Spectrum-Energy-Efficient Mode Selection and Resource Allocation for Heterogeneous V2X Networks: A Federated Multi-Agent Deep Reinforcement Learning Approach

Jinsong Gui, *Member, IEEE*, Liyan Lin, Xiaoheng Deng, *Senior Member, IEEE, Member, ACM*, and Lin Cai, *Fellow, IEEE*

***Abstract*— Heterogeneous communication environments and broadcast feature of safety-critical messages bring great challenges to mode selection and resource allocation problem. In this paper, we propose a federated multi-agent deep reinforcement learning (DRL) scheme with action awareness to solve mode selection and resource allocation problem for ensuring quality of service (QoS) in heterogeneous V2X environments. The proposed scheme includes an action-observation-based DRL and a model parameter aggregation algorithm considering local model historical parameters. By observing the actions of adjacent agents and dynamically balancing the historical samples of rewards, the action-observation-based DRL can ensure fast convergence of each agent' individual model. By randomly sampling historical model parameters and adding them to the foundation model aggregation process, the model parameter aggregation algorithm improves foundation model generalization. The generalized model is only sent to each new agent, so each old agent can retain the personality of its individual model. Simulation results show that the proposed scheme outperforms the comparison algorithms in the key performance indicators.**

***Index Terms*— Heterogeneous V2X network, mode selection, resource allocation, spectrum-energy-efficiency, deep reinforcement learning.**

I. Introduction

TO MEET the growing requirements for enhancing vehicle users' road safety, driving experience, traffic efficiency, and infotainment experience, vehicle-to-everything (V2X) communication technologies are emerging and supporting the applications of intelligent transportation system (ITS). The dedicated short-range communication (DSRC) and cellular V2X (C-V2X) have been proposed to support V2X communications [1], [2], but C-V2X has received extensive attention because of its strong cellular infrastructure and clear evolution route [3], [4]. C-V2X is originally built on long term evolution (LTE) standards, so it is known as LTE-V2X. NR-V2X is an improved version of LTE-V2X, which operates in millimeter wave (mmWave) frequency bands. In the near future, the terahertz (THz) communication as the key technology of the sixth generation (6G) mobile communication system will be introduced to C-V2X systems, which is called THz-V2X. LTE-V2X can guarantee wider coverage, but it cannot meet ultra-high-capacity demands. In contrast, NR-V2X and THz-V2X can provide ultra-high-capacity services, but their coverage areas are limited.

In a C-V2X-based system, there are usually four communication modes (i.e., vehicle to vehicle (V2V), vehicle to infrastructure (V2I), vehicle-to-pedestrian (V2P), and vehicle-to-network (V2N)) and two types of messages that get the most attention (i.e., safety-critical messages and high-capacity messages). Usually, safety-critical messages tend to be forwarded to nearby vehicles in V2V mode due to real-time requirements, while high-capacity messages are transmitted in V2N mode because of frequent access to the Internet or V2X servers [5]. In reality, the reliability of V2V mode is not always guaranteed in dynamic vehicular networks, so other communication modes need to be considered. Because the time cost of using a well-trained deep reinforcement learning (DRL) model to make decisions is negligible, it can adapt well to dynamic changing environments. Therefore, some works [5], [6], [7], [8], [9], [10] focused on resource sharing problems among V2V pairs in different communication modes and adopted DRL tools to solve these problems. However, these efforts are limited to traditional C-V2X environments.

In the latest C-V2X environments, there are much more communication modes and more complex coupling relationships between communication mode selection and resource allocation in heterogeneous V2X networks. A central question is how to design an efficient spectrum sharing architecture and an optimal dynamic vehicular access solution when multiple C-V2X technologies coexist, which aims to achieve high spectrum-energy-efficiency. In addition, rigorous mathematical methodologies are difficult to be applied to heterogeneous V2X networks due to high mobility and environmental dynamics. In this case, machine learning (ML) is a viable

Manuscript received 8 February 2023; revised 30 December 2023; accepted 6 February 2024; approved by IEEE/ACM Transactions on Networking Editor O. Yagan. Date of publication 13 February 2024; date of current version 18 June 2024. This work was supported in part by the National Natural Science Foundation of China under Grant 62272484 and Grant 62172441. *(Corresponding author: Xiaoheng Deng.)*

Jinsong Gui and Xiaoheng Deng are with the School of Electronic Information, Central South University, Changsha 410075, China (e-mail: jsgui2010@csu.edu.cn; dxh@csu.edu.cn).

Liyan Lin is with the School of Computer Science and Engineering, Central South University, Changsha 410083, China (e-mail: 214711026@csu.edu.cn).

Lin Cai is with the Department of Electrical and Computer Engineering, University of Victoria, Victoria, BC V8P 5C2, Canada (e-mail: cai@ece.uvic.ca).

Digital Object Identifier 10.1109/TNET.2024.3364161

tool. However, supervised learning and unsupervised learning require a large number of offline training samples, which cannot be applied to heterogeneous V2X networks due to the lack of prior datasets. The DRL approach can be adopted without any prior training samples. In addition, due to the consideration for bandwidth overhead and privacy issues, the reference [5] integrated the DRL technique with the federated learning (FL) [11] framework to design mode selection and resource allocation scheme.

Although the combination of DRL and FL is promising, its application in heterogeneous V2X networks still faces some new challenges. First, how to make full use of multiple communication modes in heterogeneous V2X networks to achieve optimal system spectrum-energy efficiency is a new challenge. Second, complex and diverse resource requirements of different message types and time-varying resource occupation status will cause unpredictable co-channel interference. Finally, how to design an efficient model training architecture by combining DRL with FL, which can both train a foundation model with good generalization and customize each local model, is still an open issue. We address these challenges and make the following key contributions.

1) We model communication mode selection and resource allocation problem in a system with more heterogeneous cellular interface technologies and diverse quality of service (QoS) requirements, which aims to optimize system spectrum-energy efficiency. The above system mainly involves three cellular interface technologies, three basic V2X modes (i.e., V2V, V2I, and V2N), and two types of messages (i.e., safety-critical and high-capacity messages). Unlike the most relevant work [5], we allow V2V links to reuse both uplink and downlink cellular resources, and consider data transmission in broadcast rather than assuming point-to-point V2V mode.

2) We propose an action-observation-based DRL to solve the above optimization problem, which can adapt to heterogeneous dynamic V2X environments by deploying it in each vehicular user equipment (VUE) to act as an agent. By observing the actions of adjacent agents and dynamically balancing the historical samples with positive and negative reward values, the convergence of individual model is accelerated. Moreover, by considering each VUE as an agent instead of regarding each V2V pair as an agent, it can substantially reduce the number of agents in heterogeneous V2X broadcast networks.

3) We propose a new framework combining FL framework with distributed training-execution multi-agent DRL framework to obtain a generalization model and keep the specialty of each local model. Random sampling of historical model parameters is added to the aggregation process to improve model generalization, while the generalized model is only provided to the new agents rather than the old agents to avoid influencing the specialty of the individual model.

4) Simulation results demonstrated that the proposed solution can improve system spectrum-energy efficiency under the constraints of data rate, delay, and reliability. Compared with the comparison algorithms, our scheme outperforms them in terms of system spectrum-energy efficiency, single-hop message satisfaction rate, and satisfaction rate of multi-hop message containing N links. Furthermore, by designing proper dynamic equilibrium strategy for training samples, the system spectrum-energy efficiency is improved by 92.17% under the same number of training epochs, and also the satisfaction rate of two types of messages is enhanced.

The rest of this paper are organized as follows. In Section II, we review the relevant research in mode selection and resource allocation. The system model and the problem statements are described in Section III, while the improved DRL is given in Section IV. Combining DRL with FL for problem solving is described in Section V. Section VI evaluates the performance with simulation, followed by concluding remarks and further research issues in Section VII.

II. Related Work

Many works have done to overcome various challenges in communication mode selection and resource allocation problems. Some works focused on mode selection or resource allocation for traditional cellular V2X (i.e., LTE-V2X) communications based on shared resource pool [12], [13], [14], [15], [16]. The authors in [12] transformed latency and reliability requirements to outage constraints, which aims to easily solve resource sharing problem between vehicular users and cellular users (or among different vehicular users). In [13], the resource allocation problem when each V2I[1] link shares spectrum with multiple V2V links was investigated, which aims to maximize the V2I links' capacity while meeting all the V2V links' reliability by the proposed centralized resource allocation and power control algorithms.

The authors in [14] investigated the impact of queue latency in LTE-V2X communications, and proposed a centralized scheme for opportunistic access control and mode selection. The authors in [15] explored the impact of delayed channel state information (CSI) in LTE-V2X communications, which aims to find the optimal resource allocation strategy to maximize all the V2I links' throughput while guaranteeing each V2V link's reliability. In [16], based on different network load scenarios, the authors studied the joint problem of power control and resource allocation mode selection, which aims to maximize overall information of mixed mode of centralized and distributed LTE-V2X communications.

Some other works focused on mode selection or resource allocation for V2X communications by reinforcement learning (RL) tools [17], [18], [19], [20], [21], where the RL models are generally deployed in a centralized server. In [17], the authors proposed a Q-learning-based route selection algorithm for multi-hop V2I communication, which aims to realize high throughput and low latency. In [18], the authors proposed Q-learning-based access mode selection and convex-optimization-based spectrum allocation algorithms to balance transmission performance and front-haul savings in fog-computing-based vehicular networks.

Although Q-learning is a simple and effective RL method, it cannot adapt to large-scale continuous state space. Therefore, DRL-based methods are applied to mode selection or resource allocation for V2X communications [19], [20], [21].

[1] The abbreviation V2I in Section II has the same meaning as the abbreviation V2N in Section I. In order to be consistent with the expression form in the original literatures, we do not replace V2I in Section II with V2N.

The authors in [19] designed a DRL-based method to optimize transmission mode selection policy for battery-powered vehicular networks. In [20], the authors designed a DRL-based method to optimize data transmission scheduling policy for cognitive-radio-based vehicular networks, which aims to minimize transmitting costs while meeting QoS requirements. In [21], the authors proposed a DRL-based method to optimize task offloading policy in vehicular networks with multiple edge servers and multiple offloading modes.

Besides the above single-agent DRL-based solutions, there are also some multi-agent DRL-based works on mode selection and resource allocation for LTE-V2X communications [5], [6], [7], [8], [9], [10]. Their common goal is to maximize the sum-rate of V2I links while simultaneously guaranteeing the latency and reliability requirements for V2V links. The authors in [5] proposed a DRL-based mode selection and resource allocation scheme to address the severe interference between V2I and V2V communications based on the shared resources of traditional frequencies, which aims to maximize V2I users' sum capacity while satisfying V2V pairs' latency and reliability requirements.

In [6], the authors developed a DRL-based decentralized approach for resource allocation in V2V communications, where each V2V transmitter serves as an autonomous agent to make decisions based on its local observations. In [7], the authors introduced a multi-agent DRL-based framework for transmission mode selection and power adaptation in V2V communications, where multiple V2V links compete for limited spectrum resources and form different transmission modes by the way they occupy spectrum resources. In [8], the authors investigated the spectrum sharing problem of vehicular networks and solved it by fingerprint-based multi-agent deep Q-network (DQN) method, where multiple V2V links reuse the frequency spectrum preoccupied by V2I links.

The authors in [9] proposed a multi-agent double deep Q-network (DDQN) scheme consisting of centralized learning and distributed implementation processes to both maximize the sum-rate of V2I links and satisfy the reliability and delay constraints of V2V links. In [10], authors used the multi-agent deep deterministic policy gradient (DDPG) method to investigate the resource allocation problem for LTE-V2X communications, where each V2V link serves as an agent and adopts non-orthogonal multiple access (NOMA) technology to share the spectrum pre-allocated to V2I links.

Based on the above review, we know that, in the existing works in terms of mode selection and resource allocation in V2X communications, the attention of communication modes is not comprehensive. Moreover, as analyzed and summarized in the introduction, there are still some main challenges in heterogeneous V2X networks, which motivate us to carry out further research in this paper.

III. System Model and Problem Statement

A. Network Architecture

We consider a heterogeneous V2X communication vehicular network, which consists of one macro base station (MBS), multiple small base stations (SBSs) and road side units (RSUs), and many VUEs. The MBS, SBSs, RSUs, and VUEs are equipped with three types of cellular interfaces (i.e., LTE, mmWave, THz). Furthermore, all the above cellular interfaces are equipped with the multiple antennas corresponding to the number of radio frequency (RF) chains. For a VUE, each cellular interface can work in the three communication modes (i.e., V2V, V2I, V2N). The MBS covers the entire area shown in Fig. 1, while each SBS covers only a portion of it.

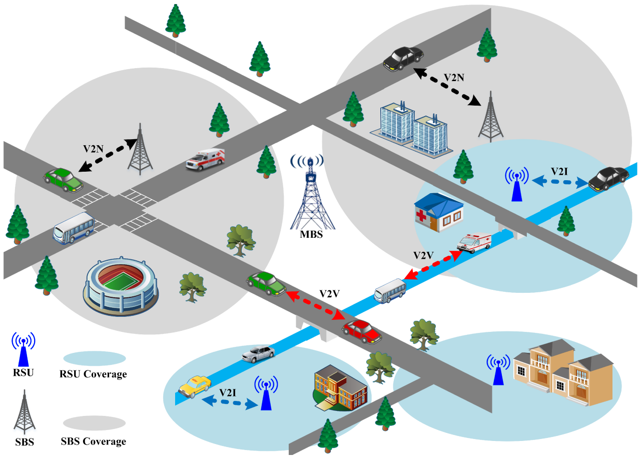


Fig. 1. Information Delivering Process in Heterogeneous V2X Systems.

