Multi-Agent Reinforcement Learning for Energy-Efficiency Edge Association in Internet of Vehicles

Yiyu Tao*§, Yan Lin*†, Yijin Zhang*§, Feng Shu‡ and Jun Li*
*School of Electronic and Optical Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
†National Mobile Communications Research Laboratory, Southeast University, Nanjing 210096, China
‡School of Information and Communication Engineering, Hainan University, Haikou 570228, China
§State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China
Email:{tyy_1103, yanlin}@njust.edu.cn, yijin.zhang@gmail.com, {shufeng, jun.li}@njust.edu.cn

GLOBECOM 2022 - 2022 IEEE Global Communications Conference | 978-1-6654-3540-6/22/$31.00 ©2022 IEEE | DOI: 10.1109/GLOBECOM48099.2022.10001200

***Abstract*—In this paper, we investigate the energy-efficiency (EE) problem in edge association for heterogeneous Internet of Vehicles (IoV), when the dynamic environmental information can not be known in advance. Aiming to maximize the long-term tradeoff between EE and handover (HO) overhead, we propose a cooperative multi-agent edge association solution, where vehicular user equipments (VUEs) make decisions cooperatively relying on their local observations under centralized training. Specifically, we first construct a multi-agent partially observable Markov decision process (MA-POMDP) problem and decompose the system value function into the local value functions for implicit individual learning. Next, through sharing learning experience and approximating the global state, each VUE is able to obtain its own optimal/suboptimal policy given its local observations and historical information. Simulation results show that the proposed solution outperforms the non-cooperative counterpart and other baselines in terms of improving EE with the most appropriate number of HOs.**

***Index Terms*—Edge association, multi-agent, handover, reinforcement learning, energy efficiency, Internet of Vehicles.**

I. Introduction

With the rapid evolution of intelligent transportation systems, Internet of Vehicles (IoV) has become an emerging paradigm for the next-generation networks [1]. In cellular vehicular communication, vehicle-to-vehicle (V2V) communication ensures real-time interactions of short-range information, while vehicle-to-infrastructure (V2I) communication makes a wide range of communication available. To combine both, heterogeneous IoV is widely deployed and becomes one of the promising solutions to support ultra-reliable and low latency communications (URLLC) services [2].

This work was supported in part by the National Natural Science Foundation of China under Grants 62001225, 62071236, 62071234, in part by the Natural Science Foundation of Jiangsu Province under Grant BK20190454, in part by the Fundamental Research Funds for the Central Universities under Grant 30920021127, in part by the open research fund of National Mobile Communications Research Laboratory, Southeast University under Grant 2022D07, in part by the Open Research Fund of State Key Laboratory of Integrated Services Networks, Xidian University, under Grant ISN22-14. (*Corresponding author: Yan Lin.*)

In heterogeneous IoV, the deployment of road side units (RSUs) tends to be more dense for guaranteeing ultra-high data rate requirement [3], and the high mobility of vehicles enables the network topology highly dynamic. As a result, the frequent handover (HO) problem emerges which may impose poor user experience and heavy signal loads [4]. Thus, the edge association problem when considering HO overhead has become a significant concern for IoV.

With the requirement of energy conservation and carbon reduction, energy efficiency (EE) problem in edge association for IoV has attracted growing research interests, where EE is relevant to the ratio between transmission rate and transmit power. Although the existing works can achieve high EE to some extent [5] [6], Sun and Khezrian *et al.* ignore the unknown and uncertainty characteristics of the vehicular mobility and the time-varying channel states. To solve this issue, some intelligent edge association solutions have been proposed with the aid of reinforcement learning (RL), which aims to maximize the long-term transmission data rate while reducing the overall HO overhead. For instance, Khan *et al.* of [7] proposed an intelligent edge association scheme for millimeter wave IoV, where RSUs make decisions for maximizing the long-term per-user average transmission rate relying on global environmental information. As a further step, our previous work of [8] investigated the intelligent edge association problem for heterogeneous IoV under user-centric clustering framework. Through interacting with the environment, the base station agent utilizes global environmental information for making the edge association decision when the long-term tradeoff between transmission rate and HO overhead is considered. To further concern on the EE performance of edge association in IoV, Pervej *et. al* of [9] proposed a distributed multi-agent resource allocation scheme to achieve the best EE in the presence of mandatory HO. However, it still requires all vehicular user equipments (VUEs) to observe almost complete environmental information for decision-making, which imposes excessive communication overhead for information sharing.

