion can achieve better real-time performance but may sacrifice accuracy. It depends on the specific demands under different traffic scenarios to determine the best deployment of fusion schemes. Take a 64-beam LiDAR as an example, early fusion and intermediate fusion will roughly require 10 to 50$M\,Bps$ communication bandwidth which is much more than the late fusion methods ($\ll 1M\,Bps$). In the meantime, early fusion/intermediate fusion methods could provide 10% to 20% accuracy improvement [115]. The decision of such a trade-off typically depends on the actual use cases. For instance, for safety-critical applications, accuracy will be placed a higher weight than communication while for some communication-critical applications, methods that consume less bandwidth would be a better choice (It is noted that the actual bandwidth consumption varies on the specific data type and packaging methods.). This section aims to give a brief landscape of how fusion schemes are considered and applied in relevant CP research. Also, we will focus more on work that has not been introduced in previous sections.

A. Early Fusion

An obvious approach is to share the raw sensor data with other PNs to expand the perceptive range and improve detection accuracy. Following this strategy, the raw sensor data from multiple PNs are projected into a unified coordinate system for further processing [98]. However, since the basic idea of early fusion is only the expansion of raw data range or density, it is inevitably sensitive to the quality of sensor data, such as sensor calibration issues and data unsynchronization [97]. Thus, early fusion can potentially provide the ideal performance only under several restricted assumptions, such as high-accurate sensor calibration and multi-source synchronization, which requires lots of effort in real-world implementations.

On the other hand, early fusion requires a large communication bandwidth to transmit a high volume of raw data. It is suitable for transmitting camera data with limited image resolution, but it may not be feasible to share real-time LiDAR data within a certain time delay (A 64-beam Velodyne LiDAR with 10Hz may generate about 20MB of data per second [48]). For V2V early fusion, it is true that communicating raw sensor data with one ego-vehicle is not an impossible solution [20], but it is definitely not feasible for large-scale V2V cooperative perception under current communication capability.

B. Late Fusion

Standing in the opposite direction compared with early fusion, late fusion chooses another natural cooperative paradigm for perception – generating perception results independently and then fusing them together. Different from early fusion, although late fusion also needs a relative position for fusing these perception results, its tolerance to calibration errors and unsynchronization issues is much higher than early fusion. One of the main reasons is that object-level fusion can be determined based on spatial and temporal constraints. For instance, Rauch et al. [74] applied EKF to jointly align the shared bounding box proposals based on spatiotemporal constraints. Additionally, Non-Maximum Suppression (NMS) [116] and other machine-learning-based proposal refining methods are widely applied in late fusion methods for object perception [21]. Recently, due to the distributed attributes of late fusion, *Federated Learning* [117] also attracts increasing popularity in perception systems [84].

C. Intermediate Fusion

The core ideology of intermediate fusion can be simply summarized as using deeply extracted features for fusion that happens at the intermediate stages of the perception pipeline. Intermediate fusion relies on hidden features mainly extracted from deep neural networks, which have higher robustness compared with raw sensor data used for early fusion. Xu et al. [16] assessed the robustness of model performance under different time delays and noises of metadata (the ego-vehicle location and heading). Different levels of errors were involved in the cooperative perception process. The evaluation results can be summarized as three points:

- With no error involved, early fusion and intermediate fusion can achieve similar performance which is better than late fusion;
- With the increase of errors, the performance of both early fusion and late fusion decreases drastically, but the performance degradation of all intermediate fusion methods [13], [16], [82], [85] is much less noticeable than early fusion and late fusion.

Additionally, feature-based fusion methods typically have only one detector for generating object perception results and thus there is no need for merging multiple proposals as required by late fusion [21], [84].

Although cooperative perception has been developed in multiple areas for several decades, deep-fusion-based cooperative perception is an emerging field. Most of the intermediate fusion methods for CP were devised in the past few years. But the related research interests wildly surged up, for example, *F-Cooper* [13] (2019), *V2VNet* [82] (2020), *OPV2V* [85], *CoFF* [118], *DiscoNet* [119](2021), *PillarGrid* [14], *PV-RCNN* [120], *CRCNet* [121], *VINet* [71] and *V2X-ViT* [16] (2022), etc. So far, most of the deep feature extraction is conducted by CNN, such as [13], [14], [82], [122], because the CNN-based feature is highly related to the local spatial information. Recently, some studies have applied transformers as the deep feature extractor [16], [85], [123] due to their capability for feature extraction with larger receptive fields.

VI. Hierarchical Cooperative Perception Framework

Based on the overview of the aforementioned literature, Three major issues can be identified for CP systems in the real world:

- **Heterogeneity**: the CP system should take advantage of both intelligent vehicles and smart infrastructures to empower the comprehensiveness of perception.
- **Scalability**: the CP system needs to be able to extend to different scales of cooperation levels, such as intersection level, corridor level, and traffic network level.
- **Dynamism**: the CP system needs to be able to dynamically cooperate with vehicle perception nodes, i.e., the I-PN should be capable of cooperating with a dynamically changed number of V-PNs.

To address the issues mentioned above, we propose a unified CP framework, called *Hierarchical Cooperative Perception* (HCP) Framework, which is demonstrated in Fig. 4. HCP aims to assimilate different CP tasks under various scenarios into a general framework. The design of the HCP framework is based on 1) the system architecture for CP as shown in Fig. 2, 2) the taxonomy of CP as shown in Fig. 3, and 3) the analysis of reviewed literature.