According to the characteristics of V2X applications, we only consider safety-critical messages and high-capacity messages for VUEs. Safety-critical messages involve beacon messages that are broadcast at regular intervals and emergency messages that are triggered by events [1]. Beacon messages are used to transmit vehicle status information (e.g., position, speed, direction), while emergency messages are used to warn of emergencies on the road (e.g., traffic accidents). Moreover, beacon messages only need to be sent to nearby VUEs via single-hop broadcast, usually V2V mode is suitable for use. Also, emergency messages may need to be sent to other VUEs outside the communication range of this VUE via multi-hop broadcast, which can be achieved by combining V2V with V2I or V2N. High-capacity messages, including high-definition electronic map download and multimedia information transfer based on infotainment, are clearly suitable for V2N transmission.

We assume that the MBS can manage a certain amount of cellular bandwidth resources. These resources include the three frequency bands (i.e., LTE, mmWave, THz) and each frequency band can be divided into multiple cellular resource blocks (RBs). The sets of cellular RBs in terms of LTE, mmWave, and THz are denoted by $\mathcal{F}_{lte} = \{1, 2, \ldots, |\mathcal{F}_{lte}|\}$, $\mathcal{F}_{mm} = \{1, 2, \ldots, |\mathcal{F}_{mm}|\}$, and $\mathcal{F}_{thz} = \{1, 2, \ldots, |\mathcal{F}_{thz}|\}$, respectively. For the convenience of the following elaboration, let $\mathcal{F} = \mathcal{F}_{lte} \cup \mathcal{F}_{mm} \cup \mathcal{F}_{thz}$. All the SBSs reuse the above cellular bandwidth resources and allocate cellular RBs to the VUEs that are requesting V2N services.

To make full use of cellular frequency band resources, each vehicle transmitter can choose a cellular RB occupied (or unoccupied) by a cellular user, and multiple vehicle transmitters can reuse the same cellular RB with each other. Usually, V2V, V2P, and V2I in C-V2X can use the ITS frequency bands, and also they can multiplex cellular RBs. To simplify the problem without losing generality, the ITS frequency bands are not considered in this paper. In addition, in this paper, once a VUE has successfully applied for a RB, it can use the RB for both uplink and downlink communications in time division mode. The allocated RB is not automatically released until the VUE's transmission task is completed or the physical link is disconnected.

The set of SBSs is denoted by $\mathcal{S} = \{1, 2, \ldots, |\mathcal{S}|\}$, the set of RSUs is denoted by $\mathcal{R} = \{1, 2, \ldots, |\mathcal{R}|\}$, and the set of VUEs is denoted by $\mathcal{V} = \{1, 2, \ldots, |\mathcal{V}|\}$. When any VUE (e.g., $v \in \mathcal{V}$) is triggered to send a beacon message, it should preferentially use its V2V long-distance communication mode (i.e., LTE-V2X) for omnidirectional broadcasting. If the above radio interface is occupied, anyone of its V2V short-distance communication modes (i.e., THz-V2X, NR-V2X) is considered. If the quality of the above direct communication link cannot be guaranteed, V2I communication mode will be considered. These measures ensure that VUE v has always a radio interface to send a beacon message even if it is receiving another message.

When VUE v is triggered to send an emergency message, it should preferentially use its V2V long-distance communication mode for omnidirectional broadcasting. If the above radio interface is occupied, VUE v uses anyone of its V2I short-distance communication modes (i.e., THz-V2X, NR-V2X) to forward this emergency message to any SBS (e.g., $s \in \mathcal{S}$), and then SBS s should preferentially use its V2I long-distance communication mode (i.e., LTE-V2X) for omnidirectional broadcasting. If the above radio interface is occupied, SBS s will use anyone of its V2I short-distance communication modes to broadcast this emergency message. Each VUE that receives an emergency message continues to broadcast it by using one of its available radio interfaces, where its V2V long-distance communication mode is preferentially considered.

When VUE v needs to send (or receive) high-capacity messages, it will preferentially select one of its V2N communication modes (e.g., THz-V2X, NR-V2X, LTE-V2X). Because cellular RBs will be multiplexed by V2V (or V2I) links to transmit safety-critical messages, the cellular RBs allocated for V2N links should be suspended at any time according to the guaranteed delay and reliability requirements for V2V (or V2I) links.

Because of VUEs' high mobility, only large-scale channel gains are easily obtained by MBS, SBSs (or RSUs), and VUEs, which mainly include shadow fading (or slow fading) and path loss. For a cellular RB (e.g., $f \in \mathcal{F}$), the channel gains from a VUE (e.g., $v \in \mathcal{V}$) to another VUE (e.g., $v' \in \mathcal{V} \backslash v$), a SBS (e.g., $s \in \mathcal{S}$), and a RSU (e.g., $r \in \mathcal{R}$) are denoted by $h^f_{v,v'}$, $h^f_{v,s}$, and $h^f_{v,r}$, respectively.

In addition, we use a binary indicator variable (e.g., $l^f_{v,s}$) to record whether a SBS allocates a cellular RB to a V2N link beforehand. Also, we use a binary indicator variable (e.g., $l^f_{v,v'}$ (or $l^f_{v,r}$)) to record whether a cellular RB is reused by a V2V (or V2I) link. For example, if the cellular RB f is allocated to the V2N link $v \to s$ beforehand, $l^f_{v,s} = 1$, otherwise, $l^f_{v,s} = 0$. Similarly, if it is reused by the V2V link $v \to v'$ (or the V2I link $v \to r$), $l^f_{v,v'} = 1$ (or $l^f_{v,r} = 1$), otherwise, $l^f_{v,v'} = 0$ (or $l^f_{v,r} = 0$).

B. Communication Modes for VUEs

We take VUE v as an example to illustrate the communication modes it can adopt and the corresponding performance estimation formulas.

1) Cellular V2N communication mode

This mode refers to communications between a SBS and a VUE. When VUE v sends high-capacity data to SBS s via any RB f, the uplink signal-to-interference plus noise ratio (SINR) at SBS s is estimated by

$$\gamma^{(N)}_{v,s} = \frac{l^f_{v,s} p^f_{v,s} h^f_{v,s}}{\begin{pmatrix} \sum_{\hat{v} \in \mathcal{V}} \sum_{\hat{s} \in \mathcal{S}} l^f_{\hat{v},\hat{s}} p^f_{\hat{v},\hat{s}} h^f_{\hat{v},s} + \\ \sum_{\hat{v} \in \mathcal{V}} \sum_{v' \in \mathcal{V}} l^f_{\hat{v},v'} p^f_{\hat{v},v'} h^f_{\hat{v},s} + \\ \sum_{\hat{v} \in \mathcal{V}} \sum_{r \in \mathcal{R}} l^f_{\hat{v},r} p^f_{\hat{v},r} h^f_{\hat{v},s} \end{pmatrix} + \sigma^2_s} \tag{1}$$

where σ^2_s is the noise power at SBS s; $p^f_{v,s}$ is the transmission power of VUE v to SBS s at f; $p^f_{\hat{v},\hat{s}}$ is the transmission power of VUE $\hat{v}$ to SBS $\hat{s}$ at f; $p^f_{\hat{v},v'}$ is the transmission power of VUE $\hat{v}$ to VUE v' at f; $p^f_{\hat{v},r}$ is the transmission power of VUE $\hat{v}$ to RSU r at f. If the bandwidth for each cellular RB is w_f, the corresponding data rate is estimated by

$$R^{(N)}_{v,s} = \sum\nolimits_{f \in \mathcal{F}} w_f \log_2 \left(1 + \gamma^{(N)}_{v,s}\right) \tag{2}$$

When SBS s sends data to VUE v via f, the downlink SINR at VUE v is denoted by $\gamma^{(N)}_{s,v}$ and the corresponding data rate is denoted by $R^{(N)}_{s,v}$. It is easy to derive their estimating formulas by referring to the formulas (1) and (2).

2) Cellular V2I communication mode

This mode refers to communications between a RSU and a VUE. When VUE v sends data to RSU r via f, the uplink SINR at RSU r is estimated by

$$\gamma^{(I)}_{v,r} = \frac{l^f_{v,r} p^f_{v,r} h^f_{v,r}}{\begin{pmatrix} \sum_{\hat{v} \in \mathcal{V}} \sum_{s \in \mathcal{S}} l^f_{\hat{v},s} p^f_{\hat{v},s} h^f_{\hat{v},r} + \\ \sum_{\hat{v} \in \mathcal{V}} \sum_{v' \in \mathcal{V}} l^f_{\hat{v},v'} p^f_{\hat{v},v'} h^f_{\hat{v},r} + \\ \sum_{\hat{v} \in \mathcal{V}} \sum_{\hat{r} \in \mathcal{R}} l^f_{\hat{v},\hat{r}} p^f_{\hat{v},\hat{r}} h^f_{\hat{v},r} \end{pmatrix} + \sigma^2_r} \tag{3}$$

where σ^2_r is the noise power at RSU r; $p^f_{v,r}$ is the transmission power of VUE v to RSU r at f; $p^f_{\hat{v},\hat{r}}$ is the transmission power of VUE $\hat{v}$ to RSU $\hat{r}$ at f; $p^f_{\hat{v},s}$ is the transmission power of VUE $\hat{v}$ to SBS s at f. Based on $\gamma^{(I)}_{v,r}$, the corresponding data rate is estimated by

$$R^{(I)}_{v,r} = \sum\nolimits_{f \in \mathcal{F}} w_f \log_2 \left(1 + \gamma^{(I)}_{v,r}\right) \tag{4}$$

When RSU r sends data to VUE v via f, the downlink SINR at VUE v is denoted by $\gamma^{(I)}_{r,v}$ and the corresponding data rate is denoted by $R^{(I)}_{r,v}$. It is easy to derive their estimating formulas by referring to the formulas (3) and (4).

3) Cellular V2V communication mode

When VUE v broadcasts to nearby vehicles via f, among the receiving VUEs, we take VUE v' as an example to estimate the SINR at VUE v' as follows.

$$\gamma_{v,v'}^{(V)} = \frac{l_{v,v'}^{f} p_{v,v'}^{f} h_{v,v'}^{f}}{\left(\begin{array}{l} \sum\limits_{\hat{v}\in\mathcal{V}\setminus v}\sum\limits_{\hat{v}'\in\mathcal{V}} l_{\hat{v},\hat{v}'}^{f} p_{\hat{v},\hat{v}'}^{f} h_{\hat{v},v'}^{f} + \\ \sum\limits_{\hat{v}\in\mathcal{V}}\sum\limits_{s\in\mathcal{S}} l_{\hat{v},s}^{f}\left(p_{\hat{v},s}^{f} h_{\hat{v},v'}^{f} + p_{s,\hat{v},}^{f} h_{s,v'}^{f}\right) + \\ \sum\limits_{\hat{v}\in\mathcal{V}}\sum\limits_{r\in\mathcal{R}} l_{\hat{v},r}^{f}\left(p_{\hat{v},r}^{f} h_{\hat{v},v'}^{f} + p_{r,\hat{v}}^{f} h_{r,v'}^{f}\right) \end{array}\right) + \sigma_{v'}^{2}} \tag{5}$$

where $\sigma_{v'}^{2}$ is the noise power at VUE v'. From the formula (5), we show that any cellular V2V link can multiplex both V2N uplink and downlink time slot resources, where the uplink and downlink communication resources can be divided by time division multiplexing mode after a cellular RB is allocated to a V2N link. Based on $\gamma_{v,v'}^{(V)}$, the corresponding data rate is estimated by

$$R_{v,v'}^{(V)} = \sum\nolimits_{f\in\mathcal{F}} w_f \log_2\left(1 + \gamma_{v,v'}^{(V)}\right) \tag{6}$$

It is worth noting that the above formulas are derived from the premise that the MBS can coordinate the synchronization of uplink and downlink communications of all the SBSs to avoid mutual interference between them.

C. QoS Requirements

As mentioned above, this paper considers the two types of V2X applications for VUEs. VUEs undertake high-capacity applications by sending (or receiving) the corresponding messages in V2N mode. Therefore, QoS requirements of each VUE's V2N links are defined as the minimum data rate requirements to ensure the users' comfortable experience. Meanwhile, each VUE should transmit its safety-critical messages in a real-time manner, which usually ensures reliable and timely transmission of such type of messages through V2V links (or the combination of V2V and V2I links). We take VUE v as an example to detail the formal expressions of these QoS requirements as follows.

1) Data rate requirements of V2N links

The data rate requirement of VUE v's V2N links is given by

$$R_{v}^{v2n} = \max_{s\in\mathcal{S}}\left\{R_{v,s}^{(N)}\right\} \geq R_{\min}^{v2n} \tag{7}$$

where $R_{\min}^{v2n}$ is the minimum data rate requirement of V2N links and R_{v}^{v2n} is VUE v's currently reachable data rate by using V2N. For simplicity, we assume that the data rate requirements are the same for all VUEs. The expression (7) indicates that at least one V2N link meets the minimum data rate requirements.

