World Electric Vehicle Journal

MDPI

Article

Multi-Cell Cooperative Resource Allocation and Performance Evaluation for Roadside-Assisted Automated Driving

Shu Yang [1], Xuanhan Zhu [2,*], Yang Li [2], Quan Yuan [2,*] and Lili Li [3]

[1] China Communications Information & Technology Group Co., Ltd., Beijing 101399, China; yangshu8@ccccltd.cn
[2] State Key Laboratory of Networking and Switching Technology, Beijing University of Posts and Telecommunications, Beijing 100876, China; leeyang866@bupt.edu.cn
[3] Talent Exchange Center, Ministry of Industry and Information Technology, Beijing 100846, China; lilili@miitec.org.cn
* Correspondence: xuanhan.zhu@bupt.edu.cn (X.Z.); yuanquan@bupt.edu.cn (Q.Y.)

Abstract: The proliferation of wireless technologies, particularly the advent of 5G networks, has ushered in transformative possibilities for enhancing vehicular communication systems, particularly in the context of autonomous driving. Leveraging sensory data and mapping information downloaded from base stations using I2V links, autonomous vehicles in these networks present the promise of enabling distant perceptual abilities essential to completing various tasks in a dynamic environment. However, the efficient down-link transmission of vehicular network data via base stations, often relying on spectrum sharing, presents a multifaceted challenge. This paper addresses the intricacies of spectrum allocation in vehicular networks, aiming to resolve the thorny issues of cross-station interference and coupling while adapting to the dynamic and evolving characteristics of the vehicular environment. A novel approach is suggested involving the utilization of a multi-agent option-critic reinforcement learning algorithm. This algorithm serves a dual purpose: firstly, it learns the most efficient way to allocate spectrum resources optimally. Secondly, it adapts to the ever-changing dynamics of the environment by learning various policy options tailored to different situations. Moreover, it identifies the conditions under which a switch between these policy options is warranted as the situation evolves. The proposed algorithm is structured in two layers, with the upper layer consisting of policy options that are shared across all agents, and the lower layer comprising intra-option policies executed in a distributed manner. Through experimentation, we showcase the superior spectrum efficiency and communication quality achieved by our approach. Specifically, our approach outperforms the baseline methods in terms of training average reward convergence stability and the transmission success rate. Control-variable experiments also reflect the better adaptability of the proposed method as the environmental conditions change, underscoring the significant potential of the proposed method in aiding successful down-link transmissions in vehicular networks.

Keywords: vehicular network communications; performance evaluation; spectrum resource allocation; multi-agent reinforcement learning; option-critic architecture

Citation: Yang, S.; Zhu, X.; Li, Y.; Yuan, Q.; Li, L. Multi-Cell Cooperative Resource Allocation and Performance Evaluation for Roadside-Assisted Automated Driving. *World Electr. Veh. J.* **2024**, *15*, 253. https://doi.org/10.3390/wevj15060253

Academic Editor: Joeri Van Mierlo

Received: 16 April 2024
Revised: 4 June 2024
Accepted: 7 June 2024
Published: 11 June 2024

1. Introduction

In recent years, there has been significant progress in the development of autonomous driving and advanced auxiliary driving technologies [1,2]. However, a critical challenge impeding the widespread adoption of autonomous driving is the issue of safety [3–5]. Existing autonomous driving technologies primarily focus on enhancing the intelligence of individual vehicles [6]. While deep learning technologies have contributed to the designing of more efficient resource allocation algorithms for autonomous driving systems with multiple vehicles and roadside base stations, the advent of wireless communication technologies such as 5G and Cellular Vehicle-to-Everything (C-V2X) has ushered in novel prospects

for autonomous driving satisfying the need for reliable and fast down-link sensory data transmission from roadside base stations with wide-range coverage [7]. These technologies facilitate the reliable real-time exchange of perception information among the mix of vehicles and roadside base stations, thereby augmenting the distant perceptual capabilities of each vehicle and enhancing the safety of autonomous-driving vehicles [6,8]. This collaborative paradigm between intelligent vehicles and roadside base stations effectively addresses the limitations stemming from sight-blocking scenarios that an individual vehicle may encounter [9], ultimately enhancing the feasibility of implementing safer driving protocols.

On the other hand, persistent challenges exist. The foremost challenge revolves around the issue of coupling. Base stations share spectrum resources for communication, resulting in cross-station interference and resource allocation coupling challenges that cannot be effectively resolved by optimizing each base station independently [10,11]. It is imperative to adopt a holistic perspective and resolve these issues through collaborative efforts among various base stations. Centralized or single-level MARL solutions for collaboration entail significant computational overhead and a lack of fine policy granularity, rendering them unsuitable for rapidly evolving environments [12,13]. Consequently, there is a compelling need to develop a distributed execution framework for collaborative resource allocation [14].

Furthermore, the distinct channel states and communication requirements of associated vehicles necessitate the adoption of diverse resource allocation policies at different times and for different individuals [15]. Meanwhile, to circumvent the pitfalls of local optima, base stations must rely not only on their local observations but also on information from other entities. Effectively leveraging this information to accommodate various resource allocation patterns poses a significant challenge [16]. As the demand for underlying communication patterns evolves, the collaboration policies among multiple base stations must adapt accordingly. This underscores the requirement for a centralized structure to assess policy changes, integrate global information and offer guidance for base stations [17].

Resource allocation has gained significant research interest within the realm of conventional methodologies, which encompass heuristic-driven algorithms, optimization-oriented approaches [18], evolutionary algorithms [19] and algorithms grounded in game theory and graph theory [20,21]. In more recent investigations, there has been a focus on the exploration of resource allocation algorithms rooted in multi-agent reinforcement learning (MARL) using deep neural networks. This endeavor aims to advance automated and finely grained strategies for enhancing resource allocation efficiency in intricate environments.

This paper proposes a two-level MARL framework for guiding the resource allocation of base station agents. In sum, the major contribution of this article includes three aspects:

- Proposal of a communication resource allocation framework based on a hierarchical MARL algorithm named multi-agent option-critic architecture for addressing the problem of resource allocation. The architecture has a hierarchical structure for the control of agents and the execution of their actions. Additionally, the option-critic architecture adopted will be more thoroughly introduced in the Related Work and Methodology Sections.
- Creation of a specialized reinforcement learning algorithm designed explicitly for the multi-agent option-critic framework. This algorithm centrally trains agent policies to facilitate collaboration and achieves the autonomous distributed allocation of communication resources.
- We performed several rounds of experiments, each tailored to specific communication demand patterns and environmental parameters. We then conducted thorough comparisons and analyses, considering both baseline methods and alternative approaches. Our observations revealed a noteworthy enhancement in system performance and its ability to adjust to diverse demand patterns when utilizing our algorithm.

The rest of this paper will first present related work that has provided insight into the current research field in spectrum resource allocation for vehicular networks. This is followed by an explanation of the system model, including architecture for the targeted

application scenario, the modeling of communication links and problem formulation. Plus, an exhaustive explanation of the backbone option-critic framework regarding its mathematical mechanisms and training algorithm is given. Lastly, the simulation results using the baseline and proposed methods are presented and analyzed.

2. Related Work

The following sections review the related work concerning two fields relevant to the proposed method. A review of hierarchical reinforcement learning and its subordinating method, the option-critic framework adopted by the method proposed in this paper, is first given. The second subsection reviews the resource allocation problem in vehicular networks researched by previous studies with a focus on the approaches used. It can be seen, as presented by the reviewed studies, that the approaches for solving the spectrum allocation problem are abundant. However, most of the reviewed work in Section 2.2 uses single-level agents, which are not capable of learning different fine-grained policies for adapting pertinently to specific scenarios encountered in a complicated environment. The separation of the two subsections aims to first put a clear emphasis and introduction on the option-critic framework used by the method proposed in this paper. The option-critic framework is a promising hierarchical reinforcement learning approach to be adopted for better adaptation to the environment. Lastly, the second subsection aims to offer a comprehensive view on the research trend of road-infrastructure resource allocation problems using MARL approaches.

2.1. Hierarchical Reinforcement Learning and Option-Critic Framework

Currently, a range of hierarchical reinforcement learning algorithms have been introduced. These algorithms can be categorized into two groups based on their problem-solving approaches. The first category involves deep hierarchical reinforcement learning frameworks that rely on options. In this framework, a lower-level network learns a set of skills, referred to as "options", which are invoked by an upper-level network. Different combinations of these skills are employed to tackle downstream tasks. For instance, Bacon et al. [22] introduced the option-critic framework in light of the actor-critic model. By combining intrinsic and extrinsic rewards and utilizing a gradient-based option learner, the option-critic architecture can effectively learn internal policies, termination conditions and the policy options without the need for additional rewards or sub-goals, showcasing its flexibility and efficiency in training. In an extension of the vanilla option-critic architecture, Riemer et al. [23] proposed a hierarchical option-critic architecture. This architecture enables learning in multiple resolutions at a time by considering an arbitrarily deep hierarchy of options where high-level options are composed of lower-level options with finer resolutions in time. Such a design allows for learning the internal policies, termination conditions and hierarchical compositions of options without the need for intrinsic rewards or sub-goals. Policy gradient theorems are derived for a deep hierarchy of options, including internal policies and termination conditions.

In the context of deep hierarchical reinforcement learning with sub-targets, neural networks are employed to extract state features. The higher-tier network acquires the ability to produce sub-goals, whereas the lower-tier network endeavors to accomplish these sub-goals by utilizing internal mechanisms. Schaul et al. [24] introduced the concept of universal value function approximators (UVFAs) in reinforcement learning. UVFAs are a form of generalization that extend to both states and goals, where they aim to approximate the future reward over a state–goal combination. An innovative supervised learning approach has been proposed for UVFAs, which involves decomposing observed values into distinct embedding vectors for states and goals, followed by the acquisition of a function that maps state–goal pairs to these vectors. Additionally, UVFAs have the capability to be updated exclusively based on received rewards within a reinforcement learning framework, demonstrating an efficient generalization to novel goals. However, it is important to note that hierarchical problems in reinforcement learning pose challenges related to defining

appropriate sub-objectives, and it is crucial to make rational choices in this regard to achieve favorable outcomes.

