Deep reinforcement learning techniques for vehicular networks: recent advances and future trends towards 6G

Abdelkader Mekrache, Abbas Bradai, Emmanuel Moulay, Samir Dawaliby

Deep reinforcement learning techniques for vehicular networks: recent advances and future trends towards 6G

Abdelkader Mekrache[a,*], Abbas Bradai[a], Emmanuel Moulay[a], Samir Dawaliby[b]

[a] *XLIM (UMR CNRS 7252), Université de Poitiers, 11 bd Marie et Pierre Curie, 86073 Poitiers Cedex 9, France*
[b] *Audensiel, 93 rue nationale, 92100 Boulogne-Billancourt, France*

Abstract

Employing machine learning into 6G vehicular networks to support vehicular application services is being widely studied and a hot topic for the latest research works in the literature. This article provides a comprehensive review of research works that integrated reinforcement and deep reinforcement learning algorithms for vehicular networks management with an emphasis on vehicular telecommunications issues. Vehicular networks have become an important research area due to their specific features and applications such as standardization, efficient traffic management, road safety, and infotainment. In such networks, network entities need to make decisions to maximize network performance under uncertainty. To achieve this goal, Reinforcement Learning (RL) can effectively solve decision-making problems. However, the state and action spaces are massive and complex in large-scale wireless networks. Hence, RL may not be able to find the best strategy in a reasonable time. Therefore, Deep Reinforcement Learning (DRL) has been developed to combine RL with Deep Learning (DL) to overcome this issue. In this survey, we first present vehicular networks and give a brief overview of RL and DRL concepts. Then we review RL and especially DRL approaches to address emerging issues in 6G vehicular networks. We finally discuss and highlight some unresolved challenges for further study.

Keywords: Vehicular networks, reinforcement learning, deep reinforcement learning, 6G wireless networks

*Corresponding author

Email addresses: abdelkader.mekrache@univ-poitiers.fr (Abdelkader Mekrache), abbas.bradai@univ-poitiers.fr (Abbas Bradai), emmanuel.moulay@univ-poitiers.fr (Emmanuel Moulay), s.dawaliby@audensiel.fr (Samir Dawaliby)

Contents

1. Introduction

In the last two decades, vehicular networks rapidly emerged as one of the hottest topics over which research community is focusing to exploit the wide range of applications ranging from road safety improvements to traffic efficiency optimization and from autonomous driving to ubiquitous internet access on vehicles [1]. This new generation of networks will significantly impact society and the daily lives of individuals across the world. Lately, various problems have arisen in vehicular communications and have been addressed by the research community such as clustering and routing [2, 3], processing large volume of data [4], content distribution [5], and data forwarding [6]. Moreover, vehicular networks bring unprecedented challenges unseen in conventional wireless networks [7] due to:

- Highly dynamic mobility scenarios ranging from low-speed vehicles (for example, less than 60km/h) to high-speed cars/trains (for example, 500km/h or higher).
- Various data services with different Quality of Service (QoS) requirements in terms of reliability, latency, and data rates, for example in-vehicle multimedia entertainment, video games, ultra-reliable and low-latency delivery of safety messages, high-precision map downloads, etc.
- Expected explosive growth of vehicular communication devices amid an increasingly fragmented and congested spectrum.

Meanwhile, with the help of high-performance computing and storage facilities, as well as various advanced on-board sensors such as lidar, radar, and cameras, vehicles will be more than just a simple means of transportation. They are generating, collecting, storing, processing, and transmitting large

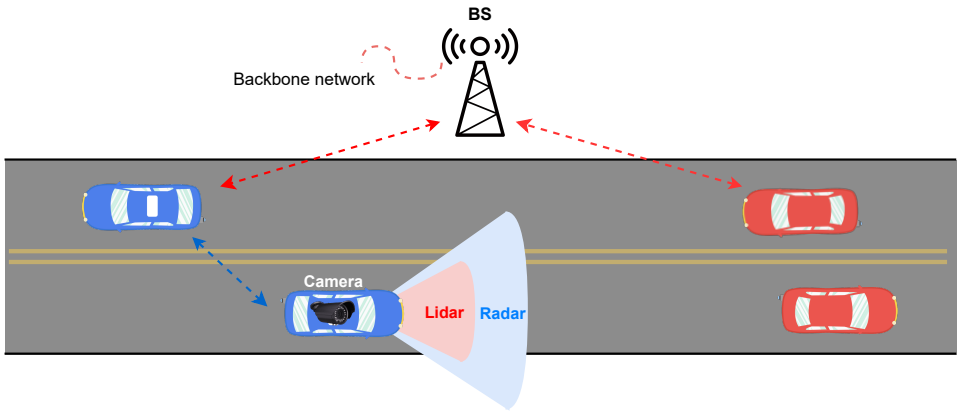

Figure 1: An illustrative structure of vehicular networks [1].

amounts of data, which are used to make driving safer and more convenient [1], as shown in Figure 1. Exploiting this data, machine learning algorithms, especially Reinforcement Learning (RL) algorithms, have been used to solve the above-mentioned challenges, because traditional communication strategies are not meant to handle such rich information.

As an important supporting technology of artificial intelligence, machine learning has been successfully applied in many fields, including computer vision, medical diagnosis, search engines, and speech recognition [8]. It is a field of research that allows computers to learn without explicit programming. Machine learning techniques can generally be divided into supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the purpose of the learning agent is to learn to map inputs to general rules with example inputs and the expected output provided, these example inputs constitute a labeled data set. In unsupervised learning, there is no need to label the data, and the agent tries to find some structure from its inputs. In reinforcement learning, the agent continuously interacts with the dynamical environment and tries to develop a good strategy based on the immediate reward/cost of environmental feedback.

Reinforcement learning has recently been used in vehicular networks as an evolving method to solve different problems and challenges effectively. In order to achieve the goals of various networks, including, for example, throughput maximization and energy consumption minimization, network entities such as vehicles, and base stations need to make local and autonomous decisions, e.g., spectrum access, data rate selection, transmission power control, and base station association, under unpredictable stochastic

conditions. Modern networks, however, are large-scale and complex, and thus the computational complexity of the techniques quickly becomes unmanageable. As a result, to overcome the challenge, deep reinforcement learning has been developing to be an alternative solution [9].

Furthermore, 6G aims to connect any smart device, from smartphones to intelligent vehicles, to the Internet. It will deliver innovative and high-quality services like holographic communication, augmented reality/virtual reality, and many others [10]. In addition, It will focus on Quality of Experience (QoE) to provide rich experiences from 6G technology. Notably, 6G technology will face complex issues and challenges in vehicular networks, which RL and DRL algorithms will be crucial in resolving.

Although there are some surveys related to machine learning and vehicular networks, they do not focus on the recent advances of the applications of RL and DRL for vehicular networks management. In [1], the authors presented some examples of application of machine learning to solve problems in vehicular networks, but they have not focused on RL algorithms where only seven articles have been discussed. In [11], *Yuan et al.* presented machine learning techniques for Next-Generation Intelligent Transportation Systems (ITS). Researchers presented in [9] only the application of DRL in communication and networking in general. A survey on resource allocation in vehicular networks was presented in [7]. Still, it did not focus on the part of machine learning where only seven papers were discussed. Finally, only Multi-Agent Reinforcement Learning (MARL) methods for vehicular networks were discussed in [12]. There are also some existing surveys on 6G integration with vehicular networks. In [13], *Tang et al.* provided a survey on various ML techniques applied to communication, networking, and security parts in vehicular networks and envision the ways of enabling AI toward a future 6G vehicular network. They concentrated on presenting supervised learning techniques and not RL where only four works were reviewed. In [14], the authors provided an overview on the recent advances of machine learning in 6G vehicular networks. They have also identified a number of key enabling technologies and revolutionary elements of next-generation 6G-V2X networks. However, they did not focus on RL techniques where only few works were reviewed. A comprehensive survey on green Unmanned Aerial Vehicles (UAV) communications for 6G was presented in [15]. Specifically, the typical UAVs and their energy consumption models are introduced. Then, the typical trends of green UAV communications are provided. Still, it only discusses a single work that applies RL.

In our paper, we provide a comprehensive survey of the current state-of-the-art of application of RL and DRL in vehicular networks and present

open issues in this area. We introduce these works from two aspects: vehicular resource management and vehicular infrastructure management. A classification of reviewed works is shown in Figure 2. The contributions of

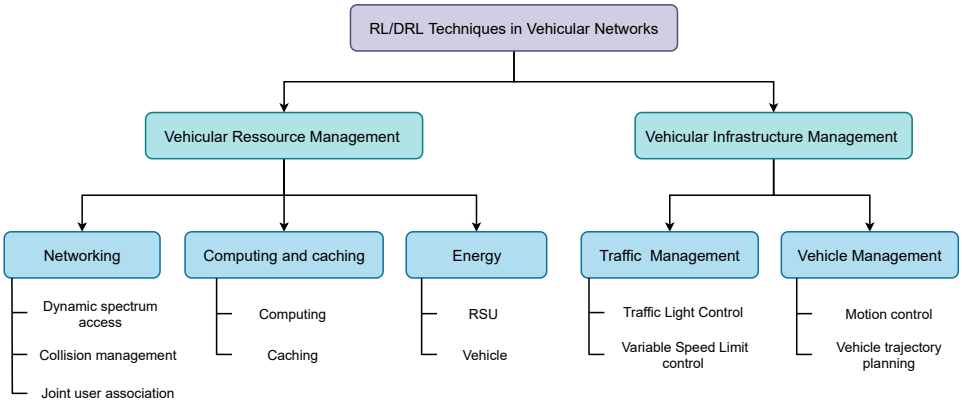

Figure 2: A taxonomy of the application of RL and DRL in vehicular networks.

our survey can be summarized as follows:

- first, we present an overview of vehicular networks;
- second, we present concepts of RL and DRL;
- third, we provide a detailed review of the current state-of-the-art on the application of RL and DRL to vehicular networks;
- finally, we discuss future trends and research directions on how RL and DRL can be applied to benefit future 6G vehicular networks.

The structure of the paper is as follows. We start our discourse in section 2 with a high-level overview of vehicular networks. Section 3 presents the introduction to RL and DRL techniques. Section 4 discusses the application of RL/DRL for vehicular resource management. Section 5 reviews the application of RL/DRL for vehicular infrastructure management. Important challenges, open issues and future directions are outlined in Section 6. Section 7 concludes the paper. The list of abbreviations commonly appeared in this paper is given in Table 1.

2. Vehicular networks

Vehicular networks will enable a wide set of applications and services within the ITS, for improving road safety, traffic efficiency, infotainment,

Abbreviation	Description
3DQN	Double Dueling Deep Q Network
AC/A3C/SAC	Actor Critic/Asynchronous Actor Critic/Soft Actor Critic
AV	Autonomous Vehicle
BS	Base Station
C-V2X	Cellular Vehicle-to-Everything
CNN/DNN/RNN	Convolutional Neural Network/Deep Neural Network/Recurrent Neural Network
CW	Contention Window
DDPG	Deep Deterministic Policy Gradient
DQL	Deep Q-Learning
DQN/DRQN/DDQN	Deep Q-Network/Deep Recurrent Q-Network/Double Deep Q-Network
DSRC	Dedicated Short-Range Communication
EMS	Energy Management System
FNN	Feed-forward Neural Networks
GVRP	Green Vehicle Routing Problem
HetNet/HetVNet	Heterogeneous Networks/Heterogeneous Vehicular Networks
HEV/P-HEV	Hybrid Electric Vehicles/Plug-in Hybrid Electric Vehicles
IoT/IoV	Internet of Things/Internet of Vehicles
ITS	Intelligent Transportation System
LSTM	Long Short Term Memory
LTE	Long Term Evolution
MAC	Meduim Access Control
MEC	Mobile Edge Computing
MDP/POMDP	Markov Decision Process/ Partially Observable MDP
PDR	Packet Delivery Ratio
QoS/QoE	Quality of Service/Quality of Experience
RB	Ressource Block
RL/DRL	Reinforcement Learning/Deep Reinforcement Learning
RSU	Roadside units
SARL/MARL	Single-Agent Reinforcement Learning/Multi-Agent Reinforcement Learning
SNR/SINR	Signal to Noise Ratio/Signal to Interference plus Noise Ratio
TDD	Time Division Duplex
TLC	Traffic Light Control
UAV/USV	Unmanned Aerial Vehicles/Unmanned Surface Vehicles
UE	User Equipment
UL/DL	Uplink/Downlink
URLLC	Ultra-Reliable and Low Latency Communications
V2I	Vehicle-to-Infrastructure
V2N	Vehicle-to-Network
V2P	Vehicle-to-Pedestrian
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-Everything
VANET	Vehicular Ad-Hoc Network
VRU	Vulnerable Road User
VSL	Variable Speed Limit
VU/VUE-pairs	Vehicle User/Vehicle User Equipment-pairs

Table 1: List of abbreviations and notations.

and supporting autonomous driving. To support these services, Vehicle-to-Everything (V2X) communications enable the exchange of information between vehicles, infrastructure, and pedestrians, using different wireless communication technologies [16]. In this section, we first present the vehicular communication modes, and a classification of the use cases of vehicular networks. We then present the two main vehicular communication technologies DSRC and C-V2X followed by a focus on the second as it is the main interest of the article. For more details, a survey on vehicular networks for smart roads is presented in [17].

2.1. Vehicular communication modes

As shown in Figure 3, 3GPP identifies four types of V2X communication modes: Vehicle-to-Vehicle (V2V), Vehicle-to-Pedestrian (V2P), Vehicle-to-Infrastructure (V2I) and Vehicle-to-Network (V2N) [18]. Through the use

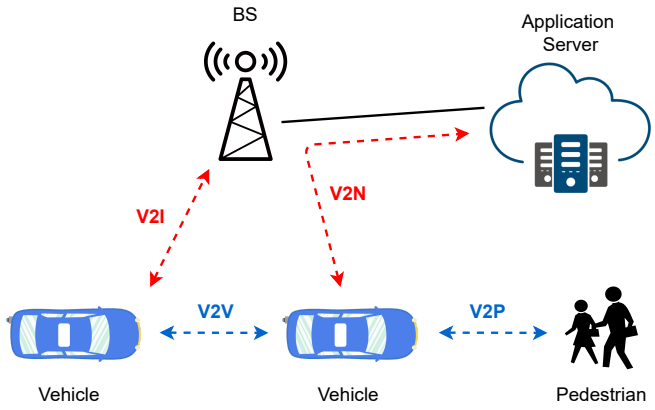

Figure 3: Types of V2X application support in 3GPP [19].

of "cooperation awareness", the above four types of V2X applications can be used jointly to provide end users with smarter services. For example, vehicles, pedestrians, application servers, and road infrastructure can obtain local environmental information by receiving messages from sensors in nearby or other vehicles, enabling smarter services such as autonomous driving, vehicle warnings, and enhanced traffic management [20, 21]. The four types of V2X communication modes are as follows:

- *V2V and V2P* modes cover direct communication between vehicle User Equipment (UE) and between vehicles and Vulnerable Road Users (VRU), such as pedestrians, cyclists, motorcycles, and wheelchair users;

- *V2I* refers to the communication between the vehicle and the roadside infrastructure, for example, a RSU implemented in an eNodeB or as a standalone fixed UE;

- *V2N* enables vehicular UEs to communicate with an application server that supports V2N applications, which provides centralized control and distribution of traffic, road, and service information.

2.2. Vehicular networks use cases

The wide set of vehicular use cases have been categorized in [20] into four categories as follows:

- *Safety and traffic efficiency.* V2V/V2P event-driven and periodic messages hold the transmitting vehicle's location and kinematics parameters, allowing other vehicles and VRUs to sense the environment and support applications such as:
  - forward collision warning, used to notify the driver of an imminent tail collision with the vehicle ahead;
  - cooperative adaptive cruise control system that allows a group of nearby vehicles to share the same path (also called platooning);
  - VRU safety to alert vehicles the presence of VRUs.

- *Autonomous driving.* In order for autonomous cars to be truly autonomous for navigation, it is essential that the vehicle is aware of its position, surrounding environment and nearby vehicles [22]. In addition, these vehicles may be very close to each other and drive at higher speeds (up to 200km/h).

- *Tele-operated driving.* Tele-operated driving will be used in environments that are dangerous or uncomfortable for people, such as nuclear accidents, earthquakes, road construction and snow removal, the drones on the wheels may be used by the driver to perform driving tasks. In fact, the driver will be located outside the vehicle and will control it using the camera, status and sensor data.

- *Vehicular Internet and infotainment.* These applications are intended for the comfort of the driver and passengers. They essentially provide services such as mobile Internet access, messaging, discussion between vehicles, collaborative network games. [23].

- *Remote diagnostics and management.* The V2X application server owned by the car manufacturer or the vehicle diagnostic center can retrieve the information periodically sent by the vehicle in V2N mode to track its status for remote diagnosis.