2) Delay requirements of V2V links, V2I-assisted V2V paths, and V2N-assisted V2V paths

For cellular RBs, whether the base stations (e.g., MBS, SBSs) are responsible for resource scheduling or the vehicle nodes (e.g., VUEs) autonomously manage resource scheduling, only transmission delay (without additional grant-based resource scheduling delay for simplicity) is considered for V2V links, V2I-assisted V2V paths, and V2N-assisted V2V paths. The actual delays are estimated by

$$\begin{cases} T_{v,v'}^{(V)} = \dfrac{\mathcal{L}_v}{R_{v,v'}^{(V)}} & \text{(8a)} \\ T_{v,v'}^{(I)} = \dfrac{\mathcal{L}_v}{R_{v,r}^{(I)}} + \dfrac{\mathcal{L}_v}{R_{r,v'}^{(I)}} & \text{(8b)} \\ T_{v,v'}^{(N)} = \dfrac{\mathcal{L}_v}{R_{v,s}^{(N)}} + \dfrac{\mathcal{L}_v}{R_{s,v'}^{(N)}} & \text{(8c)} \end{cases}$$

where $\mathcal{L}_v$ is safety-critical message size in bits, while $T_{v,v'}^{(V)}$, $T_{v,v'}^{(I)}$, $T_{v,v'}^{(N)}$ are transmission delays of the message $\mathcal{L}_v$ at V2V link $v \to v'$, V2I-assisted V2V path $v \to r \to v'$, V2N-assisted V2V path $v \to s \to v'$, respectively. Based on the above, the delay requirement of VUE v is given by

$$T_{v}^{viv} = \min\left\{\begin{array}{l} \min\limits_{v'\in\mathcal{V}\setminus v}\left\{T_{v,v'}^{(V)}\right\}, \min\limits_{r\in\mathcal{R},v'\in\mathcal{V}\setminus v}\left\{T_{v,v'}^{(I)}\right\}, \\ \qquad\qquad\qquad \min\limits_{s\in\mathcal{S},v'\in\mathcal{V}\setminus v}\left\{T_{v,v'}^{(N)}\right\} \end{array}\right\} \leq T_{\max}^{viv} \tag{9}$$

where $T_{\max}^{viv}$ is the maximum tolerable delay of safety-critical beacon messages, while T_{v}^{viv} is VUE v's current delay measure. For simplicity, we assume that the delay requirements are the same for all the VUEs' safety-critical applications. The expression (9) indicates that at least one communication mode meets the maximum delay requirement.

3) Reliability requirements of V2V links, V2I-assisted V2V paths, and V2N-assisted V2V paths

We use bit error rate (BER) to measure reliability of V2V links, V2I-assisted V2V paths, and V2N-assisted V2V paths. The BER values are closely related to the SINR values of V2V links, V2I-assisted V2V paths, and V2N-assisted V2V paths. With the approximate relation expression between BER and SINR described in [22], the reliability measurements in terms of the link $v \to v'$ as well as the paths $v \to r \to v'$ and $v \to s \to v'$ are estimated by

$$\begin{cases} B_{v,v'}^{(V)} = e^{-0.5\gamma_{v,v'}^{(V)}} & \text{(10a)} \\ B_{v,v'}^{(I)} = 1 - \left(1 - e^{-0.5\gamma_{v,r}^{(I)}}\right)\left(1 - e^{-0.5\gamma_{r,v'}^{(I)}}\right) & \text{(10b)} \\ B_{v,v'}^{(N)} = 1 - \left(1 - e^{-0.5\gamma_{v,s}^{(N)}}\right)\left(1 - e^{-0.5\gamma_{s,v'}^{(N)}}\right) & \text{(10c)} \end{cases}$$

where $B_{v,v'}^{(V)}$, $B_{v,v'}^{(I)}$ and $B_{v,v'}^{(N)}$ are the BER values when the message $\mathcal{L}_v$ is sent at the link $v \to v'$ as well as the paths $v \to r \to v'$ and $v \to s \to v'$, respectively. Based on the above, the reliability requirement of VUE v is given by

$$B_{v}^{viv} = \min\left\{\begin{array}{l} \min\limits_{v'\in\mathcal{V}\setminus v}\left\{B_{v,v'}^{(V)}\right\}, \min\limits_{r\in\mathcal{R},v'\in\mathcal{V}\setminus v}\left\{B_{v,v'}^{(I)}\right\}, \\ \qquad\qquad\qquad \min\limits_{s\in\mathcal{S},v'\in\mathcal{V}\setminus v}\left\{B_{v,v'}^{(N)}\right\} \end{array}\right\} \leq B_{\max}^{viv} \tag{11}$$

where $B_{\max}^{viv}$ is the maximum tolerable BER of V2V links, V2I-assisted V2V paths, and V2N-assisted V2V paths, while B_{v}^{viv} is VUE v's current BER measure. For simplicity, we assume that the reliability requirements are the same for all the VUEs' safety-critical applications. The expression (11) indicates that at least one communication mode meets the reliability requirements.

Unlike the delay and reliability requirements of beacon messages that are only related to each single link or one-hop relay via RSUs (or SBSs), those of emergency messages are related to each multi-hop path consisting of multiple links. If a transmission path pa of VUE v contains $\mathcal{N}$ links, the corresponding path delay $T_{pa}^{viv,v}$ and path BER $B_{pa}^{viv,v}$ are estimated by

$$T_{pa}^{viv,v} = \sum\nolimits_{n=1}^{\mathcal{N}} T_{pa,n}^{viv,v} \tag{12}$$

$$B_{pa}^{viv,v} = 1 - \prod\nolimits_{n=1}^{\mathcal{N}} \left(1 - B_{pa,n}^{viv,v}\right) \tag{13}$$

where $T_{pa,n}^{viv,v}$ is delay value of n-th link and $B_{pa,n}^{viv,v}$ is BER value of n-th link. Only when pa is the longest message transmission path as well as $T_{pa}^{viv,v}$ and $B_{pa}^{viv,v}$ are not more than T_{max}^{viv} and B_{max}^{viv} respectively, the transmission delay and reliability requirements of emergency messages can be guaranteed. For convenience, in the following text, $T_{pa}^{viv,v}$ and $B_{pa}^{viv,v}$ are regarded as the delay and reliability measurements on the longest transmission path of VUE v.

D. Problem Statement

In this paper, the global goal is to find the optimal mode selection and resource allocation scheme that maximizes the system spectrum-energy efficiency under the constraints of data rate requirements of each V2N link, the delay and reliability requirements of each V2V link (or V2I/V2N-assisted V2V path), and the delay and reliability requirements of the longest transmission path of each VUE. To obtain the expression of system spectral energy efficiency, we first give the estimation formulas of the sum data rate R, the sum power consumption P, and the sum frequency band resource consumption W as follows.

$$\begin{cases} R = \sum\nolimits_{v\in\mathcal{V}} R_v & (14a) \\ P = \sum\nolimits_{v\in\mathcal{V}} P_v & (14b) \\ W = \sum\nolimits_{v\in\mathcal{V}} W_v & (14c) \end{cases}$$

where R_v, P_v, and W_v denote the data rate, power consumption, and frequency band resource consumption associated with VUE v respectively, which are estimated by

$$R_v = \left(\sum_{s\in\mathcal{S}} \left(R_{v,s}^{(N)}\right) + \sum_{r\in\mathcal{R}} \left(R_{v,r}^{(I)}\right) + \sum_{v'\in\mathcal{V}\backslash v} \left(R_{v,v'}^{(V)}\right) \right) \tag{15}$$

$$P_v = \left(\begin{array}{c} \sum\limits_{s\in\mathcal{S}} \sum\limits_{f\in\mathcal{F}} l_{v,s}^f p_{v,s}^f + \sum\limits_{r\in\mathcal{R}} \sum\limits_{f\in\mathcal{F}} l_{v,r}^f p_{v,r}^f + \\ \sum\limits_{v'\in\mathcal{V}\backslash v} \sum\limits_{f\in\mathcal{F}} l_{v,v'}^f p_{v,v'}^f \end{array} \right) \tag{16}$$

$$W_v = \left(\begin{array}{c} \sum\limits_{s\in\mathcal{S}} \sum\limits_{f\in\mathcal{F}} l_{v,s}^f w_f + \sum\limits_{r\in\mathcal{R}} \sum\limits_{f\in\mathcal{F}} f(v,f) l_{v,r}^f w_f + \\ \sum\limits_{v'\in\mathcal{V}\backslash v} \sum\limits_{f\in\mathcal{F}} g(v,f) l_{v,v'}^f w_f \end{array} \right) \tag{17}$$

In (17), $f(*,*)$ and $g(*,*)$ are the binary indicator functions. For $f(v,f)$, it means that its value is 1 if no SBS allocates f to VUE v. Otherwise, $f(v,f) = 0$.

$$f(v,f) = \begin{cases} 1 & , \quad \sum\nolimits_{s\in\mathcal{S}} l_{v,s}^f = 0 \\ 0 & , \quad Otherwise \end{cases} \tag{18}$$

Similarly, for $g(v,f)$, it means that its value is 1 if no SBS allocates f to VUE v and also no RSU uses f to communicate with VUE v. Otherwise, $g(v,f) = 0$.

$$g(v,f) = \begin{cases} 1 & , \quad \sum\nolimits_{s\in\mathcal{S},r\in\mathcal{R}} \left(l_{v,s}^f + l_{v,r}^f\right) = 0 \\ 0 & , \quad Otherwise \end{cases} \tag{19}$$

Based on the formula (14), the system spectrum-energy efficiency is expressed by

$$S_{EE} = \frac{R}{P \cdot W} \tag{20}$$

The optimization problem in terms of the system spectrum-energy efficiency is formulated by

$$\begin{array}{l} \max\limits_{v,v'\in\mathcal{V},r\in\mathcal{R},s\in\mathcal{S}} S_{EE} \\ \left\{ \begin{array}{ll} st. \quad C1: & R_v^{v2n} \geq R_{\min}^{v2n} \\ \quad\quad C2: & T_v^{viv} \leq T_{\max}^{viv}, T_{pa}^{viv,v} \leq T_{\max}^{viv} \\ \quad\quad C3: & B_v^{viv} \leq B_{\max}^{viv}, B_{pa}^{viv,v} \leq B_{\max}^{viv} \\ \quad\quad C4: & \sum\limits_{f\in\mathcal{F}} l_{v,s}^f \leq 1, l_{v,s}^f \in \{0,1\} \\ \quad\quad C5: & \sum\limits_{f\in\mathcal{F}} l_{v,r}^f \leq 1,\ l_{v,r}^f \in \{0,1\} \\ \quad\quad C6: & \sum\limits_{f\in\mathcal{F}} l_{v,v'}^f \leq 1,\ l_{v,v'}^f \in \{0,1\} \\ \quad\quad C7: & 0 \leq p_{v,s}^f, p_{v,r}^f,\ p_{v,v'}^f \leq\ p_{\max}^{lte}, \quad f \in \mathcal{F}_{lte} \\ \quad\quad C8: & 0 \leq\ p_{v,s}^f, p_{v,r}^f,\ p_{v,v'}^f \leq\ p_{\max}^{mm}, f \in \mathcal{F}_{mm} \\ \quad\quad C9: & 0 \leq p_{v,s}^f, p_{v,r}^f,\ p_{v,v'}^f \leq\ p_{\max}^{thz}, f \in \mathcal{F}_{thz} \end{array} \right. \end{array} \tag{21}$$

where the constraints $C1 \sim C3$ are the data rate, delay and reliability requirements of each VUE, respectively. The constraint $C4$ denotes that a SBS at most allocates one RB to a VUE. The constraint $C5$ show that a RSU at most adopts one RB to communicate with a VUE. The constraint $C6$ denotes that a VUE's V2V link can only use one RB. The constraints $C7 \sim C9$ show that transmission powers of each type of radio interfaces of VUE v cannot be higher than the corresponding maximum transmission power, respectively.

The formulated problem (21) is a mixed-integer nonlinear programming problem, which is difficult to solve due to the following reasons. The constraints $C1 \sim C3$ and $C7 \sim C9$ make the problem nonconvex, while the constraints $C4 \sim C6$ result in a combinatorial problem. Therefore, rigorous mathematical modeling tools are difficult to cope with this problem. As mentioned earlier, ML is a viable tool, but the ML models that require prior data are not suitable for network environments that lack prior data. The DRL-based approach does not require any prior dataset, but it can converge to a solution through iterative feedback from dynamic vehicular environments, which can be applied to this problem. In addition, due to the difficulty of acquiring global CSI and the huge computational complexity in heterogeneous V2X networks, the problem is more suitable to be formulated as a multi-agent DRL model. Here, multiple VUEs concurrently make mode selection and resource allocation decisions based on their own observations for the purpose of optimizing the overall long-term reward (i.e., system spectrum-energy-efficiency) and individual rewards (i.e., data rate, delay and reliability).

