future internet

MDPI

Article

A Hybrid Multi-Agent Reinforcement Learning Approach for Spectrum Sharing in Vehicular Networks

Mansoor Jamal [1], Zaib Ullah [2,*], Muddasar Naeem [2], Musarat Abbas [1] and Antonio Coronato [2]

[1] Department of Electronics, Quaid-i-Azam University, Islamabad 44000, Pakistan
[2] Artificial Intelligence and Robotics Lab, Università Telematica Giustino Fortunato, 82100 Benevento, Italy
* Correspondence: z.ullah@unifortunato.eu

Abstract: Efficient spectrum sharing is essential for maximizing data communication performance in Vehicular Networks (VNs). In this article, we propose a novel hybrid framework that leverages Multi-Agent Reinforcement Learning (MARL), thereby combining both centralized and decentralized learning approaches. This framework addresses scenarios where multiple vehicle-to-vehicle (V2V) links reuse the frequency spectrum preoccupied by vehicle-to-infrastructure (V2I) links. We introduce the QMIX technique with the Deep Q Networks (DQNs) algorithm to facilitate collaborative learning and efficient spectrum management. The DQN technique uses a neural network to approximate the Q value function in high-dimensional state spaces, thus mapping input states to (action, Q value) tables that facilitate self-learning across diverse scenarios. Similarly, the QMIX is a value-based technique for multi-agent environments. In the proposed model, each V2V agent having its own DQN observes the environment, receives observation, and obtains a common reward. The QMIX network receives Q values from all agents considering individual benefits and collective objectives. This mechanism leads to collective learning while V2V agents dynamically adapt to real-time conditions, thus improving VNs performance. Our research finding highlights the potential of hybrid MARL models for dynamic spectrum sharing in VNs and paves the way for advanced cooperative learning strategies in vehicular communication environments. Furthermore, we conducted an in-depth exploration of the simulation environment and performance evaluation criteria, thus concluding in a comprehensive comparative analysis with cutting-edge solutions in the field. Simulation results show that the proposed framework efficiently performs against the benchmark architecture in terms of V2V transmission probability and V2I peak data transfer.

Keywords: vehicular networks; multi-agent reinforcement learning; deep Q-networks (DQN); QMIX; spectrum management; deep reinforcement learning

Citation: Jamal, M.; Ullah, Z.; Naeem, M.; Abbas, M.; Coronato, A. A Hybrid Multi-Agent Reinforcement Learning Approach for Spectrum Sharing in Vehicular Networks. *Future Internet* **2024**, *16*, 152. https://doi.org/10.3390/fi16050152

Academic Editors: Stefano Rinaldi and Alan Oliveira De Sá

Received: 24 March 2024
Revised: 20 April 2024
Accepted: 23 April 2024
Published: 28 April 2024

1. Introduction

Intelligent Transportation Systems (ITSs) are integral to our daily lives, thus impacting road safety, security, and efficiency [1]. ITSs encompass advanced applications that aim to provide intelligent services for various modes of transport and traffic control. V2X communication, a critical component of the ITS, enables vehicles to interact with infrastructure, machinery, grids, and other vehicles. This interaction can significantly enhance road safety, traffic efficiency, and energy preservation, as well as pave the way for self-driving vehicles. Vehicle-to-Everything (V2X) communication is envisioned for users in a new era of intelligent transportation systems, where connected vehicles seamlessly integrate with networks to meet safety and operational requirements [2].

Following the recent release of the 3rd Generation Partnership Project (3GPP) specifications, Vehicular Communication Networks (VCNs) have garnered increased attention from both industries and academia [3]. The underpinning of V2X technology lies in a heterogeneous network environment that leverages Long-Term Evolution (LTE) and 5G cellular communication technologies, which are renowned for their efficacy in high-speed data

transfer [4,5]. In [6], the authors explored challenges in V2X communications, specifically in metropolitan environments with signal obstruction due to installations. A Reconfigurable Intelligent Surface (RIS) was suggested as a key to improving communication performance by adjusting signal reflection. While a RIS shows benefits, traditional solutions have coverage constraints. Researchers have suggested simultaneously transmitting and reflecting the RIS (STAR-RIS), thus enabling 360° coverage and leveraging RIS advantages. They explored a STAR-RIS-assisted V2X system, thereby developing an optimization problem to maximize the data rate for V2I users while satisfying V2V latency and reliability conditions. They organized the problem into two sub-problems using a combined DDQN with an attention mechanism and an optimization-based technique for real-time solutions. Numerical results show substantial performance gains over established RIS solutions.

With the exponential growth of connected devices and data-intensive applications like V2X communication, the demand for high bandwidth and reliable data transmission is exceeding the capacity of existing spectrum allocations [7]. This paper explores the challenges and potential solutions for spectrum access in V2X networks, thus focusing on achieving the high data rates and reliability required for critical safety and information exchange applications. The 3GPP facilitates V2X (Vehicle-to-Everything) services within the framework of 5G cellular networks. The 3GPP Release 16 specifically presents connectivity solutions tailored for Robotics and Autonomous Systems (RASs) [8]. Due to the constant movement of vehicles and varying environments (buildings, tunnels, etc.), the signal quality (channel conditions) in V2X communication experiences significant and rapid changes. This introduces significant uncertainty when allocating resources like bandwidth and power, thereby challenging conventional optimization methods in solving V2X resource allocation problems.

Our emphasis lies in the 3GPP cellular V2X technology, where a reservoir of radio resources is at the disposal of vehicles, thus enabling them to judiciously select optimal resources for their communication needs [9]. The spectrum is currently allocated for V2I (Vehicle-to-Infrastructure) links utilizing OFDM, thus incorporating essential interference management and power control mechanisms. For V2V (Vehicle-to-Vehicle) links, effective strategies are required to efficiently share the preoccupied spectrum with V2I. This involves the selection of optimal sub-bands and power levels. Similarly, in [10], a task offloading strategy has been developed to maximize computation efficiency between vehicles.

This work introduces a novel Multi-Agent Reinforcement Learning (MARL) algorithm with QMIX cooperation for dynamic resource allocation in heterogeneous V2X networks [11]. It tackles the challenging scenario of multiple V2V links having a pool of radio resources that users can autonomously select for V2V communication. By selecting the optimal power level and spectrum sub-band, the algorithm dynamically adapts to the changing vehicular environment, thereby aiming to achieve increased network efficiency, an enhanced data rate, and improved reliability. Compared to existing methods, our QMIX-based MARL excels in promoting V2V cooperation, real-time adaptation through an online MARL approach, and joint network optimization for superior efficiency and fairness. This innovative solution presents significant contributions to dynamic resource allocation in V2X networks.

