# HierCoop: Hierarchical Cooperative Resource Allocation for Vehicular Communications in Heterogeneous Networks

Yixuan Feng, Quan Yuan*, Guiyang Luo, and Jinglin Li

State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications, Beijing, 100876, China
Email: {fengyixuan, yuanquan, luoguiyang, jlli}@bupt.edu.cn

***Abstract*—Heterogeneous Networks (HetNets) can provide intelligent vehicles with high-bandwidth access and support large-scale vehicle-infrastructure cooperation. However, the spectrum allocation of HetNets faces the problem of mutual coupling between macrocell base station (MBS) and intensively deployed small-cell base stations (SBSs) which share spectrum band. While recent research has used Multi-Agent Deep Reinforcement Learning (MADRL) to allocate resources for vehicle-infrastructure cooperation, it has focused solely on intra-cell resource distribution, overlooking inter-cell resource cooperation. Inter-cell resource cooperation in HetNets faces significant communication overhead due to real-time link state interaction between base stations. Additionally, timely updates to inter-cell resource allocation policies are critical when vehicular communication patterns undergo shifts. In this paper, we propose a hierarchical cooperative resource allocation method (HierCoop) for addressing the aforementioned challenges in vehicular communications in HetNets. Each base station is modeled as an agent with hierarchical structure which combines coarse-grained inter-cell collaboration policies and fine-grained inner-cell resource allocation. Specifically, the upper layer of each agent is deployed centrally, thus enabling inter-cell collaboration policies to be based on the abstract channel state features extracted from the lower layers instead of the original channel state information. Meanwhile, the lower layer of each agent is deployed locally in base station, which executes actual resource allocation action based on the policy formulated by the upper layer. The experimental results demonstrate that HierCoop can significantly enhances spectral efficiency and communication quality.**

***Index Terms*—heterogeneous networks, vehicular communications, resource allocation, hierarchical multi-agent reinforcement learning**

## I. Introduction

Cellular vehicle-to-everything (C-V2X) communications is a crucial technology for autonomous driving. It enables real-time information sharing to achieve collaborative perception and control, supporting the services required by intelligent and connected vehicles (ICVs), including digital twin rendering, navigation, AR/VR-based driving assistance and perceived data sharing, etc, which are extensive delay-sensitive and computation-intensive and put forward more stringent requirements at vehicles. [1] Recently, heterogeneous cellular networks (HetNets) have been introduced into C-V2X to provide ubiquitous coverage and high transmission rates to vehicles. Multiple tiers of base stations are utilized to improve the network's spectral efficiency and system capacity [2]. The multi-tier cellular system comprises a macrocell base station (MBS) and multiple small-cell base stations (SBSs). Deploying low-power SBSs can eliminate the coverage gap between MBSs and improve the capacity of hot spots. Due to their lower transmission power and smaller physical size, SBSs can be deployed flexibly and share the spectrum with MBS.

With the rise of intelligent vehicles, there's a growing need for better service quality, particularly in heterogeneous networks where spectrum allocation is complex. This complexity arises from cross-tier interference and coupling resource allocation challenges when macro base stations (MBS) and small base stations (SBS) share spectrum. Optimizing individual base stations alone cannot address these issues. [3] A holistic, global approach is required, but centralized solutions are computationally expensive and slow to adapt to dynamic environments. Additionally, the varying roles and communication requirements of SBSs, driven by diverse channel states and associated vehicles, create inherent heterogeneity. To prevent local optimization problems, SBSs need not only local observations but also information from other SBSs, resulting in significant information overhead. Adapting collaboration policies between base stations to changing communication patterns requires a centralized structure that integrates global information and guides policies for lower-tier independent base stations.

The issue of resource allocation has received substantial attention from researchers who have employed various traditional approaches, such as heuristic-based algorithms, optimization-based algorithms [4], evolutionary algorithms [5], and also game-theory and graph-theory based algorithms [6,7]. Deep reinforcement learning excels at acquiring highly effective state representations for tackling demanding

This paper is supported in part by the National Key Research and Development Program of China under Grant 2022YFB4300402, the Natural Science Foundation of China under Grant 62272053, Grant 62102041, and in part by the Young Elite Scientists Sponsorship Program by China Association for Science and Technology (CAST) under Grant 2022QNRC001.

tasks and addressing real-world issues. In recent years, researchers have investigated resource allocation algorithms based on multi-agent deep reinforcement learning (MADRL) to create automated and more precise schemes with the aim of enhancing resource allocation efficiency in complex environments.

Nevertheless, algorithms that demonstrate good performance in simple environments encounter challenges in heterogeneous network environments with a MBS and intensively deployed SBSs. Collecting complete and precise knowledge of wireless environments, which is essential to these algorithms, is complex due to the system's scale, ultra density, and heterogeneity. However, most of these algorithms presume that all base stations are homogeneous and possess statistically equal capabilities. This assumption is inconsistent with HetNets, where network entities are tremendously heterogeneous in terms of communication and computational capabilities. Moreover, owing to the extensive mobility and constant interaction among vehicles on the road, the dynamics of the vehicular environment are undergoing swift transformations such as weather factors and congestion.