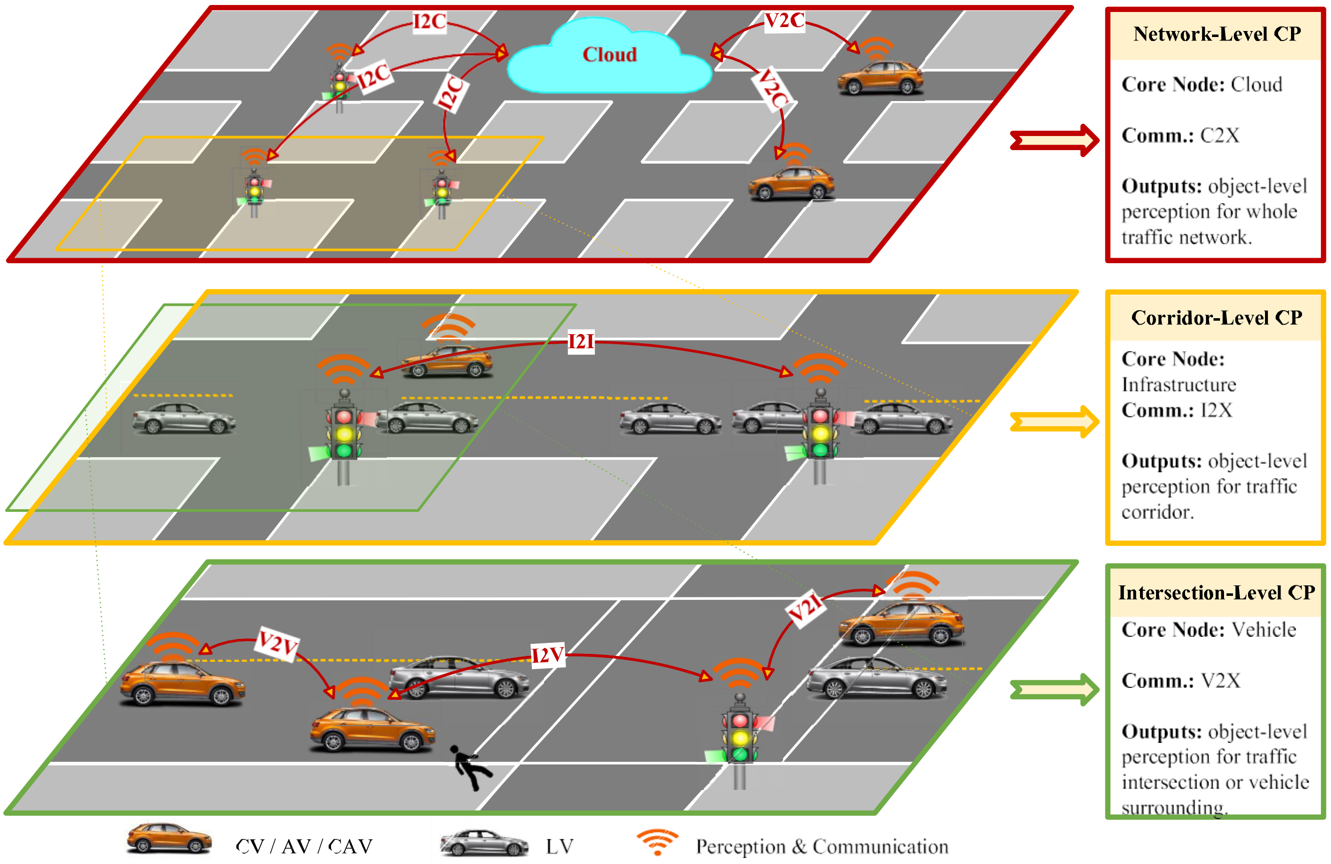

I2C
I2C
I2C
Cloud
V2C
V2C
Network-Level CP
Core Node: Cloud
Comm.: C2X
Outputs: object-level perception for whole traffic network.
I2I
Corridor-Level CP
Core Node: Infrastructure
Comm.: I2X
Outputs: object-level perception for traffic corridor.
V2V
I2V
V2I
Intersection-Level CP
Core Node: Vehicle
Comm.: V2X
Outputs: object-level perception for traffic intersection or vehicle surrounding.
CV / AV / CAV
LV
Perception & Communication

Fig. 4. The schematic diagram of the HCP framework.

In this paper, the HCP framework mainly focuses on the intersection scenarios and consists of threelevels: 1) Intersection-Level CP, 2) Corridor-Level CP, and 3) Network-Level CP, which will be introduced from several perspectives including core node, communication types, and perception outputs, respectively.

A. Intersection-Level CP

As shown in the bottom part of Fig. 4, intersection-level CP aims to perceive the object-level traffic condition around an intersection. V-PNs are designed as the core perception node at this level. For vehicles that are equipped with powerful onboard processors such as CAVs, features can be shared via V2V communication and processed onboard. The perception results from I-PN can act as auxiliary data to augment the CAV's perception results by late fusion. Most of the previous V2V CP work [13], [82], [85] can be integrated into our HCP framework from this perspective.

Since the edge processor can be deployed at the I-PN for processing the roadside sensor data and the data received from intelligent vehicles via V2I communication, vehicles are not necessarily required to be equipped with a powerful onboard processor for processing the whole perception pipeline. Lightweight computing units can be deployed for only extracting the feature. Deep features from multiple vehicles can be transmitted to the I-PN for intermediate fusion to generate perception results. The I-PN then broadcasts the perception results to vehicles within its own communication range. Recent infrastructure-enabled CP methods can be regarded as a specific version of the intersection-level CP [14], [16]. Intersection-level CP is a crucial component for unlocking the current bottleneck (in terms of efficiency, safety, and sustainability) for CDA in a mixed traffic environment [7].

B. Corridor-Level CP

As shown in the middle of Fig. 4, corridor-level CP aims to expand the perception based on the connectivity of multiple smart infrastructures in which the core node is I-PN. Currently, I2I communication (via cable or optical fiber) has a much higher capacity compared with wireless communication. For instance, optical fiber can achieve over 40GB/s communication speed with low latency and even commercial optical-fiber internet can achieve 1GB/s [124]. Theoretically, empowered by high-speed communication, I2I-based CP is capable of applying all aforementioned fusion schemes based on specific scenarios. For instance, raw data sharing can be a typical style for I2I-based CP [21].

Practically, however, the computational bottleneck on the computers at each end of the communication pipeline will occur when the approach requires the computer to encode and decode massive amounts of data, which will become the bottleneck in real-world applications. Thus, the capacity of the data encoding/decoding should also be carefully considered.

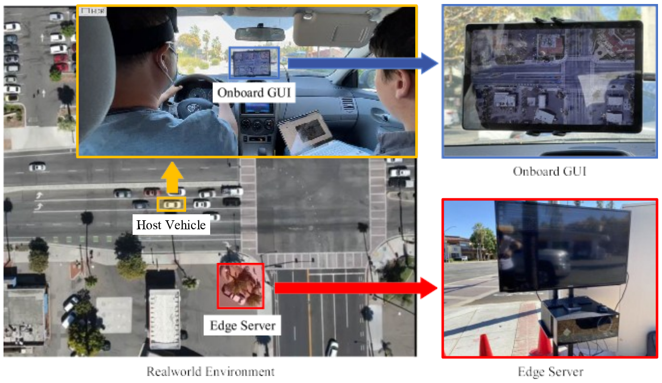


Fig. 5. Illustration of CMM field operational test from different views from a drone, host vehicle, onboard GUI, and edge server.

Meanwhile, by sharing feature-level data with corridor-level I-PNs, the CP system can generate object-level perception information with high perception accuracy to further assist road users or enhance traffic management [37].

C. Network-Level CP

As shown at the top of Fig. 4, network-level CP aims to perceive the object-level traffic condition for the whole traffic network. The cloud server is the core node to link all distributed intersections and CAVs that are out of the I-PN range. The cost-effective way for network-level CP is late fusion – retrieving perception information from I-PNs and CAVs and then merging those results for distribution. Furthermore, feature-level data can be also transmitted to the cloud server and a unified detector can be designed to generate the perception results.

It is noted that the main purpose of the HCP framework is to explore a high-level system design in which the cooperative perception could be implemented and integrated into various transportation scenarios seamlessly. Following this framework, several studies have been conducted to enable cooperative perception from different perspectives. For example, *PillarGrid* [14] is proposed to integrate heterogeneous sensing data from a vehicle PN and an infrastructure PN, which belongs to one of the key challenges in the intersection-level CP in our framework. Additionally, another key challenge in the intersection-level CP is the variety of communication capacity among perception nodes. To support CP under dynamic communication conditions, a dynamic feature-sharing strategy [125] is proposed to dynamically adjust feature sharing based on their allowable communication capacity. In addition, to improve perception accuracy after reducing the sharing features, *Pillar Attention Encoder* [126] is proposed to provide a strong representation of the sensor data. From the perspective of a vehicle PN in an intersection-level CP system, a case study is conducted to demonstrate how a vehicle PN can benefit from other vehicle PNs and infrastructure PNs [127].