The article layout is organized as follows: Section 2 discusses the related studies in detail, and Section 3 describes the system model with various subsections to explicitly talk about the necessary techniques. Section 4 explains the construction of the hybrid model and respective algorithms, and Section 5 presents the simulation results of the proposed framework. Section 6 concludes the article.

2. Related Work

In V2X communication, particularly for coordinating both V2V and V2I links, an effective resource allocation mechanism is crucial to ensure the efficient use of limited spectrum resources.

To address the challenge of uncertain optimal resource allocation due to fast-changing channel conditions in vehicular networks, The authors of [12] proposed an optimization framework for wireless charging, power allocation, and resource block assignment in cellular V2X communications. It includes a network model where roadside objects use wireless power from RF signals of electric vehicles for charging and information processing. In [13], the study addressed D2D-based V2X communication with imperfect CSI, thus aiming to maximize the VUEs' sum ergodic capacity. It proposed a low-complexity power control approach and an enhanced Gale–Shapley algorithm for spectrum resource allocation, thus showing improved efficiency and effectiveness in enhancing the VUEs' sum ergodic capacity. Similarly, the authors of [14] maximized V2X throughput via power allocation. It proposed a Weighted Minimum Mean Square Error (WMMSE) algorithm and a deep learning method. The WMMSE uses block coordinate descent, while the deep learning approach leverages a supervised Deep Neural Network (DNN) trained on WMMSE outputs. The DNN achieved accurate power allocation with significantly reduced computational overhead.

Furthermore, in [15], a novel approach for cellular V2V was proposed to enhance reliability and reduce latency. In this approach, V2V links are controlled by eNodeB, with each link monitoring packet lifetime and waiting for eNodeB responses. In contrast to centralized resource allocation, which can be vulnerable to single-point failures like eNodeB outages, decentralized or distributed schemes hold significant promise for V2X communication. In these distributed approaches, vehicles autonomously negotiate and select optimal resources through direct communication, even when traditional infrastructure is unavailable. This enhances network resilience and enables flexible and scalable resource utilization. The article [16] investigated a distributed approach for transmission mode selection and resource allocation in V2X networks, thus also formulating the problem as a Markov Decision Process (MDP).

The authors in [17] exploreed challenges in mode settings and resource allotment in heterogeneous V2X conditions. They suggest a federated multi-agent Deep Reinforcement Learning (DRL) strategy with action awareness to guarantee Quality of Service (QoS). The technique includes an action–observation-based DRL and a prototype parameter accumulation algorithm. By observing adjacent agents' actions and dynamically balancing documented rewards, a fast convergence of individual models is guaranteed. The model parameter aggregation technique enhances generalization by sampling historical parameters and maintaining individual model qualities.

Furthermore, the authors in [18] considered V2X resource allocation with Deep Reinforcement Learning (DRL). A Deep Q Network (DQN) and Deep Deterministic Policy Gradient (DDPG) DQN–DDPG combo handles the sub-band and power, while meta-DRL enables fast adaptation in dynamic environments, thus outperforming quantized power approaches. Investigations into unreliable V2V connections and signaling overhead for joint model selection have been discussed in [19]. To address these challenges, a two-time scale federated DRL approach was proposed, thus leveraging clustering techniques for enhanced efficiency. Similarly, the work in [20] addressed spectrum scarcity in V2X communications by proposing a joint Resource Block (RB) sharing and power allocation method via online distributed MARL, which optimizes the V2I data rate and V2V reliability.

Traditional approaches rely on complete knowledge of the communication channels, thereby offering centralized control. However, these methods struggle with high communication overhead and increased latency, especially in scenarios with heavy traffic. To address this, we propose a distributed system where each agent (vehicle) independently selects the optimal sub-band and power level for communication. This distributed approach aims to maximize network performance by eliminating the need for constant information exchange.

3. System Model

Consider a cellular-based vehicular communication network illustrated in Figure 1. The vehicular network comprises M V2I links and K V2V links, thus offering support for

high-bandwidth entertainment services and dependable regular safety message exchange for advanced driving applications. We note that all the vehicles are capable of doing both V2I and V2V communication equipped with different multiple radios.

The V2I links in our vehicular communication network employ cellular interfaces, thus establishing connections between vehicles and the Base Station (BS) to facilitate high-speed data transmission services. Conversely, V2V links utilize sidelink interfaces, thereby enabling localized D2D communications for periodically disseminating generated safety messages.

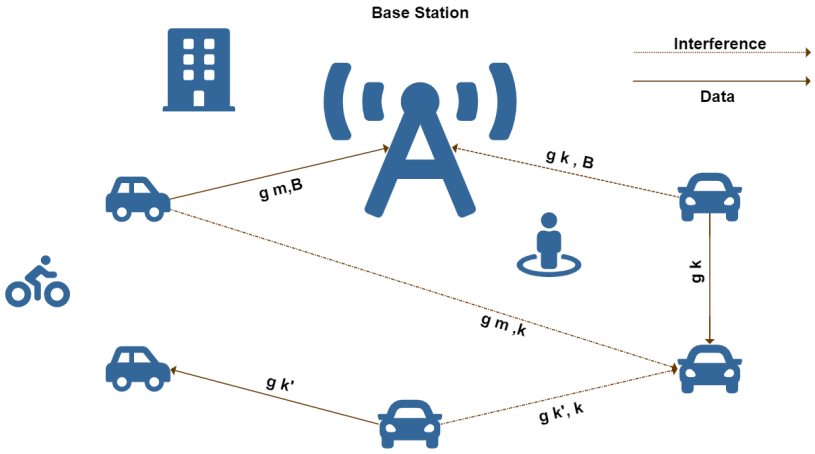


Figure 1. An illustrative structure of vehicular communication networks. B is the base station, m is V2I link, k is the V2V link, k' is another V2V link, $g_{m,B}$ is the power gain of V2I link, and $g_{k,B}$ is the interference power gain of V2I link.

The set represents V2I links in the network $\mathcal{M} = \{1,2,\ldots,M\}$, while the V2V links are represented by $\mathcal{K} = \{1,2,\ldots,K\}$. This setup enables efficient data exchange and seamless communication, thereby ensuring dependable safety message sharing and high-data-rate entertainment services among vehicles.