In the research works on multi-agent DRL for mixed cooperation and competition tasks, two typical frameworks have attracted much attention. One is the centralized training with decentralized execution (CTDE), and the other is the decentralized one with networked agents (DONA) [23]. The former results in a very large number of model parameters, which makes it difficult to train models for large-scale networks, especially when it comes to competitive tasks with discrete action space. The latter can not only reduce the size of model parameters through distributed training mode, but also enhance the stability of the model through the communication between agents to expand the observation range of agents. Therefore, based on the DONA framework, we propose the solution to the problem (21), which can be applied to large-scale networks.

IV. DONA-Based Scheme

The formulated mode selection and resource allocation problem can be regarded as a DONA-based Markov decision process (DONA-MDP). According to [24] and [25], the DONA-MDP can be characterized by a tuple $(\mathbb{S}, \mathbb{A}, \mathbb{P}, \mathbb{E}, \{\mathcal{G}_t\}_{t\geq 0})$. Here, $\mathbb{S}$ is the global state space, $\mathbb{A}$ is the joint action space of all agents, $\mathbb{P}$ is the state transition probability set, and $\mathbb{E}$ is the local reward function set. $\mathbb{S}$ is detailed as $\{\mathbb{S}^v\}_{v\in\mathcal{V}}$, where $\mathbb{S}^v$ is the local observation space of VUE agent v. $\mathbb{A}$ is detailed as $\prod_{v=1}^{|\mathcal{V}|}\mathbb{A}^v$, where $\mathbb{A}^v$ is the action space of VUE agent v. $\mathbb{P}$ is detailed as $\mathbb{P} : \mathbb{S}\times\mathbb{A}\times\mathbb{S} \rightarrow [0, 1]$. $\mathbb{E}$ is detailed as $\{\mathbb{E}^v\}_{v\in\mathcal{V}}$, where $\mathbb{E}^v : \mathbb{S}^v \times \mathbb{A}^v \rightarrow \mathbb{R}$ is the local reward function of VUE agent v. $\{\mathcal{G}_t\}_{t\geq 0}$ denotes a time-varying network.

A. DONA-MDP Model Design

On the basis of independent DDQN, an action-observation-based DDQN (AO-DDQN) is proposed to adapt to heterogeneous environments, which is deployed in each vehicle to act as an agent. We divide time $\mathbb{T}$ into series of equal time steps. In time step t, each agent independently selects an action and annunciates it among its neighbors. Because there may be some time differences when the agents choose actions, the agents that choose actions later can observe the declared results of the agents that have chosen the actions ahead of them within the receiving range. The observed action selection results are added to the respective observation set. The corresponding online neural network parameters θ_v are updated by minimizing the following loss function at the batch size.

$$L(\theta_v) = \left(y_t^v - Q(\mathbb{s}_t^v, \mathbb{a}_t^v|\theta_v)\right)^2 \tag{22}$$

where $y_t^v = \mathbb{e}_t^v + \gamma\max_{\mathbb{a}_{t+1}^v}\hat{Q}(\mathbb{s}_{t+1}^v, \mathbb{a}_{t+1}^v|\theta_v^-)$. Here, θ_v^- is the target neural network parameters, which is periodically copied from θ_v and stays the same over multiple iterations. $\mathbb{s}_t^v$ is the local observation of VUE agent v at time step t. $\mathbb{a}_t^v$ is its action at time step t. $\mathbb{e}_t^v$ is the immediate reward of VUE agent v at time step t. γ is the discount factor. The three key elements of the DONA-MDP model specific to the problem considered in this paper are detailed as follows.

1) Action space

For each VUE agent $v \in \mathcal{V}$ at time step t, its action $\mathbb{a}_t^v$ includes the following three parts.

$$\begin{cases} \vec{f} \in \mathcal{F}_{lte} \times \mathcal{F}_{mm} \times \mathcal{F}_{thz} & \text{(23a)} \\ \vec{m} \in m_{V2N} \times m_{V2I} \times m_{V2V} & \text{(23b)} \\ \vec{p} \in \left(\begin{array}{l} \cup_{i=1}^{N_p^l} \dfrac{i}{N_p^l} p_{max}^{lte} \times \\ \cup_{i=1}^{N_p^m} \dfrac{i}{N_p^m} p_{max}^{mm} \times \\ \cup_{i=1}^{N_p^t} \dfrac{i}{N_p^t} p_{max}^{thz} \end{array} \right) & \text{(23c)} \end{cases}$$

where (23a) indicates frequency band selection results, (23b) indicates the communication modes (i.e., V2N, V2I, V2V) selected by an agent, and (23c) indicates power selection results. m_{V2N}, m_{V2I}, and m_{V2V} are the indicator variables with on-negative integers. Here, $m_{V2N} \in \mathcal{S}$ if V2N mode is selected, otherwise, $m_{V2N} = 0; m_{V2I} \in \mathcal{R}$ if V2I mode is selected, otherwise, $m_{V2I} = 0$; $m_{V2V} \in \{0, 1\}$, $m_{V2V} = 1$ if V2V mode is selected, otherwise, $m_{V2V} = 0$. We assume that all the VUEs have the same number of transmission power levels at the same radio interface technology. However, different radio interface technologies may have different power levels, so they are represented by different symbols (i.e., N_p^l, N_p^m, N_p^t). From the action space, we can infer that the solution space size of problem (21) is $\left((|\mathcal{S}| + |\mathcal{R}| + 1)|\mathcal{F}|(N_p^l + N_p^m + N_p^t)\right)^{|\mathcal{V}|}$, which is also one of the important reasons why traditional optimization methods cannot be used to solve it. Therefore, the set of actions managed by VUE agent v and the joint actions of the $|\mathcal{V}|$ VUE agents at time step t can be respectively expressed by

$$\mathbb{a}_t^v = \left\{\vec{f}, \vec{m}, \vec{p}\right\}, \ \mathbb{a}_t^v \in \mathbb{A}^v \tag{24}$$

$$\mathbb{a}_t = \{\mathbb{a}_t^v| \quad \forall v \in \mathcal{V}\}, \ \mathbb{a}_t \in \mathbb{A} \tag{25}$$

2) Local observation space

The local observation of VUE agent v at time step t (i.e., $\mathbb{s}_t^v \in \mathbb{S}^v$) includes the five parts. The first part is the large-scale channel gains on each RB of each link between VUE v and its potential communication ends at the current time step, which is described in (26a). The second part indicates whether there are the remaining messages that need to be received by VUE v at the current time step, which is described in (26b). Take $L_t^{v,v',c}$ as an example, if there is a message that is about to send to VUE v via RB c, $L_t^{v,v',c} = 1$, otherwise $L_t^{v,v',c} = 0$. The third part is the observable actions of other agents within the receiving range of VUE v, which is described in (26c). Here, the receiving range of VUE v is denoted by $Rge\left(v\right) = \mathcal{V}_v$.

$$\begin{cases} h_t^v = \left\{ \begin{aligned} & \cup_{s\in\mathcal{S},c\in\mathcal{F}} \left\{ h_t^{v,s,c} \right\}, \cup_{r\in\mathcal{R},c\in\mathcal{F}} \{ h_t^{v,r,c} \}, \\ & \cup_{v'\in\mathcal{V}\backslash v,c\in\mathcal{F}} \left\{ h_t^{v,v',c} \right\} \end{aligned} \right\} & \text{(26a)} \\ L_t^v = \left\{ \begin{aligned} & \cup_{s\in\mathcal{S},c\in\mathcal{F}} \left\{ L_t^{v,s,c} \right\}, \cup_{r\in\mathcal{R},c\in\mathcal{F}} \{ L_t^{v,r,c} \}, \\ & \cup_{v'\in\mathcal{V}\backslash v,c\in\mathcal{F}} \left\{ L_t^{v,v',c} \right\} \end{aligned} \right\} & \text{(26b)} \\ O_t^v = \cup_{v'\in\mathcal{V}_v\backslash v} \{ \mathbb{a}_t^{v'} \} & \text{(26c)} \end{cases}$$

The fourth part is the remaining time to meet the delay threshold at the current time step, which is denoted by T_t^v. The fifth part is a triple to indicate the types of messages that VUE v needs to send at the current time step, which is denoted by $Y_t^v = (M_B, M_E, M_H)$. Here, M_B, M_E, and M_H are the binary indicator variables, and they respectively indicate whether VUE v transmits beacon, emergency, and high-capacity messages at current time step. Take M_B as an example, if VUE v transmits beacon messages at current time step, $M_B = 1$, otherwise $M_B = 0$. Therefore, if $Y_t^v = (1, 1, 0)$, VUE v will transmit beacon messages and emergency message messages at current time step, but it will not transmit any high-capacity message at current time step. Based on the above, the state space observed by VUE agent v can be defined by

$$\mathbb{s}_t^v = \{ h_t^v, L_t^v, O_t^v, T_t^v, Y_t^v \} \tag{27}$$

3) Immediate reward

When all the VUE agents take the joint action $\mathbb{a}_t$ on the heterogeneous V2X environment, they will receive an immediate reward. Recall that the design goal in this paper is to maximize the system spectrum-energy efficiency while meeting each VUE agent's requirements in terms of data rate, delay and reliability. To this end, a sum common reward function is proposed to measure the total performance of $|\mathcal{V}|$ VUE agents, which aims to maximize the system spectrum-energy efficiency. On the other hand, an individual reward that measures the behavior of each VUE agent is proposed to guarantee the basic performance requirements of individuals. Therefore, we propose the following immediate reward function for VUE v at time step t.

$$\mathbb{e}_t^v = \delta_1 \mathcal{H}_1 \left(\sum_{v\in\mathcal{V}} \frac{R_v}{P_v W_v} \right) + \mathbb{ie}_t^v \tag{28}$$

In (28), the first part corresponds to the spectrum-energy efficiency of the $|\mathcal{V}|$ VUEs, while the second part represents the individual reward, which is defined by

$$\mathbb{ie}_t^v = \begin{pmatrix} \delta_2 \mathcal{H}_2 \left(R_v^{v2n} - R_{min}^{v2n} \right) + \\ \delta_3 \mathcal{H}_3 \left(T_{max}^{viv} - T_v^{viv}, B_{max}^{viv} - B_v^{viv} \right) \end{pmatrix} \tag{29}$$

In (29), the individual reward consists of two parts. Here, the first part means both the rewards and punishments for minimum data rate requirements of V2N link, while the second part means the impacts of the delay and reliability requirements. The weights of the above parts are denoted by δ_1, δ_2 and δ_3 in turn, which aim to balance the revenue and penalty. $\mathcal{H}_1(*)$, $\mathcal{H}_2(*)$, and $\mathcal{H}_3(*, *)$ are all piecewise functions, which are defined by

$$\mathcal{H}_1(x) = \begin{cases} x, & if\ \mathbb{ie}_t^v > 0 \\ 0, & otherwise \end{cases} \tag{30}$$

$$\mathcal{H}_2(x) = \begin{cases} A, & x > 0 \\ x, & x \le 0 \end{cases} \tag{31}$$

$$\mathcal{H}_3(x, y) = \begin{cases} B, & x > 0, y > 0 \\ x, & x \le 0, y > 0 \\ y, & x > 0, y \le 0 \\ x + y, & x \le 0, y \le 0 \end{cases} \tag{32}$$

In (30)-(32), $\mathcal{H}_1(x)$ means that the spectrum-energy efficiency reward will be obtained only when the individual reward is met at the same time, otherwise, no reward will be obtained; $\mathcal{H}_2(x)$ means that the punishment increases with the degree of violation of the constraint, but only one constant reward is given when the constraint is satisfied; $\mathcal{H}_3(x, y)$ means that the reward will be given only when the delay and reliability requirements are met at the same time, otherwise, punishment will be given according to the violation of delay and reliability. In addition, A and B are two non-negative numbers, which represent the reward values when the constraints are met.

The multiple constraint equations aim to strengthen the constraints and give the specific rewards and punishments for each agent's actions. The differentiated real-time rewards with multiple constraints can guide the model training of agents more accurately.

B. Dynamic Equilibrium Strategy for Training Samples

Most famous models in the multi-agent DRL field [26], [27] are usually used in game scenarios. In a game scenario, an agent needs several actions (i.e., one turn) to determine whether it wins or loses. However, for communication scenarios, improper resource allocation will lead to the failure of message transmission, so the merits of the action cannot be judged by an episode due to the one-time operational characteristics of action. Furthermore, due to the sparsity of non-negative rewards and the lack of coordination between agents in the initial exploration, it will be difficult for agents to learn appropriate strategies. To address the above problems, we design a dynamic equilibrium strategy for communication scenarios, which will dynamically adjust the proportion of good and bad samples in each batch to accelerate the convergence of the model.

The DDQN model uses the experience replay pool (ERP) mechanism to reduce the data correlation and improve the model training efficiency. However, in our environment, the proportion of non-negative rewards in the replay buffer is very low in the early stages of the model training process, which also leads to the imbalance of model training. Therefore, we design the DDQN model with double replay buffers. Since spectrum-energy efficiency is constrained by individual rewards, we store non-negative individual rewards as good samples in the balanced buffer (BB), and the rest of the samples in the common buffer (CB). During training, the sampling ratio of double buffers will be dynamically adjusted according to the actual observation results.