2.3. Vehicular communication technologies

So far, there are two main methods of V2X communications: Dedicated Short-Range Communication (DSRC) and cellular-based vehicular communication. DSRC is supported by a series of standards, including the IEEE 802.11p amendment for Wireless Access in a Vehicular Environments (WAVE), and the IEEE 1609.1-4 standard for resource management, security, network services, and multi-channel operation [24]. Moreover, it is known that Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) medium access technique used in IEEE 802.11p is unsuitable for critical communication scenarios [25], i.e. QoS in vehicular networks applications cannot be guaranteed for safety-critical messages and other real-time transmissions. On the other hand, 3GPP has been developing cellular vehicular communications, also known as Cellular Vehicle-to-Everything (C-V2X), aimed at operating on cellular networks such as Long Term Evolution (LTE) and 5G New Radio (5G NR) [7] that can offer high data rate services and wide coverage. An overview of vehicular networks is shown in Figure 4. Both V2X technologies have their own advantages and limitations

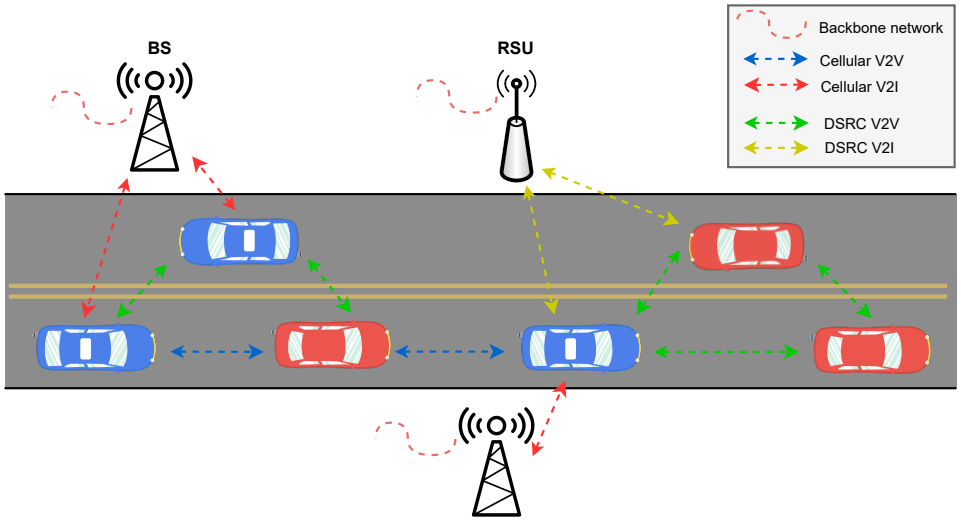

Figure 4: Overview of vehicular networks [7].

when used in a vehicular environment. As a result, it has been proposed to integrate it into heterogeneous vehicular networks to exploit their unique advantages while addressing their individual disadvantages. The main differences between C-V2X and DSRC-V2X is summarized in Table 2.

	C-V2X	**DSRC-V2X**
Synchronization	Synchronous	Asynchronous
Resource Multiplexing Across Vehicles	Frequency Division Multiplexing (FDM) and Time Division Multiplexing (TDM)	TDM Only
Channel Coding	Turbo	Convolutional
Waveform	Single Carrier FDM (SC-FDM)	Orthogonal FDM (OFDM)
Retransmission mechanism	Hybrid Automatic Repeat Request (HARQ)	No HARQ
Resource Selection	Semipersistent transmission with relative energy-based selection	Carrier Sense Multiple Access with Collision Avoidance (CSMA-CA)
Advantages	Wide coverage and high data rate services	Support high-density vehicular communications
Disadvantages	Do not support decentralized communication as the networks may become easily overloaded in situation with very high vehicle density, e.g. traffic jams	Limited coverage, low data rate, limited QoS guarantee, unbounded channel access delay

Table 2: Comparison of C-V2X and IEEE 802.11p [19, 7].

Since the interest of this article is 6G vehicular networks, we will give more details about Cellular Vehicle-to-Everything (C-V2X) in particular beyond 5G and 6G V2X in the next subsection.

2.4. Cellular vehicular networks

In this subsection, we first show the evolution of cellular vehicular networks. Next, we talk about the need to migrate beyond 5G/6G vehicular networks to deploy future services. Finally, we present the key upcoming technologies that will be used, as they form the basis for high data transmissions between vehicles and infrastructure.

2.4.1. Evolution of cellular vehicular networks

3GPP Release 12 (Rel. 12) was the first standard to introduce direct Device-to-Device (D2D) communications using cellular technologies for proximity services (ProSe) [26]. This work was used by 3GPP to develop LTE V2X, the first cellular V2X (C-V2X) standards based on the 4G Long Term Evolution (LTE) air interface. LTE V2X was developed under Release 14

(Rel. 14) [27]. Release 14 focuses on providing data transport services for fundamental road safety services such as Cooperative Awareness Messages (CAM), Basic Safety Messages (BSM), or Decentralized Environmental Notification Messages (DENM). And it was further enhanced in Release 15 (Rel. 15) (known as LTE-eV2X) in terms of higher reliability (employing transmit diversity), lower latency (with the aid of resource selection window reduction), and higher data rates (using carrier aggregation and higher order modulation e.g., 64-QAM) [14]. 5G New Radio (5G NR) V2X technology was also launched in Release 15, announced in 2019, to support advanced V2X services such as vehicle platooning, advanced driver assistance, remote driving, and extended sensors [28]. Note that in Release 16, the 3GPP announced the second phase of 5G NR, which intends to improve Ultra Reliable Low Latency Communication (URLLC) and throughput. The evolution of V2X communications is summarized in Figure 5.

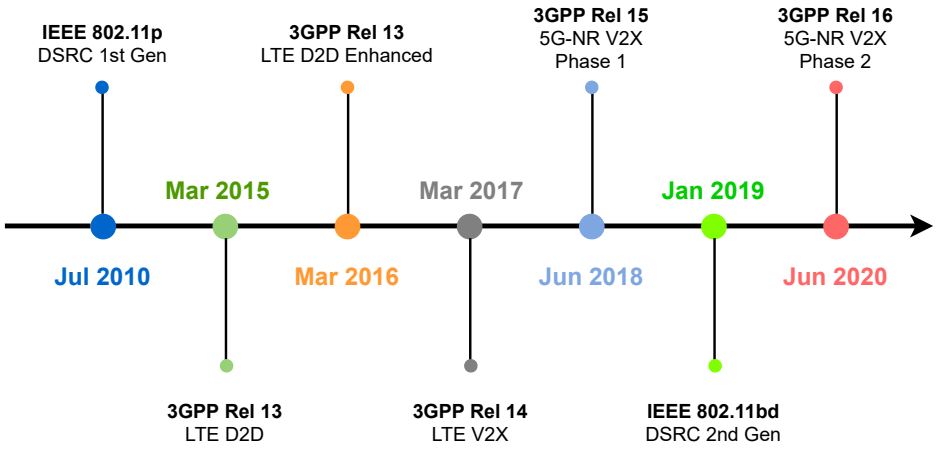

Figure 5: Evolution of V2X communications [14].

2.4.2. Migration to beyond 5G/6G vehicular networks

Although 5G-NR V2X provides better performance with advanced services, it does so by investing more in spectrum and hardware resources while inheriting the underlying mechanisms and system architectures of LTE-based V2X [29]. Meanwhile, due to urbanization, increased living standards, and technology improvements, it is expected that the number of autonomous vehicles will significantly expand in the future. Furthermore, the increasing need for a variety of future services in autonomous vehicles, such as 3D displays, holographic control display systems, immersive entertainment,

and better in-car infotainment, will provide new communication problems to the V2X network [30]. All of these developments will present new scientific and technical challenges for vehicular networks in terms of data rate, latency, spectral/energy/cost efficiency, coverage, intelligence level, networking, and security, among other things [13]. As a result, a major paradigm shift away from traditional communication networks and toward more versatile and diverse network approaches is required. Such a transformation is expected to begin with the recently proposed beyond 5G/6G wireless communication network, which aims to combine terrestrial and several non-terrestrial communication networks. This will enable truly intelligent and ubiquitous V2X systems with significantly improved reliability and security, significantly higher data rates, massive and hyper-fast wireless access, and much smarter, longer, and greener three-dimensional (3D) communication coverage. It will be similar to 5G but with higher speed, lower latency, and much-improved bandwidth [31]. It is foreseen that 6G will work in conjunction with machine learning (ML) not only to reveal the full capacity of radio signals by evolving into intelligent and autonomous radios, but also to introduce a series of new features such as enhanced context-awareness, self-aggregation, adaptive coordination and self-configuration [32].

2.4.3. Essentiel technologies used in cellular vehicular networks

The overall 5G network architecture includes various essential technologies like Device-to-Device (D2D) communications, Non Orthogonal Multiple Access (NOMA), Ultra-Dense Network (UDN), Small Cell Access (SCA), Multiple-Input Multiple-Output (MIMO), massive MIMO, and Cognitive Radio (CR) [33]. These technologies will focus on meeting all 5G requirements which are considered the minimum requirements for 2020. D2D communications solves the cellular system's network capacity problem by allowing direct device connectivity without involving the network. NOMA offers higher spectral efficiency by using the same frequency resource for multiple users. UDN is useful for managing ultra-high density of users while increasing network capacity. The SCA network increases coverage and offloads data traffic. MIMO is useful for increasing the diversity gain in order to handle a maximum number of users. Massive MIMO is an extension of traditional MIMO systems and is useful for enabling massive connectivity in the network. Cognitive Radio is another useful technology for maximizing the use of available radio bands by adjusting various parameters for simultaneous transmission. In 5G NR, two frequency bands have been specified according to 3GPP release 15. These two frequency bands are FR1 (under 6GHz) and FR2 (over 24GHz) [34]. The millimeter-wave band (mmWave)

is included in the FR2 band, which uses very high frequencies for enhancing data rates. Due to the large propagation losses associated with high frequencies, mmWave is limited to small areas. The mmWave limitation can be overcome by boosting the antenna gain via beamforming. This latter is a technique of focusing the maximum power of signal towards the direction of the user. Various other techniques of beamforming like beam merging and beam broadening help in extending the concept of beamforming in 5G [35]. There are various added features of 5G NR overviewed in [36].

To achieve the ambitious goals mentioned in 2.4.2, beyond 5G/6G will require the integration of a range of disruptive technologies including more robust and efficient air interfaces, resource allocation, decision makeing, and computing. These technologies are introduced in details in [14]. The authors classified them into two categories: *revolutionary V2X technologies* and *evolutionary V2X technologies*. Strength, open challenges, maturity, and enhancing areas of these technologies are summarized in Table I of [14].

3. Reinforcement learning: an overview

Reinforcement learning (RL) is one of the most successful artificial intelligence frameworks and the most similar machine learning paradigm to human learning. In this section, we first cover the basics of Markov decision processes, Reinforcement Learning (RL), and Deep Learning (DL) techniques, which are important branches of machine learning. Next, we discuss Deep Reinforcement Learning (DRL) that combines DL and RL to produce more effective and stable function approximations, especially for high-dimensional and infinite-state problems.

3.1. Markov decision process

Markov Decision Processes (MDP) is a discrete-time stochastic control process [37] that provides a mathematical framework for modeling decision-making problems. RL is formally defined as an MDP, which consists of:

- S denotes a set of states plus a distribution of starting states $p(s_0)$;
- A denotes a set of actions;
- transition dynamics $T(s_{t+1}|s_t, a_t)$ that map a state-action pair at time t onto a distribution of states at time $t + 1$;
- an immediate/instantaneous reward function $R(s_t, a_t, s_{t+1})$;

- a discount factor $\gamma \in [0, 1]$, where lower values place greater importance on immediate rewards.

In general, policy π is a mapping from states to a probability distribution over actions $\pi : S \mapsto p(A = a \mid S)$. If the MDP is episodic, i.e. the state is reset after each episode of duration T, then the sequence of states, actions, and rewards in an episode constitute the policy's trajectory or rollout. Every policy's trajectory accumulates rewards from the environment, resulting in the return $R = \sum_{t=0}^{T-1} \gamma^t r_{t+1}$. The aim of RL is to find an optimal policy π^* to maximize the expected return from all states:

$$\pi^* = \underset{\pi}{\operatorname{argmax}}\, \mathbb{E}[R \mid \pi]. \tag{1}$$

Non-episodic MDPs with $T = \infty$ can also be considered. In this case, $\gamma < 1$ prevents an infinite sum of rewards from being accumulated. Furthermore, methods based on complete trajectories are no longer valid, but those based on a finite set of transitions are. The discount factor γ determines the importance of future rewards compared with the immediate reward. If $\gamma = 0$, the agent is "myopic", meaning it only considers maximizing its immediate reward. In contrast, if $\gamma \to 1$, the agent will aim for a long-term higher reward.

The Markov property, which guarantees that only the current state influences the next state, is a key concept underlying RL. In other words, the future is conditionally independent of the past provided the present state. While the majority of RL algorithms make this assumption, it is rather impractical since it needs the states to be completely observable. Partially Observable MDPs [38] (POMDPs) are a generalization of MDPs in which the agent receives an observation $o_t \in \Omega$, where the distribution of the observation $p(o_{t+1} \mid s_{t+1}, a_t)$ is determined by the current state and the previous action. Given the previous belief state, the action taken, and the current observation, POMDP algorithms usually maintain a belief over the current state. A more common approach in DL is to use Recurrent Neural Networks (RNNs), which are dynamical systems as opposed to Feed-forward Neural Networks (FNNs) [39].

3.2. Reinforcement learning

RL is an important branch of ML that is commonly used in the literature to solve MDPs. An agent can learn its optimal policy π^* through interaction with its environment in a RL process. At each timestamp t, the agent observes the state s_t of its environment and performs an action a_t,

resulting in a new state s_{t+1} and receiving its immediate reward r_{t+1} as seen in Figure 6. The observed information, i.e. the immediate reward and new

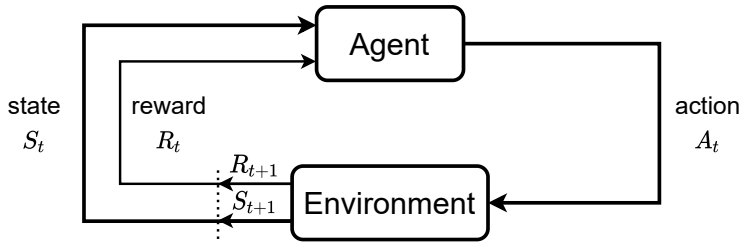

Figure 6: Reinforcement learning.

state, is used to adjust the agent's policy π, and this process will be repeated until the agent's policy approaches to the optimal policy π^*, i.e. $\pi \rightarrow \pi^*$. RL algorithms can be split into two main kinds of methods as shown in Figure 7: (i) methods based on value functions "Value-based algorithms" and (ii) methods based on policy search "Policy-based algorithms". There is also hybrid actor-critic approaches that employs both value functions and policy search.

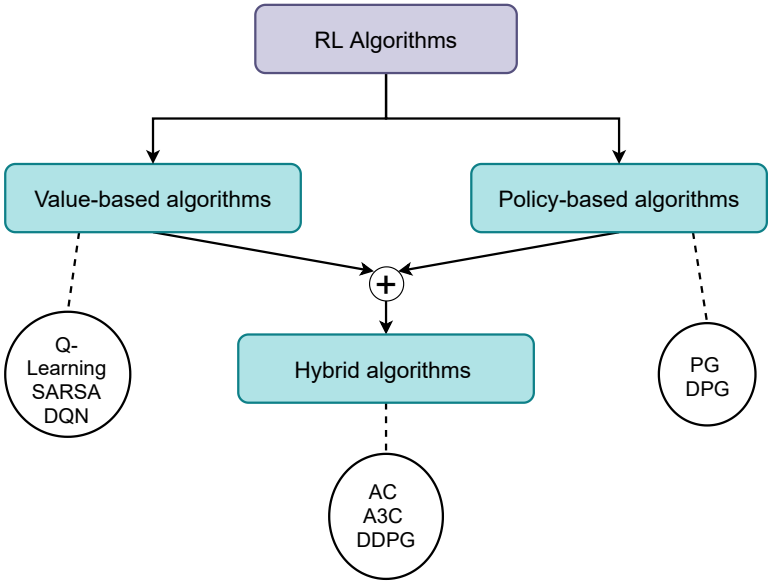

Figure 7: Kinds of RL Algorithms.

- *Value-based algorithms* are based upon temporal difference learning to

obtain the value function V_π which estimates returns when starting in state s and following π: $V^\pi(s) = \mathbb{E}[R \mid s, \pi]$. The optimal policy π^* has a corresponding state-value function $V^*(s) = max_\pi V^\pi(s)$ for all $s \in S$. If $V^*(s)$ is available then the optimal policy π^* can be retrieved by choosing among all actions available at s_t and picking the action a that maximizes $\mathbb{E}_{s_{t+1}\ T(s_{t+1}|s_t,a)}[V^*(s_{t+1})]$. Since the transition dynamics T are unavailable in the RL setting, another function called the state-action value $Q^\pi(s, a)$ is built, which is similar to V^π except that the initial action a is provided and r is only followed from the succeeding state onward: $Q^\pi(s, a) = \mathbb{E}[R \mid s, a, \pi]$. For instance Q-Learning, SARSA, and DQL are three typical value-based RL algorithms.

- *Policy-based algorithms* directly learn optimal policy π^* or try to obtain an approximate optimal policy based on the observation. Typically, a parameterized policy π_θ is chosen whose parameters are updated to maximize the expected return $\mathbb{E}[R \mid \theta]$ using either gradient-based or gradient-free optimization. Gradients can provide a strong learning signal on how to improve a parameterized policy. For example, Policy Gradients (PG), Proximal Policy Optimization(PPO), and Trust Region Policy Optimization (TRPO) are typical policy-based RL algorithms [40, 41].

- *Hybrid algorithms* combine value-based algorithms with policy-based algorithms. Their goal is to represent the policy function by policy-based algorithms where updates of policy functions depend on value-based algorithms. For example, Actor Critic (AC), Asynchronous Actor-Critic Agents (A3C), Deterministic Policy Gradients (DPG) and Deep Deterministic Policy Gradients (DDPG) are typical hybrid algorithms.

3.3. Deep learning

DL is legendary in many fields, and its success mostly relies on Artificial Neural Networks (ANNs) [42]. The latter has become a standard tool for data representation. It consists of a network of interconnected nodes that are built to mimic the functioning of the human brain. Each node features a weighted connection to a large number of nodes in neighboring layers. Individual nodes take the input received from connected nodes and calculate output values using weights and a simple function. Because of their high flexibility, non-linearity, and data-driven model building, ANNs, especially Deep Neural Networks (DNNs), have become attractive inductive approaches.

