INVITED PAPER

Mobile Edge Intelligence and Computing for the Internet of Vehicles

This article overviews the edge information system (EIS), including edge caching, edge computing, and edge AI, which will enable a plethora of new exciting intelligent IoV applications.

By JUN ZHANG, *Senior Member IEEE*, AND KHALED B. LETAIEF, *Fellow IEEE*

ABSTRACT | The Internet of Vehicles (IoV) is an emerging paradigm that is driven by recent advancements in vehicular communications and networking. Meanwhile, the capability and intelligence of vehicles are being rapidly enhanced, and this will have the potential of supporting a plethora of new exciting applications that will integrate fully autonomous vehicles, the Internet of Things (IoT), and the environment. These trends will bring about an era of intelligent IoV, which will heavily depend on communications, computing, and data analytics technologies. To store and process the massive amount of data generated by intelligent IoV, onboard processing and cloud computing will not be sufficient due to resource/power constraints and communication overhead/latency, respectively. By deploying storage and computing resources at the wireless network edge, e.g., radio access points, the edge information system (EIS), including edge caching, edge computing, and edge AI, will play a key role in the future intelligent IoV. EIS will provide not only low-latency content delivery and computation services but also localized data acquisition, aggregation, and processing. This article surveys the latest development in EIS for intelligent IoV. Key design issues, methodologies, and hardware platforms are introduced. In particular, typical use cases for intelligent vehicles are illustrated, including edge-assisted perception, mapping, and localization. In addition, various open-research problems are identified.

KEYWORDS | Autonomous driving; edge AI; Internet of Vehicles (IoV); mobile edge computing (MEC); vehicular communications; wireless caching

I. INTRODUCTION

The automobile industry has been one major economic sector for over a century, and its economical and societal impacts continue to expand. For example, automakers and their suppliers are responsible for 3% of the United States GDP, and no other manufacturing sector generates as many jobs in the United States [1]. Due to the increasing convenience, comfort, low cost, and fuel efficiency, there have been more and more vehicles on the road. In 2018, China and the United States, the two largest automobile markets, sold 23.2 and 17.2 million passenger cars, respectively [2]. The increasing number of vehicles has caused various issues, such as traffic congestion, accidents, and air pollution. According to the Global Status Report on Road Safety 2018 of the World Health Organization (WHO) [3], the number of road traffic deaths has reached 1.35 million in 2016. In other words, there are nearly 3700 fatalities on the world's roads every day.

Significant efforts have recently been spent on improving vehicle safety and efficiency. In particular, information and communication technologies have been regarded as promising tools to revolutionize vehicular networks. Connecting vehicles via vehicular *ad hoc* networks (VANETs) represents an early attempt to support safety-related applications, such as accident warning, crash notification, and cooperative cruise control [4], [5]. With VANET, neighboring vehicles are allowed to communicate with each other via vehicle-to-vehicle (V2V) communications, which helps to improve driving safety. Vehicles can

Manuscript received June 2, 2019; revised September 30, 2019; accepted October 12, 2019. Date of publication October 28, 2019; date of current version January 22, 2020. This work was supported in part by the General Research Funding from the Research Grants Council of Hong Kong under Project No. 16209418 and in part by a start-up fund of The Hong Kong Polytechnic University under Project ID P0013883. *(Corresponding author: Jun Zhang.)*

J. Zhang is with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong (e-mail: jun-eie.zhang@polyu.edu.hk).

K. B. Letaief is with the Department of Electronic and Computer Engineering, The Hong Kong University of Science and Technology, Hong Kong, and also with the Peng Cheng Laboratory, Shenzhen 518066, China (e-mail: eekhaled@ust.hk).

Digital Object Identifier 10.1109/JPROC.2019.2947490

also communicate with the roadside infrastructure via vehicle-to-infrastructure (V2I) communications to collect road and traffic-related information. These communication links are enabled by dedicated short-range communications (DSRCs) or cellular-enabled vehicle-to-everything (V2X) communications [6]–[8].

Internet access is not fully available in VANET, which limits the scope of its applications. To extend the capabilities of VANET, the Internet of Vehicles (IoV) has been proposed to form a global network of vehicles, evoking collaborations between heterogeneous communication systems to provide reliable Internet services [9], [10]. IoV will have communications, processing, storage, and learning capabilities. In particular, with IoV, vehicles will be able to leverage resources, such as cloud storage and computing. Besides vehicle driving and safety, IoV will also facilitate urban traffic management, vehicle insurance, road infrastructure construction and repair, logistics and transportation, and so on. As a special case of the Internet of Things (IoT), IoV shall be integrated with other systems, such as the smart city.

Meanwhile, we are witnessing the growing intelligence of vehicles, due to recent advancements in embedded systems, navigation, sensors, data acquisition and dissemination, and big data analytics [11], [12]. It started with assisted-driving technologies, namely, advanced driver-assistance systems (ADAS), including emergency braking, backup cameras, adaptive cruise control, and self-parking systems [13]. Around the world, the number of ADAS systems rose from 90 million units in 2014 to about 140 million in 2016, a 50% increase in just two years [14]. According to the definitions of autonomous vehicles laid out by the Society of Automotive Engineers (SAE) International, the above-mentioned systems mainly belong to Level 1 and Level 2 of automation. Tesla's Autopilot system also falls in this category [15]. Automotive manufacturers and technical companies, such as Google, Uber, Tesla, and Mobileye, are investing heavily on higher levels of driving autonomation. The 2018 Audi A8 is the first Level 3 self-driving car available in production [16]. Predictions vary, but many forecast that autonomous vehicles with Level 4 and Level 5 will be available in the market within a decade.

The upcoming intelligent IoV will need support from various sectors, including automobile, transportation, wireless communications, networking, security, and robotics, as well as regulators and policymakers. This survey shall provide a perspective from information and communication technologies on intelligent IoV. In particular, we will advocate that the integration of storage, communications, computing, and data analytics at the wireless network edge, e.g., radio access points, provides an effective framework in addressing the data acquisition, aggregation, and processing challenges for intelligent IoV. In the following, we first elaborate on the big data challenges in intelligent IoV and motivate the need for an edge information system (EIS).

A. Big Data in Intelligent IoV

The advancements in information technologies, including communication, sensing, data processing, and control, are transforming the transportation system from conventional technology-driven systems into more powerful data-driven intelligent transportation systems [17]. This trendy movement will generate a tremendous amount of data. Over the past two decades, the wireless industry has been struggling with the mobile data explosion brought by smartphones [18]. Such a struggle will be dwarfed by the expected huge amount of data to be generated by intelligent IoV. Intelligent vehicles are equipped with multiple cameras and sensors, including radar, light detection and ranging (LiDAR) sensors, sonar, and global navigation satellite systems (GNSS). It is predicted that there will be more than 200 sensors in the future vehicles [19], with total sensor bandwidth reaching 3 Gb/s ($\sim$1.4 TB/h) to 40 Gb/s ($\sim$19 TB/h) [20]. As estimated by Intel, each autonomous vehicle will be generating approximately 4000 GB of data a day, equivalent of the mobile data generated by almost 3000 people. Assuming a mere one million autonomous cars worldwide, then automated driving will be equivalent to the data of three billion people. Due to this huge surge, Brian Krzanich, CEO of Intel, remarked that "data is the new oil in the future of automated driving" [21].

The big data generated by intelligent IoV will place unprecedented pressure on communication, storage, and computing infrastructures. While onboard computing and storage capabilities are increasing rapidly, they are still limited compared with the scale of data to be stored and processed. For example, NVIDIA's self-driving learning data collection system adopts solid-state drive (SSD) as the external storage, up to several terabytes, which will be filled within hours by sensing data. Furthermore, the computation needed to process these data will easily exhaust the onboard computing resources. A car equipped with ten high-resolution cameras can generate 2 Gpixels/s of data. Processing that amount of data through multiple deep neural networks (DNNs) converts to approximately 250 trillion operations per second (TOPS) [22]. Meanwhile, to achieve better safety than the best human driver who takes action within 100–150 ms, autonomous driving systems should be able to process real-time traffic conditions within a latency of 100 ms [23], which demands a significant amount of computing power. While power-hungry accelerators, such as graphics processing units (GPUs), can provide low-latency computation, their high power consumption, further magnified by the cooling load to meet the thermal constraints, can significantly degrade the driving range and fuel efficiency of the vehicle [23].

There have been many proposals for using cloud computing [24] to help intelligent vehicles, and some of them have already been implemented, such as cloud-based software update or training powerful deep learning models [25]. Cloud computing platforms will certainly be

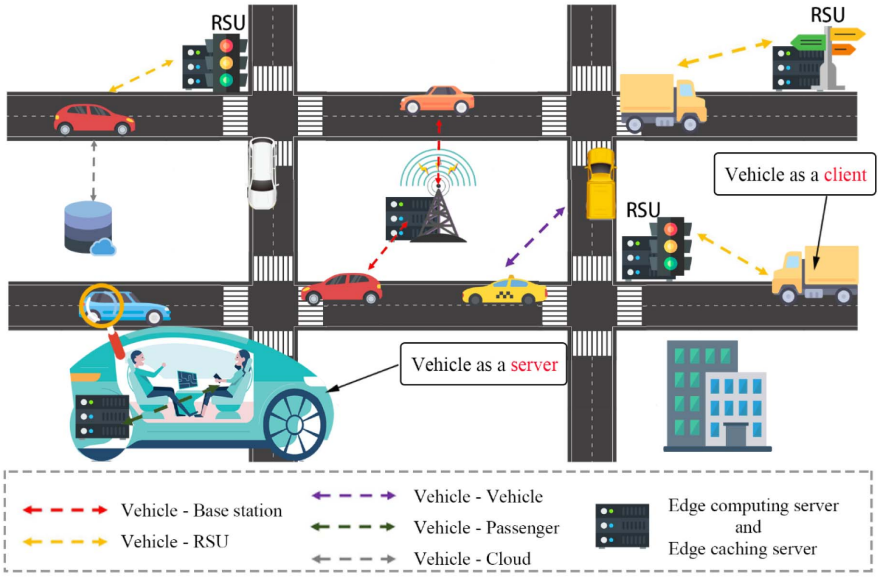

Fig. 1. ***Intelligent IoV supported by the EIS.***

an important supporter of IoV, but they are not sufficient. While cost and power consumption are the main limiting factors for onboard computation, the long latency and the massive data transmission are the bottlenecks for cloud-based processing [26]. The round trip time from a mobile client to a cloud center may easily be longer than 100 ms [27]. Moreover, such latency highly depends on the wireless channel condition, the bandwidth of the network, and traffic congestion, so real-time processing and reliability cannot be ensured. As shown in [27], considering the latency requirement, if the speech recognition task of a driving assistance system is to be offloaded, the server has to be located at the nearby base station (BS), i.e., at the edge of the wireless network. This is aligned with the recent trend to deploy computing resources at the edge of wireless networks [28]–[31], to be elaborated in the following.