Against this background, our paper aims for investigating

the edge association problem in heterogeneous IoV to strike an EE-HO tradeoff, provided that VUEs can only observe local environmental information. With the aid of cooperative multi-agent RL, all VUEs are able to make their own edge association decisions cooperatively to optimize the system reward. The main contributions of this paper are summarized as follows:

- We formulate an intelligent edge association problem for striking a long-term EE-HO tradeoff. It is modeled as a multi-agent partially observable Markov decision process (MA-POMDP) within a finite time interval, where VUEs can only observe local environmental information.
- We propose a cooperative multi-agent RL-based edge association solution by utilizing the decomposition of the system value function under centralized training framework. Through sharing learning experience and approximating the global state, VUEs can obtain their own optimal/suboptimal policy relying on their local observations and historical information.
- Simulation results show that the proposed cooperative solution outperforms the non-cooperative counterpart and other baselines in terms of improving EE with the most appropriate number of HOs, in the context of local observations and low communication overhead requirements.

II. System Model and Problem Formulation

In this section, we first introduce the system model and then formulate our optimization problem as a MA-POMDP problem.

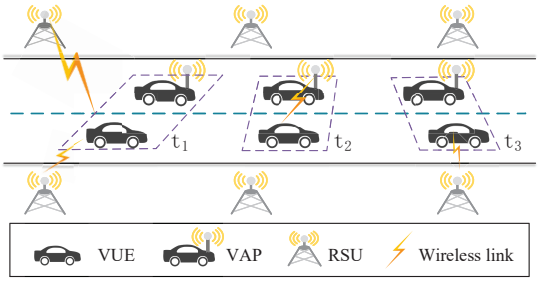


Fig. 1. An example of the one-way multi-lane downlink V2X network.

A. System Model

As shown in Fig. 1., we consider a one-way multi-lane downlink vehicle-to-everything (V2X) network system, which includes U VUEs, R RSUs and V vehicular access points (VAPs). RSUs are uniformly distributed on both sides of the road. VAPs and VUEs travel along the road without conflicting with each other by independently following a Gauss-Markov mobility model [10]. The sets of RSU, VUE, and VAP are denoted by $\mathcal{R} = \{1, ..., R\}$, $\mathcal{U} = \{1, ..., U\}$, and $\mathcal{V} = \{1, ..., V\}$, respectively. The time horizon is divided into discrete time slots (TSs), denoted by $\mathcal{T} = \{1, ..., T\}$.

Both RSUs and VAPs, referred to as access points (APs), can serve VUEs within a limited coverage radius. For improving connectivity, each VUE can be served by at most $\bar{S}$ RSUs via cooperative V2I communication or connects to one adjacent VAP via V2V communication. Moreover, each RSU can connect to multiple VUEs simultaneously, while each VAP can only access one VUE for transmission.

We assume that the inter-user interference has been mitigated through allocating orthogonal resource blocks. Let $\mathcal{G}_{i,t}$ be the AP group that connects to VUE i in TS t. Let $h_{i\leftarrow g,t}$ denote the channel gain from AP g to VUE i in TS t, including the small-scale fading and the path loss. Let $\bar{A}$ denote the minimum achievable transmission rate of VUEs. Let p_g be the transmit power of AP g and $\bar{P}$ be the minimum transmit power. Accordingly, the achievable V2I/V2V downlink transmission rate of VUE i in TS t is represented as

$$A_{i,t} = log_2(1 + \frac{\sum_{g\in\mathcal{G}_{i,t}} p_g h_{i\leftarrow g,t}}{\sigma^2}), \tag{1}$$

where σ^2 is the additive Gaussian white noise variance. Additionally, the EE of VUE i in TS t is expressed as the ratio between its transmission rate and its associated AP group's total transmit power, given by

$$E_{i,t} = \frac{A_{i,t}}{\sum_{g\in\mathcal{G}_{i,t}} p_g}. \tag{2}$$

Then, due to the mobility of VUEs and VAPs, HOs will occur when VUE changes its associated AP group for transmission, including three cases: 1) HOs between RSUs, 2) HOs between VAPs, and 3) HOs between RSUs and VAPs. The HO overhead, measured by the number of HOs $H_{i,t}$, can be obtained by

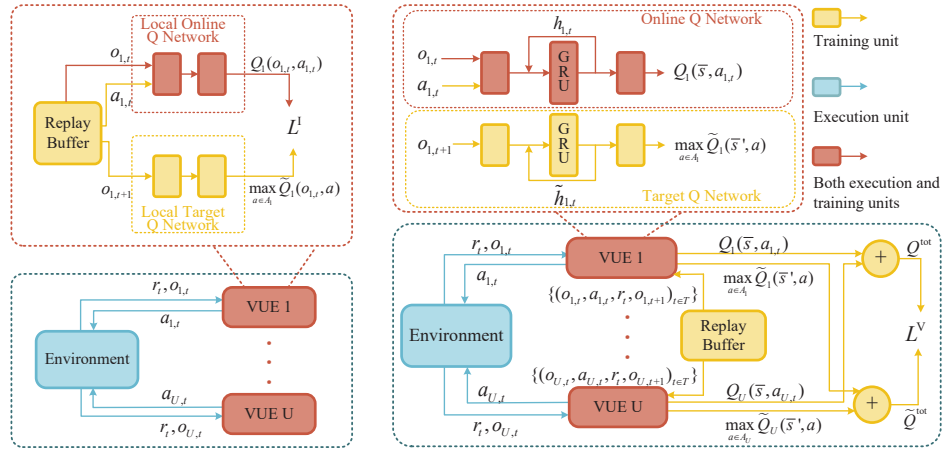
$$H_{i,t} = \max\{|\mathcal{G}_{i,t\text{-}1}|, |\mathcal{G}_{i,t}|\} - |\mathcal{G}_{i,t\text{-}1} \cap \mathcal{G}_{i,t}|. \tag{3}$$