The interference to the V2I links consists of two parts: the background noise and the signal from the V2V links sharing the same sub-band. The Signal-to-Interference-plus-Noise Ratio (SINR) for the mth V2I is given by Equation (1) as follows.

$$\gamma_m = \frac{P_m g_{m,B}}{\sigma^2 + \sum_{k \in \mathcal{K}} \rho_k P_k g_{k,B}}, \tag{1}$$

where P_m is the transmission power of the V2I link, and P_k is the transmission power of the V2V link.

σ^2 in Equation (1) is the noise power, and $g_{m,B}$ is the power gain of the channel corresponding to the mth V2I link, where B refers to the BS (Base Station), and $g_{k,B}$ is the interference power gain of the kth V2V link, which reflects the impact of interference from the V2V link on the BS. ρ_k is the spectrum indicator, with 1 when the kth V2V user reuses the spectrum and 0 otherwise.

The capacity of the mth V2I link can be expressed as given in Equation (2).

$$C_m = W \log(1 + \gamma_m), \tag{2}$$

where W in Equation (2) signifies the bandwidth associated with each spectrum sub-band.

Similarly, we can formulate the SINR for the kth V2V link, which might utilize the shared spectrum as given in Equations (3)–(5).

$$\gamma_k = \frac{P_k g_k}{\sigma^2 + G_k + G_{\hat{k}}}, \tag{3}$$

with

$$G_k = P_m g_{m,k}, \tag{4}$$

and

$$G_{\hat{k}} = \sum_{k' \neq k} \rho_{k'} P_{k'} g_{k',k} g_{k'}, \tag{5}$$

where g_k is the gain of the kth V2V link; $g_{m,k}$ is the interference from the V2I link; $\rho k'$ is the spectrum indicator, with 1 when k' the V2V user reuses the spectrum and 0 otherwise; k' is the other V2V user sharing the spectrum with kth V2V user; and $g_{k',k}$ and $g_{k'}$ are the interference from another k'th V2V links using the same resource block. The capacity of the kth V2V can be expressed as given in Equation (6).

$$C_k = W \log(1 + \gamma_k^d). \tag{6}$$

This paper aims to optimize the V2I link capacity for high-quality entertainment services, thus meeting low-latency and high-reliability requirements for V2V links to provide realistic and dependable information to vehicle users in road traffic. To achieve the first objective, the sum rate of the V2I links must be maximized. For the second objective, V2V users need to successfully transmit packets of size D within a finite time T_{max}, as modeled in Equation (7).

$$\Pr\left\{ \sum_{t=1}^{T_{\max}} C_k[t] \geq D/\Delta T \right\} \tag{7}$$

where P_r denotes the probability, $C_k[t]$ represents the achievable data rate of the V2V user K at time slot t, D is the packet size, ΔT is the transmission time of one packet, and $T_{\max}$ is the maximum allowed transmission time.

3.1. Environment Modeling

In Figure 1, we present a dynamic resource sharing scenario where multiple V2V links strive to use the spectrum occupied by the V2I links. This competitive environment underscores the suitability of a MARL approach, thereby treating each V2V link as an autonomous agent navigating the uncertain landscape to optimize both individual and collective performance.

Expanding on the resource sharing scenario in Figure 2, we conceptualize the interactions among numerous V2V links as a Multi-Agent MDP (MA-MDP). This framework views each V2V link as an independent agent, where the system reward is accessible to each V2V agent, which then adjusts its actions toward an optimal policy by updating its own DQN. The MA-MDP approach captures the inherent complexity of this dynamic environment, thus facilitating the use of MARL techniques for efficient resource allocation and coordination among V2V links.

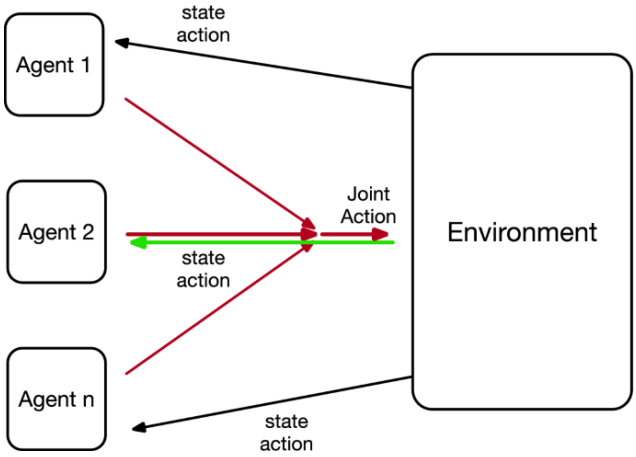


Figure 2. Multi-Agent Reinforcement Learning [21].

In this multi-agent setting, at each timestep t, each Vehicle-to-Vehicle (V2V) agent k explores the unknown environment. They receive their observations and then take action, thereby collectively forming a joint action. This joint action leads to a transition of the environment to a new state. Subsequently, the agents receive new observations based on this new state. To promote cooperation among agents, all agents share the same reward signal.

3.2. Deep Q Networks

In scenarios with limited observation space and a small lookup table, Q learning excels. However, challenges arise when confronted with larger state and action spaces, thus resulting in a growing lookup table and non-stationarity. The latter occurs when many Q values remain unchanged for extended periods, thus significantly slowing down the learning process after multiple repetitions. To overcome these challenges, DQNs introduce a paradigm shift by integrating Q values into a sophisticated neural network, which is termed the DNN, as illustrated in Figure 3. This network is characterized by the parameter set θ, thus representing its weights. The key concept involves approximating and determining these weights θ. Once established, the DNN efficiently addresses the mapping of the Q values, thereby enhancing stability in $Q(s_t, a_t)$ [22].

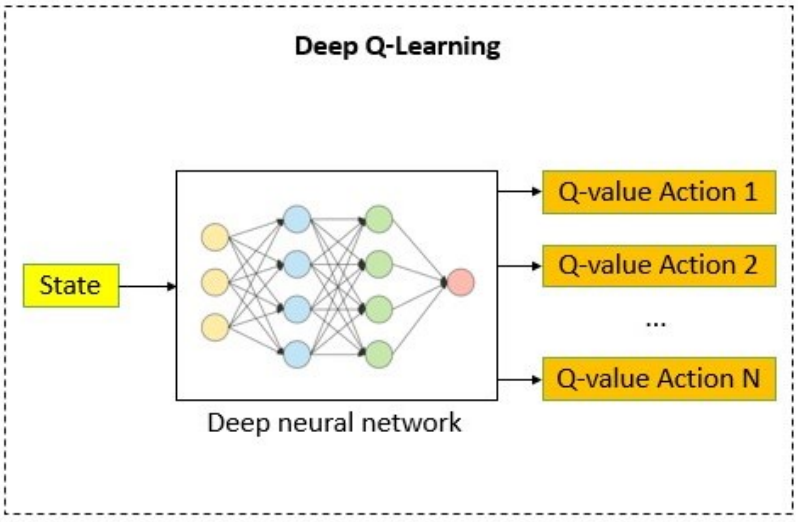


Figure 3. Generalized structure of Deep Q Networks [23].