B. Living on the Edge

To overcome the limited capabilities of onboard computing, communication, storage, and energy while avoiding excessive latency in cloud computing, deploying resources at the wireless network edge has received significant attention from both academia and industry. Popular content, such as video files, which dominates mobile data traffic, is likely to be repeatedly requested by different users, and such requests are predictable. Thus, deploying storage units and caching popular content at the wireless network edge, i.e., wireless edge caching, stands out as a promising solution for efficient content delivery [32]–[34]. Meanwhile, the revival of artificial intelligence (AI) and the emergence of intelligent mobile applications demand platforms that can support computation-intensive and delay-sensitive mobile computing. Mobile edge computing (MEC) is an emerging technology that has the potential to unite telecommunications with cloud computing to deliver cloud services directly from the network edge and support delay-critical mobile applications. This is achieved by placing computer servers at the BSs or radio access points [28]–[31], [35]. Edge caching and computing platforms further enable edge AI, which trains and deploys powerful machine learning models at the edge servers and mobile devices, and has been regarded as a key supporting technology for IoT [36], [37]. Edge AI is changing the landscape of the semiconductor industry. In 2018, shipment revenues from edge AI reached $1.3 billion, and by 2023, this figure is forecast to reach $23 billion [38]. Collectively, these platforms will be referred to as the EIS in this article.

EIS is a perfect fit for intelligent IoV. It is able to assist the key functionalities of intelligent vehicles, from data acquisition (for situational and environmental awareness), data processing (for navigation and path planning), to actuation (maneuver control), as illustrated in Fig. 1. Processing the data at the network edge can save a significant amount of communication bandwidth and also satisfy the low latency requirement for mission-critical tasks. Content in IoV is typically with high spatial locality, e.g., road conditions and map information are mainly used locally, and temporal locality, e.g., traffic conditions in the morning will be of little relevance for the evening. Moreover, vehicles are only interested in content itself, not its provenance. These key features make cache-assisted content-centric dissemination and delivery highly effective for IoV [39], [40]. On the other hand, with big sensing data, intelligent vehicles are facing tremendous computation burdens. For example, computational capability remains the bottleneck that prevents vehicles from benefiting from the high system accuracy enabled by higher resolution cameras. Specifically, the convolution tasks in the powerful

convolutional neural network (CNN) [41] for vision-based perception and the feature extraction tasks for vision-based localization are highly complicated [42]. Offloading such computation-intensive tasks to the proximate MEC servers will enable powerful machine learning methods to assist key tasks of intelligent vehicles [43].

While this article focuses on EIS, intelligent IoV shall leverage different available information processing platforms.

1) *Onboard processing* is used for highly latency-sensitive tasks, such as real-time decision-making for vehicle control, and for preprocessing sensing data to reduce communication bandwidth.
2) *Edge severs* are for latency-sensitive and computation-intensive tasks, such as localization and mapping, and for aggregating and storing local information, such as the area's high-definition (HD) map.
3) *Cloud computing* is for training powerful deep learning models with massive data sets, acts as a nonreal-time aggregator for wide-area information, and stores valuable historic data for continuous learning.

EIS shall play a vital and unique role in the information infrastructure for intelligent IoV. There has been no systematic survey covering this important topic. We intend to fill the gap, with a comprehensive and in-depth introduction of EIS for intelligent IoV. Specifically, Section II presents a general introduction of EIS for IoV. Section III introduces edge caching systems for IoV, focusing on cache placement and delivery, as well as cache-enabled applications for intelligent IoV. Section IV introduces MEC platforms for IoV and MEC-enabled applications. Section V describes edge AI frameworks and illustrates how EIS helps key tasks in intelligent IoV. Finally, Section VI concludes this article.

II. EDGE INFORMATION SYSTEM FOR INTERNET OF VEHICLES

This section first introduces EIS for intelligent IoV, considering two scenarios: Vehicle as a Client (VaaC) and Vehicle as a Server (VaaS). Then, three key tasks for intelligent vehicles are presented as key application cases, including perception, HD mapping, and simultaneous localization and mapping (SLAM).

A. Edge Information System

EIS helps to acquire, aggregate, and process data for intelligent vehicles. Within IoV, it acts as an intermediary platform between onboard processors and the remote cloud data center. As shown in Fig. 1, an EIS contains the following main components.

Edge Servers: These are computer servers deployed at BSs or roadside units (RSUs), equipped with storage units, such as SSD, and computing units, such as GPUs or edge tensor processing units (TPUs) [44]. They are connected to the backbone network with high-capacity links and are capable to communicate with vehicles within their coverage ranges via V2I communications. BS servers have larger coverage ranges and, thus, have a better capability for mobility management. On the other hand, RSU servers are closer to vehicles and, thus, can support lower latency. Both types of edge servers are aware of the local environment, can collect and process data from vehicles passing by, e.g., to update HD maps or monitor traffic, and disseminate content such as road condition, traffic information, and HD maps.

Vehicles: Equipped with various sensors, communication modules, and onboard units with computing and storage capabilities, intelligent vehicles are powerful nodes [45]. The onboard computing stack must simultaneously achieve high performance, consume minimal power, and have low thermal dissipation, at an acceptable cost [42]. Vehicles are able to communicate with each other via V2V communications or with the RSUs via V2I communications. The throughput of V2I communications is typically larger than that of V2V communications [46].

User Devices: User devices are of various types, e.g., passengers' smartphones and wearable devices. They are typically with limited computing power, storage space, and power supply. High-end devices may be equipped with AI chips, but the computing power is still limited compared with onboard or edge server processors. These devices will generate user-specific data that can be used to improve driving experience and safety.

V2X Communications: The communication module is an essential component of EIS. There two types of communication standards specialized for IoV, i.e., WLAN- and cellular-based protocol families [6]–[8]. The specification of WLAN-based vehicular communication protocols, called DSRC, has gained great attention from academia and industry for nearly a decade. Lately, with the fast development of mobile communications, V2X has been regarded as an indispensable application scenario for cellular networks, which is called C-V2X since 3GPP Release 14. The timeline of C-V2X includes three phases. Phase 1 is developed over LTE technologies, and Phase 2 and Phase 3 are based on 5G NR in order to support more advanced onboard applications. A brief comparison between DSRC and C-V2X is shown in Table 1.

In the remainder of this article, we shall consider two different scenarios, depending on the role of vehicles, as specified in the following.

1) Vehicle as a Client: First, vehicles may act as clients to access the edge resources at the RSUs or BSs. The key idea is to colocate data acquisition and processing. Edge servers act as anchor nodes for data acquisition and then process data for local applications. For example, they can collect mapping data from vehicles passing by to build and update HD maps and can actively monitor the road condition and traffic in the local area. Such applications are highly relevant for IoV.

2) Vehicle as a Server: Vehicles can also work as mobile service providers for vehicle passengers, third-party recipients, and other vehicles, which can improve user

Table 1 Different V2X Technologies

	Specification	Time	Bandwidth	Latency	Reliability	Use Cases
WLAN-Based	DSRC: SAE J2735, IEEE 1609 series, IEEE 802.11-2012	2012	10/20 MHz	50 ms	Out of coverage	Road safety, infotainment, traffic efficiency.
Cellular-Based	Phase 3 (3GPP Release 16)	2019	10/20 MHz and wideband (e.g., 40/60 MHz)	Ultra-low latency (∼1ms)	Ultra-high reliable	Advanced V2X use cases, e.g., vehicles platooning, extended sensors, advanced driving, remote driving.
	Phase 2 (3GPP Release 15)	2018	10/20 MHz	3∼10 ms	High reliable ($1-10^{-5}$)	
	Phase 1 (3GPP Release 14)	2016	10/20 MHz	100 ms/20 ms	Out of coverage	Safety-related (e.g., collision warning), non-safety-related (cooperative adaptive cruise control), enhanced positioning.

experience, e.g., to enable personalized driving experience via driver identification and to enable rich infotainment applications. Compared with edge server-based approaches, VaaS suffers less from mobility. Moreover, it also allows cooperation among neighboring vehicles, e.g., to enable cooperative perception and cooperative driving.

B. Key Tasks of Intelligent Vehicles

To introduce EIS for intelligent IoV, we shall focus on a few key tasks of intelligent vehicles. The foundation of intelligent vehicles is the ability to understand the environment. Different onboard sensors are employed for different perception tasks, e.g., object detection/tracking, traffic sign detection/classification, and lane detection. The *a priori* knowledge, e.g., *a priori* maps, is also exploited. Based on the sensing data and perception outputs, localization and mapping algorithms are applied to calculate the global and local location of the vehicle and map the environment. The results from these tasks are then used for other functions, including decision-making, planning, and vehicle control, as illustrated in Fig. 2. In this article, we shall focus on perception, HD mapping, and SLAM as the main tasks for intelligent vehicles, as summarized in Table 2. More details on edge-assisted approaches for performance enhancement will be provided in Section V.

1) Perception: There are various kinds of onboard sensors, with different features and serving for different perception tasks. Cameras are used for object detection/classification, e.g., via the powerful CNN [41], and can also be used for vision-based localization [47]; LiDAR is applied for 3-D mapping and localization [48]; radar and sonar are used for obstacle detection. Each type of sensors has its limitations. Cameras and stereovision are computationally expensive compared with active sensors, such as LiDAR and radar, while LiDAR and radar are poor in classification and poor for very near (< 2 m) measurement, and sonar has a poor angular resolution.

Considering the budget and features of different technologies, developers may choose different combinations of sensors for an intelligent vehicle. For example, Waymo is using LiDAR-based technology, while Mobileye and Tesla are relying on cameras and sensors [49]. Currently, the limitations and high costs of available onboard sensors are one main reason that commercial vehicles only achieve Level 1 to Level 2 automation [50]. The perception of intelligent vehicles faces a few main challenges, such as perception in poor weather and lighting conditions, or in complex urban environments, and limited perception ranges. Techniques such as sensor fusion can be used to compensate for shortcomings of individual sensors, by exploiting sensing data from different sensors [51]–[53]. However, this will significantly increase the onboard computation.

