HierNet: A Hierarchical Resource Allocation Method for Vehicle Platooning Networks

Xiaoyuan Fu, Quan Yuan, Guiyang Luo, Nan Cheng, Yang Li, Junfei Wang, and Jianxin Liao

***Abstract*—Vehicle platooning is a promising traffic model in intelligent transportation systems (ITS), which can effectively improve resource utilization and reduce traffic congestion. The resource allocation for vehicle-to-everything (V2X) communications that consist of intra-platoon communications and inter-platoon communications is crucial for safe operation of multiple vehicular platoons. Considering dynamic coordination pattern of vehicular platoons and layered architecture of vehicle platooning networks, a hierarchical resource decision-making framework is proposed in this paper. In the proposed framework, the resource decision-making process is divided into two levels. The high level that generates and distributes coordination meta policy is deployed on base station (BS), and the low level that generates ego resource decisions is deployed in each platoon. To deal with optimization of resource allocation for multi-platoon V2X communications, a hierarchical reinforcement learning method (HierNet) is designed based on the proposed hierarchical decision-making framework. In HierNet, meta policy of the high level can be preserved, and needs to be updated only when cooperative conditions of multiple platoons undergo distinct changes. Simulation experiments have demonstrated that our proposed method not only optimizes resource efficiency, but also reduces the communication costs for resource decision-making of vehicle platooning networks.**

***Index Terms*—vehicular networks, resource allocation, vehicle platooning, hierarchical reinforcement learning**

I. Introduction

With the rapid development of the Internet of vehicles (IoV) and autonomous driving technologies [1], intelligent autonomous driving platoons have received increasing attention [2]. Vehicle platooning is a promising application of autonomous driving, in the form of a collection of vehicles travelling in the same direction and speed [3]. A platoon consists of a trustworthy leading vehicle (LV) and its accompanying following vehicles (FVs). Platooning formation requires the LV and FVs to drive at a unified speed in a queue with the same distance between adjacent vehicles in dynamic road and traffic conditions. Vehicle platooning is an efficient operating form of intelligent transportation systems (ITS), which can effectively improve resource utilization, including reducing vehicle fuel consumption, lowering overall vehicle control costs, and improving traffic management efficiency [4]. However, the maintenance and mobility of platoons involve collaborative control of multiple platoons, bringing challenges to resource allocation of multi-platoon systems [5].

The control of platoons requires the assistance of vehicle-to-everything (V2X) communications [6]. The transmission of safety messages is completed through communications between LVs and their FVs, as well as communications between LVs. LV is responsible for managing FVs in terms of regularly transmitting control and coordination information to maintain vehicle platooning formation. Assuming that intra-platoon communications are generally completed through broadcasting by LVs, and platooning coordination messages can be transmitted through wireless communications between vehicles. LV also needs to perceive environmental information such as road conditions and traffic flow, and to provide timely feedback on changes in environmental conditions through inter-platoon communications. The inter-platoon coordination is carried out through the communications between LVs and base station (BS). The interactions of BS and multiple platoons form a hierarchical V2X communication architecture. As a result, the V2X resource allocation of multiple platoons in vehicle platooning networks could be regard as a hierarchical decision-making structure.

Recently, multi-agent reinforcement learning (MARL) has received widespread attention in V2X resource allocation in IoV [7]–[9]. In centralized RL-based decision-making, central nodes could make resource allocation decisions based on global observation information, and distribute them to each vehicle [10]. When resource decisions are distributed without reference to current environment conditions, it will cause decision-making timeliness problems. In distributed MARL decision-making, each agent makes separate decisions without considering collaboration between multiple agents, which also leads to decision-making mistakes [11]. Meanwhile, it is noteworthy that frequent communications between multiple platoons in multi-agent RL methods can waste a significant amount of communication resources. Hierarchical RL (HRL) is designed to solve large dimensionality problem and training difficulties of RL algorithms, which shows great superiority in large-scale and complex scenarios [12].

There are several research works applying HRL to deal with

This work was supported in part by the National Key Research and Development Program of China under Grant 2023YFB4301904, in part by the Natural Science Foundation of China under Grant 62272053 and Grant 62102041, in part by the Beijing Nova Program under Grant 20230484364, in part by the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) under Grant 2022QNRC001, and in part by the BUPT Innovation and Entrepreneurship Support Program under Grant 2024-YC-A086. (*Corresponding author: Quan Yuan.*)

Xiaoyuan Fu and Quan Yuan are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China, and also with the State Key Laboratory of Integrated Services Networks (ISN), Xidian University, Xi'an, Shaanxi, 710071, China (e-mail: fuxiaoyuan@bupt.edu.cn; yuanquan@bupt.edu.cn).

Guiyang Luo, Yang Li, and Jianxin Liao are with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing, 100876, China (e-mail: luoguiyang@bupt.edu.cn; leeyang866@bupt.edu.cn; liaojx@bupt.edu.cn).

Nan Cheng is with State Key Lab. of ISN and with School of Telecommunications Engineering, Xidian University, Shaanxi, 710071, China (e-mail: dr.nan.cheng@ieee.org).

complex decision-making problems. Zhang *et al.* proposed an HRL-based energy matching scheme to reduce computational complexity by decomposing the decision-making space [13]. Kim *et al.* separated decision-making policies with HRL for band selection and beam management in multi-band communications [14]. In [15], Yu *et al.* utilized a two-layer decision-making framework to mitigate bus bunching through an HRL-based method. In vehicle platooning networks, the combination of platoons is stochastic and dynamic. Also, the environmental observations of each LV and FVs are constantly changing for the mobility of vehicles. The cooperation pattern between platoons needs to timely adapt to the dynamic changes of the environment. LVs can obtain their current observations though they cannot obtain overall environment states from a global perspective.

Therefore, a hierarchical resource decision-making framework is proposed to be applicable to the hierarchical V2X structure in vehicle platooning networks. BS is applied to generate coordination information, and Mode 4 is adopted, i.e. the final allocation of communication resources are completed by LVs. As shown in Fig. 1, vehicles in different colors represent the LVs of different platoons. The coordination strategies issued by BS are called meta policy, and the resource strategies issued by LVs are called ego policy. During the hierarchical decision-making process, each vehicle feedbacks channel state information (CSI) to the BS. BS generates meta policy based on the global state to guide low level decision-making. When the cooperation pattern of the platoon has not changed, resource decisions can be generated using the preserved meta policy and ego policy. When the environmental conditions of some platoons undergo significant changes, the coordination of multiple platoons will also transform to other patterns. In this situation, only meta policy needs to be updated. In summary, the key contributions of this work are listed as follows:

- A hierarchical system framework for resource allocation in vehicle platooning has been proposed. Considering dynamic nature of vehicular platoons and limited resources of vehicles, the decision-making process is divided into two levels. The high level is deployed on BS, and the low level is deployed on the LV of each platoon. The high level is able to perceive global information and provide guidance for decision-making of the low level.
- The multi-objective resource optimization problem is formulated for multi-platoon networks, in which intra-platoon communications and inter-platoon communications are both optimized to improve safety of vehicle platooning systems.
- An HRL method (HierNet) is proposed to solve the multi-objective problem based on the proposed hierarchical system framework. The meta agent generates coordinate information (meta policy) based on global perspective, and distributes them to multiple ego agents. Ego agents generate resource allocation decisions according to their local observations and the distributed meta policy.
- Extensive simulation experiments have demonstrated the effectiveness of the proposed method. The proposed method not only has better performance than existing MARL methods, but also reduces costs of vehicular communications for resource decision-making.

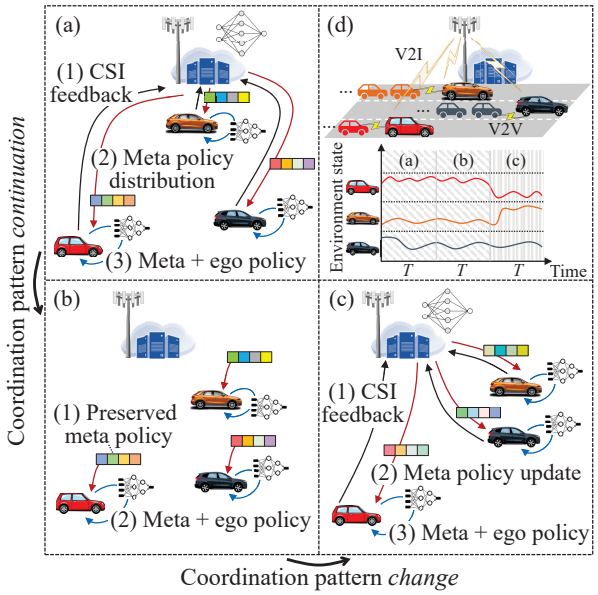

Fig. 1: The hierarchical decision-making process to deal with the change of coordination pattern. (a) Training of hierarchical decision-making process with generation and distribution of ego policy and meta policy. (b) Using the preserved meta policy when coordination pattern is not changed. (c) The update of meta policy when coordination pattern changes. (d) The obvious changes of environment state brings change of coordination pattern.

The reminder of this paper is organized as follows. The related works are presented in Section II. The system model and problem formulation are described in Section III. In Section IV, we transform the formulated problem to Markov game in vehicle platooning scenarios. Next, we introduce HierNet, and illustrate training of meta layer and ego layer in HierNet. Section V presents the simulation settings and shows the simulation results of our method. Finally, the paper is concluded, and future works are proposed in Section VI.

II. Related Works

Vehicle platooning has received extensive attention recently due to its potential to increase road capacity, reduce traffic congestion, and enhance transportation safety [16]. In vehicle platooning networks, a virtual chain-like combination of vehicles, also known as a vehicular platoon, is formed by vehicles with similar driving patterns [17]. Each platoon consists of a LV that collects and disseminates road and control information, along with a group of controlled FVs. To ensure the safety of multiple platoons, LVs must generate and distribute cooperative awareness messages (CAM) promptly to other platoons, and guide their corresponding FVs [18]. Given the limited availability of wireless resources, designing efficient resource allocation methods for V2X communications of vehicle platooning networks is essential to meet quality of service (QoS) requirements and bolster safety [19].

