

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

Zhengwei Bai¹, *Student Member, IEEE*, Guoyuan Wu¹, *Senior Member, IEEE*,

Matthew J. Barth¹, *Fellow, IEEE*, Yongkang Liu, Emrah Akin Sisbot, Kentaro Oguchi, Zhitong Huang

Abstract—Perceiving the environment is one of the most fundamental keys to enabling Cooperative Driving Automation (CDA), 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 inevitably 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, to the best of our knowledge, this is the first time to propose a unified CP framework. 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. Node structure, sensor modality, and fusion schemes are reviewed and analyzed with comprehensive literature to provide detailed explanations of specific methods. A Hierarchical CP framework is proposed, followed by a review of existing Datasets and Simulators to sketch an overall landscape of CP. 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 between 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 is attracting 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 the 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]. RADAR sensor has been an integral part of safety-critical applications in the automotive industry due to its robust performance in variable conditions [11].

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 the range and occlusion by other objects. 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 and sometimes large blind zones. Thus, neither onboard sensors nor infrastructure-based sensors alone can outbreak the physical limitations and achieve satisfactory perception performance.

Empowered by mobile connectivity, Connected Vehicles (CVs) and Connected and Automated Vehicles (CAVs) have the capability to grasp perception information from others who are equipped with perception systems and connectivity, such as smart infrastructures or other CAVs. It is reasonable to combine sensing information from spatially separated nodes

Corresponding Author: Zhengwei Bai, E-mail: zbai012@ucr.edu.

Zhengwei Bai, Guoyuan Wu, and Matthew J. Barth are with the Department of Electrical and Computer Engineering, the University of California at Riverside, Riverside, CA 92507 USA .

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

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

TABLE I
RELATIONSHIP BETWEEN CLASSES OF CDA COOPERATION AND LEVELS OF AUTOMATION [5].

		SAE Driving Automation (DA) Levels					
		Level 0: No DA	Level 1: Driver Assistance	Level2: Partial DA	Level3: 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 roadusers and operators		
	Class B: intent-sharing	e.g., Turn Signal	Limited Cooperation (only longitudinal OR alteral)	Limited Cooperation (both longitudinal AND alteral)	Improved C-ADS situational awareness through prediction reliability		
	Class C: agreement-sharing	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		

to overcome the occlusion or perception range. Thus, Cooperative Perception naturally attracts fast-increasing attention to Driving Automation. Many kinds of research have been conducted from different aspects, such as perception nodes (vehicle [13] or infrastructure [14]), sensor modalities (Camera [15] or Lidar [16]), and fusion schemes (early fusion [17], late fusion [18], or deep fusion [19]). Although a recent survey conducted by Caillot et al. [20] reviewed the cooperative perception in an automotive context, their focus is mainly on the ego-vehicle, such as localization, map generation, etc. Thus, a comprehensive overview of CP and a general CP framework for handling heterogeneity and scalability in mixed traffic are still missing.

In this paper, the CP-based object perception methods are reviewed, which aims to establish an overall landscape for cooperative perception based on different 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 different perspectives mentioned above and to provide inspiration for future research.

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 including node structure, sensor modality, and fusion scheme are reviewed in Section III to V, respectively, followed by the presentation of available Datasets and Simulators. The hierarchical cooperative perception framework is proposed and discussed in Section VI. Section VIII highlights the current challenges and future trends, followed by Section IX that concludes the paper.

II. ARCHITECTURE AND TAXONOMY

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* [21], 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], Cooperative Driving Automation (CDA) enables communication and cooperation between equipped vehicles, infrastructure, and other road users, which will, in turn, improve the safety, mobility, and sustainability of 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. Led by the Federal Highway Administration, the CARMA program [22] is one of the state-of-the-art (SOTA) projects that aim to support and enable research and testing for CDA. Based on the analysis of SAE standards and the CARMA program, the architecture and taxonomy of CP are proposed and described in the following sections.

A. Architecture

In cooperative driving automation (CDA), the fidelity and range of perception information have a significant impact on the system performance of subsequent cooperative maneuvers. Fig. 1 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) *Edge Processing*; 3) *Cloud Computing*; and 4) *Message Distribution*.

1) *Information Collection*: Collecting raw data of traffic information lays the foundation for downstream perception tasks. In the development of transportation, various kinds of sensors are implemented aiming at different tasks and scenarios. In the context of traffic surveillance, several traditional sensors are widely applied, such as *Loop Detectors* and *Microwave RADAR*, for dynamic traffic management [23]. However, the main capacity of these traditional sensors is to

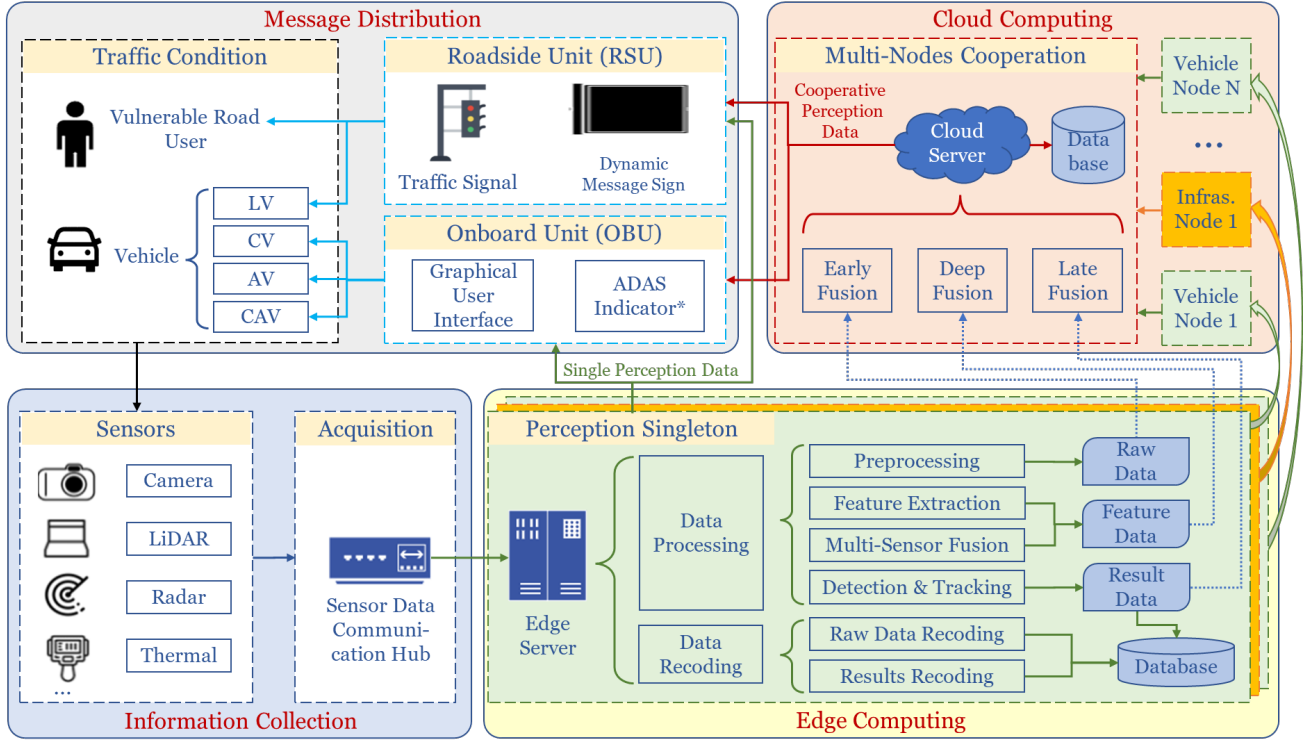


Fig. 1. Systematic architecture for cooperative perception system (*: Other non-visual driving advisory signals for Advanced driver-assistance systems (ADAS), such as audial, haptic, or even control commands.).

provide mesoscopic traffic information, such as traffic volume or queue length. To support cooperative driving automation, 3D object-level information is required, which can be generated from high-resolution sensors, such as cameras, LiDAR, etc.

Dated back to a couple of decades ago, due to the limitation of computational power and development of the computer vision field, high-resolution-sensors-based object perception is barely developed for intelligent transportation systems [24]. Although some vision-based methods are developed, very limited performance can be achieved [25]. Thanks to the quick advancement in high-performance computation and the surge of artificial intelligence (AI) [26], high-resolution sensors are able to provide object-level perception results, which can be equipped on vehicles or roadside infrastructures to perceive the environment and transmit collected data to its processing server via a communication hub for further processing.

2) *Edge Processing*: Unlike traditional traffic surveillance systems which do not need high-frequency and low-latency processing, CDA generally requires perception data with a minimal 1 – 10Hz frequency and time delay of less than 100ms [27]. Considering that using limited bandwidth to transmit a large volume of raw data (e.g., point cloud data) may cause an unacceptable time delay (especially in some safety-critical scenarios), information collected from sensors may be processed on edge servers equipped on vehicles or infrastructures. In this paper, the singleton empowered with perception and communication capabilities is regarded as a *Perception Node (PN)*. Generally, there are six main steps for processing the raw sensing data at a single PN [16], as shown

below:

- *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 task 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, 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), edge 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.