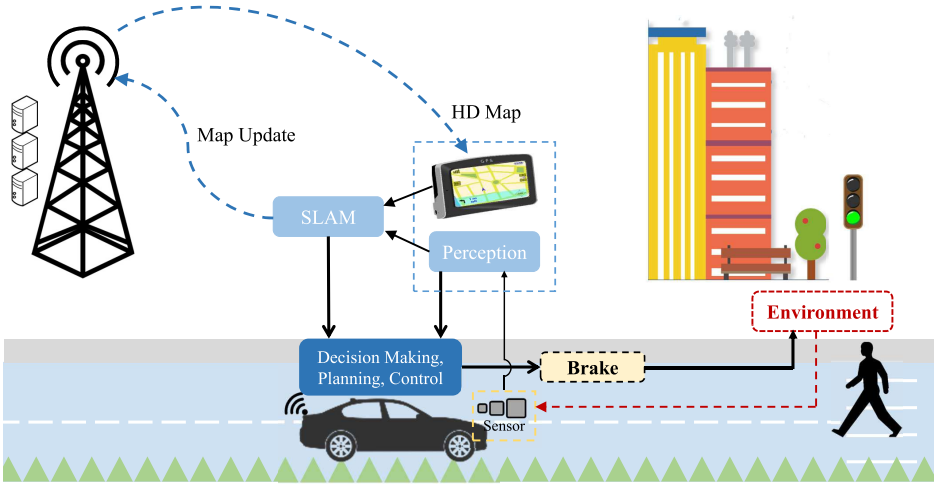

Fig. 2. ***Illustration of key tasks for intelligent vehicles, and how they act together for autonomous driving.***

Table 2 Key Tasks for Intelligent Vehicles

Tasks	Functions	Challenges	Edge-Assisted Approaches
Perception	Estimate the environment model with on-board sensors, e.g., object detection and tracking, lane detection.	High computation intensity for deep learning models; limited perception range.	Cooperative perception; edge-assisted deep learning for vision-based localization; edge-assisted multi-sensor fusion.
HD Mapping	For three-dimensional representation of all crucial aspects of a roadway.	Intensive works for building and updating maps; large storage space; significant communication overhead for dissemination.	Edge-assisted map building and update; caching-assisted data aggregation and map dissemination; multi-vehicle crowdsourced mapping.
SLAM	Simultaneous estimation of the location of a vehicle and the construction of the map.	High computational intensity; real-time execution.	Edge-assisted SLAM, multi-vehicle SLAM.

By providing additional proximate computation and storage resources, EIS is capable of enhancing perception capability. It can help to improve the sensing accuracy of cameras and stereovision, e.g., via powerful deep learning techniques, and enable sophisticated multisensor fusion, by offloading computation-intensive subtasks to edge servers. Furthermore, by sharing onboard sensing and computing power, assisted by V2V and V2I communications and coordinated by edge servers, cooperative perception can significantly improve the sensing robustness and accuracy and extend the perception range [54].

2) HD Mapping: Mapping is fundamental for any mobile robotics applications, and it is especially important for autonomous driving. Recently, HD mapping has received a lot of attention. An HD map models the surface of the road to an accuracy of 10–20 cm. It contains the 3-D representation of all crucial aspects of a roadway, e.g., slope and curvature, lane marking types, and roadside objects. Localization with HD maps can achieve centimeter-level precision. It uses onboard sensors to compare a vehicle's perceived environment with the corresponding HD map. This can overcome the limitations of GNSS-based methods (e.g., GPS), including low positioning accuracy and the varying availability. It is expected that HD map-based localization will be a common approach in Level 4 and Level 5 autonomous driving systems.

Great efforts from the industry have been put forward to build HD maps, e.g., TomTom [55] and HERE [56]. Nevertheless, there are practical implementation difficulties. Creating an HD map is time-consuming. To generate the HD map, specialized mapping vehicles equipped with a mobile mapping system (MMS) are needed, and the whole process involves three procedures: Data acquisition to acquire the mapping data, data accumulation to accumulate features acquired by the mapping vehicles, and data confirmation to manually refine and confirm the map. Moreover, HD maps are dynamic and need to be updated with timely changes [57]. Some HD map suppliers work with automakers to get fresh map data from intelligent vehicles, but this will substantially increase the onboard processing burden of vehicles [58].

Furthermore, the HD map data are of very large size, due to the high precision, and the rich geometric information and semantics [59]. This causes difficulties to transmit and store the HD maps. They are usually being served from the cloud, with a few nearby small areas of the map downloaded to the vehicle. The amount of data to download for a 3-D HD map with a centimeter-level accuracy can reach 3–4 GB/km. This not only introduces latency for the data download from the cloud but also causes a heavy burden on the backbone network.

With the inherent geographic locality, EIS will play an important role in HD mapping for intelligent IoV. Different edge-assisted approaches can be employed. Edge caching can help HD map dissemination as well as mapping data aggregation. Edge computing can assist map building and map change detection, by exploiting locally cached data. Edge servers can also coordinate vehicles driving through the area for crowdsourced mapping. In this way, by keeping and processing the data locally, and constructing the map where it is needed, more efficient HD mapping can be achieved.

3) SLAM: Map-based localization is effective for driving along roads that do not change often. However, if drastic changes occur, the loss in the accuracy may affect driving safety. SLAM comprises the simultaneous estimation of the state of a vehicle and the construction of a map of the environment [60]–[63]. It does not rely heavily on *a priori* information and allows vehicles to continuously observe the environment and readily adapt to new situations. To achieve full autonomy, it is a necessity for an intelligent vehicle to be able to perform accurate SLAM within its environment [61]. SLAM has been regarded as a key enabling technique for autonomous driving, and vehicles from the 2007 DARPA Urban Challenge have already used SLAM-based methods [64].

While many SLAM algorithms have been developed, they are mainly for indoor, highly structured environments. Autonomous vehicles need to operate in outdoor, variably lit, road-based environments, and thus, faster and more efficient algorithms are needed. For challenges of SLAM for autonomous driving, the readers are referred to [63] for a more detailed discussion. We shall mainly discuss the aspects that are related to EIS. Particularly, the computation demand for SLAM will be highly intensive for autonomous driving; 1 h of drive time can generate one terabyte of data, and interpreting one terabyte of collected data by means of high computing power requites two days to come up with usable navigation data [58]. Moreover, for real-time execution, latency must be

lower than 10 ms, which puts high pressure for onboard computing.

While cloud-based SLAM algorithms have been proposed to alleviate the computation burden of vehicles, the propagation latency will not meet the real-time execution requirement. The edge computing platform can resolve this difficulty, and it can help handle part of the computation-intensive subroutines [65]. Multivehicle SLAM with cooperation among vehicles can also help to improve the performance of SLAM [63]. More discussions will be provided in Section V.

III. EDGE CACHING FOR INTELLIGENT IoV

Edge caching has been adopted in IoV to assist content delivery by storing or prefetching content at edge servers [39]. The design of caching algorithms in IoV is more challenging compared with traditional networks, caused by the high mobility of vehicles, frequently changing content requirements, and harsh communication environments. In this section, we survey the existing research on caching for IoV with vehicles taking different roles, i.e., VaaC and VaaS. In the former scenario, vehicles act as content consumers in cache-enabled IoV to access the desired content from edge servers. In the latter scenario, vehicles also act as content providers by caching content at their storage units. We focus on cache content placement, i.e., to determine which content should be cached, and divide existing studies into three categories: 1) temporal-locality-aware caching, i.e., accounting for the temporal variation of the importance/popularity of contents; 2) spatial-locality-aware caching, i.e., considering different importance/popularity of the same content in different regions; and 3) mobility-aware caching, i.e., alleviating the effect of vehicle mobility on content caching and delivery.

A. Vehicle as a Client

With the wide deployment of edge servers, vehicles that drive through their coverage areas can receive timely content delivery services. The representative studies from traditional IoV to intelligent IoV are discussed in detail in the following.

1) Temporal-Locality-Aware Caching: In edge caching systems, temporal locality includes two aspects: The freshness of cached content and the temporal variation of user requests. By considering the characteristics of communications between RSUs and vehicles, the temporal data dissemination problem was shown in [66] as an NP-hard problem. The authors then exploited a heuristic scheduling algorithm according to the requirements of user requests (e.g., the time bound) in order to improve the request service chance. Focusing on temporal information services in IoV, a distributed edge caching mechanism was proposed in [67] based on the cooperation of RSUs and vehicles in order to optimize both the temporal data and real-time requests.

2) Spatial-Locality-Aware Caching: In vehicular environments, some information, e.g., traffic information, is associated with the location of vehicles. Therefore, the freshness of content in IoV may vary in different road segments. In [68], a deep-learning-based caching scheme was developed to optimize caching decisions for intelligent IoV, aiming at reducing the delivery delay of entertainment content. In this scheme, the content is cached at the edge servers in some specific areas by detecting the ages and genders of passengers with a CNN and predicting a proper content with a multilayer perception (MLP). The vehicles then determine which content could be accessed from edge servers based on a k-means algorithm and binary classification. Similarly, a data service scheme with the assistance of caching was developed in [69] by considering location-based vehicular services. The concept of road caching spot was applied to meet the service requirements and reduce the caching cost.

3) Mobility-Aware Caching: To deliver large-size content (e.g., videos, music, and HD maps) to moving vehicles is challenging because of limited network capacities and intermittent connections. To minimize the downloading time of vehicles, the problem of how to place large-size content at the edge servers was investigated in [70]. Three algorithms were developed to alleviate the impact of the mobility of vehicles on caching performance. In [71], a caching strategy was proposed to minimize the caching service delay in a multiaccess EIS. Specifically, the mobility of vehicles was predicted by a long short-term memory (LSTM) network over a time sequence. Based on the prediction, a deep reinforcement learning algorithm was then used to develop a proactive caching strategy. To tackle the mobility of vehicles and meet the service deadline, a new research direction was to integrate edge caching and computing in EIS. In this case, how to effectively allocate limited resources is an important problem. Focusing on resource allocation in the integrated architecture, two joint optimization models were formulated in [72] and [73] to determine the optimal caching and computing decisions, which were then solved by deep reinforcement learning-based methods.

B. Vehicle as a Server

Edge server-based caching is limited by the coverage range and unreliable connections with vehicles. Caching content on moving vehicles serves as a good complement. By exploiting the mobility of vehicles, edge caching can provide more cost-effective and utility-enhanced services. Existing studies related to this direction are presented as follows.

1) Temporal-Locality-Aware Caching: For vehicle caching, the temporal locality of content impacts not only caching services but also the implementation of other functions on vehicles due to the limited onboard storage resources. Therefore, to determine how long the

Table 3 Use Cases of Cache-Assisted Perception and Localization

Use Cases	Functions	KPIs	Cached Content
Autonomous Overtake	Guarantee safety distance between the overtaking vehicle and oncoming vehicle on two-way roads.	E2E delay: 10 ms, Reliability: 10^{-5}, Positioning accuracy: 30 cm.	Road information and vehicle intention.
Cooperative Collision Avoidance	Identify collision risks (such as in intersections) and inform vehicles in advance when traffic control mechanism fails.	E2E delay: 10 ms, Reliability: 10^{-3}, Positioning accuracy: 30 cm.	Vehicle driving trajectories.
See-Through	Obtain the view of blind spots obstructed by nearby vehicles in AR/VR.	E2E delay: 50 ms, Data rate: 10 Mbps.	Surveillance video/image for road segments.
Bird's Eye View	Provide the surveillance content of special road segments for approaching vehicles.	E2E delay: 50 ms, Data rate: 40 Mbps.	Surveillance video/image for road segments.
High Density Platooning	Form multiple vehicles into a linear chain by cooperative driving.	E2E delay: 10 ms, Reliability: 10^{-5}, Positioning accuracy: 30 cm.	Vehicle driving behaviors, situational awareness, lane changing information.
VRU Discovery	Detect vulnerable users by exchanging the localization information of vehicles and users.	Positioning accuracy: 10 cm.	Localization information of vehicles and pedestrians.

content will be cached is a vital problem. In [74], edge caching according to the content size was proposed. A new caching method, named Hamlet, was proposed to generate content diversity among adjacent nodes by determining caching updating frequency for large- and small-sized contents. Based on the proposal, users could receive different contents from nearby caching nodes in a short time, which improves the caching efficiency. Likewise, Malandrino *et al.* [75] focused on the freshness of the content. By studying the impact of the number of caching nodes on users, the authors optimized both the content freshness and the user downloading experiences.