More studies have been conducted on resource allocation for vehicle platooning recently, and various methods have been proposed [20]. Li *et al.* [21] designed a relay-based chained data dissemination protocol for intra-platoon communications, and proposed a dynamic programming algorithm to optimize queue-based resource allocation. Cao *et al.* [22] applied two communication resource allocation techniques for vehicle platooning: the improved random selection scheme and deep reinforcement learning (DRL), both grounded in the semi-persistent scheduling stated in the 3rd Generation Partnership Project (3GPP). Furthermore, Chai *et al.* [23] proposed a platoon-based dynamic multicast cluster formation scheme, where the resource allocation issue was designed as a distributed robust optimization problem aiming to maximize vehicle-to-infrastructure (V2I) capacities while securing the dependability of intra-platoon communications. A joint resource allocation and coding rate optimization algorithm was proposed in [24]. Gao *et al.* [25] addressed changes in platoon topologies, and proposed a spectrum sensing scheduling scheme utilizing a greedy algorithm to facilitate resource allocation.

With the development of multi-agent systems and DRL [26], there has been a significant rise in the popularity and application of MARL in recent years. Compared to traditional methods, MARL approaches require less prior knowledge and have lower time overhead. As a result, many studies have utilized MARL in resource allocation problems in vehicular networks, specifically focusing on the optimization of both vehicle-to-vehicle (V2V) and V2I communications. Hammami *et al.* [27] proposed an advantage actor-critic based MARL approach, including shared-critic-shared-reward and non-shared-critic-shared-reward, to solve the resource allocation problem for V2I and V2V links while meeting their QoS requirements. FedMARL [28] utilized dueling double deep Q-networks for V2V channel selection and power control, which incorporates federated learning to mitigate instability issues arising from the multi-agent environment. In contrast to the previous studies, Xiang *et al.* [29] used a CSI-independent observation space, which utilizes interference power measurements. Xu *et al.* [30] conducted a study with the aim of maximizing the transmission success ratio of intra-platoon communications and the mean opinion score of the V2I communication links. This multi-objective optimization problem was decomposed into multiple scalar optimization sub-problems using the contribution-based dual-clip proximal policy optimization algorithm. In [31], the authors leveraged the age of information (AoI) for controlling the message sending rate of CAM. In addition, Parvini *et al.* [11] focused on minimizing AoI and maximizing the probability of CAM transmission, where two MARL frameworks were introduced based on deterministic policy gradient methods.

Hierarchical decision-making is considered as a very promising framework in reinforcement learning. Due to issues such as large action space, long trajectories, and sparse rewards, traditional reinforcement learning often lacks exploration capabilities and has poor performance when facing long-term and complex tasks. In HRL framework, a long-term reinforcement learning task is decomposed into a hierarchy of sub-problems or sub-tasks which are easier to deal with [32]. In recent years, several studies have utilized HRL in resource allocation of wireless networks. Ye *et al.* [33] proposed a slicing scheme based on HRL. In this scheme, the high-level controller allocates resource blocks for each slice based on the traffic state at a large time scale, while the low-level controller allocates resources for each user based on the physical link information at a small time scale. Geng *et al.* [34] investigated a two-level HRL approach to minimize outage probability in a two-hop cooperative relay network, which hierarchically selects relays and power levels as a result of shrinking the action space. In Wi-Fi networks, HRL-TPCCCA [35] decomposed the problem of optimizing clear channel assessment threshold and transmission power into two steps, and solved it by an HRL-based method. Additionally, in intelligent reflecting surface aided wireless networks, AoI minimization and transmission were decided at a large time scale, and the uplink information transmission and downlink energy transfer to all nodes were decided at a small time scale [36].

In vehicle platooning networks, the roles of vehicles are dynamic, and combination pattern of vehicles is constantly changing as well. The mode of cooperation between multiple platoons needs to constantly adapt to the dynamic changes of the environment. In distributed MARL methods, each LV usually makes decisions independently without timely considering the collaboration of platoons. The change of traffic characteristics will lead to the invalidation of resource allocation decisions. At the same time, frequent communications between platoons in communication-enabled MARL could cause a waste of communication resources. As a result, an HRL-based resource allocation method for V2X communications of vehicle platooning is proposed in this work.

III. System Model And Problem Formulation

In this section, we first describe system scenarios and introduce communication model, and then we present the formulation of resource optimization problem.

A. System Description and Communication Model

We consider a vehicle platooning system, mainly consisting of one BS and multiple platoons, based on a hierarchical resource decision-making framework. Each LV is an ego resource allocation agent that maintains connections with a central meta agent at BS. The resource decision-making is divided into two levels including coordination guidance in high level and resource allocation decision in low level. The central node in BS generates coordination information at a certain interval to guide resource allocation decisions of low level LVs. It is able to perceive global state and then generate coordination information for multiple platoons. The coordination information, i.e. meta policy, implies observations and tactics of other agents, fitting a cooperative pattern of multiple ego agents. With the aid of meta policy, intelligent ego agents at low level can make better resource decisions using coordination information and personal observations. The environment state is dynamically changing, leading to dynamic changes in cooperation pattern of multiple platoons. The meta

policy is responsible for capturing the changing patterns of multi-platoon cooperation and providing guidance for multi-platoon collaboration. When cooperation pattern of platoons changes, only training of the central node needs to be updated, reducing the cost of retraining the underlying networks in LVs.

Each platoon includes one LV that engages in both V2I and V2V communications, as well as a set of FVs involved in V2V communications. Specifically, LV collects environment states using various types of integrated sensors and high-definition cameras, and transmits security-related messages to FVs through V2V multicast. The infotainment traffic for FVs can be transmitted from LV to FV via V2V links as well. To gather global environmental conditions and potential resource tactics of other platoons, LV should also receives cooperation guidance generated from BS through V2I links. To fully utilize limited spectrum resources, we assume that the V2I and V2V links reuse radio channels with a total bandwidth of W Hertz. The system bandwidth is divided into M orthogonal sub-bands. The sub-band can be indexed by $m \in \mathcal{M} = \{1, 2, \cdots, M\}$. Given high-mobility of vehicles and high-dynamic of CSI, the primary challenge in implementing a vehicle platooning system is to design an robust resource sharing strategy that meets both V2I and V2V link requirements. Furthermore, the randomness in the combination of multiple platoons exacerbates the difficulty of solving the resource allocation problem in vehicle platooning systems.

We assume that there are P platoons in this system, represented by $\mathcal{P} = \{1, 2, \cdots, P\}$. For platoon $\mathcal{P}_i$, its LV is denoted by i, and the set of FVs is denoted as $L_i = \{1, 2, \cdots, N_i\}$, where N_i is the number of FVs in $\mathcal{P}_i$. The power gain of the mth sub-band from the transmitter of LV i to the receiver of FV j is modelled as:

$$g_{i,j}[m] = \alpha_{i,j}\beta_{i,j}[m], \tag{1}$$

where $\alpha_{i,j}$ denote the large-scale fading factor describing primarily transmission path loss, and $\beta_{i,j}[m]$ represent the effect of small-scale fading comprised of signal scattering and Doppler effect on the mth sub-band. We assume $\beta_{i,j}[m]$ is fitted by an exponential function. Similarly, over the mth sub-band, the power gain from LV i to BS e is defined as $g_{i,e}[m]$, the interfering channel from LV i' to BS e is defined as $g_{i',e}[m]$, and the interfering channel from LV i' to FV j is defined as $g_{i',j}[m]$. We can express the signal power from LV i to BS e on the sub-band m by:

$$S_{i,e}[m] = g_{i,e}[m]p_i[m], \tag{2}$$

where $p_i[m]$ is the transmission power of LV i over the allocated sub-band m. The interference power at the receiver of BS e can be express as follow:

$$I_e[m] = \sum_{i' \neq i} g_{i',e}[m]p_{i'}[m]. \tag{3}$$

The signal power of the mth sub-band from LV i to its FV j and the interference power at the FV j over sub-band m are defined as $S_{i,j}[m]$, $I_j[m]$ in the same way, respectively. Combined with equation (2) and (3), the capacity of the V2I links from LV i to BS e can be determined by the Shannon's capacity theorem:

$$c_{i,e}[m] = \frac{W}{M}\log_2\left(1 + \frac{S_{i,e}[m]}{\sigma^2 + I_e[m]}\right), \tag{4}$$

where σ^2 is the noise power. The capacity of the V2V links between LV i and its FV j depends on the V2V link with the worst signal-to-interference-plus-noise ratio (SINR), which is expressed as:

$$c_{i,j}[m] = \min_{j, j \in L_i}\left\{\frac{W}{M}\log_2\left(1 + \frac{S_{i,j}[m]}{\sigma^2 + I_j[m]}\right)\right\}, \tag{5}$$

where $g_{i,j}[m]$ is sub-band power gain from LV i to its follower j on the mth sub-band. Assuming that multiple sub-bands can be allocated for intra-platoon communications, the achievable capacity of the V2V link from the LV i to FV j is denoted as $C_{i,j}$, which can be calculated by summing up the capacities across all allocated sub-bands [37]:

$$C_{i,j} = \sum_{m=1}^{M} \boldsymbol{\kappa}_i[m]c_{i,j}[m]. \tag{6}$$

$\boldsymbol{\kappa}_i$ is a sub-band allocation vector. When allocating the mth sub-band to LV i, the $\boldsymbol{\kappa}_i[m]$ is set to 1, otherwise, $\boldsymbol{\kappa}_i[m]$ is set to 0.

TABLE I: The definition of symbols

Symbol	Description
W	The total bandwidth of spectrum resource
$\mathcal{M}/M/m$	Set/Number/Index of sub-bands
$\mathcal{P}/P/i$	Set/Number/Index of platoons
i	The LV of platoon $\mathcal{P}_i$
$L_i/N_i/j$	Set/Number/Index of the FVs in platoon $\mathcal{P}_i$
ξ	Receiver of BS e or FV j, where $j \in L_i$
$g_{i,\xi}[m]$	Power gain from LV i to ξ on mth sub-band
$g_{i',\xi}[m]$	Power gain from LV $i', i' \neq i$ to ξ on mth sub-band
$S_{i,\xi}[m]$	Signal power from LV i to ξ on mth sub-band
$I_\xi[m]$	Interference power at ξ on mth sub-band
$p_i[m]$	Power of LV i on mth sub-band
$p_i^{\max}$	Max power of LV i
σ^2	Noise power
$c_{i,\xi}[m]$	Capacity from LV i to ξ on mth sub-band
$C_{i,j}$	Capacity of the V2V link from LV i to FV j
A_i^t	AoI of the ith V2I communications at time t
$c_{i,e}^{\min}$	Minimum capacity requirement of V2I link from i to e
η_i^t	The resource sharing switch for LV i at time slot t
ψ_b	Required workload of service b
Δ_T	Sub-band coherence period
δ	Length of a time slot
t	Index of a time slot
τ	Meta policy update interval
ρ_b	Success delivery probability of a service b
$\boldsymbol{\kappa}_i$	Sub-band allocation vector for LV i
$\mathbf{u}/u_i$	Set of meta policies / meta-policy for ego agent i

B. Problem Formulation

The definitions of symbols are listed and explained in Table I. Generally, in a vehicle platooning system, BS is responsible for producing cooperative guidance information with a global

perspective. LVs play a crucial role in collecting and transmitting environmental states to their FVs to maintain the stability of platoons. Additionally, LVs can also provide infotainment content downloaded from the BS to their FVs. Intra-platoon messages can vary in terms of frequencies and quantities, depending on the specific vehicular service requirements. η_i^t is resource sharing switch for LV i at time slot t. It is a binary variable that is used to denote the resource allocation mode of a platoon i at time step t. When allocating sub-bands to intra-platoon links of a platoon i at time step t, η_i^t is set to 1, otherwise, η_i^t is set to 0. The successful delivery of service b for FV j in a time period T is modelled as:

$$\sum_{t=1}^{T} \eta_i^t C_{i,j}^t \geq \psi_b / \Delta_T, \tag{7}$$

where ψ_b denotes the required workload of service b during the sub-band coherence period Δ_T. The success delivery probability of a service workload from LV i to its follower j is defined as

$$\rho_{i,j}^T = \Pr\{\sum_{t=1}^{T}\sum_{m=1}^{M} \eta_i^t \boldsymbol{\kappa}_i^t[m] c_{i,j}^t[m] \geq \psi_b/\Delta_T\}. \tag{8}$$

Road conditions and driving information of other platoons are essential for vehicular decision-making. We assume that each LV is responsible for collecting such information via inter-platoon links. In order to capture the timeliness of information [38], we introduce the concept of AoI. AoI describes the time elapsed since the last successful V2I communication, taking into account the transmission time of packets. Given this regard, the value of AoI for the ith V2I communications at time step t is described as:

$$A_i^t = \begin{cases} \delta, & (1-\eta_i^t)\, c_{i,e}[m] \geq c_{i,e}^{\min}, \\ \delta + A_i^{t-1}, & otherwise, \end{cases} \tag{9}$$

where δ denotes the length of a time slot, and $c_{i,e}^{\min}$ denotes the minimum capacity requirement for the V2I link between LV i and BS e.

Thereafter, the resource optimization problem for vehicle platooning networks can be formulated as a multi-objective function, which is denoted as:

$$\begin{aligned} &\max_{\lambda,\boldsymbol{\kappa},p} \left\{ \rho_{i,j}^T, -\frac{1}{T}\sum_{t=1}^{T} A_i^t, -\frac{1}{T}\sum_{t=1}^{T}\sum_{m=1}^{M} p_i^t[m] \right\}, \\ \textbf{s.t.}\quad & C1: c_{i,e}^t[m] \geqslant c_{i,e}^{\min}, \\ & C2: \lambda_i^t, \boldsymbol{\kappa}_i[m] \in \{0,1\}, \\ & C3: \sum_{m=1}^{M} \boldsymbol{\kappa}_i[m] \leq M, \\ & C4: p_i^t[m] \leq p_i^{\max}. \end{aligned} \tag{10}$$

In this function, the total transmission power of a LV is also minimized to further improve resource efficiency. The constraint $C1$ means the capacity from LV i to BS e is not less than the minimum capacity $c_{i,e}^{\min}$. The constraint $C2$ shows that λ_i^t and $\boldsymbol{\kappa}_i[m]$ take the value of 0 or 1. The constraint $C3$ means that allocated number of sub-bands could not exceed M. The constraint $C4$ means the transmission power of the LV i could not be more than the maximum power $p_i^{\max}$.

IV. Hierarchical Resource Allocation Method

A. Markov Game Transformation

In vehicle platooning networks, acquiring accurate CSI can be challenging due to rapidly changing IoV conditions. When scales of a vehicle platooning system increases, making centralized decisions only at BS can become excessively difficult. MARL proves to be an effective method to deal with this problem. To address the non-stationary problem caused by dynamic characteristics of agents, we propose a two-layer hierarchical resource allocation method called HierNet. In this method, the centralized high layer adopts a global perspective and generates coordination guidance, while the decentralized low layer takes actions to allocate resources for multiple platoons. To model the proposed resource allocation problem as an HRL problem, a double-layer Markov game is formulated, and key components including the definition of agents, states, actions, and the reward function, are described as follows.

Agent: A meta agent is designed to provide global guidance to ego agents. The meta agent deployed on BS collects local observations from LVs and feeds back meta policies via V2I links. In addition, we consider a LV of each platoon as an ego agent. Each ego agent can access to its local observation of vehicle platooning networks. Consequently, in the low level, the Markov game for P platoons can be transformed into a partially observable Markov decision process (POMDP) involving multiple agents, which can be denoted as $\langle \mathcal{O}, \mathcal{U}, \mathcal{A}, \mathcal{R}, \mathcal{G} \rangle$. $\mathcal{O} = \{o_1, o_2, \cdots, o_P\}$ represents the local observations of the environment, which collectively describe vehicle platooning networks. $\mathcal{U} = \{u_1, u_2, \cdots, u_P\}$ represents the meta policy for all platoons, which is generated and distributed by the meta agent. $\mathcal{A} = \{a_1, a_2, \cdots, a_P\}$ denotes the action set of ego agents, $\mathcal{R}$ represents the immediate reward for ego agents, and $\mathcal{G}$ denotes the transition probability distribution over $\mathcal{O}$ and $\mathcal{U}$. In the high level, $\mathcal{U}$ can be regarded as actions of the meta agent. Therefore, the Markov game for meta agent can be denoted as $\left\langle \mathcal{S}, \mathcal{U}, \bar{\mathcal{R}}, \hat{\mathcal{G}} \right\rangle$, where $\bar{\mathcal{R}}$ represents the immediate reward for the meta agent, and $\hat{\mathcal{G}}$ denotes the transition probability distribution of meta agent over $\mathcal{S}$.

State: As mentioned previously, meta agent distributes global perspectives to ego agents. In detail, meta agent utilizes the aggregated local observations from all ego agents as its global observations. At the start of time slot t, LVs gather CSI and compute interferences from other platoons in the previous time step, denoted as I_i^{t-1} for LV i. The value of AoI, which measures the message timeliness from LVs to BS via V2I links, is also included in the state space, denoted as A_i^t. In order to optimize QoS for intra-platoon communications, the remaining time for transmitting a fixed service payload, denoted as T_i^t, and the remaining payload, denoted as B_i^t, are also included in the state space. Thus, the local observation obtained by an LV i at time slot t can be described as $o_i^t = \left\{ g_{i,\xi}^t, I_i^{t-1}, A_i^t, T_i^t, B_i^t \right\}$. $g_{i,\xi}^t$ denotes the power gain of the V2X link that is allocated sub-bands in time step t. We assume that the high layer and the low layer are training in different time scales. For ego agent i, meta policy u_i^τ is

continuously effective until meta agent distributes a new meta policy. In high layer, the observations of LVs are transmitted to the meta agent through V2I links periodically. Subsequently, the meta agent update its knowledge of environment by integrating all local observations of LVs. In summary, the observation of ego agent i at time slot t is denoted as:

$$Z_i^t = \{o_i^t, u_i^t\}, \tag{11}$$

where u_i^t is the most recent meta-policy distributed to ego agent i, and the state of meta agent is represented as:

$$\mathcal{S}^t = \{o_1^t, o_2^t, \cdots, o_P^t\}. \tag{12}$$

Action: The action of the meta agent is a set of meta policies u_i^t, which can be denoted as

$$\mathcal{U}^t = \{u_1^t, u_2^t, \cdots, u_P^t\}. \tag{13}$$

The actions of a LV include making decisions regarding transmission power, resource sharing switch, and sub-band allocation. The resource sharing switch determines whether a sub-band is assigned for a V2I link or a V2V link. The sub-band allocation decision involves assigning one sub-band for a V2I link or multiple sub-bands for a V2V link. The action space of the ego agent for LV i at time slot t can be described as:

$$a_i^t = \left\{p_i^t, \eta_i^t, \boldsymbol{\kappa}_i^t\right\}, \tag{14}$$

where p_i^t and η_i^t indicate the continuous power and resource sharing switch, respectively. $\boldsymbol{\kappa}_i^t$ is the sub-band allocation vector, which decides the sub-bands allocated to a V2X link. We assume that a V2I link is allocated one sub-band but a V2V link can be allocated multiple sub-bands.

Reward: Both AoI of V2I links and QoS of vehicle platooning networks are taken into account in the reward function. The optimization problem is then transformed into a normalized reward function for ego agents, which is given as follows:

$$r_i(t) = x\rho_{i,j}^t + yA_i^t + z\sum_{m=1}^{M} p_i^t[m]. \tag{15}$$

To achieve multi-objective optimization, parameters x, y, and z are designed to ensure that the values of multiple functions fall within the same range.

In order to coordinate the cooperation of ego agents, the reward of meta agent is defined as an average value of rewards obtained by all ego agents in an update interval of τ.

$$\bar{r}(t) = \frac{1}{\tau P}\sum_{i=1}^{P}\sum_{t'=t}^{t+\tau-1} r_i(t'), \tag{16}$$

B. Hierarchical Reinforcement Learning Method

Fig. 2 shows the hierarchical architecture of the proposed HierNet method. HierNet is composed of one meta agent and n ego agents. The meta agent is deployed on BS and the ego agents is deployed on the LV of each platoon. Proximal policy optimization (PPO) [39] is applied to generate meta policy for multi-platoon cooperation, and multi-agent deep deterministic policy gradient (MADDPG) [40] is applied to generate ego policies for resource allocation.

1) PPO for Meta Policy Generation: The goal of reinforcement learning is to maximize the expectation of cumulative reward for each platoon. The key idea of policy gradient is to adjust the parameters of strategies according to the gradient direction of the reward function. Policies are randomly generated, and actions are sampled according to conditional probabilities. The PPO algorithm is a typical policy gradient algorithm. Standard policy gradient methods perform one gradient update per data sample. PPO designs a novel objective function that enables multiple epochs of mini-batch updates. Originated from the trust region policy optimization (TRPO) algorithm [41], PPO introduces a clipped surrogate objective function.

Let π_ϕ denotes meta policy from actor, and V_μ represents a learned state-value function simplified as an advantage-function, in which ϕ and μ are the parameters of the actor network and the critic network, respectively. The clipped surrogate objective function in PPO is defined with the following function:

$$L_{\text{clip}}^{\pi}(\phi) = \mathbb{E}[\min(\lambda_t(\phi)\widehat{A}_t, \text{clip}(\lambda_t(\phi), 1-\epsilon, 1+\epsilon)\widehat{A}_t)], \tag{17}$$

where $\mathbb{E}[\cdot]$ is the expectation of the average value for mini-batch samples in trajectory memory. Here, $\lambda_t(\phi)$ denotes the probability ratio, and ϵ is a hyperparameter for clip fraction. $\lambda_t(\phi)$ is defined as:

$$\lambda_t(\phi) = \frac{\pi_\phi(\mathbf{u}_t|\mathbf{o}_t)}{\pi_{\phi_{\text{old}}}(\mathbf{u}_t|\mathbf{o}_t)}, \tag{18}$$

where $\lambda_t(\phi_{\text{old}}) = 1$ and $\pi_{\phi_{\text{old}}}$ is the meta policy before one update process. The clipping method function $\text{clip}(\lambda_t(\phi), 1-\epsilon, 1+\epsilon)\widehat{A}_t)$ is to avoid excessive modification of the objective value. $\widehat{A}_t$ makes use of generalized advantage estimator with learned $V_\mu(\mathbf{o}_t)$ to compute the advantage function, which is formulated by:

$$\widehat{A}_t \approx \bar{r}(\mathbf{o}, \mathbf{u}) + \gamma V_\mu(o_{t+1}) - V_\mu(o_t). \tag{19}$$

The adjustment of the parameter μ in critic network is through minimizing the loss of squared error, which is defined as follows:

$$L^V(\mu) = \mathbb{E}[(V_\mu(o_t) - y_t)^2], \tag{20}$$

where the target value $y_t = \bar{r}_t + \gamma V_\mu(o_{t+1})$.

The augmentation of the overall objective function applies an entropy bonus $S[\pi_\mu]$ to ensure sufficient exploration. Combining these terms mentioned above, we obtain the following objective, which is maximized in each update:

$$L_\mu = \mathbb{E}\left[L_{\text{clip}}^{\pi}(\phi) - c_1 L^V(\mu) + c_2 S[\pi_\phi](\mathbf{o}_t)\right] \tag{21}$$

where c_1 and c_2 are the coefficients, and $S[\pi_\phi](\mathbf{o}_t)$ denotes the entropy of the meta policy when offered an input state $\mathbf{o}_t$.

2) MADDPG with Meta Policy for Ego Policy Generation: Compared with policy gradient algorithm, deterministic policy gradient (DPG) can better deal with challenging high-dimensional problems. Without requiring the complete exploration of the state space and action space of agents, DPG uses policy $\pi_\theta : o_t \to a_t$ to learn a state-value function:

$$V(o_t, a_t) = \mathbb{E}\left[(r_t|o_t, a_t) + V(o_{t+1}, a_{t+1})\right]. \tag{22}$$

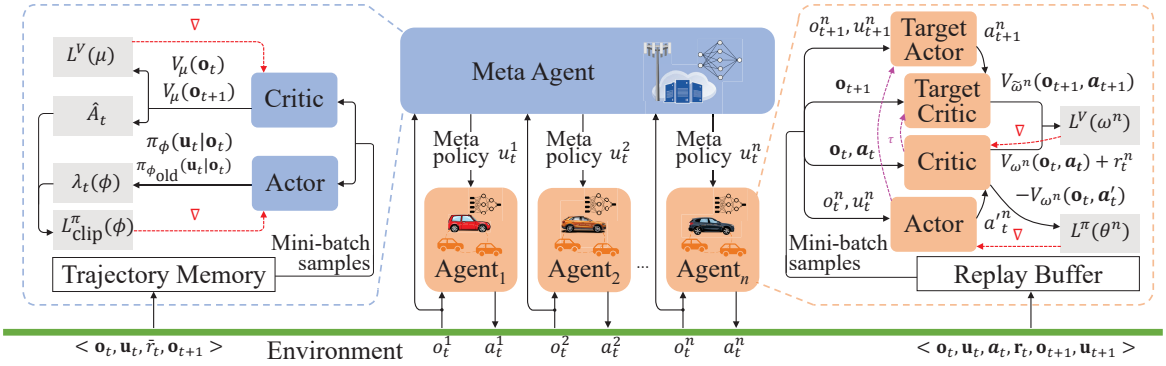

Fig. 2: Hierarchical architecture of HierNet.

Just as deep Q-learning uses deep neural networks (DNN) to estimate Q-learning functions, deep deterministic policy gradient (DDPG) uses DNN to approximate actor networks and critic networks in DPG. DDPG algorithm transforms two neural networks in actor-critic algorithm into four neural networks, using copies of the actor evaluation network and critic evaluation network to update the target actor network and the target critic network, respectively. This is because actor and critic influence each other, and the update of their parameters is also interdependent, which results in convergence problem. Updating the target network by slowly tracking the learned network can improve the stability of learning. DDPG is an off-policy algorithm, an experience replay buffer is used to store training transitions $\{o_t, u_t, a_t, r_t, o_{t+1}, u_{t+1}\}$. The parameter updating for the actor evaluation network is performed by maximizing the follow function:

$$J_\omega(\theta_t) = \mathbb{E}\left[V_\omega(o_t, a_t) \mid a_t = \nu(o_t, u_t)\right], \tag{23}$$

where $\nu(\cdot)$ is the deterministic policy function of the actor denoted as $\pi : \{o_t, u_t\} \rightarrow a_t$. The loss function for the update of the actor evaluation network is defined as:

$$L^\pi(\theta) = -J_\omega(\theta_t). \tag{24}$$

The optimized loss function to update the critic evaluation network in t transition is:

$$L^V(\omega) = \mathbb{E}\left[\left(V_\omega(o_t, a_t) + r_t - \gamma V_\omega(o_t, a_t')\right)^2\right], \tag{25}$$

where a_t' denotes ego policy from the evaluation actor network.

C. The Training Algorithm of HierNet

The training process of the resource allocation method based on HierNet is given in Algorithm 1. In this method, we assume that one episode consists of 100 time steps. The initialization of vehicle platooning environment is performed at the first step. Number of platoons, mobility model for the vehicles and communication model of cellular network are determined (Line 1). Next, a series of ego agents are initialized with network parameters, experience replay buffers, discount rates, and other hyperparameters (Line 2). A meta agent is initialized with network parameters as well (Line 3).

At the beginning of each episode, we initialize the trajectory memory for the meta agent (Line 5). The meta agent collects initial observations and output meta-policies at the start of episode (Line 6-8). We utilize $\hat{t}$ to distinguish it from time step t to emphasize that meta agent generates policies at a larger time scale compared to ego agents. The value of $\hat{t}$ increases with the update number of meta policies (Line 21). Subsequently, at each time step, every ego agent generates exploratory actions based on policy π_{θ^i} using local observations and the most recent meta policy (Line 11-12). The actions are executed by LVs, and then rewards from environment are received (Line 12). In every τ time steps, the meta agent collects the new local observations from LVs, computes average reward of all ego agent over τ time steps, and stores a transition that includes the last global observation, new global observations, reward, and last meta policy into trajectory memory $\hat{D}_k$ (Line 17-19). The meta agent then generates and distributes new meta policies through π_ϕ to support the subsequent decision-making of ego agents (Line 20). The ego agent observes the new environmental state and stores $\{o_t^i, u_t^i, a_t^i, r_t^i, o_{t+1}^i, u_{t+1}^i\}$ in its experience replay buffer D^i (Line 24-25). As mentioned earlier, meta-policy u_{t+1}^i remains unchanged, that is the same as u_t^i if it has not been updated yet. We select mini-batch samples to update the actor and critic networks of ego agents (Line 26). After a certain time interval, the network parameters of the meta agent are updated (Line 30). In this method, the update of the target networks applies the soft update method (Line 29).

The computation complexity of HierNet depends on the state space, the action space and the scale of neural networks in the training process and executing process. In detail, the length of input vectors of agents (a meta agent and multiple ego agents), their action space and the number of agents determine the training parameters of HierNet. In addition, the number of hidden layers in DNN and the number of neurons in different hidden layers determine the training process of HierNet. HierNet is composed of a high-layer meta agent and multiple low-layer ego agents that are training independently. The number of training parameters in π_ϕ is $O(P^2 D N_u)$, where D represents the local observation dimension, and N_u is the length of a meta policy vector for an ego agent. For the critic network V_μ, the number of training parameters is $O(P(D + N_u))$. The number of training parameters for ego agents based on MADDPG also involves respective actor networks and critic networks. The actor network π_θ exhibits a

Algorithm 1 The training process of HierNet.

Initialize vehicle platooning system model
Initialize replay buffer D^i, actor networks π_{θ^i}, $\pi_{\tilde{\theta}^i}$ and critic networks V_{ω^i}, $V_{\tilde{\omega}^i}$ for each ego agent i
Initialize the actor network π_ϕ and the critic network V_μ of the meta agent
for each episode k **do**
  Initialize trajectory memory $\hat{D}_k$
  Meta agent get global state $S_{\hat{t}}$ where $\hat{t} = 0$
  Update platoon locations and large-scale fading
  Initialize meta-policies $\mathbf{u}$ by π_ϕ for all ego agents
  for time step $t = 0, 1, 2, \cdots$ **do**
    for each ego agent i **do**
      Get observation $Z_t^i = \{o_t^i, u_t^i\}$
      Get action a_t^i by $\pi(Z_t^i|\theta^i)$ with exploration
    end for
    All ego agents execute $\boldsymbol{a}_t$ and receive reward $\boldsymbol{r}_t$
    Update small-scale fading
    if time to update meta-policies **then**
      Meta agent gets global observations $\mathbf{o}_{\hat{t}+1}$
      Compute the average reward $\bar{r}_{\hat{t}}$ according to (16)
      Store $\{\mathbf{o}_{\hat{t}}, \mathbf{u}_{\hat{t}}, \bar{r}_{\hat{t}}, \mathbf{o}_{\hat{t}+1}\}$ into $\hat{D}_k$
      Update and distribute meta-policies $\mathbf{u}_{\hat{t}+1}$ by π_ϕ
      Let $\hat{t} = \hat{t} + 1$
    end if
    for each ego agent i **do**
      Get new local observation o_{t+1}^i and meta-policy u_{t+1}^i
      Store $\{o_t^i, u_t^i, a_t^i, r_t^i, o_{t+1}^i, u_{t+1}^i\}$ into D^i
      Update π_{θ^i} and V_{ω^i} with mini-batch from D^i by (24) and (25)
    end for
  end for
  Update target networks $\pi_{\tilde{\theta}^i}$, $V_{\tilde{\omega}^i}$ with soft update method
  Update π_ϕ and V_μ with a mini-batch from $\hat{D}_k$ by (17) and (20)
end for

parameter space of $O(P^2(D+N_u)N_a)$, and the critic network V_ω shows $O(P(D+N_a))$, where N_a denotes the length of an ego action vector. As for neural networks, two full-connected hidden layers of neural networks have 512 and 256 neurons are used in both actor and critic networks.

V. Simulation Results and Discussions

In this section, we first introduce simulation settings and comparison baselines of HierNet, and then present the simulation results and discussions.

A. Simulation Settings

We consider a vehicular network using 2 GHz carrier frequency. The time and frequency resources are divided into multiple resource blocks, each with a bandwidth of 180 kHz. The velocity of platoons is set uniformly distributed between 10 and $15m/s$. The inter-vehicle distance in a platoon is $gap = 10m$. The number of platoons is set from $\#Platoon = 2$ to $\#Platoon = 7$, and the number of FVs in each platoon is $\#FV = 4$. The average size of V2V messages follows a continuous uniform distribution from $W = \mathcal{U}(1, 3)$ to $W = \mathcal{U}(11, 13)$. The meta policy interval is set from 1 to 5. To realize efficient resource utilization, the V2V links between LVs and their FVs and the V2I links between LVs and BS share the same pool of resource blocks. The number of resource blocks is set between $\#RB = 3$ and $\#RB = 8$. The parameter values of communication channel model and vehicle mobility model in this work are set by reference to [42]. In our simulation, the buffer size for experience replay is set to $50,000$. The learning rates are set to 0.0001 for communication and actor modules, 0.001 for critic modules.

To verify the efficiency of HierNet, it is compared with the following MARL baselines: MADDPG [40], MADDPG with QMIX [43] (MADDPG+QMIX), multi-agent twin delayed deep deterministic policy gradient (MATD3) [44], and multi-agent proximal policy optimization method (MAPPO) [39]. In recent works, an MADDPG-based method was proposed to optimize wireless resource allocation for unmanned aerial vehicle (UAV) communication networks [45]. A communication-QMIX method was applied to deal with multi-agent task assignment for mobile crowdsensing in [46]. In [47], transmit power allocation for downlink transmission of UAV-networks has been solved with an MATD3-based method. In addition, as a popular MARL baseline, MAPPO was proposed to solve distributed underwater searching and efficient data collection in [48]. In summary, both of the selected MARL baselines aim to solve resource allocation decision-making problems that are NP-hard in communication networks, which are similar to our formulated resource optimization problem in vehicle platooning networks. In detail, the comparison baselines are introduced as follows.

MADDPG: This method is based on the actor-critic framework. Each agent is equipped with a couple of actor and critic networks. The actor network utilizes the Q-value estimated by the critic to update network parameters, enabling the training of resource allocation strategies that take into account a global perspective.

MADDPG+QMIX: In this method, the actor module of each agent is the same as MADDPG. The critic module is attached with the QMIX network to achieve global optimization fitting, in which QMIX training combines the Q-values of multiple agents.

MATD3: This method adds a set of target critics on the basis of MADDPG. When calculating the target value, the smaller value is taken from the output of the two target critics to suppress the overestimation of the Q-value. And a delayed update strategy is adopted, i.e. the critic network performs multiple updates before updating the actor. In addition, when calculating the target value, a perturbation is added to the next action to make the value assessment more accurate.

MAPPO: This method also builds upon the actor-critic framework and incorporates the concepts of centralized training and distributed execution. Unlike MADDPG, MAPPO is an on-policy algorithm that requires limiting the step size of model updates and utilizing importance sampling to enhance

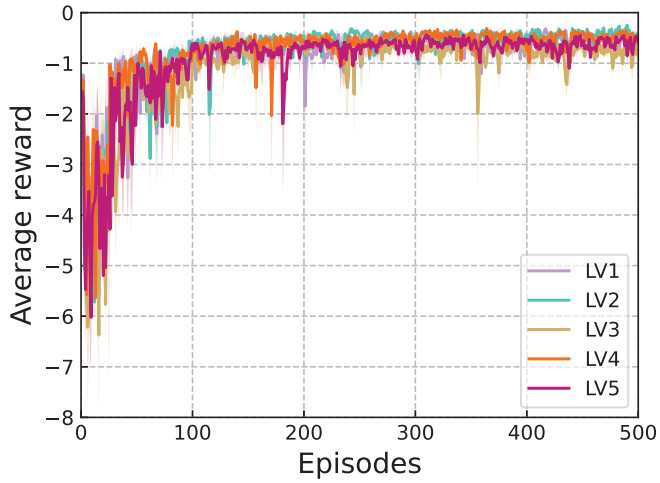

Fig. 3: Average training reward for each agent when $\#Platoon = 5$.

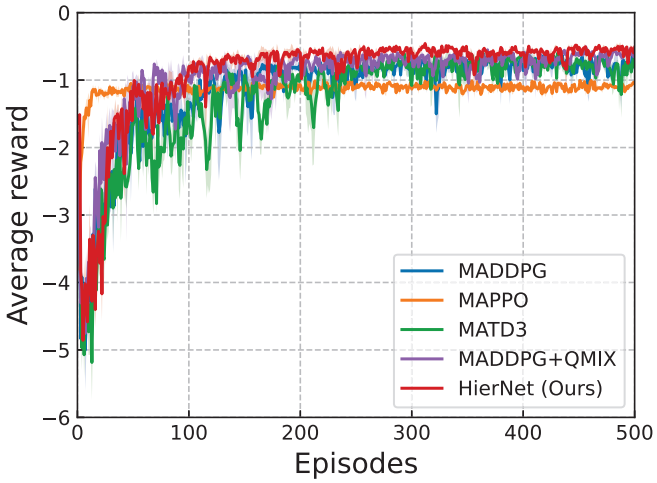

Fig. 4: Average training reward when $\#Platoon = 5$.

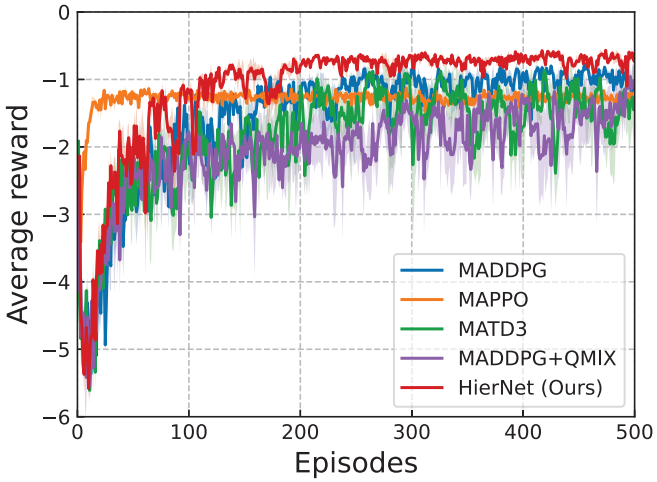

Fig. 5: Average training reward when $\#Platoon = 6$.

training efficiency.

B. Results and Discussions

Fig. 3 shows the changes of average reward for each agent with the increasing of training episodes when $\#Platoon = 5$. As the training continues, the average reward gradually stabilizes and reaches convergence. In the early stages of training, agents focus on exploration, leading to significant fluctuations in average rewards. As the number of episodes increases, the average rewards of increases rapidly and approaches a convergence value within 100 episodes. During the subsequent training process, the fluctuation of average rewards gradually decreases. Fig. 3 also shows that the performance of each agent are balanced and the average reward is simultaneously maximized. In other words, each agent successfully learns the optimized resource allocation strategies.

Fig. 4, Fig. 5 and Fig. 6 show changes of average rewards with different methods when $\#Platoon = 5$, $\#Platoon = 6$ and $\#Platoon = 7$, respectively. With different platoon numbers, the average reward in different methods shows a similar trend. In the early stages of training, due to the limited training data, decision-making policies are comparable to that of a random method. The increase of platoon numbers raises the probability of conflicts in resource allocation, resulting in a decrease in reward values. Additionally, intense channel competition imposes higher demands on method performance, requiring more training episodes, which is represented as a decrease in convergence speed in these figures. Among these comparison methods, MAPPO shows the fastest convergence speed. It completed convergence in about 30 episodes, but the convergence average reward is performing less than optimally. The training process of MADDPG+QMIX method is less stable when $\#Platoon = 6$ and $\#Platoon = 7$, showing fluctuations in large degree. The performance of the proposed method in terms of convergence speed is similar to methods based on MADDPG, but shows a quite improvement in the convergence average reward. It proves that our method can reach to a better optimization. Although a global critic is used to learn from global states and strategies, each agent still knows nothing about the situations of other agents during execution. Agents need to infer global strategies under a dynamic environment only by local observation. In the proposed method, high-layer agent provides timely and explicit global coordination information to each agents. This results in a better performance in convergence speed and convergence value.

Fig. 7 shows the changes of average reward with the increase of V2V messages. During the training phase, we maintained a fixed V2V message size. In the testing phase, we modified the V2V message size to evaluate generalization abilities when transmitting V2V messages of varying sizes. As V2V message size increases, the average reward of each method decreases to varying extents. With larger message sizes, even at the same transmission rate, it requires more

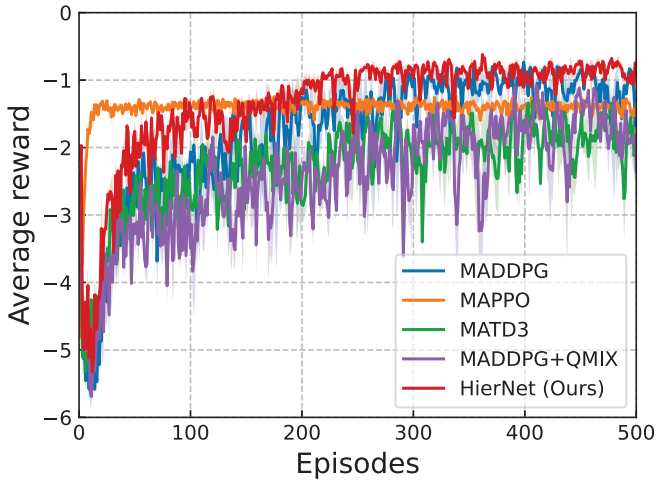

Fig. 6: Average training reward when $\#Platoon = 7$.

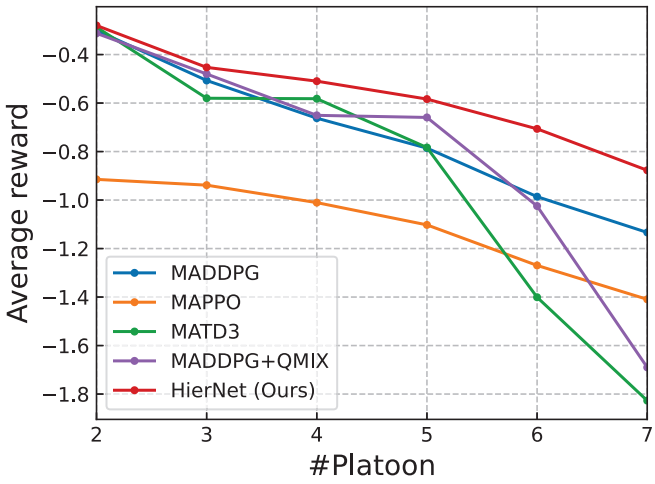

Fig. 8: Average reward *w.r.t.* Number of platoons in different methods.

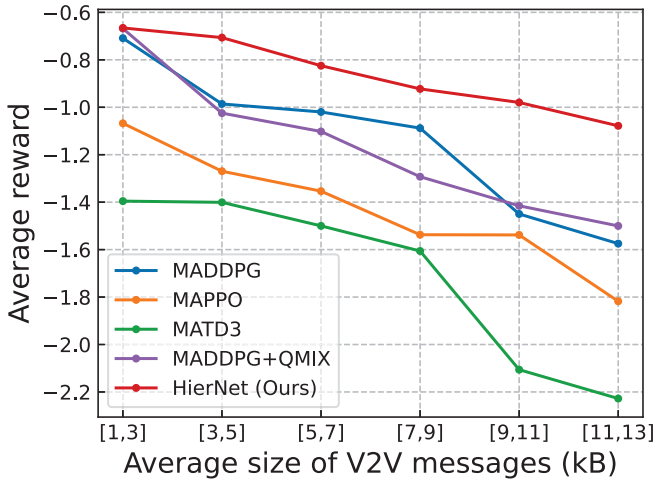

Fig. 7: Average reward *w.r.t.* Average size of V2V messages in different methods.

time to fully transmit the V2V message. Consequently, the probability of successfully transmitting the V2V message within the time budget diminishes. Moreover, larger message sizes result in increased channel occupation time, intensifying competitions for network resources and leading to greater interference between different V2V sub-bands. Consequently, the transmission rate decreases when the size of V2V messages increases. In comparison to other methods, our method demonstrates achieves better average reward with a narrower decreasing range.

Fig. 8 shows the average reward with the increase of platoons compared with baseline methods. The expansion in the number of platoons results in a greater number of V2V and V2I connections, consequently increasing the demand for network resources from multiple platoons. Through increasing the number of platoons, we comprehensively test the robustness of resource allocation methods under higher network loads. It is evident that for a small number of platoons, methods other than MAPPO exhibit similar performance. However, when we increase the number of platoons to 6, both the MATD3 and MADDPG+QMIX methods experience a significant drop in average rewards and lag behind MADDPG. Although the average reward of MAPPO is relatively lower, it exhibits a smaller decrease in average reward as the number of platoons increases. From the figures, the proposed method demonstrates a more stable average reward with different number of platoons and maintains better performance.

Fig. 9 shows the average reward with the increase of resource blocks compared with the baseline methods. Resource blocks are the most limited resource in this problem, showing a crucial impact on the performance of resource allocation methods. A larger number of resource blocks can alleviate competitions of agents and reduce interference between different V2X communications. This, in turn, increases sub-band capacities, leading to an increase in the average rewards. As shown in this figure, there is a significant increase in the average reward of each method as the number of resource blocks increases. Additionally, it is evident that our method outperforms other comparison methods with different resource block numbers.

Fig. 10 shows the average reward of HierNet with the increase of resource blocks when $\#Platoon = 2$, $\#Platoon = 4$ and $\#Platoon = 6$. As the number of resource blocks increases, the average rewards show an increasing trend. This trend becomes more significant when there are a larger number of platoons involved. Additionally, in situations where the availability of resources is relatively sufficient, the level of conflict and interference in the sub-band decreases. Consequently, the average reward approaches its maximum value within the given environment. As a result, the average reward exhibits less variation when the number of platoons changes.

Fig. 11 shows the average reward of HierNet with the

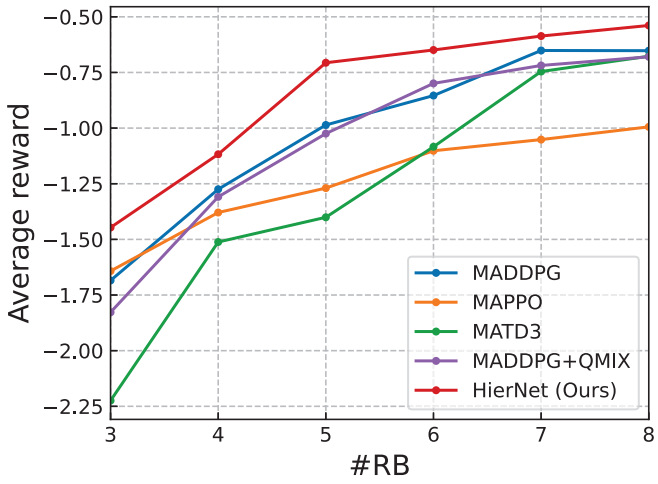

Fig. 9: Average reward *w.r.t.* Number of resource blocks in different methods.

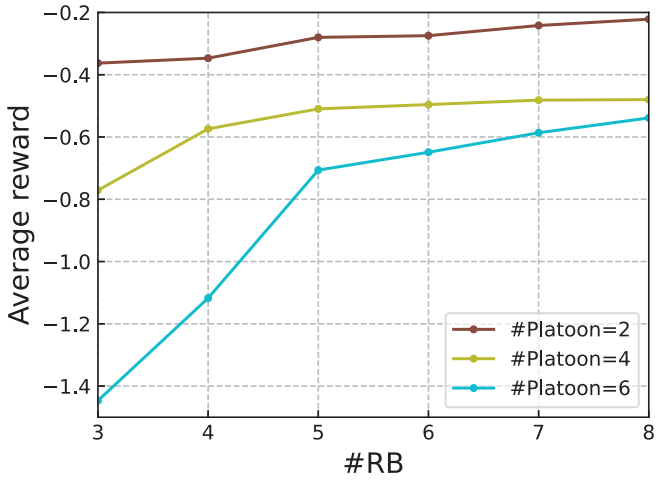

Fig. 10: Average reward *w.r.t.* Number of resource blocks when $\#Platoon = 2$, $\#Platoon = 4$ and $\#Platoon = 6$.

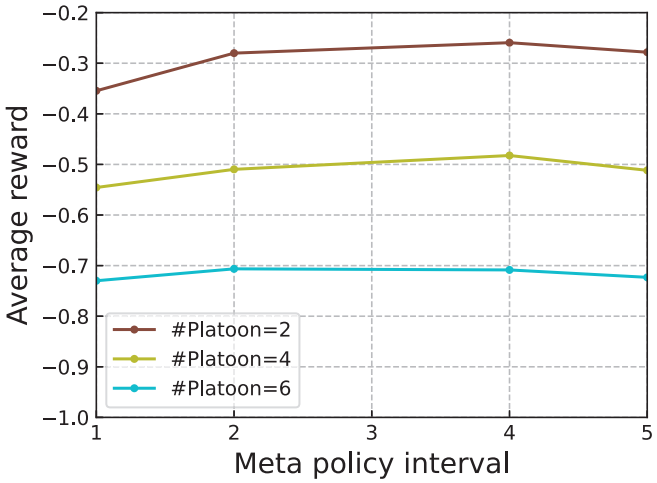

Fig. 11: Average reward *w.r.t.* Size of meta policy intervals when $\#Platoon = 2$, $\#Platoon = 4$ and $\#Platoon = 6$.

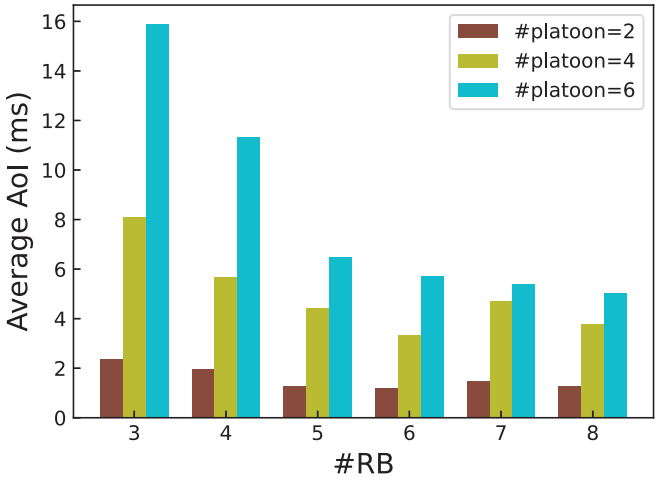

Fig. 12: Average AoI *w.r.t.* Number of resource blocks.

increase of meta policy intervals when $\#Platoon = 2$, $\#Platoon = 4$ and $\#Platoon = 6$. In HierNet, high-level meta agent execute actions on longer time scales. In other words, the meta-policy issued by the high-level meta agent will have a lasting effect for a certain duration, and intervals of updating the meta-policy is predetermined at the beginning of the experiment as a hyperparameter. The figure indicates that, for varying platoon numbers, the average rewards initially increases and then decreases with the increase of meta-policy intervals. However, the overall change is not significant. Considering that frequent meta-policy updates result in higher communication overhead, a larger meta-policy update interval can be employed in practical applications.

Fig. 12 shows the average AoI of HierNet with the increase of resource blocks when $\#Platoon = 2$, $\#Platoon = 4$ and $\#Platoon = 6$. When the number of platoons is small, competition for resources is not intense in the environment, resulting in no significant impact on AoI value with an increase of resource blocks. However, as the number of platoons increases, the influence of number of resource blocks on AoI becomes more apparent. Specifically, when there are 6 platoons and only 3 resource blocks, the figure clearly demonstrates resource scarcity, leading to a significantly higher AoI compared to other scenarios. In addition, as the number of resource blocks increases, the AoI decreases significantly, and becomes more stable when the number of resource blocks reaches 5 in different platoon scenarios.

Fig. 13 shows the average transmission success rate of HierNet with the increase of resource blocks when $\#Platoon = 2$, $\#Platoon = 4$ and $\#Platoon = 6$. Our method shows relatively constant with various platoon numbers and resource

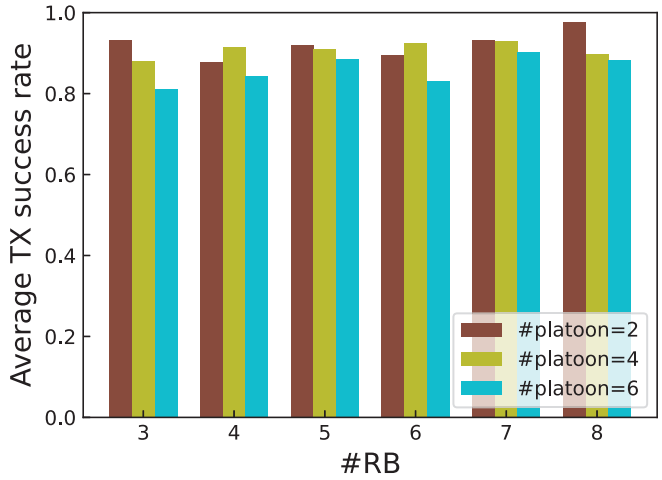

Fig. 13: Average transmission success rate *w.r.t.* Number of resource blocks.

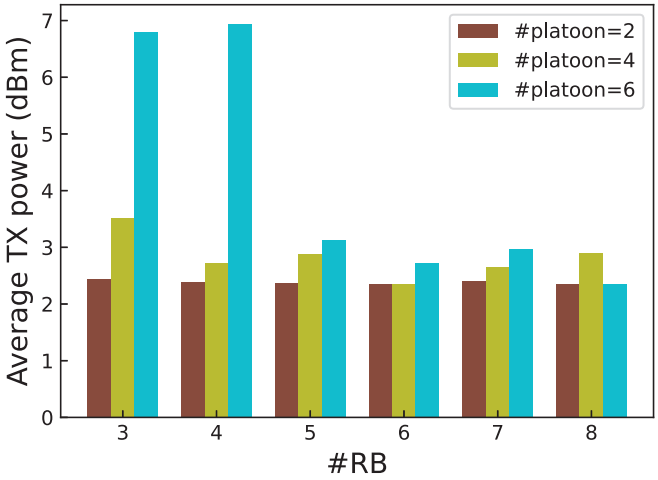

Fig. 14: Average transmission power *w.r.t.* Number of resource blocks.

blocks. In Fig. 12, it is evident that the AoI consistently decreases when the number of resource blocks increases. This observation implies that the agent prioritizes V2V transmission, allocating limited resources to ensure efficient transmission of cooperative awareness messages. In addition, large number of platoons brings competitions for resources, which results in lowest transmission successful rate when $\#Platoon = 6$.

Fig. 14 shows the average transmission power of HierNet with the increase of resource blocks when $\#Platoon = 2$, $\#Platoon = 4$ and $\#Platoon = 6$. As presented in this figure, we can find the resource shortage situation caused by the low number of resource blocks and the high number of platoons, and the sufficient resource situation with stable average transmission power. In the scenario of sufficient resources, which is depicted in most instances in the figure, cooperation mode gains the upper hand, resulting in scattered sub-band selection and minimal interference within sub-bands. It is not necessary for agents to resort to higher power levels to satisfy their communication requirements, leading to a situation of relatively lower power levels. However, when resources are scarce, agents are compelled to increase their transmission power in order to send more data within the limited sub-band resources. Combining this analysis with Fig. 13, it can be inferred that agents employ higher transmission power to achieve a V2V transmission success rate that is comparable to the rate achieved when resources are sufficient. This demonstrates that our method can adapt to high-load scenarios with tight resources, and is robust to the dynamic number of resource blocks and platoons.

VI. Conclusion

In this paper, a hierarchical system framework for resource allocation in vehicle platooning has been proposed, in which BS acts as a central coordination agent and LVs act as ego resource decision agents. The multi-objective resource optimization problem is formulated for multi-platoon networks, in which intra-platoon communications and inter-platoon communications are both optimized to improve safety of vehicle platooning service. The HRL method called HierNet is proposed to solve this multi-objective problem based on the proposed hierarchical system framework. Finally, simulations have demonstrated the effectiveness of HierNet. The dynamic change of update intervals for meta policy could be a future research direction.

References

[1] Z. Sun, G. Sun, Y. Liu, J. Wang, and D. Cao, "BARGAIN-MATCH: A game theoretical approach for resource allocation and task offloading in vehicular edge computing networks," *IEEE Transactions on Mobile Computing*, vol. 23, no. 2, pp. 1655–1673, 2024.

[2] T. Xiao, C. Chen, Q. Pei, Z. Jiang, and S. Xu, "SFO: An adaptive task scheduling based on incentive fleet formation and metrizable resource orchestration for autonomous vehicle platooning," *IEEE Transactions on Mobile Computing*, pp. 1–18, 2023.

[3] J. Yang, D. Chu, J. Yin, D. Pi, J. Wang, and L. Lu, "Distributed model predictive control for heterogeneous platoon with leading human-driven vehicle acceleration prediction," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–16, 2023.

[4] Q. Wen, D. Liu, J. Wang, and S. Baldi, "An adaptive longitudinal platooning design based on concurrent learning," *IEEE Control Systems Letters*, vol. 8, pp. 303–308, 2024.

[5] X. Gu, J. Peng, L. Cai, W. Liu, X. Zhang, and Z. Huang, "Resource reservation coordination for vehicle platooning in C-V2X networks," *IEEE Transactions on Wireless Communications*, pp. 1–14, 2023, dOI: 10.1109/TWC.2023.3326852.

[6] L. Wang, H. Liang, and D. Zhao, "Deep reinforcement learning-based computation offloading and power allocation within dynamic platoon network," *IEEE Internet of Things Journal*, vol. 11, no. 6, pp. 10 500–10 512, 2024.

[7] Y. Yuan, G. Zheng, K.-K. Wong, and K. B. Letaief, "Meta-reinforcement learning based resource allocation for dynamic V2X communications," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 9, pp. 8964–8977, 2021.

[8] K. Yu, H. Zhou, Z. Tang, X. Shen, and F. Hou, "Deep reinforcement learning-based RAN slicing for UL/DL decoupled cellular V2X," *IEEE Transactions on Wireless Communications*, vol. 21, no. 5, pp. 3523–3535, 2022.

[9] B. Y. Yacheur, T. Ahmed, and M. Mosbah, "Efficient DRL-Based selection strategy in hybrid vehicular networks," *IEEE Transactions on Network and Service Management*, vol. 20, no. 3, pp. 2400–2411, 2023.

[10] T. Şahin, R. Khalili, M. Boban, and A. Wolisz, "Reinforcement learning scheduler for vehicle-to-vehicle communications outside coverage," in *2018 IEEE Vehicular Networking Conference*, 2018, pp. 1–8.

[11] M. Parvini, M. R. Javan, N. Mokari, B. Abbasi, and E. A. Jorswieck, "AoI-aware resource allocation for platoon-based C-V2X networks via multi-agent multi-task reinforcement learning," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 8, pp. 9880–9896, 2023.

[12] T. Zhang, C. Xu, Y. Lian, H. Tian, J. Kang, X. Kuang, and D. Niyato, "When moving target defense meets attack prediction in digital twins: A convolutional and hierarchical reinforcement learning approach," *IEEE Journal on Selected Areas in Communications*, vol. 41, no. 10, pp. 3293–3305, 2023.

[13] N. Zhang, J. Yan, C. Hu, Q. Sun, L. Yang, D. W. Gao, J. M. Guerrero, and Y. Li, "Price-matching-based regional energy market with hierarchical reinforcement learning algorithm," *IEEE Transactions on Industrial Informatics*, pp. 1–12, 2024.

[14] D. Kim, M. R. Castellanos, and R. W. Heath, "Joint band assignment and beam management using hierarchical reinforcement learning for multi-band communication," *IEEE Transactions on Vehicular Technology*, pp. 1–16, 2024.

[15] M. Yu, T. Yang, C. Li, Y. Jin, and Y. Xu, "Mitigating bus bunching via hierarchical multi-agent reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–18, 2024.

[16] D. Jia, K. Lu, J. Wang, X. Zhang, and X. Shen, "A survey on platoon-based vehicular cyber-physical systems," *IEEE Communications Surveys & Tutorials*, vol. 18, no. 1, pp. 263–284, 2016.

[17] C. Wu, Z. Cai, Y. He, and X. Lu, "A review of vehicle group intelligence in a connected environment," *IEEE Transactions on Intelligent Vehicles*, pp. 1–25, 2023.

[18] L. Lei, T. Liu, K. Zheng, and L. Hanzo, "Deep reinforcement learning aided platoon control relying on V2X information," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 6, pp. 5811–5826, 2022.

[19] Y. Zhang, N. Cheng, Y. Dai, Z. Yin, W. Quan, Y. Zhou, and N. Zhang, "Resource scheduling for eMBB and URLLC multiplexing in NOMA-Based VANETs: A dual time-scale approach," *IEEE Transactions on Vehicular Technology*, pp. 1–14, 2023.

[20] Z. Sun, Y. Liu, J. Wang, G. Li, C. Anil, K. Li, X. Guo, G. Sun, D. Tian, and D. Cao, "Applications of game theory in vehicular networks: A survey," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 4, pp. 2660–2710, 2021.

[21] K. Li, W. Ni, E. Tovar, and M. Guizani, "Optimal rate-adaptive data dissemination in vehicular platoons," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 10, pp. 4241–4251, 2020.

[22] L. Cao, S. Roy, and H. Yin, "Resource allocation in 5G platoon communication: Modeling, analysis and optimization," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 4, pp. 5035–5048, 2023.

[23] G. Chai, W. Wu, Q. Yang, M. Qin, Y. Wu, and F. R. Yu, "Platoon partition and resource allocation for ultra-reliable V2X networks," *IEEE Transactions on Vehicular Technology*, pp. 1–16, 2023.

[24] Z. Dong, X. Zhu, Y. Jiang, H. Zeng, Z. Wei, F.-C. Zheng, and K.-C. Leung, "Dynamic manager selection assisted resource allocation in URLLC with finite block length for 5G-V2X platoons," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 11, pp. 11 336–11 350, 2022.

[25] W. Gao, C. Wu, L. Zhong, and K.-L. A. Yau, "Communication resources management based on spectrum sensing for vehicle platooning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 2, pp. 2251–2264, 2023.

[26] N. Cheng, F. Lyu, W. Quan, C. Zhou, H. He, W. Shi, and X. Shen, "Space/aerial-assisted computing offloading for IoT applications: A learning-based approach," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 5, pp. 1117–1129, 2019.

[27] N. Hammami and K. K. Nguyen, "Multi-agent actor-critic for cooperative resource allocation in vehicular networks," in *2022 14th IFIP Wireless and Mobile Networking Conference*, 2022, pp. 93–100.

[28] X. Li, L. Lu, W. Ni, A. Jamalipour, D. Zhang, and H. Du, "Federated multi-agent deep reinforcement learning for resource allocation of vehicle-to-vehicle communications," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 8, pp. 8810–8824, 2022.

[29] P. Xiang, H. Shan, M. Wang, Z. Xiang, and Z. Zhu, "Multi-agent RL enables decentralized spectrum access in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10 750–10 762, 2021.

[30] Y. Xu, K. Zhu, H. Xu, and J. Ji, "Deep reinforcement learning for multi-objective resource allocation in multi-platoon cooperative vehicular networks," *IEEE Transactions on Wireless Communications*, vol. 22, no. 9, pp. 6185–6198, 2023.

[31] J. M. I Parella, O. T. Ajayi, and Y. Cheng, "Adaptive messaging based on the age of information in VANETs," in *2022 IEEE Global Communications Conference*, 2022, pp. 1235–1240.

[32] S. Pateria, B. Subagdja, A.-h. Tan, and C. Quek, "Hierarchical reinforcement learning: A comprehensive survey," *ACM Comput. Surv.*, vol. 54, no. 5, 2021.

[33] F. Ye, J. Wang, J. Li, P. Zhu, D. Wang, and X. You, "Intelligent hierarchical network slicing based on dynamic multi-connectivity in cell-free distributed massive MIMO systems," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 9, pp. 11 855–11 870, 2023.

[34] Y. Geng, E. Liu, R. Wang, and Y. Liu, "Hierarchical reinforcement learning for relay selection and power optimization in two-hop cooperative relay network," *IEEE Transactions on Communications*, vol. 70, no. 1, pp. 171–184, 2022.

[35] Y. Huang and K.-W. Chin, "A hierarchical deep learning approach for optimizing CCA threshold and transmit power in Wi-Fi networks," *IEEE Transactions on Cognitive Communications and Networking*, vol. 9, no. 5, pp. 1296–1307, 2023.

[36] S. Gong, L. Cui, B. Gu, B. Lyu, D. T. Hoang, and D. Niyato, "Hierarchical deep reinforcement learning for age-of-information minimization in IRS-aided and wireless-powered wireless networks," *IEEE Transactions on Wireless Communications*, pp. 1–1, 2023.

[37] S. Jangsher, H. Zhou, V. O. K. Li, and K.-C. Leung, "Joint allocation of resource blocks, power, and energy-harvesting relays in cellular networks," *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 3, pp. 482–495, 2015.

[38] X. Cao, J. Wang, Y. Cheng, and J. Jin, "Optimal sleep scheduling for energy-efficient AoI optimization in industrial Internet of things," *IEEE Internet of Things Journal*, vol. 10, no. 11, pp. 9662–9674, 2023.

[39] C. Yu, A. Velu, E. Vinitsky, Y. Wang, and Y. Wu, "The surprising effectiveness of MAPPO in cooperative, multi-agent games," *arXiv preprint arXiv:2103.01955*, 2021.

[40] R. Lowe, Y. Wu, A. Tamar, and J. Harb, "Multi-agent actor-critic for mixed cooperative-competitive environments," *arXiv preprint arXiv:1706.02275*, 2017.

[41] J. Schulman, S. Levine, P. Moritz, M. I. Jordan, and P. Abbeel, "Trust region policy optimization," *Computer Science*, pp. 1889–1897, 2015.

[42] M. Parvini, M. R. Javan, N. Mokari, B. Abbasi, and E. A. Jorswieck, "AoI-aware resource allocation for platoon-based C-V2X networks via multi-agent multi-task reinforcement learning," *IEEE Transactions on Vehicular Technology*, pp. 1–17, 2023.

[43] T. Rashid, M. Samvelyan, C. D. Witt, G. Farquhar, J. Foerster, and S. Whiteson, "QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning," *arXiv preprint arXiv:1803.11485*, 2018.

[44] J. Ackermann, V. Gabler, T. Osa, and M. Sugiyama, "Reducing overestimation bias in multi-agent domains using double centralized critics," *arXiv preprint arXiv:1910.01465*, 2019.

[45] R. Ding, F. Zhou, Q. Wu, and D. W. K. Ng, "From external interaction to internal inference: An intelligent learning framework for spectrum sharing and UAV trajectory optimization," *IEEE Transactions on Wireless Communications*, 2024.

[46] C. Xu and W. Song, "Decentralized task assignment for mobile crowdsensing with multi-agent deep reinforcement learning," *IEEE Internet of Things Journal*, vol. 10, no. 18, pp. 16 564–16 578, 2023.

[47] S. Zhou, Y. Cheng, X. Lei, Q. Peng, J. Wang, and S. Li, "Resource allocation in UAV-assisted networks: A clustering-aided reinforcement learning approach," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 11, pp. 12 088–12 103, 2022.

[48] B. Jiang, J. Du, C. Jiang, Z. Han, and M. Debbah, "Underwater searching and multiround data collection via AUV swarms: An energy-efficient AoI-aware MAPPO approach," *IEEE Internet of Things Journal*, vol. 11, no. 7, pp. 12 768–12 782, 2024.

Xiaoyuan Fu received the Ph.D. degree in computer science and technology from Beijing University of Posts and Telecommunications (BUPT) in 2019. In 2018, she visited the Department of Systems and Computer Engineering, Carleton University, Ottawa, Canada. She is currently working as an associate research fellow at the State Key Laboratory of Networking and Switching Technology, BUPT. Her current research interests include vehicular networks, virtual distributed ledger technology (vDLT), and deep reinforcement learning.

Quan Yuan received his Ph.D. degree in computer science and technology from BUPT, China, in 2018. He is currently an associate professor at the State Key Laboratory of Networking and Switching Technology, BUPT. His current research interest includes mobile computing, crowdsensing, and vehicular networks.

Guiyang Luo is currently an Assistant Professor with the Computer Science Department, Beijing University of Posts and Telecommunications, Beijing, China (BUPT). From 2020 to 2022, he was a Postdoctoral Fellow with Computer Science Department, BUPT. His research interests include multi-agent systems and machine-type communications.

Nan Cheng received the Ph.D. degree from the Department of Electrical and Computer Engineering, University of Waterloo in 2016, and B.E. degree and the M.S. degree from the Department of Electronics and Information Engineering, Tongji University, Shanghai, China, in 2009 and 2012, respectively. He worked as a post-doctoral fellow with the Department of Electrical and Computer Engineering, University of Toronto, from 2017 to 2019. He is currently a Full Professor with State Key Lab. of ISN and with School of Telecommunications Engineering, Xidian University, Shaanxi, China. He has published over 70 journal papers in IEEE Transactions and other top journals. He serves/served as guest editors for several journals and serves as associate editors for IEEE Transactions on Vehicular Technology, IEEE Open Journal of Vehicular Technology, and Peer-to-Peer Networking and Applications. His current research focuses on B5G/6G, space-air-ground integrated network, big data in vehicular networks, and self-driving system. His research interests also include applying AI techniques for future networks.

Yang Li received the B.Eng in intelligent science and technology from North China Electric Power University in 2023. He is currently pursuing the Master degree at the State Key Laboratory of Networking and Switching Technology, BUPT. His research interests include the areas of intelligent transportation, swarm intelligence, and deep reinforcement learning.

Junfei Wang is currently pursuing the bachelor's degree at school of computer science (National Pilot Software Engineering School), BUPT. His research interests include Internet of vehicles and deep reinforcement learning.

Jianxin Liao received the Ph.D. degree from the University of Electronics Science and Technology of China, Chengdu, China, in 1996. He is currently the Dean of the Network Intelligence Research Center and a Full Professor with the State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications. He has authored or coauthored 100 research papers and several books. He has won a number of prizes in China for his research achievements, which include the Premiers Award of Distinguished Young Scientists from National Natural Science Foundation of China in 2005, and the specially invited Professor of the "Yangtse River Scholar Award Program" by the Ministry of Education in 2009. His main research interests include cloud computing, mobile intelligent networks, service network intelligent, networking architectures and protocols, and multimedia communication.