2.2. Resource Allocation for Vehicular Networks and MARL Solutions

In the context of addressing the optimization problem associated with the allocation of communication resources, a majority of conventional algorithms, including direct solutions, heuristic techniques and graph-based methods, predominantly employ centralized control strategies [15]. These strategies necessitate the presence of roadside base stations or cooperative decision-making vehicles serving as central nodes within a multi-vehicle group. However, this centralized approach encounters difficulties in adapting to evolving environmental conditions. Furthermore, as the network scale and complexity have expanded, the computational demands placed on the central controller have escalated significantly.

Recent advancements in artificial intelligence have facilitated the application of complex deep neural network (DNN) models to address optimization challenges within C-V2X networks [25,26]. Deep reinforcement learning (DRL) techniques have proven effective for resource allocation in C-V2X networks. For instance, Ye et al. [27] proposed a novel decentralized resource allocation mechanism for vehicle-to-vehicle (V2V) communications based on deep reinforcement learning. This mechanism can be applied to both unicast and broadcast scenarios, where each autonomous agent makes decisions independently to find the optimal sub-band and power level for transmission without requiring global information. The proposed method in [27] utilizes deep Q-networks to approximate the Q-function, enabling the selection of optimal actions in a large state–action space. By using deep reinforcement learning, the agents can learn to optimize their actions asynchronously, mitigating issues related to incomplete environmental characterization and enabling coordinated decision-making among neighboring agents.

However, it is worth noting that as actions are updated asynchronously, there are non-stationary problems [28] where the state transitions and rewards are influenced by the joint actions of all agents, leading to an ever-moving target for each agent. Moreover, the states observed by each V2V link may not fully capture the environment, impacting the overall decision-making process.

Multi-agent DRL has gained significant traction in various research areas. For instance, Sangwon et al. [29] proposed a multi-agent deep reinforcement learning (MADRL)-based resource allocation method for multi-cell wireless powered communication networks. The method involves multiple hybrid access points wirelessly charging energy-limited users to collect data. The proposed method aims to achieve comparable performance in terms of the sum data rate to centralized algorithms without the need for global information, but relies on only locally observable states for each agent. Thus, a distributed reinforcement learning strategy implemented utilizing the actor-critic method within the MADRL framework is designed, where hybrid access points independently determine the time and power allocation variables.

A noteworthy study in [30] introduced an MADRL approach to control traffic lights in a vehicular network, aiming to reduce traffic congestion at multiple intersections by making intelligent decisions based on the current traffic environment. Each intersection is controlled by an agent using the deep Q-learning (DQN) algorithm, where decisions are made independently without exchanging information between agents, relying on a greedy algorithm to find the optimal traffic light control policy. Furthermore, [31] proposed a method that uses deep reinforcement learning (DRL) to allocate communication resources in vehicular networks, aiming to manage interference optimally in high-mobility environments and ensure reliable and low-latency services. To enhance system capacity and reduce energy consumption caused by periodic message traffic, a vehicle clustering technique is introduced, which groups vehicles to minimize communication overheads. The proposed method also includes remote radio head (RRH) grouping, which helps in managing the communication resources more efficiently by considering the similarities of neighboring RRHs and forming clusters based on these similarities. Yet striking a balance between

maximizing power efficiency and ensuring extensive message reach presents compromises between the two, which could hinder the achievement of optimal outcomes. Other studies, such as [32] and Sahin [33], addressed the union optimization problem of mode selection and channel allocation in the device-to-device (D2D) communication of heterogeneous cellular networks that use both millimeter wave and traditional cellular bands. These studies aimed to maximize the system sum rate while meeting the QoS requirements of both cellular and D2D users using MARL-based approaches. It is important to note that these studies did not incorporate considerations of half-duplex communication and collision issues into their reward functions.

Prior research efforts by Liang et al. [34] and Vu et al. [35] proposed multi-agent deep reinforcement learning algorithms leveraging a deep Q-network (DQN) and a fingerprint replay buffering mechanism, respectively, to optimize vehicular networking and address issues in spectrum resource allocation and power allocation in V2V communication. In particular, Liang et al. extended the work initially presented in [27] by introducing a method based on MARL to address the problem of spectrum sharing in vehicular networks, where multiple V2V links reuse the frequency spectrum already occupied by vehicle-to-infrastructure (V2I) links. For V2I links, the goal is to maximize their sum capacity, which means ensuring they can efficiently handle as much data as possible. As for V2V links, the goal is to increase the success probability of delivering their payloads within a given time constraint, measured by the packet reception ratio, which is the number of packets received over the number of packets sent. The approaches proposed in [34,35] incorporated innovative view-based location distribution as a means of characterizing the system state. Similarly, Gündoğan et al. [36] examined the challenge of optimizing the total throughput of cognitive users while adhering to energy efficiency constraints. To address the non-stationarity challenge posed by concurrent learning agents, ref. [36] proposed a method to optimize the reward in a series of episodes by comparing different strategies, such as greedy and the proposed DIRAL, to find the optimal approach for maximizing episode rewards over time. Specifically, DIRAL involves adjusting the strategy based on the outcomes of previous episodes, ensuring continuous improvement and better performance over time. The strategy also considers factors like the distance between resources and their reuse, making the training process both effective and resource-efficient. Lastly, He et al. [37] introduced a spectrum allocation framework termed "neighbor agent actor-critic" (NAAC), which employs a combination of centralized training and distributed execution, resulting in enhanced generalization and scalability.

Nevertheless, these prior contributions exhibit certain limitations. For example, the algorithm proposed in [34] by Liang et al. overlooks the half-duplex issues in TDD-based C-V2X radio systems within its system reward design. Additionally, most of these approaches do not adequately account for the evolving dynamics of the environment, which may influence algorithmic performance and possess a limited range of scenarios applicable to state representation.

This paper introduces a spectrum resource allocation methodology grounded in the integration of MARL with hierarchical reinforcement learning, namely, the option-critic architecture. The proposed approach empowers individual agents to make distributed decisions and optimizations based on partial observational data and accumulated training experience, thereby enhancing their capacity to adapt to dynamic environmental changes. In comparison to conventional algorithms, the proposed method exhibits lower algorithmic complexity and superior real-time performance. Meanwhile, compared to prior approaches, the option-critic architecture in the proposed approach helps the agents to better adapt to the dynamically changing environment by flexibly using different policy options. The combination of the two branches can significantly improve the efficiency of spectrum allocation and, therefore, the quality of data transmission in vehicular networks. Additionally, this paper presents a more carefully designed system reward with finer granularity, which can help account for a more complicated and wider range of environmental dynamics in

comparison to previous studies. A summary of the major studies on resource allocation reviewed is listed in Table 1.

Table 1. Comparison of various methods using multi-agent reinforcement learning for resource allocation.

Paper	Collaboration	Hierarchical Model	Global Observation	Use Options	Distributed Execution
[27]	✓	-	-	-	✓
[28]	✓	-	-	-	✓
[29]	✓	-	-	-	✓
[30]	✓	✓	✓	-	✓
[31]	✓	✓	-	-	✓
[32]	✓	-	-	-	✓
[33]	✓	-	✓	-	-
[34]	✓	-	-	-	✓
[35]	✓	-	✓	-	✓
[36]	✓	✓	-	-	✓
Proposed	✓	✓	✓	✓	✓

Note: The checkmarks (✓) indicate the presence of the specified feature in the respective paper.

3. System Model

This section includes the description of the composition and structure of our system's model, the model for establishing communication links between base stations and vehicles, and our objectives for resource allocation. These aspects are discussed thoroughly in this section in terms of the denotations and the mathematical models used.

Overall, the communication network comprises multiple vehicles and homogeneous roadside base stations that could establish communication links in a down-link manner with the vehicles in the network based on C-V2X. The model has a hierarchical structure in which base stations are agents equipped with a two-layer control network. The upper layer is responsible for selecting good policies for making resource allocation decisions and the lower layer is responsible for executing the policies given to it in a distributed manner by allocating communication resources to the end-user vehicles. Resource sharing is achieved by the design, which allows the agents to use the same reservoir of spectrum resources for allocation with the consideration of noise, interference between links and channel fading effects.

3.1. System Architecture

The vehicular communication network primarily comprises multiple base stations, denoted as $\mathcal{N} = \{1, 2, \ldots, n\}$, where n is the total number of base stations. As the network model is assumed to be homogeneous, all the base stations are considered to have the same specifications. The V2I links in the network established between vehicles and base stations are denoted by the set $\mathcal{L} = \{1, \ldots, l\}$.

Figure 1 illustrates that our model adopts a two-level structure for communication. This hierarchical design divides the tasks of selecting policies for resource allocation and the execution of the strategies into two different layers, with one on top of another inside each agent. The upper layer is a centrally trained network for policy selection. During the application of the model, this upper layer of the model is deployed in a distributed manner in all agents and is identical for all agents. The upper layer controls the lower layer of agents (i.e., base stations) by giving them different options of policies to execute for communication resource allocation, facilitating better adaptation of agents to the environment.