2) Spatial-Locality-Aware Caching: The VaaS mode will improve the caching performance for the location-based services, due to the flexible mobility of vehicles and multihop data transmissions. Caching services in hot spot areas were investigated in [76]. Specifically, the urban areas were divided into many hot regions based on the dynamic mobility and density of vehicles. The driving traces of vehicles in the near future were predicted by partial matching based on the history data. By incorporating the vehicles that visit those hot regions frequently in a cooperative caching scheme, the optimal utility of caching services could be obtained. To mitigate the impact of mobility and communications vulnerability on caching services, a dynamic relay strategy for in-vehicle caching was developed in [77]. With the caching scheme and inter-vehicle communications, the survival of content in hot regions could be maintained.

3) Mobility-Aware Caching: The predictable mobility of vehicles can be exploited to improve the efficiency of cache-assisted content delivery. Mobility-aware caching in conventional device-to-device networks has been well studied, e.g., [78]–[80], and such methods have recently been extended to vehicular networks. A new type of caching services was explored in [81] and [82], where the content cached in vehicles could be requested by moving or static users within the communication range. In this scenario, the relationship between caching vehicles and moving users was the key to design the caching policy. A 2-D Markov process was proposed in [81] to model the interactions of caching vehicles and moving users, in order to determine the network availability of mobile users.

C. Cache-Enabled Applications

Besides typical content sharing and delivery services, there have been great interests in developing new applications enabled by edge cache servers. This section first introduces cache-assisted perception and localization, followed by other applications in IoV and intelligent transportation systems.

1) Cache-Assisted Perception and Localization: Cache-assisted perception includes such functions as autonomous overtake, cooperative collision avoidance, see-through, and bird's eye view, in which edge caching provides perception content for vehicles to assist driving and improve traffic safety. On the other hand, cache-assisted localization includes vulnerable road user (VRU) discovery, in which edge caching improves the cooperation of RSUs, vehicles, and pedestrians by caching positioning information. Table 3 shows the detailed description of these use cases as well as their key performance indicators (KPIs) in terms of end-to-end (E2E) delay, reliability, data rate, and positioning accuracy [83].

2) Other Applications: New applications enabled by edge caching keep emerging in IoV, some of them are introduced in the following.

InfoRank: For efficient urban sensing, Khan *et al.* [84] developed an information-based ranking (InfoRank) algorithm. This algorithm selects and ranks a part of intelligent vehicles to undertake urban sensing tasks. The vicinity monitoring of those vehicles, thus, could be completed with a small cost. In the algorithm, vehicles act as data cache servers to store the sensing data in order to alleviate the burden of edge servers.

Over-The-Top (OTT): A new OTT content prefetching system was designed in [85] by implementing an edge caching mechanism. The connections of vehicles and

RSUs are predicted based on a real-world test bed. A content popularity estimation scheme is also developed to estimate the content requests of users. After that, the requested content of users is proactively prefetched at edge servers.

Secure Information Sharing: Data sharing is an efficient way to reduce the data loss caused by unreliable sensor systems and to overcome the limited sensing range in autonomous vehicles. Data security, thus, becomes an important task, and Chowdhury *et al.* [86] designed a secure information sharing system for autonomous vehicles. The system aims to improve data security in two scenarios: false data dissemination and vehicle tracking.

Traffic Control: To analyze the impact of edge caching on traffic control, an edge caching-based transportation control scheme was developed in [87]. Traditionally, it is difficult to obtain the optimal state of the transportation system since drivers are selfish. As such, the optimal state of transportation networks and user equilibrium are contradictory. A communication cost model for cache-enabled vehicles was proposed to uncover the relationship between the user equilibrium and system optimal state. With the proposal, transportation networks could be optimized from a communication aspect with the assistance of edge caching.

IV. EDGE COMPUTING FOR INTELLIGENT IoV

With edge servers in close proximity to mobile users, MEC brings a number of important benefits, including ultralow latency, reduced mobile energy consumption, and enhanced privacy and security. This section introduces edge computing platforms for intelligent IoV. Available hardware platforms are introduced first, and key design problems are then discussed. Examples of MEC-enabled applications in IoV are also presented.

A. Vehicle as a Client

We first consider the VaaC scenario, where vehicles act as clients to access the computation resources at edge servers.

1) Edge Computing Platforms for Intelligent IoV: The concept of MEC was first proposed by ETSI in 2014 [28], and it has been regarded as a key component of the upcoming 5G network. Edge computing servers, equipped with GPUs or Edge TPUs [44], can be deployed at different edge nodes, and we mainly consider BSs and RSUs. The implementation of edge servers at BSs relies on several key techniques of 5G networks, such as the network virtualization architecture, network function virtualization, and virtual machine (VM) [28]. The virtualization layer aggregates the geographically distributed computing resources and presents them as a single resource pool for use by applications in the upper layers. Different applications share aggregated computation resources via VMs. Meanwhile, the capability of RSUs has been improved significantly. They normally adopt powerful multicore CPU and massive storage units.

There have been a lot of efforts in developing edge computing platforms specialized for vehicular data analytics. Open Vehicular Data Analytics Platform (OpenVDAP) [26] is an open-source platform. It is a full-stack edge-based platform, including an onboard computing/communication unit, an isolation-supported and security and privacy-preserved vehicle operation system, an edge-aware application library, as well as an optimal workload offloading and scheduling strategy. To evaluate different edge computing platforms, CAVBench was proposed in [88]. It is a benchmark suite for edge computing in connected and autonomous vehicles, including six applications: SLAM, objective detection, object tracking, battery diagnostics, speech recognition, and edge video analysis.

2) Resource Management: Effective resource allocation is essential for computation offloading in MEC. Such allocation faces a few critical issues. First, the stochastic nature of wireless channels and task arrivals should be considered [35], [89]. Second, the limited radio and computing resources are shared by multiple users, both of them will affect the computation latency [90], [91]. Finally, the mobility of vehicles will affect the task offloading and result feedback [92].

Stochastic optimization for resource management was considered in [93]. It minimizes the cost of both vehicles and the MEC server by jointly optimizing the offloading decision and local CPU frequency on the vehicle side, and the radio resource allocation and server provisioning on the server side. A contextual architecture was proposed in [94] for MEC in vehicular networks, which evaluates the available resources in real time and assigns the most logical and feasible resource to tasks.

Computation offloading decision-making among multiple vehicles was investigated in [95] by formulating it as a multiuser computation offloading game. The existence of the Nash equilibrium (NE) of the game was proven, and a distributed computation offloading algorithm was proposed to compute the equilibrium. If too many tasks were offloaded to the same edge server, the performance gain will be degraded. Load balancing among edge servers when designing the offloading decision was investigated in [96]. A joint load balancing and offloading problem was formulated as a mixed-integer nonlinear programming problem to maximize the system utility.

Mobility-aware resource management for MEC has also received a lot of attention. In [92], an online energy-aware mobility management scheme was developed, accounting for the radio handover and computation migration cost. An effective mobility-aware offloading decision algorithm was proposed in [97] by integrating mobility prediction. In [98], performance optimization under a long-term cost budget constraint was investigated. The Lyapunov optimization was applied, while the task migration cost was accounted for. Compared with the above-mentioned

results, the study in [99] was specifically for vehicle networks. It considered mobile devices in legacy vehicles, running infotainment applications, and offloading some computation to nearby intelligent vehicles. Due to high mobility, it is not efficient to always offload to one vehicle. It, thus, proposed an edge server relaying scheme to better utilize the computation resources on the road.

B. Vehicle as a Server

With powerful onboard processing capabilities, vehicles can act as servers to provide computation services for passengers or cooperate with other vehicles. For this purpose, incentive mechanisms are needed to encourage vehicles to share resources. Related studies are surveyed in the following.

1) Vehicular Cloud: Onboard computing is getting more and more powerful for intelligent vehicles [42]. In particular, different platforms have been developed for autonomous driving. For example, the NVIDIA DRIVE platform includes an in-vehicle computer (DRIVE AGX) and a complete reference architecture (DRIVE Hyperion), as well as data center-hosted simulation (DRIVE Constellation) and a DNN training platform (DGX) [100]. DRIVE AGX is built on NVIDIA Xavier, the world's first processor designed for autonomous driving. Six types of processors work together inside Xavier: an image signal processor, a video processing unit, a programmable vision accelerator, a deep learning accelerator, a CUDA GPU, and a CPU. Together, they process nearly 40 TOPS, among them 30 trillion operations are for deep learning alone.

Inspired by the success and flexibility of cloud computing in providing on-demand resources and services, the concept of vehicular cloud arises, which is to leverage onboard vehicular resources, such as network connectivity, computational power, storage, and sensing capability [101]. It can enable various applications, such as traffic management, urban surveillance, and emergency management. High traffic mobility is a major challenge in implementing a vehicular cloud. By analyzing the existing traffic models, it was shown in [102] that vehicular cloud computing is technologically feasible in dynamic scenarios, e.g., highways. Given the potential applications of vehicular cloud, a lot of studies have been carried to address its design challenges and implementation issues. Different resource management problems have been studied, including scheduling [103], VM migration [104], and computation resource allocation [105].

2) Incentive Mechanisms: To utilize onboard computation resources of intelligent vehicles to assist passenger devices or other vehicles, effective incentive mechanisms are needed. There have been many studies in incentivizing players to share resources in other domains [106], and recently, extensions to intelligent vehicles have been investigated.

The RSU servers are sparsely deployed and constrained by their radio coverage. To overcome such limitation, it was proposed in [107] to utilize the vacant computing power at vehicles. A market mechanism was developed to incentivize nearby vehicles to contribute their computing power. Distributed task allocation algorithms for cost minimization were developed. A similar study was carried in [108], where vehicles act as server nodes to help with computation tasks for the MEC server at the BS. In [109], a market mechanism was designed for computation offloading to incentivize vehicles to share resources. A Vickrey–Clarke–Groves (VCG)-based reverse auction mechanism was developed.

C. MEC-Enabled Applications in IoV

The availability of proximate MEC servers has inspired new applications in intelligent IoV, with some examples given in the following.

1) Driver Identification: Driver-specific applications are important for shared vehicles that are used by multiple drivers. An MEC system was built in [110] to collect and analyze the in-vehicle data for driver identification. The system can be used for applications such as personalization of vehicle settings (e.g., automatically adjusting entertainment, preferred temperature, and configurations to driver preferences), automated vehicle use logs, driver-dependent pay-as-you-drive insurance, and unauthorized vehicle use detection.

2) Real-Time Traffic Estimation: Traffic estimation is an important problem in intelligent transportation systems. Existing works rely on traffic surveillance cameras, which are not available on many roads, or GPS-based speed estimation, which only provides coarse estimates. Furthermore, considering the bandwidth and latency challenges, real-time traffic estimation should be near the edge and on the vehicles themselves. With the help of a front-facing camera, an MEC-assisted automated traffic estimation framework was developed in [110] for vehicle detection, vehicle tracking, and traffic estimation. The effectiveness of the system has been tested through multiple days of roadway experiments.