3) *Cloud Computing*: Considering the large-scale implementation of cooperation, cloud computing is involved to act as the fusion center for multiple PNs. Information from heterogeneous PNs will be transmitted to the *Cloud* 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 *Cloud*. 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 [28].

Generally, three types of perception data are generated from heterogeneous 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.

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, deep 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 used for CP or other CDA tasks [16]. 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 [17]. 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, deep fusion (also named intermediate fusion) attracts increasingly popular attention due to its superiority in CP performance [14], [19]. Detailed illustration and literature review for fusion schemes are conducted in Section V.

4) *Message Distribution*: The perception information (along with advisory or actuation signals) can be distributed to road users in two major ways, depending on the connectivity status. For conventional road users without wireless connectivity, such information can be delivered to end devices at the roadside, such as *Dynamic Message Sign (DMS)* or signal head display of traffic lights via the *Traffic Management Center (TMC)*. For 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], [29]. Cooperative Perception messages can support more sophisticated cooperative maneuvers in a mixed traffic environment. For example, vulnerable road users (VRUs) and legacy vehicles (LVs) can react to the message shown on DMS [6]. Connected vehicles (CVs) can use CP information to get better situational awareness and pass through intersections in a safer manner [30]. Autonomous vehicles (AVs) and connected and automated vehicles (CAVs)

can improve their driving performance via better coordination algorithms [31].

B. Taxonomy

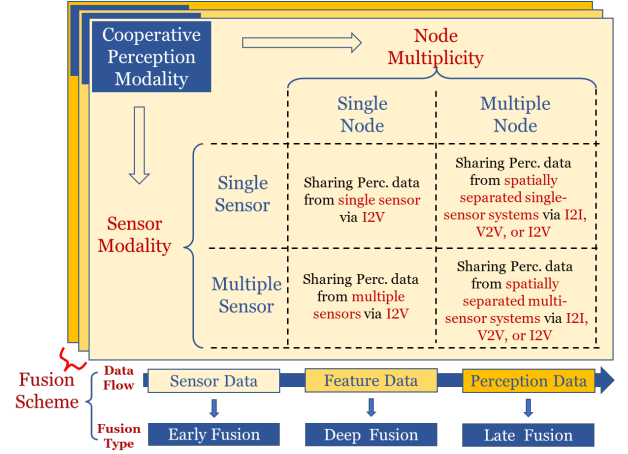


Fig. 2. Taxonomy of CP in terms of node multiplicity, sensor modality, and fusion scheme.

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

- *Single-Node Single-Mode CP (SS-CP)*: Cooperation between heterogeneous PNs by sharing perception data from the single-modal sensor(s) via infrastructure-to-everything (I2X) or vehicle-to-everything (V2X) communications.
- *Multi-Node Single-Mode CP (MS-CP)*: Cooperation between heterogeneous PNs by sharing perception data from single-modal multiple sensors perception via I2X and/or V2X communications.
- *Single-Node Multi-Mode CP (SM-CP)*: Cooperation between heterogeneous PNs by sharing perception data from multi-modal sensor perception via I2X or V2X communications.
- *Multi-Node Multi-Mode CP (MM-CP)*: Cooperation between heterogeneous PNs by sharing perception data from multi-modal sensor perception via I2X and/or V2X communications.

For each of the four CP types, three fusion schemes can be applied based on the types of perception data, which have been introduced in Section II-A3. In the following, a comprehensive literature review is conducted with detailed analyses on the aspects of node multiplicity, sensor modality, and fusion scheme, respectively.

III. NODE STRUCTURE

In this paper, we define *Node* to be a Perception Node (PN) that is capable of perceiving and communicating – as the fundamental unit for building the CP system. As mentioned

in Section 2, CP systems can be divided into Single-Node and Multi-Node CP systems. Meanwhile, in Section II-A2, the vehicle node (V-PN), and the infrastructure node (I-PN) are considered heterogeneous nodes in CP systems. For comprehensiveness and conciseness, CP is discussed from the aspect of *Node Structure* in this section: 1) I-PN-based CP, 2) V-PN-based CP, and 3) heterogeneous-PN-based CP.

A. I-PN-based CP

Object perception based on roadside sensors has a great potential to break the current bottleneck for autonomous driving, especially in a mixed traffic environment via cooperative perception [32]. This section reviews the infrastructure-based object detection and tracking approaches in the literature.

1) *Camera-based I-PN*: 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 [33]. 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 [34], 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 [35] 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 an Unmanned Aerial Vehicle (UAV) equipped with an embedded computer and a video camera. Inspired by the multi-task learning methodology, a segmentation head [36] is added to the object detector backbone [37]. Dedicated object tracking [38] 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 [39]. To identify the road boundaries, object detection and road segmentation methods are employed by utilizing the *FCN-VGG16* model [40] on the *KITTI* dataset [41] and *Faster R-CNN* [42] on *MS-COCO* dataset [43], respectively. Then, a method is designed to differentiate parked vehicles from the moving ones and then give them guidance on the nearest spot information to drivers.

For multi-camera perception systems from the roadside, 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 [18]. For

each camera, a depth image is projected to pseudo-point-cloud data [44]. Two sensor-fusion schemes are designed: early fusion and late fusion (see Fig. 3) and adapted based on Voxelnet [45]. The evaluation in a T-junction and a roundabout scenario in the CARLA simulator [46] demonstrates that the proposed method can enlarge the detection coverage without compromising accuracy.

2) *LiDAR-based I-PN*: In recent years, roadside LiDAR sensors attract 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 [47]. 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 [48]; 2) *Clustering*: To generate clusters for the laser points by implementing a DBSCAN method [49]; 3) *Classification*: To generate different labels for different traffic objects, such as vehicles and pedestrians, based on neural networks [50]; and 4) *Tracking*: To identify the same object in continuous data frames by applying a discrete Kalman filter [51]. Based on the aforementioned work, Cui et al. designed an automatic vehicle tracking system by considering vehicle detection and lane identification [52]. 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 [53]. 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 [54] and joint probabilistic data association filter [55] are adopted in the tracking flow. Finally, vehicle tracking is refined through a *Brid-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 [56]–[60].

On the other hand, using learning-based models to cope with LiDAR data is another main methodology. Bai et al. [30] proposed a deep-learning-based real-time vehicle detection and reconstruction system from roadside LiDAR data. Specifically, CARLA simulator [46] is implemented for collecting the training dataset, and *ComplexYOLO* model [61] 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. [16] 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* [62] for inference

from roadside PCD; 2) 3D multi-object tracking by improving DeepSORT [63] to support 3D tracking, and 3) 3D reconstruction by geodetic transformation and real-time onboard *Graphic User Interface (GUI)* display.

By combining traditional and deep learning algorithms Gong et al. [64] proposed a roadside LiDAR-based real-time detection approach. Several techniques are designed to guarantee real-time performance, including the application of Octree with region-of-interest (ROI) selection, and the development of an improved Euclidean clustering algorithm with an adaptive search radius. The roadside system is equipped with *NVIDIA Jetson AGX Xavier*, achieving the inference time of 110 ms per frame.

B. Vehicle Nodes

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. [65] proposed a multi-UAV CP system based on a distributed-centralized CP framework (similar to the current “edge-cloud” 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 send to a central server for further fusion by utilizing a probabilistic model.

For on-road vehicles, Rockl et al. [66] 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. [67] in 2012. A Car2X-based module was proposed to cooperate the perception results jointly 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. Machine et al. [68] 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. [69] 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 autonomous driving system (ADS) was implemented by Kim et al. [70], 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 for enabling driving automation in multiple tasks, such as the forward collision warning, overtaking/lane-changing assistance, automated lane-change capability, etc. Experiments demonstrated that by enabling ego-vehicle with expanded perception information, the potential of driving automation can be significantly improved.

For CP system based on LiDAR data, Chen et al. proposed an early fusion method (*Cooper* [17]) 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 [45], Sparse Convolution [71], and Region Proposal Network (RPN) [72]. 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. [73] 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 [74] (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 [75], 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* [76], is involved for cooperative perception to generate a PCD-based V2V dataset in a more realistic manner.

Zhang et al. [77] 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. [78] 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 [79] is applied for multi-node feature fusion and an open-source simulation-based dataset is developed and implemented for model training and validation.

C. Heterogeneous 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 by 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

TABLE II
SUMMARY OF DIFFERENT NODE STRUCTURES FOR COOPERATIVE PERCEPTION.

Structure	Modality	Pros. and Cons.	Highlighted Features	Author
Single-Node	Infrastructure	Pros: Higher location with flexible pose leads to less occlusion and system-level cost-effective. Cons: Need infrastructure support.	Infrastructure assisted high-fidelity traffic surveillance	Bai et al. [16]
	Vehicle	Pros: Low latency perception for ego-vehicle. Cons: Easily occluded by the surrounding vehicles or buildings.	Everything on the vehicle side: sensing, processing, analysis.	Arnol et al. [10]
Multi-Node	Vehi. + Vehi.	Pros: Extend perception range from vehicle side. Cons: Occlusion by other vehicles.	Sharing features generated from convolutional neural networks.	Chen et al. [13]
	Infra. + Infra.	Pros: Extend perception range from infrastructure side. Cons: Have blind zone under the sensor.	Sharing preprocessed RGB data among all roadside sensors.	Arnold et al. [18]
	Infra. + Vehi.	Pros: Achieve a comprehensive range and field of view (FOV) for perception. Cons: Require heterogeneity of the model.	Considering asynchronous information sharing, pose errors, and heterogeneity of V2X components.	Xu, et al. [19]

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) [80], Xu et al. [19] 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) [81] 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.