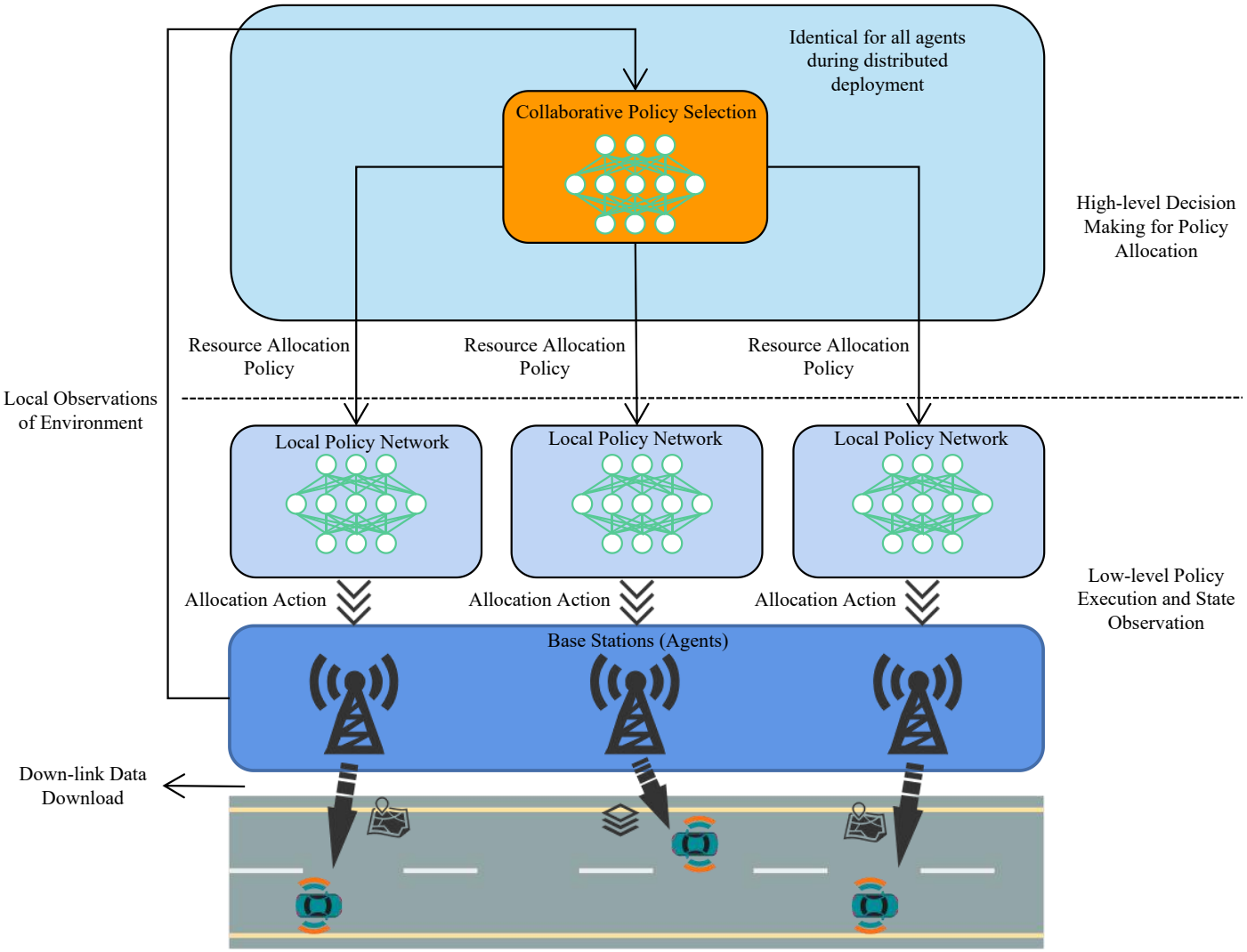


Figure 1. Illustration of the application scenario for the multi-agent option-critic spectrum resource allocation model.

3.2. Communication Links

The model considers the enabling technology for vehicular communication to be C-V2X, through which the V2I links established between base stations and vehicles are assumed to be down-linked without loss of generality. The assumption of a down-link network orients scenarios with needs such as high-resolution map downloading, data acquisition from roadside sensory facilities, etc. As depicted in Figure 1, each C-V2X end user (i.e., vehicle) keeps moving and only receives data packets sent from base stations at every time slot t. For the creation of V2I links, it is posited that the process of orthogonal frequency division multiplexing (OFDM) converts wireless channels with frequency selectivity into uniform channels that run in parallel across various subcarriers, rendering a reservoir of spectrum resources available for allocation. Each sub-band consists of several consecutive subcarriers and the set of sub-bands is denoted as $\mathcal{X} = \{1, \ldots, x\}$, where x is the total number of resource blocks. Moreover, this resource reservoir is designed to be jointly accessible by all base stations. By using this sharing of resources, the aim is to enhance the efficiency of spectrum resource usage. However, this sharing of communication resources inevitably requires an interference control algorithm, which is a main problem that will be discussed in the following sections. Lastly, it is assumed that the V2I links utilize cellular interfaces (i.e., Uu) to enable high data rate transmission and reception between the base stations and vehicles.

Multiplexing the spectrum resource and time resource creates multiple resource blocks for allocation. Assuming the channel fading remains uncorrelated across varying sub-bands, the V2I links using different channels (i.e., resource blocks) are free from mutual interference, while V2I links using the same resource block may create interference. Therefore, this study primarily focused on formulating a proficient algorithm for the allocation of spectra among base stations. The overarching aim is to mitigate interference amid V2I links within an environment characterized by vehicles exhibiting pronounced mobility and robust dynamics. This undertaking strives to optimize the possibility of vehicles realizing their transmission objectives.

The total power of interference on link l is defined as the sum of interference of all other links using the same channel as link l, as shown by the equation below. The interference from another link is computed by the transmit power $P_{l'}[x]$ of the transmitter of the other interfering link l' times the power gain $g_{l,l'}[x]$ from V2I transmitter l' on V2I receiver l on channel x and the binary indicator of spectrum allocation $\rho_{l'}[x]$.

$$I_l[x] = \sum_{l' \neq l} \rho_{l'}[x] P_{l'}[x] g_{l,l'}[x], \tag{1}$$

$g_{l,l'}[x]$ follows the formula for the channel power gain on channel x of the lth V2I link within a coherent time slot:

$$g_l[x] = \alpha_l h_l[x], \tag{2}$$

In (2), the variable h_l corresponds to the power component associated with small-scale fading, an assumption that posits an exponential distribution. Additionally, the parameter α_l encompasses the characteristics of large-scale fading, which comprises path loss and shadowing effects, not subject to the frequency utilized.

With (1) and (2), the received signal-to-interference-plus-noise ratio (SINR) of the lth V2I link occupying channel x is formulated as follows:

$$\gamma_l[x] = \frac{P_l[x]\hat{g}_l[x]}{\sigma^2 + I_l[x]}, \tag{3}$$

where σ^2 represents the power of noise.

It can be further deduced based on Shannon's theorem that the capacities of V2I links can be expressed as

$$C_l[x] = Wlog(1 + \gamma_l[x]), \tag{4}$$

where W denotes the bandwidth for sub-bands.

3.3. Problem Formulation

As mentioned previously, the first objective of resource allocation is to mitigate interference amid V2I links within an environment characterized by fast and flexibly moving vehicles. The second objective of resource allocation is to optimize the possibility of successful transmissions in the vehicular network. More specifically, the objectives are defined as maximizing the sum capacity $\sum_l C_l[x]$ and the possibility of successful deliveries characterized by

$$Pr\sum_{t=1}^{T}\sum_{l=1}^{L} \rho_l[x] C_l[x,t] \geq B \times L/\Delta_T,\ l \in L, \tag{5}$$

in which B represents the bit size of V2I packets, and T signifies the duration of time a packet can live. The variable Δ_T represents the temporal duration of channel coherence. The subscript t is introduced within $C_l[x,t]$ to denote the capacity of the lth V2I link at a distinct coherence time slot t.

Based on the above analysis and delineation of our system model, it can be concluded that our two-level hierarchical model for distributed communication resource allocation

has two main objectives. The first is to find the maximization of the sum capacity of all V2I links C^*:

$$C^* = max \sum_{l} C_l[x]. \tag{6}$$

The second is to find the maximization of the possibility of successful data packet delivery:

$$Pr^*(s) = max(Pr \sum_{t=1}^{T} \sum_{l=1}^{L} \rho_l[x] C_l[x,t] \leq B/\Delta_T),\ l \in L, \tag{7}$$

where s denotes the event of successful transmission.

In the next section, a reinforcement learning framework is further proposed for solving the optimization objectives mentioned in this section.

4. Methodology

The base stations in our model share the same reservoir of time-frequency resources. They are responsible for allocating resources to nearby vehicles when there is a need to send sensory information or high-resolution maps auxiliary to autonomous driving in a downlink manner to vehicles. The optimization goals of maximizing both the overall link capacity and the possibility of successful transmissions can be modeled as an MARL problem.

Base stations are modeled as agents making decisions on resource-allocating actions and acquiring local state observations. The actions taken by agents in the MARL framework rely on each agent's local observation and the policy allocated to the agent. All agents' local observations serve as a reference for the allocation of lower-level policies. This design enables the cooperative sharing of local observations, which helps achieve a more steadily increasing understanding of the dynamic environment and faster convergence to Nash equilibrium for agents with the same optimization goals.

Specifically, an option-critic framework was designed and used as the backbone for the proposed MARL framework. The reasoning for using such an option-critic framework lies in that the policies adopted by agents need to be changed corresponding to the underlying communication pattern. A set of different policies should be learned to cope with different communication situations. The policies with the best chance of eliminating interference and achieving successful data transmission should be selected for the agents. Moreover, the option-critic framework implements the key concept of an option defined as a temporally extended action. The typical composition of an option comprises two policies: the intra-option policy for guiding resource allocation actions and the termination function that determines the timing of the termination of a currently adopted option to be replaced by a new option. The original definition was modified to exclude the use of the termination function. Instead, the selection of new intra-option policies was set to happen at the beginning of each time step.

The full implementation of the multi-agent option-critic framework involved centralized training and distributed execution. Particularly during the centralized training phase, where the focus was on optimizing system performance-oriented rewards, a shared network architecture was employed. This shared network served as a collective repository that gathered observations from all participating agents. Subsequently, the information amassed within this shared network was utilized to update the individual networks associated with each agent. The training process was guided by the QMIX training algorithm [38].

For the distributed execution stage, individual agents received local observations of the environment's current state. Based on these local observations, each agent made decisions by selecting actions through its trained local network. This decision-making process occurred over a temporal scale that was congruent with the rapid fluctuations in small-scale channel fading, indicating quick responsiveness to dynamic changes in the environment. By leveraging their local networks and real-time observations, agents effectively adapted their actions to the ever-changing conditions of the environment, allowing them to navigate

challenges and fulfill tasks within the temporal context of small-scale channel fading. Below are the related factors in detail.

4.1. Observation Space, Action and Reward

Mathematically, the interactive process between agents and the environment, whose transition function is unknown, can be viewed as a Markovian Decision Process (MDP) [39]. S_t is defined as the current environment state at time step t comprising global channel conditions and all agents' actions. Further, the local observation $Z_t^{(k)}$ of agent k is defined as a result of the local observation function $O(S_t, k)$, by which $Z_t^{(k)} = O(S_t, k)$. The observation space of the observation function includes the following:

- Local Channel Information (I): interference from non-local V2I links over the local V2I link (current time).
- Local Resource Block Allocation Matrix (Γ): observations of the local resource block allocation matrix at time $t-1$.
- Vector of Remaining Payload of All V2I Links (E): remaining payload sizes for all V2I links at the current time.
- Remaining Time (V): Remaining time for transmission.

