electronics

Article

Multi-Agent Reinforcement Learning for Efficient Resource Allocation in Internet of Vehicles

Jun-Han Wang [1,*], He He [1], Jaesang Cha [1], Incheol Jeong [2] and Chang-Jun Ahn [1]

[1] Graduate School of Engineering, Chiba University, Chiba 263-8522, Japan; chajaesang@gmail.com (J.C.); junny@faculty.chiba-u.jp (C.-J.A.)
[2] Department of Computer and Communication Engineering, Sungkonghoe University, Seoul 08359, Republic of Korea; jeong@skhu.ac.kr
* Correspondence: cdda5211@chiba-u.jp

Abstract: The Internet of Vehicles (IoV), a burgeoning technology, merges advancements in the internet, vehicle electronics, and wireless communications to foster intelligent vehicle interactions, thereby enhancing the efficiency and safety of transportation systems. Nonetheless, the continual and high-frequency communications among vehicles, coupled with regional limitations in system capacity, precipitate significant challenges in allocating wireless resources for vehicular networks. In addressing these challenges, this study formulates the resource allocation issue as a multi-agent deep reinforcement learning scenario and introduces a novel multi-agent actor-critic framework. This framework incorporates a prioritized experience replay mechanism focused on distributed execution, which facilitates decentralized computing by structuring the training processes and defining specific reward functions, thus optimizing resource allocation. Furthermore, the framework prioritizes empirical data during the training phase based on the temporal difference error (TD error), selectively updating the network with high-priority data at each sampling point. This strategy not only accelerates model convergence but also enhances the learning efficacy. The empirical validations confirm that our algorithm augments the total capacity of vehicle-to-infrastructure (V2I) links by 9.36% and the success rate of vehicle-to-vehicle (V2V) transmissions by 6.74% compared with a benchmark algorithm.

Keywords: IoV; reinforcement learning; prioritized experience replay; power control; resource allocation

Academic Editor: Hung-Yu Chien

Received: 10 December 2024
Revised: 1 January 2025
Accepted: 3 January 2025
Published: 5 January 2025

Citation: Wang, J.-H.; He, H.; Cha, J.; Jeong, I.; Ahn, C.-J. Multi-Agent Reinforcement Learning for Efficient Resource Allocation in Internet of Vehicles. *Electronics* **2025**, *14*, 192. https://doi.org/10.3390/electronics14010192

1. Introduction

Vehicles, fundamental to modern transportation systems, provide substantial comfort and convenience [1–4]. Yet, the recent surge in vehicle numbers has highlighted significant challenges, particularly in traffic safety and the integration of advanced driving technologies. To address these issues, an intelligent transportation system (ITS) has been advocated, leveraging sophisticated control, sensing, and communication technologies alongside computer systems to enhance the overall management of transportation networks [5,6]. Within this framework, the IoV has emerged as a critical component, demanding immediate integration within various sectors to ensure robust, reliable communication and an enhanced user experience. In this context, the allocation of resources plays a pivotal role. By allocating resources effectively, the stability and reliability of IoV systems can be substantially improved [7].

Conventional optimization approaches have traditionally been employed in designing algorithms for power control to address the issues of resource distribution and interference management within vehicular communications. The authors of [8] explored energy efficiency (EE) enhancement via adaptive power control in a cognitive vehicular network under simulated attacks by authorized users. The environment surrounding these users was segmented into three zones, with unauthorized users employing varied communication techniques depending on their proximity. Nevertheless, this study failed to account for the high mobility characteristic of vehicles. In [9], the design of a three-slot one-way system for vehicular communications was presented, aiming to boost the efficacy of cooperative systems in non-orthogonal, cognitive vehicular self-organizing networks, achieving optimal power control by reducing the probability of a system outage. However, this research is limited as it focused solely on vehicle-to-vehicle (V2V) interactions. Further extending the network, the authors of [10] incorporated satellites into the cognitive vehicle network, creating a cognitive satellite-vehicle framework. This approach examined the spectral efficiency (SE) and EE of communications within vehicles, introducing an EE-SE trade-off index with a preference factor to refine power allocation strategies. Although this setting mirrors more realistic vehicular communication scenarios, the proposed algorithms fell short in adapting to the dynamic nature of these environments. To better suit actual communication scenarios, certain scholars have suggested modeling cognitive vehicular networking based on cellular orthogonal frequency division multiple access (OFDMA), optimizing both subcarrier distribution and transmission power to maximize the total system capacity [11]. Moreover, to enhance energy conservation, these models integrate energy harvesting techniques, utilizing RF energy from the network of authorized users to energize unauthorized users and optimizing energy utilization through a coordinated strategy of time and power allocation [12]. However, the issue of multi-user coexistence remained unaddressed in these studies.

Despite significant research applying conventional optimization strategies to resource allocation in the IoV [13], substantial challenges remain unresolved. The rapid variation in channel conditions due to high vehicle mobility creates profound uncertainties in resource distribution. For example, errors in estimating channel state information can lead to notable performance degradation [14]. Additionally, as the deployment of services like onboard entertainment and autonomous driving expands, the demand for enhanced in-vehicle communication grows. Addressing these requirements while ensuring the robustness of vehicular communication systems is a complex issue, complicating accurate and straightforward mathematical modeling.

Machine learning (ML), particularly deep reinforcement learning (DRL), has emerged as a robust solution to overcome the inherent limitations of traditional optimization methods in resource allocation [15–17], which include slow convergence rates, excessive iteration requirements, and delayed processing of information. In vehicular network environments, which are characterized by high levels of uncertainty and variability, DRL has demonstrated superior performance [18]. Algorithms based on the deep Q-network (DQN) framework specifically address challenges such as dynamic topologies and evolving spectral states by learning optimal scheduling strategies within cognitive, radio-enhanced vehicular networks [19]. The authors of [20] employed a strategy of centralized learning with distributed execution to enhance resource selection through a multi-agent reinforcement learning approach. Additionally, a decentralized approach using a DQN algorithm optimized the sub-band and power level selections in various V2X communication scenarios, as discussed in [21]. To ensure the reliable delivery of safety-critical messages over V2V links, the authors of [22] introduced a semi-decentralized DRL algorithm which operates on dual time scales to fine-tune the selection of transmission modes and resources.

Although DRL offers a significant advantage in performance and complexity over traditional methods, it faces challenges in real-time decision making within V2X communications. The primary technique, the DQN, addresses the discrete-continuous hybrid action spaces through discrete actions, which may introduce errors and degrade performance. Additionally, optimizing power control across multiple users to enhance overall network performance is critical; overlooking this in multi-user scenarios can result in an unbalanced resource distribution and reduced efficiency. To manage continuous action spaces more effectively, the authors of [23] proposed a parameterized deep Q-network framework which categorizes actions hierarchically, avoiding the pitfalls of continuous action discretization. Utilizing the actor-critic framework, the authors of [24] introduced a power control strategy which employs the asynchronous advantage actor-critic (A3C) algorithm and distributed proximal policy optimization (DPPO), which addresses power discretization issues effectively. Despite these advancements, the centralized nature of the solutions in [23,24] can lead to significant overhead and suboptimal resource allocation in dynamic V2X communication environments. Furthermore, DRL-based algorithms often operate under the assumption that training and operational environments are static, a premise which does not hold in the highly mobile and variable conditions typical of V2X scenarios [25,26]. Nevertheless, recent works have further explored advanced resource allocation techniques to tackle the inherent complexities of V2V communications in highly dynamic settings. For instance, an intelligent resource allocation scheme focusing on maximizing both spectral and energy efficiency was proposed in [27], while a hybrid NOMA approach combined with multi-agent reinforcement learning demonstrated promising performance in vehicular network spectrum allocation [28]. In addition, the authors of [29] proposed a Q-learning network combined with long short-term memory (LSTM) which utilizes hysteretic Q-learning and concurrent experience replay trajectories to solve the non-smoothness problem caused by concurrent learning of multiple intelligences. Although these studies reveal the potential of modern DRL-based solutions to overcome the critical limitations of real-time V2X communications, several challenges remain underexplored, particularly in optimizing power control and addressing large-scale multi-user scenarios. Consequently, the complexity arising from dynamic topologies, heterogeneous traffic demands, and stringent latency requirements calls for more robust and decentralized learning frameworks which maintain high performance in ever-changing vehicular environments.

To address the power selection challenge in V2X spectrum access, the multi-agent reinforcement learning approach is employed. The analysis begins with the Internet of Vehicles (IoV) communication system model, where training rewards are designed based on the system state, action space, and the ultimate optimization objective. Unlike conventional resource allocation strategies based on single-agent reinforcement learning, this study proposes a distributed reinforcement learning algorithm, where each vehicle is considered an individual agent. Through continuous interactions with its local environment, each agent independently learns and improves its resource allocation strategy to achieve optimal power control. The key contributions of this paper are outlined as follows:

(1) This paper proposes a resource allocation algorithm for vehicular networks which utilizes multi-agent deep reinforcement learning, specifically designed for the communication scenario where V2V links are multiplexed with V2I links. The primary goal is to enhance the success rate of V2V transmissions and increase the capacity of the V2I channel. The proposed algorithm employs a multi-agent reinforcement learning framework, where each vehicle is treated as an individual agent. This framework accounts for the interference and collaborative dynamics between vehicles in the power allocation process, thereby improving the overall communication performance in intelligent transportation systems.

(2) The resource allocation problem is modeled as a Markov decision process and addressed using the multi-agent deep deterministic policy gradient (MADDPG) algo-

rithm. In this approach, each vehicle is treated as an individual agent, and a weighted reward function is constructed to balance the completion rate of V2V link transmissions with the overall capacity of V2I link transmissions. By maximizing the expected cumulative reward, the algorithm updates its strategy to determine the optimal resource allocation decisions.

(3) In this work, a new algorithm, referred to as the prioritized experience replay for multi-agent cooperation (PER-MAC), is introduced by integrating the attention mechanism with prioritized experience replay. Within the multi-agent framework, the algorithm determines the priority of each experience based on the temporal difference (TD) error, with higher-priority experiences being selected to update the network during each sampling. By enhancing the quality of the training data, the PER-MAC algorithm accelerates the convergence of the model and improves the agents' cooperation performance. Additionally, the algorithm employs a critic network which incorporates the attention mechanism, enabling the agents to dynamically select relevant surrounding information. This dynamic selection process contributes to improving the scalability of the multi-agent system.

The structure of this paper is as follows. In Section 2, a simulation environment based on a Manhattan grid is created, and the resource allocation challenge is reformulated as a mathematical optimization model. Section 3 provides a comprehensive explanation of the proposed PER-MAC resource allocation algorithm. Section 4 presents a comparison of the performance of the proposed approach with other existing methods. Section 5 analyzes the limitations and proposes directions for future research. Finally, Section 6 concludes the paper with a summary of the key findings.

2. System Model

This study proposes a continuous, action-based power transmission algorithm specifically designed for V2V links. The primary objective is to enhance the channel capacity and transmission success rate of both V2I and V2V links, particularly in high-mobility vehicular environments. The V2V link is responsible for transmitting periodically generated security data through device-to-device (D2D) communication, while the V2I link utilizes cellular technology to establish a connection between the vehicle and the base station for high data rate services.

2.1. Channel Gain and Path Loss

In the proposed scenario, it is assumed that the region accommodates M V2I links and N V2V links, with $M = \{1, 2, \ldots, m\}$ and $N = \{1, 2, \ldots, n\}$ denoting the respective sets of links. Furthermore, it is assumed that each V2I link, being responsible for data upload, has been allocated a specific orthogonal sub-band, with the mth V2I link occupying the corresponding sub-band s.

In wireless communication, channel gain is the combined attenuation due to path loss and multipath effects when the signal is transmitted from the transmitter to the receiver. The channel gain for the nth V2V link, corresponding to the mth V2I link's frequency spectrum, is determined based on the coherence time specified in 3GPP TR 36.885 [30]:

$$g_n[m] = \delta_n h_n[m]. \tag{1}$$

With this equation, the channel characteristics between vehicles can be modeled. Here, δ_n and $h_n[m]$ denote the components of large-scale fading and small-scale fading, respectively, which encompass both the shadowing effects and path loss. The path loss for large-scale fading is modeled using two distinct path loss models: line of sight (LOS) and non-line of sight (NLOS). The expressions for these two path loss models, specifically for the V2V link, are provided in [30]:

$$P^V_{LOS} = \begin{cases} 22.7log_{10}(3) + 20log_{10}(\frac{f_c}{5}) + 41, & d < 3 \\ 22.7log_{10}(d) + 20log_{10}(\frac{f_c}{5}) + 41, & d < d'_{BP}, \\ 40log_{10}(d) - 17.3log_{10}(h'_{BS}h'_{MS}) + 2.7log_{10}(\frac{f_c}{5}) + 9.45, & else \end{cases} \tag{2}$$

$$P^V_{NLOS}(d_1, d_2) = P^V_{LOS}(d_1) + 10n_j log_{10}(d_2) + 20 - 12.5n_j + 3log_{10}(\frac{f_c}{5}). \tag{3}$$

This formula is based on the 3GPP standard path loss model for describing the line-of-sight scenario path loss for signal propagation in a typical urban environment. Here, f_c denotes the center frequency, while h'_{BS} and h'_{MS} represent the heights of the base station and mobile station, respectively. The variables d_1 and d_2 correspond to the horizontal and vertical distance differences, respectively, and the total distance is defined as $d = \sqrt{d_1^2 + d_2^2}$. The breakpoint distance is represented by d'_{BP}, and n_j is a parameter used to adjust the path loss index, reflecting the propagation characteristics of non-line-of-sight distances at varying ranges. The expressions for d'_{BP} and n_j are as follows:

$$d'_{BP} = 4\frac{(h'_{BS} - 1)(h'_{MS} - 1)f_c}{c}, \tag{4}$$

$$n_j = max(2.8 - 0.0024d_2, 1.84). \tag{5}$$

Among these, c represents the speed at which light propagates through free space. The final path loss is determined dynamically, with the model selecting between the LOS or NLOS path loss models based on the distance between the transmitter and the receiver:

$$PL^V = \begin{cases} P^V_{LOS}, & min(d_1, d_2) \\ min(P^V_{NLOS}(d_1, d_2), P^V_{NLOS}(d_2, d_1)), & else \end{cases}. \tag{6}$$

Moreover, as specified in 3GPP TR 36.855, within the region, only the LOS path loss model is considered for V2I links. This means that no NLOS path loss is accounted for in the communication between V2I links in the area:

$$PL^I = 128.1 + 37.6 + log_{10}(\sqrt{d^2 + \frac{(h'_{BS} - h'_{MS})^2}{1000}}). \tag{7}$$

Alternatively, the shadowing effect is modeled in accordance with 3GPP TR 36.885, and this is expressed through the following equations:

$$S(n) = e^{-\frac{\Delta d}{D}S(n-1)} + S^*(n)\sqrt{1 - e^{-2\frac{\Delta d}{D}}}, \tag{8}$$

where $S^*(n)$ represents a matrix of independent and uniformly distributed normal variables. It is assumed that the shadowing effects between two vehicles are reciprocal, meaning that they are the same in both directions. The parameter Δd denotes the distance between the front and rear of the vehicles, taking into account the positions of each vehicle after a time gap. The parameter D, which represents the decorrelation distance, varies between the V2I and V2V links. For small-scale fading, the assumption is made that the mean is exponentially distributed, which simplifies the model construction process.

2.2. Performance Indicators

In vehicular networks, due to the high mobility of vehicles and the dynamic nature of road environments, the interference and noise conditions become more complex. The analysis and assurance of the signal-to-interference-plus-noise ratio (SINR) are particularly critical for system performance. Generally, a higher SINR indicates that the target signal received at the receiver is stronger relative to the interference and noise, leading to higher communication reliability and a lower bit error rate. The SINR for the mth V2I link on the sth spectral sub-band can be expressed as follows:

$$\varphi_m^I[s] = \frac{P_m^I \hat{g}_{m,B}[s]}{\sigma^2 + \sum_s \xi_n[s] P_n^V g_{n,B}[s]}. \tag{9}$$

In this expression, P_m^I represents the transmission power of the mth V2I link, while P_n^V denotes the transmission power of the nth V2V link. The parameter σ^2 accounts for the noise present in the environment. A binary variable $\xi_n[s]$, which can take values of one or zero, is used to indicate whether the sth spectral sub-band is occupied by the V2V link. The terms $\hat{g}_{m,B}[s]$ and $g_{n,B}[s]$ denote the sum of the channel gains for the mth V2I link and the nth V2V link, respectively, with both gains calculated from the transmitter to the base station on the sth sub-band.

The SINR of the nth V2V link over the sth spectral sub-band can be expressed as follows:

$$\varphi_n^V[s] = \frac{P_n^V g_n[s]}{\sigma^2 + I_n[s]}, \tag{10}$$

$$I_n[s] = P_m^I \hat{g}_{m,n}[s] + \sum_{n' \neq n} \xi_{n'[s]} P_{n'}^V[s] g_{n',n}[s]. \tag{11}$$

In this expression, $g_n[s]$ refers to the channel gain between the transmitter of the V2V link and the receiver of the target vehicle, utilizing the sth spectral sub-band. The term $I_n[s]$ represents the total interference, which includes contributions from the V2I link transmitter which shares the same spectral sub-band and the transmitters of other V2V links. Additionally, $g_{n',n}[s]$ denotes the combined channel gain of the V2V link n' as it travels through sub-band s to reach the receiver of the nth V2V link.

The channel capacity of both the V2I and V2V links can be determined based on Shannon's theorem, as defined in the following:

$$C_m^I[s] = Wlog_2(1 + \varphi_m^I[s]), \tag{12}$$

$$C_n^V[s] = Wlog_2(1 + \varphi_n^V[s]), \tag{13}$$

where W denotes the bandwidth of the spectral sub-band.

As previously outlined, the primary objective of this study is to optimize the total channel capacity of the V2I link. This is achieved by simultaneously selecting the appropriate spectral sub-bands and adjusting the transmission power, while ensuring that the data transmission success rate over the V2V link remains high. Therefore, a suitable approach is to maximize the overall capacity, which is represented by the summation $\sum_s C_m^I[s]$. The success rate of data transmission in V2V links is evaluated by comparing the transmitted data volume within a specific time delay to a predefined threshold. This leads to the following formulation:

$$Pr\left\{ \sum_{t=1}^{T} \sum_{s=1}^{S} \xi_n[s] C_n^V[s,t] \geq \frac{B_n}{\triangle T} \right\}, \tag{14}$$

In this equation, B_n represents the predetermined amount of secure data to be transmitted periodically in bits, while $\triangle T$ denotes the channel coherence time. At this point, the two performance indicators have been simply derived, and after balancing, they can be incorporated within the multi-agent reinforcement learning training process in the form of a reward function.

3. Proposed Power Control Algorithm

Equation (21) involves sequential decision making over multiple coherent time slots, all constrained by the time limit T. The exponential complexity of this process presents a significant challenge for traditional algorithms. To address this issue, the problem is reformulated as a Markov decision process. In response, this paper introduces a distributed algorithm for controlling the transmit power of V2V links, leveraging the MAAC framework. The proposed solution is structured in two distinct phases: a centralized training phase and a distributed execution phase [31]. During the training phase, agents collect system rewards, after which they train the critic and actor networks using a centralized learning approach. In the execution phase, each agent, based on the received information, selects the action to perform by utilizing its trained actor network. The specific network structure is shown in Figure 1.

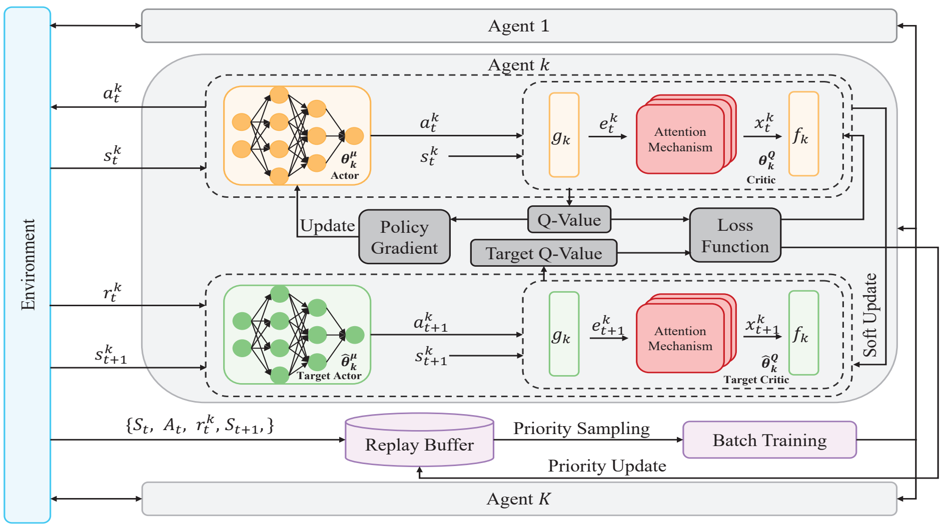


Figure 1. Network architecture and data flow of the proposed PER-MAC method.

3.1. The Environmental Status

Assume that the state space of multiple agents is S_t. When each agent obtains its respective action, these actions form a joint action A_t. Then, the agents collectively perform the action and obtain the reward R_t. The state S_t^k in which agent k is at time t is transferred with probability $Pr(S_t^k \mid S_{t+1}^k, A_t^k)$ to the state S_{t+1}^k at the next moment. An individual agent k only has access to the state S_{t+1} and action A_{t+1} of its local environment. In contrast, the global channel conditions and the actions of other agents are unknown. For ease of understanding, we assume that the kth agent occupies the kth V2V link. The local channel information observed by a single agent k includes its channel gain $g_k[s]$; the interference channel gain $g_{k',k}[s]$ from other V2V link transmitters; the interference channel gain $g_{k,B}[s]$ from its transmitter to the base station; and the interference channel gain $\hat{g}_{m,k}[s]$ from all V2I transmitters. Therefore, the set of states observed by agent k at the current time step can be represented by

$$S_t^k = \{g_k[s], g_{k',k}[s], g_{k,B}[s], \hat{g}_{m,k}[s]\}. \tag{15}$$

3.2. Action Space

As previously described, the resource allocation challenge in vehicular networks is modeled as the continuous control of V2V link transmit power. The authors of [32,33] showed that the transmit power of a V2V link transmitter is a continuous variable, which takes values in the range of $0 \le P_k^V[s] \le P_{max}^V$. Thus, the action decision of the agent k can be expressed as follows:

$$A_t^k = \left\{ P_k^V[s] \right\}, \tag{16}$$

where $P_k^V[s]$ denotes the transmission power of the nth V2V link on the sth spectral sub-band.

3.3. Reward Function

The reward function not only defines the learning goals of the agent but also guides the agent to adjust and optimize its behavioral strategies through the feedback mechanism. Considering that one of the objectives is the data transmission success rate over V2V links, for each agent k, we assign the reward R_k to the transmission rate of the V2V link [34]. When the agent completes all of the data transfer tasks, we let the reward be a constant β greater than the maximum V2V link transfer rate. Therefore, the reward for successful delivery of the V2V link payload at time step t is set to be

$$R_k(t) = \begin{cases} \sum_{s=1}^{S} \xi_k[s] C_k^V[s,t], & B_k \ge 0 \\ \beta, & B_k < 0 \end{cases}, \tag{17}$$

where B_k denotes the remaining data to be transmitted on the kth V2V link. Additionally, the total capacity of the V2I link between all agents and the base station at each time step is also utilized as a reward value:

$$R_m(t) = \sum_{s=1}^{S} C_m^I[s,t]. \tag{18}$$

For the agent to better achieve the intended learning effect, the two reward functions are finally weighted and combined:

$$R_t = \lambda R_k(t) + (1-\lambda) R_m(t). \tag{19}$$

3.4. The Structure of PER-MAC

The multi-agent deep deterministic policy gradient (MADDPG) algorithm [35] is a centralized training approach applied to distributed execution environments, where each agent possesses its own critic and actor networks. This method is often employed in both cooperative and competitive multi-agent settings. However, as the number of agents increases, the input dimensions of the centrally trained critic network expand significantly, causing challenges in convergence. The MAAC algorithm [31] addresses this limitation by introducing an attention mechanism to the centralized critic network. This allows each agent to dynamically select relevant information at each time step, partially resolving the issue.

PER-MAC utilizes the same structure, where each agent has its independent actor network with parameters θ_k^μ. It calculates the action a_t^k based on the current observation s_t^k of the agent k. The critic network of agent k outputs a Q-value to evaluate the goodness of action a_t^k, improving the actor network's performance. The Q-value of the critic network is denoted by

$$Q_k(S_t, A_t|\theta_k^Q) = f_k(g_k(s_t^k, a_t^k), x_t^k), \tag{20}$$

where g_k is a single-layer perceptron embedding function and f_k is a two-layer multi-layer perception (MLP). A_t and S_t denote the joint action and joint state of all of the agents at time step t. At the beginning of training, the current observation s_t^k and the action a_t^k of the agent k are inputted into the embedding function to obtain the embedding e_t^k of the agent k. Then, the embedding of other agents is subjected to multi-head self-attention to find the sum of each agent's contribution:

$$x_t^k = \sum_{k' \neq k} w_{k'} v_{k'} = \sum_{k' \neq k} w_{k'} h(V e_t^{k'}), \tag{21}$$

where V represents the shared matrix of all agents and h is the nonlinear activation function, while $v_{k'}$ can be regarded as a function of agent k''s embedding. The attention weight $w_{k'}$ from agent k to agent k' is obtained by comparing the similarity between the encoded information e_t^k and $e_t^{k'}$ as shown in Equation (22). W_q and W_k are the linear transformation matrices of e_t^k and $e_t^{k'}$, respectively:

$$w_j \propto exp((e_t^{k'})^T W_k^T W_q e_t^k). \tag{22}$$

Then, the target Q-value for agent k is given by the following:

$$\begin{aligned} y_t^k = r_t^k + \gamma \mathbb{E}_{A_{t+1} \sim \pi_{\hat{\theta}_k^Q}(S_{t+1})} \\ [\hat{Q}_k(S_{t+1}, A_{t+1}) - \alpha log \pi_{\hat{\theta}_k^Q}(a_{t+1}^k | s_{t+1}^k)], \end{aligned} \tag{23}$$

where γ is a discount factor in the range of $[0,1]$, which balances the significance of immediate and future rewards. Here, r_t^k represents the immediate reward acquired by the agent k at time step t, π represents the agent's strategy for choosing actions in different states, and α represents the entropy parameter which determines the balance between diversification and reward maximization of the strategy. Then, the loss function is defined by the Q-value and the target Q-value, which is the TD error:

$$J(\theta_k^Q) = \mathbb{E}_{(S_t, A_t, r_t^k, S_{t+1}) \sim D}[w_i (Q_k(S_t, A_t | \theta_k^Q) - y_t^k)^2], \tag{24}$$

where D is a replay buffer which stores past experiences and w_i is the importance sampling weight, which is used to correct the bias due to non-uniform sampling. Its calculation formula is as follows:

$$w_i = (\frac{1}{|D|} \cdot \frac{1}{P_{t,i}})^\rho, \tag{25}$$

where $|D|$ denotes the size of the experience pool, $P_{t,i}$ denotes the probability that the ith experience is sampled at moment t, and ρ controls the strength of the correction, with $\rho = 1$ denoting full correction. $P_{t,i}$ can be expressed as follows:

$$P_{t,i} = \eta \frac{p_i^\zeta}{\sum_d^{|D|} p_d^\zeta} + \varphi, \tag{26}$$

$$p_i = |J(\theta_k^Q)| + \varepsilon. \tag{27}$$

The parameter ζ determines the extent to which prioritization affects the sampling process. When ζ equals zero, the sampling process becomes uniform. The term p_i refers to the priority associated with the ith experience. The combination of parameters η and φ incorporates elements from both the greedy preference algorithm and random selection, ensuring that the probability of updating experience tuples remains monotonic. Furthermore, this combination guarantees that experiences with higher priority are more likely to update the experience playback cache pool, while those with lower priority still have a chance to be included, albeit with a reduced probability. To prevent the sampling probability of experience tuples from approaching zero as the TD error nears zero, the positive constant ε is introduced. This ensures that even experiences with minimal TD errors maintain a nonzero sampling probability.

The actor network gradient formula for agent k is defined as follows:

$$\begin{aligned}&\bigtriangledown \theta_k^\mu J(\theta_k^\mu) = \mathbb{E}_{S_t \sim D, A_t \sim \pi} \\ &[\bigtriangledown \theta_k^\mu log(\pi_{\theta_k^\mu}(a_t^k|s_t^k))(Q_k(S_t, A_t) - \alpha log(\pi_{\theta_k^\mu}(a_t^k|s_t^k)))].\end{aligned} \tag{28}$$

To stabilize the learning, the target actor network and the target critic network of agent k are constructed with parameters $\hat{\theta}_k^\mu$ and $\hat{\theta}_k^Q$, respectively, and the parameters are soft updated:

$$\hat{\theta}_k^\mu = \tau \theta_k^\mu + (1-\tau)\hat{\theta}_k^\mu, \tag{29}$$

$$\hat{\theta}_k^Q = \tau \theta_k^Q + (1-\tau)\hat{\theta}_k^Q, \tag{30}$$

where τ is the soft update factor. Algorithm 1 summarizes the training process of the algorithm.

Algorithm 1 The proposed PER-MAC algorithm.

Input: Environmental status information s_t

Output: Parameters of the trained actor and critic networks θ_k^μ, θ_k^Q

Initialization:
Initialize each agent k's actor network θ_k^μ and critic network θ_k^Q, along with their target networks $\hat{\theta}_k^\mu$ and $\hat{\theta}_k^Q$.
Initialize the global replay buffer D.
Set hyperparameters for prioritized experience replay $(\rho, \eta, \zeta, \varphi, \varepsilon)$ and optional attention mechanism parameters.
for each episode **do**
    for each time step t **do**
        1. Interact with the environment:
        for each agent k **do**
            obtain current status information s_t.
            Select action $a_t^k = \theta_k^\mu(s_t^k)$ based on its policy and execute it.
            Interact with the environment to obtain reward r_t^k and next state s_{t+1}^k.
        end for
        Store (S_t, A_t, R_t, S_{t+1}) into replay buffer D.
        2. Sample prioritized experiences:
        Sample a batch of experiences (s_t, a_t, r_t, s_t+1) from D using priorities $P_{t,i}$:
        Calculate sampling probability: $P_{t,i} = \eta \frac{p_i^\zeta}{\sum_d^{|D|} p_d^\zeta} + \varphi$.
        Compute importance sampling weight: $w_i = (\frac{1}{|D|} \cdot \frac{1}{P_{t,i}})^\rho$.
        3. Update critic networks:
        for each agent k **do** Calculate the target action a_{t+1}^k
            Process joint information with attention: $w_j \propto exp((e_t^{k'})^T W_k^T W_q e_t^k)$.

Algorithm 1 *Cont.*

Calculate Q-value and target Q-value:

$$Q_k(S_t, A_t|\theta_k^Q) = f_k(g_k(s_t^k, a_t^k), x_t^k),$$

$$y_t^k = r_t^k + \gamma \mathbb{E}_{A_{t+1} \sim \pi_{\hat{\theta}_k^Q}(S_{t+1})}[\hat{Q}_k(S_{t+1}, A_{t+1}) - \alpha log \pi_{\hat{\theta}_k^Q}(a_{t+1}^k|s_{t+1}^k)].$$

Compute TD error: $J(\theta_k^Q) = \mathbb{E}_{(S_t, A_t, r_t^k, S_{t+1}) \sim D}[w_i(Q_k(S_t, A_t|\theta_k^Q) - y_t^k)^2]$.

Update θ_k^Q's parameters using gradient descent.

end for

4. Update actor networks:

for each agent k **do** Calculates the current actor's action a_t^k

Actor loss: $\bigtriangledown\theta_k^\mu J(\theta_k^\mu) = \mathbb{E}_{S_t \sim D, A_t \sim \pi}[\bigtriangledown\theta_k^\mu log(\pi_{\theta_k^\mu}(a_t^k|s_t^k))(Q_k(S_t, A_t) - \alpha log(\pi_{\theta_k^\mu}(a_t^k|s_t^k)))]$.

Update θ_k^μ's parameters using gradient ascent.

end for

5. Update target networks:

for each agent k **do**

Perform soft updates for actor and critic target networks:

$$\hat{\theta}_k^\mu = \tau\theta_k^\mu + (1-\tau)\hat{\theta}_k^\mu,$$

$$\hat{\theta}_k^Q = \tau\theta_k^Q + (1-\tau)\hat{\theta}_k^Q.$$

end for

end for

end for

4. Simulation

This section provides an overview of the simulation environment, key parameters, and performance metrics used to evaluate the effectiveness of the proposed PER-MAC algorithm in V2I and V2V communication scenarios. The core simulation framework is introduced first, detailing the vehicle mobility model and the Manhattan grid layout to accurately reflect urban traffic conditions characterized by multiple lanes and high vehicle densities. This challenging environment is well suited for testing multi-agent reinforcement learning algorithms under dynamic and complex channel conditions, thereby assessing their adaptability and robustness. As this research builds upon the MADDPG algorithm, it is regarded as the baseline algorithm and is compared with other reinforcement learning algorithms. Finally, this section explains the key performance metrics, including the convergence speed, average reward, and stability of the learned policies, which are used to quantitatively measure the improvements provided by PER-MAC.

4.1. Simulation Environment

The wireless communication framework adopted in this research follows the guidelines set forth by 3GPP TR 36.885 [30], encompassing models for vehicles, lanes, and wireless communication networks. The simulation scenario represented a cellular-based vehicular network featuring both V2I and V2V communication links, structured according to a Manhattan city grid layout, as illustrated in Figure 2. It consisted of horizontal and vertical roadway intersections which formed rectangular blocks, and the base station was located in the center of the block. The primary simulation parameters are presented in Table 1. The training of the algorithm spanned 2000 episodes. Throughout the training period, the exploration probability gradually decreased, starting from 1.0 at the initial episode and reducing to 0.02 by the 1600th episode, maintaining this lower value for the remaining training episodes. During the training phase, the V2V link transmission packet size remained fixed, whereas it changed stepwise in the evaluation phase to assess the robustness of each strategy. Table 2 lists the hyperparameters used throughout the training process.

Table 1. Simulation parameter settings.