C. Algorithm Description and Performance Analysis of AO-DDQN

On the basis of DONA-MDP model and dynamic equilibrium strategy, the specific process of AO-DDQN is shown in

Algorithm 1 AO-DDQN

```
Run at any VUE $v$
Input: $\varepsilon_1, \varepsilon_2, \gamma, \mathbb{T}, N_r, \theta$
Output: $\theta_v$
$t = 0$
$\theta_v = \theta; \theta_v^- = \theta_v$
$\Delta\varepsilon = \frac{N_r(\varepsilon_1 - \varepsilon_2)}{\mathbb{T}};\ \varepsilon = \varepsilon_1$
Initialize replay memory BB, CB to capacity $M$
While $t < \mathbb{T}$ do
    Update $h_t^v$ and $L_t^v$ based on local observation
    Get $\mathcal{V}_v$ based on local observation
    Compute $\cup_{v' \in \mathcal{V}_v \setminus v}\{\mathbb{a}_t^{v'}\}$ and update $O_t^v$
    Update $T_t^v$ and $Y_t^v$ based on local observation
    Get state space $\mathbb{s}_t^v = \{h_t^v, L_t^v, O_t^v, T_t^v, Y_t^v\}$
    Select a random action with probability $\varepsilon$
    Otherwise get action $\mathbb{a}_t^v = argmax_{\mathbb{a}_t^v} Q(\mathbb{s}_t^v, \mathbb{a}_t^v|\theta_v)$
    Broadcast the action selection result to neighborhood
    Observe the new state space $\mathbb{s}_{t+1}^v$
    Receive the immediate reward $\mathbb{e}_t^v$
    If $\mathbb{e}_t^v > 0$ then store $(\mathbb{s}_t^v, \mathbb{a}_t^v, \mathbb{e}_t^v, \mathbb{s}_{t+1}^v)$ to BB ERP
    Else store it to CB ERP End if
    If the spaces for BB ERP and CB ERP are full then
        Randomly sample $bs$ experiences as $E$ according
        to the double buffer sampling ratio
        For each experience in $E$ do
            $y_t^v = \mathbb{e}_t^v + \gamma max_{\mathbb{a}_{t+1}^v} \hat{Q}(\mathbb{s}_{t+1}^v, \mathbb{a}_{t+1}^v|\theta_v^-)$
            Compute $L(\theta_v)$ according to the formula (22)
        End for
        Update $\theta_v$ by executing mini-batch gradient
        descent
        If $(t \ mod \ N_r) == 0)$ then $\theta_v^- \leftarrow \theta_v$ End if
    End if
    $t++$
    $\varepsilon = \varepsilon - \Delta\varepsilon$
End while
return $\theta_v$
```

Algorithm 1. An approximate optimal solution to the optimization problem (21) can be obtained by having each VUE run the AO-DDQN algorithm independently. By analyzing the performance of the AO-DDQN algorithm, we can get the following three propositions.

Proposition 1: The total reward of the system will not diverge infinitely if there is no dynamic replacement of VUE agents in the system.

Proof. From the formulas (28)-(32), we know that a VUE agent (e.g., v) can obtain its immediate reward (e.g., $\mathbb{e}_t^v$) only when its individual performance requirements are met (i.e., $\mathbb{i}\mathbb{e}_t^v > 0$). In this case, the prerequisite for each VUE to obtain a positive reward is to strive for resources to ensure that its performance indicators meet the constraints. Here, frequency bands are the critical resources that affect rewards. When the supply of frequency bands is reduced and the demand of frequency bands is increasing, some or all frequency bands need to be shared or reused, resulting in co-channel interference. In this case, according to Shannon formula, take formula (2) for example, there is the possibility of positive rewards through finding a combination of transmission powers (see $p_{v,s}^f, \bigcup_{\hat{v} \in \mathcal{V}, \hat{s} \in \mathcal{S}} p_{\hat{v},\hat{s}}^f, \bigcup_{\hat{v}, v' \in \mathcal{V}} p_{\hat{v},v'}^f, \bigcup_{\hat{v} \in \mathcal{V}, r \in \mathcal{R}} p_{\hat{v},r}^f$ in formula (1)) among the concurrent transmitters. The greater the number of VUEs that share a frequency band resource, the more difficult it is to find a combination of transmission powers that makes each VUE's reward positive. However, the use of additional frequency bands can effectively alleviate the above difficulty. In the case of excessive supply of frequency band resources relative to the number of VUEs, although it is easier to meet the requirement that each VUE's reward is positive, the reward may be difficult to increase effectively according to formula (28). Therefore, the supply should be reduced appropriately. If dynamic replacement of VUE agents happens all the time, the number of VUEs on each shared frequency band may be changing, which causes the total reward to change all the time. Otherwise, the system always gets a convergent reward through the coordination of power resources under a certain frequency band resource supply. This completes the proof.

Proposition 2: There may be at least an effective solution to the optimization problem (21) as long as sufficient frequency band resources are available.

Proof. When very sufficient frequency band resources are available, the AO-DDQN algorithm can pick out some RBs from the set $\mathcal{F}$ to eliminate co-channel interferences of each type of links, including those of V2N links (see formula (1)), V2I links (see formula (3)), V2V links (see formula (5)). Then, it can select a reasonable combination of communication modes to lay the foundation for problem (21). In the absence of co-channel interference, the system spectrum-energy efficiency S_{EE} is only treated as the function of transmission powers according to formula (20). Therefore, formula (20) can be roughly reduced to $S_{EE} = \frac{\mathcal{D}_1 log_2(1+\mathcal{D}_2 P)}{P}$, where $\mathcal{D}_1$ and $\mathcal{D}_2$ are treated as two positive constants. The first-order derivative of this simplified formula is expressed by $\frac{\partial S_{EE}}{\partial P} = \frac{\mathcal{D}_1 \mathcal{D}_2}{P(1+\mathcal{D}_2 P)\ln 2} - \frac{\mathcal{D}_1 log_2(1+\mathcal{D}_2 P)}{P^2}$. When $P > 0$, $\frac{\partial S_{EE}}{\partial P} < 0$, from which it can be inferred that S_{EE} is a monotonically decreasing function. Therefore, the smaller the transmission powers, the higher the system spectrum-energy efficiency. However, to meet the constraints $C1 \sim C3$ of problem (21), the transmission powers cannot be infinitely reduced. Therefore, considering the above factors, there is theoretically a suitable set of transmission powers to optimize S_{EE} while meeting the constraints of problem (21). Moreover, since transmission powers have been discretized in this paper, an exhaustive search method can be adopted to find this specific set of transmission powers, which constitutes an effective solution to problem (21) together with the selected frequency band resources and communication modes. This completes the proof.

Proposition 3: The convergence speed of the AO-DDQN algorithm increases with the range of notification information received by a VUE agent from other VUE agents.

Proof: When a VUE agent has the larger receiving range, it can observe the more other agents' action choices. In this way, the AO-DDQN algorithm can reduce the number of selecting infeasible actions and thus get the desired action results that approximates the effective solution mentioned in Proposition 2 faster. Take VUE agent v at for example, its receiving range $\mathcal{V}_v$ determines the amount of information about the historical actions of other VUE agents included in its state space. We know from equation (26c) that, when all the VUE agents are included in $\mathcal{V}_v$, VUE agent v can avoid choosing actions that may result in utility decline based on the currently observable actions. It will increase the number of samples in the balanced buffer pool and raise the probability of positive samples with high spectrum-energy efficiency. This will also have a positive impact on model training and form a positive cycle. On the contrary, with the limited observations from other agents, each individual agent operates in a fully distributed manner, making it more difficult to choose a set of actions that are less conflict with each other. This, in turn, generates a small cumulative number of positive samples in the balanced buffer pool (or those with poor spectrum-energy efficiency), resulting in an under-expressed model for positive samples. Therefore, more iterations are needed to approximate the effective solution in Proposition 2. This completes the proof.

V. FL-Dona-Based Algorithm

Although the AO-DDQN algorithm can achieve model convergence based on local observation, the highly dynamic vehicle communication environment means that a new vehicle entering a particular area needs a suitable foundation model. By starting with a foundation model, a VUE agent using the AO-DDQN algorithm can get an individual model quickly. Although it is useful to apply transfer learning to save the learning time of individual models [28], each new VUE agent also needs to choose a suitable foundation model. FL framework allows multiple devices to be loosely federated under the coordination of a central server to participate in global foundation model training [11]. Its original intention is to protect the privacy of training datasets, but it also reduces model training burden on a centralized server. Although the distributed trained local model parameters need to be transferred to a centralized server to update the global foundation model parameters, the communication overhead is negligible when compared to aggregating the decentralized raw datasets to the centralized server [29]. Therefore, FL framework is a suitable framework for designing the foundation model training scheme.

In this paper, the MBS acts as the central parameter aggregation server of FL framework, while all the VUEs act as the client devices to perform local model training. More precisely, the VUEs that are only new to the MBS coverage need to request a foundation model from the MBS, while the VUEs that have started local model training only use the personalized models that they have trained themselves. However, the models locally trained to a certain accuracy are pooled into the MBS to perform federated averaging. Different from the existing typical FL process, this paper does not require a process in which the parameter server periodically distributes the currently aggregated model parameters to the local model trainer. The federated averaging algorithm [11] is used in our FL framework, where the minibatch-based stochastic gradient descent method is adopted. To improve the generalization of the aggregated model, the MBS should store the history model parameters to participate in the aggregation of subsequent model parameters after random sampling. With the local models of $|\mathcal{V}|$ VUE agents, the parameters of the corresponding global model can be updated by

$$\begin{cases} \theta^{r+1} \leftarrow \sum_{v \in \mathcal{V}} \left(\varphi_0 \frac{\theta_v^{r+1}}{|\mathcal{V}|} + \ldots + \varphi_l \frac{\theta_v^{r+1-l}}{|\mathcal{V}|} \right) \\ \varphi_0 + \ldots + \varphi_l = 1 \end{cases} \tag{33}$$

where θ^{r+1} is the parameters of the global Q network updated by the MBS, while θ_v^{r+1} is the parameters of the local Q network trained by VUE v at coordination round $r+1$; l is the historical sample length after random sampling; $\varphi_0, \ldots, \varphi_l$ are weight coefficients, where the single value range is between 0 and 1, but the sum is equal to 1. Fig. 2 shows the overall framework of FL-aided DONA model. In addition, the pseudo-code of the proposed FL-DONA-based algorithm is described in Algorithm 2. From Propositions 1 and 2, we know that, as long as the number of VUE agents in the MBS coverage area as shown in Fig. 1 is relatively stable, the AO-DDQN algorithm can obtain an approximate optimal solution. Even if any new VUE agent is added, its personalized model will be trained quickly with the assistance of a foundation model. Therefore, it is feasible to obtain the approximate optimal solution in urban vehicle networks.

Algorithm 2 FL-DONA-Based Algorithm

```
MBS Initializes the online Q network model with $\theta$
MBS distributes this model to each VUE $v \in \mathcal{V}$
For each coordination round $r = 1, 2, \ldots$ do
    For each VUE $v \in \mathcal{V}$ do
        VUE $v$ executes Algorithm 1
        VUE $v$ uploads model parameters $\theta_v$ to the MBS
        MBS computes the global model parameters
        according to the formula (33) and distributes it to
        each new VUE
    End for
    If there is any VUE that needs more newly global
    model then
        This VUE requests MBS to distribute this global
        model to it
    End if
End for
```

VI. Performance Evaluation

A. Experimental Parameter Settings and Comparison Schemes

We consider a 3 × 3 km area, which consists of four intersections. Each intersection is deployed with a SBS. The MBS is located in the center of the area. Each road contains two lanes in each direction. The vehicles are generated by spatial Poisson process, and they are equipped with LTE, mmWave and THz radio interfaces. The simulation parameters for communication environment are listed in Table I. We use a DDQN network with two fully connected layers, where the number of neurons in hidden layer is 128. Pytorch (i.e., a deep learning framework) is adopted to build the above DDQN network, where Relu activation function and RMSprop optimizer are used. The simulation parameters for DRL are listed in Table II.