Based on above analysis, to solve the challenge of resource allocation in internet of vehicles (IOV) based HetNets, we develop a hierarchical MADRL algorithm to allocate resource efficiently, combining coarse-grained collaboration policies between base stations and fine-grained poicies within base stations. The key contributions of this paper are as follows:

- First, in order to tackle resource management and collaboration challenges in HetNets, we propose an resource allocation architecture based on the hierarchical MADRL algorithm called HierCoop, which composed of two levels and improves spectrum efficiency.
- Second, we developed a specific multi-agent hierarchical reinforcement learning algorithm which aids agents in decision-making and learning at different levels, while enabling information transmission and collaboration between levels.
- Third, we conducted multiple sets of experiments targeting different communication demand patterns and environmental parameters, compared and analyzed them with other methods and noticed that our algorithm can significantly improve the performance of the system and its adaptability to different demand patterns.

## II. Related Work

### A. Resource Allocation in HetNets

There are extensive research on interference aware resource allocation in HetNets with shared spectrum. Traditional methods include heuristic search algorithms, graph based and game theory based algorithms, algorithm based on matching theory and so on. Zhu et al. [8] proposes a game-theory based resource allocation algorithm where SBS users and MBS users compete with each other to maximize their own performance. Kim et al. [9] propose a resource allocation scheme for dense deployed SBSs to mitigate co-tier interference, assuming cross-tier interference as white noise. Abdelnasser et al. [10] propose a joint resource allocation and access control method to optimize the performance of MBS users and SBSs users by formulating an optimization problem and obtain results through convex relaxation and employing dual decomposition technique. Liu et al. [11] propose a heuristic subchannel assignment algorithm and a Taylor series and successive convex approximation-based power allocation algorithm to solve a mixed integer non-linear programming optimization problem in cognitive satellite-unmanned aerial vehicle (UAV)terrestrial network.

Recent research focus on deep reinforcement learning based methods, which have good performance in making judicious control decisions in uncertain wireless environments. In order to enable network entities conduct optimization actions locally, only local observation is required. Because usually DRL based methods uses centralized training and distributed execution policies. When training, agents use the full connection layer of neural network for information transmission and fusion or use global state information to train the value function network of each agent respectively, promoting the update of the policy network of each agent. When the external environment changes, the intelligent agent can quickly respond and converge based on the trained policy. This intelligent dynamic spectrum allocation method has great advantages over traditional algorithms in adaptability to dynamic environment and real-time performance of spectrum reallocation [12].

Liang et al. [13] proposed a DQN based distributed algorithm to optimize the joint optimization of spectrum resources and power allocation of V2X network, and proposed a fingerprint based replay buffer mechanism to solve the non-stationary problem. Vu et al. [14] also proposes a DQN algorithm based on fingerprint replay buffering mechanism to solve the spectrum resource allocation problem of the C-V2X fleet driving system. Gundogan, et al. [15] proposes an algorithm based on MARL to optimize the allocation of V2V communication spectrum resources in congested scenarios. They propose a view based position distribution as a special state representation of agents to cope with non-stationary characteristics. Although these works [13,14] have shown some improvements of performance, there are still some shortcomings. Most of the above works ignore the impact of channel fading interference brought by the high mobility feature of vehicular networks. For example, the authors of [14,15] did not consider the impact of changes in environmental dynamic characteristics on algorithm performance, while the state representation proposed by Gundogan, et al. [15] is only applicable to scenarios where vehicles move in a single direction.

### B. Hierarchical Reinforcement Learning

Currently, numerous hierarchical reinforcement learning algorithms exist, which fall into two categories based on their original concepts: option-based deep hierarchical reinforcement learning, where the lower network acquires skills,

and the upper network combines these skills for downstream task solutions. PL et al. [16] proposed the option-critic (OC) framework which adopts the similar structure as the actor-critic framework. Based on the policy gradient theorem, it maximizes the expected return by optimizing the internal policy and interruption function of option. On the basis of OC, Riemer et al. [17] proposes a layered option-critic algorithm, which extends the two-layer OC framework to the multi-layer OC framework. They applies the gradient theorem of the option's internal policy and the gradient theorem of the interruption function to all levels, selects the option from top to bottom, judges the interruption function from bottom to top, and calls or interrupts the option layer by layer. While another deep hierarchical reinforcement learning framework is based on sub-objectives, where state features are extracted using neural networks, and are taken as the sub-objectives space. The upper level network learning to generate sub-objectives, while the lower level network achieves sub-objectives based on internal drivers. Schaul et al. [18] is using an unified value function, which adds goal as input on top of the original value function, making the value function become the value under a certain state (or state action) and a certain goal. Directly fitting the corresponding hidden variables through any state and target and integrating them into the state target value plays a generalization role. Such hierarchical problems face prior setting problems and is necessary to select sub-objectives reasonably to achieve good results.

## III. System Model

We consider a vehicular communication heterogeneous network based on C-V2X. As Fig. 1 shows, each C-V2X user keeps moving and requests to send data packets to base stations. An MBS and $n$ SBSs denoted as $N = \{1, 2, \ldots, n\}$. V2I links leverage cellular (i.e., Uu) interfaces to connect vehicles to BSs for high data rate services. At every time slot $t$, SBSs receive requests and select for vehicles to connect to MBS or itself, and reallocate spectrum resources for vehicles which send requests to it. We assume the set of V2I links in our vehicular network model are denoted by $L = \{1, \ldots, l\}$ respectively.

In this HetNet model, MBS has a pool of spectrum resources that it can select for connected vehicles to construct V2I communications. Such resource pool is shared between an MBS and all SBSs within the coverage range of MBS for better spectrum utilization with the premise of necessary interference control algorithm which we will study in this work. Furthermore, we assume that any of the V2I links (without loss of generality, we choose uplink) may use any of the spectrum resource. Therefore, the main challenge of the work is to design an efficient spectrum sharing algorithm for base stations to minimize the interference between V2I links in the environment containing vehicles of high mobility and strong dynamics and maximize the vehicles' possibility of achieving their transmission goals.

As shown in Fig. 1, our architecture is divided into upper level and lower level within agents and improves overall efficiency through hierarchical resource allocation. Specifically, the upper layer of agents is deployed on the edge cloud of MBS, and the lower layer of agents will be deployed in each SBS separately. We have designed a collaborative mechanism in the upper layer to achieve resource sharing between agents, and transmitted the selected policies to the lower layer, enabling agents to adaptively execute actions based on policies in a distributed manner at the lower level.

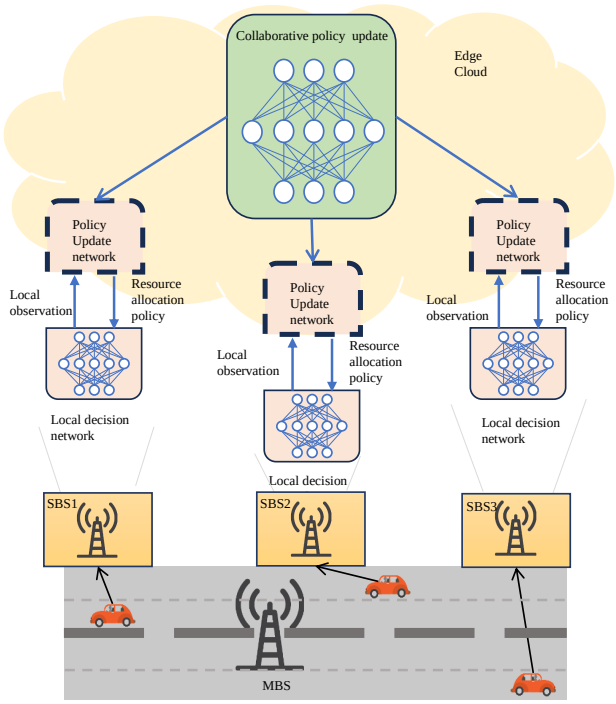


Fig. 1. Framework of HierCoop.

The channel model is based on orthogonal frequency division multiplexing (OFDM) technology, which converts the frequency selective wireless channels into multiple parallel flat channels over different subcarriers. Several consecutive subcarriers are grouped to form a spectrum sub-band. The sub-bands in this model are denoted as $M = \{1, \ldots, m\}$. Therefore, V2I links allocated to different channels do not have interference between each other. During a coherence time slot $t$, the channel power gain $g_l[m]$ on channel $m$ of $l$th V2I link is defined as follows:

$$g_l[m] = \alpha_l h_l[m], \tag{1}$$

where $h_l$ is small-scale fading power component, which is assumed to be exponentially distributed and $\alpha_l$ is large-scale fading parameter, including path loss and shadowing, and is frequency independent. Interference exists only at the time when two V2I link are assigned to same resource block. Similarly, we denote the interfering channel from the $l'$th V2I transmitter to the $l$th V2I receiver over the $m$th sub-band as $g_{l,l'}[m]$.

Therefore, the received signal-to-interference-plus-noise ratios (SINRs) of the $l$th V2I link over resource block $m$ are expressed as

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

where $P_l$ denotes transmit power of the $l$th V2I transmitter over sub-band $m$, $\sigma^2$ is the noise power, and

$$I_l[m] = \sum_{l' \neq l} \rho_{l'}[m] P_{l'}[m] g_{l,l'}[m], \tag{3}$$

denotes the interference power, where $\rho_{l'}[m]$ is a binary spectrum allocation indicator. $\rho_{l'}[m] = 1$ implies the $l'$th V2I link uses the $m$th sub-band and otherwise.

Capacities of the $l$th V2I link over sub-band $m$ are therefore derived as follows based on Shannon's theorem:

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

where $W$ is the bandwidth of each spectrum sub-band.

As mentioned before, the fundamental purpose is to make V2I links support mobile high data rate services such as digital twins construction, and meanwhile ensure the reliability of data package delivery to provide smooth and reliable advanced driving service. Therefore the design objective is to maximize the sum capacity and the success possibility of data package delivery. The sum capacity is defined as $\sum_l C_l[m]$. And the success possibility of data package delivery can be represented using the following formula:

$$Pr\{\frac{\sum_{t=1}^{T}\sum_{l=1}^{L}\rho_l[m]C_l[m,t]}{T} \leq \frac{(B \times L)}{\Delta_T}\}, l \in L, \tag{5}$$

where $B$ denotes the size of V2I packages in bits and $T$ denotes the lifetime of the packet, the $\Delta_T$ is channel coherence time, and the index $t$ is added in $C_l[m,t]$ to indicate the capacity of the $k$th V2I link at different coherence time slots.

From the above analysis, it can be concluded that our work's resource allocation problem can be stated as follows: to design a V2I spectrum resource allocation algorithms to let base stations select spectrum for its V2I links to maximize the sum capacity of all V2I links and the success possibility of data package delivery defined in (5).

Therefore, an improved multi-agent reinforcement learning framework is proposed in this paper to help solve the problem of distributed spectrum allocation of base stations.

## IV. Method

In our resource allocation scenario described in section III, an MBS and multiple SBSs share spectrum resource and SBSs receive nearby vehicles' requests and reallocate resource at each time slot for requesting vehicle respectively. Specifically, SBSs choose whether to connect to MBS or itself for the vehicle and a particular frequency band. This scenario can be modeled as a multi-agent reinforcement learning problem. We model each SBS as an agent which interacts with the unknown communication environment to gain experiences and use the experiences to guide the directions for agents to optimize their action selection. At every step, multiple agents will take joint actions, get local observations and exchange information to establish a deep understanding of the environment and therefore to improve their spectrum allocation policies. This resource allocation problem is fully cooperative because each agent in the environment use common optimization objective.

In this paper, we adopt the multi-agent reinforcement learning framework. The entire process include centralized training and distributed execution. In the training phase when system performance-oriented reward is accessible, agents have a shared network which collects the observation of all agents and update the network of each agent. In execution phase, each agent receives local observations of environment state and selects an action according to its trained local network on a time scale on par with the small-scale channel fading. The specific essential factors of multi-agent reinforcement learning are defined as follows.

### A. Observation Space, Action and Reward

As mentioned earlier, each SBS is modeled as an agent and concurrently explores the unknown environment. We can model the scenario as an MDP mathematically. At each coherence time step $t$, let the current environment state be $S_t$ which include global channel conditions and all agents' behaviors and is unknown to each individual agent. Due to the fact that each SBS agent can only acquire knowledge of the underlying environment through the output of local observation function, let each agent $k$'s observation is determined by the observation function $O$ as $O(S_t, k)$ and take action $A_t^{(k)}$. After that, agents receive reward $R_{t+1}$ and the environment evolves to the next state $S_{t+1}$ with the probability of $p(S_{t+1}, R|S, A)$ and the new observation $O(S_{t+1}, k)$ is received by each agent. It deserves to be emphasized that all agents share the same reward and are encouraged to collaborate with each other.

The observation space of agent $k$ contains local channel information, which includes local channel information $I$, i.e. the interference from V2I links connected to other agents over local V2I links at time $t-1$, a local observation of resource block allocation matrix $RA_{t-1}$ at time $t-1$, the remaining payload of every V2I links $B_t$ and remaining time $T_t$ at time $t$. Wherein local channel information refers to the interference channels from other V2I transmitters to local receiver. Such information can be accurately estimated by the receiver at the beginning of each time slot. Therefore, the observation function for agent $k$ is summed up as follows:

$$O(S_t, k) = (B_t, T_t, RA_{t-1}, I_{t-1}). \tag{6}$$

The action of resource allocation can be divided into two parts: connect base station selection and resource block selection. Each SBS will first select for its connected vehicle whether to connect to MBS or simply connect to local SBS and then choose the spectrum sub-band for the vehicle. Therefore, the dimension of the action space is $2 \times M \times V$,

wherein $V$ denotes the number of vehicles connected to the base station.

The setting of reward, i.e., the optimization objective, is an essential part for reinforcement learning algorithms because the feedback from environment decides the direction of network updating. With more reasonable rewards, the performance of the system can be improved. In our model above, we have two main objectives: maximizing the sum V2I capacity and meanwhile increasing the success probability of payload delivery within a certain time constraint $T$.

For the first objective, we directly compute the instantaneous sum capacity of all V2I links, $\sum_{l\in L} C_l[m,t]$ as defined in (3) in the reward of each time step $t$. As for the latter objective, we adopt a piece wise function $TC_l$, which is set as 0 when a packet is transmitting and is set to a certain constant number $\phi$ when transmitting is complete. When transmission is failed, $TC_l$ is set to a negative constant number to discourage these behaviours. Therefore, the expression of $TC_l$ at every time step $t$ is as follows:

$$TC_l = \begin{cases} 0, B_l > 0 \\ \phi, T_l \geq 0 \, and \, B_l \leq 0 \\ -\phi, T_l < 0 \, and \, B_l > 0 \end{cases} \tag{7}$$

The goal of reinforcement learning is to find an optimal policy $\pi^*$, in other words, to find a mapping from states to probabilities of selecting each action that maximizes the expected return from any initial state, where the return, denoted by $G_t$,is defined as the cumulative discounted rewards with a discount rate $\gamma$ as follows:

$$G_t = \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}, 0 \leq \gamma \leq 1. \tag{8}$$

The $\gamma$ is discount rate which indicates the importance attached to future rewards. If setting $\gamma$ to 1, the system will attach more importance to data package delivery.

Considering the ultimate goal and learning efficiency, we set the overall reward as follows to achieves a balance between these two objectives.

$$R_{t+1} = \sum_{l} (\lambda_c C_l[m,t] + \lambda_d TC_l(t)). \tag{9}$$

where $\lambda_c$ and $\lambda_d$ are positive weights to balance two objectives.

### *B. Architecture and Learning Algorithm*

At each episode, we reset the environment state and set the payload to size $B$ for transmission, and lasts until the step reach to max steps in an episode. The resource allocation state cause the change of small-scale channel fading and therefore triggers a transition of the environment state and causes each individual agents to adjust their resource allocation actions.

By combining the characteristics of centralized and distributed DRL methods, we conduct an effective resource allocation framework to adapt to highly changing environments and meet the different QoS requirements of vehicles. We develop a new reinforcement learning framework called HierCoop to train the multiple agents to learn effective resource allocation policies.

As mentioned earlier, each SBS is modeled as an agent with upper and lower layers, with the upper layer selecting abstract policies for the lower layer and the lower layer selecting specific actions based on the policies. Therefore, HierCoop extends the option-critic architecture with hierarchical structure and option mechanism to multi-agent scenarios.

The option-critic architecture [14] abstracts the timing of actions using the concept of option. An option $\omega$ is represented as a triplet $(I_\omega, \pi_\omega, \beta_\omega)$ which representing the initial state set of options, the policy for selecting actions, and the termination function which describe whether the upper layer needs to replace the current option respectively, where the option is corresponding to the upper layer policies and the $\pi_\omega$ represents the lower layer policies. To be more concrete, in the above HetNet secnario, the conception of option is corresponding to the abstract policy selected by upper layer of SBS agents to instruct selection of actions of lower layer. The upper layers of SBS agents take global state $s$ and value function of lower layer as input and output the option for lower layer centrally. While the lower layers of SBS agents then utilize option as well as local observation as input to select specific resource allocation actions.

Our ultimate objective is to optimize the global discounted reward. Based on (8) and combined with option-critic, $G_t$ can be further represented as:

$$G_t = \sum_{k=0}^{n} \mathbb{E}_{\theta_k, \omega_k} [\sum_{t=0}^{\infty} \gamma^t r_{t+1} | s_0^k, \omega_0^k], \tag{10}$$

where $n$ represents the number of agents. For each agent $k$, local observation $O_k$ is used to approximate the current state $s_k$ and apply it to the following text. $G_t$ depends on policy over options and the parameters of the option policies and termination functions of all agents. First, the option-value function is represented as follows:

$$Q_\Omega(s,\omega) = \sum_{a} \pi_{\omega,\theta}(a|s) Q_U(s,\omega,a), \tag{11}$$

where $a$ represents the action of SBSs, $Q_U$ is the value function of conducting an action with local observation and option selected by $Q_\Omega(s,\omega)$, i.e., the upper layer of SBSs:

$$Q_U(s,\omega,a) = r(s,a) + \gamma \sum_{s'} P(s|s,a) U(\omega, s'), \tag{12}$$

where the function $U$ is option-value function which means the expected cumulative reward when choosing option $\omega$ on state $s'$ and is represented as follows:

$$U(\omega, s') = (1 - \beta_{\omega,\theta_2}(s')) Q_\Omega(s', \omega) + \beta_{\omega,\theta_2}(s') V_\Omega(s'), \tag{13}$$

where the termination function $\beta$ is independent from option-selection function, therefore uses a set of parameters $\theta_2$. The probability of transition in (12) from $(s_t, \omega_t)$ to $(s_{t+1}, \omega_{t+1})$ is represented as follows:

$$P(s_{t+1}, \omega_{t+1}|s_t, \omega_t) = \sum_a \pi_{\omega_t,\theta}(a|s_t)P(s_{t+1}|s_t, a)$$
$$((1 - \beta_{\omega_t,\theta_2}(s_{t+1})\mathbf{1}_{\omega_t=\omega_{t+1}} + \beta_{\omega_t,\theta_2}(s_{t+1})\pi_\Omega(\omega_{t+1|s_{t+1}}). \tag{14}$$

In order to update the parameters of the network, we need the gradient of $G_t$ over parameters $\theta$ of option-selection policies. According to above formulas, it can be concluded that:

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

Similarly, it can be concluded that the gradient of expected cumulative reward $U(s', \omega)$ over parameters $\theta_2$ of termination functions $\beta_{\omega,\theta_2}$ can be represented as follows:

$$\begin{aligned}\frac{\partial U(s',\omega)}{\partial\theta_2} =& -(\sum_{s,\omega}\sum_{t=0}^{\infty}\gamma^t P(s_{t+1} = s', \omega_t = \omega|s_1, \omega_0)) \\ &\cdot(\frac{\partial\beta_{\omega,\theta_2}(s')}{\partial\theta_2})A_\Omega(s',\omega),\end{aligned} \tag{16}$$

where $A_\Omega$ is the advantage function over options $A_\Omega(s', \omega) = Q_\Omega(s', \omega) - V_\Omega(s')$.

In HierCoop, for each agent $k$, there is a local network to represent the option value function $Q^k_\Omega(s^k, \omega^k)$ and the actor policy $\omega_k$ and triplet $(I^k_\omega, \pi^k_\omega, \beta^k_\omega)$. The actor policy consists of intra-option policies, termination functions and policy over options. We obtain $\omega_k$ through $\epsilon$-greedy policy over option-value function $Q^k_\Omega$ and use gradient descent backpropagation algorithm to update option actors.

In order to make agents collaborate with each other instead of making decisions based solely on the local action observation history of each agent, we introduce a centralized a value function $Q_{tot}$ which conditions on the global state and the joint action in HierCoop to optimize the option-value function of each agent. By using a set of hypernetworks and mixing networks, $Q_{tot}$ is supposed to have uniform monotonicity with the local option-value function of agents, i.e.

$$\frac{\partial Q_{tot}}{\partial Q^k_\Omega} \geq 0, k \in 1, \ldots, n. \tag{17}$$

Therefore, each agent $k$ is allowed to participate in a decentralised execution solely by choosing greedy actions with respect to its $Q^k_\Omega$. As shown in Fig. 2, the mixing network takes outputs from the agents as input and mixes them monotonically, producing the values of $Q_{tot}$, with the weights of mixing network being restricted as non-negative. The weights of the mixing network are produced by independent hypernetworks which take global state $S$ as input and generates the weights and the biases of the mixing network.

Then the following formula is used to compute the loss and update the local option-value function in critic using backpropagation methods:

$$\begin{aligned}L(\theta) &= \sum_{i=1}^{b}[(y_i^{tot} - Q_{tot}(S, \Omega; \theta))^2] \\ y_i^{tot} &= r + \gamma max_{\Omega'} Q_{tot}(S', \Omega'; \theta^-),\end{aligned} \tag{18}$$

where $b$ is the batch size of transitions sampled from the replay buffer, $S$ is the global state and $\Omega$ is the set of agents options in corresponding state. Then we can train the critic network using DQN-like method.

In our proposed architecture, we have effectively addressed the challenges of resource allocation in heterogeneous network scenarios. Specifically, regarding the issue of shared spectrum resources and mutual coupling between base stations, the global state needs to be obtained through the collaboration of agents. The policy network module in the upper layer aggregates the environmental status information collected by each base station in a centralized manner, making joint decisions on the selection of options for each base station. In order to maximize the benefits of the entire system, we have introduced a decentralized QMIX algorithm. The models of the QMIX algorithm include agent networks, mixing networks, and hyper networks. The agent network receives observations from the agent as input and outputs the corresponding Q value. The Mixing network simultaneously receives the Q value output by all agent networks and the current global state as input, outputting the behavioral utility value Q of all agents' joint behavior $u$ in the current state $Q_{tot}$. The weights and biases of the middle layer neurons in the Mixing network are generated by the parameter generation network's hyper network receiving the current global state.

Meanwhile, in response to the challenge of diverse communication demand patterns and collaborative resource allocation modes, the lower level of agents is distributed decision-making, where each base station makes specific decisions based on their individual options provided by the upper layer, and updates the network with global rewards after the decision. Therefore, multi-agent hierarchical reinforcement learning algorithm based on option-critic is used to train multiple agents, corresponding to the decision-making module of each SBS. The model structure is shown in Fig. 2.

## V. Simulation Results

This section presents simulation results for validating our proposed resource allocation algorithm using multi-agent hierarchical reinforcement learning for vehicular networks. The parameters and environment definitions follow the urban case evaluation methodology defined in Annex A of 3GPP TR 36.885 [11]. Major simulation parameters and the channel models for V2I links are described in Table I.

During training, each agent shares the network parameters, which means the network can gather information from all agents and update from the data of all agents. At the lower

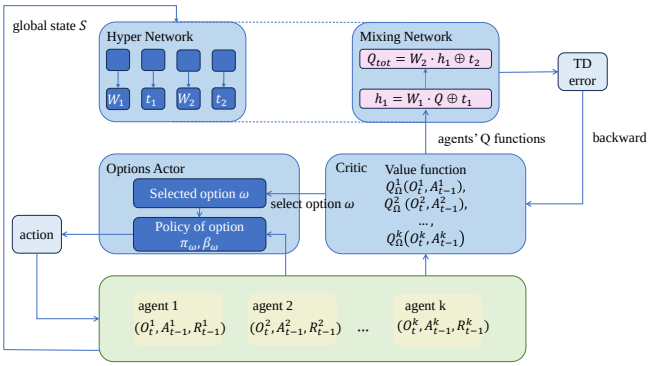


Fig. 2. Network structure of HierCoop

layer of agent, we use RMSProp optimizer with learning rate set to 0.001. We train each agent's network for a total of 7,000 episodes, annealing the exploration rate from 1 to 0.02 linearly over the first 2,400 episodes and keeping it constant thereafter. Additionally, for each training episode, we set the position and large-scale fading, and compute small-scale fading at each step. This aids the network in acquiring more knowledge about underlying fading dynamics and facilitates stable training.

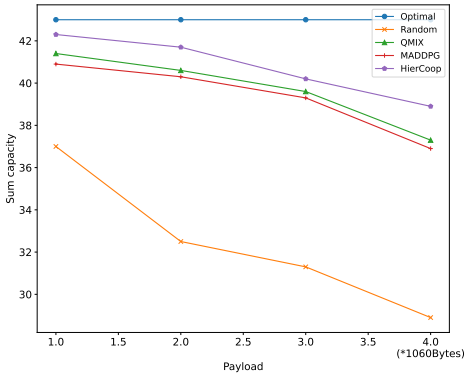


Fig. 3. Sum capacity v.s. payload sizes

In Fig. 3 and Fig. 4, we compare the instantaneous sum capacity and the success probability of payload delivery of proposed multi-agent hierarchical reinforcement learning algorithm with following baselines:

1) Random baseline which chooses randomly in the range of the action space of a base station agent at each time step.

2) The optimal baseline which based on greedy algorithm and allocate resource for agents to ensure maximum sum rate and ignore the interference between V2I links. Although the idealistic baseline cannot be reached in realistic, it provide a upper bound that illustrate how closely other algorithms can approach the limit.

Furthermore, we adopt widely used MADRL algorithm as comparison algorithms:

1) Multi-agent reinforcement learning algorithm QMIX, compared to our proposed algorithm, it does not have

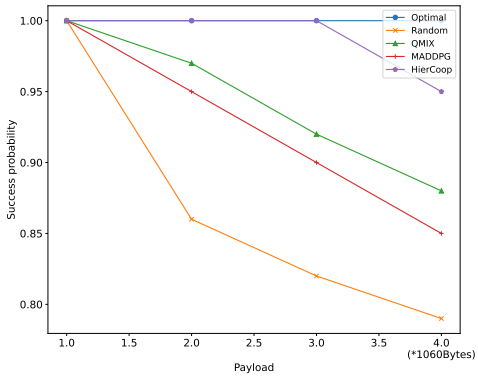


Fig. 4. Success probability v.s. payload size

hierarchical structure like option-critic module and every agent have shared central mixing network to optimize the network of each agent.

2) Multi-agent reinforcement learning algorithm MADDPG, which is based on actor-critic structure, using a centralized critic and decentralized actors to train.

Fig. 3 shows the variation of sum capacity with the payload sizes B increase. It can be indicated from the figure that all algorithms' performance drop with growing payload sizes. Because longer transmission time caused by larger payload will inevitably lead to more conflict between V2I links and thus influence the sum capacity. From Fig. 4, we observe that our proposed HierCoop method achieve better performance than two baselines and two MADRL algorithms, representing the robustness against data packet payload variation.

Fig. 4 shows the variation of success probability of payload delivery with the payload sizes B increase. It can be indicated from the figure that all algorithms' performance drop with growing payload sizes, except that the optimal baseline can achieve 100 percent packet delivery throughout the tested cases. The performance of our proposed HierCoop method achieves significantly better performance than other two baseline and MADRL methods and is close to 100 percent success rate.

In Fig. 5, we show the rewards of every training episode with increasing training iterations to compare the convergence behavior of the above methods. It can be indicated from the figure that the cumulative rewards per episode improve faster as training continues and rewards of proposed algorithm is more than others', demonstrating the efficiency of our method, although there are some fluctuations due to mobility-induced channel fading in vehicular environments.

In order to observe the changes in the policy of our proposed HierCoop algorithm with communication mode, we conducted further experiments. Fig. 6 shows the changes of the output policy from the upper layer to the lower layer under different V2I packet loads. The figure shows the difference of distributions of the upper-level strategy, as delineated by the fluctuations of packet load $B$.

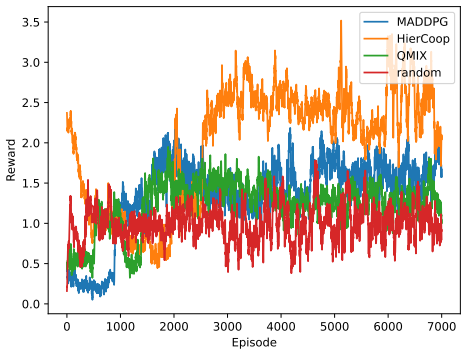


Fig. 5. Rewards

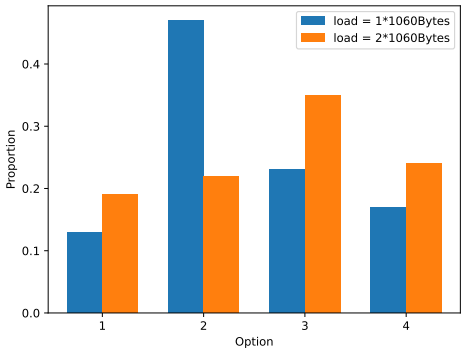


Fig. 6. Option distribution

## VI. Conclusion

In this paper, we have developed a resource allocating algorithm based on hierarchical MADRL for vehicular networks with heterogeneous base stations. By modeling SBSs as agents with hierarchical structure, we combine coarse-grained collaboration policies between base stations and fine-grained resource allocation within base stations to deal with the challenges. We also adopt the centralized training and distributed implementation framework. Finally, we prove the efficiency and the significance of our algorithm compared to other commonly used MADRL methods. Future work may do further research on resource allocation algorithm in HetNets with more complex environment and add V2V links to environment settings to make the scenarios more general.