D. Summary

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. SENSORS MODALITY

For the CP system, sensors are the most fundamental modules due to their roles in raw data collection. Thus, this section overviews typical types of sensors that are utilized in transportation systems from different perspectives.

A. Configuration and Performance

For sensors equipped on current ADS, the most popular ones are cameras, LiDAR, and radar. Onboard radar has been deployed on vehicles to mainly achieve ADAS functionalities for many years [82], such as Adaptive Cruise Control (ACC),

Collision Avoidance, etc [83]. For the onboard cameras, different ADS may take different configurations including single-camera ADS and multi-camera ADS. Single-camera ADS, such as the ADS developed by *Comma.ai* [84], only deploys a camera-based perception system at the middle-top of the windshield. Multi-camera ADS, such as *Waymo ADS* [85], utilizes multiple cameras installed around the top-surrounding positions of the vehicle. In most common cases, LiDAR sensors, due to their capability of panoramic FOV, are mainly configured on the top of the vehicle. In some ADS, e.g., *Waymo ADS*, auxiliary LiDAR systems are installed for complementing the blind zone of the top LiDAR [86].

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 [16], [30], [47].

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

TABLE III
PERFORMANCE MATRIX FOR DIFFERENT SENSORS UTILIZED FOR INFRASTRUCTURE-BASED PERCEPTION (RATING RANGE FROM 1 TO 3 STARS).

Capabilities	Camera	LiDAR	RADAR	Thermal	Fisheye	Loop
Privacy-safe data	★	★★★	★★★	★★	★	★★★
Accurately detects and classifies objects	★★	★★★	★	★★	★★	★
Accurately measures object speed and position	★★	★★★	★★★	★★	★★	★★
Extensive FOV	★	★★★	★★	★	★★★	★
Reliability across changes in lighting, sun, temperature	★★	★★★	★★★	★★	★★	★★★
Ability to read signs and differentiate color	★★★	★★	★	★	★★★	★
Cost for deployment and maintenance	★★★	★	★★	★★	★★	★

provides resilience to lighting changes.

- Fisheye Camera: 360-degree full field-of-view (FOV) for detection. Requires a high-accurate calibration matrix to account for distortion.
- Loops: Measuring traffic counts and speed. Costly to install and maintain due to intrusiveness.

In terms of the number of sensors applied, a systematic operational pipeline of object perception can be divided into two main categories, i.e., single-sensor-based and multi-sensor-based, as shown in Fig. 3.

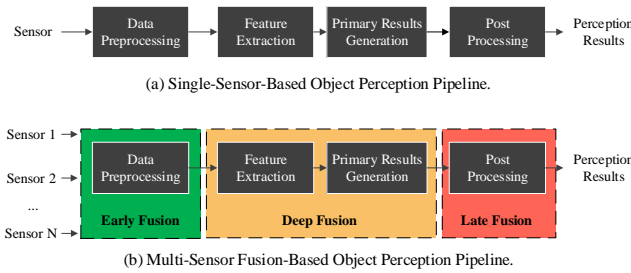


Fig. 3. Systematic diagram of operational pipeline for: (a) single-sensor-based perception model; and (b) multi-sensor-based perception model.

B. Single-Sensor Perception

Single-sensor-based object perception systems have been widely developed and applied in the real-world transportation system whose main pipeline is demonstrated in Fig. 3 (a). Data collected from the sensor is first *preprocessed* to reduce noise, filter unrelated data, and properly reformat for downstream modules. Then, *feature extraction* is applied to calculate predefined features by mathematical models (if based on traditional methods) or to generate hidden features by neural networks (if based on deep learning). Detection and tracking results are generated by the *perception* module and are fed into the *post-processing* module to further clean the perception outputs (e.g., filtering overlapped bounding boxes and predictions with scores under the threshold).

In this section, we briefly cover major milestones of single-sensor perception chronologically from two perspectives – the traditional approach and the deep-learning approach – for cameras and LiDARs, respectively.

1) *Camera*: Approximately twenty years ago, Viola and Jones [87] proposed a method for real-time detection of human faces without any constraints. This algorithm outperformed any of other contemporary algorithms in terms of real-time performance, without compromising detection accuracy. In 2005, Dalal and Triggs [88] proposed the Histogram of Oriented Gradients (HOG) feature descriptor which provided significant improvement of the scale-invariant feature transform [89] and shapes context [90]. The HOG detector has been regarded as the cornerstone for many subsequent object detectors and implemented in various real-world applications [91], [92]. Deformable Part-based Model (DPM) proposed by Felzenszwalb et al. [91] consecutively won the Pascal Visual Object Classes (VOC)-07, -08, and -09 detection challenges [93]. Due to their dominant performance, DPM and its variants [92] are widely regarded as the pinnacle of traditional object detection methods [12].

Benefiting from the increased computational power, convolutional neural networks (CNNs) [94] started to be widely used in 2012. Two years later, Girshick et al. proposed the Regions with CNN features (R-CNN) for object detection and completely unfolded the advantage of deep learning [95]. In the same year, Spatial Pyramid Pooling Networks (SPPNet) proposed by He et al. was able to generate feature representation regardless of the image size, and run 20 times faster than R-CNN without compromising accuracy [96]. In 2015, multiple renowned detectors were proposed by researchers: 1) Fast R-CNN [97] – over 200 times faster than R-CNN – proposed by Girshick; 2) Faster R-CNN [42] – the first end-to-end, and the first near-realtime deep learning detector – proposed by Ren et al.; 3) You Only Look Once (YOLO) [98] – the first one-stage detector in the deep learning era with extremely fast speed (45 - 155 fps) – proposed by Joseph et al.; and 4) Single Shot MultiBox Detector (SSD) [99] – the second one-stage detector but with significantly improved accuracy – proposed by Liu et al. In 2017, Lin et al. proposed Feature Pyramid Networks (FPN) [100] based on Faster R-CNN, which achieved the SOTA object detection performance and has become a fundamental building block for various object perception models. In recent years, *Transformers* [79] embedded with the mechanism of attention have been leading the trend in the majority of object perception tasks, such as the Vision Transformer (ViT) proposed by Dosovitskiy et al. [80], and Swin-Transformer proposed by Liu et al. [101].

2) *LiDAR*: Before 2015, one of the most popular methodologies for solving PCD from LiDAR sensors is the bottom-up pipeline based on traditional methods, such as “*Clustering [102]→Classification [103]→Tracking [51]*”. Due to its explanation, interpolation, and free from data labeling, traditional bottom-up methodologies are still popular in current infrastructure-based LiDAR perception tasks [47], [60], [104].

With the great success achieved by CNNs in image-based object perception, PCD quickly became the upcoming target for CNNs. However, PCD has a totally different data format compared with RGB images, which brings lots of challenges for applying existing CNN technologies to 2D vision tasks. Point-wise manipulation is considered straightforward for extracting features from PCD for object detection. *Point RCNN [105]* was proposed by Shi et al. to aggregate the point features via a *PointNet++ [106]* encoder. Endowed with the natural fit, point-based methods provide dominant performance in detection accuracy, however, under the sacrifice of computational efficiencies, such as *PV-RCNN [107]* (Shi et al. 2020).

Since PCD is in 3-dimension and sparse data, Wang et al. [108] creatively cut the whole 3D point cloud into 3D voxels grids. Then a feature vector was designed to represent each voxel, which was fed into a linear SVM [109] for classification results. A specific voting scheme was designed and mathematically proved to be able to act as sparse convolution [110] and the method was named *Vote3D* in 2015. Two years later, Engelcke et al. [111] proposed the *Vote3Deep*, which improved *Vote3D* by involving sparse convolution directly into the voting scheme. Furthermore, Rectified Linear Unit (ReLU) [112] and L_1 regularisation [111] (or named *L1 Norm*) were involved to boost the learning process based on large sparse data, like PCD. In 2018, *VoxelNet [45]* was proposed by Zhou et al., which introduced a learnable voxel encoder to generate hidden features for voxels. Specifically, 3D convolution was applied as a 3D backbone for 3D voxel feature extraction, and 2D CNN-based Region Proposal Network (RPN) [72] was designed as a 2D backbone. This voxelization mechanism has been widely used in the following work, such as *SECOND [71]* (Yan et al. 2018), *PointPillar [62]* (Lang et al. 2019), *Voxel RCNN [113]* (Deng et al. 2021), etc.

Starting in 2018, projecting PCD into a 2D BEV feature map has quickly become a popular methodology. Inspired by YOLO, Simony et al. [61] proposed *ComplexYolo* which projected PCD into three manually defined feature channels, and then the BEV feature map was fed into a 2D backbone for generating detection results. Since the BEV scheme provides a straightforward way for solving 3D data in 2D manners, lots of BEV-based methods have emerged such as *PIXOR [114]* (Yang et al. 2018), *SCANet [115]* (Lu et al. 2019), *BEVFusion [116]* (Liu et al. 2022), etc.

C. 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 when compared with single-sensor-based perception

systems. In this section, three popular multi-sensor perception schemes based on high-resolution sensors are discussed in this paper, i.e., *Camera+Camera*, *Camera+LiDAR*, and *LiDAR+LiDAR*.

1) *Cam + Cam*: The multi-camera system has been developed for decades and lots of applications have been designed and implemented in our current transportation systems [117], 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. [65] 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. [18] 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 for generating 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 [118]. 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 [119]. 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 [120], the particle filter [121], 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 disjoint camera system, appearance cues are designed for capturing the common features between multiple views by integrating spatial-temporal information [122]. 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 [123], unsupervised learning network [124], etc., have been implemented to dynamically update the model.

2) *Cam + Lidar*: As different sensor modalities, camera and LiDAR seem to be a naturally complementary couple for perception. For instance, the camera is good at perceiving the vision information but lacking 3D distance data, while the LiDAR excels at collecting 3D information but lacking vision data.

TABLE IV
SUMMARY OF DIFFERENT SENSOR MODALITIES FOR COOPERATIVE PERCEPTION.

Structure	Modality	Pros. and Cons.	Highlighted Features	Author
Single-Sensor	Camera	Pros: abundant vision data with cost-effective system.	Using shifted window multi-head attention	Liu et al. [101]
		Cons: difficult to provide high-fidelity 3D information and significantly impact by lighting condition.		
	Lidar	Pros: capable to provide high-fidelity 3D information with panoramic FOV.	Encoding point cloud into voxelized pillars	A. Lang, et al. [62]
		Cons: sparse data without vision information.		
Multi-Sensor	Cam. + Cam.	Pros: expand the FOV and perception area.	Projecting RGB camera data into pseudo-LiDAR point cloud	E. Arnold et al. [18]
		Cons: difficult to provide high-fidelity 3D information and significantly impact by lighting condition.		
	Lidar + Lidar	Pros: expand the FOV and increase the density of the point cloud.	Considering heterogeneous perception nodes, e.g., vehicle and infra.	Bai et al. [14]
Cons: sparse data without vision information.				
	Cam. + Lidar	Pros: taking advantage of both camera and Lidar.	Capturing BEV features from both sensors via CNN.	Liu et al. [116]
		Cons: totally different information modality, thus difficult to fuse the data effectively.		

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 [15]. 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 CBNNet [125] 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 [126] for roadside sensor data.

Since 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. A popular way is to project camera information into LiDAR data to endow PCD with vision information. *PointPainting* [127], 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* [62] for generating object detection results. Recently, Liu et al. [116] 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. The performance of *BEVFusion* is the current SOTA for 3D object detection.

3) *Lidar + Lidar*: 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 [17]. 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], [75]. Marvasti et al. [73] 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. [78] proposed a V2V-based cooperative object detection method. A similar CNN process [62] 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 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. [19] 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.

D. Summary

Table IV summarizes the advantages and disadvantages of different sensor modalities in the CP system. Different high-resolution sensors have different strengths. The camera is good

TABLE V
SUMMARY OF DIFFERENT FUSION SCHEMES FOR COOPERATIVE PERCEPTION.

Fusion Scheme	Methodology	Pros. and Cons.	Highlighted Features	Author
Early Fusion	Deep Learning	Pros: Raw data is shared and gathered to form a holistic view.	Raw point cloud data is compressed to fit the limited bandwidth.	Chen et al. [17]
		Cons: Low tolerance to the noise and delay of the transmitted data; potentially constrained by the communication bandwidth.		
Deep Fusion	Deep Learning	Pros: High tolerance to the noise, delay, and difference between different nodes and sensor models.	Deep neural features are extracted and fused based on spatial correspondence.	Bai et al. [14]
		Cons: Require training data and hard to find a systematic way for model design.		
Late Fusion	Traditional	Pros: Easy to design and deploy in real-world system.	A late-fusion is proposed based on joint re-scoring and non-maximum suppression.	Zhang et al. [77]
		Cons: Significantly limited by the wrong perception results or the difference between sources.		

at capturing vision information while LiDAR is excellent for collecting 3D information. Simultaneously taking advantage of these sensors in a complementary scheme is regarded as a promising solution to improving the perception accuracy of surveillance systems.

V. FUSION SCHEME

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 data at the preprocessing stage; 2) *Deep Fusion* – to fuse features at the feature extraction stage; and 3) *Late Fusion* – to fuse perception results at the post-processing stage. Different fusion schemes both have advantages and disadvantages in terms of different perspectives. For instance, Early Fusion and Deep 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. 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

It is intuitive to share the raw sensor data with other PNs for expanding the perceptive range and improving detection accuracy. Following this strategy, the raw sensor data from multiple PNs are projected into a unified coordinate system for further processing [119]. 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 [118]. 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 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 [41]). For V2V early fusion, it is true that communicating raw sensor data with one ego-vehicle is not an impossible solution [17], 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 fusion 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. [67] applied EKF to jointly align the shared bounding box proposals based on spatiotemporal constraints. Additionally, Non-Maximum Suppression (NMS) [128] and other machine-learning-based proposal refining methods are widely applied in late fusion methods for object perception [18]. Recently, due to the distributed attributes of late fusion, *Federated Learning* [129] also attracts increasing popularity in perception systems [77].

C. Deep Fusion

The core ideology of deep fusion (also named *Intermediate Fusion*) can be simply summarized as using deeply extracted features for fusion that happens at the intermediate stages of the perception pipeline. Deep 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. [19] 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:

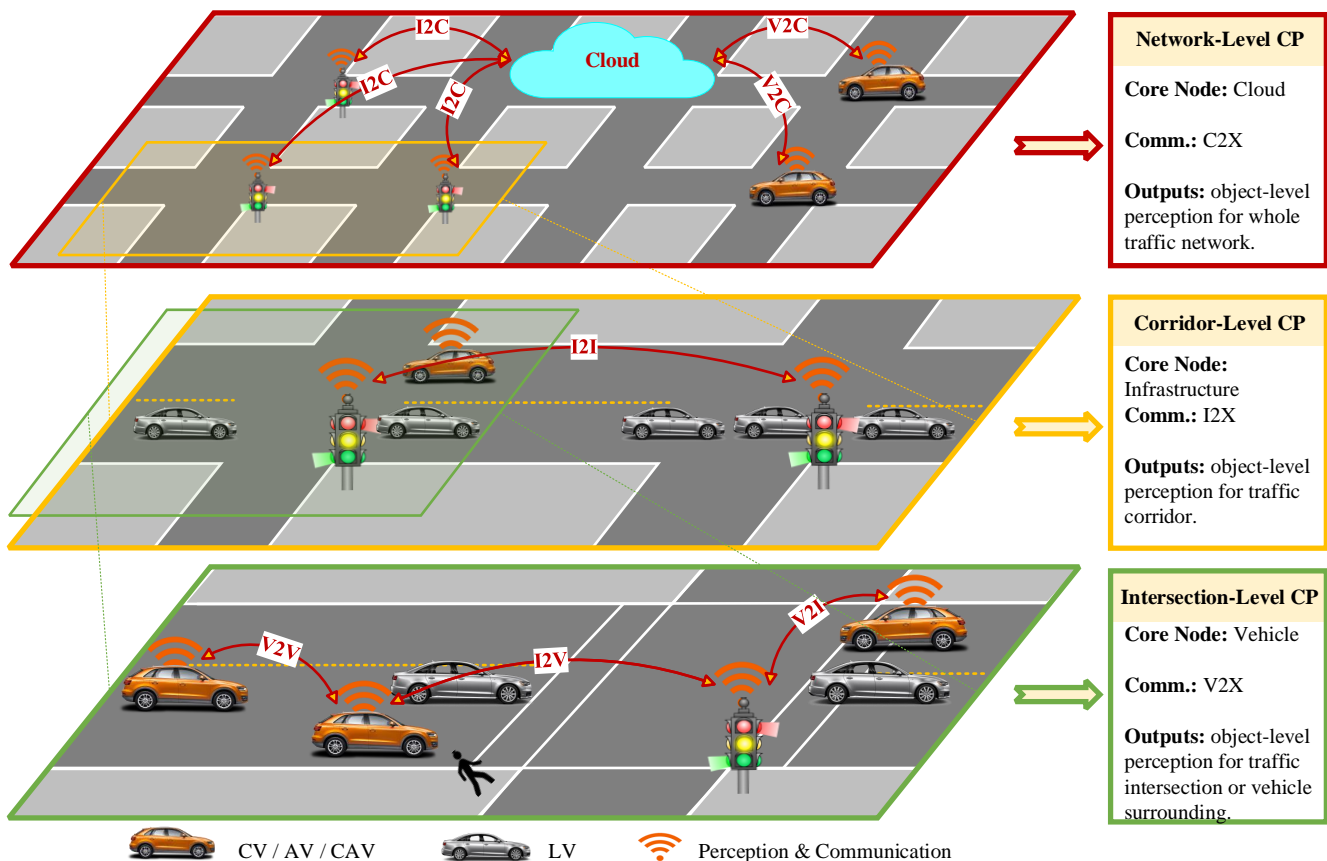


Fig. 4. The schematic diagram of the HCP framework.

- With no error involved, early fusion and deep 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 deep fusion methods [13], [19], [75], [78] 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 [18], [77].

Although cooperative perception has been developed in multiple areas for several decades, deep-fusion-based cooperative perception is still in its infancy. Most of the deep fusion methods for CP were devised in the past few years, such as *F-Cooper* [13] (2019), *V2V Net* [75] (2020), *OPV2V* [78] (2021), *PillarGrid* [14] and *V2X-ViT* [19] (2022), etc. So far, most of the deep feature extraction is conducted by CNN, such as [13], [14], [75], because the CNN-based feature is highly related to the local spatial information. Very recently, some studies have applied transformers as the deep feature extractor [19], [78] due to their capability for panoramic feature extraction.

D. Summary

Table V summarizes the advantages and disadvantages of different sensor fusion schemes for CP systems. Early fusion only needs the calibration for aligning multi-source data into a unified coordinate system but requires a large communication bandwidth for transmitting data. Late fusion mainly focuses on how to merge the proposals generated from multiple perception pipelines, which is straightforward but suffered from limited accuracy. Deep fusion is quickly becoming a transformable solution for CP due to its capabilities of low-communication requirements and high accuracy.

VI. HIERARCHICAL COOPERATIVE PERCEPTION FRAMEWORK

Based on the overview of the aforementioned literature, Three major issues can be identified for CP system 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 consecutively 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. 1, 2) the taxonomy of CP as shown in Fig. 2, and 3) the analysis of reviewed literature.

In this paper, the HCP framework mainly focuses on the intersection scenarios and consists of three-level: 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.

1) *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-PN and I-PN 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], [75], [78] 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 deep fusion to generate perception results. The I-PN then broadcasts the perception results to vehicles within its own communication range. Recent V2I-based CP can be regarded as a specific version of the intersection-level CP [14], [19]. Intersection-level CP is a crucial component for unlocking the current bottleneck (in terms of efficiency, safety, and sustainability) for cooperative driving automation in a mixed traffic environment [7].

2) *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. The core is the infrastructure node, i.e., 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 $1GB/s$ [130], which is enough for transmitting intersection-level data between intersections.

Empowered by high-speed communication, I2I-based CP is capable of applying all fusion schemes based on specific scenarios. Raw data sharing can be a typical style for I2I-based CP [18]. 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 improve traffic management [30].

3) *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 for generating the perception results.

VII. DATASETS AND SIMULATORS

In this section, we briefly introduce the tools that support the development of cooperative perception including datasets and simulators. We hope this section can give researchers a quick glance at the foundations that can possibly enable their relevant research.

A. Datasets

1) *General Object Perception*: Owing to prevailing needs in autonomous driving for surrounding perception, most real-world datasets for object detection and tracking are collected from onboard sensors. Several widely used datasets (both supporting Camera and LiDAR) for driving automation are briefly introduced as follows:

- *KITTI*: one of the most popular datasets, which consists of hours of traffic scenarios recorded with a variety of sensor modalities for mobile robotics and autonomous driving [41].
- *NuScenes*: the first dataset to carry the fully autonomous vehicle sensor suite: 6 cameras, 5 radars, and 1 LiDAR, all with a full 360-degree field of view [131].
- *Waymo Open Dataset*: a large-scale, high-quality, diverse dataset that consists of 1150 scenes captured across a range of urban and suburban geographical terrains [85].

2) *Infrastructure-based Perception*: Because roadside perception has great potential to promote the development of CDA, there are immediate demands for establishing a roadside sensor-based dataset for various infrastructure-based object perception tasks. In 2021, *BAAI-VANJEE Roadside Dataset* was published by Deng et al. to support the Connected Automated Vehicle Highway technologies [132]. The BAAI-VANJEE Roadside Dataset consists of LiDAR data and RGB images collected by a roadside data-collection platform and contains 2500 frames of LiDAR data, and 5000 frames of RGB images which includes 12 classes of objects, 74K 3D object annotations, and 105K 2D object annotations.

3) *Cooperative Perception*: Although various kinds of real-world datasets have been collected for training models for different perception tasks. Before 2022, there is no available open-sourced cooperative perception dataset for real-world data. Thus, to overcome this issue, researchers mainly follow two ways of dataset acquisition. The most popular way is to build cooperative perception scenarios in a high-fidelity simulator and then collect multi-node multi-sensor data from

the environment. Several datasets have been built based on CARLA [46]. For instance, to enable V2V cooperative perception, *OpenV2V* Dataset [78] is collected by attaching LiDAR sensors to multiple vehicles in the CARLA simulator under different scenarios.

Additionally, for heterogeneous perception nodes, the *CARTI* dataset [14], [133] has been collected by deploying LiDAR and camera sensors to both vehicles and infrastructure in CARLA environments. Specifically, the raw data, sensor calibration, and ground truth label are designed following the same format as *KITTI* dataset. Thus *CARTI* dataset is readily integrated into the current deep learning codebase for quick development, such as *MMDetection3D* [134].

In 2022, *DAIR-V2X* [135], the first real-world cooperative perception dataset comes to the stage, which is a large-scale, multi-node, multi-modality CP dataset. Specifically, *DAIR-V2X* contains 39k images, 39k PCD frames, and 10 classes of ground truth labels with synchronized time stamps. Sensors are collected from both vehicle nodes and infrastructure nodes.

B. Simulators

In the context of cooperative perception using simulation, high-fidelity sensors are inevitably required due to the dependence on high-resolution sensor data [136]. Although traditional microscopic traffic simulators can also emulate the behavior at the object level, such as SUMO [137], VIS-SIM [138], AIMSUN [139], etc, they can not provide high-resolution sensor data with high fidelity. Thus, in recent years, several autonomous driving simulators have been developed to enable high-fidelity modeling of the surrounding environment and sensor capability by utilizing game engines, such as Unity [140] and Unreal Engine [141]. Specifically, several representative simulators are CARLA [46], SVL [142], AirSim [143], etc.

- CARLA is an open-source simulator for autonomous driving and supports flexible specifications of sensor suites and environmental conditions. In addition to open-source codes and protocols, CARLA provides open digital assets (e.g., urban layouts, buildings, and vehicles) that can be used in a friendly manner for researchers.
- SVL is a high-fidelity simulator for driving automation, which provides end-to-end and full-stack simulation ready to be hooked up with several open-source autonomous driving stacks, such as Autoware [144] and Apollo [145].
- Besides high-fidelity sensors and environments, AirSim includes a physics engine that can operate at a high frequency for real-time hardware-in-the-loop (HIL) simulations with the support for popular protocols, such as MavLink [146].

These simulators are all open-source with detailed tutorials and can provide high-resolution and high-fidelity sensor data, such as cameras and LiDARs. These simulators can provide a highly customized and cost-effective way for collecting training datasets and traffic scenarios, and thus are widely applied in learning-based object perception tasks [10], [18].

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 [16], [34], [35], [56]. From the perspective of cooperative automated driving, V2V-based cooperative perception is also a promising solution to enabling the ego-vehicle with the capability of *seeing through* [66]–[68], [70].

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 [147]. Due to the heterogeneity of the perception singleton, only recently few studies have considered the cooperation between vehicle nodes and infrastructure nodes [14], [19]. Thus, vehicle-infrastructure cooperation is one of the most significant opening tasks for cooperative perception.

2) *Sensor System for Fidelity*: To the most extent, the capability of the sensor system can be regarded as the stepping stone of an intelligent transportation system. Since the perception data generated from sensor systems is the foundation of the downstream modules, such as prediction, decision-making, and actuation [30]. Thus for cooperative perception, cameras and LiDAR are widely applied to accessing high-fidelity sensing data.

However, in most of the research, these two kinds of high fidelity sensors work separately – a cooperative perception system only equipped with one kind of sensor – such as multi-camera-based CP [15], [18] and multi-LiDAR-based CP [14], [75]. According to the analysis in Section IV-C, cameras and LiDAR are naturally complimentary to each other, and the camera-LiDAR-based perception method can also achieve the SOTA performance in general object detection [116]. Thus, developing multi-modality sensors for cooperative perception is an important way to improve the overall fidelity of the perception results.

On the other hand, although the infrastructure plays a key role in cooperative perception, current roadside-sensor-based perception methods are, to most extent, applied directly from general perception methods, i.e., onboard sensor-based model. 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 [47], [53], [57], [59], [60], which has a performance gap compared with the SOTA methods [45], [62]. Since the sensing methodologies of roadside sensors are different from onboard perceptions, one of the major challenges is roadside data acquisition and annotation for promoting the deep learning-based research of infrastructure-based perception systems.

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 [17], [18]. 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 [19]. Late fusion-based research has been widely applied to various kinds of cooperative perception tasks since decades ago [18], [65], [77]. Late fusion has less requirement for communication but its performance also suffers from the merging of the object proposals from multiple sources [18].

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], [19]. However, due to the deeply coupled feature and model complexity, large-scale extension is an inevitable challenge for deep fusion-based cooperative perception.

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 cooperate the information from vehicles with infrastructures are the keys to unlocking a new era of perception for cooperative driving automation.

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 [148] 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 [116]. 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 [19], 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, it could be extremely costly and energy inefficient to empower every single vehicle with a high-performance computation system for perception. Therefore, by only deploying lightweight onboard computation modules on the vehicle side, such as feature map extraction, it becomes much more cost-effective to 1) enable local deep-fusion-based cooperative perception [14] or 2) retrieve perception results from infrastructure-based high-performance nodes for a wider range of perceptions [16].

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), a 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.

IX. CONCLUSIONS

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 the cooperative perception system. Cooperative perception methods are then introduced with detailed literature reviews from three perspectives: node structure, sensor modality, and fusion scheme. The proposed hierarchical cooperative perception framework is analyzed from the various 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.

ACKNOWLEDGMENTS

This research was funded by Toyota Motor North America, InfoTech Labs. The contents of this paper 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. D. of Transportation, "Overview of motor vehicle crashes in 2019," Available: <https://crashstats.nhtsa.dot.gov/Api/Public/Publication/813060>, 2020.

- [2] INRIX, "Inrix: Congestion costs each american 97 hours, \$1,348 a year," Available: <https://inrix.com/press-releases/scorecard-2018-us/>, 2018.
- [3] U. D. of Energy, "Fotw #1204: Fuel wasted due to u.s. traffic congestion in 2020 cut in half from 2019 to 2020," Available: <https://www.energy.gov/eere/vehicles/articles/fotw-1204-sept-20-2021-fuel-wasted-due-us-traffic-congestion-2020-cut-half>, 2021.
- [4] D. J. Fagnant and K. Kockelman, "Preparing a nation for autonomous vehicles: opportunities, barriers and policy recommendations," *Transportation Research Part A: Policy and Practice*, vol. 77, pp. 167–181, 2015.
- [5] S. of Automotive Engineers (SAE), "Taxonomy and definitions for terms related to cooperative driving automation for on-road motor vehicles," Available: https://www.sae.org/standards/content/j3216_202107, 2021.
- [6] J. Wu, H. Xu, Y. Zhang, and R. Sun, "An improved vehicle-pedestrian near-crash identification method with a roadside lidar sensor," *Journal of safety research*, vol. 73, pp. 211–224, 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 Transactions on Intelligent Transportation Systems*, pp. 1–14, 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 Intelligent Transportation Systems Magazine*, vol. 12, no. 1, pp. 4–24, 2020.
- [9] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International journal of computer vision*, vol. 128, no. 2, pp. 261–318, 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 Transactions on Intelligent Transportation Systems*, vol. 20, no. 10, pp. 3782–3795, 2019.
- [11] A. Manjunath, Y. Liu, B. Henriques, and A. Engstle, "Radar based object detection and tracking for autonomous driving," in *2018 IEEE MTT-S International Conference on Microwaves for Intelligent Mobility (ICMIM)*, 2018, pp. 1–4.
- [12] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.
- [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 *Proceedings of the 4th ACM/IEEE Symposium on Edge Computing*, ser. SEC '19. New York, NY, USA: Association for Computing Machinery, 2019, p. 88–100. [Online]. Available: <https://doi.org/10.1145/3318216.3363300>
- [14] Z. Bai, G. Wu, M. J. Barth, Y. Liu, A. Sisbot, and K. Oguchi, "Pillargrid: Deep learning-based cooperative perception for 3d object detection from onboard-roadside lidar," *arXiv preprint arXiv:2203.06319*, 2022.
- [15] 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, 2021.
- [16] Z. Bai, S. P. Nayak, X. Zhao, G. Wu, M. J. Barth, X. Qi, Y. Liu, and K. Oguchi, "Cyber mobility mirror: Deep learning-based real-time 3d object perception and reconstruction using roadside lidar," *arXiv preprint arXiv:2202.13505*, 2022.
- [17] Q. Chen, S. Tang, Q. Yang, and S. Fu, "Cooper: Cooperative perception for connected autonomous vehicles based on 3d point clouds," in *2019 IEEE 39th International Conference on Distributed Computing Systems (ICDCS)*. IEEE, 2019, pp. 514–524.
- [18] E. Arnold, M. Dianati, R. de Temple, and S. Fallah, "Cooperative perception for 3d object detection in driving scenarios using infrastructure sensors," *IEEE Transactions on Intelligent Transportation Systems*, 2020.
- [19] R. Xu, H. Xiang, Z. Tu, X. Xia, M.-H. Yang, and J. Ma, "V2x-vit: Vehicle-to-everything cooperative perception with vision transformer," *arXiv preprint arXiv:2203.10638*, 2022.
- [20] A. Caillot, S. Ouerghi, P. Vasseur, R. Boutteau, and Y. Dupuis, "Survey on cooperative perception in an automotive context," *IEEE Transactions on Intelligent Transportation Systems*, 2022.
- [21] S. of Automotive Engineers (SAE), "Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles," Available: https://www.sae.org/standards/content/j3016_202104/, 2021.
- [22] F. H. A. USDOT, "Carma," Available: <https://highways.dot.gov/tags/carma>, 2021.
- [23] K. Nellore and G. P. Hancke, "A survey on urban traffic management system using wireless sensor networks," *Sensors*, vol. 16, no. 2, p. 157, 2016.
- [24] S. Y. Cheung, S. C. Ergen, and P. Varaiya, "Traffic surveillance with wireless magnetic sensors," in *Proceedings of the 12th ITS world congress*, vol. 1917, 2005, p. 173181.
- [25] B. Coifman, D. Beymer, P. McLauchlan, and J. Malik, "A real-time computer vision system for vehicle tracking and traffic surveillance," *Transportation Research Part C: Emerging Technologies*, vol. 6, no. 4, pp. 271–288, 1998. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0968090X98000199>
- [26] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [27] S. Zhang, J. Chen, F. Lyu, N. Cheng, W. Shi, and X. Shen, "Vehicular communication networks in the automated driving era," *IEEE Communications Magazine*, vol. 56, no. 9, pp. 26–32, 2018.
- [28] 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," *Transportation Research Part C: Emerging Technologies*, vol. 68, pp. 168–184, 2016.
- [29] 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 Transactions on Intelligent Vehicles*, vol. 2, no. 4, pp. 266–277, 2017.
- [30] Z. Bai, G. Wu, X. Qi, K. Oguchi, and M. J. Barth, "Cyber mobility mirror for enabling cooperative driving automation: A co-simulation platform," *arXiv preprint arXiv:2201.09463*, 2022.
- [31] M. Shan, K. Narula, Y. F. Wong, S. Worrall, M. Khan, P. Alexander, and E. Nebot, "Demonstrations of cooperative perception: Safety and robustness in connected and automated vehicle operations," *Sensors*, vol. 21, no. 1, p. 200, 2020.
- [32] 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*, p. 100057, 2021.
- [33] R. Ojala, J. Vepsäläinen, J. Hanhiova, V. Hirvisalo, and K. Tammi, "Novel convolutional neural network-based roadside unit for accurate pedestrian localisation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 9, pp. 3756–3765, 2020.
- [34] E. Guo, Z. Chen, S. Rahardja, and J. Yang, "3d detection and pose estimation of vehicle in cooperative vehicle infrastructure system," *IEEE Sensors Journal*, vol. 21, no. 19, pp. 21759–21771, 2021.
- [35] 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 Sensing*, vol. 13, no. 4, p. 573, 2021.
- [36] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.
- [37] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.
- [38] D. S. Bolme, J. R. Beveridge, B. A. Draper, and Y. M. Lui, "Visual object tracking using adaptive correlation filters," in *2010 IEEE computer society conference on computer vision and pattern recognition*. IEEE, 2010, pp. 2544–2550.
- [39] E. Cicek and S. Gören, "Fully automated roadside parking spot detection in real time with deep learning," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 23, p. e6006, 2021.
- [40] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.
- [41] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
- [42] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *Advances in neural information processing systems*, vol. 28, pp. 91–99, 2015.
- [43] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *European conference on computer vision*. Springer, 2014, pp. 740–755.
- [44] C. Glennie and D. D. Lichti, "Static calibration and analysis of the velodyne hdl-64e s2 for high accuracy mobile scanning," *Remote sensing*, vol. 2, no. 6, pp. 1610–1624, 2010.