3) Public Safety: Video analytics for public safety (VAPS) is an important application case for IoT. Due to the limitations in budgets, size, weight, and power, as well as the complexity of public safety operations, analytic integration and optimization for VAPS is a significant challenge. In [111], built upon the OpenVDAP platform [26], an IoT-enabled public safety service, called AutoVAPS, was proposed. It integrates body-worn cameras and other public safety sensors and consists of three layers: a data layer for data management, a model layer for edge intelligence, and an access layer for privacy-preserving data sharing and access.

V. EDGE AI FOR INTELLIGENT IoV

In intelligent IoV, edge AI involves training powerful machine learning models and data analytics for key tasks

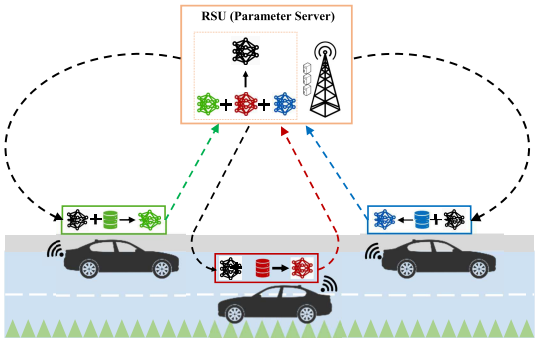

Fig. 3. *Federated learning for collaborative privacy-preserving machine learning.*

of intelligent vehicles. This section first introduces two edge AI frameworks for collaborative training and joint inference. Then, key use cases in intelligent IoV are illustrated, including edge-assisted perception, mapping, and SLAM.

A. Edge AI Frameworks

In this part, we introduce two edge AI frameworks to illustrate the training and inference stages, respectively.

1) Federated Learning: One main challenge of edge AI is to train machine learning models by aggregating a large amount of data that are distributed at different edge devices, including vehicles and onboard devices. Directly moving data to a central server for training, e.g., at a cloud server, will introduce prohibitive communication overhead. Moreover, many types of data contain personal information and, thus, are privacy-sensitive. Federated learning [112]–[114] is a recently proposed machine learning paradigm that allows to collaboratively train a shared model for many users without direct access to the raw data. As shown in Fig. 3, each user trains a local machine learning model on the local data set and uploads it to a server for a global model aggregation. In this way, distributed data on mobile devices can be well exploited without leak of privacy. Federated learning may help to train machine learning models for some privacy-sensitive tasks of IoV, e.g., speech recognition in the driving assistant system, and infotainment applications.

One difficulty of cloud-based federated learning [112] is that the model update introduces significant communication overhead to the backbone network and long latency. The communication overhead is proportional to the machine learning model size, making it expensive to be applied to powerful deep learning models. Edge servers, acting as an intermediary between clients and the cloud server, can be exploited to reduce the communication overhead. Specifically, we can first perform multiple local aggregations at each edge server and then apply one global aggregation at the cloud. As shown in a recent study [115], such edge-assisted hierarchical federated learning reaches the desired model performance with much less communication.

2) Joint Device-Edge Inference: To apply computation-intensive DNN models, running them directly on vehicles will take too much computation resource and also consume a lot of energy, while offloading them to the edge server may suffer from time-varying wireless fading channels, which leads to excessive latency when the offloading data size is large. These powerful techniques are important for vehicle perception, as shown in Section II, and thus, how to exploit the edge resources for efficient execution is of critical importance. Recently, the joint device-edge inference has been proposed to address this challenge. As illustrated in Fig. 4, such techniques will partition the computation of DNN models and offload part of the computation to the edge server. After being processed at the device, the amount of data to be offloaded can be reduced, and thus, the offloading will be more efficient.

Neurosurgeon [116] is a framework that can automatically partition DNN computation between the client and server at the granularity of the neural network layers. It adapts to various DNN architectures, hardware platforms, wireless networks, server load levels, and intelligently partitions computation for the best latency or the best mobile energy. Both latency reduction and energy saving have been demonstrated. A similar framework, called Edgent, was proposed in [117]. It partitions DNN computation between the device and edge and further adopts an early exit mechanism at a proper intermediate DNN layer to further reduce the computation latency.

B. Vehicle as a Client

In this part, we present edge-assisted approaches for the three key tasks of intelligent vehicles, as illustrated in Fig. 5.

1) Edge-Assisted Perception: Perception tasks, such as object detection and tracking, can be assisted by edge servers, especially when powerful deep learning models are used. Two different offloading models can be considered: binary and partial offloading [31]. For binary offloading, an offloading controller will determine whether the task will be executed by the onboard unit of the vehicle or be offloaded to the edge server, depending on

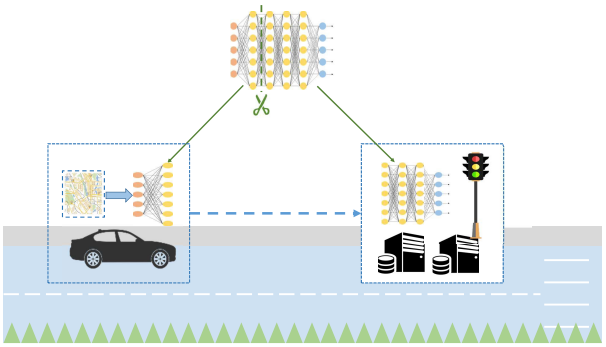

Fig. 4. *Joint device-edge inference for intelligent vehicles.*

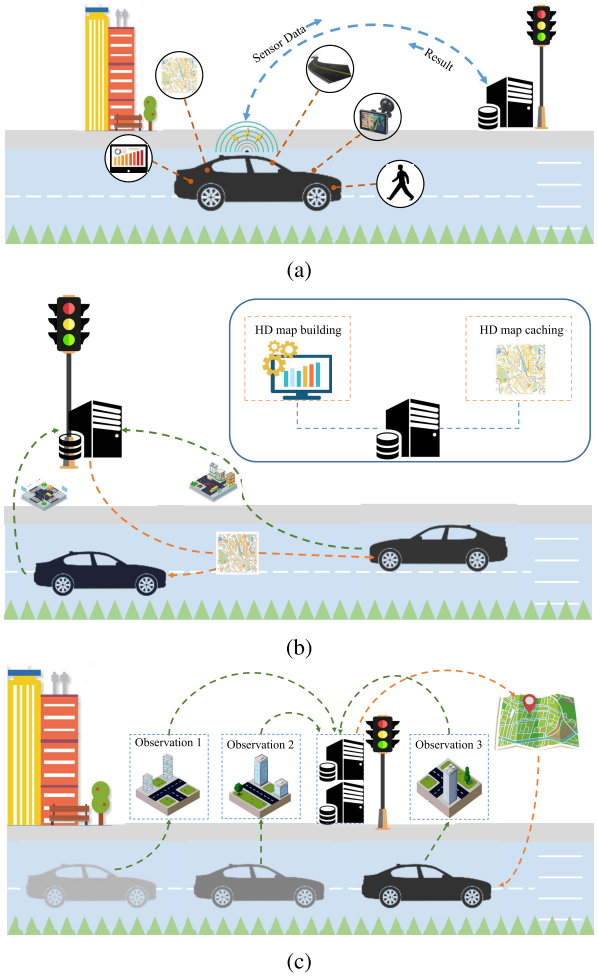

Fig. 5. *Edge-assisted approaches for three key tasks for intelligent vehicles. (a) Edge-assisted perception. (b) Edge-assisted HD mapping. (c) Edge-assisted SLAM.*

factors such as the channel state, workload at the server, and required computation intensity. For partial offloading, the computation task is partitioned and executed at both the device and the edge server, and thus, joint device-edge inference frameworks, e.g., [116] and [117], can be applied.

When designing offloading strategies, features of the specific algorithm should be carefully taken into account. For example, for localization, vision-based approaches enjoy highly parallel data-processing stages, such as feature extraction, disparity map generation, optical flow, feature match, and the Gaussian blur [118]. Thus, they are more amenable for the joint client–server processing by offloading part of the tasks to the edge server, which may provide abundant GPU resources. In comparison, LiDAR-based localization heavily uses the iterative closest point algorithm, which is hard to parallelize [119] and, thus, may be difficult to offload.

2) Edge-Assisted Map Update: To account for the dynamics of the environment, HD maps have to be refreshed timely, for which we need to collect fresh HD mapping data and detect changes from the fresh data. To build or update an HD map, the data collected in one area will be used only for that particular area. Thus, to alleviate the storage and communication burden, data collected by vehicles should be aggregated at the nearby edge servers, assisted by V2I communications. After enough fresh data are available, road change detection and road events (e.g., road closure) detection can be performed at the edge servers [59], e.g., applying the DNN-based methods [57]. Once each server builds its own map, multisession mapping can be employed to combine multiple maps in a common metrical coordinate system [120]. The updated map will be cached at the edge servers, which then help to notify the changes and distribute the map to vehicles in the coverage area.

3) Edge-Assisted SLAM: SLAM is a key technology for autonomous driving [63]. Compared with mobile robots that are normally in indoor environments, SLAM for autonomous driving is more challenging. To overcome the limitation of onboard computation, there have been studies on cloud-based SLAM, i.e., to offload part of the computation load to the cloud server. For example, cloud framework for cooperative tracking and mapping (C^2TAM) [65] is a distributed framework where the expensive map optimization and storage are performed on the cloud, while a light camera tracking client runs on a local computer. It applies the parallel tracking and mapping (PTAM) algorithm, which has two parallel threads. On one hand, a geometric map is computed by nonlinear optimization over a set of selected key frames. This background process is able to produce an accurate 3-D map at a low frame rate. On the other hand, a foreground tracking process is able to estimate the camera location at the frame rate assuming a known map. While the framework was developed assuming the cloud platform, the experiment in the article used a desktop as the "cloud." Hence, it essentially is an edge-based SLAM method. Thus, this article demonstrated the feasibility of edge-assisted SLAM.

Edge servers may also assist cooperation among multiple vehicles for SLAM, which is called as centralized SLAM [63]. In this case, an edge server acts as a controller to aggregate and fuse the data before sending the results back to vehicles, via V2I communications. Different approaches have been proposed. Vehicles can first build submaps by themselves, which are then fused at the nearby edge server [121]. For this method, the relative locations of vehicles must be known. In another proposal [122], a multirobot visual SLAM method was proposed, where data from multiple robots, i.e., observations and visual descriptors, are sent to a central agent to build a map, using a Rao–Blackwellized particle filter to estimate both the map and the trajectories of the robots.

C. Vehicle as a Server

With effective V2V communications, intelligent vehicles can share their sensing information and perception outputs. This enables various cooperative driving

techniques [123], for which edge servers at BSs or RSUs may act as coordinators.