Herein, the first term refers to the maximum number of HOs supported during TS t and TS t-1, while the second term is the corresponding number of the same associated RSUs.

B. Problem Formulation

The goal of this paper is to make the edge association decision for maximizing the long-term system EE with the most appropriate number of HOs. Particularly, VUEs can only observe partial environmental information in the face of environmental uncertainty, e.g. its own location and adjacent APs' location. Thus, the commonly-used conventional convex-optimization techniques, Markov decision process (MDP) methods and single-agent RL algorithms cannot be applied in solving our problem. As such, the EE edge association problem can be constituted a MA-POMDP problem, where each VUE plays the role of an agent for selecting its associated AP group based on its local observations.

To be specific, the MA-POMDP can be defined by a tuple of $\{\mathcal{S}, \mathcal{A}, T, \mathcal{O}, Z, E, \gamma\}$, where $\mathcal{S}$ is the global state space, $\mathcal{A}$ is the joint action space and $\mathcal{O}$ is the joint observation space. Since the state transition probability T and the observation transition probability Z are difficult to obtain, we consider the VUEs can find the optimal/suboptimal policy by interacting with the environment without modeling the environment beforehand. Moreover, E is the instant system reward, and γ is the discount factor that measures the impact of the future rewards.



(A) IDQN-based non-cooperative multi-agent solution. (B) VDN-based cooperative multi-agent solution.

Fig. 2. Comparison of multi-agent edge association solution for heterogeneous IoV.

1) Observation: Assuming that the real-time locations of both the vehicles and the RSUs can be obtained by the sensors and positioning technology, the observation of VUE i in TS t is denoted as

$$\boldsymbol{o}_{i,t} = [\boldsymbol{d}^{\mathrm{u}}_{i,t}, \boldsymbol{d}^{\mathrm{r}}_{i,t}, \boldsymbol{d}^{\mathrm{v}}_{t}, \boldsymbol{d}_{i,t\text{-}1}], \tag{4}$$

where $\boldsymbol{d}^{\mathrm{u}}_{i,t}$ is the position vector of VUE i in TS t, $\boldsymbol{d}^{\mathrm{r}}_{i,t}$ is the position vector of $\bar{S}$ nearest RSUs that connects to VUE i in TS t, $\boldsymbol{d}^{\mathrm{v}}_{t}$ is the position vector of all VAPs in TS t and $\boldsymbol{d}_{i,t\text{-}1}$ is the position vector of the AP group associated with VUE i in TS t-1. Let $\mathcal{O}_i = \{\boldsymbol{o}_{i,t}\}$ denote the observation set of the i-th VUE, and then the joint observation space is denoted by $\mathcal{O} = \mathcal{O}_1 \times \cdots \times \mathcal{O}_U$.

2) Action: The action of VUE i in TS t is denoted as

$$\boldsymbol{a}_{i,t} = [\boldsymbol{c}^{\mathrm{r}}_{i,t}, \boldsymbol{c}^{\mathrm{v}}_{i,t}], \tag{5}$$

where $\boldsymbol{c}^{\mathrm{r}}_{i,t} = [c^1_{i,t}, ..., c^{\bar{S}}_{i,t}]$ and $\boldsymbol{c}^{\mathrm{v}}_{i,t} = [c^1_{i,t}, ..., c^V_{i,t}]$ denote the association status of $\bar{S}$ nearest RSUs and of all VAPs for VUE i in TS t, respectively. More explicitly, if AP g is selected by VUE i in TS t, we have $c^g_{i,t} = 1$, otherwise $c^g_{i,t} = 0$. Let $\mathcal{A}_i = \{\boldsymbol{a}_{i,t}\}$ denote the action set of the i-th VUE, and then the joint action space is denoted by $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_U$.