The Temporal Difference (TD) error, denoted by δ, is fundamental to the learning process and is defined as given in Equation (8).

$$\delta = R + \gamma \max_{a'} Q(s', a'; \theta') - Q(s, a; \theta) \tag{8}$$

where δ represents the TD error, R is the immediate reward, γ is the discount factor (between 0 and 1), $max_{a'} Q(s', a'; \theta')$ represents the maximum expected future reward the agent can achieve from the next state (s'), and $Q(s, a; \theta)$ represents estimated Q value for taking the chosen action a in the current state s, which is also based on the current policy parameters θ'. The loss function used during training, denoted by L, is based on the Mean Squared Error (MSE) between predicted and target Q values as given in Equation (9).

$$L = \frac{1}{2}\left(Q_{\text{target}} - Q_{\text{pred}}\right)^2 \tag{9}$$

where Q_{target} is calculated using the TD error and represents the target Q value. Q_{pred} is the predicted Q value obtained from the DNN.

3.3. QMix Integration for Multi-Agent Reinforcement Learning

Our work seamlessly integrates QMix, a potent algorithm designed for cooperative multi-agent systems, into our spectrum allocation framework for vehicular communication. V2V agents, functioning as independent learners in an uncertain environment, collaborate to dynamically explore and determine the optimal spectrum choices.

The challenges introduced by multiple agents actively exploring the environment extend beyond traditional learning paradigms. To address these complexities, we model the environment as fully cooperative, thus fostering collaborative adaptation among V2V agents under dynamic conditions.

QMix, also known as "Deep Q-Mix", strategically addresses the hurdles of cooperative learning by introducing a central mixing network [24,25]. This network efficiently blends the individual Q values of agents, thereby enabling seamless information sharing and coordinated decision making.

In Figure 4, we illustrate the decentralized Q values integrating into a centralized mixing network. Each node represents the Q value estimation network for an individual agent $(Q_1(s_1, a_1), \ldots, Q_N(s_N, a_N))$, where $o_1 \ldots o_N$ represents the agent observation, and $a_1 \ldots a_N$ represents the corresponding action. These networks estimate the Q values for each agent's actions in a given state, and the Q values flow through the central "Mixing Network" block.

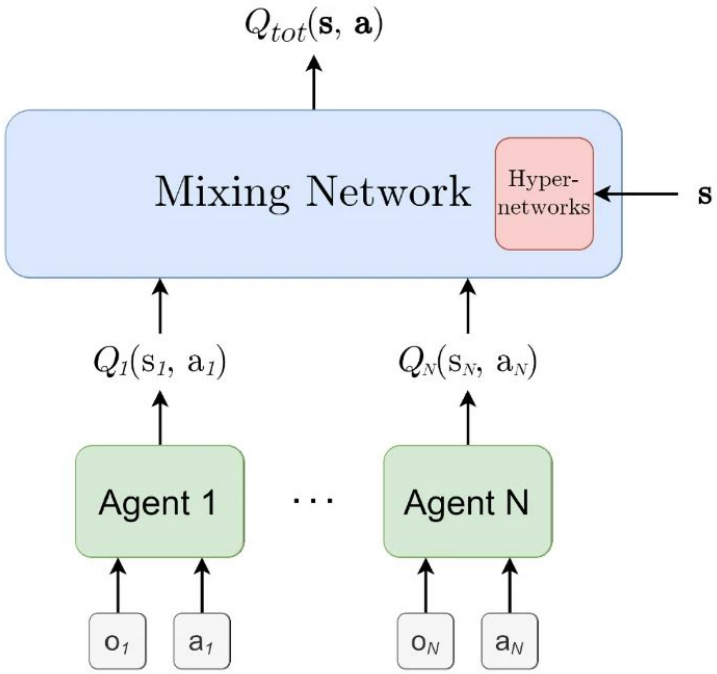


Figure 4. Generalized illustration of the QMix Structure [26].

Agents actively share their Q values, thus contributing to the mixing network. This network utilizes hyper-networks to assign mixing weights to each agent's Q value, thereby determining their influence on the final joint action value function. The output of the network, depicted as the node, represents this joint action value function $(Q_{tot}(s, a))$. It encapsulates the combined value for all agents taking a specific joint action in the current state.

This collaborative decision-making process guides agents toward optimal spectrum allocations, thereby ultimately enhancing communication and network efficiency.

$$Q_{\text{tot}}(s, a) = \sum_i w_i(s) \cdot Q_i(s_i, a_i) + V(s) \tag{10}$$

where:

$Q_{\text{tot}}(s, a)$ is the joint action value function for state s and joint action a.
$w_i(s)$ are the mixing weights for agent i, which are determined by the hyper-networks.
$Q_i(s_i, a_i)$ are the individual Q values for agent i in state s and action a_i.
$V(s)$ is the state-dependent bias term in Equation (10) respectively.

3.4. State and Action Space

In the multi-agent resource sharing problem, numerous V2V agents navigate the environment. At each timestep, individual agent k processes observation o_k and takes action a_k, thus collectively forming a joint action. The action a_k involves the selection of the power level and optimal sub-band for data transmission.

4. Hybrid Training

In our approach, agents start with decentralized learning, where they individually interact with the environment to learn. They then share this information with a central network, thus forming a hybrid training process. Decentralized learning lets V2X agents adapt independently based on local observations, but it may not always optimize collaboration.

To enhance collaboration, we use QMix, which is a centralized technique. QMix's central mixer network processes individual Q values from agents, thereby considering relationships and dependencies, and generates final Q values [27]. This incentivizes agents to consider team success, thus fostering efficient and collaborative decision making in V2X communication. Our deployment integrates both decentralized and centralized learning for optimal outcomes. The learning and deployment procedures are summarized in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 Collaborative spectrum sharing with multi-agent RL and QMix integration.

```
Start environment simulator, generating vehicles and links
Initialize Q networks for all V2V agents randomly
for each episode do
    Update vehicle locations and large-scale fading
    Reset parameters for all V2V agents
    for each step t do
        for each V2V agent k do
            Observe the environment at time t
            Choose an action using a policy
            Share Q values with QMix
        end for
        Update channel small-scale fading
        All agents take actions and receive rewards
        for each V2V agent k do
            Observe the environment at time t + 1
            Store experience for learning
        end for
    end for
    for each V2V agent k do
        Sample mini-batches for learning
        Share learned Q values with QMix
        Optimize Q network and QMix using gradient descent
    end for
end for
```

Algorithm 2 Deployment scheme with QMix-enhanced agents.

```
Procedure Testing
Start Run Environment
Load QMix-Enhanced Models
Activate all links
for each test_step do
  for each agent do
    Update Environment
    Update Demand
    Select Optimal link using QMix-Enhanced Model
  end for
end for
```

4.1. Decentralized Learning

In the initial step, Q networks serve as the learning mechanisms for individual agents and are randomly initialized. Each V2V agent has a dedicated DQN that takes the current observation as input and outputs the value functions corresponding to all actions. The training process unfolds within a simulated V2V communication scenario shown in each training episode. During an episode, dynamic adjustments are made to vehicle locations and channel characteristics. Agents interact with the environment, receive observation, make decisions based on their policies, and participate in decentralized learning. We train Q networks through multiple episodes and at each episode, and all V2V agents explore sate–action space with some policies. This decentralized learning is the key aspect of the training phase, thus enabling agents to adapt to dynamic conditions.

4.2. Centralized Learning (QMIX)

In the centralized learning phase a central body, the QMIX, facilitates centralized learning by collecting and integrating relevant information from all agents that coordinate the learning process. Agents share their learning with the QMIX, thus fostering collaborative learning. The QMIX combines these decentralized learnings, thus resulting in a joint action value function [28]. This joint action value function guides the actions of all agents to achieve a common goal enhancing collaboration among them. Moreover, the small-scale characteristics of the communication channel are updated, and agents store their experiences. These experiences serve as valuable data for optimizing learning through gradient descent, thus further improving decision making over time.

4.3. Deployment

In the deployment phase, the pre-trained models acquired from the training phase are applied in real-time scenarios. During each timestep t, individual Vehicle-to-Vehicle (V2V) agent k estimate their local channel state. They then leverage their observations of the surrounding environment, incorporating the knowledge gained through the training process, to select actions based on their trained Q networks. These models encapsulate learned behaviors and optimized decision-making strategies cultivated during the training phase. The deployment procedure in Algorithm 2 is executed online and, based on environment conditions, the trained DQN of the agent only updates when it experiences significant changes in the environment.

In the context of live communication instances, V2V agents employ these pre-trained models to estimate local channel characteristics in real time. Leveraging these observations and the acquired decision-making strategies helps agents make informed decisions to optimize communication links. The overarching goal is to ensure efficient spectrum sharing among V2V entities, thereby contributing to the overall effectiveness of the communication network.

5. Simulation Results and Findings

This section presents the simulation results, thus demonstrating the effectiveness of our proposed multi-agent RL-based spectrum sharing scheme. We utilized a custom-built simulator adhering to the evaluation methodology outlined in Annex *A* of 3GPP TR 36.885, thus considering critical parameters such as vehicle drop models, traffic densities, speeds, directional movements, and channel characteristics for comprehensive analysis.

Table 1 lists the simulation parameters reflecting realistic vehicular communication scenarios. The parameters like power levels, frequencies, noise characteristics, and speed were precisely chosen for their alignment with real-world scenarios of vehicular communication. Additionally, Table 2 specifies the channel models for both V2I and V2V links. Moreover, in Table 2, "B1 Manhattan" refers to the typical environment or scenario being considered, especially the metropolitan environment of Manhattan. In the context of wireless communication modeling, assuming the unique characteristics of urban surroundings such as Manhattan is fundamental due to aspects like high population density, tall buildings, and complex propagation conditions.

All the parameters in Tables 1 and 2 were set to default values for the simulation. However, meticulous attention to specific configurations for each figure ensures precision and coherence in the results. The simulations were carefully designed to authenticate the effectiveness and efficiency of our proposed resource sharing scheme in dynamic and realistic vehicular environments.

Table 1. Simulation parameters [29].

Parameter	Value
Number of V2I links M	4
Number of V2V links K	4
Carrier frequency	2 GHz
Bandwidth	4 MHz
BS antenna height	25 m
BS antenna gain	8 dBi
BS receiver noise figure	5 dB
Vehicle antenna height	1.5 m
Vehicle antenna gain	3 dBi
Vehicle receiver noise figure	9 dB
Absolute vehicle speed v	36 km/h
Vehicle drop and mobility model	Urban case of A.1.2 in [29]
V2I transmit power P^c	23 dBm
V2V transmit power P^d	[23, 10, 5, 0] dBm
Noise power σ^2	-114 dBm
Time constraint of V2V payload T	100 ms
V2V payload size B	$[1, 2, \ldots] \times 10^6$ bytes

Table 2. Modeling the wireless channels for V2I and V2V communications [29].

Parameter	V2I Link	V2V Link
Path loss model	$128.1 + 37.6\log_{10} d$, d in km	LOS in WINNER + B1 Manhattan
Shadowing distribution	Log-normal	Log-normal
Shadowing standard deviation ξ	8 dB	3 dB
Decorrelation distance	50 m	10 m
Path loss and shadowing update	A.1.4 100 ms	A.1.4 100 ms
Fast fading	Rayleigh fading	Rayleigh fading
Fast fading update	Every 1 ms	Every 1 ms

The neural architecture for each V2V agent's DQN comprises three fully connected hidden layers with 500, 250, and 120 neurons. Inside the hidden layers, the ReLU activation function, expressed as $f(x) = \max(0, x)$ is utilized. This activation function was selected to introduce non-linear transformations, thereby enabling the network to effectively capture

intricate patterns and features from the input data. Throughout the training process, the network parameters were updated using the RMSProp optimizer with a learning rate set at 0.001. Additionally, the network was trained using the Huber loss function, thus enhancing its ability to handle outliers and noisy data during learning. In Figures 5 and 6, we compared the QMIX resource sharing scheme with the following three baseline methods:

(1) The multi-agent RL-based algorithm where multiple agents update their actions in a distributed way.
(2) The single-agent RL-based algorithm where only a single agent at each moment updates its action, while the other agent's actions remain unchanged.
(3) The random method, which chooses the action randomly for all V2V agents.