1) Cooperative Perception: As discussed in Section II, the vehicle sensing capability is fundamentally constrained by the inherent characteristic of each sensor, as well as the budget limit. With IoV, vehicles are connected with each other, which provides the opportunity to cooperate for better sensing capabilities. The cooperative perception among multiple vehicles [54] improves situation awareness and perception capability and provides traffic information beyond line of sight and field of view. It can be useful for situations such as hidden obstacle avoidance, safe lane-changing/overtaking, and smooth braking/acceleration.

Cooperative perception has received a lot of attention, first, for mobile robots for environmental surveillance [124] and, recently, for intelligent vehicles, e.g., cooperative localization [125] and cooperative mapping [126]. In [127], a cooperative driving system based on cooperative perception was tested. A multimodal cooperative perception method was first developed, which provides a far-sight see-through, lifted-seat, satellite, or all-around view to a driver. Cooperative driving by a see-through forward collision warning, overtaking/lane-changing assistance, and automated hidden obstacle avoidance was then tested through real-world experiments using four vehicles on the road. Recently, in [128], augmented vehicular reality (AVR) was proposed to broaden the vehicle's visual range by sharing instantaneous 3-D views of the surroundings with other nearby vehicles, assisted by effective V2V communications.

2) Crowd-Sourced Mapping: While HD mapping is playing a significant role in autonomous driving, it is tedious and costly to construct an HD map. While autonomous driving companies could rely on their autonomous vehicles being tested on road to collect fresh data, the coverage is still limited. It is more efficient to work with automakers to get fresh map data from intelligent vehicles equipped with various sensors in a crowd-sourced manner. For example, with a large number of vehicles equipped with necessary hardware components for self-driving capability, it took Tesla only around 4 h to collect one million miles of data [129]. In comparison, after years' driving tests, Waymo's self-driving fleet accumulated ten million miles of data by October 2018 [130], which Tesla could acquire within two days. While this is a rough comparison, the message is clear. The cache and computing resources at intelligent vehicles will be the basis for crowd-sourced map construction, while edge servers can act as local aggregators.

3) Multivehicle SLAM: Cooperation between the vehicles can be exploited to address the computational challenges in SLAM. Besides edge-assisted centralized SLAM mentioned in Section V-B, decentralized SLAM has also been proposed, where each vehicle builds its own decentralized map while communicating with the other vehicles [131]. In this way, vehicles can quickly update maps in case of sudden changes or to anticipate dynamic conditions and can be robust to error/failure of any one of the vehicles. Decentralized SLAM is more challenging than centralized multivehicle SLAM, with difficulties, including estimating relative poses of robots, uncertainty of the relative poses, updating maps and poses, and complexity and communications issues. Existing studies on decentralized SLAM are mainly for mobile robotics.

For decentralized SLAM, each vehicle has to integrate multiple local maps provided by the other vehicles to generate a global map. This is a challenging task as the required alignments or transformation matrices to relate different maps are, in general, unknown. One key principle is to identify the relative pose between spatial information from different vehicles, which can be handled by map merging, or map fusion. Different approaches have been proposed, e.g., [132] and [133]. To achieve real-time decentralized SLAM, communications among vehicles should be carefully considered. Depending on the available bandwidth, different contents can be exchanged, e.g., graph-based representations [134] or topological maps [135].

VI. CONCLUSION

This article introduced an EIS for intelligent IoV, including edge caching, edge computing, and edge AI. Platforms, design methodologies, and key use cases were presented. The unique advantages of the EIS make it a key component of the information infrastructure that is needed to support intelligent IoV. There are, therefore, abundant opportunities but also significant challenges. Efforts are needed from different players, including researchers, entrepreneurs, governments, policymakers, and standardization bodies, to help create the technologies and policies to meet the challenges ahead. ■

REFERENCES

[1] American Automotive Policy Council. (2018). *U.S. Economic Contribution Report*. [Online]. Available: http://www.americanautocouncil.org/us-economic-contributions

[2] Association of the Automotive Industry. (2019). *International Automotive Business Produces Solid Results in 2018*. [Online]. Available: https://www.vda.de/en/press/press-releases/20190116-international-automotive-business-produces-solid-results-in-2018.html

[3] World Health Organization (WHO). (2018). *Global Status Report on Road Safety 2018*. [Online]. Available: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/

[4] N. Lu, N. Cheng, N. Zhang, X. Shen, and J. W. Mark, "Connected vehicles: Solutions and challenges," *IEEE Internet Things J.*, vol. 1, no. 4, pp. 289–299, Aug. 2014.

[5] J. E. Siegel, D. C. Erb, and S. E. Sarma, "A survey of the connected vehicle landscape—Architectures, enabling technologies, applications, and development areas," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 8, pp. 2391–2406, Aug. 2018.

[6] K. Abboud, H. A. Omar, and W. Zhuang, "Interworking of DSRC and cellular network technologies for V2X communications: A survey," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9457–9470, Dec. 2016.

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

[8] L. Liang, H. Peng, G. Y. Li, and X. Shen, "Vehicular communications: A physical layer perspective," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10647–10659, Dec. 2017.

[9] E.-K. Lee, M. Gerla, G. Pau, U. Lee, and J.-H. Lim, "Internet of Vehicles: From intelligent grid to autonomous cars and vehicular fogs," *Int. J. Distrib. Sensor Netw.*, vol. 12, no. 9, pp. 1–14, Sep. 2016.

[10] W. Xu *et al.*, "Internet of vehicles in big data era," *IEEE/CAA J. Autom. Sinica*, vol. 5, no. 1, pp. 19–35, Jan. 2018.

[11] J. Levinson *et al.*, "Towards fully autonomous driving: Systems and algorithms," in *Proc. IEEE Intell. Vehicles Symp.*, Jun. 2011, pp. 163–168.

[12] T. Luettel, M. Himmelsbach, and H.-J. Wuensche, "Autonomous ground vehicles—Concepts and a path to the future," *Proc. IEEE*, vol. 100, Special Centennial Issue, pp. 1831–1839, May 2012.

[13] K. Bengler, K. Dietmayer, B. Farber, M. Maurer, C. Stiller, and H. Winner, "Three decades of driver assistance systems: Review and future perspectives," *IEEE Intell. Transp. Syst. Mag.*, vol. 6, no. 4, pp. 6–22, Oct. 2014.

[14] K. Heineke, P. Kampshoff, A. Mkrtchyan, and E. Shao, "Self-driving car technology: When will the robots hit the road?" McKinsey, New York, NY, USA, Tech. Rep., May 2017.

[15] Tesla. (2019). *Tesla Autopilot: Full Self-Driving Hardware on All Cars*. [Online]. Available: https://www.tesla.com/autopilot

[16] C. Gauthier, "The 2018 Audi A8 is the first Level 3 self driving car," *Autonation Drive*, Aug. 2017. [Online]. Available: https://autonationdrive.com/2018-audi-a8-first-level-3-self-driving-car/

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

[18] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[19] (2017). *Automotive Sensors and Electronics Expo 2017*. [Online]. Available: http://www.automotivesensors2017.com/

[20] S. Heinrich, "Flash memory in the emerging age of autonomy," in *Proc. Flash Memory Summit*, Santa Clara, CA, USA, Aug. 2017, pp. 1–10.

[21] B. Krzanich. (2016). Data is the new oil in the future of automated driving. Intel. [Online]. Available: https://newsroom.intel.com/editorials/krzanich-the-future-of-automated-driving/

[22] *Self Driving Safety Report*, Nvidia, Santa Clara, CA, USA, 2018.

[23] S.-C. Lin *et al.*, "The architectural implications of autonomous driving: Constraints and acceleration," in *Proc. 23rd ACM ASPLOS*, Williamsburg, VA, USA, Mar. 2018, pp. 751–766.

[24] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud computing: State-of-the-art and research challenges," *J. Internet Services Appl.*, vol. 1, no. 1, pp. 7–18, May 2010.

[25] B. Kehoe, S. Patil, P. Abbeel, and K. Goldberg, "A survey of research on cloud robotics and automation," *IEEE Trans. Autom. Sci. Eng.*, vol. 12, no. 2, pp. 398–409, 2015.

[26] Q. Zhang *et al.*, "OpenVDAP: An open vehicular data analytics platform for CAVs," in *Proc. IEEE Int. Conf. Distrib. Comput. Syst. (ICDCS)*, Vienna, Austria, Jul. 2018, pp. 1–11.

[27] A. Cartas *et al.*, "A reality check on inference at mobile networks edge," in *Proc. EdgeSys*, Dresden, Germany, Mar. 2019, pp. 54–59.

[28] Y. C. Hu, M. Patel, D. Sabella, N. Sprecher, and V. Young, "Mobile edge computing—A key technology towards 5G," ETSI, Sophia Antipolis, France, White Paper 11, Sep. 2015, pp. 1–16.

[29] W. Shi, J. Cao, Q. Zhang, Y. Li, and L. Xu, "Edge computing: Vision and challenges," *IEEE Internet Things J.*, vol. 3, no. 5, pp. 637–646, Oct. 2016.

[30] P. Mach and Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628–1656, 3rd Quart., 2017.

[31] Y. Mao, C. You, J. Zhang, K. Huang, and K. B. Letaief, "A survey on mobile edge computing: The communication perspective," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2322–2358, 1st Quart., 2017.

[32] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[33] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[34] M. Ji, G. Caire, and A. F. Molisch, "Fundamental limits of caching in wireless D2D networks," *IEEE Trans. Inf. Theory*, vol. 62, no. 2, pp. 849–869, Feb. 2016.

[35] Y. Mao, J. Zhang, Z. Chen, and K. B. Letaief, "Dynamic computation offloading for mobile-edge computing with energy harvesting devices," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 12, pp. 3590–3605, Dec. 2016.

[36] S. Wang *et al.*, "When edge meets learning: Adaptive control for resource-constrained distributed machine learning," in *Proc. IEEE Int. Conf. Comput. Commun. (INFOCOM)*, Honolulu, HI, USA, Apr. 2018, pp. 63–71.

[37] Z. Zhou, X. Chen, E. Li, L. Zeng, K. Luo, and J. Zhang, "Edge intelligence: Paving the last mile of artificial intelligence with edge computing," *Proc. IEEE*, vol. 107, no. 8, pp. 1738–1762, Aug. 2019.

[38] D. H. Deans. (2019). How 'AI at the edge' is creating new semiconductor demand. CloudTech. [Online]. Available: https://www.cloudcomputing-news.net/news/2019/mar/26/ai-at-the-edge-creates-new-semiconductor-demand/

[39] M. Amadeo, C. Campolo, and A. Molinaro, "Information-centric networking for connected vehicles: A survey and future perspectives," *IEEE Commun. Mag.*, vol. 54, no. 2, pp. 98–104, Feb. 2016.

[40] R. W. L. Coutinho, A. Boukerche, and A. A. F. Loureiro, "Design guidelines for information-centric connected and autonomous vehicles," *IEEE Commun. Mag.*, vol. 56, no. 10, pp. 85–91, Oct. 2018.

[41] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[42] S. Liu, J. Tang, Z. Zhang, and J.-L. Gaudiot, "Computer architectures for autonomous driving," *Computer*, vol. 50, no. 8, pp. 18–25, 2017.