Overall, it renders

$$O(S_t, k) = (E_t, V_t, \Gamma_{t-1}, I_{t-1}). \tag{8}$$

The formula above represents the concatenation of these four matrices or values into a single observation vector.

The local observation $Z_{t+1}^{(k)}$ of agent k depends on all agents' actions taken at time step t defined as a vector A_t, in which each agent k's action is denoted by $A_t^{(k)}$. The action vector $A_t^{(k)}$ denotes, at time t, the identifiers of the resource blocks allocated by base station k (agent k) to the vehicles it is linked with, and its length corresponds to the count of vehicles associated with the base station. The state transition function is expressed in the form of a conditional probability, $P(S_{t+1}|S_t, A_t)$, which illustrates the process by which agents execute actions and engage with the environment, thereby determining the state for the subsequent time step.

Regarding the reward, it is crucial as it directly impacts the direction of all agents' updates. The reward is configured according to the formula below, which matches the two aforementioned optimization objectives:

$$R_t = \sum_l (\lambda_c C_l[x, t] + \lambda_d f_l(t)). \tag{9}$$

Herein, λ_c and λ_d denote positive coefficients for scaling the dual objectives.

To recap, the first aim is to maximize the sum capacity of V2I links in the network. This first aim is reflected by the first term in (9), which computes the sum capacity of all V2I links based on (4). At the same time, the second objective of maximizing the probability of successful data package delivery is reflected by the second term in (9), where $f_l(t)$ for link l is a segmented function in the following form:

$$f_l = \begin{cases} \chi, V_l \geq 0 \, and \, E_l \leq 0, \\ -\chi, V_l < 0 \, and \, E_l > 0, \\ 0, otherwise. \end{cases} \tag{10}$$

with E_l denoting the remaining bit size of V2I packets to be transmitted on link l and V_l denoting the time left for transmission on link l.

The idea behind $f_l(t)$ is to specify the rewards for different outcomes of a transmission. The termination of an episode, which corresponds to a single data packet transmission cycle, occurs under two conditions: the data packet payload has been completely transmitted within the stipulated timeframe, resulting in a zero residual amount of E_l; the set time limit

has been reached and there is still payload that has not been transmitted. If a transmission is completed at any time step t signified by $E_l \leq 0$ within the time constraint V of link l, a positive constant χ is given as the reward to encourage the actions that led to this outcome by adding to the return, denoted as G_t, and defined as cumulative discounted rewards by (11) with a discount rate of γ, which represents the importance of future rewards. On the other hand, the negative value of χ is rewarded as a penalty to a transmission failure on V2I link l signified by the condition $E_l \geq 0$ when the transmission time expires (i.e., $V_l = 0$).

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, 0 \leq \gamma \leq 1. \tag{11}$$

Notably, in the scenario of our model, all agents have the same objectives and, hence, the agents are fully cooperative, collaborating to optimize their individual rewards to achieve a global optimal return. The value for each agent's return depends on not only its own actions $A_t^{(i)}$, but also on all the other agents' actions $\{A_t^{(j)}\}_{i \neq j}$.

4.2. Option-Critic Architecture in MARL Scenario

The option-critic architecture was originally proposed in [22]. The original option-critic was modified and further extended to the problem settings of our MARL communication resource allocation scenario for vehicular networks. To simplify the training process, the termination function of options was replaced by setting the selection of new options at a fixed rate at the beginning of every time step. The extension made on the original option-critic architecture is a hierarchical structure bearing features of centralized training and distributed execution for multi-agents. The acquired results in Section 5 show it to be effective for solving our optimization objectives and enabling the system to adapt to a highly dynamic environment with various QoS requirements. The key ideas and formulas related to the option-critic architecture in a multi-agent scenario will now be delineated as illustrated in Figure 2. Specifically, the upper and lower levels correspond to the higher-level decision-making for policy allocation and low-level policy execution in Figure 1, respectively.

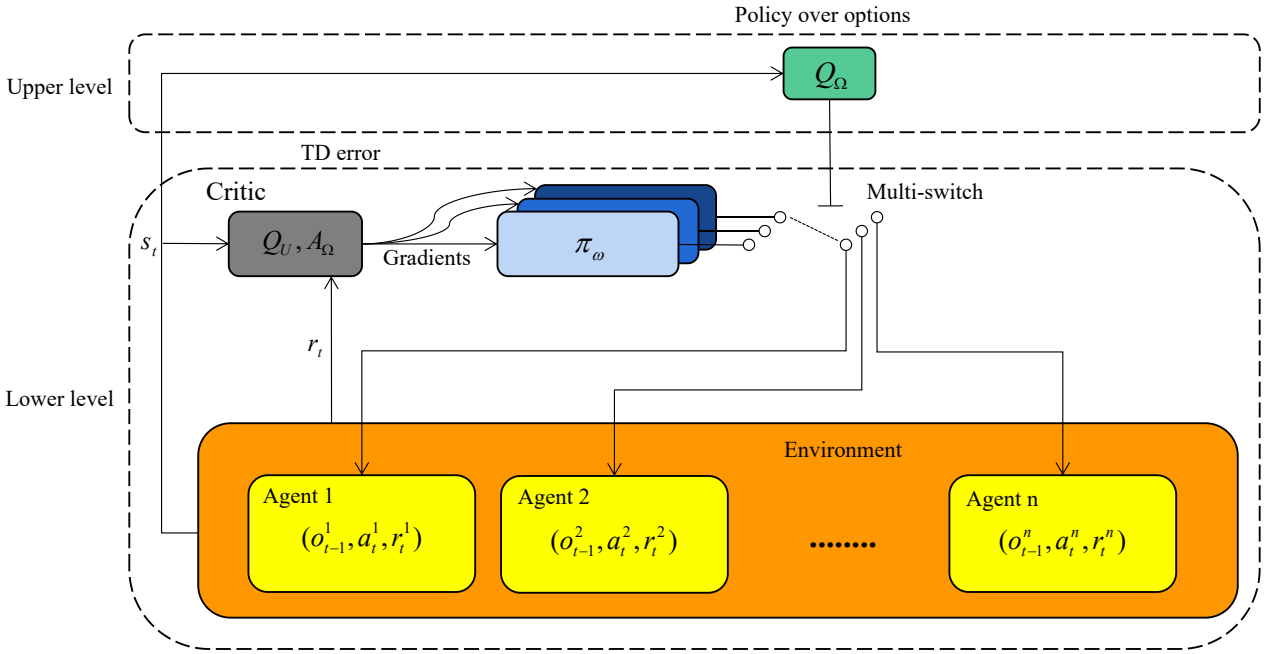


Figure 2. Model structure.

The concept of an option is defined as a temporally extended action [22]. Each Markovian option $\omega \in \Omega$ is mathematically expressed as the combination of a set of

initiation states $I_\omega \subseteq S$, an intra-option policy π_ω that guides the taking of actions before an option terminates. As with the majority of option learning approaches, it is assumed that $s \in I_\omega$ for $\forall s \in S$ and $\forall \omega \in \Omega$, which essentially means that all options are available to be chosen from any initiation state. In our two-level model, options are selected by the upper layer whose intra-option policies are executed in a distributed manner. The option-selecting decisions are made centrally based on global state s and the lower-layer agents' value functions. Then, the lower layer of the model utilizes the given option to guide resource allocation with local observations as input.

The learning of options stems from the direct optimization of the expectation of the global return in a form similar to (11) but starts from the appointed state s_0 and option ω_0 and is the sum of all agents' reward trajectories:

$$G = \sum_{k=0}^{n} \mathbb{E}_{\theta_k, \omega_k}[\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0^k, \omega_0^k], \tag{12}$$

Notably, the global return depends on the policy functions, plus the specific intra-option policies. Thus, the gradients of the optimization objective concerning the parameters of the intra-option policies θ need to be derived. To perform this, the corresponding option-value function for an MDP endowed with a set of options (i.e., a Semi-Markov Decision Process) is firstly found as

$$Q_\Omega(s, \omega) = \sum_a \pi_{\omega,\theta}(a|s) Q_U(s, \omega, a), \tag{13}$$

where a stands for the lower-level actions and Q_U represents the value function of conducting an action under the guidance of option ω, producing a local observation in a certain state s. The mathematical expression for Q_U is

$$Q_U(s, \omega, a) = r(s, a) + \gamma \sum_{s'} Prob(s'|s, a) U(\omega, s'), \tag{14}$$

The gradient of the expected discounted return for the parameters of the intra-option policies θ can then be computed to be

$$\begin{aligned} \frac{\partial Q_\Omega(s, \omega)}{\partial \theta} =& (\sum_{s,\omega} \sum_{t=0}^{\infty} \gamma^t Prob(s_t = s, \omega_t = \omega | s_0, \omega_0)) \cdot \\ & (\sum_a \frac{\partial \pi_{\omega,\theta}(a|s)}{\partial \theta} Q_U(s, \omega, a)). \end{aligned} \tag{15}$$

In our multi-agent model, employing the option-critic architecture means using a local network for each agent representing the option value function $Q_\Omega^k(o^k, a^k)$ as the critic used in both the training and execution phases, and the intra-option policy ω_k as the actor.

4.3. Learning Algorithm and Training Setup

We then designed the execution and training steps of the algorithm. Following the previous methods of resource allocation in vehicular networks [40–42], we set a data packet transmission task as an episode and set the maximum transmission time duration as the maximum time step for each episode. It is worth mentioning that, for the sake of easier comparison with previous methods, we adopted the same execution and training framework and a similar MDP transition process. However, the multi-agent option-critic reinforcement learning algorithm designed in this paper is an innovative hierarchical reinforcement learning strategy that incorporates the particularities of the vehicular network environment, enabling it to adapt to rapidly changing environments and efficiently generate corresponding resource selection plans.

At the commencement of every episode, the process is initiated by resetting the state of the environment and configuring the payload size to a designated value, denoted as E,

to facilitate transmission. This setup remains effective until the progression of steps within the episode reaches the maximum predefined threshold. The alteration in the resource allocation state leads to variations in the small-scale channel fading phenomenon [43], consequently instigating a transformation in the environmental state and prompting each individual agent to adapt their resource allocation strategies accordingly. The training procedure is summarized in Algorithm 1.