3) Reward: We craft the system reward in TS t as the sum of all VUEs' reward in terms of the normalized tradeoff between the HO overhead and the EE, denoted by

$$r_t = \sum_{i=1}^{U} \left[\beta E_{i,t} \bar{P}/\bar{A} - (1-\beta) H_{i,t}/\bar{S}\right], \tag{6}$$

where $\beta \in (0, 1)$ is the weight parameter of the EE. Explicitly, when β is larger, the EE part contributes more and the HO part contributes less.

Thus, with the goal of maximizing the long-term EE-HO tradeoff, the optimization objective of our MA-POMDP problem becomes maximizing the expectation of discounted future cumulative reward, represented as

$$\pi = \arg\max \ \mathbb{E}\Big[\sum_{q=0}^{\infty} \gamma^q r_{t+q}\Big]. \tag{7}$$

III. Multi-agent EE Edge Association Solution: Non-cooperative versus Cooperative

To solve the above model-free MA-POMDP problem, multi-agent RL can be employed by enabling each VUE as an agent for maximizing the system reward. In this section, we will first introduce the independent deep Q network (IDQN) based non-cooperative multi-agent solution. Then, we present the proposed value decomposition network (VDN) based cooperative multi-agent solution [11], where VUEs share their learning experience under centralized training so that they can make their own decisions cooperatively relying on local observations and historical information.

A. IDQN-based Non-Cooperative Multi-agent Solution

In the IDQN-based non-cooperative multi-agent solution, each VUE has to make decisions independently and complete the training process using the single-agent deep Q network (DQN) algorithm [12]. The optimization goal becomes seeking the optimal policy with the maximum Q value, i.e. the action-value function, which is the expectation of discounted future cumulative reward after taking joint action $\boldsymbol{a}$ under global state $\boldsymbol{s}$ with policy π, denoted by

$$Q(\boldsymbol{s}, \boldsymbol{a}) = \mathbb{E}_{\pi}\Big[\sum_{q=0}^{\infty} \gamma^q r_{t+q} | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}\Big]. \tag{8}$$

The framework of the IDQN-based non-cooperative multi-agent edge association solution is illustrated in Fig. 2(A). In order to improve the stability of training, all VUEs are equipped with their own local replay buffer and a pair of neural networks, namely local online Q network and local target Q network. More explicitly, the local online Q network is used to obtain the actual Q value, while the local target Q network yields the target Q value for training.

Next, let us introduce the workflow from the perspective of VUE i. Firstly, VUE i obtains its observation $\boldsymbol{o}_{i,t}$ from the environment in TS t, and then gets $Q_i(\boldsymbol{o}_{i,t}, \boldsymbol{a})_{\boldsymbol{a}\in\mathcal{A}_i}$ from its

Algorithm 1 Training Process for the VDN-based Cooperative Multi-agent Solution

Initialize the online Q network $Q_i(\boldsymbol{o}_i, \boldsymbol{a}_i|\theta_i)$, the target Q network $\widetilde{Q}_i(\boldsymbol{o}_i, \boldsymbol{a}_i|\widetilde{\theta}_i)$ and the learning parameters.
for each episode $j = 1, 2, ..., J$ **do**
  Initialize $\{\boldsymbol{h}_{i,1}\}_{i\in\mathcal{U}}$ and $\{\widetilde{\boldsymbol{h}}_{i,1}\}_{i\in\mathcal{U}}$.
  for $t = 1, ..., T$ **do**
    for VUE $i = 1, ..., U$ **do**
      Observe $\boldsymbol{o}_{i,t}$, and the online Q network yields $Q_i(\bar{\boldsymbol{s}}, \boldsymbol{a})_{\boldsymbol{a}\in\mathcal{A}_i}$.
      Choose action $\boldsymbol{a}_{i,t}$ according to eq. (9).
      Update exploration probability ϵ.
    end for
    Execute action $\{\boldsymbol{a}_{i,t}\}_{i\in\mathcal{U}}$ and each VUE receives system reward r_t.
  end for
  Store experience $\{(\boldsymbol{o}_{i,t}, \boldsymbol{a}_{i,t}, r_t, \boldsymbol{o}_{i,t+1})_{t\in\mathcal{T}}\}_{i\in\mathcal{U}}$ in shared replay buffer $\mathcal{P}$.
  Sample a mini-batch of experiences from $\mathcal{P}$.
  Obtain $Q_i(\bar{\boldsymbol{s}}, \boldsymbol{a}_{i,t})$ and $\widetilde{Q}_i(\bar{\boldsymbol{s}}', \boldsymbol{a})_{\boldsymbol{a}\in\mathcal{A}_i}$ from the online Q network and the target Q network, respectively.
  Calculate Q^{tot} and $\widetilde{Q}^{\text{tot}}$ according to eq. (14) and (17).
  Update the online Q network according to eq. (15).
  Update the target Q network every T_u times.
end for

local online Q network. Second, VUE i gets its current action $\boldsymbol{a}_{i,t}$ according to the ϵ-greedy strategy, denoted by

$$\boldsymbol{a}_{i,t} = \begin{cases} \arg\max_{\boldsymbol{a}\in\mathcal{A}_i} Q_i(\boldsymbol{o}_{i,t}, \boldsymbol{a}), & \text{with probability } 1-\epsilon, \\ \text{random action}, & \text{with probability } \epsilon. \end{cases} \tag{9}$$