[43] K. Zhang, Y. Mao, S. Leng, Y. He, and Y. Zhang, "Mobile-edge computing for vehicular networks: A promising network paradigm with predictive off-loading," *IEEE Veh. Technol. Mag.*, vol. 12, no. 2, pp. 36–44, Jun. 2017.

[44] Google. *Edge TPU: Google's Purpose-Built ASIC Designed to Run Inference at the Edge*. [Online]. Available: https://cloud.google.com/edge-tpu/

[45] S. Abdelhamid, H. Hassanein, and G. Takahara, "Vehicle as a resource (VaaR)," *IEEE Netw.*, vol. 29, no. 1, pp. 12–17, Jan. 2015.

[46] B. Yu and F. Bai, "ETP: Encounter transfer protocol for opportunistic vehicle communication," in *Proc. IEEE Int. Conf. Comput. Commun.*, Shanghai, China, Apr. 2011, pp. 2201–2209.

[47] R. W. Wolcott and R. M. Eustice, "Visual localization within LIDAR maps for automated urban driving," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Sep. 2014, pp. 176–183.

[48] S. Thrun, W. Burgard, and D. Fox, *Probabilistic Robotics*. Cambridge, MA, USA: MIT Press, 2005.

[49] M. Barnard. (2016). Tesla & Google disagree about LIDAR 'which is right?' CleanTech. [Online]. Available: https://cleantechnica.com/2016/07/29/tesla-google-disagree-lidar-right/

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

[51] G. Bresson, M.-C. Rahal, D. Gruyer, M. Revilloud, and Z. Alsayed, "A cooperative fusion architecture for robust localization: Application to autonomous driving," in *Proc. IEEE 19th Int. Conf. Intell. Transp. Syst. (ITSC)*, Rio de Janeiro, Brazil, Nov. 2016, pp. 859–866.

[52] M. H. Daraei, A. Vu, and R. Manduchi, "Velocity and shape from tightly-coupled LiDAR and camera," in *Proc. IEEE Intell. Vehicles Symp.*, Los Angeles, CA, USA, Jul. 2017, pp. 60–67.

[53] D. Gruyer, R. Belaroussi, and M. Revilloud, "Accurate lateral positioning from map data and road marking detection," *Expert Syst. Appl.*, vol. 43, pp. 1–8, Jan. 2016.

[54] S.-W. Kim, W. Liu, M. H. Ang, E. Frazzoli, and D. Rus, "The impact of cooperative perception on decision making and planning of autonomous vehicles," *IEEE Intell. Transp. Syst. Mag.*, vol. 7, no. 3, pp. 39–50, Jul. 2015.

[55] TomTom. (2017). *TomTom HD Map for Autonomous Driving Extends to Japan*. [Online]. Available: https://corporate.tomtom.com/news-releases/news-release-details/tomtom-hd-map-autonomous-driving-extends-japan?releaseid=1045730

[56] (2016). *HERE Introduces HD Live Map to Show the Path to Highly Automated Driving*. [Online]. Available: https://360.here.com/2016/01/05/here-introduces-hd-live-map-to-show-the-path-to-highly-automated-driving/

[57] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Auto. Robots*, vol. 42, no. 7, pp. 1301–1322, Oct. 2018.

[58] H. G. Seif and X. Hu, "Autonomous driving in the iCity—HD maps as a key challenge of the automotive industry," *Engineering*, vol. 2, no. 2, pp. 159–162, Jun. 2016.

[59] J. Jiao, "Machine learning assisted high-definition map creation," in *Proc. IEEE 42nd Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, Tokyo, Japan, Jul. 2018, pp. 367–373.

[60] J. J. Leonard and H. F. Durrant-Whyte, "Simultaneous map building and localization for an autonomous mobile robot," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, Nov. 1991, pp. 1442–1447.

[61] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I," *IEEE Robot. Autom. Mag.*, vol. 13, no. 2, pp. 99–110, Jun. 2006.

[62] C. Cadena *et al.*, "Past, present, and future of simultaneous localization and mapping: Toward the robust-perception age," *IEEE Trans. Robot.*, vol. 32, no. 6, pp. 1309–1332, Dec. 2016.

[63] G. Bresson, Z. Alsayed, L. Yu, and S. Glaser, "Simultaneous localization and mapping: A survey of current trends in autonomous driving," *IEEE Trans. Intell. Veh.*, vol. 2, no. 3, pp. 194–220, Sep. 2017.

[64] C. Urmson *et al.*, "Tartan racing: A multi-modal approach to the DARPA urban challenge," Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep., Apr. 2007.

[65] L. Riazuelo, J. Civera, and J. M. M. Montiel, "C^2TAM: A cloud framework for cooperative tracking and mapping," *Robot. Auto. Syst.*, vol. 62, no. 4, pp. 401–413, 2014.

[66] K. Liu, V. C. S. Lee, J. K.-Y. Ng, J. Chen, and S. H. Son, "Temporal data dissemination in vehicular cyber–physical systems," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 6, pp. 2419–2431, Dec. 2014.

[67] P. Dai, K. Liu, L. Feng, Q. Zhuge, V. C. S. Lee, and S. H. Son, "Towards real-time and temporal information services in vehicular networks via multi-objective optimization," in *Proc. IEEE LCN*, Nov. 2016, pp. 671–679.

[68] A. Ndikumana, N. H. Tran, and C. S. Hong, "Deep learning based caching for self-driving car in multi-access edge computing," 2018, *arXiv:1810.01548*. [Online]. Available: https://arxiv.org/abs/1810.01548

[69] S. Abdelhamid, H. S. Hassanein, and G. Takahara, "On-road caching assistance for ubiquitous vehicle-based information services," *IEEE Trans. Veh. Technol.*, vol. 64, no. 12, pp. 5477–5492, Sep. 2015.

[70] R. Ding, T. Wang, L. Song, Z. Han, and J. Wu, "Roadside-unit caching in vehicular ad hoc networks for efficient popular content delivery," in *Proc. IEEE WCNC*, Mar. 2015, pp. 1207–1212.

[71] L. Hou, L. Lei, K. Zheng, and X. Wang, "A *Q*-learning-based proactive caching strategy for non-safety related services in vehicular networks," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4512–4520, Jun. 2018.

[72] L. T. Tan and R. Q. Hu, "Mobility-aware edge caching and computing in vehicle networks: A deep reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 10190–10203, Nov. 2018.

[73] T. He, N. Zhao, and H. Yin, "Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 1, pp. 44–55, Jan. 2018.

[74] M. Fiore, C. Casetti, and C.-F. Chiasserini, "Caching strategies based on information density estimation in wireless ad hoc networks," *IEEE Trans. Veh. Technol.*, vol. 60, no. 5, pp. 2194–2208, Jun. 2011.

[75] F. Malandrino, C. Casetti, C. F. Chiasserini, C. Sommer, and F. Dressler, "Content downloading in vehicular networks: Bringing parked cars into the picture," in *Proc. IEEE PIMRC*, Sep. 2012, pp. 1534–1539.

[76] L. Yao, A. Chen, J. Deng, J. Wang, and G. Wu, "A cooperative caching scheme based on mobility prediction in vehicular content centric networks," *IEEE Trans. Veh. Technol.*, vol. 67, no. 6, pp. 5435–5444, Jun. 2018.

[77] B. Hu, L. Fang, X. Cheng, and L. Yang, "In-vehicle caching (IV-Cache) via dynamic distributed storage relay (D^2SR) in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 843–855, Jan. 2019.

[78] R. Wang, X. Peng, J. Zhang, and K. B. Letaief, "Mobility-aware caching for content-centric wireless networks: Modeling and methodology," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 77–83, Aug. 2016.

[79] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility-aware caching in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5001–5015, Aug. 2017.

[80] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Exploiting mobility in cache-assisted D2D networks: Performance analysis and optimization," *IEEE Trans. Wireless Commun.*, vol. 17, no. 8, pp. 5592–5605, Aug. 2018.

[81] Y. Zhang, C. Li, T. H. Luan, Y. Fu, W. Shi, and L. Zhu, "A mobility-aware vehicular caching scheme in content centric networks: Model and optimization," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3100–3112, Apr. 2019.

[82] L. Vigneri, T. Spyropoulos, and C. Barakat, "Low cost video streaming through mobile edge caching: Modelling and optimization," *IEEE Trans. Mobile Comput.*, vol. 18, no. 6, pp. 1302–1315, Jun. 2019.

[83] 5G-PPP. *5G Automotive Vision*. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Automotive-Vertical-Sectors.pdf

[84] J. A. Khan, Y. Ghamri-Doudane, and D. Botvich, "Autonomous identification and optimal selection of popular smart vehicles for urban sensing—An information-centric approach," *IEEE Trans. Veh. Technol.*, vol. 65, no. 12, pp. 9529–9541, Dec. 2016.

[85] Z. Zhao, L. Guardalben, M. Karimzadeh, J. Silva, T. Braun, and S. Sargento, "Mobility prediction-assisted over-the-top edge prefetching for hierarchical VANETs," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1786–1801, Aug. 2018.

[86] M. Chowdhury, A. Gawande, and L. Wang, "Secure information sharing among autonomous vehicles in NDN," in *Proc. IEEE/ACM IoTDI*, Apr. 2017, pp. 15–26.

[87] T. Liu, A. A. Abouzeid, and A. Julius, "Traffic flow control in vehicular multi-hop networks with data caching," *IEEE Trans. Mobile Comput.*, to be published.

[88] Y. Wang, S. Liu, X. Wu, and W. Shi, "CAVBench: A benchmark suite for connected and autonomous vehicles," in *Proc. ACM/IEEE SEC*, Bellevue, WA, Oct. 2018.

[89] W. Zhang, Y. Wen, K. Guan, D. Kilper, H. Luo, and D. O. Wu, "Energy-optimal mobile cloud computing under stochastic wireless channel," *IEEE Trans. Wireless Commun.*, vol. 12, no. 9, pp. 4569–4581, Sep. 2013.

[90] C. You, K. Huang, H. Chae, and B.-H. Kim, "Energy-efficient resource allocation for mobile-edge computation offloading," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1397–1411, Mar. 2017.

[91] Y. Mao, J. Zhang, S. H. Song, and K. B. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5994–6009, Sep. 2017.

[92] Y. Sun, S. Zhou, and J. Xu, "EMM: Energy-aware mobility management for mobile edge computing in ultra dense networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 11, pp. 2637–2646, Nov. 2017.

[93] J. Du, F. R. Yu, X. Chu, J. Feng, and G. Lu, "Computation offloading and resource allocation in vehicular networks based on dual-side cost minimization," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1079–1092, Feb. 2019.

[94] Z. W. Lamb and D. P. Agrawal, "Analysis of mobile edge computing for vehicular networks," *Sensors*, vol. 19, no. 6, p. 1303, Mar. 2019.

[95] Y. Liu, S. Wang, J. Huang, and F. Yang, "A computation offloading algorithm based on game theory for vehicular edge networks," in *Proc. IEEE ICC*, Kansas City, MO, USA, May 2018, pp. 1–6.