As seen in Table 3, the three major types of neural networks are Fully-connected Neural Networks (FNNs), Convolutional Neural Networks (CNNs), and Recurrent Neural Networks (RNNs). CNNs excel at visual tasks such as exploiting fundamental spatial properties of images and videos. RNNs can effectively classify data's temporal correlations, making them superior for time series tasks. Long Short-Term Memory (LSTM) methods, which use RNNs as units, can learn order dependence in sequence prediction problems. Graph Neural Networks (GNNs) [43] is a type of graph structure that models a collection of nodes (entities) and edges (relationship). Euclidean data is used to train FNNs, CNNs, and RNNs. GNNs, on the other hand, use non-Euclidean data structures for deep learning [11]. Generative Adversarial Networks (GAN) [44] are a clever way of training a generative model by framing the problem as a supervised learning problem with two sub-models: the generator model, which we train to generate new examples, and the discriminator model, which tries to classify examples as real or fake.

Type	**Entities**	**Relations**	**Scenario**
FNN	Units	All-to-all	-
CNN	Grid elements	Local	Spatial correlation
RNN	Time steps	Sequential	Time correlation

Table 3: Neural networks comparison [11].

3.4. Deep reinforcement learning

DRL is an advanced model of RL technique first introduced by DeepMind in [45], in which DL is used as an effective tool to increase the learning rate for RL algorithms [39]. Specifically, during the real-time learning process, the obtained experiences will be saved and used to train the neural network. The latter will be then used to assist the agent in making optimal decisions in real-time. It should be noted that, in contrast to DL techniques, the neural network in the DRL will be trained on a regular basis based on new experiences obtained through real-time interactions with surrounding environments. Figure 8 describes some RL and DRL algorithms. The Q-values can be expressed as a table for a finite number of discrete states and actions. For continuous state spaces, however, function approximators such as DNNs are needed to represent the Q-values. In DQL, for example, a DNN maps from the continuous state space to the Q-values of a set of actions. All available actions' Q-values must be expected. This prohibits the use of very large or continuous action spaces. The use of continuous action spaces

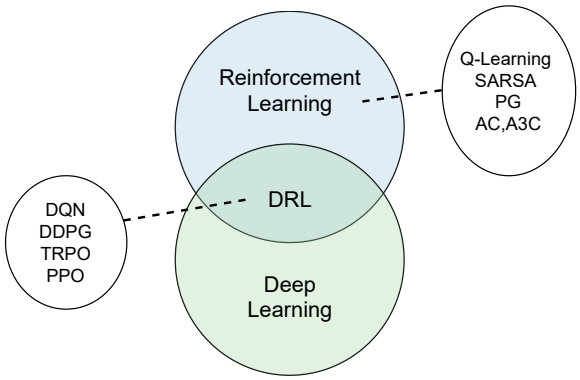

Figure 8: RL, DL, and DRL.

is encouraged by PG methods. In the Soft Actor-Critic (SAC) algorithm, for example, the critic DNN is trained to predict the Q-values for a state-action tuple, and the actor is a second DNN that is used to approximate the Boltzmann distribution over the expected Q-values of available actions.

In this survey paper, some modern DRL algorithms are cited. We list these algorithms in Table 4 and give references for each technique to get details if needed.

	Techniques	**Ref**
Advanced DQL models	Deep Q-Learning	[45]
	Double Deep Q-Learning	[46]
	Deep Q-Learning With Prioritized Experience Replay	[47]
	Dueling Deep Q-Learning	[48]
	Asynchronous Multi-Step Deep Q-Learning	[49]
	Distributional Deep Q-Learning	[50]
	Deep Q-Learning With Noisy Nets	[51]
	Rainbow Deep Q-Learning	[52]
DRL for extensions of MDPs	Deep Deterministic Policy Gradient	[53]
	Deep Recurrent Q-Learning	[54]
	Deep SARSA Learning	[55]

Table 4: Modern DRL algorithms.

4. Vehicular resource management

Resource-intensive use cases, e.g. on-demand multimedia video and live traffic reports, require efficient resource allocation. In support of these use-cases, efficient and intelligent management of local and shared resources is required. To address these issues, RL/DRL was applied to resource management. Next, RL/DRL-based resource management techniques are reviewed considering each resource category: networking, computing and caching, energy.

4.1. Networking

Entities in vehicular networks must make independent decisions, such as channel and Base Station (BS) selections, in order to achieve their own objectives, such as throughput maximization. However, due to the dynamic and unpredictability of network status, this is difficult. Learning algorithms like RL/DRL allow network entities to learn and build knowledge about the networks they are in, allowing them to make the best decisions possible. In this subsection, we look at how RL/DRL can be used in vehicular networks to address the following issues:

- *Dynamic spectrum access.* It allows users to select channels locally to maximize their throughput. These users may not have complete observations of the system, such as channel states. As a result, RL/DRL can be a useful tool for dynamic spectrum access.

- *Collision management.* To improve the performance of data transmission in DSRC vehicular networks, RL/DRL techniques were used for contention window adjustment to minimize the average network delay in congested infrastructure-less Vehicular Ad-Hoc Networks (VANETs).

- *Joint user association and beamforming.* User association shall be established to determine which user to be assigned to which BS. The problems are typically combinatorial and non-convex, requiring almost complete and accurate network information to achieve an optimal strategy. RL/DRL is capable of providing solutions that can be used effectively to address these issues.

4.1.1. Dynamic spectrum access

Deep Q-Learning (DQL) has been extensively adopted in joint channel assignment, power allocation design, and transmission mode selection.

In [56], the authors have developed a new decentralized resource allocation mechanism for cellular V2V communication based on DRL that can be applied to both unicast and broadcast scenarios. The study focuses on resource allocation for V2V links under V2V link latency constraints and minimized interference with V2I links. In order to get the optimal policy, DQL is implemented in both unicast and broadcast scenarios. In the unicast scenario, the structure of RL for V2V links is shown in Figure 9. While the agent corresponds to each V2V link, it interacts with the environment that includes different components beyond the V2V links. The state for characterizing the environment is defined as a set of the instantaneous channel information of the V2V link and V2I link, the remaining amounts of traffic, the remaining time to meet the latency constraints, and the interference level and selected channels of neighbors in the previous time slot. An action refers to the selection of the sub-band and transmission power. The reward is calculated by the capacity of V2I links and the V2V latency. In the broadcast scenario, each vehicle is considered as an agent in the system. In addition to the states of the unicast system, they included the number of times that the message have been received by the vehicle, and the minimum distance to the vehicles that have broadcast the message. The action includes determining the massages for broadcasting and the sub-channel for transmission. The reward function consists of three parts: the capacity of V2I links, the capacity of V2V links, and the latency condition. Part of this work has been published in [57, 58] for unicast and broadcast, respectively. In their implementation, the time for each selection is 2.4×10^{-4}s, using GPU 1080 Ti. This speed can be minimized using methods that reduce the

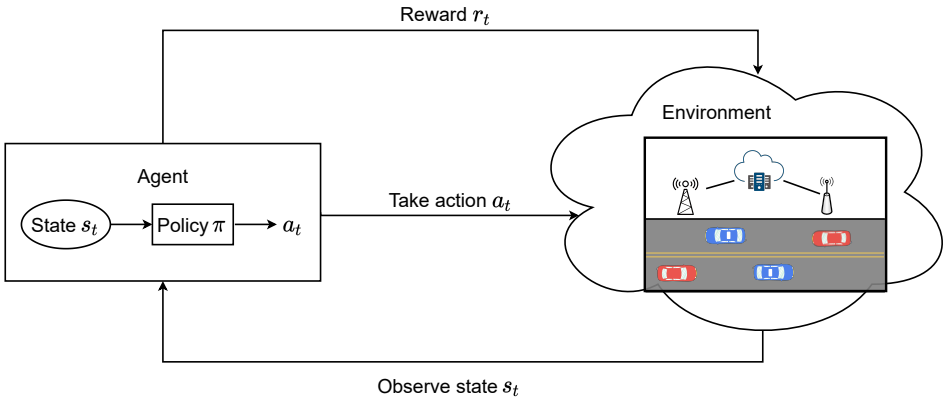

Figure 9: SARL for V2V communications [56].

computation complexity of the DNNs, such as binarizing the weights of the network [56].

In [59], authors tackled a Multi-Agent RL (MARL) problem, which is then solved using a fingerprint-based DQN method that is amenable to a distributed implementation, to improve spectrum and power allocation in order to maximize the sum capacity of V2I links and the success probability of V2V payload delivery. Each V2V link acts as an agent as illustrated in Figure 10, concurrently exploring the unknown environment. The observa-

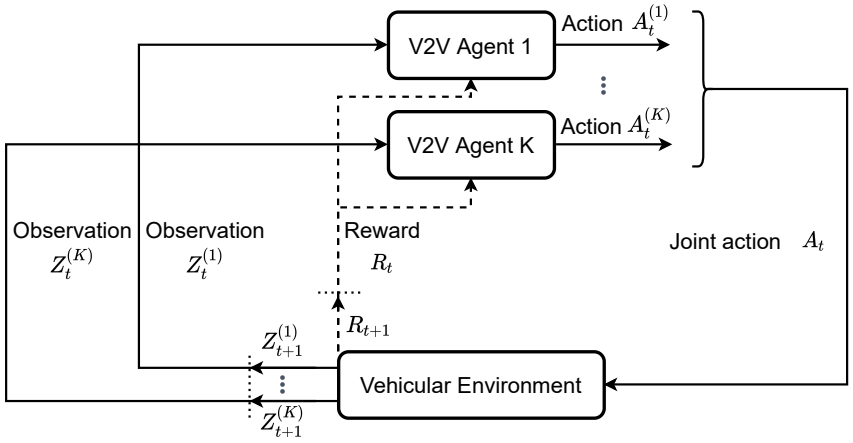

Figure 10: MARL for V2V communications [59].

tion space of an individual V2V agent contains local channel information, including its own channel interference from other V2V transmitters, the interference channel from its own transmitter to the BS and the interference channel from all V2I transmitters. Each action corresponds to a specific combination of a spectrum sub-band and a power selection. They limited the power control options to four levels, i.e. $\{23, 10, 5, -100\}$. As a result, the dimension of the action space is $4 \times M$ where M is the number of disjoint sub-bands. The reward function consists of two parts: the capacity of V2I links, instantaneous sum capacity of all V2I links, and the effective V2V transmission rate. From the simulation results, the proposed MARL method achieves significantly better performance than the Single-Agent RL-based resource sharing scheme (SARL) proposed in [56], but suffers from noticeable degradation when the payload size grows beyond 4×1060 bytes.

The network trainer in [56] trained a single DQN using global states obtained from all agents, whereas in [59], only local states were needed to train the DQN at each agent with limited parameter exchange. As a result, the

latter approach is more efficient in training. In the context of platoon-based C-V2X systems, the researchers in [60] used a MARL with a separate DQN at each agent for joint channel assignment and power allocation, similar to [59]. However, they used a different reward function design that makes it possible to improve the sum-rate of V2I links compared to the reward design offered in [59] while satisfying the high probability of successful delivery of V2V payloads.

The aforementioned works do not take into account the vehicle mobility, which not only affects the channel qualities but also provides the possibility of frequency sharing among different groups of Vehicle User Equipment-pairs (VUE-pairs). Considering vehicle mobility, *Chen et al.* in [61] formulated the age of information-aware radio resource management problem in a Manhattan grid V2V network as a single-agent MDP where the RSU makes decisions regarding frequency band allocation and packet scheduling over time in order to optimize the expected long-term performance for all VUE-pairs. However, the local network state space of all VUE-pairs is huge with the consideration of vehicle mobility, therefore to overcome the partial observability and the curse of high dimensionality, they resorted to the Long Short Term Memory technique (LSTM) and the DQN, and propose a proactive algorithm based on the Deep Recurrent Q-Network (DRQN).

In addition, the authors of [62] modeled the channel allocation problem in V2X communication networks as a decentralized MDP, in which each V2V agent decides independently its channel and power level based on the local environmental observations and global network reward. The best joint resource allocation solution is then derived using a multi-agent distributed channel resource multiplexing framework based on DRL. Furthermore, the prioritized Double DQN (DDQN) algorithm is used to provide a more accurate estimation target for the action evaluation and can effectively minimize overestimation of Q-values.

However, the aforementioned works in this subsection addressed the resource allocation in reuse mode, without jointly considering transmission mode selection for further performance improvement. In [63], the authors investigated the joint optimization problem of access mode selection and spectrum allocation in fog computing-based vehicular networks. They proposed a Q-learning-based access mode selection algorithm and a convex optimization-based spectrum allocation algorithm. Nevertheless, the large-scale continuous state space generated by multiple sensing components and realistic channel gains makes Q-Learning ineffective. Therefore, DQL algorithm was used is [64, 65] to investigate jointly communication mode selection, Resource Block (RB) assignment, and power control in V2V with the

purpose of guaranteeing the strict Ultra-Reliable and Low Latency communications (URLLC) requirements of V2V links while maximizing the sum capacity of V2I links. In each time step, the agent (V2I link) selects the communication mode, the RB assignment, and the transmit power level. Different from [64], the authors in [65] considered resource sharing among V2V pairs in different transmission modes. However, since the DQL framework may not always be suitable to deal with continuous-valued state and action spaces in Internet of Vehicles (IoV) networks, the researchers proposed in [64] a decentralized Actor-Critic (AC) RL model with a new reward function to learn the policy by interacting with the environment. The AC approach can efficiently deal with the continuous-valued state and action spaces, where the actor is used to exploit the stochastic actions and the critic is applied to estimate the state-action value function.

The multi-agent DQN approach was also of interest in [66] under the optimization objective of maximizing the V2I link capacity while ensuring the transmission delay of V2V links. Each V2V link acted as an agent collectively interacted with the environment so as to find the optimal cellular/D2D transmission mode and transmit power level.

4.1.2. Collision management

In [67], *Choe et al.* proposed a self-adaptive MAC layer algorithm employing DQN with a novel contention information-based state representation to improve the performance of the V2V safety packet broadcast for infrastructure-less congested VANET. They evaluated the algorithm with two criterions: Packet Delivery Ratio (PDR) and end-to-end delay. They proposed a fully informative state representation with the contention information. Table 5 shows the proposed state representation. During Control CHannel Interval (CCHI), vehicles broadcast safety packets appending their node ID, selected Contention Window (CW), and the corresponding success rate calculated by broadcast results of the selected CW. According to this, vehicles establish the contention information-based state. The action

Vehicle ID	**Current CW**	**Frequency**	**Success Rate**
Itself	15	5	0.32
2	255	10	0.73
17	31	3	0.6
51	15	17	0.37

Table 5: The proposed state representation [67].

definition consists of three components: Keep (K), Increase (I), and Decrease (D), In addition, they presented the transition rule of CW following two different spaces: discrete and continuous changes as illustrated in Figure 11. The authors utilized a binary reward function: vehicles will receive

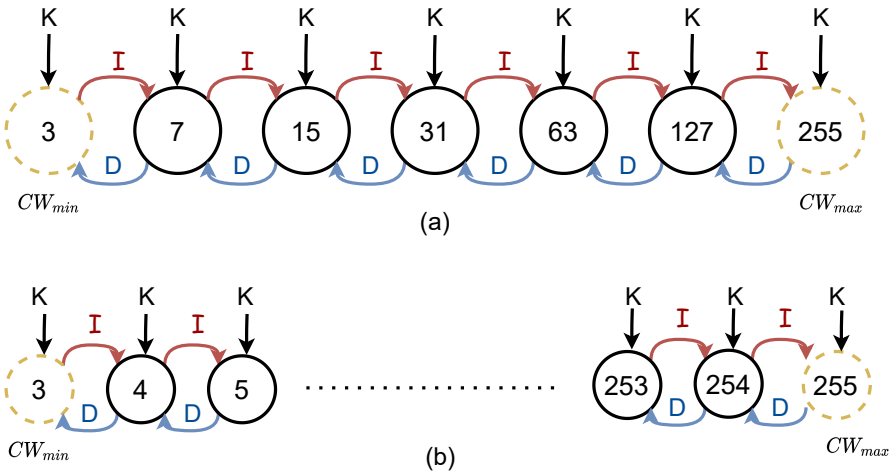

Figure 11: (a) Discrete CW space. (b) Continuous CW space [67].

the positive reward 1 or the negative reward –1 from successful and failed broadcast, respectively. They evaluated by simulations considering various levels of traffic congestion. From simulation results, it is confirmed that there is a clear trade-off between PDR and latency. Although the degradation, the proposed algorithm satisfies the latency requirement of VANET safety applications when the number of vehicles is lower than 125. However, the performance of the end-to-end delay degrades in the highest congestion level. Thus, it is required to study an adaptive MAC algorithm that can further improve PDR and latency performance for severe traffic congestion conditions.

There have been other works proposed based on RL as a V2V communications solution in congested infrastructure-less VANETs [68, 69, 70, 71]. A Q-Learning-based MAC algorithm is proposed that defines each vehicle as a single agent and improves the performance of data transmission, but it only considers V2V unicast case [68]. In order to enhance the V2V broadcast performance, a Q-Learning-based MAC protocol is proposed, and the authors demonstrate the performance improvement of V2V broadcasts from various experiments [69, 70]. However, [68, 70] use only the original Exponential Increase Exponential Decrease (EIED) [72] to deal with the channel

efficiency problem. In [71], Q-Learning is used to control the contention window through a hybrid back-off that combines EIED and Linear Increase Linear Decrease (LILD [73]) back-off as vehicles in the network become agents. However, comparing to [67], these works assume single-channel operation of the control channel and did not consider the multi-channel operation of DSRC standard.