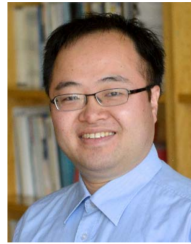
- [45] Y. Zhou and O. Tuzel, "Voxelnet: End-to-end learning for point cloud based 3d object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4490–4499.
- [46] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.
- [47] J. Zhao, H. Xu, H. Liu, J. Wu, Y. Zheng, and D. Wu, "Detection and tracking of pedestrians and vehicles using roadside lidar sensors," *Transportation research part C: emerging technologies*, vol. 100, pp. 68–87, 2019.
- [48] J. Wu, H. Xu, and J. Zheng, "Automatic background filtering and lane identification with roadside lidar data," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017, pp. 1–6.
- [49] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *kdd*, vol. 96, no. 34, 1996, pp. 226–231.
- [50] J. Li, J.-h. Cheng, J.-y. Shi, and F. Huang, "Brief introduction of back propagation (bp) neural network algorithm and its improvement," in *Advances in computer science and information engineering*. Springer, 2012, pp. 553–558.
- [51] G. Welch, G. Bishop *et al.*, "An introduction to the kalman filter," 1995.
- [52] Y. Cui, H. Xu, J. Wu, Y. Sun, and J. Zhao, "Automatic vehicle tracking with roadside lidar data for the connected-vehicles system," *IEEE Intelligent Systems*, vol. 34, no. 3, pp. 44–51, 2019.
- [53] J. Zhang, W. Xiao, B. Coifman, and J. P. Mills, "Vehicle tracking and speed estimation from roadside lidar," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 13, pp. 5597–5608, 2020.
- [54] S. J. Julier and J. K. Uhlmann, "Unscented filtering and nonlinear estimation," *Proceedings of the IEEE*, vol. 92, no. 3, pp. 401–422, 2004.
- [55] Y. Bar-Shalom, F. Daum, and J. Huang, "The probabilistic data association filter," *IEEE Control Systems Magazine*, vol. 29, no. 6, pp. 82–100, 2009.
- [56] L. Zhang, J. Zheng, R. Sun, and Y. Tao, "Gc-net: Gridding and clustering for traffic object detection with roadside lidar," *IEEE Intelligent Systems*, 2020.
- [57] 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. 184 426–184 436, 2020.
- [58] M. Gouda, B. Arantes de Achilles Mello, and K. El-Basyouny, "Automated object detection, mapping, and assessment of roadside clear zones using lidar data," *Transportation research record*, vol. 2675, no. 12, pp. 432–448, 2021.
- [59] Z. Zhang, J. Zheng, X. Wang, and X. Fan, "Background filtering and vehicle detection with roadside lidar based on point association," in *2018 37th Chinese Control Conference (CCC)*, 2018, pp. 7938–7943.
- [60] Z. Zhang, J. Zheng, H. Xu, X. Wang, X. Fan, and R. Chen, "Automatic background construction and object detection based on roadside lidar," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4086–4097, 2019.
- [61] 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 *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018, pp. 0–0.
- [62] A. H. Lang, S. Vora, H. Caesar, L. Zhou, J. Yang, and O. Beijbom, "Pointpillars: Fast encoders for object detection from point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 697–12 705.
- [63] B. Veeramani, J. W. Raymond, and P. Chanda, "Deepsort: deep convolutional networks for sorting haploid maize seeds," *BMC bioinformatics*, vol. 19, no. 9, pp. 1–9, 2018.
- [64] Z. Gong, Z. Wang, B. Zhou, W. Liu, and P. Liu, "Pedestrian detection method based on roadside light detection and ranging," *SAE International Journal of Connected and Automated Vehicles*, vol. 4, no. 12-04-04-0031, 2021.
- [65] 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," *Journal of Field Robotics*, vol. 23, no. 3-4, pp. 165–184, 2006.
- [66] M. Rockl, T. Strang, and M. Kranz, "V2v communications in automotive multi-sensor multi-target tracking," in *2008 IEEE 68th Vehicular Technology Conference*. IEEE, 2008, pp. 1–5.
- [67] A. Rauch, F. Klanner, R. Rasshofer, and K. Dietmayer, "Car2x-based perception in a high-level fusion architecture for cooperative perception systems," in *2012 IEEE Intelligent Vehicles Symposium*, 2012, pp. 270–275.
- [68] Z. Y. Rawashdeh and Z. Wang, "Collaborative automated driving: A machine learning-based method to enhance the accuracy of shared information," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, 2018, pp. 3961–3966.
- [69] Z. Xiao, Z. Mo, K. Jiang, and D. Yang, "Multimedia fusion at semantic level in vehicle cooperative perception," in *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2018, pp. 1–6.
- [70] S.-W. Kim, B. Qin, Z. J. Chong, X. Shen, W. Liu, M. H. Ang, E. Frazzoli, and D. Rus, "Multivehicle cooperative driving using cooperative perception: Design and experimental validation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 2, pp. 663–680, 2014.
- [71] Y. Yan, Y. Mao, and B. Li, "Second: Sparsely embedded convolutional detection," *Sensors*, vol. 18, no. 10, p. 3337, 2018.
- [72] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.
- [73] 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 *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*. IEEE, 2020, pp. 1–7.
- [74] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, and Y. Bengio, "Maxout networks," in *International conference on machine learning*. PMLR, 2013, pp. 1319–1327.
- [75] 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 *European Conference on Computer Vision*. Springer, 2020, pp. 605–621.
- [76] S. Manivasagam, S. Wang, K. Wong, W. Zeng, M. Sazanovich, S. Tan, B. Yang, W.-C. Ma, and R. Urtasun, "Lidarsim: Realistic lidar simulation by leveraging the real world," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 167–11 176.
- [77] Z. Zhang, S. Wang, Y. Hong, L. Zhou, and Q. Hao, "Distributed dynamic map fusion via federated learning for intelligent networked vehicles," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2021, pp. 953–959.
- [78] R. Xu, H. Xiang, X. Xia, X. Han, J. Liu, and J. Ma, "Opv2v: An open benchmark dataset and fusion pipeline for perception with vehicle-to-vehicle communication," *arXiv preprint arXiv:2109.07644*, 2021.
- [79] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [80] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *ICLR*, 2021.
- [81] Z. Hu, Y. Dong, K. Wang, and Y. Sun, "Heterogeneous graph transformer," in *Proceedings of The Web Conference 2020*, 2020, pp. 2704–2710.
- [82] A. Ziebinski, R. Cupek, H. Erdogan, and S. Waechter, "A survey of adas technologies for the future perspective of sensor fusion," in *International Conference on Computational Collective Intelligence*. Springer, 2016, pp. 135–146.
- [83] 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 IV2003 Intelligent Vehicles Symposium. Proceedings (Cat. No. 03TH8683)*. IEEE, 2003, pp. 304–309.
- [84] E. Santana and G. Hotz, "Learning a driving simulator," *arXiv preprint arXiv:1608.01230*, 2016.
- [85] P. Sun, H. Kretschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 2446–2454.
- [86] Waymo, "Introducing the 5th-generation waymo driver: Informed by experience, designed for scale, engineered to tackle more environments." Available: <https://blog.waymo.com/2020/03/introducing-5th-generation-waymo-driver.html>, 2022.
- [87] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

- [88] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, vol. 1. Ieee, 2005, pp. 886–893.
- [89] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.
- [90] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *IEEE transactions on pattern analysis and machine intelligence*, vol. 24, no. 4, pp. 509–522, 2002.
- [91] P. Felzenszwalb, D. McAllester, and D. Ramanan, "A discriminatively trained, multiscale, deformable part model," in *2008 IEEE conference on computer vision and pattern recognition*. Ieee, 2008, pp. 1–8.
- [92] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester, "Cascade object detection with deformable part models," in *2010 IEEE Computer society conference on computer vision and pattern recognition*. Ieee, 2010, pp. 2241–2248.
- [93] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.
- [94] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.
- [95] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142–158, 2015.
- [96] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.
- [97] R. Girshick, "Fast R-CNN," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.
- [98] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 779–788.
- [99] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European conference on computer vision*. Springer, 2016, pp. 21–37.
- [100] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 2117–2125.
- [101] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.
- [102] S. M. Ahmed and C. M. Chew, "Density-based clustering for 3d object detection in point clouds," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10608–10617.
- [103] Z. Zhang, L. Zhang, X. Tong, P. T. Mathiopoulos, B. Guo, X. Huang, Z. Wang, and Y. Wang, "A multilevel point-cluster-based discriminative feature for als point cloud classification," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 54, no. 6, pp. 3309–3321, 2016.
- [104] Z. Zhang, J. Zheng, H. Xu, and X. Wang, "Vehicle detection and tracking in complex traffic circumstances with roadside lidar," *Transportation research record*, vol. 2673, no. 9, pp. 62–71, 2019.
- [105] S. Shi, X. Wang, and H. Li, "Pointcnn: 3d object proposal generation and detection from point cloud," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 770–779.
- [106] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "Pointnet++: Deep hierarchical feature learning on point sets in a metric space," *Advances in neural information processing systems*, vol. 30, 2017.
- [107] S. Shi, C. Guo, L. Jiang, Z. Wang, J. Shi, X. Wang, and H. Li, "Pv-rcnn: Point-voxel feature set abstraction for 3d object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10529–10538.
- [108] D. Z. Wang and I. Posner, "Voting for voting in online point cloud object detection," in *Robotics: Science and Systems*, vol. 1, no. 3. Rome, Italy, 2015, pp. 10–15.
- [109] S. Suthaharan, "Support vector machine," in *Machine learning models and algorithms for big data classification*. Springer, 2016, pp. 207–235.
- [110] B. Liu, M. Wang, H. Foroosh, M. Tappen, and M. Pensky, "Sparse convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 806–814.
- [111] M. Engelcke, D. Rao, D. Z. Wang, C. H. Tong, and I. Posner, "Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 1355–1361.
- [112] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *ICML*, 2010.
- [113] J. Deng, S. Shi, P. Li, W. Zhou, Y. Zhang, and H. Li, "Voxel rcnn: Towards high performance voxel-based 3d object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 2, 2021, pp. 1201–1209.
- [114] B. Yang, W. Luo, and R. Urtasun, "Pixor: Real-time 3d object detection from point clouds," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, 2018, pp. 7652–7660.
- [115] H. Lu, X. Chen, G. Zhang, Q. Zhou, Y. Ma, and Y. Zhao, "Scanet: Spatial-channel attention network for 3d object detection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 1992–1996.
- [116] Z. Liu, H. Tang, A. Amini, X. Yang, H. Mao, D. Rus, and S. Han, "Befusion: Multi-task multi-sensor fusion with unified bird's-eye view representation," *arXiv preprint arXiv:2205.13542*, 2022.
- [117] 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.
- [118] X. Wang, "Intelligent multi-camera video surveillance: A review," *Pattern recognition letters*, vol. 34, no. 1, pp. 3–19, 2013.
- [119] R. Eshel and Y. Moses, "Homography based multiple camera detection and tracking of people in a dense crowd," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.
- [120] I. Mikic, S. Santini, and R. Jain, "Video processing and integration from multiple cameras," in *Proceedings of the 1998 Image Understanding Workshop, Morgan-Kaufman, San Francisco*, vol. 6. Citeseer, 1998.
- [121] K. Kim and L. S. Davis, "Multi-camera tracking and segmentation of occluded people on ground plane using search-guided particle filtering," in *European Conference on Computer Vision*. Springer, 2006, pp. 98–109.
- [122] B. Song and A. K. Roy-Chowdhury, "Robust tracking in a camera network: A multi-objective optimization framework," *IEEE Journal of Selected Topics in Signal Processing*, vol. 2, no. 4, pp. 582–596, 2008.
- [123] T. Huang and S. Russell, "Object identification in a bayesian context," in *IJCAI*, vol. 97. Citeseer, 1997, pp. 1276–1282.
- [124] K.-W. Chen, C.-C. Lai, Y.-P. Hung, and C.-S. Chen, "An adaptive learning method for target tracking across multiple cameras," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.
- [125] Y. Liu, Y. Wang, S. Wang, T. Liang, Q. Zhao, Z. Tang, and H. Ling, "Cbnet: A novel composite backbone network architecture for object detection," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 34, no. 07, 2020, pp. 11653–11660.
- [126] A. Farhadi and J. Redmon, "Yolov3: An incremental improvement," in *Computer Vision and Pattern Recognition*. Springer Berlin/Heidelberg, Germany, 2018, pp. 1804–2767.
- [127] S. Vora, A. H. Lang, B. Helou, and O. Beijbom, "Pointpainting: Sequential fusion for 3d object detection," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4604–4612.
- [128] A. Neubeck and L. Van Gool, "Efficient non-maximum suppression," in *18th International Conference on Pattern Recognition (ICPR'06)*, vol. 3. IEEE, 2006, pp. 850–855.
- [129] Y. Liu, A. Huang, Y. Luo, H. Huang, Y. Liu, Y. Chen, L. Feng, T. Chen, H. Yu, and Q. Yang, "Fedvision: An online visual object detection platform powered by federated learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 08, 2020, pp. 13172–13179.
- [130] F. e. Poletti, N. Wheeler, M. Petrovich, N. Baddela, E. Numkam Fokoua, J. Hayes, D. Gray, Z. Li, R. Slavík, and D. Richardson, "Towards high-capacity fibre-optic communications at the speed of light in vacuum," *Nature Photonics*, vol. 7, no. 4, pp. 279–284, 2013.
- [131] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nusenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11621–11631.
- [132] D. Yongqiang, W. Dengjiang, C. Gang, M. Bing, G. Xijia, W. Yajun, L. Jianchao, F. Yanming, and L. Juanjuan, "Baai-vanjee roadside dataset: Towards the connected automated vehicle highway technologies in challenging environments of china," *arXiv preprint arXiv:2105.14370*, 2021.

- [133] Z. Bai, “Carti dataset for cooperative perception,” Available: https://github.com/zwbai/CARTI_Dataset, 2022.
- [134] M. Contributors, “MMDetection3D: OpenMMLab next-generation platform for general 3D object detection,” <https://github.com/open-mmlab/mmdetection3d>, 2020.
- [135] H. Yu, Y. Luo, M. Shu, Y. Huo, Z. Yang, Y. Shi, Z. Guo, H. Li, X. Hu, J. Yuan *et al.*, “Dair-v2x: A large-scale dataset for vehicle-infrastructure cooperative 3d object detection,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 21 361–21 370.
- [136] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez, “Deep reinforcement learning for autonomous driving: A survey,” *IEEE Transactions on Intelligent Transportation Systems*, 2021.
- [137] M. Behrisch, L. Bieker, J. Erdmann, and D. Krajzewicz, “Sumo-simulation of urban mobility: an overview,” in *Proceedings of SIMUL 2011, The Third International Conference on Advances in System Simulation*. ThinkMind, 2011.
- [138] M. Fellendorf and P. Vortisch, “Microscopic traffic flow simulator vissim,” in *Fundamentals of traffic simulation*. Springer, 2010, pp. 63–93.
- [139] J. Barceló and J. Casas, “Dynamic network simulation with aimsun,” in *Simulation approaches in transportation analysis*. Springer, 2005, pp. 57–98.
- [140] A. Juliani, V.-P. Berges, E. Teng, A. Cohen, J. Harper, C. Elion, C. Goy, Y. Gao, H. Henry, M. Mattar *et al.*, “Unity: A general platform for intelligent agents,” *arXiv preprint arXiv:1809.02627*, 2018.
- [141] B. Karis and E. Games, “Real shading in unreal engine 4,” *Proc. Physically Based Shading Theory Practice*, vol. 4, no. 3, 2013.
- [142] G. Rong, B. H. Shin, H. Tabatabaee, Q. Lu, S. Lemke, M. Možeiko, E. Boise, G. Uhm, M. Gerow, S. Mehta *et al.*, “Lgsvl simulator: A high fidelity simulator for autonomous driving,” *arXiv preprint arXiv:2005.03778*, 2020.
- [143] S. Shah, D. Dey, C. Lovett, and A. Kapoor, “Airsim: High-fidelity visual and physical simulation for autonomous vehicles,” in *Field and service robotics*. Springer, 2018, pp. 621–635.
- [144] S. Kato, S. Tokunaga, Y. Maruyama, S. Maeda, M. Hirabayashi, Y. Kitsukawa, A. Monrroy, T. Ando, Y. Fujii, and T. Azumi, “Autoware on board: Enabling autonomous vehicles with embedded systems,” in *2018 ACM/IEEE 9th International Conference on Cyber-Physical Systems (ICCPs)*. IEEE, 2018, pp. 287–296.
- [145] F. Graf, *Apollo*. Routledge, 2008.
- [146] A. Koubâa, A. Allouch, M. Alajlan, Y. Javed, A. Belghith, and M. Khalgui, “Micro air vehicle link (mavlink) in a nutshell: A survey,” *IEEE Access*, vol. 7, pp. 87 658–87 680, 2019.
- [147] J. Zhang, F.-Y. Wang, K. Wang, W.-H. Lin, X. Xu, and C. Chen, “Data-driven intelligent transportation systems: A survey,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.
- [148] 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,” *Computers & Graphics*, vol. 99, pp. 153–181, 2021.



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 a Ph.D. student in electrical and computer engineering at 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 in Urban Transportation Systems and Mobility.



Guoyuan Wu (Senior Member, IEEE) received his Ph.D. degree in mechanical engineering from the University of California, Berkeley in 2010. Currently, he holds an Associate Researcher and an Associate Adjunct Professor position at Bourns College of Engineering – Center for Environmental Research & Technology (CE-CERT) and Department of Electrical & Computer Engineering in the University of California at Riverside. development 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. Dr. Wu serves as Associate Editors for a few journals, including IEEE Transactions on Intelligent Transportation Systems, SAE International Journal of Connected and Automated Vehicles, and IEEE Open Journal of ITS. 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 is a recipient of Vincent Bendix Automotive Electronics Engineering Award.



Matthew J. Barth (Fellow, IEEE) received the M.S. and Ph.D degree 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 serving as the Director for 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. Dr. Barth has been active in the IEEE Intelligent Transportation System Society for many years, serving as a Senior Editor for both the Transactions of ITS and the 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 Ph.D. and M.S. degrees in electrical engineering from the University of Texas at Dallas in 2021 and 2017, respectively. He is currently a Research Engineer at Toyota Motor North America, InfoTech Labs. 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 Postdoctoral Research Fellow at LAAS-CNRS, Toulouse, France, and at the University of Washington, Seattle. He is currently a Principal Engineer with Toyota Motor North America, InfoTech Labs, Mountain View, CA. 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 a Director at Toyota Motor North America, InfoTech Labs. Oguchi's 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. Prior, he worked as a senior researcher at Toyota Central R&D Labs in Japan.



Zhitong Huang is senior transportation research scientist and Analysis, Simulation, and Modeling program manager at Leidos. He has 17 years of research experience and conducted dozens of research projects in the field of transportation engineering. His main focus is on transportation simulation and modeling, connected and automated vehicle (CAV) systems, traffic operation and management, and digital twin, etc.