To scale up the system, *VINet* [71] is proposed to handle the heterogeneity and scalability of the CP system at both the intersection level and corridor level by designing a two-stream neural network. Besides, VINet introduces a lightweight encoder design to alleviate the computation requirement for the whole CP system.

A real-world operational system (*Cyber Mobility Mirror*, shown in Fig. 5) is developed by using infrastructure-to-cloud and cloud-to-vehicle cooperation, as shown in the network-level CP above.

Following all these studies, we believe that the HCP framework can provide more inspiring ideas for other researchers in this multidiscipline field.

VII. Open-Source Datasets and Platforms

This section aims to provide some open-source datasets and platforms that can support the development of cooperative perception. We hope this section can help researchers expedite the onboarding processes for conducting their own research in this field.

A. Open-Source CP Datasets

Owing to prevailing needs in automated driving for surrounding perception, most real-world datasets for object detection and tracking are collected from onboard sensors from the perspective of a single PN, such as *KITTI* [48], *NuScenes* [128] and *Waymo Open Dataset* [93]. Training CP models usually requires datasets collected from multi- PN systems, which are missing in the early stage of CP research. To train the multi-PN CP models, researchers came out with ideas to emulate multi-PN datasets from the single- PN datasets [13], [20] by aligning data frames collected at different times.

However, it is nearly impossible for such a synthetic dataset to fully represent a real multi-PN dataset. As shown in Fig. 1, before 2022, there is no available open-sourced cooperative perception dataset collected from real-world data. To move forward, researchers tried to collect multi-PN datasets from high-fidelity 3D simulators, such as *CARLA* [65]. Empowered by advanced 3D modeling and graphic computing power, these simulators can generate vivid scenarios with nearly realistic sensor outputs. Due to the cost-effectiveness and high fidelity, multi-PN datasets were collected quickly and significantly expedited the development of CP methods, such as *OPV2V* [85] for supporting V2V-based CP models and *V2X-Sim* [129] for enabling CP models considering infrastructure-based sensors.

In 2022, *DAIR-V2X* [130], the first real-world cooperative perception dataset came to the stage, which is a large-scale, multi-node, multi-modality CP dataset. Specifically, *DAIR-V2X* contains 39*k* images, 39*k* PCD frames, and 10 classes of ground truth labels with synchronized time stamps. Sensor measurements are collected from both vehicle nodes and infrastructure nodes. One year later, as an upgraded version of the previous dataset, *V2X-Seq* [131] was published which includes data frames, trajectories, vector maps, and traffic lights captured from natural scenery to support V2X-based cooperative perception and forecasting tasks.

Concurrently, the *V2V4Real* [115] dataset was published for enabling the V2V-based CP tasks, which was collected by two vehicles simultaneously providing multi-view sensor data streams including 410 km of the driving area, 20K LiDAR data frames, 40K RGB camera frames, and 240K annotated 3D bounding boxes across 5 vehicle classes.

B. Open-Source CP Platforms

Instead of collecting datasets, various CP platforms are developed to support the development of CP methods. For CP model training and testing, *OpenCOOD* platform [85] provides a high-level codebase to support the design and benchmark of CP models for both simulation datasets and real-world datasets. For dataset flexibility, *CARTI* platform [132] is developed to enable researchers to customize their own cooperative perception scenarios and collect the customized dataset for training and testing the CP models.

In recent few years, several platforms were developed to enable the development of the CP model as well as its subsequent tasks, such as decision-making, planning, control, etc, which ends up with the CDA system mentioned earlier. For instance, *AutoCastSim* [133] was developed not only to support sensor data sharing and fusion for CP problems but also to enable low-level tasks such as vehicle control. *OpenCDA* [134] and *CARMA* [135] were developed to provide comprehensive capabilities to enable full-stack CDA system development.

VIII. Discussion

Although cooperative perception is an emerging research area, it is playing an increasingly significant role in promoting the perception capabilities for CDA applications. Many studies have been conducted to lay the foundation and provide inspiration for future work. In this section, we present our insights concerning the current states, open problems, and future trends in cooperative perception for CDA applications.

A. Current States and Open Challenges

1) Perception Singleton for Heterogeneity: The most common perception agents in transportation are intelligent vehicles and smart infrastructure which can be regarded as heterogeneous perception singletons. Since roadside sensors have more flexible locations and pose for data acquisition, one typical way of cooperative perception is to transmit information from the infrastructure side to road users [18], [41], [42], [60]. From the perspective of cooperative automated driving, V2V-based cooperative perception is also a promising solution to enable the ego-vehicle with the capability of *seeing through* [73], [74], [75], [77].

However, none of them can make an epochal revolution if they do not cooperate together in a deep manner, because the evolution of intelligent transportation systems is always highly coupled with the cooperation between vehicles and infrastructures [136]. Due to the heterogeneity of the perception singleton, only recently few studies have considered the cooperation between vehicle nodes and infrastructure nodes [14], [16]. Thus, vehicle-infrastructure cooperation is one of the most significant opening tasks for cooperative perception.

2) Sensor System for Fidelity: Generally speaking, the capability of the sensor system can be regarded as the foundation of subsequent applications in intelligent transportation systems. Since the perception data generated from sensor systems is the foundation of the downstream modules, such as prediction, decision-making, and actuation [37], for cooperative perception, cameras and LiDAR are widely applied to accessing high-fidelity sensing information.

However, in most research, these two types of high-fidelity sensors work separately – a cooperative perception system only equipped with one kind of sensor – such as multicamera- based CP [17], [21] and multi-LiDAR-based CP [14], [82]. According to the analysis in Section IV-B, fusing data from complementary sensors tends to significantly improve the object perception performance, such as Camera+LiDAR [95], Camera+RADAR [112], etc. Thus, developing multi-modality sensors for cooperative perception is an important way to improve the overall fidelity of the perception results.

Moreover, while infrastructure is a crucial component of cooperative perception systems, the existing perception methods employed by roadside sensors largely rely on general perception approaches designed for onboard sensors. Comparing the methods reviewed in Section IV and Section III, there is an evident gap between general object perception and cooperative perception. For instance, the core methodologies of a large portion of the existing roadside LiDAR-based detection approaches are based on DBSCAN for clustering [51], [57], [61], [63], [64], which has a performance gap compared with the SOTA methods [67], [70]. However, due to differences in sensor data distributions between roadside systems and onboard systems, datasets collected dedicated to roadside sensors are crucial for training roadside perception models. Additionally, to make use of the massive amount of onboard sensor-based datasets for roadside perception training, investigating the transferability of models that can be trained on the onboard datasets but implemented on roadside datasets is another key challenge to the improvement of the I-PN-based CP system.

Additionally, to the best of authors' knowledge, adverse environmental conditions are still lacking consideration for CP research. Thus, to improve the robustness of the sensing system is still an open challenge for CP systems to be able to be implemented in real-world conditions.