Parameter	Value
V2I link number M	4
V2V link number K	4
V2I link transmission power P^I	23 dBm
V2V link transmission power P^V	[23, −100] dBm
Carrier frequency	2 GHz
Bandwidth	4 MHz
Base station antenna height	25 m
Base station antenna gain	8 dBi
Base station receiver noise figure	5 dB
Vehicle antenna height	1.5 m
Vehicle antenna gain	3 dBi
Vehicle receiver noise figure	9 dB
Vehicle speed v	10 m/s
Gaussian white noise σ^2	−114 dBm
Time limit T	100 ms

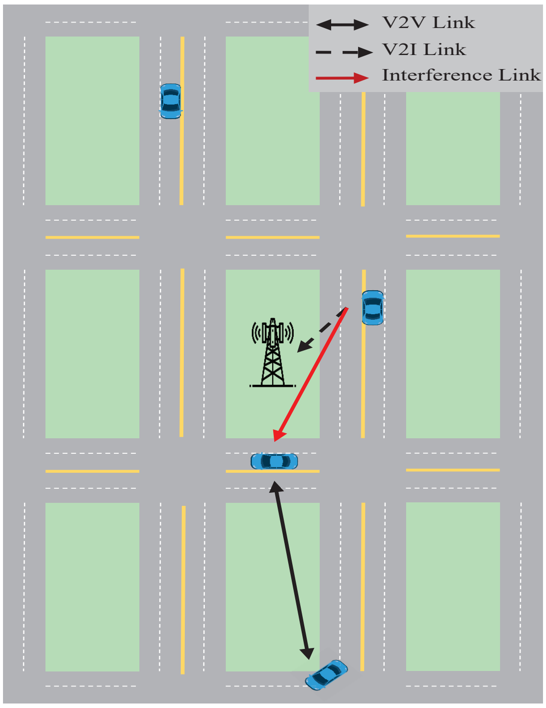


Figure 2. Schematic of the Manhattan grid layout.

Table 2. Training parameter settings.

Parameter	Value
Layers of the network	3
Number of neurons	[256, 64, 16]
Soft update learning rate τ	0.01
Activation function	ReLU
Training episode	2000
Exploration rate	$1 \longrightarrow 0.02$
Discount factor γ	0.99
Reward weight λ	0.9

4.2. Simulation Results

In addition to the previously discussed MAAC and MADDPG algorithms, this section also includes comparisons with two additional algorithms, providing a detailed performance evaluation:

(1) Deep deterministic policy gradient (DDPG) [36]: This algorithm features a single agent equipped with a dual-network structure, comprising an evaluation network and a target network. The agent's perception is limited to local observations and rewards associated with a single vehicle.
(2) Multi-agent deep Q network (MADQN) [37]: In this algorithm, each agent selects discrete actions from the available action set, based on the current state of the vehicular network environment. Unlike the DDPG algorithm, the MADQN uses a single type of neural network, specifically the Q-network. It also incorporates experience replay along with a dual-network structure, consisting of both the current and target networks. This approach follows a centralized training and distributed execution model, similar to the DQN algorithm.

To assess the convergence performance of the proposed algorithms, Figure 3 displays the progression of reward accumulation per episode for each algorithm. Each episode was composed of 100 training steps, with the cumulative rewards from these steps representing the episode's total reward. It is observable that the average rewards for all five algorithms exhibited an increasing trend as the episode count grew, signifying their gradual improvement in strategy optimization through the training process. Nevertheless, notable variances are apparent in the convergence rates and the ultimate reward values among the algorithms. PER-MAC distinguished itself by outshining its counterparts, reaching convergence near 300 episodes and subsequently maintaining the highest average reward level. Its rapid convergence and sustained stability demonstrate its effectiveness and resilience in managing power allocation challenges within complex scenarios. Both the MAAC and MADDPG algorithms showed comparable performance, but the MAAC algorithm attained a marginally superior final reward compared with the MADDPG algorithm. The DDPG algorithm's average reward, although consistently improving, remained somewhat below that of the MAAC and MADDPG algorithms and converged more slowly. MADQN exhibited the least favorable performance, achieving a significantly lower ultimate reward, which underscores its limited adaptability in this context. Moreover, the high-speed motion of the vehicle led to rapid changes and fading in the channel, which may cause fluctuations in algorithmic values even post convergence. PER-MAC's smaller reward fluctuations suggest a more stable learning process, highlighting its robust learning capabilities in dynamic settings. While the MAAC algorithm reached a high average reward ultimately, it maintained a local optimum until 1500 episodes, illustrating that the influence of short-term experiences can sometimes destabilize the final training outcomes.

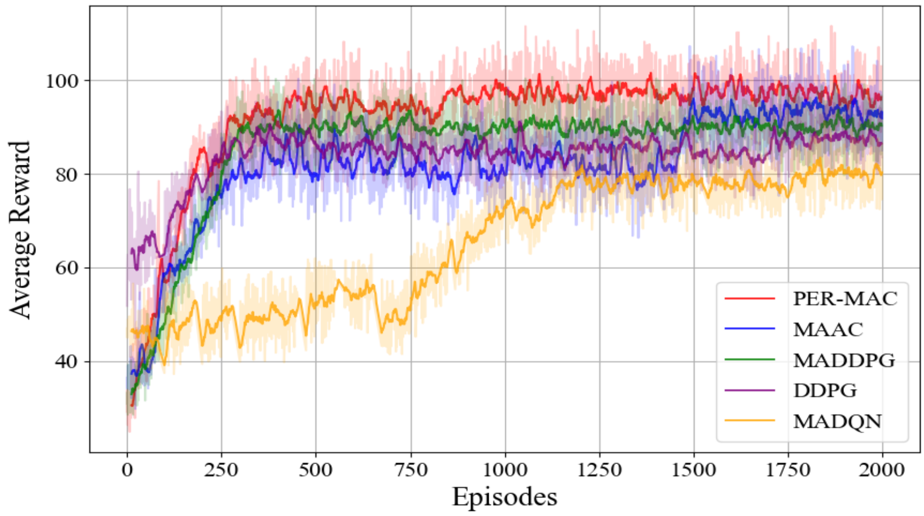
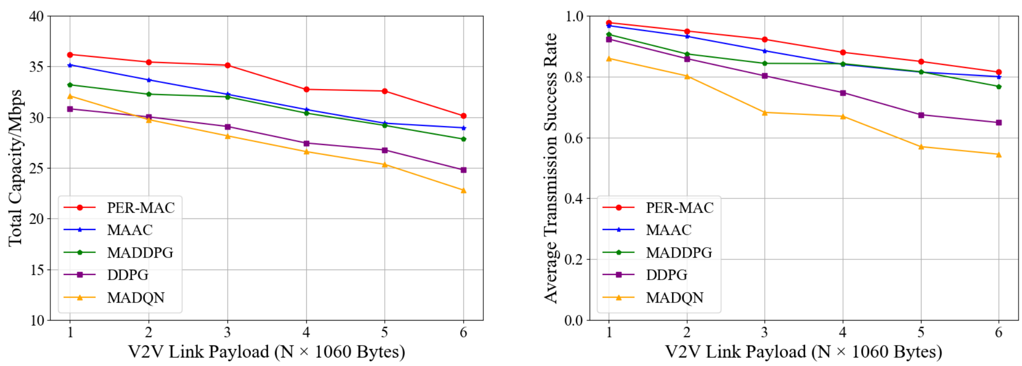


Figure 3. Total reward change curve for each episode of the training phase.

Figure 4a illustrates a comparison of the V2I link capacity across different algorithms with varying V2V link payload sizes. As the V2V link payload increased, a downward trend was observed in the V2I link capacity for all algorithms. This decline can be attributed to the fact that larger V2V link payloads result in longer transmission times, thereby increasing the transmit power required for the V2V link. This, in turn, causes greater interference for the V2I link, leading to a reduction in the overall communication capacity. Notably, the MADDPG algorithm demonstrated superior performance over the MADQN strategy, highlighting the effectiveness of the actor-critic architecture in more complex environments. Furthermore, a comparison with the DDPG strategy confirmed that algorithms designed specifically for multi-agent environments can better adapt to such environments. Meanwhile, the MAAC strategy showed performance similar to the MADDPG algorithm, though it outperformed the MADDPG algorithm in both the low- and high-payload regions. In the low-payload range, the MAAC algorithm appeared to be more capable of achieving near-optimal performance with fewer resources and less interference, whereas in the high-payload range, the attention mechanism employed by the MAAC algorithm helped dynamically adjust its strategy based on the importance of the agents, which reduced interference and enhanced the channel capacity. The proposed algorithm consistently surpassed the performance of all other algorithms across all payload sizes, and the gap between it and the others became more pronounced in the high-payload region, demonstrating the robustness of the proposed method.



(**a**) (**b**)

Figure 4. Performance of each model. (**a**) Total V2I link channel capacity under different payloads. (**b**) V2V link data transmission completion rate under different loads.

Figure 4b illustrates the relationship between the V2V link transmission success rate and payload for five different strategies. The primary goal of this study was to ensure the successful transmission of V2V link packets while maintaining a sufficient level of channel capacity for V2I links. As a result, the rewards for the V2V link were given greater weight, which narrowed the disparity in packet success transmission rates between the V2V and V2I links. To more clearly highlight the differences between the strategies, the V2V link packet success transmission rate is expressed as the ratio of the average time steps required for completing a link's transmission to the total time steps available. It was observed that all curves exhibited a downward trend, which can be attributed to the increased difficulty of completing transmissions with larger payloads given fixed transmission rates and time slots. The two graphs demonstrate that the PER-MAC algorithm excelled in terms of both the V2I link capacity and the V2V transmission success rate, highlighting its advantages in interference mitigation and ensuring link reliability.

To gain a deeper insight into the factors contributing to the superior performance of the proposed algorithms, an episodic comparison was conducted and analyzed. In order to highlight the differences between the algorithms more effectively, all V2V links were set to a data size of 6360 bytes. As illustrated in Figure 5, the MAAC algorithm allowed V2V link 4 to complete data transmission relatively early, while the remaining three links finished transmission at approximately the same time. In contrast, the PER-MAC algorithm showed a gap in completion times between links, except for links 2 and 4, which finished at nearly the same time. Nevertheless, the overall completion time for PER-MAC was shorter. The DDPG algorithm demonstrated early completion of transmissions for three links, but link 2 required a longer duration, which negatively impacted the overall performance. For MADQN, although two links completed the transmission later, the algorithm failed to ensure data transmission for the other two links, resulting in a failure to complete secure data transmission.

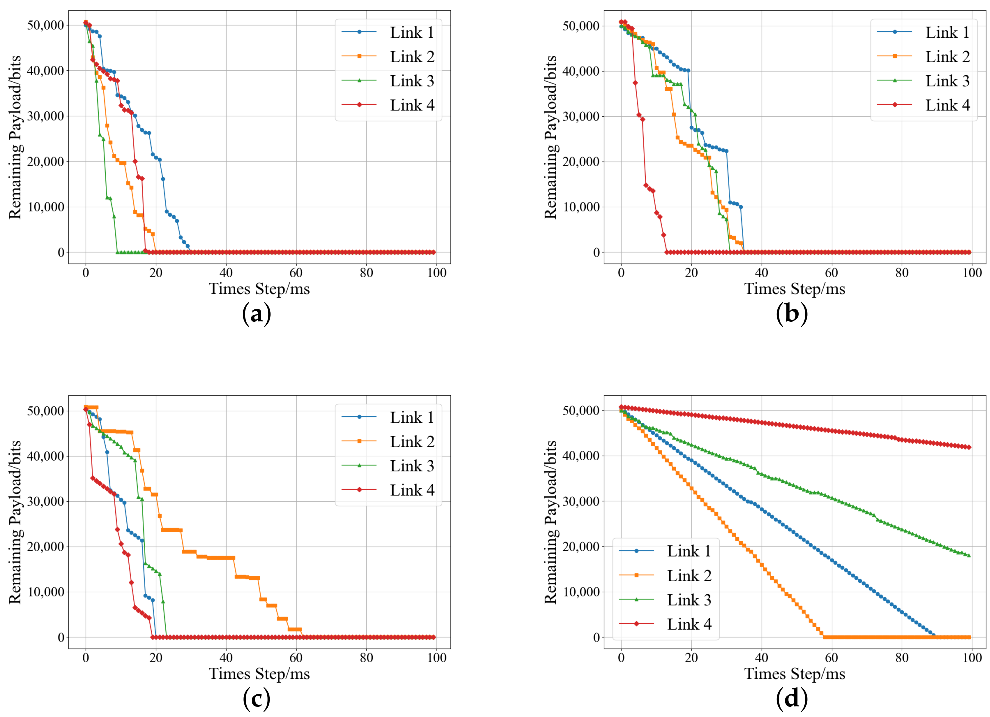


Figure 5. Variation in remaining payload over time step for each V2V link for each algorithm. (**a**) The remaining payload of PER-MAC at each time step. (**b**) The remaining payload of the MADDPG algorithm at each time step. (**c**) The remaining payload of the DDPG algorithm at each time step. (**d**) The remaining payload of MADQN at each time step.

Figure 6 illustrates the instantaneous transmission rates of all V2V links under various resource allocation schemes at each step, using the same scenario depicted in Figure 5. As shown in Figure 6a, the proposed method enabled V2V link 3 to fully capitalize on favorable channel conditions, achieving an exceptionally high transmission rate early on, which allowed for the completion of data transmission ahead of schedule and minimized interference with other links thereafter. Link 4 initially operated at a reduced transmission rate to ensure that the more vulnerable links 1 and 2 benefited from a relatively favorable transmission environment. Once the transmission of link 2 neared completion, link 4 rapidly increased its rate to finish transmitting the remaining data. A closer look at the transmission rates of links 1 and 2 revealed a strategic approach where both links rotated their transmissions efficiently. Figure 6b shows a similar pattern, but with the MADDPG strategy prioritizing the completion of transmission for link 4 first. This decision resulted in some competition for spectrum resources between links 2 and 3, causing a slight delay in overall transmission completion. The competition was more intense in the DDPG method, where link 2, experiencing a suboptimal transmission environment, struggled to transmit data securely and ended up finishing transmission later. In contrast, MADQN took an overly cautious approach with links 1 and 2, failing to promptly adjust the transmission rates of other links once the former two finished, leading to an eventual transmission failure.

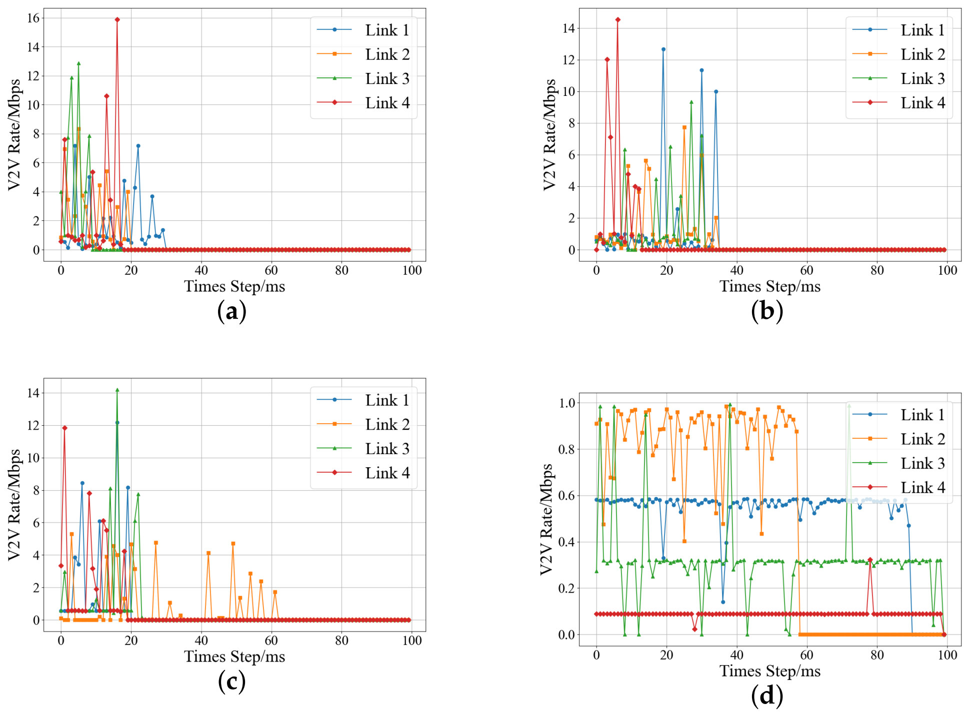


Figure 6. Variation in transmission rate over time step for each V2V link for each algorithm. (**a**) The transmission rate of PER-MAC at each time step. (**b**) The transmission rate of the MADDPG algorithm at each time step. (**c**) The transmission rate of the DDPG algorthm at each time step. (**d**) The transmission rate of MADQN at each time step.

5. Discussion

This study proposed the PER-MAC algorithm, a novel reinforcement learning framework designed to optimize power control in vehicular networks. While earlier studies have explored the contribution of MARL to power allocation in vehicular networks, there is no explicit discussion of the impact of significant experience on learning in complex environments. The simulation results demonstrated that the proposed framework effectively improved the SINR, reduced interference, and scaled well with increasing vehicular nodes, providing a robust solution for V2V and V2I communication scenarios. Although the proposed algorithm achieved promising results, it also highlighted several areas where further refinements are necessary to enhance its robustness and adaptability.

5.1. Performance Metrics

While this study primarily optimized the SINR as a key metric for vehicular communication performance, several related factors, such as the error rate, data types, and fault injection, have significant implications for the robustness and reliability of the proposed framework.

The error rate is a fundamental indicator of communication reliability.The PER-MAC algorithm sidesteps the error rate problem by optimizing the SINR to improve the reliability of data transmission as well as support different transmission requirements. Future work should consider explicitly incorporating the error rate into the reward function to explore its direct impact on vehicle communication performance. In addition, vehicular communication encompasses multiple data types, each with unique requirements for reliability, latency, and throughput. The PER-MAC framework, while not explicitly modeled in this study, can be extended to consider data prioritization. By utilizing its attention mechanism, the algorithm can allocate higher power to safety-critical data while conserving resources for lower-priority transmissions.

5.2. Bridging the Gap Between Simulation and the Real World

Transitioning from simulated to real-world environments represents a dynamic area of research in deep reinforcement learning (DRL). By performing domain adaptation to feature and pixel dimensions, OpenAI [38] skillfully trained a robotic arm in a gym environment to be able to perform grasping actions in the real world without the need for supplemental real-world training. This work depends on a domain randomization strategy which entails integrating randomness into the simulation training environment to encompass potential real-world scenarios, thus bridging the gap between simulated and actual environments.

While simulated environments provide a safe and controlled learning environment, the complexity and unpredictability of the real world far exceed those of simulated environments. In order to ensure model performance and stability under real road conditions, a series of targeted technical approaches and strategies must be employed. Training should cover a wide range of driving scenarios, such as adverse weather conditions, diverse traffic configurations, and countless emergencies, to confirm the model's adaptability to real-world complexity.

Although our current research on generalization focuses on adaptation to different scenarios, in real-world in-vehicle networks, failure conditions such as communication delays, packet loss, or power anomalies are common and can severely affect system performance. For example, an increase in delay may reduce the effectiveness of real-time power control decisions.The cooperative nature of PER-MAC allows the agent to adapt to dynamic environments, suggesting that the algorithm is inherently robust to certain faults. However, further research, particularly fault injection testing, is needed to test this hypothesis.

One of the limitations of this study is the lack of explicit fault injection experiments to evaluate the robustness of the algorithm under extreme conditions, such as severe communication outages or hardware failures. However, the current experimental set-up incorporates dynamic channel conditions and random vehicle behavior which partially simulates real-world uncertainty.

Future work will focus on fault injection testing to systematically evaluate the algorithm's ability to adapt to realistic failure scenarios such as communication delays or hardware failures. In addition, integrating fault tolerance mechanisms into the PER-MAC framework can further enhance its robustness.

5.3. Safety

A key challenge in in-vehicle communications is ensuring the security and reliability of multi-agent systems. Mobile agents in a reinforcement learning framework are vulnerable to threats such as tampering, malicious interference, and data poisoning. These risks are particularly acute in decentralized in-vehicle networks, where agent interactions are complicated by high mobility and frequent topology changes.

In [39], two approaches for training safety compliance agents were outlined: incorporating the safety layer into the decision-making process or modifying the optimization criteria. Within the domain of DRL applications for power control, the security layer is responsible for managing data privacy and authentication to prevent forged information from interfering with vehicle communications, in addition to dynamically adjusting communication parameters to minimize interference and ensure reliable transmission of information. As an integral part of the system, the security layer not only continuously monitors and evaluates potential security risks but also corrects or replaces policies as needed to avoid dangerous situations. For example, Gu et al. [40] detailed a state-based approach for enhancing the safety of autonomous driving, which significantly improves performance and safety in complex highway environments by combining dynamic goal setting with adaptive safety constraints through hierarchical reinforcement learning.

In reinforcement learning, directly constraining an agent's actions to prevent dangerous behaviors often leads to sparse rewards, complicating the learning process. This challenge is particularly relevant in the PER-MAC framework, where agents must balance efficient power allocation with avoiding potential risks, such as interference or inefficient resource usage.

To address this issue, input preprocessing can be an effective strategy for indirectly guiding agents toward safer behaviors without imposing hard constraints on their action choices. By leveraging techniques such as feature engineering and state filtering, it is possible to identify and mitigate state elements which contribute to hazardous or inefficient actions. For instance, in vehicular networks, input preprocessing could involve filtering out noisy or anomalous channel conditions which might mislead the agent into making suboptimal power allocation decisions. The fundamental principle behind this approach is to optimize the agent's input data, steering the learning process away from unsafe or redundant actions while preserving the agent's capacity for comprehensive environmental exploration. This aligns well with the PER-MAC framework's reliance on attention mechanisms, as the attention weights can be adjusted to deprioritize potentially dangerous or unreliable input features.

However, excessive input constraints could impair the agent's exploration capabilities, potentially hindering its ability to discover optimal strategies in complex and dynamic environments. To mitigate this risk, input preprocessing strategies could be selectively applied during specific training phases:

(1) Pretraining phase: During the initial stages of training, preprocessing could help the agent avoid exploring excessively unsafe regions of the action space.
(2) Threshold-based activation: After the agent has reached a certain performance threshold, preprocessing can be introduced to refine its behavior and reinforce safe practices.

For PER-MAC, this dual-phase approach ensures that the algorithm maintains a balance between safety and exploration, optimizing both its learning efficiency and robustness in practical vehicular network scenarios.

6. Conclusions

In this research, a multi-agent reinforcement learning approach known as PER-MAC was introduced to address the challenging problem of resource allocation in vehicular networks. The algorithm incorporates both a prioritized experience replay mechanism and an attention-based strategy, which enables the system to prioritize learning from crucial experiences and mitigate the issue of information redundancy, which arises as the number of agents increases. The simulation results demonstrate that PER-MAC not only achieved faster convergence and greater stability compared with existing methods but also substantially enhanced both the V2I link channel capacity and the success rate of V2V link transmissions, particularly under conditions of dynamic high mobility. The proposed algorithm augmented the total capacity of V2I links by 9.36% and the success rate of V2V transmissions by 6.74% compared with the benchmark algorithm, which demonstrates the effectiveness of experience-first replay and attention mechanisms. In addition to this, by analyzing the transmission rate of the V2V link, it was demonstrated that the proposed multi-agent reinforcement learning can indeed adopt a clever cooperation strategy in complex environments. Nevertheless, further investigation into multi-agent collaboration mechanisms, the development of more sophisticated reward functions, and the application of alternative reinforcement learning algorithms in low-latency and high-reliability environments will be the key directions for future research. In addition, the security of the algorithms needs to be addressed. Introducing adversarial scenarios for training, randomly simulating failure scenarios such as node failure, communication link interruption, or data poisoning, as well as protecting the agent's data integrity and decision logic through hardware security modules, such as the trusted platform module, are likewise some of the focuses of future work.

Author Contributions: J.-H.W. was the main author. H.H., J.C., and I.J. provided the methodology, software, and contructive feedback on every part of this manuscript. C.-J.A. was our supervisor. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by JST BOOST, Japan (grant number JPMJBS2413).

Data Availability Statement: Data available on request due to restrictions. The data presented in this study are available on request from the corresponding author due to sponsor limitations.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Al-Fuqaha, A.; Guizani, M.; Mohammadi, M.; Aledhari, M.; Ayyash, M. Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications. *IEEE Commun. Surv. Tutor.* **2015**, *17*, 2347–2376. [CrossRef]
2. Lin, J.; Yu, W.; Zhang, N.; Yang, X.; Zhang, H.; Zhao, W. A Survey on Internet of Things: Architecture, Enabling Technologies, Security and Privacy, and Applications. *IEEE Internet Things J.* **2017**, *4*, 1125–1142. [CrossRef]
3. Zhang, F.; Luo, C.; Xu, J.; Luo, Y.; Zheng, F.-C. Deep Learning-based Automatic Modulation Recognition Algorithm in Internet of Things. In Proceedings of the 2019 IEEE 2nd International Conference on Electronic Information and Communication Technology (ICEICT), Harbin, China, 20–22 January 2019; pp. 576–579.
4. Paden, B.; Čáp, M.; Yong, S.Z.; Yershov, D.; Frazzoli, E. A survey of motion planning and control techniques for self-driving urban vehicles. *IEEE Trans. Intell. Veh.* **2016**, *1*, 33–55. [CrossRef]
5. Liang, L.; Peng, H.; Li, G.Y.; Shen, X. Vehicular Communications: A Physical Layer Perspective. *IEEE Trans. Veh. Technol.* **2017**, *66*, 10647–10659. [CrossRef]
6. Peng, H.; Liang, L.; Shen, X.; Li, G.Y. Vehicular communications: A network layer perspective. *IEEE Trans. Veh. Technol.* **2018**, *68*, 1064–1078. [CrossRef]
7. Liang, L.; Li, G.Y.; Xu, W. Resource Allocation for D2D-Enabled Vehicular Communications. *IEEE Trans. Commun.* **2017**, *65*, 3186–3197. [CrossRef]