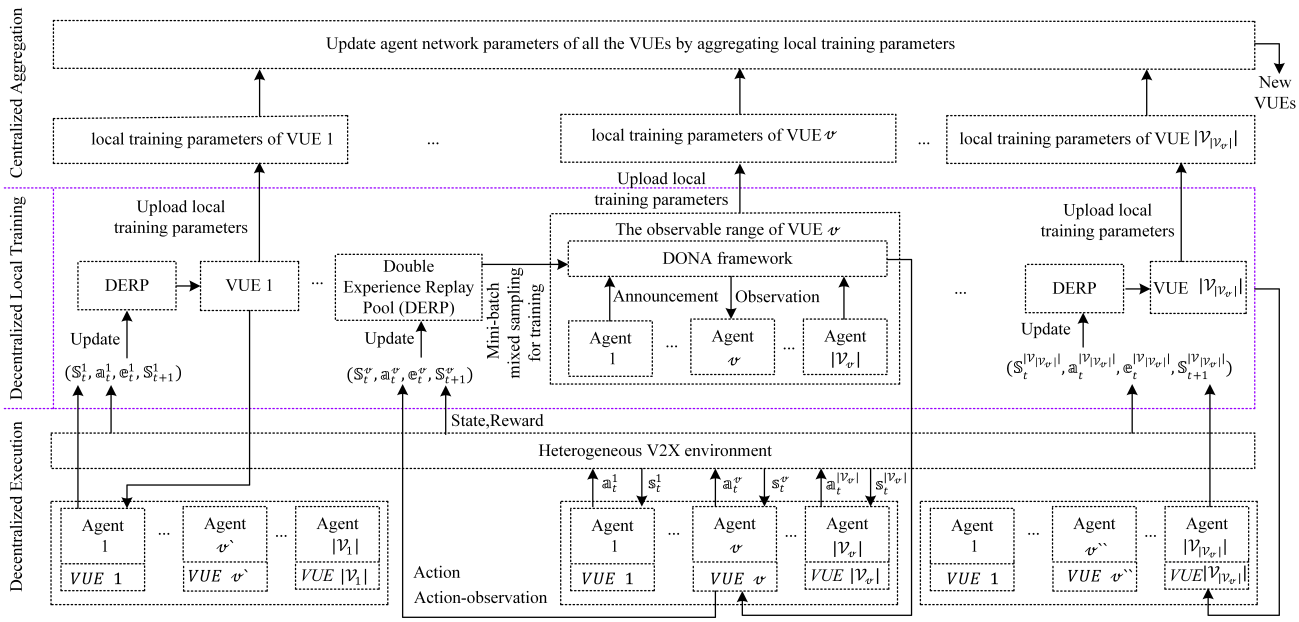

Centralized Aggregation
Update agent network parameters of all the VUEs by aggregating local training parameters
New VUEs
local training parameters of VUE 1
...
local training parameters of VUE v
...
local training parameters of VUE $|\mathcal{V}_{|\mathcal{V}_v|}|$
Decentralized Local Training
Upload local training parameters
DERP
VUE 1
...
Double Experience Replay Pool (DERP)
Mini-batch mixed sampling for training
Upload local training parameters
The observable range of VUE v
DONA framework
Announcement
Observation
Agent 1
Agent v
Agent $|\mathcal{V}_v|$
...
Upload local training parameters
DERP
VUE $|\mathcal{V}_{|\mathcal{V}_v|}|$
Update
$(\mathbb{s}_t^1, \mathbb{a}_t^1, \mathbb{e}_t^1, \mathbb{s}_{t+1}^1)$
Update
$(\mathbb{s}_t^v, \mathbb{a}_t^v, \mathbb{e}_t^v, \mathbb{s}_{t+1}^v)$
Update
$(\mathbb{s}_t^{|\mathcal{V}_{|\mathcal{V}_v|}|}, \mathbb{a}_t^{|\mathcal{V}_{|\mathcal{V}_v|}|}, \mathbb{e}_t^{|\mathcal{V}_{|\mathcal{V}_v|}|}, \mathbb{s}_{t+1}^{|\mathcal{V}_{|\mathcal{V}_v|}|})$
State,Reward
Decentralized Execution
Heterogeneous V2X environment
$\mathbb{a}_t^1$
$\mathbb{s}_t^1$
$\mathbb{a}_t^v$
$\mathbb{s}_t^v$
$\mathbb{a}_t^{|\mathcal{V}_v|}$
$\mathbb{s}_t^{|\mathcal{V}_v|}$
Agent 1
VUE 1
Agent v`
VUE v`
Agent $|\mathcal{V}_1|$
VUE $|\mathcal{V}_1|$
Action
Action-observation
Agent 1
VUE 1
Agent v
VUE v
Agent $|\mathcal{V}_v|$
VUE $|\mathcal{V}_v|$
Agent 1
VUE 1
Agent v``
VUE v``
Agent $|\mathcal{V}_{|\mathcal{V}_v|}|$
VUE$|\mathcal{V}_{|\mathcal{V}_v|}|$

Fig. 2. FL-aided DONA framework.

TABLE I
SIMULATION PARAMETERS FOR COMMUNICATION ENVIRONMENT

Symbol	Description	Value
$p_{\max}^{lte}$	Maximum transmission power for LTE interface of VUEs, SBSs and RSUs	23 dBm
$p_{\max}^{mm}$	Maximum transmission power for cellular mmWave interface of VUEs, SBSs and RSUs	23 dBm
$p_{\max}^{thz}$	Maximum transmission power for THz interface of VUEs, SBSs and RSUs	23 dBm
$w_{f\in\mathcal{F}_{lte}}$	LTE resource block	5 MHz
$w_{f\in\mathcal{F}_{mm}}$	Cellular mmWave resource block	1 GHz
$w_{f\in\mathcal{F}_{thz}}$	Cellular THz resource block	5 GHz
f_c^{lte}	Carrier frequency for LTE	2 GHz
f_c^{mm}	Carrier frequency for cellular mmWave	30 GHz
f_c^{thz}	Carrier frequency for THz	0.3 THz
$\|\mathcal{F}_{lte}\|$	Number of LTE RBs	3-12
$\|\mathcal{F}_{mm}\|$	Number of mmWave RBs	3-12
$\|\mathcal{F}_{thz}\|$	Number of THz RBs	3-12
$\|\mathcal{S}\|$	Number of SBSs	4
$\|\mathcal{V}\|$	Number of VUEs	5-50
σ_v^2	Noise power of each VUE v	-114 dBm
N_p^l, N_p^m, N_p^t	Transmission power levels	4
MM	Emergency message size	512 Byte
MB	Beacon message size	190 Byte
DB	Maximum tolerable delay for beacon messages	100 ms
DE	Maximum tolerable delay for emergency messages	20 ms
TB	Transmission interval for beacon messages	100 ms
TE	Transmission interval for emergency messages	3 s

To verify the efficiency of our AO-DDQN algorithm, we consider five comparison algorithms in our simulation experiments. The first is the DDQN-based algorithm without action observation (WO-DDQN). WO-DDQN is the improved algorithm based on the idea in [5], since the original algorithm cannot be directly applied to solve the problem in this paper.

TABLE II
SIMULATION PARAMETERS FOR DRL

Symbol	Description	Value
γ	Discount factor	0.99
ξ	Learning rate	0.0005
M	Experience replay pool size	500
bs	Mini-batch size	64
N_r	Target Q Network update frequency	200
$\mathbb{T}$	Number of time steps each epoch	4000
$\delta_1, \delta_2, \delta_3$	Weights in reward function	1,1,1
N_e	Neurons for hidden layers	128
ε_1	Start epsilon for ε-greedy	1
ε_2	End epsilon for ε-greedy	0.05

In [5], the communication occurs only between bound pairs of vehicles, which only considers the V2V and V2I modes between vehicle pairs. In order to make WO-DDQN suitable for broadcast and multi-type messages environments, we have extended the original algorithm from the following aspects. First, we have extended the state space by adding the channel gains from the transmitter to the first k potential vehicle receivers and all the SBSs as well as message type indicators. Second, we cancel the fixed cellular users in [5], since this type of users can correspond to the V2N message-type users in our environment. Finally, we change the reward function to

$$\begin{aligned} r_t = &\sum_{v\in\mathcal{V}} c_1 R_v + \sum_{v\in\mathcal{V}} c_2 \mathcal{H}_2\left(R_v^{(N)} - R_{min}\right) \\ &+ \sum_{v\in\mathcal{V}} c_3 \mathcal{H}_2\left(\gamma_v^{(V)} - \gamma_{eff}\right) \\ &+ \sum_{v\in\mathcal{V}} c_4 \mathcal{H}_2\left(R_v^{(V)} - \frac{L_v}{T_v}\right) \end{aligned} \quad (34)$$

where $\gamma_v^{(V)}$ is the SINR value of safety-critical messages received by VUE v, while γ_{eff} is the effective outage threshold; $R_v^{(V)}$ is the data rate of safety-critical messages received by VUE v, while $R_v^{(N)}$ is the data rate of high-capacity messages received by VUE v. c_1, c_2, c_3 and c_4 are weights of each part to balance the revenue and penalty.

The other four comparison algorithms are the multi-agent DDPG (MADDPG) [30], value decomposition network (VDN) [31], QMIX [26], and the random selection algorithm, respectively. MADDPG is a policy-based cooperative learning algorithm, while VDN and QMIX are value-based cooperative learning algorithms. However, they are based on CTDE framework, which use the same Markov decision process as that in this paper. In addition, the Gumbel-Softmax estimator is adopted to make MADDPG suitable for discrete action space. In the random selection algorithm, each vehicle randomly chooses an action from its available actions.

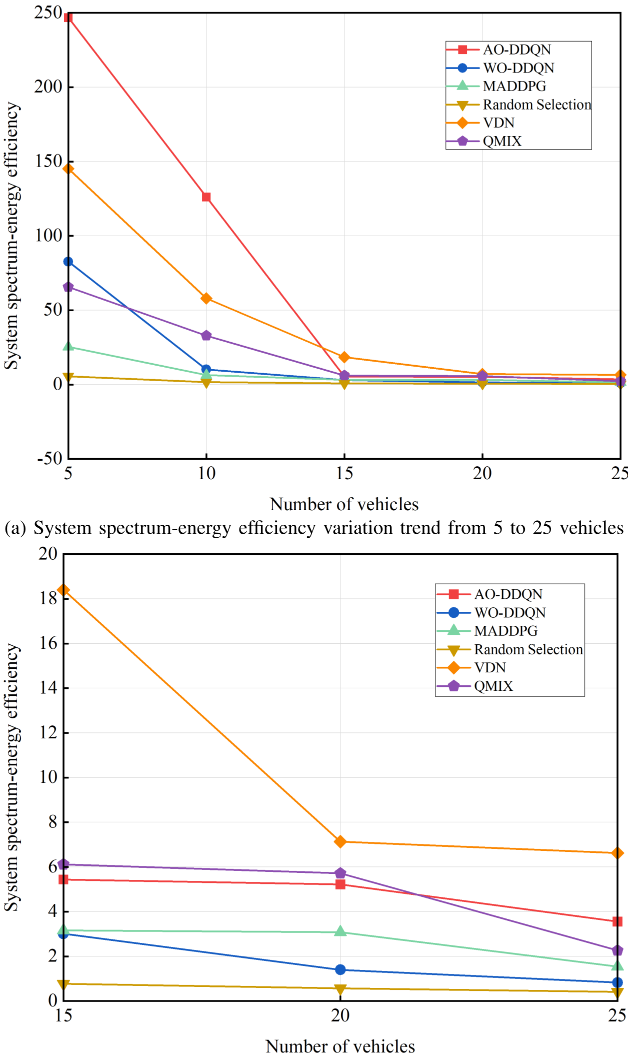
B. Network Performance Versus Number of Vehicles

In this sub-section, we compare our AO-DDQN algorithm with the other five algorithms (i.e., WO-DDQN, MADDPG, VDN, QMIX, and Random Selection) in terms of the following performance indicators with the different number of vehicles: system spectrum-energy efficiency, single-hop message satisfaction rate, and satisfaction rate of multi-hop messages containing $\mathcal{N}$ links. In particular, the two satisfaction rate indexes all indicate the proportion of messages meeting QoS requirements in the total number of sent messages.

From Fig. 3-Fig. 5, we can see that the proposed AO-DDQN always performs best among the six algorithms in terms of the two message satisfaction rates, where the number of RBs is set to 9. In terms of system spectrum-energy efficiency, AO-DDQN and VDN are generally superior to the other four algorithms. Although VDN outperforms AO-DDQN when the number of vehicles is relatively large, AO-DDQN has an advantage over VDN when the number of vehicles is relatively small. Overall, AO-DDQN still performs well. This is because each agent in AO-DDQN can be aware of the actions taken by its neighboring agents through observing the results of action choices made by them. In this case, more different channels are more likely to be chosen by different agents and then less co-channel interference would be caused naturally. Therefore, system throughput can be greatly improved.

Meanwhile, driven by the target of spectrum-energy efficiency, AO-DDQN does not blindly add new frequency band resources, but it quickly searches the appropriate transmission power parameters with the assistance of observing its neighboring agents' actions to achieve the target of improving spectrum-energy efficiency. Besides, since a dynamic equilibrium strategy and a multi-constraint reward function are adopted in AO-DDQN, it also contributes to finding better sub-optimal values for all the aforementioned indicators under the same number of training epochs.

In addition, we can see from Fig. 3(a) that, when the number of vehicles is less than 15, AO-DDQN is clearly superior to the other five algorithms in terms of system spectrum-energy efficiency. This is because AO-DDQN can more effectively coordinate the use of RBs to reduce co-channel interference when the number of vehicles is small. With the increasing number of vehicles, the fixed 9 RBs can no longer meet the



(a) System spectrum-energy efficiency variation trend from 5 to 25 vehicles

(b) System spectrum-energy efficiency variation trend from 15 to 25 vehicles

Fig. 3. System spectrum-energy efficiency versus number of vehicles.