4.1.3. Joint user association and beamforming

In [74], the RL approach is used to develop the user association algorithm for load balancing in heterogeneous vehicular networks. Considering data flow (generated from vehicular networks) characteristics of the spatial-temporal dimension, a two-step association algorithm is proposed. The initial association decision is based on a one-step RL approach. Subsequently, the BS uses historical association patterns to make association decisions, as illustrated in Figure 12. In addition, the BS, as a learning agent, keeps ac-

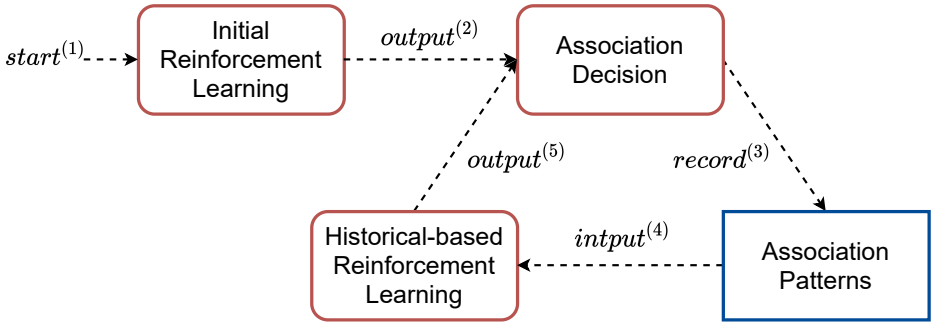

Figure 12: The architecture of proposed approach in [74].

cumulating feedback information and updates the association results in an adaptive manner. By communicating periodically with other BSs, each BS can maintain a Signal to Interference plus Noise Ratio (SINR) matrix and an association matrix as a state space. Action is defined as the BS trying to create associations with certain vehicles. The reward is defined as a reciprocal of the difference in the average service rate for all users. While each BS runs the proposed algorithm in a distributed manner, it is shown in the long run that both the real-time feedback and the regular traffic association patterns help the algorithm to manage network changes.

Rather than allowing only one vehicle to be served from only one Access Point (AP), the authors of [75] proposed the concept of virtual cell formation so that a user could be served from multiple APs simultaneously. A

single-agent Q-Learning algorithm was then developed to optimize the efficiency of joint user associations and power allocations in a highly mobile vehicular network. The RL agent needs to take two action sets: vehicle-AP associations and beamforming weights. The same authors considered energy consumption issues in vehicular edge networks in [76]. They expressed a joint virtual cell formation and power allocation problem for highly mobile Vehicle Users (VUs) in a sophisticated Software-Defined (SD) environment. They used a model-free Distributed MARL (D-MARL) solution that can effectively formulate the virtual cell and slice the resources.

In [77], the authors proposed a Deep Deterministic Policy Gradient (DDPG) based beam tracking approach which extracts information and hence achieves the URLLC requirements in typical V2X networks. It is shown that conventional EKF and PF-based [78] approaches performance in non-stationary channels are not satisfactory in terms of average packet latency due to overhead channel training and transmission failures, while a DRL-based approach can reduce the delay to about 6ms.

In [79], a vertical hand-off strategy has been devised using a fuzzy Q-Learning approach for heterogeneous vehicular networks consisting of a cellular network with global coverage complemented by V2I. Four input parameters are sent to the RSU side: the received signal strength value, the vehicle speed, the data quantity, and the number of users associated with the targeted network. The RSU then considers the information supplied as well as the traffic load, i.e. the number of users associated with the target network, and makes hand-off decisions using the fuzzy Q-Learning method.

4.2. Computing and caching

Mobile Edge Computing (MEC) dramatically increases energy efficiency and QoS for applications that require intensive computations and low latency by deploying both computing resources and caching capabilities near to end-users. With limited computation, memory and power supplies, network operators in vehicular networks, such as vehicles, become the bottleneck to support advanced applications. To address such a challenge, network entities can offload the computational tasks to nearby MEC servers, integrated with the BSs, and even neighboring vehicles. As a result, data and computation offloading can potentially reduce the processing delay, save the battery energy, and even enhance security for computation-intensive vehicular applications. In-network caching, as one of the main features of information-centric networking, will effectively eliminate duplicated content transmissions. Studies on wireless caching have shown that caching contents in wireless devices can greatly reduce access delays, energy consumption, and

overall traffic. In this subsection, we review the modeling and optimization of computation and data offloading, and caching policies in vehicular networks by leveraging the RL/DRL framework.

4.2.1. Computation and data offloading

By applying a DQL approach, [80] proposed optimal target MEC server determination and transmission mode selection schemes, which maximize the utilities of the offloading system under given delay constraints in a heterogeneous vehicular network. They focused on reliable offloading in presence of task transmission failure, and propose an adaptive redundant offloading algorithm to ensure offloading reliability while improving system utility.

In [81], the authors constructed an offloading framework for 5G-enabled vehicular networks, by jointly utilizing licensed cellular spectrum and unlicensed channels in order to minimize the offloading cost of vehicles while satisfying the latency constraint. The formulated problem is divided into two subproblems: the Vehicle-to-RSU (V2R) scheduling and the V2I allocation. For the first subproblem, they proposed a two-sided matching algorithm to schedule the unlicensed spectrum. For the second one, different from traditional centralized DRL where the macrocell or the MEC server is selected as the agent to make offloading decisions, they developed a distributed DRL method by considering multiple agents (i.e. V2I users) to schedule cellular channels. They simplify the system states to realize distributed traffic offloading, which can greatly decrease the communication overhead between vehicles and the macrocell. A DDQN is proposed to mitigate this problem.

Although DQN solved the high-dimensional action output of Q-Learning through a Deep Neural Network (DNN), the action spaces of DQN as well as DDQN, Dueling DQN are still discrete. However, for many problems, e.g. control tasks, the action space is continuous. If the action space is discretized to replace the continuous space, the dimension of the action is too high, which will lead to the curse of dimensionality. Therefore DDPG, which is a DRL method that relies on the actor-critic architecture was used in [82]. The authors have designed a task computation offloading model in a heterogeneous vehicular network taking into account the arrival of stochastic tasks, time-varying channel state, and the bandwidth allocation to achieve a trade-off between energy consumption cost while avoiding the curse of dimensionality induced by large action space.

Researchers in [83] proposed two typical multi-dimensional resource management frameworks with placing the MEC server at a Macro-cell BS (MBS) and an Edge Node (EN), respectively (Figure 13), to maximize the number

of offloaded tasks that meet QoS criteria by using a limited amount of available spectrum, computing, and storage resources A DDPG-based algorithm

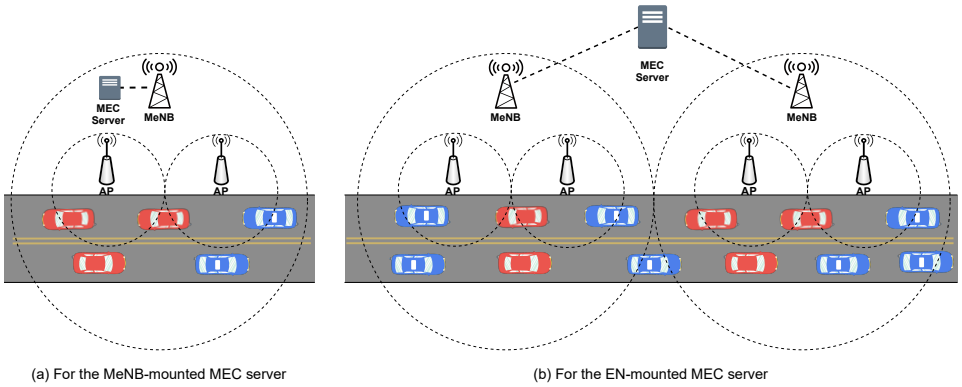

Figure 13: Dynamic spectrum management frameworks in [83].

is proposed to solve the problems. The complexity of the transformed RL problems increases with the sizes of environment state and action, a hierarchical DDPG (HDDPG)-based algorithm is developed by combining the DDPG and the hierarchical learning architecture. Each vehicle periodically sends the driving state information and task information to the MEC server. By collecting such information, the agent (i.e. MEC server) can obtain the environment state. The action space includes the spectrum slicing ratio, spectrum allocation fraction, computing resource allocation fraction, and storing resource allocation fraction.

A knowledge-driven service offloading decision framework for moving vehicles is realized in [84] using the Asynchronous Advantage Actor-Critic (A3C) algorithm. The state-space includes the task profile, the current state of the offloading destination nodes (the edge computing nodes and vehicular nodes), and the moving speed of the vehicle. The action space denoting that the task will be executed locally on the vehicle or offloaded to the accessible edge computing nodes. The reward for each decision slot is determined by task execution delay.

Collaborative computing was discussed in [85] where the authors also adopt DDPG to find the optimal offloading strategy and MEC server assignment in order to provide low-latency and reliable computing services. Instead of optimizing the computation delay for individual VUs, they adopted the optimal location-aware computing strategy to reduce the decision space and make the problem tractable. Each vehicle and RSU can only communicate with at most two RSUs with the highest Signal to Noise Ratio (SNR)

subject. Moreover, each receiver RSU has three options for forwarding the computing data to a deliver RSU: the receiver RSU processes everything or forwards a part of computing data to one of the two connected RSUs. The previous work has been extended in [86] where researchers added an algorithm to obtain the corresponding Task Partition and Scheduling Policy (TPSA) according to the server selection results calculated by the DDPG algorithm.

In [87], the authors developed an intent-based traffic control system for 5G-envisioned IoV networks, which can dynamically orchestrate edge computing and content caching to improve the profits of Mobile Network Operator(MNO). They used a DDPG based model to optimize task assignment and resource allocation in a continuous space-based.

4.2.2. Caching

The authors of [88] proposed a cooperative edge caching system to jointly optimize the placement and delivery of content in the vehicular edge computing and networks, using flexible trilateral cooperations between a macrocellular station, RSUs, and vehicles. They modeled the joint optimization problem with a double time-scale MDP, based on the fact that the content timeline changes less frequently compared to vehicle mobility and network states during the content delivery process. At the beginning of the large time-scale, the content placement/updating decision can be made based on the popularity of the content, vehicle driving routes, and resource availability. On the small time-scale, the joint vehicle scheduling and bandwidth allocation scheme is designed to minimize the content access cost while satisfying the content delivery latency constraint. In order to solve the Long-Term Mixed-Integer Linear Programming (LT-MILP) problem, a nature-inspired method based on the DDPG framework has been proposed to obtain a suboptimal solution with a low computation complexity. As compared with the non-cooperative and random edge caching schemes, the proposed cooperative caching system can reduce the system cost and content delivery while enhancing the content hit ratio.

In [89], researchers proposed secure and intelligent content caching for vehicles by integrating DRL and authorized blockchain in vehicular edge computing networks. They first proposed a distributed and secure content caching framework with a blockchain, in which vehicles perform content caching and BSs maintain an authorized blockchain to ensure content caching. Next, they exploited the advanced DRL approach to design a new DRL-inspired content caching scheme by taking vehicular mobility into account. Finally, they proposed a new block verifier selection method to enable

a fast and efficient blockchain consensus mechanism.

In [90], the authors investigated the edge caching strategy taking into account content delivery and cache replacement by leveraging distributed MARL. They first presented a hierarchical edge caching architecture for IoVs, where cooperative caching between multi-RSUs and MBSs is used to reduce content delivery costs and traffic load in the system. In addition, they formulated the corresponding optimization problem to minimize the long-term overhead of content delivery, and they extended the MDP to the multi-agent system case.

The authors in [91] formulated the resource allocation problem as a joint optimization of caching, networking, and computing, e.g. compressing and encoding operations of the video contents. The system states include the channel state information from each BS, the computational capability, and the cache size of each MEC/content server. The network operator feeds the FNN based DQN with the system state and gets the optimal policy that determines the resource allocation for each vehicle. These same authors improved Q-Learning in [92] by using CNNs in DQN to exploit spatial correlations in learning. This enables the extraction of high-level features from raw input data. They have also introduced a dueling DQN in [93] to improve the stability and performance of the ordinary DQN method. Dueling DQN is designed to avoid overestimation of Q-value in ordinary DQN. This will make the training process faster and more reliable. Dueling DQN is also integrated into the design with the intuition that it is not always necessary to estimate the reward by taking some action. The state-action Q-value is decomposed into a value function that represents the reward in the current state and the advantage function that measures the relative importance of a given action compared to other actions. Simulation results show that the proposed dueling DQN scheme outperforms the existing static scheme in terms of total utility. The previously mentioned DQL framework for VANETs, e.g. [91, 93], has also been extended to smart city applications in [94], which involve dynamic orchestration of networking, caching, and computation to meet different service requirements.

A multi-time scale DQN framework is proposed in [95] to minimize the system cost by the joint design of communication, caching, and computing in VANET while taking into account the huge action space and high complexity with the vehicle's mobility and service delay deadline Simulation results show that the proposed framework can reduce the cost up to 30% compared with the random resource allocation scheme.

4.3. Energy

Management applications also need to consider energy-efficient management of resources. Because some RSUs in vehicular networks are powered by a battery, the RL/DRL can be used to extend the battery life. Moreover, taking into account the limited energy of vehicles, the energy management of hybrid electric vehicles is an important issue that involves a trade-off between gasoline and electricity. DQN has been used in vehicular energy management for both electric and hybrid vehicles. In this subsection, we look at how RL/DRL is used in vehicular networks to:

- optimize RSU's battery usage while scheduling uplink and downlink messages. We also present some pioneering works [96, 97] which only considered scheduling without optimizing RSU power consumption.
- optimize the energy consumption of vehicles.

4.3.1. Roadside units scheduling

In [96], *Zhou et al.* proposed a RL-based resources allocation algorithm to adaptively change Time Division Duplex (TDD) configuration [98] for one channel in order to satisfy the high traffic demand in 5G vehicular networks with limited resources. TDD can change the Uplink (UL) and Downlink (DL) ratio in the same frequency band. They implemented a Q-Learning algorithm where the agent and environment are BS and vehicular network respectively. It means that BS can choose UL/DL ratio considering predicted future network situation. The authors defined the state as the percentage of UL/DL data rate and channel capacity. The action of the agent is the UL/DL ratio in each interval. It is selected from set W, which contains patterns of TDD configuration [99]. The reward is set to make the percentage of UL/DL data rate and channel capacity close to 100%. On the one hand, if the percentage is much lower than 100% then it means that there is tremendous waste in channel capacity and the TDD configuration needs to be updated. On the other hand, if it is much higher than 100% then it stands for high packet loss rate. The conducted simulation results show that the proposal outperforms the conventional TDD method in throughput and packet loss rate. A Conventional RL Q-Learning algorithm is used since the state and action spaces are small. However, in 5G Heterogeneous Networks (HetNet), conventional RL hardly handles the complex environment of high mobility and heterogeneous structure of these networks. For that, the same authors proposed a novel DRL-based intelligent TDD configuration algorithm in [100] to adaptively change TDD UL/DL ratio for 5G HetNet.

In [101], the authors presented a DRL model, namely DQN which learns an energy-efficient scheduling strategy from high-dimensional inputs corresponding to the characteristics and requirements of vehicles residing within an RSU communication range. On the one hand, VANET communications provide services related to protection. On the other hand, mobile users can access a variety of non-secure Internet services through V2I communications. As a result, a multi-objective RSU scheduling issue arises, with the goal of meeting the diverse QoS requirements of various non-safety applications while maintaining a secure driving environment in such a way that efficiently utilizes available energy and extends the lifespan of the underlying vehicular network. The authors implemented a DQL algorithm where the agent and environment are BS and vehicular network respectively. The considered network environment is illustrated in Figure 14. The agent's input

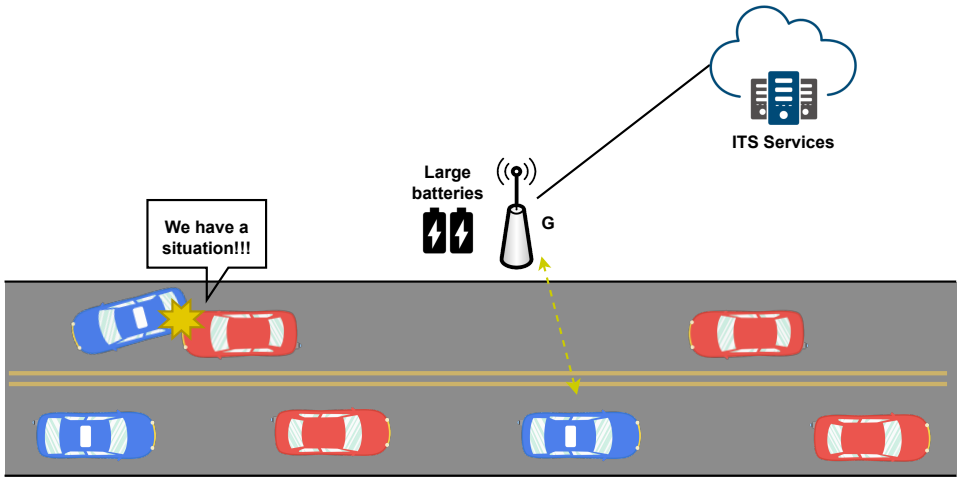

Figure 14: Energy-limited VANET [101].

from the environment (state space) is made up of: the time elapsed since the last RSU battery recharge, the remaining power in the RSU's battery, the number of vehicles residing within the communication range of point G, the remaining discrete sojourn times of each vehicle, the remaining request sizes for each vehicle, the waiting times of the safety messages in the vehicles' buffers, the separation distances between G and each of the inrange vehicles. The RSU either chooses to receive a safety message, whose existence has been announced, or to transmit data to a vehicle. Whenever the RSU chooses to transmit data to a particular vehicle, the reward received is the number of transmitted bits. In this case, the cost paid is composed of

two components: (i) the power consumed by the RSU to serve the vehicle, and (ii) the waiting time of a safety message whenever it exists. When the RSU chooses to listen to an announced safety message, the induced cost pertains to the amount of power required for the RSU to receive the safety message's data. Furthermore, whenever a vehicle departs from the RSU's coverage range with an incomplete download request, the agent is penalized by a value corresponding to the remaining number of bits that need to be downloaded in order to fulfill that vehicle's request.

The same authors extended this work in [102] in the context of IoV to multiple RSUs. They presented a DRL model that learns an energy-efficient and QoS-oriented scheduling policy, which dictates the operation of multiple energy-limited RSUs. A central ITS agent governs the operation of multiple connected RSUs deployed on a long road segment as illustrated in Figure 15. The state space is defined as a set of states of each RSU,

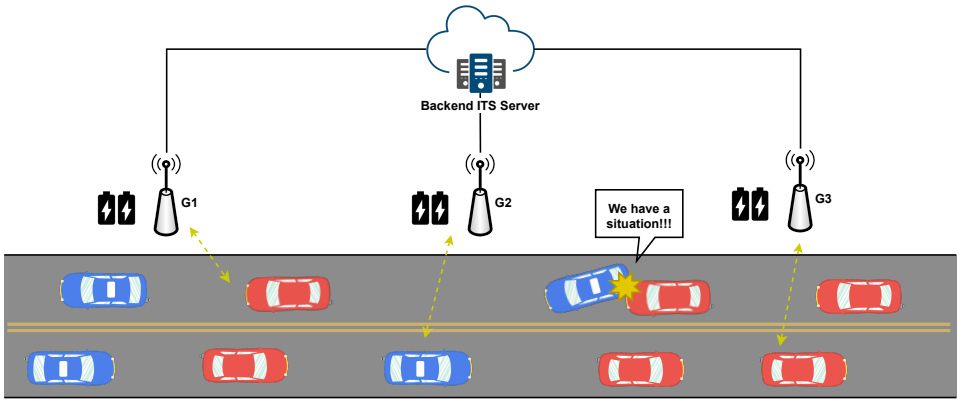

Figure 15: Energy-limited multi-RSU vehicular network [101].

and the state of each RSU is defined as in [101]. The central ITS agent defines for each RSU whether the latter will receive a security message or transmit packets to a vehicle. The immediate cost (negative reward) is the sum of the RSU rewards presented in [101]. The proposed DQN algorithm outperformed several existing scheduling benchmarks in terms of completed request percentage (average improvement between 10.9% and 21.2%), mean request delay (average improvement between 10.2% and 21.1%) and total network lifetime (improvement between 13% and 71%) under variable vehicular densities and vehicle request sizes.

The same authors have developed in [97] an MDP framework with discretized states in order to establish an optimal RSU scheduling policy whose

objective was to satisfy the maximum number of vehicle downloads requests. Therein, MDP resolution was achieved using RL techniques. However, Q-Learning with discrete states and actions has poor scalability. In [101], the state space is continuous. Therefore, conventional RL techniques are no longer feasible in this case.

4.3.2. Vehicle

Hybrid Electric Vehicles (HEVs) have been on the market in recent years to help with electricity shortages and global warming. To minimize energy consumption and pollution, they use both Internal Combustion Engines (ICEs) and Electric Motors (EMs) for propulsion. *Qi et al.* implemented a DQL based Plug-in Hybrid Electric Vehicles (P-HEV) Energy Management System (EMS) to autonomously learn the optimal fuel use from its own historical driving record in [103] . Power demand at wheel and the battery pack's state of charge are selected to form a two-dimensional state space. The action is to select Internal Combustion Engine (ICE) power level. The reciprocal of the resultant ICE energy consumption at each time step is defined as the immediate reward. However, according to [104], there are several problems in this study. The learning process is still offline. As a result, this method can be used in buses with only fixed route. The relationship between fuel economy and engine power is complex and the paper lacks the ability to justify this phenomenon. In [104], researchers developed a DRL-based control framework and an online learning architecture for an EMS for HEV, which is adapted to different driving conditions. There are many other works for energy management in HEV. A detailed survey on the application of RL in HEV is presented in [105].

A parametric study of several key factors during the development of RL-based EMS for HEV/P-HEV is presented in [106] including (1) state types and number of states, (2) states and action discretization, (3) exploration and exploitation, and (4) learning experience selection. The main results show that selecting learning experiences can effectively minimize vehicle fuel consumption. The analysis of states and action discretization reveals that as action discretization increases, vehicle fuel consumption decreases while rising state discretization decreases fuel consumption. Furthermore, the growing number of states improves fuel economy. However, DDPG was used in [107] for HEV energy control which does not necessitate the discretization of both state and action variables.

Hybrid Electric Tracked Vehicles (HETV) were discussed in [108, 109] where a DRL algorithm was used in [108] to derive energy management strategy, and an online RL algorithm based on Q-Learning was used in

[109].

There have been some works for Extended-Range Electric Vehicle (EREV). The authors in [110] have proposed an RL method and a rule-based strategy is used to improve the fuel economy of an in-use EREV used in a last-mile package delivery application. The authors used a double Q-Learning with experience replay. To remove the process of action space discretization, the same authors presented in [111] an AC-based RL framework that can dynamically update the RB vehicle parameter during a trip with uncertain remaining distance, velocity trajectory, and energy intensity. The RL framework uses real-time information collected from the vehicle and a learned strategy from historical data.

Summary

This section reviews resource management techniques based on RL/DRL in vehicular networks. We have presented these works by considering each resource category: networking, computing and caching, energy. The reviewed approaches are summarized with references in the Table 6. In the next section, we will review research works that applied RL/DRL techniques to vehicular infrastructure management.

5. Vehicular infrastructure management

Infrastructure is the key vector of interaction between applications and the vehicular environment among the various components of the vehicular networks. As a result, the objectives of vehicular applications are achieved through vehicular infrastructure management, which is mainly categorized in (*i*) traffic management and (*ii*) vehicle management.

5.1. Traffic management

The inefficient Traffic Light Control (TLC) causes numerous problems, including long delays of travelers, massive energy consumption, and deteriorating air quality. It can also lead to vehicular accidents in some situations. Conventional TLC either deploys fixed programs without consideration for real-time traffic or considers traffic only to a minimal extent. Fortunately, RL/DRL technique is a promising method for monitoring and processing the real-time road condition. Also, traffic congestion has become a common transportation problem on freeways around the world in recent decades. Congestion usually begins at the bottleneck of the road and extends upstream and downstream. As a result, Variable Speed Limit (VSL) control systems are being studied extensively as solutions for improving safety and

Resource	Issues	References	Role
Networking	Dynamic spectrum access	[56, 57, 58, 59, 60]	Channel assignment and power allocation for V2V links
		[61, 62]	Channel assignment and power allocation for V2V links taking into account vehicle mobility
		[63, 64, 65, 66]	Transmission mode selection, channel assignment and power allocation for V2V links
	Collision management	[68, 69, 70]	Contention Window control, EIED, single-channel operation
		[71]	Contention window control, LIED and EIED, single-channel operation
		[67]	Contention window control, LIED and EIED, multi-channel operation
	Joint user association and beamforming	[74]	User association for load balancing with one access point
		[75, 76]	User association for load balancing with multiple access points
		[77]	Beam tracking
		[79]	Hand-off management
Computing and caching	Computation and data offloading	[81, 82, 83, 84, 87]	Find the optimal offloading strategy
		[80, 85, 86]	Find the optimal offloading strategy and MEC server assignment
	Caching	[88, 89, 90]	Optimize content caching
		[91, 92, 93, 94, 95]	Optimize caching, networking, and computing
Energy	Roadside units scheduling	[96, 100]	Adaptively change Time Division Duplex (TDD) configuration
		[100, 102, 97, 101]	Learns the optimal energy-efficient scheduling strategy
	Vehicle	[103, 104, 106, 107]	Learn the optimal energy management strategy in HEVs
		[105]	A detailed survey on the application of RL in HEV
		[108, 109]	Learn the optimal energy management strategy in HETV
		[110, 111]	Learn the optimal energy management strategy in EREV

Table 6: Researches on RL/DRL-based resource management for vehicular networks.

throughput on urban freeways. In this subsection, we review some works that use RL/DRL for TLC and VSL control.

5.1.1. Traffic light control

Liang et al. have proposed in [112] a DRL model to solve the TLC problem for one intersection in vehicular networks. The states are two-dimension values that provide details about the vehicles' location and speed information. The actions defined by how to update the duration of every phase in the next cycle and the rewards are the cumulative waiting time difference between two cycles. To handle the complex traffic scenario in this problem, they proposed a double dueling deep Q-network (3DQN) with prioritized experience replay. The model can reduce the average waiting timing by over 20% from the start of the training.

In [113], the authors explored the capability of DRL for handling TLC systems using partial vehicle detection, see Figure 16. The designed DRL-based algorithm specifically DQL performs well under low penetration ratio and detection rates. The state contains the following information: detected

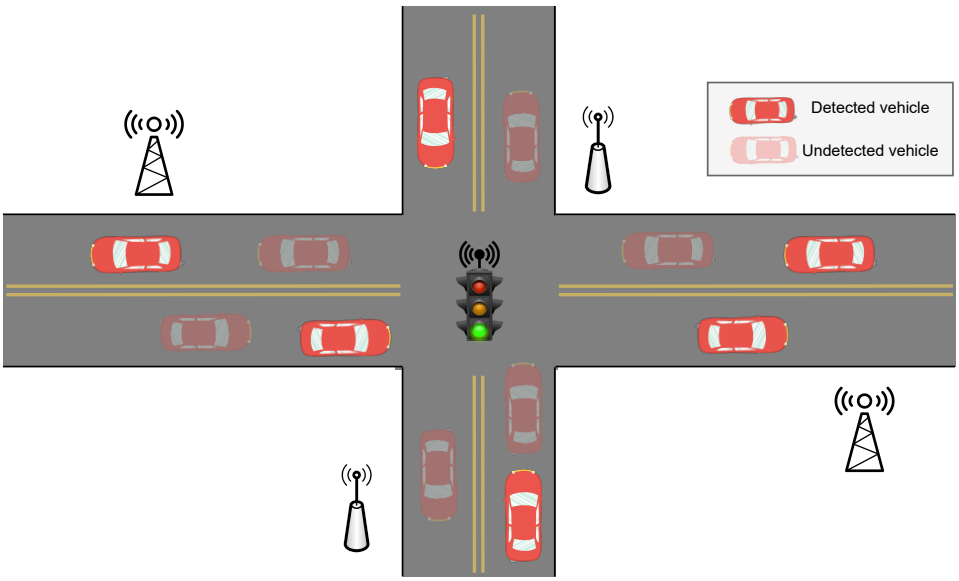

Figure 16: Illustration of partially detected ITS [113].

car count, distance to the nearest detected car, current phase time, amber phase, current time, and current phase. The action of the agent is either to keep the current traffic light phase or to switch to the next traffic light phase. The reward is the average traffic delay of commuters in the network. The results of this study show that RL is a promising new approach to optimizing traffic control problems under partial detection scenarios, such as traffic

control systems using DSRC technology. However, since multiple intersections can have an effect on each other in the real world, just considering one intersection is insufficient.

Complementary to the aforementioned study [113] where TLC with varying portions of V2I enabled vehicles ranging from 0 to 100% at a *single intersection*, the authors of [114] examined the performance impact of traffic state information collected via V2I communication for road networks consisting of *multiple intersections* considering the extreme cases of 0% of V2I enabled vehicles in [113] (equivalent to their agnostic agent) and 100% of V2I enabled vehicles in [113] (equivalent to their holistic agent). The study's key contribution is a detailed comparison of a representative state-of-the-art agnostic DRL agent that is unaware of the current traffic state versus a representative state-of-the-art holistic DRL agent that is aware of the current traffic state. States representation is shown in Table 7. The action of both agents is to decide the phase to show, and the display duration. They also compared a reward function that considers only the average vehicle velocity with a composite reward function that takes into account a weighted combination of the average vehicle velocity, vehicle flow rate, CO_2 emissions, and driver stress level. They found that the holistic system substantially

Feature	**Agnostic Agent**	**Holistic Agent**
Current phase of all traffic lights (phase ID & period ID)	✔	✔
Time passed since the last change	✔	✔
Traces of all phases	✔	✔
Positions of vehicles closest to an intersection		✔
Velocities of vehicles closest to an intersection		✔
Number of vehicles on each lane		✔
Average velocity of vehicles on each lane		✔

Table 7: State spaces of agnostic and holistic agents [114].

increases average vehicle velocities and flow rates while reducing CO_2 emissions, average wait and trip times, as well as a driver stress metric.

MARL has been utilized in [115] where researchers have proposed a Multi-Agent Recurrent DDPG algorithm to reduce traffic congestion at multiple intersections taking into consideration pedestrian and bus, which make the whole system more humanized. They utilized various road information to change the phase of multiple traffic lights in real-time. The traffic light controller at each intersection is not isolated, and it can observe the global state during the training process. Each traffic light controller can estimate the traffic light control policies of other intersections when making

decisions. Scalability analysis shows that this method is more suitable for medium-scale traffic networks.

Other recent studies have also used MARL for TLC [116, 117, 118, 119]. General related computational frameworks for RL control applications have been explored for multi-objective decision modeling in [120] and for a hybrid fuzzy and RL control in [121].

5.1.2. Variable speed limit control

Researchers in [122] proposed a single QL algorithm for VSL control system. To alleviate traffic congestion, The QL agent should keep the bottleneck density below its critical value. The state-space includes the density at the immediate downstream of the merge area, the density at the upstream mainline section, and the density on the ramp. The action is to set the speed limit. The reward is the system travel time where the aim of the agent is to minimize this reward. The results demonstrated that the QL approach outperforms feedback control methods, both in stabilized and fluctuant demand scenarios.

In [123], *Wang et al.* proposed a distributed QL algorithm to tackle the cooperative control problem in continuous traffic state space. The agents work cooperatively using the proposed distributed RL approach to maximize the freeway traffic mobility and safety benefits. A per-lane VSL control based on Lagrangian control using DRL is proposed in [124]. Since the traffic flow used in the research includes Autonomous Vehicles (AVs), the VSL controller can directly change the speed of AVs within a particular traffic lane and thus monitor remaining traffic flow rather than using traditional Vehicle Management Systems (VMSs). In [125] a MARL framework for solving the bottleneck congestion is proposed, using both VSL and Ramp Metering (RM) simultaneously. The proposed MARL outperforms base cases (independent and feedback-based VSL and RM) with respect to measured network TTS.

Researchers in [126] proposed a method for establishing MARL-based VSL using the W-Learning algorithm (WL-VSL), in which two agents monitor two segments leading up to the congested area. Each agent's reward role is determined by the agent's local output as well as the downstream bottleneck. WL-VSL is evaluated in a microscopic simulation in two traffic scenarios with dynamic and static traffic demand. They demonstrated that WL-VSL outperforms base cases (no control, single agent, and two independent agents) with the improvement of traffic parameters up to 18%. The same authors presented a comprehensive survey on the state-of-the-art of RL-VSL in [127].

5.2. Vehicle management

The management of vehicles is one of the most critical tasks for vehicular networks, especially for autonomous driving. It consists of two primary components: motion control, such as steering angle and vehicle speed control, and vehicle path/trajectory planning. The scenario of vehicle management includes diverse types of events like parking, lane changing, merging, platooning, and so on.

Our goal is to focus this survey mainly on the application of RL and DRL techniques for vehicular telecommunications. For motion planning and control in autonomous vehicle networks, There is a lot of works that apply RL and DRL in this research area, and they are reviewed in the following surveys [128, 129, 130].

Summary

This section reviews infrastructure management techniques based on RL/DRL in vehicular networks. We categorized these works into two categories: traffic management and vehicle management. The reviewed approaches are summarized with references in the Table 8. In the next section, we will outline some open issues and future trends that are likely to contribute to the continuous shaping of future 6G vehicular networks.

Resource	**Issue**	**References**	**Role**
Traffic management	Traffic light control	[112, 118, 121]	Solve the TLC problem for one intersection using full vehicle detection
		[113]	Solve the TLC problem for one intersection using partial vehicle detection
		[114, 115, 116, 117, 119]	Solve the TLC problem for multiple intersection
	Variable speed limit control	[122, 123, 124, 125, 126]	Set the speed limit to keep the bottleneck density below its critical value
		[127]	A comprehensive survey on the state-of-the-art of RL-VSL
Vehicle management	Motion control and trajectory planning	[128, 129, 130]	Surveys for vehicle management using RL/DRL

Table 8: Researches on RL/DRL-based infrastructure management for vehicular networks.

6. Challenges, open issues and future trends towards 6G

Different approaches reviewed in this survey evidently show that RL and DRL can effectively address various emerging issues for resource and

infrastructure management. There are existing challenges, open issues, and future trends which are discussed as follows.

6.1. Challenges

The complex and dynamic challenges of V2X paradigm have been addressed using different techniques of RL and DRL. In addition, the computation resources, storage resources, and the advancements in RL and DRL algorithms have established a firm path of RL based research for V2X enabled by the 5G paradigm. Yet, there are issues that require attention in this context. Some of the important challenges are presented below:

- *Training and Performance Evaluation of DRL Framework.* The majority of previous research works relies on simulated data, which raises doubts about the DRL framework's applicability in real-world systems. A specific stochastic model is frequently used to generate the simulated data set, which is a simplification of the real system and may overlook hidden patterns. As a result, a more efficient method for generating simulation data is necessary to ensure that the DRL framework's training and performance evaluation are more compatible with real-world systems [9].

- *Complexity of calculation and synchronization.* The complexity of the signal timing design, which grows exponentially with the number of traffic flow state/control actions, is one of the primary challenges that RL is confronting for traffic signal timing management. The practical implementation of RL-based congestion control approaches in VANETs necessitates the use of RSUs with GPUs. For ML algorithms, these RSUs would have to do more complex computations and operations. For RL-driven V2X applications, the deployment of Edge computing facilities is also required [131].

- *Multi-Agent RL/DRL in Dynamic HetNets.* Vehicular networks consist of hierarchically nested IoT devices/networks with fast changing service requirements and networking conditions. RL/DRL agents for individual entities must be light-weight and adaptable to changing network conditions in this situation. This implies a a reduction in the state and action spaces in learning, which however may compromise the performance of the convergent policy. Multiple agent interactions further complicate the network environment by causing a significant increase in the state space, which slows down the learning algorithms [9].

- *Fairness vs optimization in TLC.* The decision of the agent's fairness policy is a major issue in the realm of RL. A fair TLC system, for example, would ensure that all vehicles are given the same priority while crossing the intersection. However, this fairness will cause a conflict with the optimization of specific traffic metrics. These optimization metrics could be the minimum delay or maximum throughput. A future challenge will be to achieve a balance between fairness and optimization, which can be accomplished by using the appropriate reward function and other AI techniques [131].

6.2. Open issues and future trends

The RL/DRL framework has an effect on a wide variety of vehicular applications. However, we believe that existing studies do not capture the full potential of RL/DRL driven vehicular networks due to both the limitations of existing RL/DRL approaches and the changing needs of vehicular networks. Next, we discuss some open issues and future trends of 6G vehicular networks that deserve further investigation.

6.2.1. Autonomous and semi-autonomous vehicles

Autonomous and semi-autonomous driving demands unprecedented levels of reliability and low latency. However, the 5G at its current development state cannot meet these requirements [132]. Thus, we need 6G wireless networks to pave the way for connected vehicles through advances in hardware, software, and the new connectivity solutions [30].

Autonomous Vehicles (AVs) are continually being developed and are the focus of many research projects. The objective is to prevent accidents caused by human driving errors and to carry out emissions reduction [133]. RL and DRL have been recently used to learn a policy of the different tasks of autonomous driving involving lane keeping, lane change, ramp merging, overtaking, motion planning, intersections [129]. For *lane keeping*, the authors in [134] proposed a DRL system for discrete actions (DQN) and continuous actions (Deep Deterministic Actor Critic (DDAC)) to follow the lane and maximize average velocity. For *lane change*, authors in [135] used Q-Learning that make vehicles learn to perform no operation, lane change to left/right, accelerate/decelerate. For *ramp merging*, researchers in [136] applied DQL to find an optimal driving policy by maximizing the long-term reward in an interactive environment. For *overtaking*, *Ngai et al.* in [137] proposed multi-goal RL policy that is learnt by Q-Learning to determine individual action decisions based on whether the other vehicle interacts with the agent for that particular goals. For *motion planning*, [138] proposed

an improved algorithm using DQNs over image-based input obstacle map. Authors in [139] used DQN to negotiate *intersections*. The development of MARL approaches for the autonomous driving problem is also an important future challenge that has not received a lot of attention to date. It could be very beneficial in high-level decision-making and coordination between autonomous vehicles. There have been some prior works that used MARL for AVs [140, 141, 142].

The motion planning and control aspects for autonomous driving are rather disconnected from the telecom part. However, it is well known that joint optimization problems arise in telecommunications and control of autonomous vehicles as explained in [143]. In this case, RL/DRL should be able to simultaneously deal with both problems, for example, if autonomous vehicles are poorly piloted and moving too far from each other then radio links between them can be lost. We believe that this is an intriguing path for future research.

Although research is towards fully autonomous vehicles, semi-autonomous mode (or tele-operated driving) is desired when autonomous mode fails or a complicated scenario requires human intervention. It can aid as a measure for a smooth transition from non-autonomous to autonomous driving by offering a fallback in certain situations [144]. Tele-operated driving allows cars to be controlled remotely providing direct steering, acceleration and braking commands to the vehicle. It will require ultra-low latency that communicates signals and instructions between the driver and the vehicle, especially in the face of danger where an immediate response is needed. A high level of security, privacy, and network integrity is also desired [132]. RL and DRL are poorly used in the context of semi-autonomous driving in vehicular networks to allocate resources for this type of application to meet the requirement of latency and reliability. This is an interesting direction for future studies in the 6G context.

6.2.2. Brain-vehicle interfacing

In a Brain-Controlled Vehicle (BCV), the vehicle is controlled by the human mind rather than any physical contact between the human and the vehicle through the use of a brain-computer interface (BCI), which can translate brain activity signals into motion commands. For people with disabilities, BCVs hold great promise for increased independence and improved quality of life by offering an alternative interface by which they can control vehicles [145]. Current wireless connectivity and computing systems are incapable of realizing BCV because services involving brain-machine communications would necessitate ultra-high reliability, ultra-low latency, ultra-high data

rate communication, and ultra-high-speed computation. However, 6G-V2X must be capable of understanding and adapting human driver behavior [14].

There are some successful works that demonstrated the feasibility of BCV [146, 147]. However, the currently developed BCV is not a scalable solution since it would necessitate a wireless connection to facilitate brain-machine interactions with high coverage, availability, speed, and low latency to provide end-users with reliability and protection. In order to meet these requirements, RL and DRL frameworks could be used to allocate resources for these applications in future studies. Also, researchers did not pay much attention to using AI in the context of BCV, for example, [148] focused on processing of the signals received from the Electroencephalography (EEG) headset into directions (move forward, backward, left, right, or stop) using ANN. In addition, most of the existing works on BCV have been verified through simulation only. Thus, extensive real-world experiments are required to demonstrate the effectiveness of BCVs [14].

6.2.3. Green vehicular networks

6G vehicular networks will be using emerging technology such as edge and artificial intelligence in its network nodes that require high energy [10], and they will have ultra-high throughput, ultra-wide bandwidth, and ultra-large-scale ubiquitous wireless nodes, which will pose a huge challenge to energy consumption [149]. These networks may face stringent requirements to achieve green vehicular communications which enable energy conservation, decreased emissions, and lower environmental pollution [150]. This research area is still in its infancy.

Dense Heterogeneous Vehicular Networks (HetVNets) are recognized as a core technology for green vehicular networks. Higher transmission rates and lower power consumption are seen as the key concerns to achieve green communications in vehicular networks. It is therefore necessary to upgrade the current communication mode to enrich the user experience and to reduce energy consumption associated with base station access. There exist some works that apply RL/DRL in this context [102, 101] but it is still insufficient.

Also, the Green Vehicle Routing Problem (GVRP) has emerged as a key agenda item in green logistics, attracting scientific interest from researchers. Referencing to [151], most of the current GVRP research focuses on statically determined systems. Therefore, future research can target GVRP under a dynamic environment to make the research more convincing. For new energy vehicles, the construction of charging stations is a prerequisite for solving GVRP. Therefore, future studies can consider combining the vehicle routing problem of new energy vehicles with the location of charging stations.

6.2.4. Integrating unmanned aerial and surface vehicles

6G should be a ubiquitous and integrated network with broader and deeper coverage that can serve in various environments such as airspace, land, and sea, realizing a global ubiquitous mobile broadband communication system [152], as shown in Figure 17.

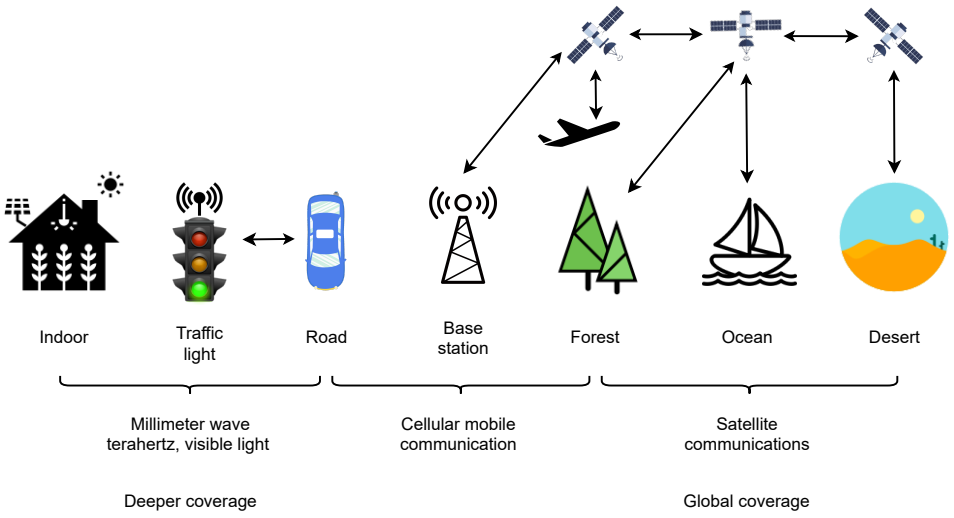

Figure 17: 6G with global and deeper coverage [152].

Unmanned Aerial Vehicles (UAVs) are aircraft that do not have a human pilot aboard. Recently, the enthusiasm for utilizing UAVs in a proliferation of fields has exploded, thanks to advanced technologies and their reduced cost [11]. For instance, Amazon and DHL are trying to use UAVs to deliver commodities to customers over the air. There are various other roles that a UAV can play, e.g. aerial traffic signals or aerial cameras. Unmanned Surface Vehicles (USVs) are boats that work without a crew on the water's surface [153]. Recently, USVs have attracted extensive research attention, due to their advantages in many applications, such as environmental monitoring, resource exploration, enhancing the efficiency and safety of water-borne transportation, and many more [154].

UAVs' high mobility leverages the importance of prediction tasks, which are useful in path planning and collision avoidance, for instance, [155, 156] highlighted the use of RL/DRL in collision avoidance, whereas [157, 158] studied path planning. Also, path planning and collision avoidance are fundamental aspects of USV guidance and navigation systems, for instance,

[159, 160] used DRL techniques for path planning whereas [161, 162] used DRL for USV collision avoidance. In addition, UAVs' high mobility will also increase the necessity of an ultra-reliable and low-latency network, leveraging the importance of management applications for traffic control and resource sharing in such networks [163, 164].

To conclude, UAVs and USVs are likely to feature the 6G vehicular networks. However, UAVs and USVs will be deployed in the air and the water respectively, in environments with different obstacles, mobility, and infrastructure from the usual road transportation. Therefore, they will require further attention to their deployment.

6.2.5. Security enhancement with blockchain

Vehicles authentication and privacy protection in 6G networks have seriously restricted the development of IoV, especially when being dependent on a centralized trusted authority to distribute identity information [165]. This problem can be solved with blockchain due to its many potential uses in IoV when including the distributed ledger to secure the network and enable autonomous communications [166], radio access network decentralization [167] and mobile service authorization [168]. The integration of both AI and blockchain is highly trending because it provides security, intelligent architecture, and flexible resource sharing [169] for IoV as well as Industrial IoT networks [170]. In fact, the integration of DRL, as well as permissioned blockchain, enabled the ability to cash a huge amount of data and multimedia content in proximity to vehicles. On one hand, vehicles perform content caching whereas base stations maintain the permissioned blockchain. Then, the DRL-based approach is exploited to design an optimal content caching scheme that takes into account the mobility parameter in the IoV network [89]. DRL also optimized the performance of blockchain-enabled IoV in terms of maximizing transactional throughput while guaranteeing security and complete decentralization of the underlying blockchain system [171].

Various research works also tried to tackle security challenges using applications based on blockchain technology [172]. However, many security and privacy issues still need to be addressed to improve vehicle authentication, privacy, and trust management in IoV networks. This can be achieved using a decentralized authentication scheme based on consortium blockchain [173] with multiple trusted authorities instead of using traditional client-server strategies for key sharing and certifications storage. The challenge here remains in achieving complete synchronization between multiple control units on the road while protecting the privacy of the vehicles using pseudo-random IDs. This security architecture provides additional advantages in terms of

reducing communication overhead and speedily updating the status of revocated vehicles in the shared blockchain ledger [174]. Moreover, various blockchain-assisted architectures were proposed to integrate and adapt the blockchain to IoV applications on edge computing [175], cloud computing [176] or software-defined VANETs [177] where controllers play the role of blockchain miners. However, reaching a consensus between controllers in such a use case is not straightforward and depends on various parameters such as the number of consensus nodes, the computational capability of the blockchain, and the trust features of vehicles and blockchain nodes [178]. In this context, the latter is provided with blockchain by enabling the ability for vehicles to validate the received messages from neighboring vehicles [179]. However, this solution is not sufficiently efficient in large-scale IoV networks, specially that 6G networks are expected to support a massive number of applications with various QoS requirements. Therefore, driven by the heterogeneous applications and the massive demands of hyper-connected vehicles, it is clear that more efforts should be placed to propose a combination of IoV-specific consensus and DRL-based optimizations for realistic IoV application scenarios.

7. Conclusion

This paper has presented a comprehensive survey of the applications of reinforcement learning and deep reinforcement learning to vehicular networks. First, we have presented an overview of vehicular networks. Then we have introduced reinforcement learning, deep learning, and deep reinforcement learning. Afterward, we have provided detailed reviews of reinforcement learning and especially deep reinforcement learning to solve different issues in vehicular networks. We have categorized these schemes into two categories, i.e. vehicular resource management and vehicular infrastructure management with an emphasis on vehicular telecommunications issues. Finally, we have outlined some open issues and future trends that are likely to contribute to the continuous shaping of future 6G vehicular networks. We expect this survey to provide a rapid and comprehensive understanding of the current state-of-the-art in vehicular network communications involving deep reinforcement learning techniques while attracting and motivating more researchers into this interesting research area.

References

[1] H. Ye, L. Liang, G. Y. Li, J. Kim, L. Lu, M. Wu, Machine learning for vehicular networks: Recent advances and application examples, IEEE Vehicular Technology Magazine 13 (2) (2018) 94–101.

[2] N. Kumar, N. Chilamkurti, J. H. Park, ALCA: agent learning–based clustering algorithm in vehicular ad hoc networks, Personal and ubiquitous computing 17 (8) (2013) 1683–1692.

[3] G. Singh, N. Kumar, A. K. Verma, Antalg: An innovative aco based routing algorithm for manets, Journal of Network and Computer Applications 45 (2014) 151–167.

[4] N. Kumar, S. Misra, J. J. Rodrigues, M. S. Obaidat, Coalition games for spatio-temporal big data in internet of vehicles environment: A comparative analysis, IEEE Internet of Things Journal 2 (4) (2015) 310–320.

[5] N. Kumar, J. J. Rodrigues, N. Chilamkurti, Bayesian coalition game as-a-service for content distribution in internet of vehicles, IEEE Internet of Things Journal 1 (6) (2014) 544–555.

[6] N. Kumar, R. Iqbal, S. Misra, J. J. Rodrigues, Bayesian coalition game for contention-aware reliable data forwarding in vehicular mobile cloud, Future Generation Computer Systems 48 (2015) 60–72.

[7] M. Noor-A-Rahim, Z. Liu, H. Lee, G. M. N. Ali, D. Pesch, P. Xiao, A survey on resource allocation in vehicular networks, IEEE Transactions on Intelligent Transportation Systems (2020).

[8] Y. Sun, M. Peng, Y. Zhou, Y. Huang, S. Mao, Application of machine learning in wireless networks: Key techniques and open issues, IEEE Communications Surveys & Tutorials 21 (4) (2019) 3072–3108.

[9] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, D. I. Kim, Applications of deep reinforcement learning in communications and networking: A survey, IEEE Communications Surveys & Tutorials 21 (4) (2019) 3133–3174.

[10] S. Nayak, R. Patgiri, 6G: Envisioning the key issues and challenges, EAI Endorsed Transactions on Internet of Things 6 (24) (2020).

[11] T. Yuan, W. B. da Rocha Neto, C. E. Rothenberg, K. Obraczka, C. Barakat, T. Turletti, Machine learning for next-generation intelligent transportation systems: A survey, Tech. rep., eprint hal-02284820 (2019).
URL https://hal.inria.fr/hal-02284820

[12] I. Althamary, C.-W. Huang, P. Lin, A survey on multi-agent reinforcement learning methods for vehicular networks, in: 15th International Wireless Communications & Mobile Computing Conference (IWCMC), IEEE, 2019, pp. 1154–1159.

[13] F. Tang, Y. Kawamoto, N. Kato, J. Liu, Future intelligent and secure vehicular network toward 6G: Machine-learning approaches, Proceedings of the IEEE 108 (2) (2019) 292–307.

[14] Z. Liu, H. Lee, M. O. Khyam, J. He, D. Pesch, K. Moessner, W. Saad, H. V. Poor, et al., 6G for vehicle-to-everything (V2X) communications: Enabling technologies, challenges, and opportunities (2020). arXiv:2012.07753.

[15] X. Jiang, M. Sheng, N. Zhao, C. Xing, W. Lu, X. Wang, Green uav communications for 6g: A survey, Chinese Journal of Aeronautics (2021).

[16] M. Muhammad, G. A. Safdar, Survey on existing authentication issues for cellular-assisted V2X communication, Vehicular Communications 12 (2018) 50–65.

[17] J. Jeong, Y. Shen, T. Oh, S. Céspedes, N. Benamar, M. Wetterwald, J. Härri, A comprehensive survey on vehicular networks for smart roads: A focus on ip-based approaches, Vehicular Communications (2021) 100334.

[18] P. K. Singh, S. K. Nandi, S. Nandi, A tutorial survey on vehicular communication state of the art, and future research directions, Vehicular Communications 18 (2019) 100164.

[19] X. Wang, S. Mao, M. X. Gong, An overview of 3GPP cellular vehicle-to-everything standards, GetMobile: Mobile Computing and Communications 21 (3) (2017) 19–25.

[20] C. Campolo, A. Molinaro, A. Iera, F. Menichella, 5G network slicing for vehicle-to-everything services, IEEE Wireless Communications 24 (6) (2017) 38–45.

[21] G. Velez, Á. Martín, G. Pastor, E. Mutafungwa, 5G beyond 3GPP release 15 for connected automated mobility in cross-border contexts, Sensors 20 (22) (2020) 6622.