8. Das, D.; Das, S. Adaptive resource allocation scheme for cognitive radio vehicular ad-hoc network in the presence of primary user emulation attack. *IET Netw.* **2017**, *6*, 5–13. [CrossRef]
9. Ji, Y.; Sun, D.; Zhu, X.; Dong, D. Power allocation for cooperative communications in non-orthogonal cognitive radio vehicular ad-hoc networks. *China Commun.* **2020**, *17*, 91–99. [CrossRef]
10. Ruan, Y.; Li, Y.; Wang, C.-X.; Zhang, R. Energy efficient adaptive transmissions in integrated satellite-terrestrial networks with SER constraints. *IEEE Trans. Wirel. Commun.* **2017**, *17*, 210–222. [CrossRef]
11. Eze, J.; Zhang, S.; Liu, E.; Eze, E. Design optimization of resource allocation in OFDMA-based cognitive radio-enabled Internet of Vehicles (IoVs). *Sensors* **2020**, *20*, 6402. [CrossRef]
12. Xiao, H.; Jiang, H.; Shi, F.; Luo, Y.; Deng, L.; Mukherjee, M. Energy-efficient resource allocation in radio-frequency-powered cognitive radio network for connected vehicles. *IEEE Trans. Intell. Transp. Syst.* **2020**, *22*, 5426–5436. [CrossRef]
13. Noor-A-Rahim, M.; Liu, Z.; Lee, H.; Ali, G.G.M.N.; Pesch, D.; Xiao, P. A Survey on Resource Allocation in Vehicular Networks. *IEEE Trans. Intell. Transp. Syst.* **2022**, *23*, 701–721. [CrossRef]
14. He, H.; Wang, J.-H.; Kojima, S.; Maruta, K.; Ahn, C.-J. Regression CNN Based Fast Fading Channel Tracking Using Decision Feedback Channel Estimation. *J. Signal Process.* **2023**, *27*, 49–57. [CrossRef]
15. Nong, H.-Q. Optimization Strategy of Cognitive Radio Network Based on Machine Learning. *Comput. Technol. Dev.* **2020**, *30*, 125–131.
16. Zhang, C.; Patras, P.; Haddadi, H. Deep learning in mobile and wireless networking: A survey. *IEEE Commun. Surv. Tutor.* **2019**, *21*, 2224–2287. [CrossRef]
17. Zappone, A.; Di Renzo, M.; Debbah, M. Wireless networks design in the era of deep learning: Model-based, AI-based, or both? *IEEE Trans. Commun.* **2019**, *67*, 7331–7376. [CrossRef]
18. Althamary, I.; Huang, C.-W.; Lin, P. A Survey on Multi-Agent Reinforcement Learning Methods for Vehicular Networks. In Proceedings of the 2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC), Tangier, Morocco, 24–28 June 2019; pp. 1154–1159.
19. Zhang, K.; Leng, S.; Peng, X.; Pan, L.; Maharjan, S.; Zhang, Y. Artificial intelligence inspired transmission scheduling in cognitive vehicular communications and networks. *IEEE Internet Things J.* **2018**, *6*, 1987–1997. [CrossRef]
20. 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]
21. 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]
22. Zhang, X.; Peng, M.; Yan, S.; Sun, Y. Deep-reinforcement-learning-based mode selection and resource allocation for cellular V2X communications. *IEEE Internet Things J.* **2019**, *7*, 6380–6391. [CrossRef]
23. Ran, Y.; Hu, H.; Wen, Y.; Zhou, X. Optimizing energy efficiency for data center via parameterized deep reinforcement learning. *IEEE Trans. Serv. Comput.* **2022**, *16*, 1310–1323. [CrossRef]
24. Zhang, H.; Yang, N.; Huangfu, W.; Long, K.; Leung, V.C.M. Power control based on deep reinforcement learning for spectrum sharing. *IEEE Trans. Wirel. Commun.* **2020**, *19*, 4209–4219. [CrossRef]
25. Wu, Q.; Zhao, Y.; Fan, Q.; Fan, P.; Wang, J.; Zhang, C. Mobility-Aware Cooperative Caching in Vehicular Edge Computing Based on Asynchronous Federated and Deep Reinforcement Learning. *IEEE J. Sel. Topics Signal Process.* **2023**, *17*, 66–81. [CrossRef]
26. Bagheri, H.; Noor-A-Rahim, M.; Liu, Z.; Lee, H.; Pesch, D.; Moessner, K.; Xiao, P. 5G NR-V2X: Toward Connected and Cooperative Autonomous Driving. *IEEE Commun. Standards Mag.* **2021**, *5*, 48–54. [CrossRef]
27. Xu, C.; Wang, S.; Song, P.; Li, K.; Song, T. Intelligent Resource Allocation for V2V Communication with Spectrum–Energy Efficiency Maximization. *Sensors* **2023**, *23*, 6796. [CrossRef]
28. Alatabani, L.E.; Saeed, R.A.; Ali, E.S.; Mokhtar, R.A.; Khalifa, O.O.; Hayder, G. Vehicular network spectrum allocation using hybrid NOMA and multi-agent reinforcement learning. In *Sustainability Challenges and Delivering Practical Engineering Solutions: Resources, Materials, Energy, and Buildings*; Springer International Publishing: Cham, Switzerland, 2023; pp. 151–158.
29. Wang, W.; Su, J.; Chen, Y.; Zhang, J.; Tang, Z. Multi-Agent Reinforcement Learning Enabled Spectrum Sharing for Vehicular Networks. *Acta Electron. Sin.* **2024**, *5*, 1690–1699.
30. *3GPP TR 36.885 V14.0.0*; Technical Specification Group Radio Access Network, Study on LTE-Based V2X Services (Release 14). 3GPP: Sophia Antipolis, France, 2016.
31. Iqbal, S.; Sha, F. Actor-Attention-Critic for Multi-Agent Reinforcement Learning. In Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, 10–15 July 2018.
32. Chen, Z.; Wang, X. Decentralized computation offloading for multi-user mobile edge computing: A deep reinforcement learning approach. *Eurasip J. Wirel. Commun. Netw.* **2020**, *2020*, 188. [CrossRef]
33. Peng, H.; Shen, X. Multi-Agent Reinforcement Learning Based Resource Management in MEC- and UAV-Assisted Vehicular Networks. *IEEE J. Sel. Areas Commun.* **2021**, *39*, 131–141. [CrossRef]

34. Liang, L.; Xie, S.; Li, G.Y.; Ding, Z.; Yu, X. Graph-Based Radio Resource Management for Vehicular Networks. In Proceedings of the 2018 IEEE International Conference on Communications (ICC), Kansas City, MO, USA, 20–24 May 2018; pp. 1–6.
35. Lowe, R.; Wu, Y.; Tamar, A.; Harb, J.; Abbeel, P.; Mordatch, I. Multi-Agent Actor-Critic for Mixed Cooperative-Competitive Environments. *arXiv* **2017**, arXiv:1706.02275.
36. Lillicrap, T.P.; Hunt, J.J.; Pritzel, A.; Heess, N.; Erez, T.; Tassa, Y.; Silver, D.; Wierstra, D. Continuous control with deep reinforcement learning. *arXiv* **2015**, arXiv:1509.02971.
37. Yanmei, C.; Lin, W.; Jiaqing, L.; Jun, Z.; Lin, Y.; Guomei, Z. Joint Resource Allocation Scheme Based Multi-agent DQN for Massive MIMO-NOMA Systems. In Proceedings of the 2022 14th International Conference on Communication Software and Networks (ICCSN), Chongqing, China, 10–12 June 2022; pp. 51–55.
38. OpenAI: Marcin Andrychowicz; Baker, B.; Chociej, M.; Jozefowicz, R.; McGrew, B.; Pachocki, J.; Petron, A.; Plappert, M.; Powell, G.; Ray, A.; et al. Learning dexterous in-hand manipulation. *Int. J. Robot. Res.* **2020**, *39*, 3–20.
39. Garcıa, J.; Fernández, F. A comprehensive survey on safe reinforcement learning. *J. Mach. Learn. Res.* **2015**, *16*, 1437–1480.
40. Gu, Z.; Gao, L.; Ma, H.; Li, S.E.; Zheng, S.; Jing, W.; Chen, J. Safe-state enhancement method for autonomous driving via direct hierarchical reinforcement learning. *IEEE Trans. Intell. Transp. Syst.* **2023**, *24*, 9966–9983. [CrossRef]