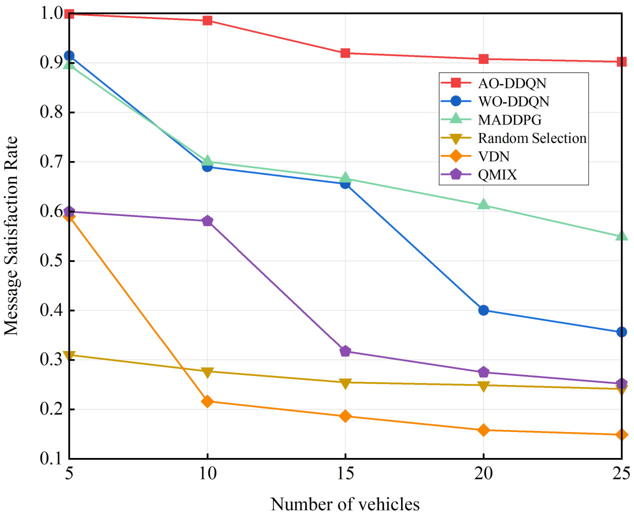


Fig. 4. Message satisfaction rate versus number of vehicles.

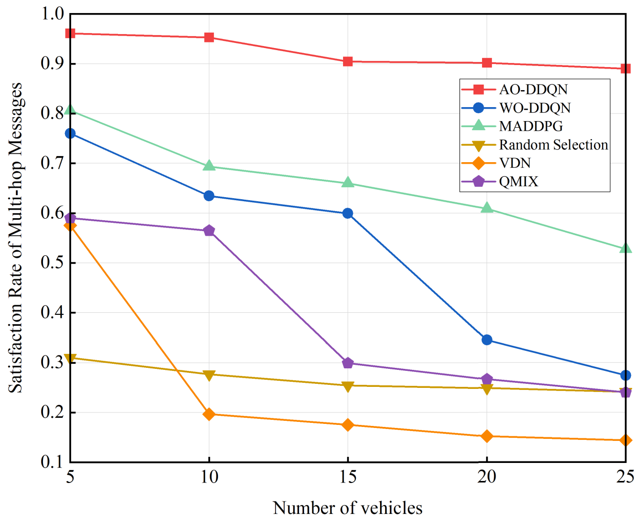


Fig. 5. Satisfaction rate of multi-hop messages versus number of vehicles.

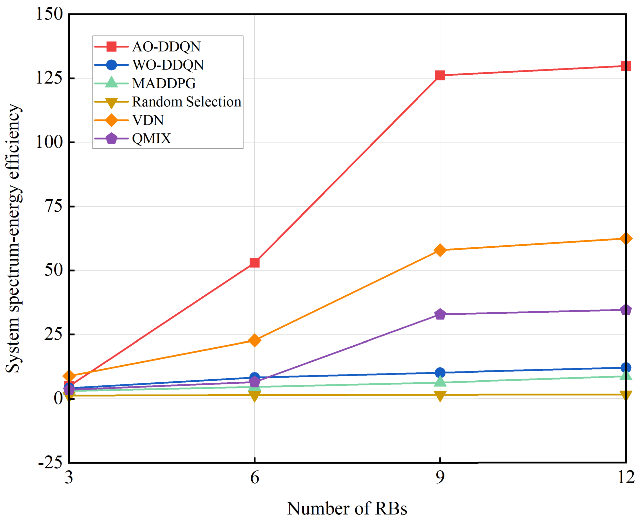


Fig. 6. System spectrum-energy efficiency versus the number of RBs.

requirement for effective coordination, so the differences of system spectrum-energy efficiency between the six algorithms decreased significantly, as shown in Fig. 3(b).

From Fig. 4 and Fig. 5, we can see that, when the number of vehicles increases, both AO-DDQN and Random Selection algorithms show a smoother decline in the two satisfaction rates than the other four algorithms. That is, AO-DDQN and Random Selection show the better stability in system performance. However, due to the blindness of the choice of actions, Random selection algorithm is only better than VDN when the number of vehicles is greater than 10, while it is always worse than AO-DDQN, WO-DDQN, MADDPG, QMIX. Therefore, we can make a conclusion that our AO-DDQN algorithm not only meets the QoS requirements mentioned in previous sections but also effectively deals with multiple broadcast scenarios with heterogeneous technologies and messages. Besides, we can see from Fig. 4 and Fig. 5 that MADDPG has relatively better performance when compared with the other four comparison algorithms. However, our AO-DDQN algorithm has improved performance by 12%-82% when compared to MADDPG. Since the model scale of MADDPG is hundreds of times larger than that of AO-DDQN proposed in this paper, it always causes huge training costs (e.g., memory usage and power consumption) and is unsuitable to handle the large number of vehicles.

VDN and QMIX are two DRL methods based on value decomposition. The key objective of these algorithms is to establish the relationship between the global value function Q_{total} and the individual value functions $[Q_1, Q_2, \ldots, Q_{|\mathcal{V}|}]$, while satisfying $\frac{\partial Q_{total}}{\partial Q_i} \geq 0$ $(i = 1, 2, \ldots, |\mathcal{V}|)$. This makes them well-suited for fully cooperative tasks. However, in this paper, each agent needs to prioritize the QoS requirements of messages before maximizing spectrum-energy efficiency. Consequently, the optimization problem (21) becomes a hybrid task, where VDN and QMIX exhibit suboptimal performance. Fig. 3 shows that they outperform the other three comparison algorithms in terms of spectrum-energy efficiency, and even exceed the AO-DDQN proposed in this paper when the number of vehicles exceeds 10. However, Fig. 4 and Fig. 5 show that they cannot consistently achieve a high message satisfaction rate. These findings demonstrate that VDN and QMIX sacrifice some of their individual performance to optimize the common goal, which is inherent to their model nature.

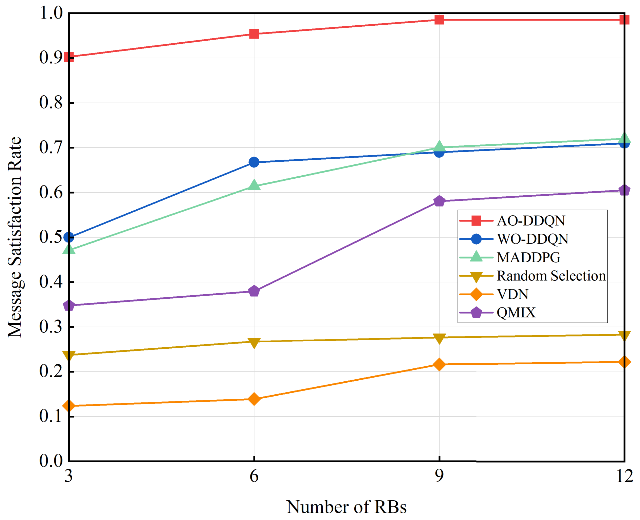


Fig. 7. Message satisfaction rate versus number of RBs.

C. Network Performance Versus Number of RBs

Fig. 6-Fig. 8 show the performance comparison under the different number of RBs, where the number of vehicles is set to 10 vehicles. From the Fig. 6, we can see that the spectrum-energy efficiency of AO-DDQN increases significantly with the number of RBs and then it tends to flatten out. Before the number of RBs exceeds 9 RBs, the increase of RBs can cause rapid growth of spectrum-energy efficiency. This is because the number of RBs is smaller than that of VUEs. In this case, the increase of RBs can ease the competition of vehicles for RBs. However, after the number of RBs exceeds 9, the amount of RBs exceeds the VUEs' demand, and thus the spectrum-energy efficiency cannot be improved significantly. We can further conclude that taking spectrum-energy

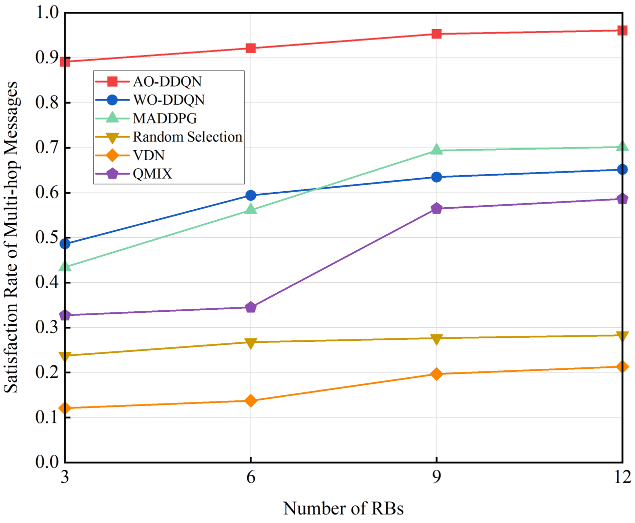


Fig. 8. Satisfaction rate of multi-hop messages versus number of RBs.

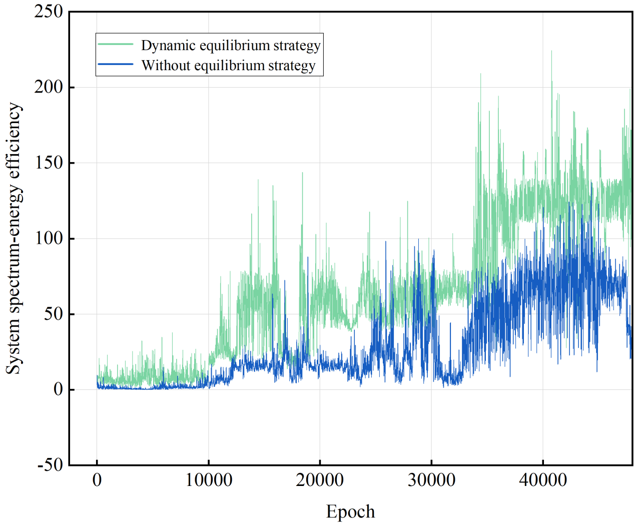


Fig. 9. System spectrum-energy efficiency variation trend in model training process.

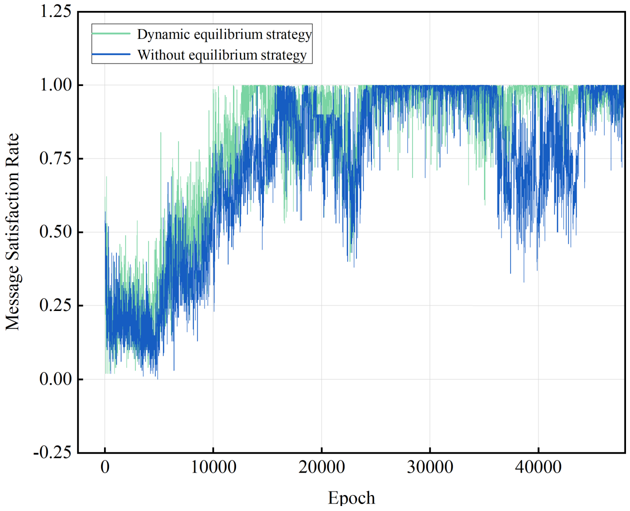


Fig. 10. Message satisfaction rate variation trend in model training process.

efficiency as an optimization goal can significantly reduce resource waste and achieve efficient utilization of resources.

From Fig. 7 and Fig. 8, we can see that the two satisfaction rate indicators show the basic similar variation trend to spectrum-energy efficiency. The main reasons are similar to the above explanation for Fig. 6. We can also observe that AO-DDQN can maintain a relatively high satisfaction rate in both single-hop sending and multi-hop forwarding. Even if the resources are far less than the VUEs' requirements, a satisfaction rate of about 90% is guaranteed. This also shows the high efficiency and stability of AO-DDQN in the case of resource shortage.

In addition, we can see from Fig. 7 and Fig. 8 that, compared with WO-DDQN and MADDPG, the proposed algorithm has improved the performance by about 36%-80%. The boost is more pronounced when resources are scarce. This is because AO-DDQN has more alternative communication modes. In particular, the high propagation loss characteristics of mmWave and THz greatly reduce co-channel interference between VUEs and further enhance the quality of vehicle communication.

Although both VDN and QMIX are value decomposition networks, the global value function Q_{total} of QMIX is generated through a neural network based on the global state $\mathbb{S}$ instead of a simple sum of $\{Q_i\}_{i=1,2,\ldots,|\mathcal{V}|}$ like VDN. Therefore, QMIX better fits the relationship between Q_{total} and $\{Q_i\}_{i=1,2,\ldots,|\mathcal{V}|}$ and outperforms VDN in terms of message satisfaction rate. For the random selection algorithm, because the strategy is purposeless, its performance is definitely worse than those of AO-DDQN, WO-DDQN, MADDPG, and QMIX. However, it still outperforms VDN.

D. Network Performance Versus Equilibrium Processing

Fig. 9-Fig. 12 show the model convergence of our AO-DDQN algorithm with and without an equilibrium strategy. From these figures, we can see that the system performance and model convergence speed are improved when using dynamic equilibrium strategy. We can also see from Fig. 9-Fig. 11 that, compared with the non-use of dynamic equalization strategy, the system spectrum-energy efficiency is improved by 92.17%, the satisfaction rate is improved by 0.68%, and the multi-hop message satisfaction rate is improved by 0.55%. From the change of individual reward in Fig. 12, we can see that, the model is more stable when using the equilibrium strategy, and the QoS requirements of various messages can be met more quickly.