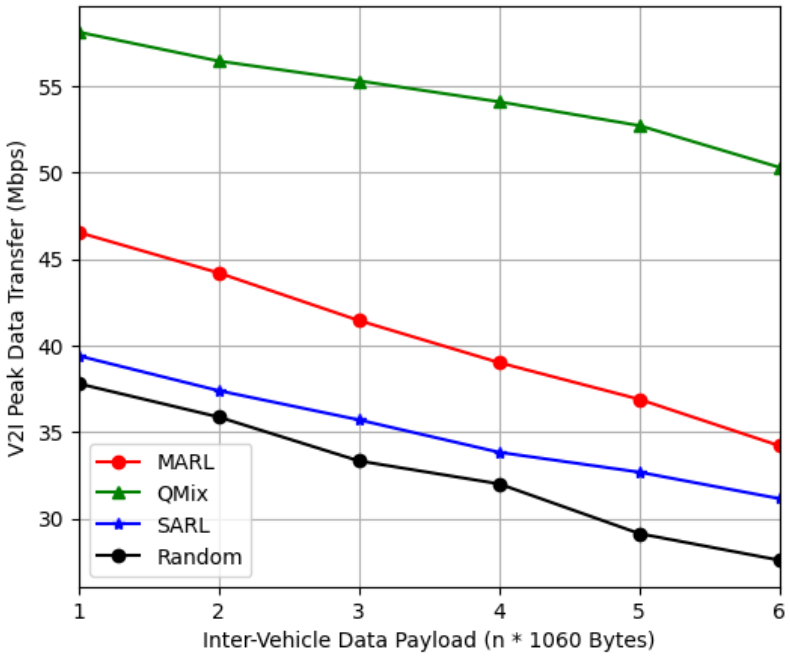


Figure 5. Effect of V2V data size on V2I total capacity performance.

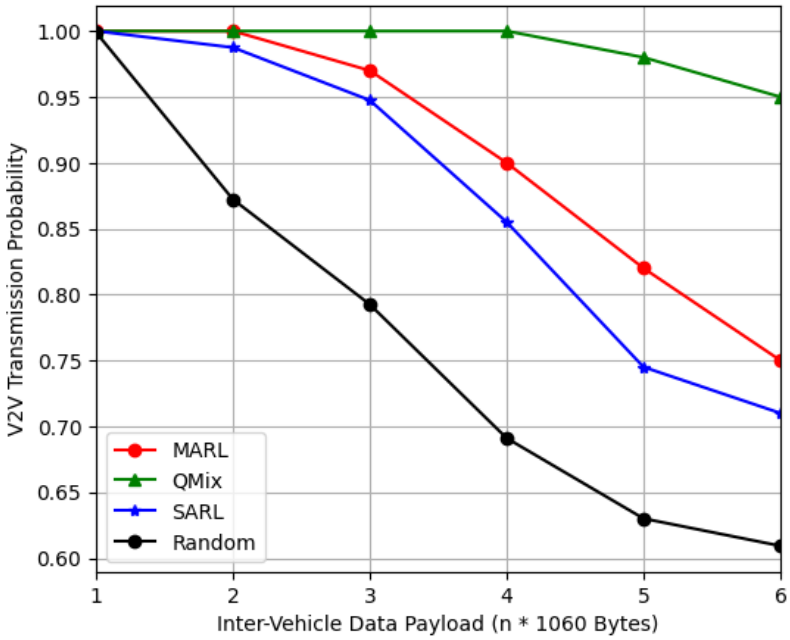


Figure 6. V2V transmission success rate across different data sizes.

5.1. V2I Data Transfer Capability

Figure 5 delivers a powerful illustration of how the proposed QMIX scheme transcends existing approaches, particularly the baseline MARL scheme, in ensuring seamless V2I performance, especially as V2V data sizes inflate. Unlike the diminishing V2I transmission rates exhibited by other methods as data loads increased, QMIX maintained an impressive consistency, thus showcasing remarkably high rates over its competitors. From Figure 5, we can observe that when the payload size was 1×1060 bytes, the corresponding improvement of the QMIX was 24.91% as compared with the benchmark MARL technique. As we increased the payload size, the efficiency of the proposed QMIX technique remained improved at 27.7%, 33.46%, 38.67%, 42.86%, and 47.17%, respectively. Figure 5 shows that the average improvement of the proposed technique over MARL was 35.795%. This unwavering performance signifies QMIX's superior adaptability and effectiveness in handling diverse data demands within V2X communication scenarios.

Furthermore, QMIX established itself as the undisputed winner across the spectrum of V2V data sizes. Regardless of the payload size, it consistently outperformed the baseline scheme, thus achieving the highest V2I transmission rates. This consistent excellence emphasizes QMIX's robustness and its ability to excel in various communication scenarios, thereby solidifying its potential for real-world applications.

By effectively addressing challenges faced by existing approaches, QMIX emerges as a frontrunner in ensuring reliable and efficient V2X communication, thus paving the way for a future of seamless data exchange and improved network performance.

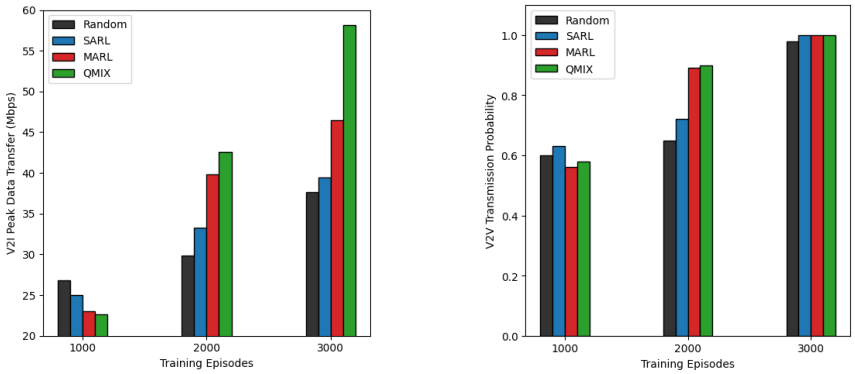
5.2. V2V Communication Reliability

Figure 6 paints a compelling picture of how data size impacts V2V transmission probability under various schemes. While benchmark schemes exhibited a predictable decline in transmission success as data size increased, the proposed QMIX scheme stood out. At small data sizes, QMIX outperformed with a remarkably higher transmission probability, thereby highlighting its effectiveness in handling light communication loads. This consistent success makes it an ideal choice for scenarios where reliable transmission of smaller packets is crucial.

From the Figure 6, it is evident that as the data payload increased, the proposed QMIX technique continued to outperform by a significant margin. For example, we have improvements of 3.09%, 8.89%, 19.5%, and 26.67% for data sizes of 3×1060, 4×1060, 5×1060, and 6×1060 bytes, respectively. Figure 6 shows that the average improvement of the proposed technique over MARL was 14% in the given scenario.

This consistent performance across varying data sizes signifies the strength and potential of QMIX. It underscores its ability to deliver enhanced transmission performance, even in the dynamic and challenging environment of V2V communication. The QMIX adaptation in V2X networks can significantly improve data communication performance.

In addition, it is important to note that the number of training episodes greatly influences the effectiveness of the QMIX method. When the training duration is very short, agents cannot adequately explore their environment, and reinforcement learning-based strategies may not perform better than random baseline approaches. This behavior is demonstrated in Figure 7, where the performance of the V2I (Figure 7a) and V2V (Figure 7b) links is shown for a fixed payload size. We can observe that, particularly in scenarios with limited training time—around 2000 training episodes—the QMIX approach offered significant benefits.



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

Figure 7. Comparison of QMIX and baseline schemes for different numbers of training episodes. (**a**) Effect of V2V data size on V2I capacity. (**b**) V2V transmission success rate.

5.3. Discussion

In this study, we investigated the challenge of multiple Vehicle-to-Vehicle (V2V) links interfering with the spectrum used by Vehicle-to-Infrastructure (V2I) links. This can significantly reduce the data capacity available for V2I communication and hinder the reliability of data transmission for V2V communication. To address this issue, we aimed to improve both spectrum efficiency and message transmission reliability using the Manhattan urban simulation model [29]. Our approach involved employing the multi-agent reinforcement learning method with QMIX, where agents autonomously learn and share insights with the central QMIX network, thereby fostering collaboration towards a common goal.

From the simulation, we observed that the RL techniques initially showed lower performance than baseline schemes. However, as iterations progressed and the RL methods learned, their performance became exceptional. Specifically, as we increased the V2V payload size, our proposed approach maintained superior performance with higher V2I data transfer capability and V2V data transmission reliability.

It is worth noting that implementing the hybrid QMIX approach requires substantial computational resources. This is due to the initial learning phase of the agents, followed by the sharing of insights with the central network, which necessitates additional computational settings and resources.

6. Conclusions

In this article, we have developed a resource sharing scheme based on multi-agent reinforcement learning in vehicular networks. We have utilized the QMIX method to address the non-stationarity and competitive behavior of multi-agent reinforcement learning. The proposed scheme comprises two stages such as hybrid learning and distributed implementation. The QMIX showed that the average improvement of the proposed technique over MARL was 35.79% in V2I data transfer capability for varying data sizes, and the average increment in message transmission probability was 14% for all data sizes.

The proposed deep reinforcement learning-based QMIX approach efficiently manages spectrum sharing. This adaptability helps QMIX to perform well in variable environments, thereby making it a good choice for spectrum sharing. The promising results as presented in Section 5 highlight the impact and difference of hybrid techniques. Moreover, the proposed QMIX method addresses network needs and uses the optimal way to send data to ensure smooth communication and better network performance.

Currently, it is challenging to realize the framework due to the high computational cost practically, but as high-performance chips for AI applications mature, the proposed framework will be easy to develop.

In future work, we plan to conduct a detailed analysis and comparison of the robustness of MARL schemes, including QMIX, and explore the combination of QMIX with other distributed methods such as the Quanitle Transfer Network (QTRAN). This integration

aims to enhance coordination and collaboration among agents in distributed systems. Additionally, we will investigate the integration of QMIX with federated learning frameworks such as Federated Q Learning and Federated Policy Gradient to enable localized decision-making and reduce communication overhead, especially in applications like UAV swarms and healthcare systems.

Author Contributions: Methodology, M.J., Z.U. and A.C.; Formal analysis, M.A.; Investigation, M.N.; Resources, M.A. and A.C.; Writing—original draft, M.N.; Supervision, A.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data can be shared upon request. The data are not publicly available due to we are currently working on this project and exploring other dimensions of the mentioned research topic.

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

Abbreviations

The following abbreviations are used in this manuscript:

Term	Abbreviation
3GPP	3rd Generation Partnership Project
BS	Base Station
CSI	Channel State Information
D2D	Device-to-Device
DNN	Deep Neural Network
DQN	Deep Q-Network
DRL	Deep Reinforcement Learning
ITS	Intelligent Transportation Systems
LTE	Long-Term Evolution
MA-MDP	Multi-Agent Markov Decision Process
MARL	Multi-Agent Reinforcement Learning
MDP	Markov Decision Process
QTRAN	Quantile Transfer Network
RAS	Robotics and Autonomous Systems
RB	Resource Block
RL	Reinforcement Learning
RSU	Roadside Unit
SINR	Signal-to-Interference-plus-Noise Ratio
TD	Temporal Difference
VCN	Vehicular Communication Networks
VN	Vehicular Networks
V2I	Vehicle-to-Infrastructure
V2V	Vehicle-to-Vehicle)
V2X	Vehicle-to-Everything communication
WMMSE	Weighted Minimum Mean Square Error

References

1. El Zorkany, M.; Yasser, A.; Galal, A.I. Vehicle to vehicle "V2V" communication: Scope, importance, challenges, research directions and future. *Open Transp. J.* **2020**, *14*, 86–98. [CrossRef]
2. Liang, L.; Peng, H.; Li, G.Y.; Shen, X. Vehicular Communications: A Physical Layer Perspective. *IEEE Trans. Veh. Technol.* **2017**, *66*, 10647–10659. [CrossRef]
3. Naeem, M.; Bashir, S.; Ullah, Z.; Syed, A.A. A near optimal scheduling algorithm for efficient radio resource management in multi-user MIMO systems. *Wirel. Pers. Commun.* **2019**, *106*, 1411–1427. [CrossRef]
4. Dhinesh Kumar, R.; Rammohan, A. Revolutionizing Intelligent Transportation Systems with Cellular Vehicle-to-Everything (C-V2X) technology: Current trends, use cases, emerging technologies, standardization bodies, industry analytics and future directions. *Veh. Commun.* **2023**, *43*, 100638.