3) Fusion Strategies for Generality: As reviewed in Section V, different fusion schemes have their specific advantages and disadvantages. Early fusion-based studies mainly require high-speed communication to enable the transmission of raw data [20], [21]. However, the reliance on raw data inevitably makes the perception model very sensitive, and small communication errors or synchronization issues can cause significant degradation in system performance [16]. Late fusion-based research has been widely applied to various kinds of cooperative perception tasks since decades ago [21], [72], [84]. Late fusion has less requirement for communication but its performance also suffers from the merging of the object proposals from multiple sources [21].

To solve the issues mentioned above, recent work has been focusing on transmitting and fusing feature-level data to gain better accuracy with higher robustness [14], [16]. However, due to the deeply coupled feature and model complexity, large- scale extension is an inevitable challenge for intermediate fusion-based cooperative perception.

4) Policies for Sustainability: Since remarkable potential has been uncovered for the CP system to improve the current transportation systems, policies and standards have been formulated and released accordingly to stimulate and standardize the development of CP technology [5], [23], [24]. Based on the review and analysis, the development of CP tends to be a path that requires extensive investments for various topics such as sensors, communication systems, roadside infrastructures, etc.

However, current research studies and industrial standards mainly focus on the technical advancement of CP development while lacking consideration of the economic challenges that come with it. Based on the real-world CP demonstration from Federal Highway Administration (FHWA) [34], implementing the CP system in the real world requires multilateral efforts such as industrial solutions for sensors and communication, policy support from the local government, or transportation agencies, and numerous funding support for system expansion and maintenance. Meanwhile, public concerns (such as privacy issues) also need to be considered. Hence, it is a key challenge to make proper policies to push the development of CP in a sustainable way.

B. Future Trends

1) Towards Heterogeneous Cooperation: Physical occlusion is considered one of the unavoidable obstacles to single-node perception, and perceiving the environment from multiple nodes can mitigate such limitations. Given that transportation is a system of systems, vehicle-infrastructure cooperation is a promising solution to many existing traffic-related issues. More specifically, vehicle-infrastructure cooperative perception can leverage the capabilities of both vehicles (as mobile perception nodes with lightweight processing power) and infrastructure (as fixed nodes but with powerful processing/storage units) to achieve much better performance. Efficient and dynamic ways to fuse the information from vehicles with infrastructures are the keys to unlocking a new era of perception for CDA.

2) Towards Multi-Modal Cooperation: A multi-sensor-based perception system has the potential to improve perceived performance by taking advantage of complementary sensor data [137] with appropriate fusion techniques. In the scope of camera and LiDAR sensors, the development of current multi-modal sensor fusion is mainly targeting general object perception by multiple sensors equipped on one single agent [95]. Specific multi-modal sensor fusion for multiple perception nodes is still a blank field, which is, however, an important way to improve the perception accuracy for the whole system.

3) Towards Scalable Cooperation: The concept of cooperative perception is never intended to be only applied to a small number of nodes, such as two vehicles [13] or one vehicle with one infrastructure [14]. Some cooperative perception methods are mainly designed for enhancing the ego-vehicle with the assistance of surrounding nodes by asking surrounding nodes to align their data based on the metadata from the ego-vehicle [16], which may cause scalability issues when numerous ego-vehicles are involved.

On the other hand, the computational power and perceptive range of perception nodes are not the same for vehicles and infrastructure. An infrastructure-based perception system is more flexible in terms of sensor equipment and capable of empowering high-computational edge processors, large data storage and wide communication bandwidth. Although the onboard device has made major strides in development recently, it could be extremely costly and energy-inefficient to empower every vehicle with a high-performance computational system for enabling CP. Therefore, by only deploying lightweight-computing modules on the V-PN side (e.g., sharing data extraction) and leaving the heavy computing parts to the I-PN side (e.g., the backbone neural network), it can be more cost-effective to 1) enable intermediate fusion-based CP approach [71] and 2) implement the CP system in real-world situations for a broader range of perceptions [18].

Considering the issues for cooperative perception in real-world development, such as scalability, dynamic environment, and heterogeneous resources (such as computational power, storage space, and communication bandwidth), the hierarchical structure, including vehicle, infrastructure, and cloud, introduced in Section VI can be a promising solution. Thus, building a unified framework will be a systematic challenge and can lay a solid foundation for further research on cooperative perception. In the meantime, communication-oriented CP [138], [139] is also a critical direction for pushing CP technologies toward real-world implementations.

4) Towards Sustainable Cooperation: To implement the CP system in real-world conditions, multilateral efforts are required, which include automakers, policymakers, industrial societies, local transportation agencies, the general public, etc. Meanwhile, due to the sophisticated system architecture that involves vehicles, infrastructures, communication, sensors, and computing systems, the development of the CP system requires careful consideration of challenges from economic effects, liability issues, security concerns, public policies, etc.

Thus, to make the development of CP feasible and sustainable, a critical future direction for CP is to make policies, strategies and standards based on comprehensive consideration of multilateral interests, such as Infrastructure as a Service (IaaS) or public-private partnership (P3) mode. Although we recognize this challenge, this paper aims to raise this concept and arouse further discussion to work this out together.

Meanwhile, AV-related aspects, such as 1) vehicular communication [140], 2) vehicle self-pose estimation [141], 3) V2X data synchronization [142], 4) vehicular control [143], and 5) cyber security for sharing safety-critical sensor data [144], are also significant future directions for making CP happen and sustainably evolve in the real world, but are not deeply investigated in this paper due to the limited structure and space of this paper.

IX. Conclusion

This paper provides a comprehensive overview and proposes a hierarchical framework for cooperative perception. The architecture and taxonomy are presented to illustrate the fundamental components and core aspects of a cooperative perception system. Cooperative perception methods are then introduced with detailed literature reviews from three perspectives: node structure, sensing modality, and fusion scheme. The proposed hierarchical cooperative perception framework is analyzed from the levels of intersection, corridor, and network respectively. Existing datasets and simulators for enabling cooperative perception are briefly reviewed to identify the gaps. Finally, this paper discusses current issues and future trends. To the best of our knowledge, this work is the first study to provide a unified framework for cooperative perception.

Acknowledgment

The contents of this article reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. The contents do not necessarily reflect the official views of Toyota Motor North America.

References

[1] U.S. Dept. Transp. (2019). *Overview of Motor Vehicle Crashes in 2019*. [Online]. Available: https://crashstats.nhtsa.dot.gov/Api/Public/Publication/813060

[2] INRIX. (2018). *Inrix: Congestion Costs Each American 97 Hours,1,348 a Year*. [Online]. Available: https://inrix.com/press- releases/scorecard-2018-us/

[3] U.S. Department of Energy. (2020). *Fotw1204: Fuel Wasted Due to U.s. Traffic Congestion in 2020 Cut in Half From 2019 to 2020*. [Online]. Available: https://www.energy.gov/eere/vehicles/articles/fotw-1204-Sep.-20-2021-fuel-wasted-due-us-traffic-congestion-2020-cut-half

[4] D. J. Fagnant and K. Kockelman, "Preparing a nation for autonomous vehicles: Opportunities, barriers and policy recommendations," *Transp. Res. Part A, Policy Pract.*, vol. 77, pp. 167–181, Jul. 2015.