Note that, ϵ can gradually decrease from 1 to 0 over time for better exploration. Third, VUE i receives the system reward r_t after executing action $\boldsymbol{a}_{i,t}$, and then obtains the new observation $\boldsymbol{o}_{i,t+1}$. Finally, VUE i stores the experience of TS t with a tuple of $(\boldsymbol{o}_{i,t}, \boldsymbol{a}_{i,t}, r_t, \boldsymbol{o}_{i,t+1})$ into its local replay buffer.

When training, VUE i samples the set $\mathcal{M}_i$ which contains M samples from its local replay buffer. Then, these samples are fed into VUE i's local online Q network and local target Q network respectively, yielding $Q_i(\boldsymbol{o}_{i,t}, \boldsymbol{a}_{i,t})$ and $\widetilde{Q}_i(\boldsymbol{o}_{i,t+1}, \boldsymbol{a})_{\boldsymbol{a}\in\mathcal{A}_i}$. Finally, the local Q online network is updated by minimizing the loss function $\mathcal{L}^{\text{I}}$, given by

$$\mathcal{L}^{\text{I}} = \frac{1}{M} \sum_{m\in\mathcal{M}_i} \left[r_t + \gamma \max_{\boldsymbol{a}\in\mathcal{A}_i} \widetilde{Q}_i(\boldsymbol{o}_{i,t+1}, \boldsymbol{a}) - Q_i(\boldsymbol{o}_{i,t}, \boldsymbol{a}_{i,t})\right]^2. \tag{10}$$

Correspondingly, the local target Q network of VUE i is updated by coping the local online Q network every T_u times of training.

After training, each VUE is able to obtain its individual policy once its own observation is obtained. However, the agents in the IDQN-based non-cooperative multi-agent solution can only receive the system reward, rather than any shared environmental information. Therefore, when the system reward fluctuates, the agents are hard to identify whether the changes are related to themselves, and then the policy may fall into the dilemma of local optimum.

B. VDN-based Cooperative Multi-agent Solution

In this subsection, we propose a cooperative multi-agent edge association solution with the aid of the VDN algorithm, as shown in Fig. 2(B). Considering the fact that our system reward is the sum of each agent's reward, the system Q value can be decomposed into a set of local Q values for each agent, which is the basic principle of the VDN algorithm. As a result, the centralized training with decentralized execution (CTDE) framework can be employed in our solution, where a central controller is responsible for learning while each agent makes its own decisions relying on its local Q value for execution, rather than the system Q value.

1) Value Decomposition: As aforementioned, the system reward obtained after executing joint action $\boldsymbol{a}$ under global state $\boldsymbol{s}$ is designed as the sum of each VUE's reward, given by $r_t = \sum_{i=1}^{U} r_{i,t}$, where $r_{i,t}$ denotes the reward of VUE i in TS t. Then, we have

$$\begin{aligned} Q(\boldsymbol{s}, \boldsymbol{a}) &= \mathbb{E}_\pi\Big[\sum_{q=0}^{\infty} \gamma^q r_{t+q} | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}\Big] \\ &= \mathbb{E}_\pi\Big[\sum_{q=0}^{\infty} \gamma^q \sum_{i=1}^{U} r_{i,t+q} | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}\Big] \\ &= \sum_{i=1}^{U} \mathbb{E}_\pi\Big[\sum_{q=0}^{\infty} \gamma^q r_{i,t+q} | \boldsymbol{s}_t = \boldsymbol{s}, \boldsymbol{a}_t = \boldsymbol{a}\Big] = \sum_{i=1}^{U} Q_i(\boldsymbol{s}, \boldsymbol{a}), \end{aligned} \tag{11}$$

where $Q_i(\boldsymbol{s}, \boldsymbol{a})$ is the local Q value of VUE i after taking joint action $\boldsymbol{a}$ under global state $\boldsymbol{s}$ with policy π.

2) Shared Online and Target Q Networks: All VUEs share an online Q network and a target Q network for cooperative decision. Different with the local Q networks established in the IDQN, the shared online/target Q networks are composed of the fully connected layers and the gate recurrent unit (GRU) that is used for collecting historical information with the hidden state $\boldsymbol{h}$ inherited from the previous TS. As such, by integrating the historical information carried in $\boldsymbol{h}$ with the local observation $\boldsymbol{o}$ with the aid of GRU, each VUE can approximate the global state $\boldsymbol{s}$ to a certain extent, thereby the negative effects caused by partial observability can be mitigated. Let $\bar{\boldsymbol{s}}$ be the approximation of global state $\boldsymbol{s}$, and eq. (11) can be further denoted by

$$Q(\boldsymbol{s}, \boldsymbol{a}) = \sum_{i=1}^{U} Q_i(\boldsymbol{s}, \boldsymbol{a}) \approx \sum_{i=1}^{U} Q_i(\bar{\boldsymbol{s}}, \boldsymbol{a}). \tag{12}$$

This implies that the system Q value can be approximated as the sum of local Q values across all VUEs.

According to eq. (12), the optimal joint policy by maximizing $Q(\boldsymbol{s}, \boldsymbol{a})$ can be obtained by maximizing $Q_i(\bar{\boldsymbol{s}}, \boldsymbol{a})$ for each VUE $i \in \mathcal{U}$, namely

$$\arg\max_{\boldsymbol{a}\in\mathcal{A}} Q(\boldsymbol{s}, \boldsymbol{a}) = \begin{pmatrix} \arg\max_{\boldsymbol{a}\in\mathcal{A}_1} Q_1(\bar{\boldsymbol{s}}, \boldsymbol{a}) \\ \vdots \\ \arg\max_{\boldsymbol{a}\in\mathcal{A}_U} Q_U(\bar{\boldsymbol{s}}, \boldsymbol{a}) \end{pmatrix}. \tag{13}$$

Note that $Q(\boldsymbol{s}, \boldsymbol{a})$ is maximal only if $Q_i(\bar{\boldsymbol{s}}, \boldsymbol{a})$ for each VUE $i \in \mathcal{U}$ are maximal simultaneously. Accordingly, the VUEs can learn the policy cooperatively by optimizing their own Q values with the shared Q networks. When executing, each VUE makes decisions relying on its local observations without information sharing, which imposes the reduction of communication overhead.

3) Shared Experience Replay: To support the centralized training, a shared replay buffer is established for storing the experiences of all VUEs which are used in training process. When all VUEs have left off the road, the experiences are stored in strict accordance with the training steps in order to better approximate the global state [13].

4) CTDE Framework: Although all VUEs are able to make decisions under local observations according to eq. (9) which is similar to the IDQN-based non-cooperative multi-agent solution, the Q-network is shared among all VUEs for obtaining local Q values. During training, the system Q value Q^{tot} can be obtained by

$$Q^{\text{tot}} = \sum_{i=1}^{U} Q_i(\bar{\boldsymbol{s}}, \boldsymbol{a}_{i,t}), \tag{14}$$

where $Q_i(\bar{\boldsymbol{s}}, \boldsymbol{a}_{i,t})$ is the local Q value of VUE i obtained from the online Q network in TS t.

Moreover, the online Q network updates by minimizing the loss function $\mathcal{L}^{\text{V}}$, denoted by

$$\mathcal{L}^{\text{V}} = \frac{1}{N} \sum_{n \in \mathcal{N}} l_n, \tag{15}$$

where $\mathcal{N}$ is the sample set which contains N samples. Herein, we use the Huber loss function l_n with the goal of enhancing the robustness of training and mitigating the impact of the outliers of the loss, that is

$$l_n = \begin{cases} \frac{1}{2}(y - Q^{\text{tot}})^2, & \text{if } |y - Q^{\text{tot}}| \le \delta, \\ \delta |y - Q^{\text{tot}}| - \frac{1}{2}\delta^2, & \text{otherwise}, \end{cases} \tag{16}$$

where the hyperparameter δ is used to control the sensitivity to outliers, and $y = r_t + \gamma \widetilde{Q}^{\text{tot}}$. Let $\bar{\boldsymbol{s}}'$ be the approximation of global state in TS t+1, and the $\widetilde{Q}^{\text{tot}}$ is obtained by

$$\widetilde{Q}^{\text{tot}} = \sum_{i=1}^{U} \max_{\boldsymbol{a} \in \mathcal{A}_i} \widetilde{Q}_i(\bar{\boldsymbol{s}}', \boldsymbol{a}), \tag{17}$$

where $\widetilde{Q}_i(\bar{\boldsymbol{s}}', \boldsymbol{a})$ is the local target Q value of VUE i obtained from the target Q network.

Through the above training process, $Q_i(\bar{\boldsymbol{s}}, \boldsymbol{a}_{i,t})$ can be learned implicitly below the surface of Q^{tot}, which assists VUE i to identify its own contribution in the system reward. As such, VUEs are able to make decisions relying on their local observations, and the optimal joint action is constituted by the optimal actions of all VUEs. For more details of the training process, please refer to Algorithm 1.

TABLE I
SIMULATIONS PARAMETERS.

System Parameters	**Value**
Number of RSUs R	30
Number of VUEs U	5
Number of VAPs V	5
Path loss	[14]
Maximum number of RSUs associated to VUE $\bar{S}$	4
Transmit power of RSUs P_r	32(dBm)
Transmit power of VAPs P_v	30(dBm)
Weighting parameter β	0.6
Minimum data rate constraint $\bar{A}$	15(b/s/Hz)
Minimum transmission power constraint $\bar{P}$	30(dBm)
Algorithm Parameters	**Value**
Learning rate	0.00002 (0.0005 in IDQN)
Buffer capacity	5000 (10000 in IDQN)
Discount factor γ	0.8
Target network update frequency T_u	200
Huber loss parameter δ	0.1
Hidden state units of GRU	64
Size of mini-batch N	32
Activation function	ReLU
Optimizer	Adam

IV. Simulation Results

In simulation settings, we set the road having a length of 1 km and a width of 7.5 m. The limited coverage radius of RSUs and VAPs are set as 200 m and 50 m, respectively. For the vehicular mobility parameters please refer to [8]. The noise power density and the carrier bandwidth are set as -174 dBm/Hz and 180 kHz, respectively. We adopt the small scale fading as the Rayleigh fading with unit variance. The other simulation parameters are listed in Table I.

For comparison, we consider the IDQN-based non-cooperative multi-agent solution and the following two non-intelligent baselines: 1) Full-connection (FC): each VUE connects to the closest $\bar{S}$ RSUs; 2) Received signal strength (RSS): each VUE connects to the RSU/VAP which has the strongest received signal power.

Fig. 3. shows the cumulative reward, the EE per VUE and the number of HOs per VUE versus the number of RSUs R. First, from the right upper subfigure, we can see that the number of HOs of the two non-intelligent solutions increases naively with the increased number of available RSUs, whilst that of both intelligent solutions fluctuates within a certain range. This phenomenon reflects the effectiveness of multi-agent framework in reducing the HO frequency. Second, from the lower right subfigure, it can be shown that the EE trends for all the solutions are increased as R increases, since the increased number of RSUs enables VUEs connect to more closer RSUs, thus significantly improving the transmission rate, which contributes to the EE. Additionally, it is clear that our proposed solution outperforms the non-cooperative counterpart and other baselines in terms of the EE owing to the cooperative multi-agent framework. Finally, the left

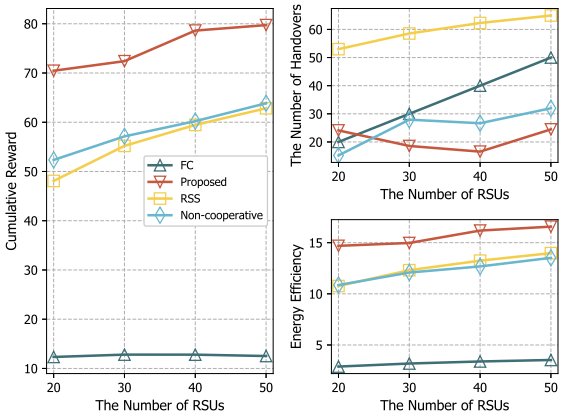


Fig. 3. Cumulative reward (left), the number of HOs (upper right) and EE (lower right) versus the number of RSUs.

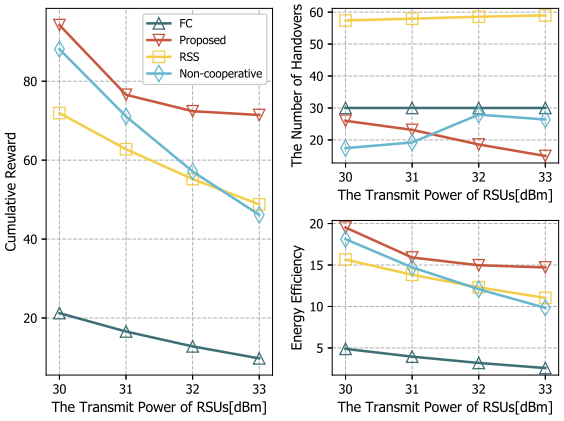


Fig. 4. Cumulative reward (left), the number of HOs (upper right) and EE (lower right) versus the transmit power of RSU.

subfigure shows that all the cumulative reward curves are increased along with the increased number of RSUs except the FC solution, and our proposed solution exhibits a significant improvement in terms of cumulative reward, which verifies the efficiency of cooperative multi-agent framework again.

Fig. 4. investigates the performance versus the transmit power of RSU P_r. First of all, we observe from the upper right subfigure, the two non-intelligent solutions are immune to the increased transmit power of RSUs since their policies remain unchanged even though the transmit power of RSUs changes. By contrast, the two intelligent solutions may adjust the edge association policy for achieving a higher EE, thus their HO frequency fluctuates dynamically. Second, the lower right subfigure shows that the EE of all solutions is reduced along with P_r increases. This is because the increase of the transmission rate is much lower than the increase of P_r. As a result, the cumulative reward is also decreased with the increase of the transmit power of RSUs, as shown in the left subfigure. Finally, observe from the left and the lower right subfigures, the proposed solution achieves both higher cumulative reward and higher EE compared with the baselines as expected. In particular, when P_r exceeds 32 dBm, the performance gap becomes more significant. This implicitly highlights the adaptability of our proposed solution in improving EE with the most appropriate number of HOs through the cooperative learning.

V. Conclusion

A cooperative multi-agent edge association solution was proposed for heterogeneous IoV in the face of unknown dynamics. Aiming for optimizing the EE-HO tradeoff, this solution enables each VUE to make its own edge-association decisions cooperatively relying on local observations under centralized training by sharing learning experience and utilizing historical information. Numerical simulations have shown that the proposed solution is superior to the IDQN-based non-cooperative multi-agent solution and other non-intelligent baselines in terms of EE in a variety of scenarios, with the most appropriate number of HOs. In future works, we will consider the joint edge-association and resource allocation optimization problem, as well as the integration of communication and computation.

References

[1] X. You *et al.*, "Towards 6G wireless communication networks: Vision, enabling technologies, and new paradigm shifts," *Sci. China Inf. Sci.*, vol. 64, no. 1, pp. 1–74, 2021.

[2] K. Zheng, Q. Zheng, P. Chatzimisios, W. Xiang, and Y. Zhou, "Heterogeneous vehicular networking: A survey on architecture, challenges, and solutions," *IEEE Commun. Surveys & Tuts.*, vol. 17, no. 4, pp. 2377–2396, 2015.

[3] S. Chen, F. Qin, B. Hu, X. Li, and Z. Chen, "User-centric ultra-dense networks for 5G: challenges, methodologies, and directions," *IEEE Wireless Commun.*, vol. 23, no. 2, pp. 78–85, 2016.

[4] E. Gures, I. Shayea, A. Alhammadi, M. Ergen, and H. Mohamad, "A comprehensive survey on mobility management in 5G heterogeneous networks: Architectures, challenges and solutions," *IEEE Access*, vol. 8, pp. 195 883–195 913, 2020.

[5] P. Sun, N. AlJeri, and A. Boukerche, "An energy-efficient proactive handover scheme for vehicular networks based on passive RSU detection," *IEEE Trans. Sustain. Comput.*, vol. 5, no. 1, pp. 37–47, 2020.

[6] A. Khezrian, T. D. Todd, G. Karakostas, and M. Azimifar, "Energy-efficient scheduling in green vehicular infrastructure with multiple roadside units," *IEEE Trans. Veh. Technol.*, vol. 64, no. 5, pp. 1942–1957, 2015.

[7] H. Khan, A. Elgabli, S. Samarakoon, M. Bennis, and C. S. Hong, "Reinforcement learning-based vehicle-cell association algorithm for highly mobile millimeter wave communication," *IEEE Trans. Cognit. Commun. Netw.*, vol. 5, no. 4, pp. 1073–1085, 2019.

[8] Y. Lin, Z. Zhang, Y. Huang, J. Li, F. Shu, and L. Hanzo, "Heterogeneous user-centric cluster migration improves the connectivity-handover trade-off in vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 16 027–16 043, 2020.

[9] M. F. Pervej and S.-C. Lin, "Eco-vehicular edge networks for connected transportation: A distributed multi-agent reinforcement learning approach," in *Proc. IEEE VTC2020-Fall*, 2020, pp. 1–7.

[10] S. Batabyal and P. Bhaumik, "Mobility models, traces and impact of mobility on opportunistic routing algorithms: A survey," *IEEE Commun. Surveys & Tuts.*, vol. 17, no. 3, pp. 1679–1707, 2015.

[11] P. Sunehag *et al.*, "Value-decomposition networks for cooperative multi-agent learning based on team reward," in *Proc. Int. Conf. Auton. Agents. Multi-Agent Syst. (AAMAS)*, Stockholm, Sweden, 2018, pp. 2085–2087.

[12] Y. Wang, H. Liu, W. Zheng, Y. Xia, Y. Li, P. Chen, K. Guo, and H. Xie, "Multi-objective workflow scheduling with deep-q-network-based multi-agent reinforcement learning," *IEEE Access*, vol. 7, pp. 39 974–39 982, 2019.

[13] M. Hausknecht and P. Stone, "Deep recurrent Q-learning for partially observable MDPs," in *Proc. AAAI Fall Symp. Series*, 2015, pp. 29–37.

[14] 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, 2019.