Algorithm 1: Resource allocation based on multi-agent option-critic reinforcement learning

```
Start environment simulator, generating vehicles and links
Initialize option-critic network for all agents and overall Q network randomly
for each episode do
    for each step t do
        for each base station agent k do
            Observe O_t^(k)
            Select option ω_t^(k) based on upper level policies
            Choose action A_t^(k) from O_t^(k) and ω_t^(k) according to ϵ-greedy policy
        end for
        All agents take action and receive reward R_t
        Update channel small-scale fading
        All agents calculate TD loss and update value function Q_U(s, ω, a)
        Store (O_{t-1}, A_{t-1}, R_{t-1}, O_t, ω_{t-1}) in the buffer
        if the buffer length is greater than the threshold value then
            Update the upper-level and low-level networks using Monte-Carlo sampling
        end if
    end for
end for
```

In our proposed framework, the challenges associated with resource allocation in networks composed of homogeneous base stations are effectively addressed. These challenges include managing shared spectrum resources and mitigating mutual coupling among base stations. To tackle these issues, it is crucial to obtain a comprehensive view of the global state through the cooperation of multiple agents. The policy network component, situated in the upper tier, plays a pivotal role by centralizing the environmental status data collected by individual base stations. This centralized approach facilitates collective decision-making, ensuring that optimal options are selected for each base station based on a holistic understanding of the network environment.

Simultaneously, the lower tier of agents is tasked with managing diverse communication demand patterns and implementing collaborative resource allocation methods through decentralized decision-making. In this tier, each base station makes specific decisions derived from the options provided by the upper tier. After these decisions are made, the network is updated with global rewards, reflecting the outcomes of the decisions. To support this dual-tier decision-making process, a multi-agent hierarchical reinforcement learning algorithm grounded in the option-critic framework is employed. This algorithm is designed to train multiple agents, aligning with the decision-making modules of each base station. By leveraging this sophisticated learning approach, our framework ensures efficient and effective resource allocation across the network, enhancing overall performance and reliability.

During the execution phase, at each time step t, each agent k estimates local channels and then generates a local observation for selecting an action, $A_t(k)$, with the maximum action value according to its trained hierarchical Q-network. Then, vehicular users use the allocated resource to download data from the connected base station.

5. Simulation Results

5.1. Simulation Environment Setup

This section presents the experiment settings and training results of the proposed method for vehicular network resource allocation in comparison with the following three baseline methods:

(1) Deep deterministic policy gradient (DDPG) baseline, which uses the same actor-critic network for every agent as in single-agent environments. A replay buffer is used for storing past experience, allowing the algorithm to benefit from learning across a range of past states and actions, improving sample efficiency.

(2) Multi-agent deep deterministic policy gradient (MADDPG) baseline in which each agent is treated as a separate learner with its own critic but can share some information with other agents. Specifically, each agent's critics are centralized, meaning they have access to the actions and states of all agents in the environment, but the agents' actors are decentralized, meaning an agent's actor only has access to its own state.

(3) QMIX baseline, which uses a monotonic central joint action-value function for all agents representing the total value of all combined actions of the agents given their collective state.

The simulation environment is self-defined in Python mainly to specifically define modeling parameters for the channel interference effects between V2I links. The parameters and environment setup for the experiments follow those advised by Annex A of 3GPP TR 36.885 [44] regarding the evaluation methodology for studying V2I link-level performance. The main parameters related to the environmental definition and antenna transmission specifications are listed in Table 2.

Table 2. Experiment parameters.

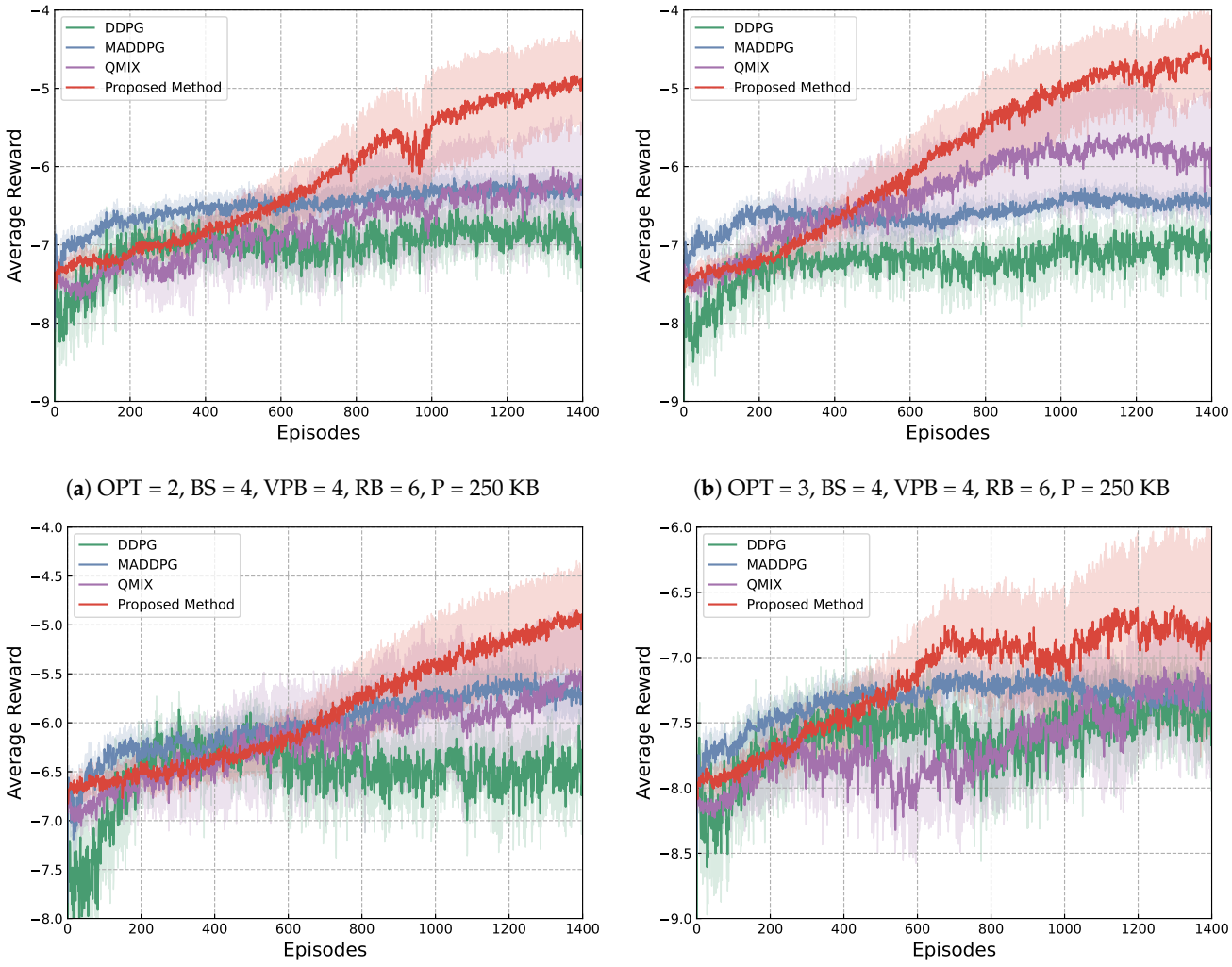
Index	Value
V2I links M	9
Carrier freq.	2 GHz
Bandwidth	4 MHz
Base station antenna height	25 m
Base station channel gain	8 dBi
Base station receiver noise figure	5 dB
Vehicle antenna height	1.5 m
Vehicle channel gain	3 dBi
Vehicle receiver noise figure	9 dB
Absolute moving speed v	15 m/s
V2I transit power P^c	23 dBm
Noise power σ^2	−114 dBm
Time constraint of payload transmission T	100 ms

5.2. Results Analysis

Regarding the results, Figure 3 illustrates the changing of the average reward during six training experiments. Darker curves of different colors are the average of the baselines and the proposed method, whereas their margins indicate the variance among the multiple experiments conducted. Here, average reward refers to the sum of rewards received by the agent(s) over a training episode and divided by the number of time steps in an episode. The training processes were executed with different setups of training parameters. Specifically, OPT indicates the number of options learned, BS indicates the number of base stations used, VPB suggests the number of vehicles that can be connected per base station, RB represents the total number of resource blocks, and lastly is the packet size, represented by P.

From Figure 3, it can be seen that the proposed method achieves a higher value of average reward faster than the other baseline methods in all different setups of the training environment, showing its superiority in stable convergence. By further comparing the

curves of the proposed method in Figure 3a,b, it can be seen that with other parameters being the same, having more available options results in a faster increase in reward, which can be viewed as proof of more options offering more flexibility in dealing with more diverse dynamics in the environment. By comparing Figure 3c,d, it can be seen that with the same setting of the number of base stations, resource blocks, connectable vehicles per station and the number of available options, a bigger packet size resulted in larger volatility during training. This difference suggests that the transmission of bigger packets is harder to complete, which hinders efficient communication in the network overall.



(**a**) OPT = 2, BS = 4, VPB = 4, RB = 6, P = 250 KB

(**b**) OPT = 3, BS = 4, VPB = 4, RB = 6, P = 250 KB

(**c**) OPT = 4, BS = 4, VPB = 4, RB = 6, P = 200 KB

(**d**) OPT = 4, BS = 4, VPB = 4, RB = 6, P = 350 KB

Figure 3. Training curves of the four experimental methods under different environmental setups.

To further investigate the influence of different training parameters, two other evaluation metrics in addition to rewards were used as listed below:

(1) Success rate: the average of the completed data transmission amount in total initiated transmissions. The calculation of the complete rate in one step of an episode follows Equation (16):

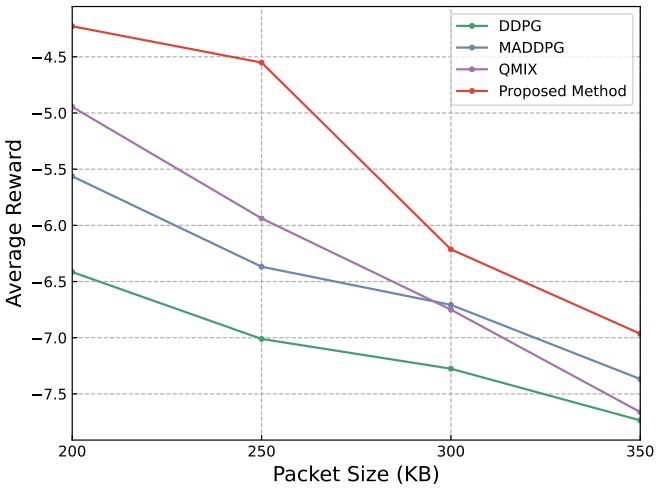
$$\eta = \frac{C}{C+U} \tag{16}$$

where η represents the complete rate reflecting the proportion of completed transmissions C in the total initiated transmissions. The total for the initiated transmissions is calculated by C plus the number of uncompleted transmissions U.

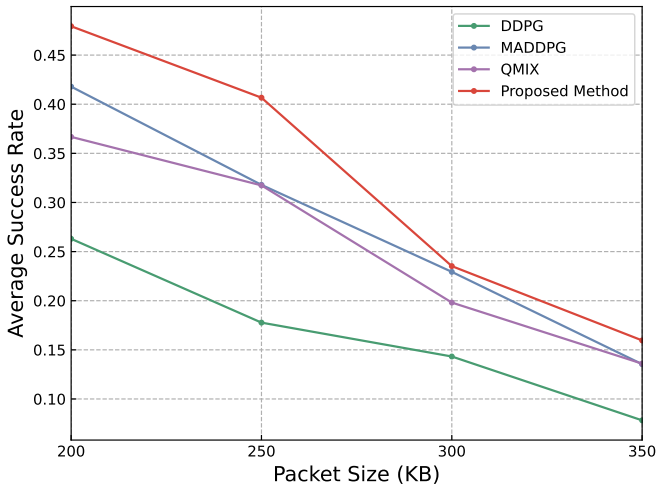
(2) V2I rate: the average of the V2I transmission speed, which is measured by the unit Mbps.

The influence of packet size on the performance of the proposed method and the baseline methods was investigated. From Figure 4, it can be seen that as the packet size increases, all three positive metrics tend to decrease, but the proposed method shows the overall best performance across all three metrics used among all experimental methods, relying on its more intricately designed hierarchical architecture and the learning of more policies for different situations.

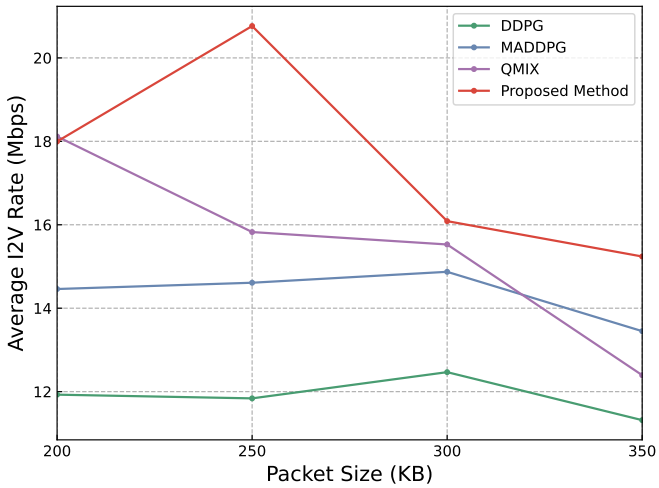
The observed decrease in performance metrics with larger packet sizes among all methods may be due to several factors. Firstly, larger packets are inherently more challenging to transmit completely within a limited time frame of 100 steps, likely leading to increased packet loss or delays. Furthermore, the transmission of larger packets typically consumes a resource block for a more extended period, reducing the flexibility in resource allocation. This extended occupation of resources can result in increased interference among concurrent transmissions, further degrading performance.



(**a**) Average reward vs. packet size.



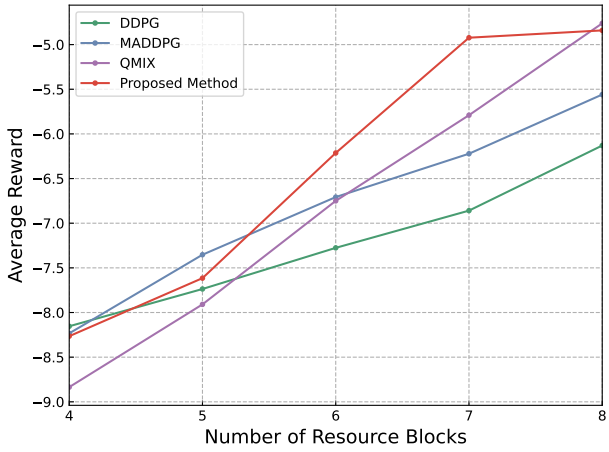
(**b**) Average success rate vs. packet size.



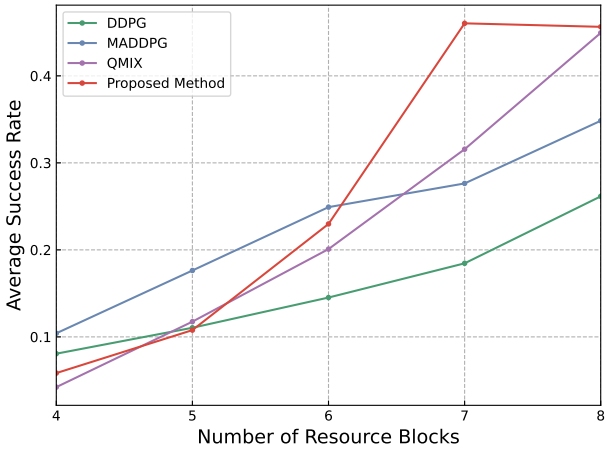
(**c**) Average I2V rate vs. packet size.

Figure 4. Results showing the influence of different packet sizes on reward, transmission success rate and speed. Other parameters are set as OPT = 3, BS = 4, VPB = 4, RB = 6.

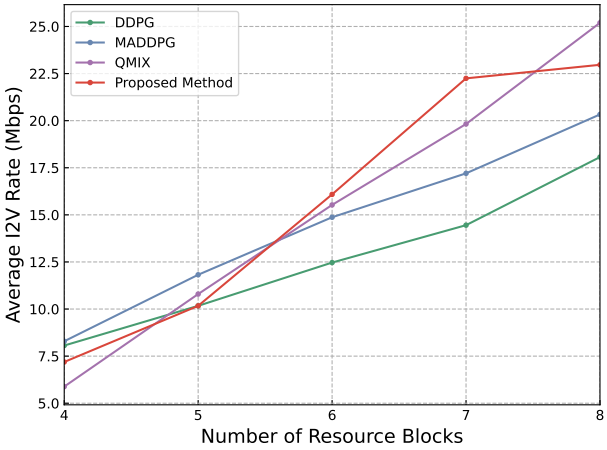
Next, the influence of the number of resource blocks on three metrics was explored as depicted in Figure 5. The general trend is that as the number of resource blocks increases, the three positive metrics also reflect better performance for all methods.



(**a**) Average reward vs. resource blocks.



(**b**) Average success rate vs. resource blocks.



(**c**) Average V2I rate vs. resource blocks.

Figure 5. Results showing the influence of different numbers of resource blocks available on reward, transmission success rate and speed. Other parameters are set to be OPT = 3, BS = 4, VPB = 4, RB = 6.

However, this correlation introduces a subtle interaction among the diverse methodologies implemented in this study. Notably, when the number of resource blocks is below six, the disparities in performance metrics among the tested methods are minimal, indicating a relatively consistent performance across methods under restricted-resource circumstances. This situation evolves as the resource blocks increase to seven. At this stage, the differences in effectiveness between each method become more evident, with the suggested approach notably surpassing others, signifying its superior scalability and efficiency in utilizing additional resources.

The performance superiority of the proposed approach is further emphasized as the resource blocks reach eight. It is at this juncture that the increase in the average reward, average success rate and V2I transmission rate starts to level off, suggesting a reduced benefit from the incremental allocation of resource blocks. This occurrence implies an optimal threshold beyond which additional resources do not result in proportional performance enhancements.

In sum, the experimental results shown in Figure 5 indicate a strategic turning point in resource allocation, where the efficacy of extra blocks starts to stabilize, and, potentially,

further allocation could lead to inefficiencies in resource utilization. Thus, when the number of resource blocks is relatively limited, the performance disparity among various methods is not substantial, reducing the significance of selecting one method over another. Nevertheless, as resources expand, choosing an optimal method becomes crucial for maximizing performance efficiency, with the proposed approach demonstrating particular potential in achieving the most favorable results across all assessed metrics.

Lastly, the relationship between packet size, the number of options learned and the performance metrics was investigated. Specifically, experiments were conducted with a varying range of one to five for the number of options and from 200 KB to 350 KB for packet size.

The trends reflected by the results shown in Figure 6 across all three positive metrics are consistent. Specifically, for smaller packet sizes, such as 200 KB and 250 KB, the performance of the proposed method initially improves as the number of options increases up to three. However, a further increase in the number of options beyond this point leads to a deterioration in performance. This indicates that while a moderate increase in options enhances the system's adaptability and efficiency, an excess may lead to complexity that outweighs these benefits, potentially due to overfitting or increased decision-making time.

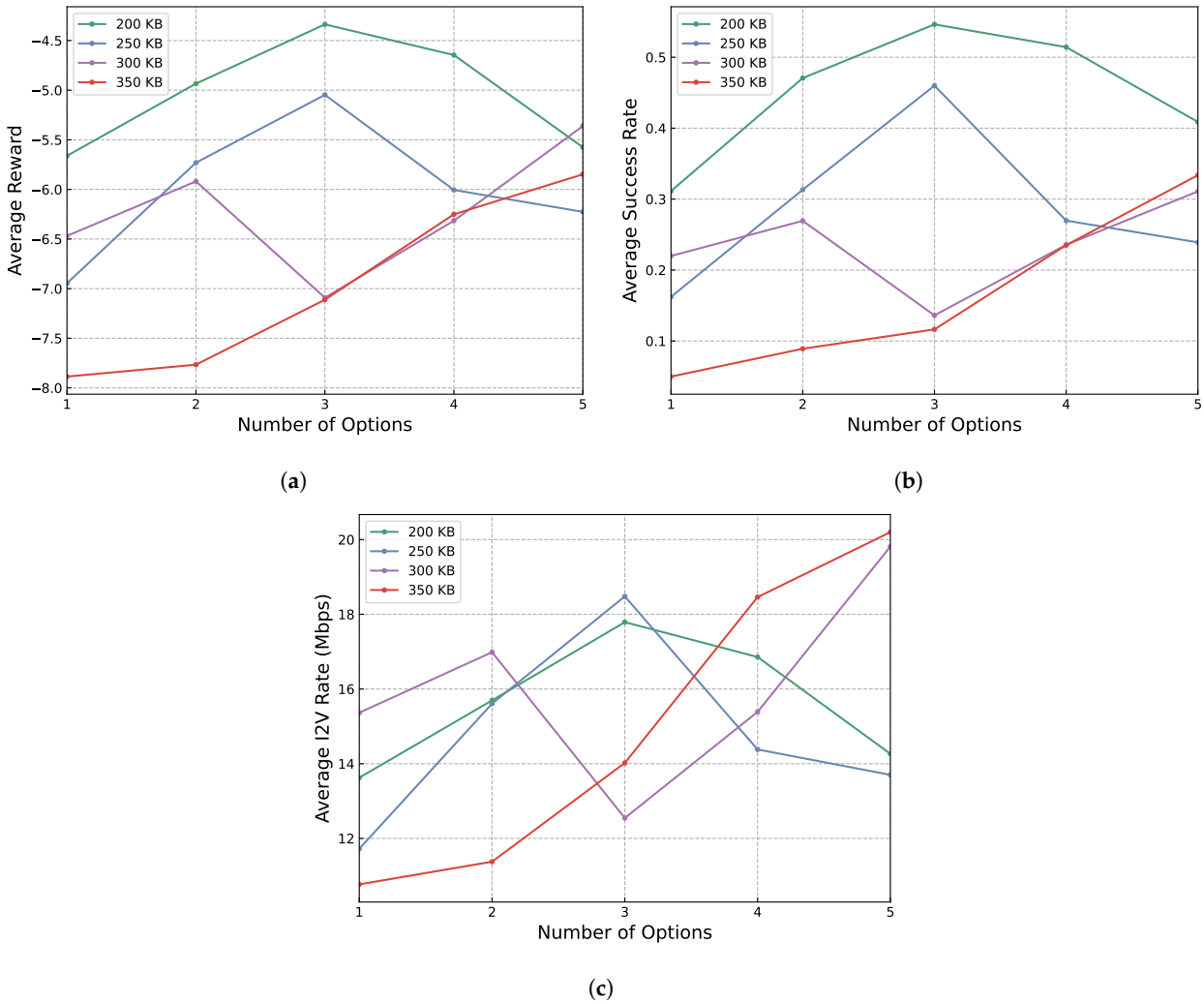


Figure 6. Results of experiments on the influence of option quantity and packet size on the performance of the proposed method. (**a**) Average reward vs. number of options learned and packet size. (**b**) Average success rate vs. number of options learned and packet size. (**c**) Average V2I rate vs. number of options learned and packet size.

Conversely, for larger packet sizes of 300 KB and 350 KB, there is a consistent improvement in performance as the number of options increases, even up to the maximum tested value of five. This trend suggests that larger packet sizes, which introduce more complexity into the network environment, require a broader range of strategies to maintain effective transmission rates. The increased number of options likely provides a more nuanced set of strategies that agents can use to adapt to the challenges presented by larger data packets.

In sum, these observations imply two critical insights: firstly, the incorporation of multiple learning options enhances the adaptability of agents to their operational environment. This adaptability is crucial for optimizing performance, particularly under varying network conditions and demands. Secondly, the complexity introduced by larger packet sizes necessitates a more flexible and strategic approach to option learning. The need for a diverse set of options becomes more pronounced as the environmental complexity increases, highlighting the importance of a scalable and versatile learning framework in managing network performance effectively.

Overall, the presentation of results indicates that the best training configuration is achieved with a moderate number of options (three) and a smaller packet size (200 KB to 250 KB). This configuration consistently yields the highest average reward, success rate and V2I transmission rate, demonstrating superior adaptability and efficiency in dynamic vehicular network environments.

6. Discussion

This article offers a comprehensive analysis of the option-critic framework's impact on spectrum resource allocation optimization using MARL in down-link vehicular networks that consist of multiple base stations and vehicles. This study assessed the effectiveness of the proposed method by examining its influence on both the reward curve and the transmission completion rate, comparing it with two fundamental baseline methods, epsilon-greedy and random. These findings provide basic but valuable insights into the potential of the option-critic framework to enhance spectrum management in autonomous vehicular communication systems within the context of 5G networks.

In contemporary research trends within the field of vehicular communication systems, there is a growing interest in employing advanced techniques like MARL to address the complex challenges posed by autonomous driving and the widespread adoption of 5G networks. These trends reflect the recognition of the imperative need for intelligent spectrum allocation methods to facilitate the reliable and efficient communication required for the safe operation of autonomous vehicles. Furthermore, research in MARL is gaining momentum due to its capacity to model intricate, dynamically changing environments with multiple decision-making agents, which aligns seamlessly with the inherently decentralized nature of vehicular networks.

Looking forward, there exist several avenues for further exploration and improvement in the domain of spectrum allocation optimization for vehicular networks. These include investigating how reinforcement learning algorithms perform under diverse traffic conditions, network densities and interference scenarios. Particularly, future research can focus on using the option-critic framework in an MARL setting for heterogeneous networks (HetNets) and consider the addition of V2V links in the network composition to approximate more complicated application scenarios. Regarding the algorithm itself, the current design uses a single critic for all agents, which is a setting with other potential alternatives that may render better performance. For example, as inspired by the traditional notion of centralized training and distributed execution, it can be designed that each agent has a private individual option critic for a distinct set of options. All option critics can be trained centrally to allow for the acquisition of global state information in the training stage and execution in a distributed manner in the execution stage. In this way, if the experiments were to be conducted in an environment with more specific regional conditions for agents at different locations, intuitively, the use of individual option critics and the learning of

a distinct set of options for each agent could be more pertinent to the specific regional requirements posed by the local environment an agent is in.

7. Conclusions

The results of our experiments suggest that the option-critic framework holds promise for spectrum allocation optimization. Notably, the reward curve illustrates the framework's ability to enhance resource allocation efficiency when compared to baseline methods. Additionally, the transmission completion rate, a critical metric in vehicular networks, displays positive trends when the option-critic framework is employed. Our findings underscore the potential of option-critic reinforcement learning to effectively adapt to the dynamic and evolving nature of the vehicular environment. This adaptability, in turn, enhances communication quality and spectral efficiency, which are vital factors for the success of autonomous vehicular communication systems.

To sum up the work in this paper, a resource allocation algorithm based on the option-critic architecture named multi-agent option-critic is formulated within the context of multi-agent deep reinforcement learning for vehicular networks featuring homogeneous base stations in down-link application scenarios. By conceptualizing base stations as autonomous agents arranged with hierarchical structures, collaboration strategies are synthetically integrated during the training stage among base stations with intricate resource allocation mechanisms occurring within individual base stations. The proposed approach also embraces a dual-stage process of implementation for the multi-agent option-critic framework encompassing centralized training and distributed execution. Ultimately, the effectiveness and significance of our algorithm in contrast to the DDPG baseline, the MADDPG baseline and the QMIX baseline methods are demonstrated empirically. Our methodology demonstrates superior performance compared to the standard methods in terms of the stability of training average reward convergence and the success rate of transmission. Furthermore, experiments with control variables show the increased adaptability of our proposed approach to changing environmental conditions. Furthermore, prospective investigations may delve into the refinement of resource allocation algorithms in HetNets situated in more intricate environmental conditions using a combination of simulation platforms, such as NS3 with Carla, while also incorporating V2V links into the environmental settings to enhance the generality of the considered scenarios.

Author Contributions: Conceptualization, S.Y.; Data curation, X.Z. and Y.L.; Formal analysis, Q.Y. and L.L.; Investigation, X.Z. and Y.L.; Methodology, S.Y.; Software, Y.L.; Validation, X.Z. and Y.L.; Writing—original draft, X.Z.; Writing—review and editing, S.Y., Q.Y. and L.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the "National Key Research and Development Program of China" under Grant 2022YFF0605205, and in part by the BUPT Innovation and Entrepreneurship Support Program under Grant 2024-YC-A086.

Institutional Review Board Statement: This study did not require ethical approval.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Conflicts of Interest: Shu Yang is an employee of China Communications Information Technology Group Co., Ltd. Lili Li is an employee of the Talent Exchange Center, Ministry of Industry and Information Technology. This paper reflects the views of the scientists and not of the companies.

References

1. Sun, J.; Fang, X.; Zhang, Q. Reinforcement Learning Driving Strategy Based on Auxiliary Task for Multi-Scenarios Autonomous Driving. In Proceedings of the 2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS), Xiangtan, China, 12–14 May 2023. [CrossRef]
2. Mishra, A.; Purohit, J.; Nizam, M.; Gawre, S.K. Recent Advancement in Autonomous Vehicle and Driver Assistance Systems. In Proceedings of the 2023 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), Bhopal, India, 18–19 February 2023. [CrossRef]