[5] Society of Automotive Engineers (SAE). (2021). *Taxonomy and Definitions for Terms Related To Cooperative Driving Automation for On-road Motor Vehicles*. [Online]. Available: https://www.sae.org/standards/content/j3216_202107

[6] J. Wu, H. Xu, Y. Zhang, and R. Sun, "An improved vehicle-pedestrian near-crash identification method with a roadside LiDAR sensor," *J. Saf. Res.*, vol. 73, pp. 211–224, Jun. 2020.

[7] Z. Bai, P. Hao, W. ShangGuan, B. Cai, and M. J. Barth, "Hybrid reinforcement learning-based eco-driving strategy for connected and automated vehicles at signalized intersections," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 9, pp. 15850–15863, Sep. 2022.

[8] Z. Wang, Y. Bian, S. E. Shladover, G. Wu, S. E. Li, and M. J. Barth, "A survey on cooperative longitudinal motion control of multiple connected and automated vehicles," *IEEE Intell. Transp. Syst. Mag.*, vol. 12, no. 1, pp. 4–24, Sep. 2020.

[9] L. Liu et al., "Deep learning for generic object detection: A survey," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 261–318, Feb. 2020.

[10] E. Arnold, O. Y. Al-Jarrah, M. Dianati, S. Fallah, D. Oxtoby, and A. Mouzakitis, "A survey on 3D object detection methods for autonomous driving applications," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 10, pp. 3782–3795, Oct. 2019.

[11] A. Manjunath, Y. Liu, B. Henriques, and A. Engstle, "Radar based object detection and tracking for autonomous driving," in *IEEE MTT-S Int. Microwave Symp. Dig.*, Apr. 2018, pp. 1–4.

[12] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," 2019, *arXiv:1905.05055*.

[13] 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.*, Nov. 2019, pp. 88–100.

[14] Z. Bai, G. Wu, M. J. Barth, Y. Liu, E. Akin Sisbot, and K. Oguchi, "PillarGrid: Deep learning-based cooperative perception for 3D object detection from onboard-roadside LiDAR," 2022, *arXiv:2203.06319*.

[15] E. Thonhofer et al., "Infrastructure-based digital twins for cooperative, connected, automated driving and smart road services," *IEEE Open J. Intell. Transp. Syst.*, vol. 4, pp. 311–324, 2023.

[16] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2X-ViT: Vehicle-to-everything cooperative perception with vision transformer," 2022, *arXiv:2203.10638*.

[17] J. Zhu, X. Li, P. Jin, Q. Xu, Z. Sun, and X. Song, "MME-YOLO: Multi-sensor multi-level enhanced YOLO for robust vehicle detection in traffic surveillance," *Sensors*, vol. 21, no. 1, p. 27, Dec. 2020.

[18] Z. Bai et al., "Cyber mobility mirror: A deep learning-based real- world object perception platform using roadside LiDAR," 2022, *arXiv:2202.13505*.

[19] W. Liu, S. Muramatsu, and Y. Okubo, "Cooperation of V2I/P2I communication and roadside radar perception for the safety of vulnerable road users," in *Proc. 16th Int. Conf. Intell. Transp. Syst. Telecommun. (ITST)*, Oct. 2018, pp. 1–7.

[20] 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. (ICDCS)*, Jul. 2019, pp. 514–524.

[21] 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.

[22] 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.

[23] *Intelligent Transport Systems (Its); Vehicular Communications; Basic Set of Applications; Analysis of the Collective Perception Service (CPS); Release 2*, Standard ETSI 103 562 v2. 1.1, 2019.

[24] *Cooperative Intelligent Transportation System—Vehicular Communication Application Layer Specification and Data Exchange Standard (phase Ii)*, Standard T/CSAE 157-2020, Chin. Soc. Automot. Eng., 2020.

[25] H.-J. Gunther, B. Mennenga, O. Trauer, R. Riebl, and L. Wolf, "Realizing collective perception in a vehicle," in *Proc. IEEE Veh. Netw. Conf. (VNC)*, Dec. 2016, pp. 1–8.

[26] 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.

[27] S. of Automotive Engineers (SAE). (2021). *Taxonomy and Definitions for Terms Related To Driving Automation Systems for On-road Motor Vehicles*. [Online]. Available: https://www.sae.org/standards/content/j3016_202104/

[28] K. Nellore and G. Hancke, "A survey on urban traffic management system using wireless sensor networks," *Sensors*, vol. 16, no. 2, p. 157, Jan. 2016.

[29] S. Y. Cheung, S. C. Ergen, and P. Varaiya, "Traffic surveillance with wireless magnetic sensors," in *Proc. 12th ITS world Congr.*, 2005, pp. 173–181.

[30] B. Coifman, D. Beymer, P. McLauchlan, and J. Malik, "A real-time computer vision system for vehicle tracking and traffic surveillance," *Transp. Res. Part C, Emerg. Technol.*, vol. 6, no. 4, pp. 271–288, Aug. 1998.

[31] E. Akleman, "Deep learning," *Computer*, vol. 53, no. 9, pp. 1–17, Sep. 2020.

[32] S. Zhang, J. Chen, F. Lyu, N. Cheng, W. Shi, and X. Shen, "Vehicular communication networks in the automated driving era," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 26–32, Sep. 2018.

[33] K. C. Dey, A. Rayamajhi, M. Chowdhury, P. Bhavsar, and J. Martin, "Vehicle-to-vehicle (V2V) and vehicle-to-infrastructure (V2I) communication in a heterogeneous wireless network—Performance evaluation," *Transp. Res. Part C: Emerg. Technol.*, vol. 68, pp. 168–184, Jul. 2016.

[34] Y. Lou et al., "Cooperative automation research: Carma proof-of-concept TSMO use case testing: Carma cooperative perception concept of operations," Leidos, Inc., Reston, VA, USA, Tech. Rep. FHWA-HRT-22-062, 2022.

[35] H. Qiu, F. Ahmad, F. Bai, M. Gruteser, and R. Govindan, "AVR: Augmented vehicular reality," in *Proc. 16th Annu. Int. Conf. Mobile Syst., Appl., Services*, 2018, pp. 81–95.

[36] O. D. Altan, G. Wu, M. J. Barth, K. Boriboonsomsin, and J. A. Stark, "GlidePath: Eco-friendly automated approach and departure at signalized intersections," *IEEE Trans. Intell. Vehicles*, vol. 2, no. 4, pp. 266–277, Dec. 2017.

[37] Z. Bai, G. Wu, X. Qi, Y. Liu, K. Oguchi, and M. J. Barth, "Cyber mobility mirror for enabling cooperative driving automation in mixed traffic: A co-simulation platform," 2022, *arXiv:2201.09463*.

[38] M. Shan et al., "Demonstrations of cooperative perception: Safety and robustness in connected and automated vehicle operations," *Sensors*, vol. 21, no. 1, p. 200, Dec. 2020.

[39] 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, Jul. 2021, Art. no. 100057.

[40] R. Ojala, J. Vepsäläinen, J. Hanhirova, V. Hirvisalo, and K. Tammi, "Novel convolutional neural network-based roadside unit for accurate pedestrian localisation," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 9, pp. 3756–3765, Sep. 2020.

[41] 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.

[42] N. Balamuralidhar, S. Tilon, and F. Nex, "MultEYE: Monitoring system for real-time vehicle detection, tracking and speed estimation from UAV imagery on edge-computing platforms," *Remote Sens.*, vol. 13, no. 4, p. 573, Feb. 2021.