[22] E. Adegoke, J. Zidane, E. Kampert, C. R. Ford, S. A. Birrell, M. D. Higgins, Infrastructure wi-fi for connected autonomous vehicle positioning: A review of the state-of-the-art, Vehicular Communications 20 (2019) 100185.

[23] M. N. Mejri, J. Ben-Othman, M. Hamdi, Survey on VANET security challenges and possible cryptographic solutions, Vehicular Communications 1 (2) (2014) 53–66.

[24] A. J. Gopinath, B. Nithya, An optimal multi-channel coordination scheme for IEEE 802.11 p based vehicular adhoc networks (VANETs), in: 2019 11th International Conference on Communication Systems & Networks (COMSNETS), IEEE, 2019, pp. 38–43.

[25] N. Gupta, A. Prakash, R. Tripathi, Medium access control protocols for safety applications in vehicular ad-hoc network: A classification and comprehensive survey, Vehicular Communications 2 (4) (2015) 223–237.

[26] X. Lin, J. G. Andrews, A. Ghosh, R. Ratasuk, An overview of 3GPP device-to-device proximity services, IEEE Communications Magazine 52 (4) (2014) 40–48.

[27] S. Chen, J. Hu, Y. Shi, Y. Peng, J. Fang, R. Zhao, L. Zhao, Vehicle-to-everything (V2X) services supported by LTE-based systems and 5G, IEEE Communications Standards Magazine 1 (2) (2017) 70–76.

[28] M. H. C. Garcia, A. Molina-Galan, M. Boban, J. Gozalvez, B. Coll-Perales, T. Şahin, A. Kousaridas, A tutorial on 5g nr v2x communications, arXiv preprint arXiv:2102.04538 (2021).

[29] F. Tariq, M. R. Khandaker, K.-K. Wong, M. A. Imran, M. Bennis, M. Debbah, A speculative study on 6G, IEEE Wireless Communications 27 (4) (2020) 118–125.

[30] M. Giordani, M. Polese, M. Mezzavilla, S. Rangan, M. Zorzi, Toward 6G networks: Use cases and technologies, IEEE Communications Magazine 58 (3) (2020) 55–61.

[31] R. Gupta, A. Nair, S. Tanwar, N. Kumar, Blockchain-assisted secure uav communication in 6g environment: Architecture, opportunities, and challenges, IET Communications (2021).

[32] K. Sheth, K. Patel, H. Shah, S. Tanwar, R. Gupta, N. Kumar, A taxonomy of ai techniques for 6g communication networks, Computer Communications 161 (2020) 279–303.

[33] A. Gupta, R. K. Jha, A survey of 5g network: Architecture and emerging technologies, IEEE access 3 (2015) 1206–1232.

[34] Y. Kim, Y. Kim, J. Oh, H. Ji, J. Yeo, S. Choi, H. Ryu, H. Noh, T. Kim, F. Sun, et al., New radio (nr) and its evolution toward 5g-advanced, IEEE Wireless Communications 26 (3) (2019) 2–7.

[35] W. Roh, J.-Y. Seol, J. Park, B. Lee, J. Lee, Y. Kim, J. Cho, K. Cheun, F. Aryanfar, Millimeter-wave beamforming as an enabling technology for 5g cellular communications: Theoretical feasibility and prototype results, IEEE communications magazine 52 (2) (2014) 106–113.

[36] A. Dogra, R. K. Jha, S. Jain, A survey on beyond 5g network with the advent of 6g: Architecture and emerging technologies, IEEE Access 9 (2020) 67512–67547.

[37] R. S. Sutton, A. G. Barto, Reinforcement learning: An introduction, MIT Press, 2018.

[38] M. T. Spaan, Partially observable markov decision processes, in: Reinforcement Learning, Springer, 2012, pp. 387–414.

[39] K. Arulkumaran, M. P. Deisenroth, M. Brundage, A. A. Bharath, Deep reinforcement learning: A brief survey, IEEE Signal Processing Magazine 34 (6) (2017) 26–38.

[40] S. Ivanov, A. D'yakonov, Modern deep reinforcement learning algorithms (2019). arXiv:1906.10025.

[41] P. Henderson, R. Islam, P. Bachman, J. Pineau, D. Precup, D. Meger, Deep reinforcement learning that matters, in: The Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 3207–3214.

[42] I. Goodfellow, Y. Bengio, A. Courville, Deep learning, MIT Press, 2016.

[43] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications (2018). arXiv:1812.08434.

[44] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, Advances in neural information processing systems 27 (2014).

[45] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, Nature 518 (7540) (2015) 529–533.

[46] H. Van Hasselt, A. Guez, D. Silver, Deep reinforcement learning with double Q-learning, in: The Thirtieth AAAI Conference on Artificial Intelligence, 2016, pp. 2094–2100.

[47] T. Schaul, J. Quan, I. Antonoglou, D. Silver, Prioritized experience replay, in: International Conference on Learning Representations (ICLR), 2016.

[48] Z. Wang, T. Schaul, M. Hessel, H. Hasselt, M. Lanctot, N. Freitas, Dueling network architectures for deep reinforcement learning, in: 33rd International conference on machine learning, PMLR, 2016, pp. 1995–2003.

[49] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, K. Kavukcuoglu, Asynchronous methods for deep reinforcement learning, in: International conference on machine learning, PMLR, 2016, pp. 1928–1937.

[50] M. G. Bellemare, W. Dabney, R. Munos, A distributional perspective on reinforcement learning, in: 34th International Conference on Machine Learning, PMLR, 2017, pp. 449–458.

[51] M. Fortunato, M. G. Azar, B. Piot, J. Menick, I. Osband, A. Graves, V. Mnih, R. Munos, D. Hassabis, O. Pietquin, et al., Noisy networks for exploration, in: Sixth International Conference on Learning Representations (ICLR), 2018.

[52] M. Hessel, J. Modayil, H. Van Hasselt, T. Schaul, G. Ostrovski, W. Dabney, D. Horgan, B. Piot, M. Azar, D. Silver, Rainbow: Combining improvements in deep reinforcement learning, in: The Thirty-

Second AAAI Conference on Artificial Intelligence, 2018, pp. 3215–3222.

[53] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, M. Riedmiller, Deterministic policy gradient algorithms, in: 31st International conference on machine learning, PMLR, 2014, pp. 387–395.

[54] M. Hausknecht, P. Stone, Deep recurrent Q-learning for partially observable MDPs (2015). arXiv:1507.06527.

[55] D. Zhao, H. Wang, K. Shao, Y. Zhu, Deep reinforcement learning with experience replay based on SARSA, in: 2016 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE, 2016, pp. 1–6.

[56] H. Ye, G. Y. Li, B.-H. F. Juang, Deep reinforcement learning based resource allocation for V2V communications, IEEE Transactions on Vehicular Technology 68 (4) (2019) 3163–3173.

[57] H. Ye, G. Y. Li, Deep reinforcement learning for resource allocation in V2V communications, in: 2018 IEEE International Conference on Communications (ICC), IEEE, 2018, pp. 1–6.

[58] H. Ye, G. Y. Li, Deep reinforcement learning based distributed resource allocation for V2V broadcasting, in: 14th International Wireless Communications & Mobile Computing Conference (IWCMC), IEEE, 2018, pp. 440–445.

[59] L. Liang, H. Ye, G. Y. Li, Spectrum sharing in vehicular networks based on multi-agent reinforcement learning, IEEE Journal on Selected Areas in Communications 37 (10) (2019) 2282–2292.

[60] H. V. Vu, Z. Liu, D. H. Nguyen, R. Morawski, T. Le-Ngoc, Multi-agent reinforcement learning for joint channel assignment and power allocation in platoon-based C-V2X systems (2020). arXiv:2011.04555.

[61] X. Chen, C. Wu, T. Chen, H. Zhang, Z. Liu, Y. Zhang, M. Bennis, Age of information aware radio resource management in vehicular networks: A proactive deep reinforcement learning perspective, IEEE Transactions on Wireless Communications 19 (4) (2020) 2268–2281.

[62] R. Hu, X. Wang, Y. Su, B. Yang, An efficient deep reinforcement learning based distributed channel multiplexing framework for V2X communication networks, in: 2021 IEEE International Conference on

Consumer Electronics and Computer Engineering (ICCECE), IEEE, 2021, pp. 154–160.

[63] S. Yan, X. Zhang, H. Xiang, W. Wu, Joint access mode selection and spectrum allocation for fog computing based vehicular networks, IEEE Access 7 (2019) 17725–17735.

[64] H. Yang, X. Xie, M. Kadoch, Intelligent resource management based on reinforcement learning for ultra-reliable and low-latency IoV communication networks, IEEE Transactions on Vehicular Technology 68 (5) (2019) 4157–4169.

[65] X. Zhang, M. Peng, S. Yan, Y. Sun, Deep-reinforcement-learning-based mode selection and resource allocation for cellular V2X communications, IEEE Internet of Things Journal 7 (7) (2019) 6380–6391.

[66] D. Zhao, H. Qin, B. Song, Y. Zhang, X. Du, M. Guizani, A reinforcement learning method for joint mode selection and power adaptation in the V2V communication network in 5G, IEEE Transactions on Cognitive Communications and Networking 6 (2) (2020) 452–463.

[67] C. Choe, J. Choi, J. Ahn, D. Park, S. Ahn, Multiple channel access using deep reinforcement learning for congested vehicular networks, in: 2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring), IEEE, 2020, pp. 1–6.

[68] C. Wu, S. Ohzahata, Y. Ji, T. Kato, A MAC protocol for delay-sensitive VANET applications with self-learning contention scheme, in: 2014 IEEE 11th Consumer Communications and Networking Conference (CCNC), IEEE, 2014, pp. 438–443.

[69] A. Pressas, Z. Sheng, F. Ali, D. Tian, M. Nekovee, Contention-based learning MAC protocol for broadcast vehicle-to-vehicle communication, in: 2017 IEEE Vehicular Networking Conference (VNC), IEEE, 2017, pp. 263–270.

[70] A. Pressas, Z. Sheng, F. Ali, D. Tian, A Q-learning approach with collective contention estimation for bandwidth-efficient and fair access control in IEEE802.11 p vehicular networks, IEEE Transactions on Vehicular Technology 68 (9) (2019) 9136–9150.

[71] D.-j. Lee, Y. Deng, Y.-J. Choi, Back-off improvement by using Q-learning in IEEE802.11p vehicular network, in: 2020 International

Conference on Information and Communication Technology Convergence (ICTC), IEEE, 2020, pp. 1819–1821.

[72] N.-O. Song, B.-J. Kwak, J. Song, M. Miller, Enhancement of IEEE 802.11 distributed coordination function with exponential increase exponential decrease backoff algorithm, in: 57th IEEE Semiannual Vehicular Technology Conference, 2003. VTC 2003-Spring., Vol. 4, IEEE, 2003, pp. 2775–2778.

[73] C.-H. Ke, C.-C. Wei, K. W. Lin, J.-W. Ding, A smart exponential-threshold-linear backoff mechanism for IEEE 802.11 WLANs, International Journal of Communication Systems 24 (8) (2011) 1033–1048.

[74] Z. Li, C. Wang, C.-J. Jiang, User association for load balancing in vehicular networks: An online reinforcement learning approach, IEEE Transactions on Intelligent Transportation Systems 18 (8) (2017) 2217–2228.

[75] M. F. Pervej, S.-C. Lin, Dynamic power allocation and virtual cell formation for throughput-optimal vehicular edge networks in highway transportation, in: 2020 IEEE International Conference on Communications Workshops (ICC Workshops), IEEE, 2020, pp. 1–7.

[76] M. F. Pervej, S.-C. Lin, Eco-vehicular edge networks for connected transportation: A distributed multi-agent reinforcement learning approach, in: IEEE 92nd Vehicular Technology Conference (VTC-Fall), 2020.

[77] Y. Liu, Z. Jiang, S. Zhang, S. Xu, Deep reinforcement learning-based beam tracking for low-latency services in vehicular networks, in: 2020 IEEE International Conference on Communications (ICC), IEEE, 2020, pp. 1–7.

[78] S. Konatowski, P. Kaniewski, J. Matuszewski, Comparison of estimation accuracy of EKF, UKF and PF filters, Annual of Navigation 23 (1) (2016) 69–87.

[79] Y. Xu, L. Li, B.-H. Soong, C. Li, Fuzzy Q-learning based vertical handoff control for vehicular heterogeneous wireless network, in: 2014 IEEE International Conference on Communications (ICC), IEEE, 2014, pp. 5653–5658.

[80] K. Zhang, Y. Zhu, S. Leng, Y. He, S. Maharjan, Y. Zhang, Deep learning empowered task offloading for mobile edge computing in urban informatics, IEEE Internet of Things Journal 6 (5) (2019) 7635–7647.

[81] Z. Ning, P. Dong, X. Wang, M. S. Obaidat, X. Hu, L. Guo, Y. Guo, J. Huang, B. Hu, Y. Li, When deep reinforcement learning meets 5G-enabled vehicular networks: A distributed offloading framework for traffic big data, IEEE Transactions on Industrial Informatics 16 (2) (2019) 1352–1361.

[82] H. Ke, J. Wang, L. Deng, Y. Ge, H. Wang, Deep reinforcement learning-based adaptive computation offloading for MEC in heterogeneous vehicular networks, IEEE Transactions on Vehicular Technology 69 (7) (2020) 7916–7929.

[83] H. Peng, X. S. Shen, Deep reinforcement learning based resource management for multi-access edge computing in vehicular networks, IEEE Transactions on Network Science and Engineering 7 (4) (2020) 2416–2428.

[84] Q. Qi, J. Wang, Z. Ma, H. Sun, Y. Cao, L. Zhang, J. Liao, Knowledge-driven service offloading decision for vehicular edge computing: A deep reinforcement learning approach, IEEE Transactions on Vehicular Technology 68 (5) (2019) 4192–4203.

[85] M. Li, J. Gao, N. Zhang, L. Zhao, X. S. Shen, Collaborative computing in vehicular networks: A deep reinforcement learning approach, in: 2020 IEEE International Conference on Communications (ICC), IEEE, 2020, pp. 1–6.

[86] M. Li, J. Gao, L. Zhao, X. Shen, Deep reinforcement learning for collaborative edge computing in vehicular networks, IEEE Transactions on Cognitive Communications and Networking 6 (4) (2020) 1122–1135.

[87] Z. Ning, K. Zhang, X. Wang, M. S. Obaidat, L. Guo, X. Hu, B. Hu, Y. Guo, B. Sadoun, R. Y. Kwok, Joint computing and caching in 5G-envisioned internet of vehicles: A deep reinforcement learning-based traffic control system, IEEE Transactions on Intelligent Transportation Systems (2020).

[88] G. Qiao, S. Leng, S. Maharjan, Y. Zhang, N. Ansari, Deep reinforcement learning for cooperative content caching in vehicular edge com-

puting and networks, IEEE Internet of Things Journal 7 (1) (2019) 247–257.

[89] Y. Dai, D. Xu, K. Zhang, S. Maharjan, Y. Zhang, Deep reinforcement learning and permissioned blockchain for content caching in vehicular edge computing and networks, IEEE Transactions on Vehicular Technology 69 (4) (2020) 4312–4324.

[90] K. Jiang, H. Zhou, D. Zeng, J. Wu, Multi-agent reinforcement learning for cooperative edge caching in internet of vehicles, in: 17th IEEE International Conference on Mobile Ad-Hoc and Smart Systems (MASS 2020), 2020, pp. 455–463.

[91] Y. He, C. Liang, Z. Zhang, F. R. Yu, N. Zhao, H. Yin, Y. Zhang, Resource allocation in software-defined and information-centric vehicular networks with mobile edge computing, in: 2017 IEEE 86th Vehicular Technology Conference (VTC-Fall), IEEE, 2017, pp. 1–5.

[92] Y. He, F. R. Yu, N. Zhao, H. Yin, A. Boukerche, Deep reinforcement learning (DRL)-based resource management in software-defined and virtualized vehicular ad hoc networks, in: 6th ACM Symposium on Development and Analysis of Intelligent Vehicular Networks and Applications, 2017, pp. 47–54.

[93] Y. He, N. Zhao, H. Yin, Integrated networking, caching, and computing for connected vehicles: A deep reinforcement learning approach, IEEE Transactions on Vehicular Technology 67 (1) (2017) 44–55.

[94] Y. He, F. R. Yu, N. Zhao, V. C. Leung, H. Yin, Software-defined networks with mobile edge computing and caching for smart cities: A big data deep reinforcement learning approach, IEEE Communications Magazine 55 (12) (2017) 31–37.

[95] R. Q. Hu, et al., Mobility-aware edge caching and computing in vehicle networks: A deep reinforcement learning, IEEE Transactions on Vehicular Technology 67 (11) (2018) 10190–10203.

[96] Y. Zhou, F. Tang, Y. Kawamoto, N. Kato, Reinforcement learning-based radio resource control in 5G vehicular network, IEEE Wireless Communications Letters 9 (5) (2019) 611–614.

[97] R. F. Atallah, C. M. Assi, J. Y. Yu, A reinforcement learning technique for optimizing downlink scheduling in an energy-limited vehicu-

lar network, IEEE Transactions on Vehicular Technology 66 (6) (2016) 4592–4601.

[98] R. Esmailzadeh, M. Nakagawa, E. A. Sourour, Time-division duplex CDMA communications, IEEE Personal Communications 4 (2) (1997) 51–56.

[99] A. Khoryaev, A. Chervyakov, M. Shilov, S. Panteleev, A. Lomayev, Performance analysis of dynamic adjustment of TDD uplink-downlink configurations in outdoor picocell LTE networks, in: 2012 IV International Congress on Ultra Modern Telecommunications and Control Systems, IEEE, 2012, pp. 914–921.

[100] F. Tang, Y. Zhou, N. Kato, Deep reinforcement learning for dynamic uplink/downlink resource allocation in high mobility 5G HetNet, IEEE Journal on Selected Areas in Communications 38 (12) (2020) 2773–2782.