3. Chib, P.S.; Singh, P. Recent Advancements in End-To-End Autonomous Driving Using Deep Learning: A Survey. *IEEE Trans. Intell. Veh.* **2023**, *9*, 103–118. [CrossRef]
4. Huang, Y.; Chen, Y.; Yang, Z. An Overview about Emerging Technologies of Autonomous Driving. *arXiv* **2023**. [CrossRef]
5. Kosuru, V.S.R.; Venkitaraman, A.K. Advancements and Challenges in Achieving Fully Autonomous Self-Driving Vehicles. *World J. Adv. Res. Rev.* **2023**, *18*, 161–167. [CrossRef]
6. Rawlley, O.; Gupta, S. Artificial Intelligence -Empowered Vision-Based Self Driver Assistance System for Internet of Autonomous Vehicles. *Trans. Emerg. Telecommun. Technol.* **2022**, *34*, e4683. [CrossRef]
7. Khan, M.J.; Khan, M.A.; Malik, S.; Kulkarni, P.; Alkaabi, N.; Ullah, O.; El-Sayed, H.; Ahmed, A.; Turaev, S. Advancing C-V2X for Level 5 Autonomous Driving from the Perspective of 3GPP Standards. *Sensors* **2023**, *23*, 2261. [CrossRef] [PubMed]
8. Zhang, S.; Wang, S.; Yu, S.; Yu, J.J.Q.; Wen, M. Collision Avoidance Predictive Motion Planning Based on Integrated Perception and V2V Communication. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 9640–9653. [CrossRef]
9. Yang, K.; Yang, D.; Zhang, J.; Li, M.; Liu, Y.; Liu, J.; Wang, H.; Sun, P.; Song, L. Spatio-Temporal Domain Awareness for Multi-Agent Collaborative Perception. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023. [CrossRef]
10. Hossain, E.; Rasti, M.; Tabassum, H.; Abdelnasser, A. Evolution toward 5G Multi-Tier Cellular Wireless Networks: An Interference Management Perspective. *IEEE Wirel. Commun.* **2014**, *21*, 118–127. [CrossRef]
11. Kafafy, M.; Ibrahim, A.S.; Ismail, M.H. Optimal Placement of Reconfigurable Intelligent Surfaces for Spectrum Coexistence with Radars. *IEEE Trans. Veh. Technol.* **2022**, *71*, 6574–6585. [CrossRef]
12. Liu, X.; Yu, J.; Feng, Z.; Gao, Y. Multi-Agent Reinforcement Learning for Resource Allocation in IoT Networks with Edge Computing. *China Commun.* **2020**, *17*, 220–236. [CrossRef]
13. Fu, J.; Qin, X.; Huang, Y.; Tang, L.; Liu, Y. Deep Reinforcement Learning-Based Resource Allocation for Cellular Vehicular Network Mode 3 with Underlay Approach. *Sensors* **2022**, *22*, 1874. [CrossRef]
14. Alyas, T.; Ghazal, T.M.; Alfurhood, B.S.; Issa, G.F.; Thawabeh, O.A.; Abbas, Q. Optimizing Resource Allocation Framework for Multi-Cloud Environment. *Comput. Mater. Contin.* **2023**, *75*, 4119–4136. [CrossRef]
15. Nurcahyani, I.; Lee, J.W. Role of Machine Learning in Resource Allocation Strategy over Vehicular Networks: A Survey. *Sensors* **2021**, *21*, 6542. [CrossRef]
16. Hong, J.-P.; Park, S.; Choi, W. Base Station Dataset-Assisted Broadband Over-The-Air Aggregation for Communication-Efficient Federated Learning. *IEEE Trans. Wirel. Commun.* **2023**, *22*, 7259–7272. [CrossRef]
17. Guo, W.; Wagan, S.A.; Shin, D.R.; Siddiqui, I.F.; Koo, J.; Qureshi, N.M.F. Periodic-Collaboration-Based Energy-Efficient Cell Dormancy in Heterogeneous Dense Networks. In Proceedings of the 2022 IEEE 23rd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), Belfast, UK, 14–17 June 2022. [CrossRef]
18. Nasir, Y.S.; Guo, D. Deep Reinforcement Learning for Joint Spectrum and Power Allocation in Cellular Networks. In Proceedings of the 2021 IEEE Globecom Workshops (GC Wkshps), Madrid, Spain, 7–11 December 2021. [CrossRef]
19. Xiu, Z.; Wu, Z. Utility- and Fairness-Based Spectrum Allocation of Cellular Networks by an Adaptive Particle Swarm Optimization Algorithm. *IEEE Trans. Emerg. Top. Comput. Intell.* **2020**, *4*, 42–50. [CrossRef]
20. Zhang, Y.; Zhou, Y. Resource Allocation Strategy Based on Tripartite Graph in Vehicular Social Networks. *IEEE Trans. Netw. Sci. Eng.* **2022**, *10*, 3017–3031. [CrossRef]
21. Qian, B.; Zhou, H.; Ma, T.; Xu, Y.; Yu, K.; Shen, X.; Hou, F. Leveraging Dynamic Stackelberg Pricing Game for Multi-Mode Spectrum Sharing in 5G-VANET. *IEEE Trans. Veh. Technol.* **2020**, *69*, 6374–6387. [CrossRef]
22. Bacon, P.-L.; Harb, J.; Precup, D. The Option-Critic Architecture. In Proceedings of the AAAI Conference on Artificial Intelligence, San Francisco, CA, USA, 4–9 February 2017; Volume 31. [CrossRef]
23. Riemer, M.; Liu, M.; Tesauro, G. *Learning Abstract Options*; Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R., Eds.; Curran Associates, Inc.: Nice, France, 2018; Volume 31.
24. Schaul, T.; Horgan, D.; Gregor, K.; Silver, D. *Universal Value Function Approximators*; Bach, F., Blei, D., Eds.; PMLR: Birmingham, UK, 2015; Volume 37, pp. 1312–1320.
25. Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A.A.; Veness, J.; Bellemare, M.G.; Graves, A.; Riedmiller, M.; Fidjeland, A.K.; Ostrovski, G.; et al. Human-Level Control through Deep Reinforcement Learning. *Nature* **2015**, *518*, 529–533. [CrossRef] [PubMed]
26. Silver, D.; Huang, A.; Maddison, C.J.; Guez, A.; Sifre, L.; van den Driessche, G.; Schrittwieser, J.; Antonoglou, I.; Panneershelvam, V.; Lanctot, M.; et al. Mastering the Game of Go with Deep Neural Networks and Tree Search. *Nature* **2016**, *529*, 484–489. [CrossRef] [PubMed]
27. Ye, H.; Li, G.Y.; Juang, B.-H.F. Deep Reinforcement Learning Based Resource Allocation for V2V Communications. *IEEE Trans. Veh. Technol.* **2019**, *68*, 3163–3173. [CrossRef]
28. Tampuu, A.; Matiisen, T.; Kodelja, D.; Kuzovkin, I.; Korjus, K.; Aru, J.; Aru, J.; Vicente, R. Multiagent Cooperation and Competition with Deep Reinforcement Learning. *PLoS ONE* **2017**, *12*, e0172395. [CrossRef]
29. Hwang, S.; Kim, H.; Lee, S.-H.; Lee, I. Multi-Agent Deep Reinforcement Learning for Distributed Resource Management in Wirelessly Powered Communication Networks. *IEEE Trans. Veh. Technol.* **2020**, *69*, 14055–14060. [CrossRef]
30. Wu, T.; Zhou, P.; Liu, K.; Yuan, Y.; Wang, X.; Huang, H.; Wu, D.O. Multi-Agent Deep Reinforcement Learning for Urban Traffic Light Control in Vehicular Networks. *IEEE Trans. Veh. Technol.* **2020**, *69*, 8243–8256. [CrossRef]

31. Park, H.; Lim, Y. Deep Reinforcement Learning Based Resource Allocation with Radio Remote Head Grouping and Vehicle Clustering in 5G Vehicular Networks. *Electronics* **2021**, *10*, 3015. [CrossRef]
32. Zhi, Y.; Tian, J.; Deng, X.; Qiao, J.; Lu, D. Deep Reinforcement Learning-Based Resource Allocation for D2D Communications in Heterogeneous Cellular Networks. *Digit. Commun. Netw.* **2021**, *8*, 834–842. [CrossRef]
33. Sahin, T.; Khalili, R.; Boban, M.; Wolisz, A. VRLS: A Unified Reinforcement Learning Scheduler for Vehicle-To-Vehicle Communications. *arXiv* **2019**, arXiv:1907.09319. [CrossRef]
34. Liang, L.; Ye, H.; Li, G.Y. Spectrum Sharing in Vehicular Networks Based on Multi-Agent Reinforcement Learning. *IEEE J. Sel. Areas Commun.* **2019**, *37*, 2282–2292. [CrossRef]
35. Vu, H.V.; Farzanullah, M.; Liu, Z.; Nguyen, D.H.; Morawski, R.; Le-Ngoc, T. Multi-Agent Reinforcement Learning for Joint Channel Assignment and Power Allocation in Platoon-Based C-V2X Systems. *arXiv* **2020**. [CrossRef]
36. Gündogan, A.; Gursu, H.M.; Pauli, V.; Kellerer, W. Distributed Resource Allocation with Multi-Agent Deep Reinforcement Learning for 5G-V2V Communication. *arXiv* **2020**, arXiv:2010.05290. [CrossRef]
37. He, H. Research on Key Technologies of Dynamic Spectrum Access in Cognitive Radio. Ph.D. Thesis, University of Electronic Science and Technology of China, Chengdu, China, 2014. [CrossRef]
38. Rashid, T.; Samvelyan, M.; Schroeder, C.; Farquhar, G.; Foerster, J.; Whiteson, S. QMIX: Monotonic Value Function Factorisation for Deep Multi-Agent Reinforcement Learning. *J. Mach. Learn. Res.* **2018**, *21*, 4292–4301.
39. Zhou, Z.; Liu, G.; Tang, Y. Multi-Agent Reinforcement Learning: Methods, Applications, Visionary Prospects, and Challenges. *arXiv* **2023**, arXiv:2305.10091. [CrossRef]
40. Hou, W.; Wen, H.; Song, H.; Lei, W.; Zhang, W. Multiagent Deep Reinforcement Learning for Task Offloading and Resource Allocation in Cybertwin-Based Networks. *IEEE Internet Things J.* **2021**, *8*, 16256–16268. [CrossRef]
41. Parvini, M.; Javan, M.R.; Mokari, N.; Abbasi, B.; Jorswieck, E.A. AoI-Aware Resource Allocation for Platoon-Based C-V2X Networks via Multi-Agent Multi-Task Reinforcement Learning. *IEEE Trans. Veh. Technol.* **2023**, *72*, 9880–9896. [CrossRef]
42. Sheikh, H.U.; Bölöni, L. Multi-Agent Reinforcement Learning for Problems with Combined Individual and Team Reward. *arXiv* **2020**, arXiv:2003.10598. [CrossRef]
43. Jang, J.; Yang, H.J. Deep Reinforcement Learning-Based Resource Allocation and Power Control in Small Cells with Limited Information Exchange. *IEEE Trans. Veh. Technol.* **2020**, *69*, 13768–13783. [CrossRef]
44. 3GPP TR 36.885. Available online: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=2934 (accessed on 26 August 2023) .