[43] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[44] A. Bochkovskiy, C.-Y. Wang, and H.-Y. Mark Liao, "YOLOv4: Optimal speed and accuracy of object detection," 2020, *arXiv:2004.10934*.

[45] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2544–2550.

[46] E. Çiçek and S. Gören, "Fully automated roadside parking spot detection in real time with deep learning," *Concurrency Computation, Pract. Exper.*, vol. 33, no. 23, Dec. 2021, Art. no. e6006.

[47] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[48] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? The KITTI vision benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 3354–3361.

[49] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 28, 2015, pp. 91–99.

[50] T. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[51] J. Zhao, H. Xu, H. Liu, J. Wu, Y. Zheng, and D. Wu, "Detection and tracking of pedestrians and vehicles using roadside LiDAR sensors," *Transp. Res. Part C, Emerg. Technol.*, vol. 100, pp. 68–87, Mar. 2019.

[52] J. Wu, H. Xu, and J. Zheng, "Automatic background filtering and lane identification with roadside LiDAR data," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Yokohama, Japan, Oct. 2017, pp. 1–6.

[53] M. Ester, "A density-based algorithm for discovering clusters in large spatial databases with noise," *Kdd*, vol. 96, no. 34, pp. 226–231, 1996.

[54] J. Li, J.-H. Cheng, J.-yY. Shi, and F. Huang, "Brief introduction of back propagation (BP) neural network algorithm and its improvement," *Adv. Comput. Sci. Inf. Eng.*, vol. 1, pp. 553–558, Sep. 2012.

[55] G. Welch and G. Bishop, "An introduction to the Kalman filter," Dept. Comput. Sci., Univ. North Carolina at Chapel Hill, Chapel Hill, NC, USA, Tech. Rep. TR 95-041, 1995.

[56] Y. Cui, H. Xu, J. Wu, Y. Sun, and J. Zhao, "Automatic vehicle tracking with roadside LiDAR data for the connected-vehicles system," *IEEE Intell. Syst.*, vol. 34, no. 3, pp. 44–51, May 2019.

[57] J. Zhang, W. Xiao, B. Coifman, and J. P. Mills, "Vehicle tracking and speed estimation from roadside lidar," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 5597–5608, 2020.

[58] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proc. IEEE*, vol. 92, no. 3, pp. 401–422, Mar. 2004.

[59] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Syst. Mag.*, vol. 29, no. 6, pp. 82–100, Dec. 2009.

[60] L. Zhang, J. Zheng, R. Sun, and Y. Tao, "GC-Net: Gridding and clustering for traffic object detection with roadside LiDAR," *IEEE Intell. Syst.*, vol. 36, no. 4, pp. 104–113, Jul. 2021.

[61] Y. Song, H. Zhang, Y. Liu, J. Liu, H. Zhang, and X. Song, "Background filtering and object detection with a stationary LiDAR using a layer-based method," *IEEE Access*, vol. 8, pp. 184426–184436, 2020.

[62] M. Gouda, B. Arantes de Achilles Mello, and K. El-Basyouny, "Automated object detection, mapping, and assessment of roadside clear zones using LiDAR data," *Transp. Res. Record: J. Transp. Res. Board*, vol. 2675, no. 12, pp. 432–448, Dec. 2021.

[63] Z. Zhang, J. Zheng, X. Wang, and X. Fan, "Background filtering and vehicle detection with roadside LiDAR based on point association," in *Proc. 37th Chin. Control Conf. (CCC)*, Jul. 2018, pp. 7938–7943.

[64] Z. Zhang, J. Zheng, H. Xu, X. Wang, X. Fan, and R. Chen, "Automatic background construction and object detection based on roadside LiDAR," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 10, pp. 4086–4097, Oct. 2020.

[65] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "CARLA: An open urban driving simulator," in *Proc. 1st Annu. Conf. Robot Learn.*, vol. 78, Nov. 2017, pp. 1–16.

[66] M. Simony, S. Milzy, K. Amendey, and H.-M. Gross, "Complex-YOLO: An euler-region-proposal for real-time 3D object detection on point clouds," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 1–28.

[67] 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. (CVPR)*, Jun. 2019, pp. 12697–12705.

[68] B. Veeramani, J. W. Raymond, and P. Chanda, "DeepSort: Deep convolutional networks for sorting haploid maize seeds," *BMC Bioinf.*, vol. 19, no. S9, pp. 1–9, Aug. 2018.

[69] C. Glennie and D. D. Lichti, "Static calibration and analysis of the velodyne HDL-64E S2 for high accuracy mobile scanning," *Remote Sens.*, vol. 2, no. 6, pp. 1610–1624, Jun. 2010.

[70] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4490–4499.

[71] Z. Bai, G. Wu, M. J. Barth, Y. Liu, E. A. Sisbot, and K. Oguchi, "VINet: Lightweight, scalable, and heterogeneous cooperative perception for 3D object detection," 2022, *arXiv:2212.07060*.

[72] L. Merino, F. Caballero, J. R. Martínez-de Dios, J. Ferruz, and A. Ollero, "A cooperative perception system for multiple UAVs: Application to automatic detection of forest fires," *J. Field Robot.*, vol. 23, nos. 3–4, pp. 165–184, Mar. 2006.

[73] M. Rockl, T. Strang, and M. Kranz, "V2V communications in automotive multi-sensor multi-target tracking," in *Proc. IEEE 68th Veh. Technol. Conf.*, Sep. 2008, pp. 1–5.

[74] 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. Vehicles Symp.*, Jun. 2012, pp. 270–275.

[75] Z. Y. Rawashdeh and Z. Wang, "Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information," in *Proc. 21st Int. Conf. Intell. Transp. Syst. (ITSC)*, Nov. 2018, pp. 3961–3966.

[76] Z. Xiao, Z. Mo, K. Jiang, and D. Yang, "Multimedia fusion at semantic level in vehicle cooperactive perception," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops*, Jul. 2018, pp. 1–26.

[77] S.-W. Kim et al., "Multivehicle cooperative driving using cooperative perception: Design and experimental validation," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 663–680, Apr. 2015.

[78] Y. Yan, Y. Mao, and B. Li, "SECOND: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, Oct. 2018.

[79] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[80] 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.*, Aug. 2020, pp. 1–7.

[81] I. Goodfellow et al., "Maxout networks," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1319–1327.

[82] 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. Eur. Conf. Comput. Vis.*, 2020, pp. 605–621.

[83] S. Manivasagam et al., "LiDARsim: Realistic LiDAR simulation by leveraging the real world," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1–24.

[84] Z. Zhang, S. Wang, Y. Hong, L. Zhou, and Q. Hao, "Distributed dynamic map fusion via federated learning for intelligent networked vehicles," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Jun. 2021, pp. 953–959.

[85] R. Xu, H. Xiang, X. Xia, X. Han, J. Li, and J. Ma, "OPV2V: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," 2021, *arXiv:2109.07644*.

[86] A. Vaswani et al., "Attention is all you need," in *Proc. Adv. Neural Inform. Process. Syst. (NIPS)*, 2017, pp. 5998–6008.