[101] R. Atallah, C. Assi, M. Khabbaz, Deep reinforcement learning-based scheduling for roadside communication networks, in: 2017 15th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), IEEE, 2017, pp. 1–8.

[102] R. F. Atallah, C. M. Assi, M. J. Khabbaz, Scheduling the operation of a connected vehicular network using deep reinforcement learning, IEEE Transactions on Intelligent Transportation Systems 20 (5) (2018) 1669–1682.

[103] X. Qi, Y. Luo, G. Wu, K. Boriboonsomsin, M. J. Barth, Deep reinforcement learning-based vehicle energy efficiency autonomous learning system, in: 2017 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2017, pp. 1228–1233.

[104] Y. Hu, W. Li, K. Xu, T. Zahid, F. Qin, C. Li, Energy management strategy for a hybrid electric vehicle based on deep reinforcement learning, Applied Sciences 8 (2) (2018) 187.

[105] X. Hu, T. Liu, X. Qi, M. Barth, Reinforcement learning for hybrid and plug-in hybrid electric vehicle energy management: Recent advances and prospects, IEEE Industrial Electronics Magazine 13 (3) (2019) 16–25.

[106] B. Xu, D. Rathod, D. Zhang, A. Yebi, X. Zhang, X. Li, Z. Filipi, Parametric study on reinforcement learning optimized energy management strategy for a hybrid electric vehicle, Applied Energy 259 (2020) 114200.

[107] Y. Li, H. He, J. Peng, H. Wang, Deep reinforcement learning-based energy management for a series hybrid electric vehicle enabled by history cumulative trip information, IEEE Transactions on Vehicular Technology 68 (8) (2019) 7416–7430.

[108] G. Du, Y. Zou, X. Zhang, T. Liu, J. Wu, D. He, Deep reinforcement learning based energy management for a hybrid electric vehicle, Energy 201 (2020) 117591.

[109] G. Du, Y. Zou, X. Zhang, Z. Kong, J. Wu, D. He, Intelligent energy management for hybrid electric tracked vehicles using online reinforcement learning, Applied Energy 251 (2019) 113388.

[110] P. Wang, Y. Li, S. Shekhar, W. F. Northrop, A deep reinforcement learning framework for energy management of extended range electric delivery vehicles, in: 2019 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2019, pp. 1837–1842.

[111] P. Wang, Y. Li, S. Shekhar, W. F. Northrop, Actor-critic based deep reinforcement learning framework for energy management of extended range electric delivery vehicles, in: 2019 IEEE/ASME International Conference on Advanced Intelligent Mechatronics (AIM), IEEE, 2019, pp. 1379–1384.

[112] X. Liang, X. Du, G. Wang, Z. Han, A deep reinforcement learning network for traffic light cycle control, IEEE Transactions on Vehicular Technology 68 (2) (2019) 1243–1253.

[113] R. Zhang, A. Ishikawa, W. Wang, B. Striner, O. K. Tonguz, Using reinforcement learning with partial vehicle detection for intelligent traffic signal control, IEEE Transactions on Intelligent Transportation Systems (2020).

[114] J. V. Busch, V. Latzko, M. Reisslein, F. H. Fitzek, Optimised traffic light management through reinforcement learning: Traffic state agnostic agent vs. holistic agent with current V2I traffic state knowledge, IEEE Open Journal of Intelligent Transportation Systems 1 (2020) 201–216.

[115] T. Wu, P. Zhou, K. Liu, Y. Yuan, X. Wang, H. Huang, D. O. Wu, Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks, IEEE Transactions on Vehicular Technology 69 (8) (2020) 8243–8256.

[116] T. Chu, J. Wang, L. Codecà, Z. Li, Multi-agent deep reinforcement learning for large-scale traffic signal control, IEEE Transactions on Intelligent Transportation Systems 21 (3) (2019) 1086–1095.

[117] A. Hussain, T. Wang, C. Jiahua, Optimizing traffic lights with multi-agent deep reinforcement learning and V2X communication (2020). arXiv:2002.09853.

[118] F. Rasheed, K.-L. A. Yau, Y.-C. Low, Deep reinforcement learning for traffic signal control under disturbances: A case study on sunway city, Malaysia, Future Generation Computer Systems 109 (2020) 431–445.

[119] N. Wu, D. Li, Y. Xi, Distributed weighted balanced control of traffic signals for urban traffic congestion, IEEE Transactions on Intelligent Transportation Systems 20 (10) (2018) 3710–3720.

[120] J. Jin, X. Ma, A multi-objective agent-based control approach with application in intelligent traffic signal system, IEEE Transactions on Intelligent Transportation Systems 20 (10) (2019) 3900–3912.

[121] N. Kumar, S. S. Rahman, N. Dhakad, Fuzzy inference enabled deep reinforcement learning-based traffic light control for intelligent transportation system, IEEE Transactions on Intelligent Transportation Systems (2020).

[122] Z. Li, P. Liu, C. Xu, H. Duan, W. Wang, Reinforcement learning-based variable speed limit control strategy to reduce traffic congestion at freeway recurrent bottlenecks, IEEE Transactions on Intelligent Transportation Systems 18 (11) (2017) 3204–3217.

[123] C. Wang, J. Zhang, L. Xu, L. Li, B. Ran, A new solution for freeway congestion: Cooperative speed limit control using distributed reinforcement learning, IEEE Access 7 (2019) 41947–41957.

[124] E. Vinitsky, K. Parvate, A. Kreidieh, C. Wu, A. Bayen, Lagrangian control through deep-rl: Applications to bottleneck decongestion, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2018, pp. 759–765.

[125] T. Schmidt-Dumont, J. van Vuuren, A case for the adoption of decentralised reinforcement learning for the control of traffic flow on south african highways, Journal of the South African Institution of Civil Engineering 61 (3) (2019) 7–19.

[126] K. Kušić, I. Dusparic, M. Guériau, M. Gregurić, E. Ivanjko, Extended variable speed limit control using multi-agent reinforcement learning, in: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2020, pp. 1–8.

[127] K. Kušić, E. Ivanjko, M. Gregurić, M. Miletić, An overview of reinforcement learning methods for variable speed limit control, Applied Sciences 10 (14) (2020) 4917.

[128] S. Grigorescu, B. Trasnea, T. Cocias, G. Macesanu, A survey of deep learning techniques for autonomous driving, Journal of Field Robotics 37 (3) (2020) 362–386.

[129] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, P. Pérez, Deep reinforcement learning for autonomous driving: A survey, IEEE Transactions on Intelligent Transportation Systems (2021).

[130] S. Aradi, Survey of deep reinforcement learning for motion planning of autonomous vehicles, IEEE Transactions on Intelligent Transportation Systems (2020).

[131] W. Tong, A. Hussain, W. X. Bo, S. Maharjan, Artificial intelligence for vehicle-to-everything: A survey, IEEE Access 7 (2019) 10823–10843.

[132] A. L. Imoize, O. Adedeji, N. Tandiya, S. Shetty, 6G enabled smart infrastructure for sustainable society: Opportunities, challenges, and research roadmap, Sensors 21 (5) (2021) 1709.

[133] B. Yang, X. Cao, K. Xiong, C. Yuen, Y. L. Guan, S. Leng, L. Qian, Z. Han, Edge intelligence for autonomous driving in 6G wireless system: Design challenges and solutions (2020). arXiv:2012.06992.

[134] A. E. Sallab, M. Abdou, E. Perot, S. Yogamani, End-to-end deep reinforcement learning for lane keeping assist, in: 30th Conference on Neural Information Processing Systems (NIPS 2016), 2016.

[135] P. Wang, C.-Y. Chan, A. de La Fortelle, A reinforcement learning based approach for automated lane change maneuvers, in: 2018 IEEE Intelligent Vehicles Symposium (IV), IEEE, 2018, pp. 1379–1384.

[136] P. Wang, C.-Y. Chan, Formulation of deep reinforcement learning architecture toward autonomous driving for on-ramp merge, in: 2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2017, pp. 1–6.

[137] D. C. K. Ngai, N. H. C. Yung, A multiple-goal reinforcement learning method for complex vehicle overtaking maneuvers, IEEE Transactions on Intelligent Transportation Systems 12 (2) (2011) 509–522.

[138] A. Keselman, S. Ten, A. Ghazali, M. Jubeh, Reinforcement learning with A* and a deep heuristic (2018). arXiv:1811.07745.

[139] D. Isele, R. Rahimi, A. Cosgun, K. Subramanian, K. Fujimura, Navigating occluded intersections with autonomous vehicles using deep reinforcement learning, in: 2018 IEEE International Conference on Robotics and Automation (ICRA), IEEE, 2018, pp. 2034–2039.

[140] P. Palanisamy, Multi-agent connected autonomous driving using deep reinforcement learning, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–7.

[141] S. Bhalla, S. G. Subramanian, M. Crowley, Deep multi agent reinforcement learning for autonomous driving, in: Canadian Conference on Artificial Intelligence, Springer, 2020, pp. 67–78.

[142] C. Yu, X. Wang, X. Xu, M. Zhang, H. Ge, J. Ren, L. Sun, B. Chen, G. Tan, Distributed multiagent coordinated learning for autonomous driving in highways based on dynamic coordination graphs, IEEE Transactions on Intelligent Transportation Systems 21 (2) (2019) 735–748.

[143] G. Karagiannis, O. Altintas, E. Ekici, G. Heijenk, B. Jarupan, K. Lin, T. Weil, Vehicular networking: A survey and tutorial on requirements, architectures, challenges, standards and solutions, IEEE Communications Surveys & Tutorials 13 (4) (2011) 584–616.

[144] K. Keller, C. Zimmermann, J. Zibuschka, O. Hinz, Trust is good, control is better-customer preferences regarding control in teleoperated and autonomous taxis, in: 54th Hawaii International Conference on System Sciences, 2021, p. 1849.

[145] Y. Lu, L. Bi, H. Li, Model predictive-based shared control for brain-controlled driving, IEEE Transactions on Intelligent Transportation Systems 21 (2) (2019) 630–640.

[146] X.-a. Fan, L. Bi, T. Teng, H. Ding, Y. Liu, A brain–computer interface-based vehicle destination selection system using P300 and SSVEP signals, IEEE Transactions on Intelligent Transportation Systems 16 (1) (2014) 274–283.

[147] A. Hekmatmanesh, P. H. Nardelli, H. Handroos, Review of the state-of-the-art on bio-signal-based brain-controlled vehicles (2020). arXiv:2006.02937.

[148] A. Kumar, A. Bhisikar, A. K. Pandit, K. Singh, A. Shitole, Brain controlled car using deep neural network, Asian Journal For Convergence In Technology 5 (1) (2019).

[149] Y. Lu, X. Zheng, 6G: A survey on technologies, scenarios, challenges, and the related issues, Journal of Industrial Information Integration (2020) 100158.

[150] Y. Su, M. LiWang, L. Huang, X. Du, N. Guizani, Green communications for future vehicular networks: Data compression approaches, opportunities, and challenges, IEEE Network 34 (6) (2020) 184–190.

[151] Y. Wang, Research review of green vehicle routing optimization, in: IOP Conference Series: Earth and Environmental Science, Vol. 632, IOP Publishing, 2021, p. 032031.

[152] S. Chen, Y.-C. Liang, S. Sun, S. Kang, W. Cheng, M. Peng, Vision, requirements, and technology trend of 6G: How to tackle the challenges of system coverage, capacity, user data-rate and movement speed, IEEE Wireless Communications 27 (2) (2020) 218–228.

[153] R.-j. Yan, S. Pang, H.-b. Sun, Y.-j. Pang, Development and missions of unmanned surface vehicle, Journal of Marine Science and Application 9 (4) (2010) 451–457.

[154] Q. Zhang, W. Pan, V. Reppa, Model-reference reinforcement learning for collision-free tracking control of autonomous surface vehicles (2020). arXiv:2008.07240.

[155] D. Wang, T. Fan, T. Han, J. Pan, A two-stage reinforcement learning approach for multi-UAV collision avoidance under imperfect sensing, IEEE Robotics and Automation Letters 5 (2) (2020) 3098–3105.

[156] G. Raja, S. Anbalagan, V. S. Narayanan, S. Jayaram, A. Ganapathisubramaniyan, Inter-UAV collision avoidance using deep-Q-learning in flocking environment, in: 2019 IEEE 10th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), IEEE, 2019, pp. 1089–1095.

[157] H. Bayerlein, M. Theile, M. Caccamo, D. Gesbert, UAV path planning for wireless data harvesting: A deep reinforcement learning approach, in: 2020 IEEE Global Communications Conference (GLOBECOM), 2020.

[158] M. Theile, H. Bayerlein, R. Nai, D. Gesbert, M. Caccamo, UAV coverage path planning under varying power constraints using deep reinforcement learning, in: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2020.

[159] S. Y. Luis, D. G. Reina, S. L. T. Marín, A deep reinforcement learning approach for the patrolling problem of water resources through autonomous surface vehicles: The Ypacarai lake case, IEEE Access 8 (2020) 204076–204093.

[160] S. Y. Luis, D. G. Reina, S. L. T. Marín, A multiagent deep reinforcement learning approach for path planning in autonomous surface vehicles: The Ypacaraí lake patrolling case, IEEE Access 9 (2021) 17084–17099.

[161] J. Woo, N. Kim, Collision avoidance for an unmanned surface vehicle using deep reinforcement learning, Ocean Engineering 199 (2020) 107001.

[162] Y. Ma, Y. Zhao, Y. Wang, L. Gan, Y. Zheng, Collision-avoidance under COLREGS for unmanned surface vehicles via deep reinforcement learning, Maritime Policy & Management 47 (5) (2020) 665–686.

[163] Y. Lin, M. Wang, X. Zhou, G. Ding, S. Mao, Dynamic spectrum interaction of UAV flight formation communication with priority: A deep reinforcement learning approach, IEEE Transactions on Cognitive Communications and Networking 6 (3) (2020) 892–903.

[164] U. Challita, W. Saad, C. Bettstetter, Deep reinforcement learning for interference-aware path planning of cellular-connected UAVs, in: 2018 IEEE International Conference on Communications (ICC), IEEE, 2018, pp. 1–7.

[165] M. Wang, T. Zhu, T. Zhang, J. Zhang, S. Yu, W. Zhou, Security and privacy in 6G networks: New areas and new challenges, Digital Communications and Networks 6 (3) (2020) 281–291.

[166] S. Dawaliby, A. Aberkane, A. Bradai, Blockchain-based IoT platform for autonomous drone operations management, in: 2nd ACM MobiCom Workshop on Drone Assisted Wireless Communications for 5G and Beyond, 2020, pp. 31–36.

[167] X. Ling, J. Wang, T. Bouchoucha, B. C. Levy, Z. Ding, Blockchain radio access network (B-RAN): Towards decentralized secure radio access paradigm, IEEE Access 7 (2019) 9714–9723.

[168] S. Kiyomoto, A. Basu, M. S. Rahman, S. Ruj, On blockchain-based authorization architecture for beyond-5G mobile services, in: 2017 12th International Conference for Internet Technology and Secured Transactions (ICITST), IEEE, 2017, pp. 136–141.

[169] Y. Dai, D. Xu, S. Maharjan, Z. Chen, Q. He, Y. Zhang, Blockchain and deep reinforcement learning empowered intelligent 5G beyond, IEEE Network 33 (3) (2019) 10–17.

[170] M. Liu, F. R. Yu, Y. Teng, V. C. Leung, M. Song, Performance optimization for blockchain-enabled industrial internet of things (iiot) systems: A deep reinforcement learning approach, IEEE Transactions on Industrial Informatics 15 (6) (2019) 3559–3570.

[171] Y. Liu, H. Yu, S. Xie, Y. Zhang, Deep reinforcement learning for offloading and resource allocation in vehicle edge computing and networks, IEEE Transactions on Vehicular Technology 68 (11) (2019) 11158–11168.

[172] L. Mendiboure, M. A. Chalouf, F. Krief, Survey on blockchain-based applications in internet of vehicles, Computers & Electrical Engineering 84 (2020) 106646.

[173] L. Wang, D. Zheng, R. Guo, C. Hu, C. Jing, A blockchain-based privacy-preserving authentication scheme with anonymous identity in

vehicular networks, International Journal of Network Security 22 (6) (2020) 981–990.

[174] N. Malik, P. Nanda, A. Arora, X. He, D. Puthal, Blockchain based secured identity authentication and expeditious revocation framework for vehicular networks, in: 2018 17th IEEE international conference on trust, security and privacy in computing and communications/12th IEEE international conference on big data science and engineering (TrustCom/BigDataSE), IEEE, 2018, pp. 674–679.

[175] D. C. Nguyen, P. N. Pathirana, M. Ding, A. Seneviratne, Privacy-preserved task offloading in mobile blockchain with deep reinforcement learning, IEEE Transactions on Network and Service Management 17 (4) (2020) 2536–2549.

[176] H. Liu, Y. Zhang, T. Yang, Blockchain-enabled security in electric vehicles cloud and edge computing, IEEE Network 32 (3) (2018) 78–83.

[177] Y. Yahiatene, A. Rachedi, Towards a blockchain and software-defined vehicular networks approaches to secure vehicular social network, in: 2018 IEEE Conference on Standards for Communications and Networking (CSCN), IEEE, 2018, pp. 1–7.

[178] D. Zhang, F. R. Yu, R. Yang, Blockchain-based distributed software-defined vehicular networks: A dueling deep Q-learning approach, IEEE Transactions on Cognitive Communications and Networking 5 (4) (2019) 1086–1100.

[179] Z. Yang, K. Yang, L. Lei, K. Zheng, V. C. Leung, Blockchain-based decentralized trust management in vehicular networks, IEEE Internet of Things Journal 6 (2) (2018) 1495–1505.