TABLE I
Simulation Parameters

| Parameter | Value |
|---|---|
| Number of V2I links M | 9 |
| Carrier frequency | 2GHz |
| Bandwidth | 4MHz |
| MBS antenna height | 25m |
| MBS antenna gain | 20dBi |
| MBS antenna receiver noise figure | 5dB |
| SBS antenna height | 6m |
| SBS antenna gain | 5dBi |
| SBS antenna receiver noise figure | 5dB |
| Vehicle antenna height | 1.5m |
| Vehicle antenna gain | 3dBi |
| Absolute vehicle speed v | 15m/s |
| V2I transit power $P^c$ | 23dBm |
| Noise power $\sigma^2$ | -114dBm |

## References


[1] Qin, P., Fu, Y., Tang, G., Zhao, X., Geng, S. (2022). Learning based energy efficient task offloading for vehicular collaborative edge computing. IEEE Transactions on Vehicular Technology, 71(8), 8398-8413.

[2] Alwarafy, Abdulmalik, et al. "The frontiers of deep reinforcement learning for resource management in future wireless HetNets: Techniques, challenges, and research directions." IEEE Open Journal of the Communications Society 3 (2022): 322-365.

[3] Hossain, Ekram, et al. "Evolution toward 5G multi-tier cellular wireless networks: An interference management perspective." IEEE Wireless communications 21.3 (2014): 118-127.

[4] Sinan Nasir, Yasar, and Dongning Guo. "Deep Reinforcement Learning for Joint Spectrum and Power Allocation in Cellular Networks." arXiv e-prints (2020): arXiv-2012.

[5] Zhang, Xiu, Xin Zhang, and Zhou Wu. "Utility-and fairness-based spectrum allocation of cellular networks by an adaptive particle swarm optimization algorithm." IEEE Transactions on Emerging Topics in Computational Intelligence 4.1 (2020): 42-50.

[6] Zhang, Yuexia, and Ying Zhou. "Resource allocation strategy based on tripartite graph in vehicular social networks." IEEE Transactions on Network Science and Engineering (2022).

[7] Qian, Bo, et al. "Leveraging dynamic stackelberg pricing game for multi-mode spectrum sharing in 5G-VANET." IEEE Transactions on Vehicular Technology 69.6 (2020): 6374-6387.

[8] Zhu, Kun, Ekram Hossain, and Alagan Anpalagan. "Downlink power control in two-tier cellular OFDMA networks under uncertainties: A robust Stackelberg game." IEEE Transactions on Communications 63.2 (2014): 520-535.

[9] Kim, Juyeop, and Dong-Ho Cho. "A joint power and subchannel allocation scheme maximizing system capacity in indoor dense mobile communication systems." IEEE Transactions on Vehicular Technology 59.9 (2010): 4340-4353.

[10] Abdelnasser, Amr, Ekram Hossain, and Dong In Kim. "Tier-aware resource allocation in OFDMA macrocell-small cell networks." IEEE Transactions on Communications 63.3 (2015): 695-710.

[11] Liu, R., Guo, K., An, K., Zhou, F., Wu, Y., Huang, Y., Zheng, G. (2023). Resource allocation for NOMA-enabled cognitive satellite-UAV-terrestrial networks with imperfect CSI. IEEE Transactions on Cognitive Communications and Networking.

[12] Alwarafy, Abdulmalik, et al. "The frontiers of deep reinforcement learning for resource management in future wireless HetNets: Techniques, challenges, and research directions." IEEE Open Journal of the Communications Society 3 (2022): 322-365.

[13] Liang, Le, Hao Ye, and Geoffrey Ye Li. "Spectrum sharing in vehicular networks based on multi-agent reinforcement learning." IEEE Journal on Selected Areas in Communications 37.10 (2019): 2282-2292.

[14] Vu, Hung V., et al. "Multi-Agent Reinforcement Learning for Channel Assignment and Power Allocation in Platoon-Based C-V2X Systems." arXiv preprint arXiv:2011.04555 (2020).

[15] Gündoğan, Alperen, et al. "Distributed resource allocation with multi-agent deep reinforcement learning for 5G-V2V communication." Proceedings of the Twenty-First International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing. (2020): 357-362.

[16] Bacon, Pierre-Luc, Jean Harb, and Doina Precup. "The option-critic architecture." Proceedings of the AAAI conference on artificial intelligence. Vol. 31. No. 1. 2017.

[17] Riemer, Matthew, Miao Liu, and Gerald Tesauro. "Learning abstract options." Advances in neural information processing systems 31 (2018).

[18] Schaul, Tom, et al. "Universal value function approximators." International conference on machine learning. PMLR, 2015.