[87] A. D. Uszkoreit et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–25.

[88] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *Proc. Web Conf.*, New York, NY, USA, Apr. 2020, pp. 2704–2710, doi: 10.1145/3366423.3380027.

[89] S. Campbell et al., "Sensor technology in autonomous vehicles : A review," in *Proc. 29th Irish Signals Syst. Conf. (ISSC)*, Jun. 2018, pp. 1–4.

[90] A. Ziebinski, R. Cupek, H. Erdogan, and S. Waechter, "A survey of ADAS technologies for the future perspective of sensor fusion," in *Proc. Int. Conf. Comput. Collective Intell.* Cham, Switzerland: Springer, 2016, pp. 135–146.

[91] S. Tokoro, K. Kuroda, A. Kawakubo, K. Fujita, and H. Fujinami, "Electronically scanned millimeter-wave radar for pre-crash safety and adaptive cruise control system," in *IEEE Intell. Vehicles Symp. Proc.*, Jul. 2003, pp. 304–309.

[92] E. Santana and G. Hotz, "Learning a driving simulator," 2016, *arXiv:1608.01230*.

[93] P. Sun et al., "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jul. 2020, pp. 1–15.

[94] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[95] Z. Liu et al., "BEVFusion: Multi-task multi-sensor fusion with unified Bird's-eye view representation," 2022, *arXiv:2205.13542*.

[96] A. S. Olagoke, H. Ibrahim, and S. S. Teoh, "Literature survey on multi-camera system and its application," *IEEE Access*, vol. 8, pp. 172892–172922, 2020.

[97] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern Recognit. Lett.*, vol. 34, no. 1, pp. 3–19, Jan. 2013.

[98] R. Eshel and Y. Moses, "Homography based multiple camera detection and tracking of people in a dense crowd," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[99] I. Mikic, S. Santini, and R. Jain, "Video processing and integration from multiple cameras," in *Proc. Image Understand. Workshop*, vol. 6, 1998, pp. 1–20.

[100] K. Kim and L. S. Davis, "Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," in *Proc. Eur. Conf. Comput. Vis.*, 2006, pp. 98–109.

[101] B. Song and A. K. Roy-Chowdhury, "Robust tracking in a camera network: A multi-objective optimization framework," *IEEE J. Sel. Topics Signal Process.*, vol. 2, no. 4, pp. 582–596, Aug. 2008.

[102] T. Huang and S. Russell, "Object identification in a Bayesian context," in *Proc. IJCAI*, vol. 97, 1997, pp. 1276–1282.

[103] K.-W. Chen, C.-C. Lai, Y.-P. Hung, and C.-S. Chen, "An adaptive learning method for target tracking across multiple cameras," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[104] Y. Liu et al., "CBNet: A novel composite backbone network architecture for object detection," in *Proc. AAAI Conf. Artif. Intell.*, Apr. 2020, vol. 34, no. 7, pp. 11653–11660.

[105] A. Farhadi and J. Redmon, "YOLOv3: An incremental improvement," in *Computer Vision and Pattern Recognition*. Cham, Switzerland: Springer, 2018, pp. 1804–2767.

[106] L. Wang et al., "Multi-modal 3D object detection in autonomous driving: A survey and taxonomy," *IEEE Trans. Intell. Vehicles*, vol. 1, no. 1, pp. 1–19, Jun. 2023.

[107] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "PointPainting: Sequential fusion for 3D object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4604–4612.

[108] Z. Wang, Y. Wu, and Q. Niu, "Multi-sensor fusion in automated driving: A survey," *IEEE Access*, vol. 8, pp. 2847–2868, 2020.

[109] F. Garcia, P. Cerri, A. Broggi, A. de la Escalera, and J. M. Armingol, "Data fusion for overtaking vehicle detection based on radar and optical flow," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2012, pp. 494–499.

[110] K.-E. Kim, C.-J. Lee, D.-S. Pae, and M.-T. Lim, "Sensor fusion for vehicle tracking with camera and radar sensor," in *Proc. 17th Int. Conf. Control, Autom. Syst. (ICCAS)*, Oct. 2017, pp. 1075–1077.

[111] B. Lagos-Álvarez, L. Padilla, J. Mateu, and G. Ferreira, "A Kalman filter method for estimation and prediction of space–time data with an autoregressive structure," *J. Stat. Planning Inference*, vol. 203, pp. 117–130, Dec. 2019.

[112] R. Nabati and H. Qi, "CenterFusion: Center-based radar and camera fusion for 3D object detection," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 1527–1536.

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

[114] Y. Kim, S. Kim, J. W. Choi, and D. Kum, "CRAFT: Camera-radar 3D object detection with spatio-contextual fusion transformer," in *Proc. AAAI Conf. Artif. Intell.*, Jun. 2023, vol. 37, no. 1, pp. 1160–1168.

[115] R. Xu et al., "V2V4Real: A real-world large-scale dataset for vehicle-to-vehicle cooperative perception," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1–18.

[116] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, 2006, pp. 850–855.

[117] Y. Liu et al., "FedVision: An online visual object detection platform powered by federated learning," in *Proc. AAAI Conf. Artif. Intell.*, vol. 34, 2020, pp. 13172–13179.

[118] J. Guo et al., "CoFF: Cooperative spatial feature fusion for 3-D object detection on autonomous vehicles," *IEEE Internet Things J.*, vol. 8, no. 14, pp. 11078–11087, Jul. 2021.

[119] Y. Li, S. Ren, P. Wu, S. Chen, C. Feng, and W. Zhang, "Learning distilled collaboration graph for multi-agent perception," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, 2021, pp. 1–20.

[120] Y. Yuan, H. Cheng, and M. Sester, "Keypoints-based deep feature fusion for cooperative vehicle detection of autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 7, no. 2, pp. 3054–3061, Apr. 2022.

[121] G. Luo, H. Zhang, Q. Yuan, and J. Li, "Complementarity-enhanced and redundancy-minimized collaboration network for multi-agent perception," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 3578–3586.

[122] 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. Vehicles*, vol. 1, no. 1, pp. 1–13, Jul. 2023.

[123] 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*.

[124] F. Poletti et al., "Towards high-capacity fibre-optic communications at the speed of light in vacuum," *Nature Photon.*, vol. 7, no. 4, pp. 279–284, Apr. 2013.

[125] Z. Bai, G. Wu, M. J. Barth, Y. Liu, E. A. Sisbot, and K. Oguchi, "Dynamic feature sharing for cooperative perception from point clouds," in *Proc. IEEE 26th Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2023, pp. 3970–3976.

[126] Z. Bai et al., "Pillar attention encoder for adaptive cooperative perception," *IEEE Internet Things J.*, vol. 11, no. 14, pp. 24998–25009, Jul. 2024.

[127] Z. Bai, J. G. Escobar, G. Wu, and M. J. Barth, "Object perception framework for connected and automated vehicles: A case study," in *Proc. IEEE Transp. Electrific. Conf. Expo. (ITEC)*, Jun. 2023, pp. 1–5.

[128] H. Caesar et al., "nuScenes: A multimodal dataset for autonomous driving," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11621–11631.