The improvements mentioned above are mainly because the dynamic equilibrium strategy can record the history of positive feedback during exploration. Thus, it ensures the balance of the samples fed into the model at the beginning of the model training, making the model training more stable and balanced. With the increase of training epochs, the frequency of non-negative rewards increases, and the strategy can dynamically adjust the sample proportion in the batch to ensure the diversity of samples. In this way, the model training will become more adequate and will not be affected by the sparseness of the non-negative reward in the early stage. We observe that the model can also converge when

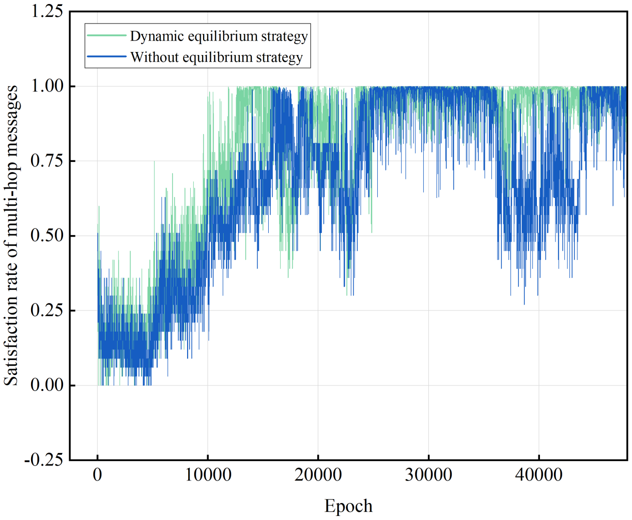


Fig. 11. Satisfaction rate variation trend of multi-hop messages in model training process.

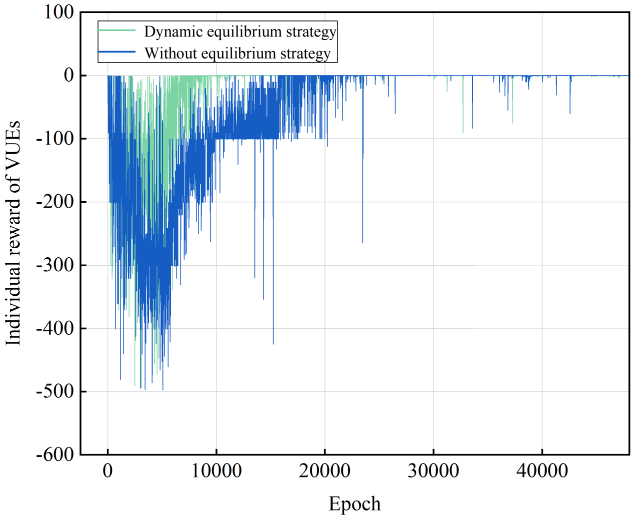


Fig. 12. VUEs' individual reward $\mathbb{ie}_t$ variation trend in model training process.

the equilibrium strategy is not used. However, due to the lack of good action selection records in the early exploration, excessive negative rewards make the model less able to express non-negative rewards at the early stage.

VII. Conclusion and Futurn Researches

In this paper, we proposed a novel communication mode selection and resource allocation scheme by combining FL framework with DONA framework. In the proposed scheme, the AO-DDQN algorithm is deployed in each VUE to act as an agent. By observing the actions of other agents and dynamically balancing the historical samples of positive and negative reward values, the AO-DDQN algorithm can obtain the fast convergence results in heterogeneous V2X broadcast networks. Also, the FL-DONA-based algorithm in our scheme can obtain a generalization model and keep the personality of each local model. Here, good model generalization is achieved by randomly sampling historical model parameters to participate in model parameter aggregation, and the generalized model is only provided to the new agents rather than the old agents to avoid influencing the personality of the individual model. Simulation experimental results showed that, in most cases, the proposed scheme has advantages over the comparison algorithms in terms of system spectrum-energy efficiency, single-hop message satisfaction rate, and satisfaction rate of multi-hop messages containing $\mathcal{N}$ links. We also observed that, when using the dynamic equilibrium strategy for training samples, the system spectrum-energy efficiency, single-hop satisfaction rate, and multi-hop message satisfaction rate are improved by 92.17%, 0.68%, and 0.55%, respectively.

However, in this paper, we assume that any message transfer task is accomplished via single communication mode selection. While this assumption is realistic for safety-critical messages with short packets, it may require multiple communication mode selections for a high-capacity message. Especially in highly dynamic vehicle environments, physical communication links may break down frequently, and the resources determined by one communication mode selection may quickly become unavailable. In the future, we plan to explore the problem. In addition, to simplify the formulas (1)-(6), we assume that the MBS can coordinate the synchronization of uplink and downlink communications of all the SBSs to avoid mutual interference between them. In our future work, we will discard this assumption and explore the communication mode selection and resource allocation problem under more complex mutual interference relations.

References

[1] Z. Pei, W. Chen, C. Li, L. Du, H. Liu, and X. Wang, "Analysis and optimization of multihop broadcast communication in the Internet of Vehicles based on C-V2X mode 4," *IEEE Sensors J.*, vol. 22, no. 12, pp. 12428–12443, Jun. 2022.

[2] J. B. Kenney, "Dedicated short-range communications (DSRC) standards in the United States," *Proc. IEEE*, vol. 99, no. 7, pp. 1162–1182, Jul. 2011.

[3] S. Gyawali, S. Xu, Y. Qian, and R. Q. Hu, "Challenges and solutions for cellular based V2X communications," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 222–255, 1st Quart., 2020.

[4] A. Bazzi, A. O. Berthet, C. Campolo, B. M. Masini, A. Molinaro, and A. Zanella, "On the design of sidelink for cellular V2X: A literature review and outlook for future," *IEEE Access*, vol. 9, pp. 97953–97980, 2021.

[5] X. Zhang, M. Peng, S. Yan, and Y. Sun, "Deep-reinforcement-learning-based mode selection and resource allocation for cellular V2X communications," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6380–6391, Jul. 2020.

[6] H. Ye, G. Y. Li, and B.-H. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[7] D. Zhao, H. Qin, B. Song, Y. Zhang, X. Du, and M. Guizani, "A reinforcement learning method for joint mode selection and power adaptation in the V2V communication network in 5G," *IEEE Trans. Cogn. Commun. Netw.*, vol. 6, no. 2, pp. 452–463, Jun. 2020.

[8] L. Liang, H. Ye, and G. Y. Li, "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2282–2292, Oct. 2019.

[9] A. D. Mafuta, B. T. J. Maharaj, and A. S. Alfa, "Decentralized resource allocation-based multiagent deep learning in vehicular network," *IEEE Syst. J.*, vol. 17, no. 1, pp. 87–98, Mar. 2023.

[10] Y.-H. Xu, C.-C. Yang, M. Hua, and W. Zhou, "Deep deterministic policy gradient (DDPG)-based resource allocation scheme for NOMA vehicular communications," *IEEE Access*, vol. 8, pp. 18797–18807, 2020.

[11] O. A. Wahab, A. Mourad, H. Otrok, and T. Taleb, "Federated machine learning: Survey, multi-level classification, desirable criteria and future directions in communication and networking systems," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 2, pp. 1342–1397, 2nd Quart., 2021.

[12] W. Sun, D. Yuan, E. G. Ström, and F. Brännström, "Cluster-based radio resource management for D2D-supported safety-critical V2X communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2756–2769, Apr. 2016.

[13] L. Liang, S. Xie, G. Y. Li, Z. Ding, and X. Yu, "Graph-based resource sharing in vehicular communication," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4579–4592, Jul. 2018.

[14] M. Peng, Y. Li, T. Q. S. Quek, and C. Wang, "Device-to-device underlaid cellular networks under Rician fading channels," *IEEE Trans. Wireless Commun.*, vol. 13, no. 8, pp. 4247–4259, Aug. 2014.

[15] L. Liang, J. Kim, S. C. Jha, K. Sivanesan, and G. Y. Li, "Spectrum and power allocation for vehicular communications with delayed CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 458–461, Aug. 2017.

[16] X. Li, L. Ma, R. Shankaran, Y. Xu, and M. A. Orgun, "Joint power control and resource allocation mode selection for safety-related V2X communication," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 7970–7986, Aug. 2019.

[17] C. Wu, T. Yoshinaga, Y. Ji, and Y. Zhang, "Computational intelligence inspired data delivery for vehicle-to-roadside communications," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12038–12048, Dec. 2018.

[18] S. Yan, X. Zhang, H. Xiang, and W. Wu, "Joint access mode selection and spectrum allocation for fog computing based vehicular networks," *IEEE Access*, vol. 7, pp. 17725–17735, 2019.

[19] R. F. Atallah, C. M. Assi, and M. J. Khabbaz, "Scheduling the operation of a connected vehicular network using deep reinforcement learning," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1669–1682, May 2019.

[20] T. Dang and M. Peng, "Joint radio communication, caching, and computing design for mobile virtual reality delivery in fog radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 7, pp. 1594–1607, Jul. 2019.

[21] K. Zhang, Y. Zhu, S. Leng, Y. He, S. Maharjan, and Y. Zhang, "Deep learning empowered task offloading for mobile edge computing in urban informatics," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7635–7647, Oct. 2019.

[22] H. Ren and M. Q.-H. Meng, "Game-theoretic modeling of joint topology control and power scheduling for wireless heterogeneous sensor networks," *IEEE Trans. Autom. Sci. Eng.*, vol. 6, no. 4, pp. 610–625, Oct. 2009.

[23] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," in *Handbook of Reinforcement Learning and Control*. Cham, Switzerland: Springer, 2021, pp. 321–384.

[24] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5872–5881.

[25] Q. Tang, R. Xie, F. R. Yu, T. Huang, and Y. Liu, "Decentralized computation offloading in IoT fog computing system with energy harvesting: A Dec-POMDP approach," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 4898–4911, Jun. 2020.

[26] T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson, "Monotonic value function factorisation for deep multi-agent reinforcement learning," *J. Mach. Learn. Res.*, vol. 21, no. 1, pp. 7234–7284, 2020.

[27] Z. Guo, Z. Chen, P. Liu, J. Luo, X. Yang, and X. Sun, "Multi-agent reinforcement learning-based distributed channel access for next generation wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 5, pp. 1587–1599, May 2022.

[28] X. Lu, L. Xiao, T. Xu, Y. Zhao, Y. Tang, and W. Zhuang, "Reinforcement learning based PHY authentication for VANETs," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3068–3079, Mar. 2020.

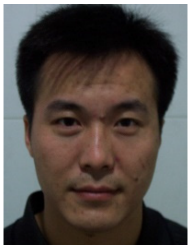
[29] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Netw.*, vol. 33, no. 5, pp. 156–165, Sep. 2019.

[30] R. Lowe, Y. I. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 6379–6390.

[31] P. Sunehag et al., "Value-decomposition networks for cooperative multi-agent learning," in *Proc. Int. Conf. Auton. Agents MultiAgent Syst. (AAMAS)*, 2018, pp. 2085–2087.

Jinsong Gui (Member, IEEE) received the B.E. degree from the University of Shanghai for Science and Technology, China, in 1992, and the M.S. and Ph.D. degrees from Central South University, China, in 2004 and 2008, respectively. He is currently a Professor with the School of Electronic Information, Central South University. He has published over 60 international journal articles and over ten international conference papers. His research interests include distributed systems and related fields, such as wireless network topology control, cloud and green computing, and network trust and security. He is a member of the China Computer Federation (CCF).

Liyan Lin is currently pursuing the master's degree with the School of Computer Science and Engineering, Central South University, China. Her research interests include the Internet of Vehicles, network simulation, and performance evaluation.



Xiaoheng Deng (Senior Member, IEEE) received the Ph.D. degree in computer science from Central South University, Changsha, Hunan, China, in 2005. His research interests include wireless communications and networking, congestion control for wired/wireless networks, cross-layer route design for wireless mesh networks and ad hoc networks, online social network analysis, and edge computing. He is a Senior Member of CCF and a member of the CCF Pervasive Computing Council and ACM.

Lin Cai (Fellow, IEEE) received the M.A.Sc. and Ph.D. degrees in electrical and computer engineering from the University of Waterloo, Waterloo, Canada, in 2002 and 2005, respectively. Since 2005, she has been with the Department of Electrical and Computer Engineering, University of Victoria, where she is currently a Professor. Her research interests include communications and networking, with a focus on network protocol and architecture design supporting emerging multimedia traffic and the Internet of Things. She is an NSERC E.W.R. Steacie Memorial Fellow, an Engineering Institute of Canada (EIC) Fellow, and a Canadian Academy of Engineering (CAE) Fellow. In 2020, she was elected as a member of the Royal Society of Canada's College of New Scholars, Artists and Scientists, and the 2020 "Star in Computer Networking and Communications" by N2Women. She has co-founded and chaired the IEEE Victoria Section Vehicular Technology and Communications Joint Societies Chapter. She was elected to serve the IEEE Vehicular Technology Society Board of Governors from 2019 to 2024 and served its VP Mobile Radio in 2023. She has been a Voting Board Member of IEEE Women in Engineering from 2022 to 2023. She has served as an Associate Editor-in-Chief for IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY; a member of the Steering Committee of IEEE TRANSACTIONS ON MOBILE COMPUTING, IEEE TRANSACTIONS ON BIG DATA, and IEEE TRANSACTIONS ON CLOUD COMPUTING; an Associate Editor of IEEE INTERNET OF THINGS JOURNAL, IEEE/ACM TRANSACTIONS ON NETWORKING, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY, and IEEE TRANSACTIONS ON COMMUNICATIONS; and the Distinguished Lecturer for the IEEE VTS Society and the IEEE Communications Society.