5. Naeem, M.; Coronato, A.; Ullah, Z.; Bashir, S.; Paragliola, G. Optimal user scheduling in multi antenna system using multi agent reinforcement learning. *Sensors* **2022**, *22*, 8278. [CrossRef] [PubMed]
6. Aung, P.S.; Nguyen, L.X.; Tun, Y.K.; Han, Z.; Hong, C.S. Deep Reinforcement Learning based Joint Spectrum Allocation and Configuration Design for STAR-RIS-Assisted V2X Communications. *IEEE Internet Things J.* **2023**, *11*, 11298–11311. [CrossRef]
7. Naeem, M.; Bashir, S.; Khan, M.U.; Syed, A.A. Performance comparison of scheduling algorithms for MU-MIMO systems. In Proceedings of the 2016 13th International Bhurban Conference on Applied Sciences and Technology (IBCAST), Islamabad, Pakistan, 12–16 January 2016; pp. 601–606.
8. Baek, S.; Kim, D.; Tesanovic, M.; Agiwal, A. 3GPP New Radio Release 16: Evolution of 5G for Industrial Internet of Things. *IEEE Commun. Mag.* **2021**, *59*, 41–47. [CrossRef]
9. Sun, S.; Hu, J.; Peng, Y.; Pan, X.; Zhao, L.; Fang, J. Support for vehicle-to-everything services based on LTE. *IEEE Wirel. Commun.* **2016**, *23*, 4–8. [CrossRef]
10. Raza, S.; Wang, S.; Ahmed, M.; Anwar, M.R.; Mirza, M.A.; Khan, W.U. Task Offloading and Resource Allocation for IoV Using 5G NR-V2X Communication. *IEEE Internet Things J.* **2022**, *9*, 10397–10410. [CrossRef]
11. Rashid, T.; Samvelyan, M.; De Witt, C.S.; Farquhar, G.; Foerster, J.; Whiteson, S. Monotonic value function factorisation for deep multi-agent reinforcement learning. *J. Mach. Learn. Res.* **2020**, *21*, 7234–7284.
12. Jameel, F.; Khan, W.U.; Kumar, N.; Jäntti, R. Efficient Power-Splitting and Resource Allocation for Cellular V2X Communications. *IEEE Trans. Intell. Transp. Syst.* **2021**, *22*, 3547–3556. [CrossRef]
13. Li, X.; Ma, L.; Xu, Y.; Shankaran, R. Resource Allocation for D2D-Based V2X Communication with Imperfect CSI. *IEEE Internet Things J.* **2020**, *7*, 3545–3558. [CrossRef]
14. Gao, J.; Khandaker, M.R.A.; Tariq, F.; Wong, K.K.; Khan, R.T. Deep Neural Network Based Resource Allocation for V2X Communications. In Proceedings of the 2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall), Honolulu, HI, USA, 22–25 September 2019; pp. 1–5. [CrossRef]
15. Abbas, F.; Fan, P.; Khan, Z. A Novel Low-Latency V2V Resource Allocation Scheme Based on Cellular V2X Communications. *IEEE Trans. Intell. Transp. Syst.* **2019**, *20*, 2185–2197. [CrossRef]
16. 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.* **2020**, *7*, 6380–6391. [CrossRef]
17. Gui, J.; Lin, L.; Deng, X.; Cai, L. Spectrum-Energy-Efficient Mode Selection and Resource Allocation for Heterogeneous V2X Networks: A Federated Multi-Agent Deep Reinforcement Learning Approach. *IEEE/ACM Trans. Netw.* **2024**. [CrossRef]
18. Yuan, Y.; Zheng, G.; Wong, K.K.; Letaief, K.B. Meta-Reinforcement Learning Based Resource Allocation for Dynamic V2X Communications. *IEEE Trans. Veh. Technol.* **2021**, *70*, 8964–8977. [CrossRef]
19. Rasheed, I. Dynamic mode selection and resource allocation approach for 5G-vehicle-to-everything (V2X) communication using asynchronous federated deep reinforcement learning method. *Veh. Commun.* **2022**, *38*, 100532. [CrossRef]
20. Li, J.; Zhao, J.; Sun, X. Deep Reinforcement Learning Based Wireless Resource Allocation for V2X Communications. In Proceedings of the 2021 13th International Conference on Wireless Communications and Signal Processing (WCSP), Changsha, China, 20–22 October 2021; pp. 1–5. [CrossRef]
21. Du, W.; Ding, S. A survey on multi-agent deep reinforcement learning: From the perspective of challenges and applications. *Artif. Intell. Rev.* **2021**, *54*, 3215–3238. [CrossRef]
22. Roderick, M.; MacGlashan, J.; Tellex, S. Implementing the deep q-network. *arXiv* **2017**, arXiv:1711.07478.
23. Luu, Q.T. Q-Learning vs. Deep Q-Learning. 2023. Available online: https://www.baeldung.com/cs/q-learning-vs-deep-q-learning-vs-deep-q-network (accessed on 19 February 2024).
24. Rashid, T.; Farquhar, G.; Peng, B.; Whiteson, S. Weighted qmix: Expanding monotonic value function factorisation for deep multi-agent reinforcement learning. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 10199–10210.
25. Yang, Y.; Hao, J.; Liao, B.; Shao, K.; Chen, G.; Liu, W.; Tang, H. Qatten: A general framework for cooperative multiagent reinforcement learning. *arXiv* **2020**, arXiv:2002.03939.
26. Li, C.; Fussell, L.; Komura, T. Multi-agent reinforcement learning for character control. *Vis. Comput.* **2021**, *37*, 3115–3123. [CrossRef]
27. Chen, X.; Xiong, G.; Lv, Y.; Chen, Y.; Song, B.; Wang, F.Y. A Collaborative Communication-Qmix Approach for Large-scale Networked Traffic Signal Control. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 3450–3455. [CrossRef]
28. Yao, X.; Wen, C.; Wang, Y.; Tan, X. Smix (λ): Enhancing centralized value functions for cooperative multiagent reinforcement learning. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *34*, 52–63. [CrossRef] [PubMed]
29. 3GPP. 3rd Generation Partnership Project; Technical Specification Group Radio Access Network; Study on LTE-based V2X Services; (Release 14). 3GPP TS 36.300 V14.0.0. 2016. Available online: https://www.etsi.org/deliver/etsi_ts/136300_136399/136361/14.00.00_60/ts_136361v140000p.pdf (accessed on 24 January 2024).