[129] Y. Li et al., "V2X-sim: Multi-agent collaborative perception dataset and benchmark for autonomous driving," *IEEE Robot. Autom. Lett.*, vol. 7, no. 4, pp. 10914–10921, Oct. 2022.

[130] 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. (CVPR)*, Jun. 2022, pp. 21361–21370.

[131] H. Yu et al., "V2X-seq: A large-scale sequential dataset for vehicle-infrastructure cooperative perception and forecasting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1–18.

[132] Z. Bai. (2022). *Carti Dataset for Cooperative Perceptiion*. [Online]. Available: https://github.com/zwbai/CARTI_Dataset

[133] H. Qiu, P.-H. Huang, N. Asavisanu, X. Liu, K. Psounis, and R. Govindan, "AutoCast: Scalable infrastructure-less cooperative perception for distributed collaborative driving," in *Proc. 20th Annu. Int. Conf. Mobile Syst., Appl. Services*, Jun. 2022, pp. 1–18.

[134] R. Xu, Y. Guo, X. Han, X. Xia, H. Xiang, and J. Ma, "OpenCDA: An open cooperative driving automation framework integrated with co-simulation," in *Proc. IEEE Int. Intell. Transp. Syst. Conf. (ITSC)*, Sep. 2021, pp. 1155–1162.

[135] Federal Highway Administration (FHW). (2021). *Carma*. [Online]. Available: https://highways.dot.gov/tags/carma

[136] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 12, no. 4, pp. 1624–1639, Dec. 2011.

[137] 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, Oct. 2021.

[138] Y.-C. Liu, J. Tian, N. Glaser, and Z. Kira, "When2com: Multi-agent perception via communication graph grouping," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4106–4115.

[139] Y. Hu, S. Fang, Z. Lei, Y. Zhong, and S. Chen, "Where2comm: Communication-efficient collaborative perception via spatial confidence maps," 2022, *arXiv:2209.12836*.

[140] Y.-C. Liu, J. Tian, C.-Y. Ma, N. Glaser, C.-W. Kuo, and Z. Kira, "Who2com: Collaborative perception via learnable handshake communication," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 6876–6883.

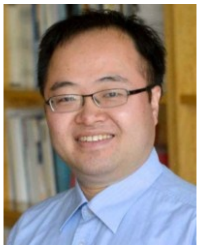
[141] H. Li and F. Nashashibi, "Multi-vehicle cooperative localization using indirect vehicle-to-vehicle relative pose estimation," in *Proc. IEEE Int. Conf. Veh. Electron. Saf. (ICVES)*, Jul. 2012, pp. 267–272.

[142] A. Petrillo, A. Salvi, S. Santini, and A. S. Valente, "Adaptive multi-agents synchronization for collaborative driving of autonomous vehicles with multiple communication delays," *Transp. Res. Part C, Emerg. Technol.*, vol. 86, pp. 372–392, Jan. 2018.

[143] J. Cui, H. Qiu, D. Chen, P. Stone, and Y. Zhu, "COOPERNAUT: End-to-end driving with cooperative perception for networked vehicles," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 17252–17262.

[144] X. Sun, F. R. Yu, and P. Zhang, "A survey on cyber-security of connected and autonomous vehicles (CAVs)," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 6240–6259, Jul. 2022.

Zhengwei Bai (Student Member, IEEE) received the B.E. and M.S. degrees from Beijing Jiaotong University, Beijing, China, in 2017 and 2020, respectively. He is currently pursuing the Ph.D. degree in electrical and computer engineering with the University of California at Riverside. His research focuses on object detection and tracking, cooperative perception, decision making, motion planning, and cooperative driving automation (CDA). He serves as a Review Editor for *Urban Transportation Systems and Mobility*.



Guoyuan Wu (Senior Member, IEEE) received the Ph.D. degree in mechanical engineering from the University of California at Berkeley in 2010. Currently, he is an Associate Researcher and an Associate Adjunct Professor with the Bourns College of Engineering, Center for Environmental Research & Technology (CECERT), and the Department of Electrical and Computer Engineering, University of California at Riverside. Develop- ment and evaluation of sustainable and intelligent transportation system (SITS) technologies, including connected and automated transportation systems (CATS), shared mobility, transportation electrification, optimization and control of vehicles, traffic simulation, and emissions measurement and modeling. He is also a member of the Vehicle-Highway Automation Standing Committee (ACP30) of the Transportation Research Board (TRB), a Board Member of Chinese Institute of Engineers Southern California Chapter (CIE-SOCAL), and a member of Chinese Overseas Transportation Association (COTA). He was a recipient of Vincent Bendix Automotive Electronics Engineering Award. He serves as an Associate Editor for a few journals, including IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, *SAE International Journal of Connected and Automated Vehicles*, and IEEE OPEN JOURNAL OF INTELLIGENT TRANSPORTATION SYSTEMS.

Matthew J. Barth (Fellow, IEEE) received the M.S. and Ph.D. degrees in electrical and computer engineering from the University of California at Santa Barbara, in 1985 and 1990, respectively. He is currently the Yeager Families Professor with the College of Engineering, University of California at Riverside, USA. He is also the Director of the Center for Environmental Research and Technology. His current research interests include ITS and the environment, transportation/emissions modeling, vehicle activity analysis, advanced navigation techniques, electric vehicle technology, and advanced sensing and control. He has been active in the IEEE Intelligent Transportation System Society for many years, serving as a Senior Editor for both IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS and IEEE TRANSACTIONS ON INTELLIGENT VEHICLES. He served as the IEEE ITSS President for 2014 and 2015 and is currently the IEEE ITSS Vice President of Education.

Yongkang Liu received the M.S. and Ph.D. degrees in electrical engineering from The University of Texas at Dallas in 2017 and 2021, respectively. He is currently a Research Engineer with Toyota Motor North America, InfoTech Laboratories. His current research interests are focused on in-vehicle systems and advancements in intelligent vehicle technologies.

Emrah Akin Sisbot (Member, IEEE) received the Ph.D. degree in robotics and artificial intelligence from Paul Sabatier University, Toulouse, France, in 2008. He was a Post-Doctoral Research Fellow with LAAS-CNRS, Toulouse, and the University of Washington, Seattle. He is currently a Principal Engineer with Toyota Motor North America, InfoTech Laboratories, Mountain View, CA, USA. His current research interests include real-time intelligent systems, robotics, and human–machine interaction.

Kentaro Oguchi received the M.S. degree in computer science from Nagoya University. He is currently the Director of Toyota Motor North America, InfoTech Laboratories. His team is responsible for creating intelligent connected vehicle architecture that takes advantage of novel AI technologies to provide real-time services to connected vehicles for smoother and efficient traffic, intelligent dynamic parking navigation, and vehicle guidance to avoid risks from anomalous drivers. His team also creates technologies to form a vehicular cloud using vehicle-to-everything technologies. He was a Senior Researcher with the Toyota Central Research and Development Laboratories, Japan.

Zhitong Huang is currently a Senior Transportation Research Scientist and the Analysis, Simulation, and Modeling Program Manager of Leidos. He has 17 years of research experience and conducted dozens of research projects in the field of transportation engineering. His main research interests include transportation simulation and modeling, connected and automated vehicle (CAV) systems, traffic operation and management, and digital twin.