[96] Y. Dai, D. Xu, S. Maharjan, and Y. Zhang, "Joint load balancing and offloading in vehicular edge computing and networks," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4377–4387, Jun. 2019.

[97] F. Yu, H. Chen, and J. Xu, "DMPO: Dynamic mobility-aware partial offloading in mobile edge computing," *Future Gener. Comput. Syst.*, vol. 89, pp. 722–735, Dec. 2018.

[98] T. Ouyang, Z. Zhou, and X. Chen, "Follow me at the edge: Mobility-aware dynamic service placement for mobile edge computing," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2333–2345, Oct. 2018.

[99] Z. Wang, Z. Zhong, D. Zhao, and M. Ni, "Vehicle-based cloudlet relaying for mobile computation offloading," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 11181–11191, Nov. 2018.

[100] NVIDIA Corp. (2019). *NVIDIA DRIVE: Scalable AI Platform for Autonomous Driving*. [Online]. Available: https://www.nvidia.com/en-us/self-driving-cars/drive-platform/

[101] T. Mekki, I. Jabri, A. Rachedi, and M. B. Jemaa, "Vehicular cloud networks: Challenges, architectures, and future directions," *Veh. Commun.*, vol. 9, pp. 268–280, Jul. 2017.

[102] A. Boukerche and R. E. De Grande, "Vehicular cloud computing: Architectures, applications, and mobility," *Comput. Netw.*, vol. 135, pp. 171–189, Apr. 2018.

[103] P. Ghazizadeh, R. Mukkamala, and S. El-Tawab, "Scheduling in vehicular cloud using mixed integer linear programming," in *Proc. 1st Int. Workshop Mobile Sens. Comput. Commun.* New York, NY, USA: ACM, Aug. 2014, pp. 7–12.

[104] T. Refaat, B. Kantarci, and H. Mouftah, "Dynamic virtual machine migration in a vehicular cloud," in *Proc. IEEE ISCC*, Funchal, Portugal, Sep. 2014, pp. 1–6.

[105] K. Zheng, H. Meng, P. Chatzimisios, L. Lei, and X. Shen, "An SMDP-based resource allocation in vehicular cloud computing systems," *IEEE Trans. Ind. Electron.*, vol. 62, no. 12, pp. 7920–7928, Dec. 2015.

[106] A.-L. Jin, W. Song, P. Wang, D. Niyato, and P. Ju, "Auction mechanisms toward efficient resource sharing for cloudlets in mobile cloud computing," *IEEE Trans. Services Comput.*, vol. 9, no. 6, pp. 895–909, Nov./Dec. 2016.

[107] Z. Su, Y. Hui, and T. H. Luan, "Distributed task allocation to enable collaborative autonomous driving with network softwarization," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 10, pp. 2175–2189, Oct. 2018.

[108] Z. Zhou, P. Liu, J. Feng, Y. Zhang, S. Mumtaz, and J. Rodriguez, "Computation resource allocation and task assignment optimization in vehicular fog computing: A contract-matching approach," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3113–3125, Apr. 2019.

[109] M. Liwang, S. Dai, Z. Gao, Y. Tang, and H. Dai, "A truthful reverse-auction mechanism for computation offloading in cloud-enabled vehicular network," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4214–4227, Jun. 2019.

[110] G. Kar, S. Jain, M. Gruteser, F. Bai, and R. Govindan, "Real-time traffic estimation at vehicular edge nodes," in *Proc. ACM/IEEE SEC*, San Jose, CA, USA, Oct. 2017, pp. 1–13.

[111] L. Liu, X. Zhang, Q. Zhang, A. Weinert, Y. Wang, and W. Shi, "AutoVAPS: An IoT-enabled public safety service on vehicles," in *Proc. SCOPE*, Montreal, QC, Canada, Apr. 2019, pp. 1–7.

[112] J. Konečnỳ, H. B. McMahan, and D. Ramage, "Federated optimization: Distributed optimization beyond the datacenter," 2015, *arXiv:1511.03575*. [Online]. Available: https://arxiv.org/abs/1511.03575

[113] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2017, *arXiv:1610.05492*. [Online]. Available: https://arxiv.org/abs/1610.05492

[114] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. Int. Conf. Artif. Intell. Stat.*, Fort Lauderdale, FL, USA, Apr. 2017, pp. 1273–1282.

[115] L. Liu, J. Zhang, S. Song, and K. B. Letaief, "Edge-assisted hierarchical federated learning with non-IID data," 2019, *arXiv:1905.06641*. [Online]. Available: https://arxiv.org/abs/1905.06641

[116] Y. Kang *et al.*, "Neurosurgeon: Collaborative intelligence between the cloud and mobile edge," in *Proc. 22nd Int. Conf. Archit. Support Program. Lang. Oper. Syst.*, 2017, pp. 615–629.

[117] E. Li, Z. Zhou, and X. Chen, "Edge intelligence: On-demand deep learning model co-inference with device-edge synergy," in *Proc. Workshop Mobile Edge Commun. (MECOMM SIGCOMM)*, Budapest, Hungary, Aug. 2018, pp. 31–36.

[118] D. Scaramuzza and F. Fraundorfer, "Visual odometry. Part I: The first 30 years and fundamentals," *IEEE Robot. Autom. Mag.*, vol. 18, no. 4, pp. 80–92, Apr. 2011.

[119] P. J. Besl and N. D. McKay, "Method for registration of 3-D shapes," *Proc. SPIE*, vol. 1611, pp. 586–607, Apr. 1992.

[120] J. McDonald, M. Kaess, C. Cadena, J. Neira, and J. J. Leonard, "6-DOF multi-session visual SLAM using anchor nodes," in *Proc. ECMR*, Örebro, Sweden, Nov. 2011, pp. 69–76.

[121] T. Tao, Y. Huang, J. Yuan, F. Sun, and X. Wu, "Cooperative simultaneous localization and mapping for multi-robot: Approach & experimental validation," in *Proc. 8th World Congr. Intell. Control Autom.*, Jinan, China, Jul. 2010, pp. 2888–2893.

[122] A. Gil, Ö. Reinoso, M. Ballesta, and M. Juliá, "Multi-robot visual SLAM using a Rao-Blackwellized particle filter," *Robot. Auton. Syst.*, vol. 58, no. 1, pp. 68–80, Jan. 2010.
[123] R. Kianfar *et al.*, "Design and experimental validation of a cooperative driving system in the grand cooperative driving challenge," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 994–1007, Sep. 2012.
[124] L. Merino, F. Caballero, J. M.-D. Dios, J. Ferruz, and A. Ollero, "A cooperative perception system for multiple UAVs: Application to automatic detection of forest fires," *J. Field Robot.*, vol. 22, nos. 3–4, pp. 165–184, Apr. 2006.
[125] H. Li and F. Nashashibi, "Cooperative multi-vehicle localization using split covariance intersection filter," *IEEE Intell. Transp. Syst. Mag.*, vol. 5, no. 2, pp. 33–44, 2013.
[126] H. Li, M. Tsukada, F. Nashashibi, and M. Parent, "Multivehicle cooperative local mapping: A methodology based on occupancy grid map merging," *IEEE Trans. Intell. Transp. Syst.*, vol. 15, no. 5, pp. 2089–2100, Oct. 2014.
[127] S.-W. Kim *et al.*, "Multivehicle cooperative driving using cooperative perception: Design and experimental validation," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 2, pp. 663–680, Apr. 2015.
[128] H. Qiu, F. Ahmad, R. Govindan, M. G. F. Bai, and G. Kar, "Augmented vehicular reality: Enabling extended vision for future vehicles," in *Proc. ACM HotMobile*, Sonoma, CA, USA, Feb. 2017, pp. 67–72.
[129] G. Giacaglia, "Self-driving cars," *Medium*, Feb. 2019. [Online]. Available: https://medium.com/@giacaglia/self-driving-cars-f921d75f46c7
[130] K. Korosec, "Waymo's self-driving cars hit 10 million miles," TechCrunch, San Francisco, CA, USA, Tech. Rep., Oct. 2018.
[131] S. Saeedi, M. Trentini, M. Seto, and H. Li, "Multiple-robot simultaneous localization and mapping: A review," *J. Field Robot.*, vol. 33, no. 1, pp. 3–46, 2016.
[132] J. L. Blanco, J. González-Jiménez, and J. A. Fernández-Madrigal, "A robust, multi-hypothesis approach to matching occupancy grid maps," *Robotica*, vol. 31, no. 5, pp. 687–701, Aug. 2013.
[133] S. Saeedi, L. Paull, M. Trentini, M. Seto, and H. Li, "Group mapping: A topological approach to map merging for multiple robots," *IEEE Robot. Autom. Mag.*, vol. 21, no. 2, pp. 60–72, Jun. 2014.
[134] M. Pfingsthorn, B. Slamet, and A. Visser, "A scalable hybrid multi-robot SLAM method for highly detailed maps," in *Proc. RoboCup*, 2007, pp. 457–464.
[135] T. A. Vidal-Calleja, C. Berger, J. Solà, and S. Lacroix, "Large scale multiple robot visual mapping with heterogeneous landmarks in semi-structured terrain," *Robot. Auton. Syst.*, vol. 59, no. 9, pp. 654–674, Sep. 2011.

ABOUT THE AUTHORS

Jun Zhang (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from The University of Texas at Austin, Austin, TX, USA, in 2009.

He is currently an Assistant Professor with the Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong. His current research interests include wireless communications and networking, mobile edge computing and edge learning, and distributed learning and optimization.

Dr. Zhang was a co-recipient of the 2019 IEEE Communications Society and Information Theory Society Joint Paper Award, the 2016 Marconi Prize Paper Award in Wireless Communications, and the 2014 Best Paper Award for the *EURASIP Journal on Advances in Signal Processing*. He received the 2016 IEEE ComSoc Asia–Pacific Best Young Researcher Award. He is also an Editor of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS and the IEEE TRANSACTIONS ON COMMUNICATIONS.

Khaled B. Letaief (Fellow, IEEE) received the B.S. (Hons.), M.S., and Ph.D. degrees in electrical engineering from Purdue University, West Lafayette, IN, USA, in 1984, 1986, and 1990, respectively.

He is currently the New Bright Professor of engineering with The Hong Kong University of Science and Technology (HKUST), Hong Kong, where he is an internationally recognized leader in wireless communications with research interests in artificial intelligence, mobile edge computing, and 5G systems and beyond. He is also with the Peng Cheng Laboratory, Shenzhen, China. Since 1993, he has been with HKUST, where he held many administrative positions, including the Dean of Engineering.

Dr. Letaief is well recognized for his dedicated service to professional societies and, in particular, IEEE, where he has served in many leadership positions, including the IEEE Communications Society's President. He was a recipient of many distinguished awards, including the 2019 Distinguished Research Excellence Award by the HKUST School of Engineering, the 2019 IEEE Communications Society and Information Theory Society Joint Article Award, the 2010 Purdue University Outstanding Electrical and Computer Engineer Award, and the 2009 IEEE Marconi Prize Award in Wireless Communications. He is also the Founding Editor-in-Chief of the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS.