IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier

A Robust Channel Access using Cooperative Reinforcement Learning for Congested Vehicular Networks

CHUNGJAE CHOE[1], JANGYONG AHN[1], JUNSUNG CHOI[1], DONGRYUL PARK[1], MINJUN KIM[2], AND SEUNGYOUNG AHN.[1], (Senior Member, IEEE)

[1]The CCS Graduate School of Green Transporation, Korea Advanced Institute of Science and Technology (KAIST), Daejeon 34141, South Korea (e-mail: cjchoe12@kaist.ac.kr)

[2]Korea Aerospace Research Institute (KARI), Daejeon 34133, South Korea

Corresponding author: Seungyoung Ahn (e-mail: sahn@kaist.ac.kr).

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.2020-0-00839,Development of Advanced Power and Signal EMC Technologies for Hyper-connected E-Vehicle. This research was supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2020-2016-0-00291) supervised by the IITP (Institute for Information & communications Technology Planning & Evaluation.

ABSTRACT Vehicular Ad-hoc Network (VANET) is an emerging technique dedicated to wireless vehicular communication to improve transportation safety by exchanging driving information between vehicles. For safety purposes, vehicles periodically broadcast a safety packet via Vehicle-to-Vehicle (V2V) communication. Accordingly, VANET safety applications demand a reliable exchange of the safety packet with high Packet Delivery Ratio (PDR), acceptable latency, and communication fairness. However, the communication performance significantly degrades due to numerous packet collisions when a large number of vehicles simultaneously access limited channel resources for the safety broadcast. In particular, the problem grows more severe in congested VANETs absent infrastructures since vehicles must control channel access using a self-adaptive scheme without external assistance. Thus, a robust and decentralized channel access protocol for VANETs is required to achieve road safety. In this paper, we propose an intelligent channel access algorithm empowered by cooperative Reinforcement Learning (RL), in which vehicles coordinate the channel access in a fully-decentralized manner. We also consider a proper interaction scheme between vehicles for enhancing the V2V safety broadcast in infrastructure-less congested VANETs. We provide evaluation results with extensive simulations according to various levels of traffic congestion. Simulations confirm the superior performance of the algorithm: the algorithm has a 20% increase in PDR compared to the latest RL-based channel access scheme. Furthermore, the algorithm satisfies the low latency requirement of VANET safety applications as well as both short-term and long-term communication fairness.

INDEX TERMS Vehicular ad-hoc network, congestion control, decentralized channel access, reinforcement learning, cooperative multi-agent systems.

I. INTRODUCTION

VEHICULAR ad-hoc Network (VANET) is an emerging technique that allows vehicular wireless communication to achieve transportation safety by exchanging traffic information between vehicles with infrastructures [1]. VANETs employ Dedicated Short-Range Communication (DSRC) protocol, which has seven communication channels in 5.9 GHz band for communication. As designated by the protocol, vehicles periodically broadcast a safety packet that contains a position, velocity, heading via Vehicle-to-Vehicle (V2V) communication. However, the broadcast performance is significantly deteriorated as the number of vehicles within a network increases owing to packet collisions caused by numerous simultaneous channel access [2]. Accordingly, the operation of safety applications that require a high Packet Delivery Ratio (PDR) with an acceptable end-to-end delay [3] will face challenges. Thus, the development of a robust channel access scheme for VANETs is a fundamental task to

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

enhance road safety.

Medium Access Control (MAC) layer is responsible for defining a feasible channel access protocol to avoid packet collisions. Since the MAC layer of VANETs is expected to simultaneously manage more than 100 interconnected vehicles, the operation of the channel access will become more intricate [4]. There are two types of conventional solutions: Centralized Channel Access (CCA) and Decentralized Channel Access (DCA) [5], [6]. In CCA-based solutions, such as Time Division Multiple Access (TDMA), infrastructures control the channel access of vehicles, which requires synchronization and connection among vehicles controlled by infrastructures [7]. Although CCA-based protocols [8] provide the highly stable performance on the safety broadcast, these may not be optimal solutions due to the following two distinctive features of VANETs: (1) infrastructures are sparsely located, and these cannot provide unlimited communication coverage [9]; (2) network topology and components are frequently changed by the dynamic movement of vehicles, so the synchronization and connection cannot be steady state [10]. Consequently, a DCA-based MAC unconstrained from the existence of infrastructures and the synchronization can be appropriate solution for VANETs.

Carrier-Sense Multiple Access with Collision Avoidance (CSMA/CA) [11] is the representative DCA protocol. Following the protocol, vehicles randomly access the channel according to the size of a selected Contention Window (CW) that provides a random amount of waiting time, *backoff*, before transmitting data. CSMA/CA-based DCA protocol can offer a proper solution with the advantages of rapid adaptability and scalability to dynamic changes of VANETs without depending on infrastructures [12]–[14]. However, the CSMA/CA-based DCA is highly challenging in congested VANETs with hundreds of vehicles that exist within a one-hop communication range, since vehicles are supposed to access a channel by a self CW-adaptation [15]. The following inherent problems can occur in this situation. First, vehicles suffer from a high packet collision rate since they are unable to obtain precise network status information of VANETs; therefore, they cannot adapt CW properly. Moreover, the large control data exchange via V2V communication to recognize the status information of VANETs can disturb the safety broadcast [12]. Finally, a communication fairness problem can arise in which vehicles cannot gain the equal opportunity of the channel access under network congestion [16]. Therefore, an intelligent MAC protocol is required that performs a suitable self CW-adaptation according to the network variation to deal with the aforementioned fundamental problems.

In this paper, we employ Reinforcement Learning (RL) [17] to develop an intelligent CSMA/CA-based DCA protocol. The pioneering studies [18]–[20] on an RL-based MAC for VANETs have demonstrated impressive functionality from solving the fundamental problems of the CSMA/CA-based DCA. Nevertheless, the previous works have not achieved sufficiently high performance of the safety broadcast due to the disregard for interactive CW-adaptation among vehicles. They ignore the interactive nature of VANETs, where the behaviors of adjacent vehicles can mutually affect broadcast success. To address the low-performance constraint, we propose a novel Cooperative RL-based MAC (CORL-MAC) algorithm using an interactive CW-adaptation policy between vehicles.

We model the VANET channel access problem as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) [21] and enhance a decentralized CW-adaptation policy with allowable cooperation among vehicles via V2V communications. To prevent disruption of the safety broadcast from a heavy extra transmission [22], we consider the constraint of imperfect communication to develop a suitable cooperation scheme contrary to previous multi-agent RL studies that assume seamless communication among agents [23], [24]. Furthermore, we deal with a stochastic communication nature of VANETs where the broadcast success is difficult to estimate owing to uncertain *backoff* selections of neighboring vehicles. In such case, a return-reward for RL training can be highly variable. Thus, we propose to exploit a distributional Deep Q Network (DQN) [25] and analyze its impact on the safety broadcast performance compared to the conventional DQN [26]. For the evaluation, we provide a highway traffic scenario to demonstrate the superior broadcast performance of the proposed algorithm in terms of PDR, end-to-end delay, and communication fairness from comparison with the latest previous work [20]. Our distinctive contributions of this paper are as follows:

- We propose a novel cooperative RL-based channel access algorithm employing Dec-POMDP modeling to achieve high PDR, acceptable end-to-end delay, and communication fairness for the safety broadcast in infraless congested VANETs.
- To improve the V2V safety broadcast performance, we present a suitable Dec-POMDP model that includes a semi-global state representation, a fairness-aware action function, and a weighted reward function. We analyze the cooperative behaviors between vehicles by applying the proposed model for each agent, which overcomes the limitation of the single-agent policy-based channel access algorithms [18]–[20]. Also, we demonstrate that a distributional DQN-based learning policy is a suitable approach for VANETs.
- The proposed algorithm enables cooperation among vehicles by exchanging minimal additional data following an alternative multi-channel protocol of DSRC, which does not interfere with the safety broadcast.

This paper is organized as follows. In Section II, we provide a literature review of related studies and the improvements in the present paper. Preliminaries of the study are presented in Section III. Section IV introduces the proposed algorithm in detail, and Section V evaluates the functionality of the algorithm. Lastly, the conclusion with a discussion of future works is drawn in Section VI.

II. RELATED WORK

The importance of an adaptive MAC protocol for VANETs has been strongly emphasized to improve the broadcast performance of safety packets [5], [6]. For this purpose, roadside infrastructures can be employed for a CCA since infrastructures can behave as a channel coordinator. Centralized TDMA-based algorithms can reduce packet collision probability significantly [8]. Besides, studies based on the channel interval decision have been conducted in [27] and [28]. An optimization algorithm for multi-channel intervals of the DSRC standard is presented to enhance the stability of the broadcast [27]. The minimization of idle service interval increases the performance of the broadcast using the reservation time mechanism [28]. Although the CCA-based protocols provide the advantage of reliable packet delivery, they are not suitable for VANETs absent infrastructures.

As a channel access solution of the V2V safety broadcast for infra-less VANETs, decentralized manners depend on self-adaptive channel access solely considering the interaction among vehicles has been presented [12], [13], [29], [30]. Clustering-based DCA algorithms [12], [29] provide a stable performance of the safety broadcast. Stable clusters boost communication performance even in dense VANETs. In addition, a hybrid CSMA and TDMA-based DCA mechanisms can guarantee fair medium sharing [13], [30]. Such schemes amend the centralized TDMA by enabling vehicles to acquire time slots more efficiently and to minimize slot conflicts. However, the aforementioned DCA schemes may paralyze channel operation and interfere with the safety broadcast due to a large control data exchange to maintain coordination between vehicles. Therefore, MAC protocols for VANETs should be studied to maximize operation efficiency with minimal burden on communication channels.

RL-based distributed MAC protocols [18]–[20] can be an optimal solution owing to its lightweight and decentralized features. Previous studies have focused on a self CW-adaptation that resolves contentions within VANETs. A Q-learning [17] MAC algorithm is proposed that defines each vehicle as a single agent and improves the performance of packet transmitting, but it only considers the V2V unicast case [18]. In order to enhance the V2V broadcast performance, a Q-learning-based MAC protocol employing a collective contention estimation-based reward function is proposed, and various experiments are executed to verify the performance improvement of the V2V broadcast [19], [20]. Large control data exchanging between vehicles is not required because the research only seek to improve the performance by transmitting minimum acknowledgements among vehicles. Accordingly, the protocol does not interfere with safety broadcast and has a strong advantage in the dynamic change of VANETs. However, the previous RL-based algorithms still show low performance of PDR in congested VANETs because they ignore the interactive nature of VANETs. Due to the unstable PDR performance, communication fairness issues may arise in highly congested VANETs with more than 100 vehicles within one-hop communication range. Furthermore, the previous RL-based protocols assume a single channel operation of the control channel (CCH) and do not consider a multi-channel operation of the DSRC standard. There is a potential to further increase the V2V broadcast performance in terms of PDR, latency, and fairness by employing a proper cooperation scheme among vehicles with an efficient multi-channel operation.

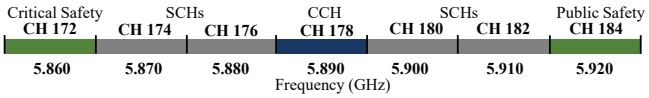


FIGURE 1: Spectrum of DSRC at 5.9GHz frequency band.

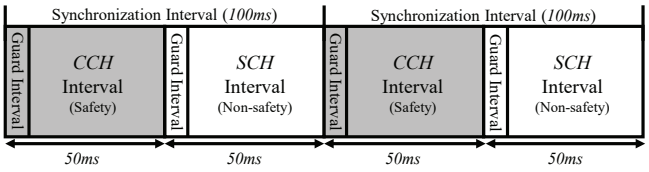


FIGURE 2: Alternating CCH and SCH intervals.

In contrast to previous studies [18]–[20] on the RL-based VANET MAC that have seldom considered interactions between vehicles, the salient contribution of this paper is that the proposed CORL-MAC has strong robustness to infra-less congested VANETs with a high level of traffic density ($40 - 120$ vehicles in a one-hop communication range). Moreover, the algorithm follows the operation standard of DSRC alternating multi-channel coordination and slightly revises the protocol with minimum additional data for the cooperation. The present paper demonstrates the performance merits of the proposed algorithm using numerous simulations in three performance metrics: average PDR, average end-to-end delay, and communication fairness.

III. PRELIMINARIES

This section provides the preliminaries of the paper. We introduce the basics of VANETs including multi-channel DSRC, various communication requirements, and the principle of CSMA/CA. Also, we present a concept of RL-based CW-adaptation with various learning policies.

A. OVERVIEW OF VANET CHANNEL ACCESS

1) DSRC Multi-Channel Operation with V2V Safety Broadcast

The DSRC standard allocates 75 MHz of spectrum in the 5.9 GHz frequency band for vehicular communications [1]. As shown in Fig. 1, the spectrum of DSRC consists of seven 10 MHz channels. Channel 178 corresponds to CCH, which is allocated for the broadcast of safety packets. There are six service channels (SCHs) primarily used for non-safety packets. Fig. 2 shows an alternation concept of DSRC multi-channel operation following IEEE 1609.4 standard [31]. The alternation is divided into a CCH interval (CCHI) and an SCH interval (SCHI), and each channel interval is 50 ms. Additionally, the IEEE 1609.4 standard typically dictate a 4ms guard interval at the start of each interval for the com-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

TABLE 1: Example of various communication requirements

Applications	Packet Size (bytes)	Latency (ms)
CCW, Lane Changing	$100-160$	≤ 100
Emergency Broadcast	$200-290$	≤ 100
CACC	$300-400$	≤ 20

pensation of interval switching time and timing inaccuracy between different DSRC devices. In principle, the DSRC standard requires the rule that every vehicle in VANETs broadcasts the safety packet for every 100 ms during CCHI. The safety packet is mainly exchanged by the V2V links between vehicles in a one-hop communication range for safety purposes. This requirement imposes a robust channel access challenge to the DSRC MAC layer, especially in infra-less congested VANETs.

2) Various Communication Requirements

Various types of VANET safety applications such as collision warning, lane-changing, and cruise control exist; communication requirements vary depending on the purpose of applications. Vehicles broadcast safety packets composed of various sizes according to specific applications. Table 1 shows the various communication requirements of VANET safety applications in previous studies [32]–[36]. For Cooperative Collision Warning (CCW) and lane-changing applications, $100-160$-bytes of safety packet should be broadcasted [32], [33]. Emergency broadcast applications demand more than 256-byte safety packet exchanges among vehicles [34]. Furthermore, Cooperative Adaptive Cruise Control (CACC) [35], [36] require more than 300-bytes packet exchanges for additional cooperation data. CACC typically demands ultra-low latency (≤ 20 ms) for directly adapting control system of vehicles whereas the other applications demand alleviated latency (≤ 100 ms) [37].

3) CSMA/CA and Limitations in congested VANETs

Basically, DSRC inherits IEEE 802.11p protocol which features a Distributed Coordination Function (DCF) for MAC layer operation. In addition, IEEE 802.11p employs CSMA/CA as the primary DCF-based channel access policy. CSMA/CA is a contention-based random channel access scheme. When a vehicle senses that the channel is idle, i.e., no-one is using the channel, it waits a random amount of time before transmitting data. This scheme is called a random *backoff* and the waiting duration, *backoff* value, is decided by the value of CW [11]. When vehicles sense the channel is idle during a DCF Interframse Space (DIFS), they draw a *backoff* value uniform randomly over the range of $[0, CW_{min}]$. *Backoff* values are decremented when the current channel is idle during a specified time-slot. If the channel becomes busy while decreasing a *backoff* value, vehicles have to wait for an Arbitation Inter-Frame Spacing (AIFS) to resume the countdown. When a *backoff* value decreases to 0, the packet is transmitted via a channel. Senders acquire

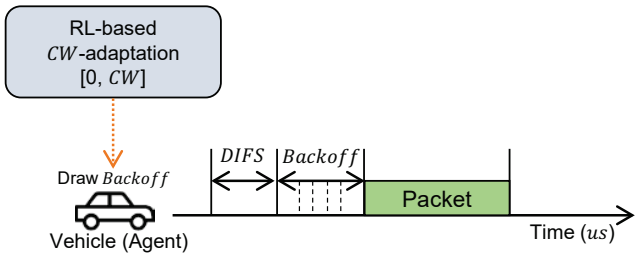


FIGURE 3: Reinforcement Learning-based CW-adaptation.

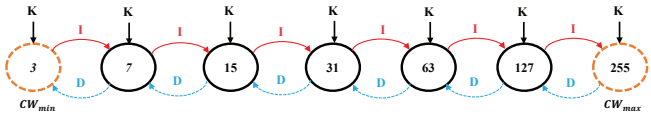


FIGURE 4: CW change according to the binary exponential backoff rule.

an acknowledgement packet (ACK) to recognize a successful transmission, whereas CW value doubles for each packet collision (no ACK) within a maximum CW (CW_{max}).

It is fundamental to adjust CW according to network density levels for a successful safety broadcast. However, the adaptation is complicated due to a clear trade-off between a successful packet delivery and a delay with transmission fairness [16]. A small CW likely to causes a packet collision in congested VANETs since two or more vehicles possibly draw the same *backoff* value. In particular, the probability of collision increases with periodic safety broadcasts. Although higher CW values can offer a stable packet delivery, these cause a huge delay as well as fairness issue. The adaptive CW decision is not simple work VANETs with dynamic nature. Besides, the problem becomes worse since IEEE 802.11p MAC prohibits the ACK transmission for broadcast packets to prevent the ACK storm phenomenon [22]. In other words, vehicles cannot obtain the information whether the safety broadcast was successful, and the adjustment of CW according to a network density is impossible. The broadcast operation of CSMA/CA becomes CSMA-like function without the collision avoidance mechanism [3]. As a result, the performance of the V2V safety broadcast is significantly degraded due to the high packet collision rate in the infra-less congested VANETs, and it interrupts the operation of V2V-based safety applications. Therefore, in order to tackle the collision and delay problems of VANETs, a robust channel access is needed that exploits a proper ACK transmission scheme to flexibly adapt the CW. We address the problem with a RL-based approach and consider a proper ACK scheme for the safety broadcast. The detailed description of our approach is drawn in Section. IV.

B. REINFORCEMENT LEARNING FOR CONTENTION WINDOW ADAPTATION

1) Basics of Reinforcement Learning Approach

RL designs a computational approach to interactive learning with an environment. An agent learns which action is most beneficial in each observed state by receiving a reward from

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

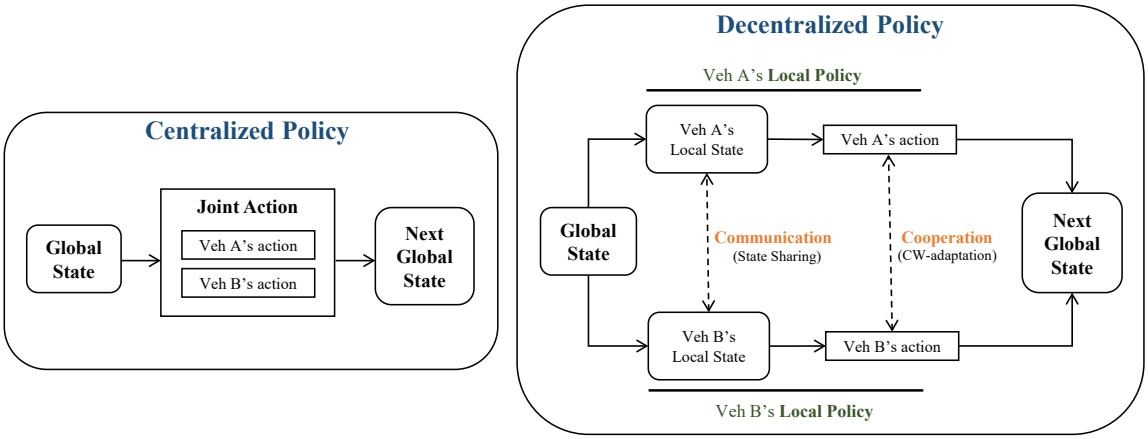


FIGURE 5: Difference between a cetralized policy and a decentralized policy. (a) centrlized CW-adaptation policy, (b) decentralized CW-adaptation policy.

an environment. Markov Decision Process (MDP) provides a mathematical foundation to represent the decision-making process of RL agents. A model-free MDP such as the VANET channel access problem consists of <s, a, r, s′> tuple. Fig. 3 illustrates the concept of RL-based CW-adaptation. Agents (vehicles) learn to control optimal CW values to select the channel access parameter *backoff*. Also, the following introduces an example concept of MDP for RL-based CW adaptation in reference to the previous works [18]–[20].

- **Agent** : All the vehicles in VANET are single agents.
- **State** (s): Agents observe the key features of the environment. The state can be expressed as a current CW value of an agent itself.
- **Action** (a): The action function adjusts the current CW value following pre-defined rules such as Keep (K), Increase (I), and Decrease (D). The previous works [18]–[20] consider the CW change regarding the binary exponential *backoff* [38] as shown in Fig. 4. There are seven CW transitions following a currently-selected action a;
- **Reward** (r): The reward function can be simplistic binary feedback. Vehicles (agents) receive 1 from a successful safety broadcast (ACK received) and -1 in the case of a failed broadcast (packet collision).
- **Next-State** (s′): The changed state after the agent takes action on state S. For example, the CW changes following the chosen action.

Agents aim to maximize the sum of cumulative rewards (future-oriented reward) that will be received until the final time-step of learning. The cumulative reward at time-step t is defined as

$$R_t = r_t + \gamma(r_{t+1} + \gamma(r_{t+2} + \ldots)) = r_t + \gamma R_{t+1}, \quad (1)$$

where the discount factor γ is for future discounting, between 0 and 1: closer to 0 is myopic and closer to 1 is far-sighted. Since the agent cannot recognize an accurate return value of the reward, the expectation value of the reward when the action is taken in the observed state is defined as the function called Q function as

$$Q(s_t, a_t) = E\big[R_{t+1}|s_t, a_t\big]. \quad (2)$$

The rule to select an action with the highest Q function value for each state is called policy $\pi(s)$,

$$\pi(s) = \underset{a}{\text{argmax}}\, Q(s, a). \quad (3)$$

The ultimate goal of the agent is to find the optimal action policy (π^*) that maximizes the cumulative reward when actions are taken in all possible states. The Bellman optimality equation [17] can be used to find actions that maximize Q(s,a) in each state, that is, the optimal action policy as

$$Q_{\pi^*}(s, a) = E\big[r_t + \gamma \underset{a'}{\text{max}} Q_{\pi^*}(s', a')|s, a\big]. \quad (4)$$

2) Decentralized Learning on Contention Window Adaptation

In the previous sub-section, we explain the basic MDP model for a RL-based CW adaptation. However, the conventional MDP model is not suitable for the VANET channel access problem. Vehicles cannot recognize the perfect status information of other vehicles in VANETs since they observe only partial information of the network due to the communication imperfection. Accordingly, the VANET channel access problem can be regarded as Partially Observable MDP (POMDP) [39]. Moreover, decentralized vehicles interact with the other communicational vehicles, so the Dec-POMDP [21] is proper for the CW adaptation problem. Without considering the interactive nature of VANETs, vehicles are likely to adjust CW that ignores the success transmission of other vehicles. Thus, it is essential to develop an effective learning method that maximizes communication performance in VANETs by considering a proper interaction of numerous vehicles.

There are two training approaches: centralized and decentralized policies to formulate the VANET channel access problem. Fig. 5 shows the difference between these two approaches. A centralized policy assumes global decision making that maximizes the communication performance by

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

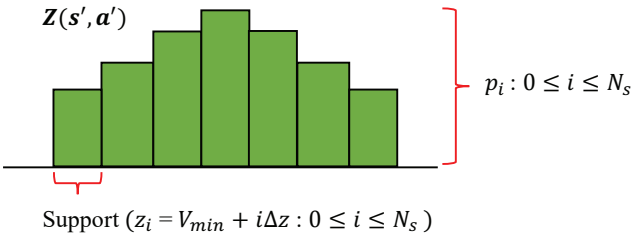


FIGURE 6: Output distribution of the distributional RL.

the centralized control. However, the centralized policy is unsuitable for infra-less VANETs since vehicles are placed in a decentralized network without the synchronization and connection. Vehicles are supposed to adjust the CW by themselves following a decentralized policy. They observe the network status only partially as a local state and choose action from each local policy. Only an acceptable level of cooperation via the V2V communication to enhance the local policy without interfering with the safety broadcast. Thus, we focus on Dec-POMDP for VANETs and enhance a decentralized CW-adaptation policy using the proposed cooperation technique.

3) A Distributional Perspective on Reinforcement Learning for Stochastic VANET Nature

Conventional RL [17] models the expectation of expected scalar return as shown in equation (2). However, from the stochastic nature of VANETs, it is hard to predict precise reward values since safety broadcast results will be different even in the same configuration of states and actions because of unpredictable CW-adaptations (actions) of other vehicles. For example, a state-action combination that introduces success broadcast in the previous time-step t-1 may fail at the current time-step t due to overlapping *backoff* values with other vehicles. Accordingly, rewards from VANETs are stochastic and can be represented by a multi-modal distribution. This phenomenon refers to a reward uncertainty. Therefore, choosing actions based on expected scalar values following the conventional RL policy may lead to suboptimal outcomes.

On the other hand, the distributional perspective on RL [25], Q-function represented in Equation (2) of the conventional DQN [26] is replaced by random variable $Z(s, a)$ as

$$Z(s, a) = r + \gamma Z(s', a'), \tag{5}$$

where Z represents the distribution of future rewards, which is not a scalar value as $Q(s, a)$. This replacement allows the vehicles to select actions based on the distribution instead of expected scalar values. Also, vehicles would choose actions with smaller variance in the future reward distributions. The multi-modal distributional approach provides more efficient exploration in stochastic environments, inducing the vehicles closer to the optimal behavioral policy. Fig. 6 illustrates the discrete probability distribution Z of each action; the x-axis is called the support zi that represents the value, and the y-axis represents the probability of support.

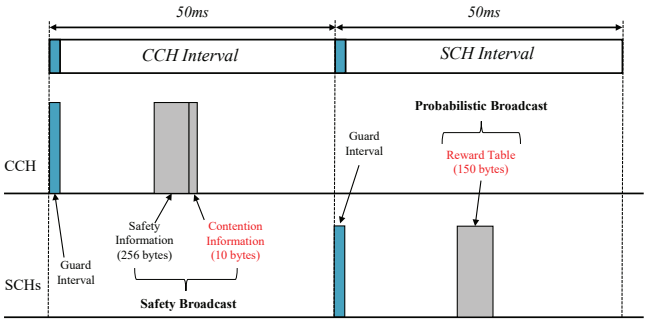


FIGURE 7: Proposed multi-channel operation.

The output of the network is the discrete probability distribution Z of each action; the x-axis is called the support z_i that represents the value, and the y-axis represents the probability of support. Accordingly, support values are divided at the same intervals according to the selected support number N_s range from the minimum value of V_{min} to the maximum value of V_{max}. Then, the Q value of each action is calculated by expectation of probabilities as

$$Q(x_t, a_t) = \sum_i z_i p_i(x_t, a_t), \tag{6}$$

where z_i and p_i correspond to the value and probability of each support, respectively. For further detailed information on the operating mechanism of the distributional approach, refer to [25]. From this reward prediction step, vehicle can effectively learn probabilistic action policy to deal with the stochastic nature of VANETs.

IV. THE PROPOSED CORL-MAC ALGORITHM

In this section, we formulate the proposed CORL-MAC algorithm with a Dec-POMDP model for a decentralized and cooperative decision-making problem. Our Dec-POMDP model is a tuple $<I, S, A, \pi_i, R>$, where I and S correspond to a finite set of agents and states, respectively. $\mathrm{A} = a_1 \times a_2 \times \cdots a_i$ is a finite set of actions of each agent I. π_i is the set of all observations for agents, and R corresponds to a reward function. At every decision time-step t, agents individually observe o_i^t and select action a_i^t. From the joint action $a_t = \{a_1^t, a_2^t, \cdots, a_N^t\}$ of agents, the environment state transitions from s_t to s_{t+1}.

A. OVERVIEW OF THE PROPOSED DEC-POMDP MODEL

Fig. 7 shows the proposed multi-channel operation. It basically follows the IEEE 1609.4 protocol [31]; we only add minimal cooperation data to transmitted packets in each channel interval. During CCHI, every vehicle broadcasts the safety packet with additional 10 bytes of contention information. Also, following a probabilistic manner, some vehicles (only 10%) broadcast 150 bytes of a reward table to distribute global acknowledgements for vehicles within a network during SCHI. With this modification, a cooperative

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

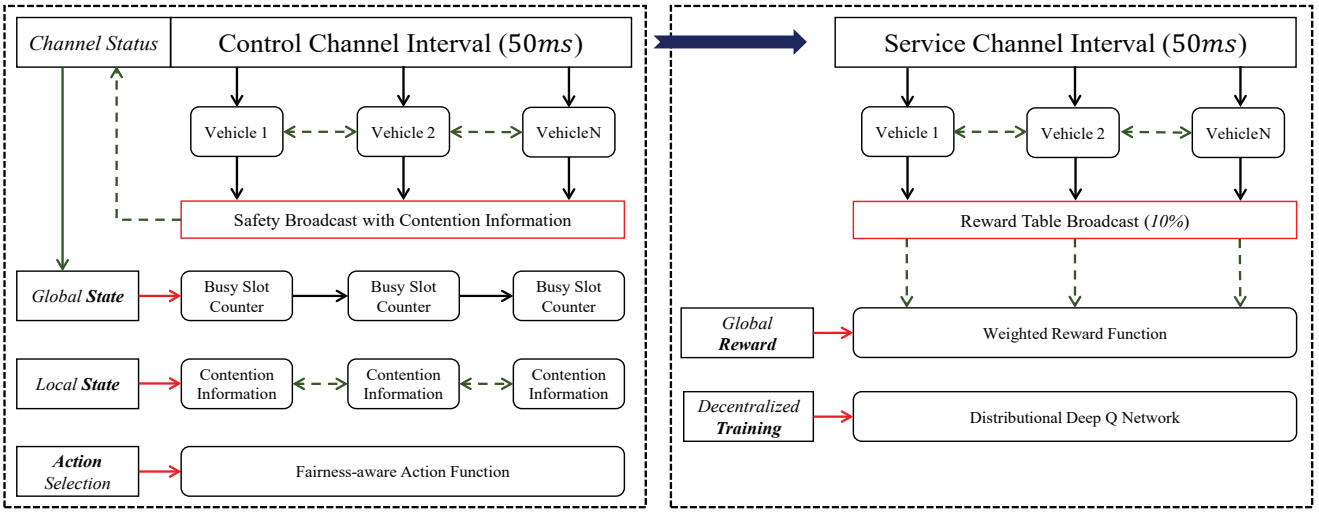


FIGURE 8: **A schematic view on the operation concept of the proposed CORL-MAC.**

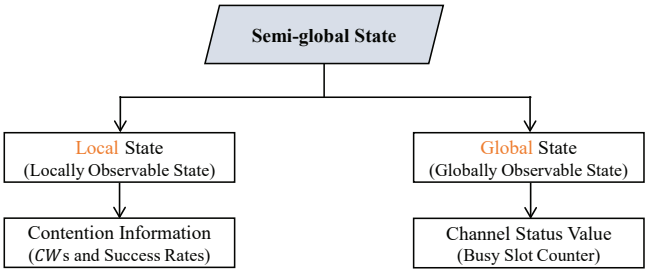
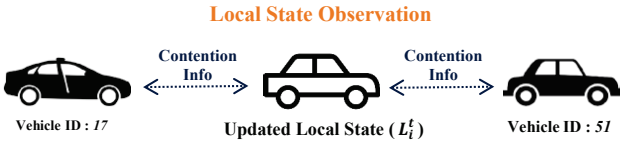


FIGURE 9: **Taxonomy of the proposed state representation.**



Contention Info = { Vehicle ID, *CW*, Success Rate }		
Vehicle ID	***CW***	**Success Rate**
17	**200**	**0.8**
51	**31**	**0.37**

FIGURE 10: **Local state observation.**

CW-adaptation protocol can be established with a small amount of additional data exchange between vehicles.

The overall operation concept of the algorithm is introduced in Fig. 8. Every vehicle in a network acts as a decentralized agent. Vehicles use the shared contention information and Busy Slot Counter (BSC) for state observation, and they choose an action for CW-adaptation during 50ms CCHI. Next, vehicles utilize a distributed reward table to recognize the broadcast success status of the network. In the next sub-sections, we provide in-depth explanation of the proposed Dec-POMDP model including the semi-global state representation, the fairness-aware action function, and the weighted reward function as well as the decentralized training network based on the distributional DQN [25].

B. STATE REPRESENTATION WITH DISTRIBUTED OBSERVATION

Fig. 9 shows the taxonomy of the proposed semi-global state representation. The state representation is divided into the local state and global state; we define a state space $S = \langle L_i, G_i \rangle$ where L_i and G_i are the local and global state of vehicle i, repectively. The local state L_i consists of collected contention information from neighboring vehicles, and it is only partially observed due to the imperfection of the safety broadcast in dense VANETs. Also, the global state G_i can be obtained by the channel status observation from a busy slot counting mechanism regardless of the safety broadcast success. The detailed configuration of the state space S is as follows.

Local State: Fig. 10 represents a local state observation, where L_i is observed from the broadcasted contention information including a vehicle ID, a currently selected CW value, and the broadcast success rate corresponding to the current CW. Vehicles partially observe the local state since every vehicle in a congested network cannot succeed the safety broadcast. Accordingly, the vehicles update the elements of local state L_i^{t-1} to L_i^t by contention information packets partially collected during CCHI. From the updated local state L_i, the indirect channel contention level can be estimated by the relationship between the collected CWs and success rates [40]. This is because in dense networks, the smaller CW values increase the likelihood of packet collision probability due to the overlapped *backoff* values among vehicles. In other words, only larger CW values guarantee the higher success rate of the safety broadcast by mitigating the *backoff* overlap problem. In sparse networks, on the other hand, even smaller CW values can yield high success rates. Therefore, this correlation allows vehicles to estimate the network congestion level.

Global State: Inspired by [41], [42], we exploit a BSC for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

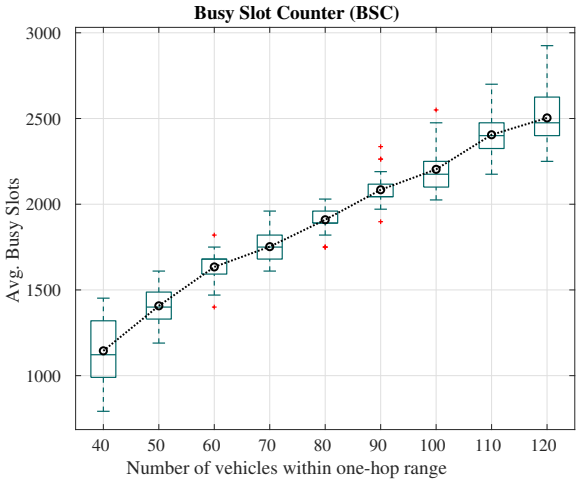


FIGURE 11: Average number of busy slots according to the number of vehicles.

the global state definition. Busy slot indicates the channel status at each time slot: busy or not. Unlike the partially observable local state, the global state is obtained by directly analyzing the status of the wireless channel regardless of whether the safety broadcast is successful or not. The global state G_i corresponds to the observed BSC of agent i. The inherent characteristics of the busy slot are as follows. The number of the busy slots of the channel increases as the number of broadcasting vehicles in VANETs grows since channel access attempts will also grow [41], [42]. Fig. 11 shows the relationship between the average BSC values and the number of vehicles. It is straightforward to recognize the upward trend due to the periodic channel access of vehicles for the safety broadcast. From the BSC observation, vehicles can straightforwadly distinguish the network congestion level. This global state can boost the learning performance of vehicles by offering the precise state information of the channel.

Distributed Observation: The observation π_i represents the current observation of each vehicle i following the state space S. It is defined as

$$\pi_i = \langle o^{L_i}, o^{G_i} \rangle, i \Subset N = \{index\ of\ vehicles\},$$

where o^{L_i} means the observed local state L_i from vehicle i. The local state also includes the self-information of vehicle i. o^{L_i} is updated by the successful broadcast of other vehicles in each time-step t. Otherwise, no vehicle succeeded in the safety broadcast, o^{L_i} remains without the update. o^{G_i} corresponds the observed global state G_i from agent i. This value can be updated in every time-step.

C. FAIRNESS-AWARE ACTION FUNCTION

Previous works [18]–[20] present an action function consisting of three choices such as keep, decrease, and variation of CW value ranging from CW_{min}= 3 to CW_{max}= 255 following the binary exponential *backoff* rule [38]. However, the previous action function has the following two limitations: 1) the limited seven transitions of CW as shown in Fig. 4 may hinder the performance of the safety broadcast; 2) short-term

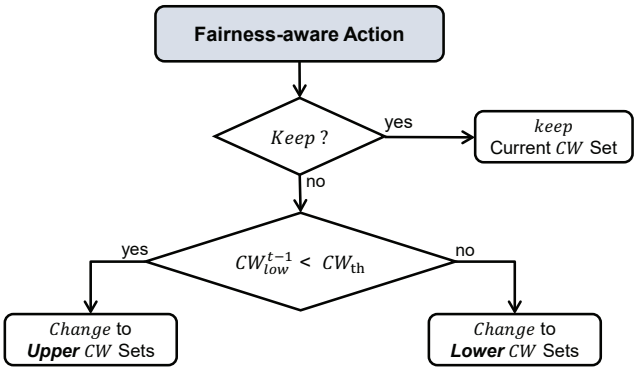


FIGURE 12: Flow chart of the fairness-aware action function.

TABLE 2: Predefined 11 action sets according to lower and upper sets

Action	Threshold	Boundary
$Keep$	-	$[CW_{low}^{t-1}, CW_{high}^{t-1}]$
$Change$-set 1	Lower set 1	$[3, 14]$
	Upper set 1	$[128, 140]$
$Change$-set 2	Lower set 2	$[15, 26]$
	Upper set 2	$[141, 153]$
$Change$-set 3	Lower set 3	$[27, 39]$
	Upper set 3	$[154, 166]$
$Change$-set 4	Lower set 4	$[40, 52]$
	Upper set 4	$[167, 179]$
$Change$-set 5	Lower set 5	$[53, 65]$
	Upper set 5	$[180, 192]$
$Change$-set 6	Lower set 6	$[66, 78]$
	Upper set 6	$[193, 205]$
$Change$-set 7	Lower set 7	$[79, 91]$
	Upper set 7	$[206, 218]$
$Change$-set 8	Lower set 8	$[92, 104]$
	Upper set 8	$[219, 231]$
$Change$-set 9	Lower set 9	$[105, 116]$
	Upper set 9	$[232, 244]$
$Change$-set 10	Lower set 10	$[117, 127]$
	Upper set 10	$[245, 255]$

fairness of packet delivery may not be achieved in highly congested networks. Due to the unfair CW selections, some vehicles can converge to select minimum CW, while the others converge to select maximum CW.

In order to improve the performance and the fairness of the safety broadcast, we propose the fairness-aware action function. We also present a new *backoff* rule enhancing the conventional CSMA/CA. Fig. 12 introduces the concept of the proposed function, where CW_{low}^{t-1} means previously selected CW_{low} and CW_{th} is a threshold value. The explanation of the proposed action function is as follows. First, the lowest CW (CW_{min}) and the highest CW (CW_{max}) are defined as 3 and 255, respectively. Second, the action function consists of 11 actions, vehicles can select $Change$-actions to decide one of 10 predefined sets of a *backoff* drawing boundary (CW_{low} and CW_{high}), or select $Keep$-action to maintain a previous boundary condition. The value of *backoff* is extracted uniform randomly within the boundary $[CW_{low}, CW_{high}]$. Vehicles decide the drawing boundary of *backoff* value from the action a_i. Furthermore, delay

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

Algorithm 1 Delay Fairness Action Function

Input
$CW_{min} = 3, CW_{max} = 255.$
Notation
a_t: selected action $(1 - 11)$.
CW_{low}: lower boundary value.
CW_{high}: upper boundary value.
CW_{th}: a threshold value $(CW_{max}/2)$.
t: time-step.

if $a_t = 1(Keep)$ **then**
  maintain the previous drawing boundary condition of $[CW_{low}^{t-1}, CW_{high}^{t-1}]$
else
  if $CW_{low}^{t-1} \leq CW_{th}$ **then**
    change $[CW_{low}, CW_{high}]$ following *upper CW sets* in Table 2
  else
    change $[CW_{low}, CW_{high}]$ following *lower CW sets* in Table 2
  end if
end if

Reward Table	
Vehicle ID	***Success***
17	**200**
51	**31**

FIGURE 13: Example of reward table definition.

discrimination is executed to satisfy communication fairness and solve the overlapped *backoff* problem. We define the boundary selection rule restricted to upper CW sets or lower CW sets based on the threshold value $CW_{th} = 127$ (CW_{max} / 2). If the selected CW_{low}^{t-1} at the previous time-step $t - 1$ less than or equal to CW_{th}, the boundary condition is drawn from the upper CW sets. Otherwise, if the previously selected CW_{low}^{t-1} is larger than CW_{th}, the drawing boundary is selected from the lower CW sets. Table 2 shows the $keep$ and 10 $change$ actions regarding the predefined boundary values of each lower and upper set. From this function, vehicles alternately select between the lower CW sets and the upper CW sets. The alternation of the *backoff* selection boundary can solve the unfairness problem that certain vehicles converge to a high or low CW selection policy. Moreover, a high success rate of the safety broadcast can be achieved by minimizing the overlapped *backoff* values among vehicles. The change of CW boundary according to a selected action is applied to both the CCHI and SCHI. Algorithm 1 provides the protocol of the proposed action function.

D. WEIGHTED REWARD FUNCTION

The acknowledgement packet provides a feedback on whether the safety broadcast was successful for vehicles

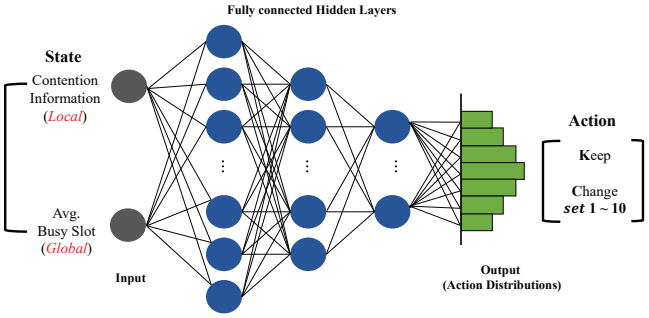


FIGURE 14: Proposed distributional DQN architecture.

to adjust CW size according to network congestion levels. Although the feedback is required, numerous ACK transmissions must be avoided to prevent an ACK storm phenomenon [22]. Thus, it is necessary to regulate an appropriate ACK scheme to define a reward function without exacerbating congestion on networks.

We propose a probabilistic broadcast ACK-based reward function to consider the interactive nature of VANETs. During SCHI, vehicles responsible for the ACK broadcast are selected uniform randomly with a 10% probability. A higher value of the probability should be avoided because it can interrupt non-safety packet transmissions. Selected vehicles (ACK broadcasters) build the reward table as shown in Fig. 13. The reward table contains Vehicle IDs and the corresponding Boolean values (0 or 1) to inform the broadcast result; 0 and 1 mean success and failure, respectively. According to the distributed reward table, each vehicle can identify whether the broadcast was successful in addition to broadcast results of other vehicles in a network. From the obtained reward tables from ACK broadcasters, vehicle i obtains a reward r as

$$r_i^t = \alpha \sum_{x=0}^{b} r_i + (1-\alpha)\left(\frac{1}{N-1} \sum_{j=0}^{N-1} \sum_{x=0}^{b}\right) r_{j,x}, \qquad (7)$$

where b is the number of broadcasted reward tables and r is the safety broadcast result (0 or 1) of each vehicle. N means the number of vehicles identified by Vehicle IDs in the reward table. Finally, α is the weight value, and we set it as 0.7 to present a higher reward to a self-success. According to the reward function, vehicles learn a cooperative action decision policy by considering not only the self-success of the safety broadcast but also that of neighboring vehicles.

E. DISTRIBUTIONAL DEEP Q NETWORK

To mitigate the reward uncertainty from the stochastic broadcast success in VANETs, we employ the distributional DQN [25]. Fig. 14 shows the network structure of the proposed distributional DQN consists of a Deep Neural Network (DNN) including three successive hidden layers with 256, 128, and 64 output dimensions, and Leaky Rectified Linear Unit (Leaky-ReLU) is used for the activation function. The observation $\pi_i = \langle o^{L_i}, o^{G_i} \rangle$ is the input, and the output is the estimated distribution for 11-actions ($Keep$, 10-$changes$). In

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

Algorithm 2 Cooperative Reinforcement Learning-based MAC

Input vehicle index i, safety packet P_{tx}, received packet P_{rx}, reward table P_{rew}, success rate of contention window R^{CW_i}, replay memory M_i, replay memory size m, number of episodes E, epsilon greedy ϵ
Notation
CW: contention window.
t: time-step.

for episode=1 to E **do**
  start CCHI
  procedure STATE-OBSERVATION(P_{rx})
    if P_{rx}**.IsBroadcast then**
      update each element in local state o^{L_i}
    end if
    update global state o^{G_i} according to BSC
  end procedure
  select an action a_i^{t+1} from Algorithm 1 according to ϵ-greedy
  $CW_i^{t+1} \leftarrow CW_{low}$
  procedure SAFETY-BROADCAST(P_{tx}, *backoff*)
    P_{tx}**.AddContentionInformation**(i,CW_i^{t+1},$R^{CW_i^{t+1}}$)
    Broadcast(*backoff*,P_{tx})
  end procedure
  start SCHI
  procedure REWARDTABLE-BROADCAST(P_{rx})
    if IsACKBroadcaster then
      Broadcast(P_{rew})
    end if
  end procedure
  procedure RECEIVED-FEEDBACK(P_{rx})
    if P_{rx}**.IsRewardTable** && P_{rx}**.ValidLatency then**
      r_i^t is calculated as (5)
    else
      $r_i^t \leftarrow 0$
    end if
    update R^{CW_i} acccording to CW_i^{t+1} and r_t
  end procedure
  procedure LEARNING
    store experience $< \pi_i^t, a_i^{t+1}, r_i^t, \pi_i^{t+1} >$ into M_i
    $t \leftarrow t+1$
    if $M > m$ **then**
      update network parameter θ_i following (8)
      update taget network parameter θ_i' according to (9)
    end if
  end procedure
end for

addition, the number of supports for the output distribution is configured as 51 following the guide on [25]. As a precaution against a local optimum problem, we also utilize the ϵ-greedy algorithm [17] to balance the exploration and exploitation in the action selection. We set the initial and minimum values of ϵ, which gradually decreases over the time-step following the discount factor. The weight parameter θ of the training network is updated along the gradient of a loss function, and the target network is defined to solve the problem in which the target that to be updated changes over time. The target network has a weight parameter θ', that is different from the θ. Consequently, the distributional DQN aims to predict the exact distribution output of the target network. The loss function J is defined as the cross-entropy of the difference between the target distribution and the estimated distribution,

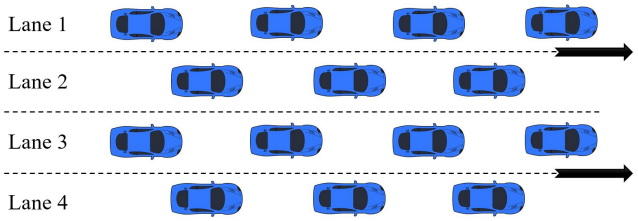


FIGURE 15: The highway traffic enviroment with four straight lanes.

and is calculated as

$$Loss = -\sum_i m_i \log p_i(x_t, a_t) \tag{8}$$

where m_i is the target distribution. Then, the network is updated along the gradient θ of the loss function. Finally, θ' is updated as according to the following soft-update policy as

$$\theta' = \alpha\theta' + (1-\alpha)\theta, \tag{9}$$

where α is an update rate. The target network parameter θ' will move slightly to the value θ. Finally, Algorithm 2 shows the learning process of the proposed distributional DQN based on our Dec-POMDP model.

V. EVALUATION

This section presents the simulation-based experiments for the performance evaluation of the proposed CORL-MAC algorithm. We compare the proposed algorithm to the latest RL-based MAC study [20] by highway scenario-based simulations according to diverse traffic density with various data sizes of the safety packet. Furthermore, we provide three performance metrics including PDR, end-to-end delay, and communication fairness. To build a simulation environment, the Simulation of Urban Mobility (SUMO) is employed to generate realistic traffic mobility. Then, the network simulator NS-3 is coupled with the generated mobility from SUMO. We also use the ns3-gym library [43] and Keras to implement an RL programming baseline.

A. SIMULATION SETUP

1) Traffic Scenario and Parameters

Fig. 15 describes the highway traffic scenario utilized for the experiments. The highway topology consists of four straight lanes with a length of 3km. In the traffic scenario, infrastructure is not deployed so vehicles are completely decentralized. Vehicles are only able to communicate by V2V without external coordination for channel access.

Table 3 shows the traffic parameters for the simulation experiments. We consider a varied number of vehicles to simulate various traffic congestion levels. The number of vehicles ranges from from 40 to 120 at 20 intervals (five congestion levels), where 120 vehicles correspond to the vehicle density during rush hour on a highway [44]. The congestion levels are expressed such as 40-VANET, 100-VANET, and 120-VANET. Simulations are alternately conducted according to the pre-defined five congestion levels. As the traffic flow consideration, vehicles follow the Krauss-

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

TABLE 3: Highway traffic parameters for the simulation

Parameter	Value
The number of vehicles	$40, 60, 80, 100, 120$
Car-following model	Krauss model [45]
Maximum velocity	16 m/s

car following model [45] with default parameters ($\sigma = 0.5, \tau = 1$) and the maximum velocity 16 m/s. Furthermore, we implement the traffic scenario following two conditions: (1) All the vehicles exist within the one-hop communication range ($\sim$ 1.2km) [46] and their position does not alter. From this condition, packet collisions can be accurately measured as the number of vehicles increases by eliminating the hidden terminal problem. (2) Since this study primarily focuses on the improvement of communication performance as the number of vehicles increases in a one-hop communication range, obstacles such as buildings are not considered.

2) VANET Parameters with Various Packet Configuration

To achieve diverse communication requirements, vehicles broadcast safety packet of various sizes according to specific applications [32]–[36]. In the experiments, the feasibility of the proposed algorithm is verified based on simulations regarding the three different size configurations of the safety packet: 128, 256, 384-byte. Accordingly, PDR, end-to-end delay and communication fairness are measured by each dedicated data size.

Table 4 summarizes the network parameters of the VANET considered for the simulation. The description of key parameters is as follows. First, we assign AIFS, CW_{min}, CW_{max} as the same transmission priority for vehicles since we focus on the broadcast function enhancement, not the priority enhancement. Second, three sizes of the safety packet are set to 128, 256, and 384 bytes including 10 bytes of the proposed contention information, and that of the non-safety packet is 400bytes. Third, we set the transmission power and the energy detection threshold as 25 dBm and -89 dBm, respectively, to cover the intended one-hop communication range ($\sim$ 1.2km). We also configure the data rate to 6 Mbps since higher data rates may not provide communication coverage up to 1 km because of the serious packet drop [47]. All the vehicles broadcast the safety packet at 100 ms intervals (10 Hz) via V2V links. Only some vehicles, decided by the probability Sn, transmit a non-safety packet at the SCHI. Simultaneously, the reward table broadcast is determined by the probability Bn at the SCHI. Lastly, we employ the Nakagami propagation model [48] to implement a realistic wireless channel considering received signal strength variations according to the distances between vehicles.

3) Channel Coordination with Assumptions

The proposed algorithm operates based on the multi-channel operation of IEEE 1609.4 [31]. The channel operation scenario dictates a vehicle switches every 50 ms between the CCH and the SCH according to the alternating channel access scheme. In the simulation experiments, we assume the following four conditions of channel coordination to operate the proposed algorithm appropriately.

(1) All the vehicles are equipped with a multi-channel DSRC device which allows switching to different channel frequencies. Also, DSRC devices equipped by vehicles are time-synchronized with Global Positioning System (GPS) based on the Coordinated Universal Time (UTC).
(2) Alternating channel access to CCH and SCH is utilized for channel coordination. The equipped DSRC devices are configured to alternatively monitor the CCH and SCH with the 50 ms interval. Also, the guard interval is defined as 4 ms to be tuned to another channel for accommodating device differences [31]. Transmission is not allowed during the guard interval. If a packet transmission is not finished at the start of the guard interval, the transmission is canceled and it is treated as a transmission failure.
(3) Vehicles exploit one SCH among six SCHs such as CH174 (5.870 GHz), dedicated to the acknowledgement (ACK) operation with the proposed reward table broadcast.
(4) Non-safety packets are generated uniform randomly in the SCHI; this condition enables selected vehicles following the uniform probability $B_n = 0.1$ to broadcast the reward table with a high success rate.

The assumptions may be not how VANETs operate in the real world, however, these offer significant insight into the broadcast performance of the safety packet when the congestion level of VANETs is varied.

TABLE 4: VANET parameters for the simulation

Parameter	Value
AIFS, CW_{min}, CW_{max}	3, 3, 255
CCH and SCH frequency	5.890 GHz and 5.870 GHz
Safety packet size	128 , 256 , 384 bytes
Reward table packet size	150 bytes
Non-safety packet size	400 bytes
Transmission power	25 dBm
Energy Detection Threshold	-89 dBm
Data rate	6 Mbps
Reward table broadcast probability B_n	0.1
Non-safety transmission probability S_n	0.2
Backoff slot time	13 us
Propagation model	Nakagami model [48]

4) Evaluation Targets and RL Parameters

We compare three types of algorithms via the simulation as follows:

- Q-MAC [20]: The simple Q-learning-based algorithm that utilizes a presented collective contention estimation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

in the latest study [20] that leads vehicles to select similar CW values with neighboring vehicles.
- Distributional DQN-based CORL-MAC (D-CORL-MAC): The proposed algorithm employing the distributional DQN for the learning process.
- Conventional DQN-based CORL-MAC (C-CORL-MAC): The modified version of the proposed algorithm, which uses the conventional DQN [26] structure without the distributional reward prediction approach.

In comparison, we evaluate that the distribution approach is suitable for VANETs from a self-comparison between the D-CORL-MAC and the C-CORL-MAC. Table 5 shows the network parameters for the learning process of the proposed algorithm. In the learning process, one episode corresponds to 10 seconds and $3,000$ training episodes are consumed for each congestion level. We set the epsilon decay rate γ as 0.9995. As we consider the soft update scheme of the target network, the update rate α is 0.001.

TABLE 5: DQN training parameters

Parameter	Value
Time-step t	$0.1\ s$
Replay memory size M	$10,000$
Mini-batch size	10
Epsilon decay rate	0.9995
Starting ϵ	1
Minimum ϵ	0.1
discount factor γ	0.99
Num of support z	51 [25]
Learning rate β	0.0001
Target network update rate α	0.001

B. SIMULATION RESULTS

We provide the detailed numerical analysis to validate if the three algorithms satisfy the various communication requirements as explained in Section III A. The numerous simulation experiments are established based on the highway traffic scenario. In order to investigate the impact of safety packet size, we adopt the three applications including application A, B, and C which is assigned 128, 256, and 384-byte safety packet, respectively. For each application, we present the five congestion levels-based evaluation results. Moreover, the evaluation results consist of the three metrics: average PDR, average end-to-end delay, and communication fairness of packet delivery. The average values of the three metrics indicate the measured numerical values from all the operating vehicles in a network. Those values are calculated by 300 test episodes after finishing the training process. We develop box plot results of the PDR and delay focusing on the proposed D-CORL-MAC for detailed analysis.

We also employ Jain's fairness index J [49] as

$$J(x_1, x_2, ..., x_n) = \frac{(\sum_{i=1}^{n} x_i)^2}{n \sum_{i=1}^{n} x_i^2}, \qquad (10)$$

where n represents the number of vehicles and x is the measured PDR value of each vehicle for the communication fairness evaluation. J is calculated over a sampling window range from 1 to 10 seconds (short-term to long-term) with a step size of 0.5 seconds following the previous work on the Q-MAC [20]. The fairness criterion is assigned as $J_c = 95\%$. We assess whether the algorithms satisfy the criterion J_c in both the short and long term according to the lowest and the highest congestion level (40-VANET and 120-VANET).

1) Average Packet Delivery Ratio

We first examine the average PDR of the V2V safety broadcast regarding the three packet size configurations. Fig. 16 shows the PDR evaluation result when the three different packet sizes are applied to the V2V safety broadcast operation. All the cases have a descending trend as vehicles increase. It represents the fundamental resource constraint of VANETs. In spite of the overall degradation, the proposed D-CORL-MAC has the highest average PDRs in all cases, while those of the C-CORL-MAC is relatively lower. From this result, we confirm that the distributional reward prediction is the suitable approach for VANETs which have the inherent uncertainty from unpredictable *backoff* selections of neighboring vehicles. Furthermore, the results show a clear performance improvement of the D-CORL-MAC over the Q-MAC since the D-CORL-MAC outperforms the Q-MAC in all cases. We present the detailed analysis of numerical results for each application as follows.

- **Application A (128-byte packets), Fig. 16(a)**: When the network congestion level is low such as the 40-VANET and 60-VANET, the average PDR of all algorithms satisfy the high performance over than 88%. In particular, the average PDR of the proposed D-CORL-MAC is higher than 95%. However, as the number of vehicles increases, the performance differences become noticeable, especially from the 80-VANET. The largest difference can be identified in the 100-VANET, where the D-CORL-MAC and C-CORL-MAC outperforms the Q-MAC. The D-CORL-MAC shows a 13% performance improvement over the Q-MAC; the D-CORL-MAC has around 91% average PDR while that of the Q-MAC is about 78%. In the highest congestion level, the 120-VANET, our D-CORL-MAC shows around 79% average PDR, and that of the Q-MAC is around 69%. In overall, the performance gap between D-CORL-MAC and C-CORL-MAC is not significant in application A.
- **Application B (256-byte packets), Fig. 16(b)**: The difference in average PDR between the D-CORL-MAC and the Q-MAC is wider compared to the case of application A. Although the three algorithms indicate the high performance of over 85% at low congestion levels below 60-VANET, the average PDR of the Q-MAC is highly deteriorated from the 80-VANET. In the case of 100-VANET, the D-CORL-MAC outperforms the Q-MAC by 18%, each algorithm satisfies the average PDR of 86% and 68%, respectively. Furthermore, we observe

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

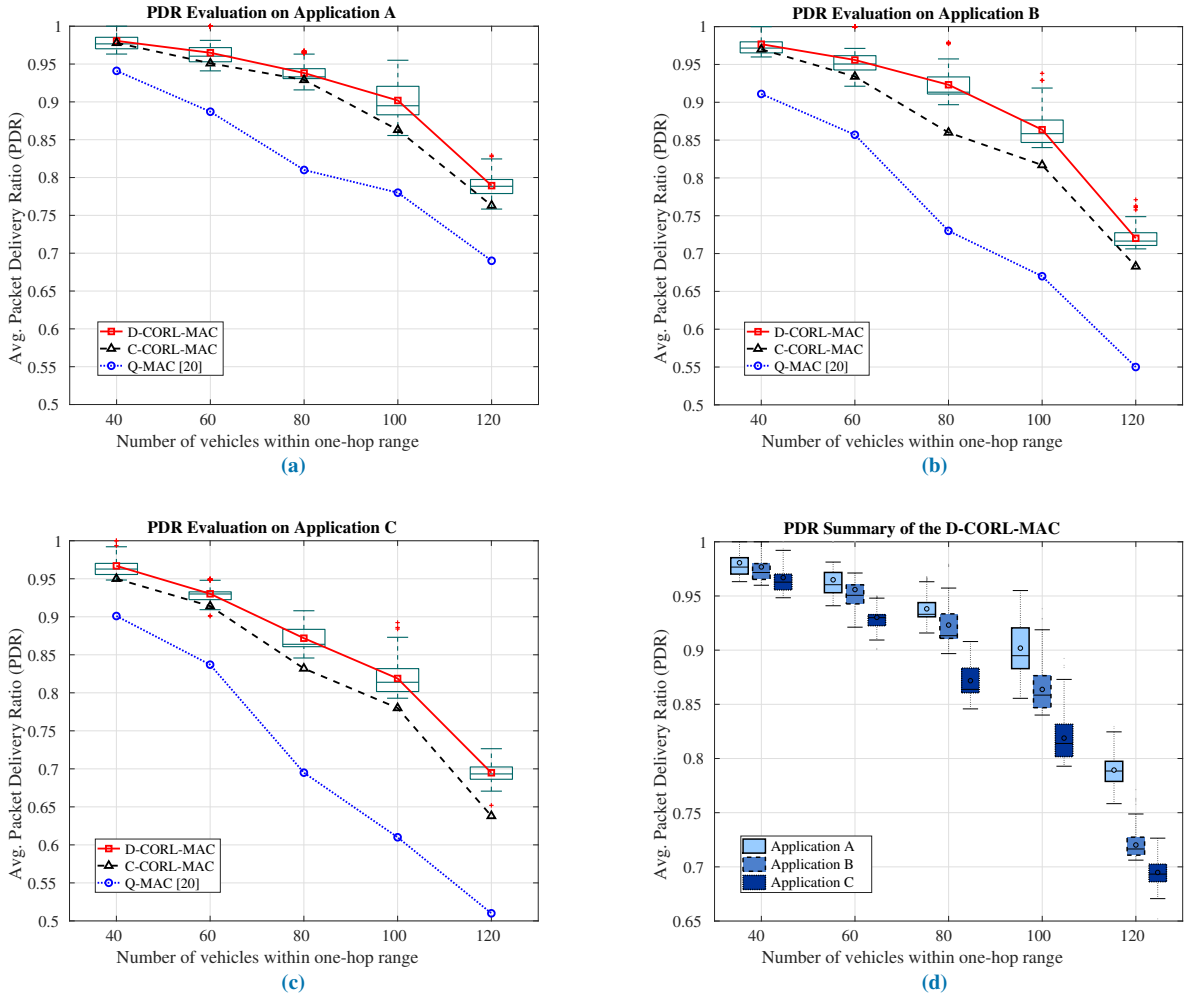


FIGURE 16: Average packet delivery ratio evaluation. (a) Evaluation result for 128-byte Packet, (b) Evaluation result for 256-byte Packet, (c) Evaluation result for 384-byte Packet, (d) Evaluation summary for the proposed D-CORL-MAC.

that the D-CORL-MAC clearly performs better than the C-CORL-MAC when the network load is increased by the larger packet size.

- **Application C (384-byte packets), Fig. 16(c)**: In the case of the largest safety packet size, the decreasing rate of average PDRs becomes larger as the network congestion increases compared to that of application A and B. Even though the network congestion becomes severe from the larger data exchange, the proposed D-CORL-MAC shows the higher robustness than the other methods. In particular, our D-CORL-MAC outperforms the Q-MAC by around 20% since the intermediate network congestion level, 80-VANET. These results establish that the cooperation-based action-policy among vehicles to adapt CW values significantly contributes to improving packet delivery performance in congested VANETs. Moreover, the D-CORL-MAC has the highest average PDRs achieving the 5% performance gain over the C-CORL-MAC. The performance differences between the two reward prediction approaches are dominant especially in the largest packet size configuration. Thus, the distributional reward prediction is more suitable for congested VANETs inherently characterized by the stochastic *backoff* selections of neighboring vehicles.
- **Performance summary of the D-CORL-MAC, Fig. 16(d)**: We observe that the packet size configuration highly affects the performance variation of the proposed D-CORL-MAC. Especially, the performance gap grows wider when network congestion is increased. The algorithm satisfies over 80% PDR in all cases up to 100-VANET. On the other hand, application C shows around 70% PDR in 120-VANET due to the tremendous network traffic load. This bottleneck may be compensated with accommodating the higher transmission data rate. In this study we configure the transmission data rate as 6 Mbps since our primary purpose to evaluate the proposed algorithm in congested VANETs depending on the number of vehicles in the one-hop communication range [47]. Since the higher data rate limits the communication range, the data rate should be adjusted according to communication requirements of various safety applications.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

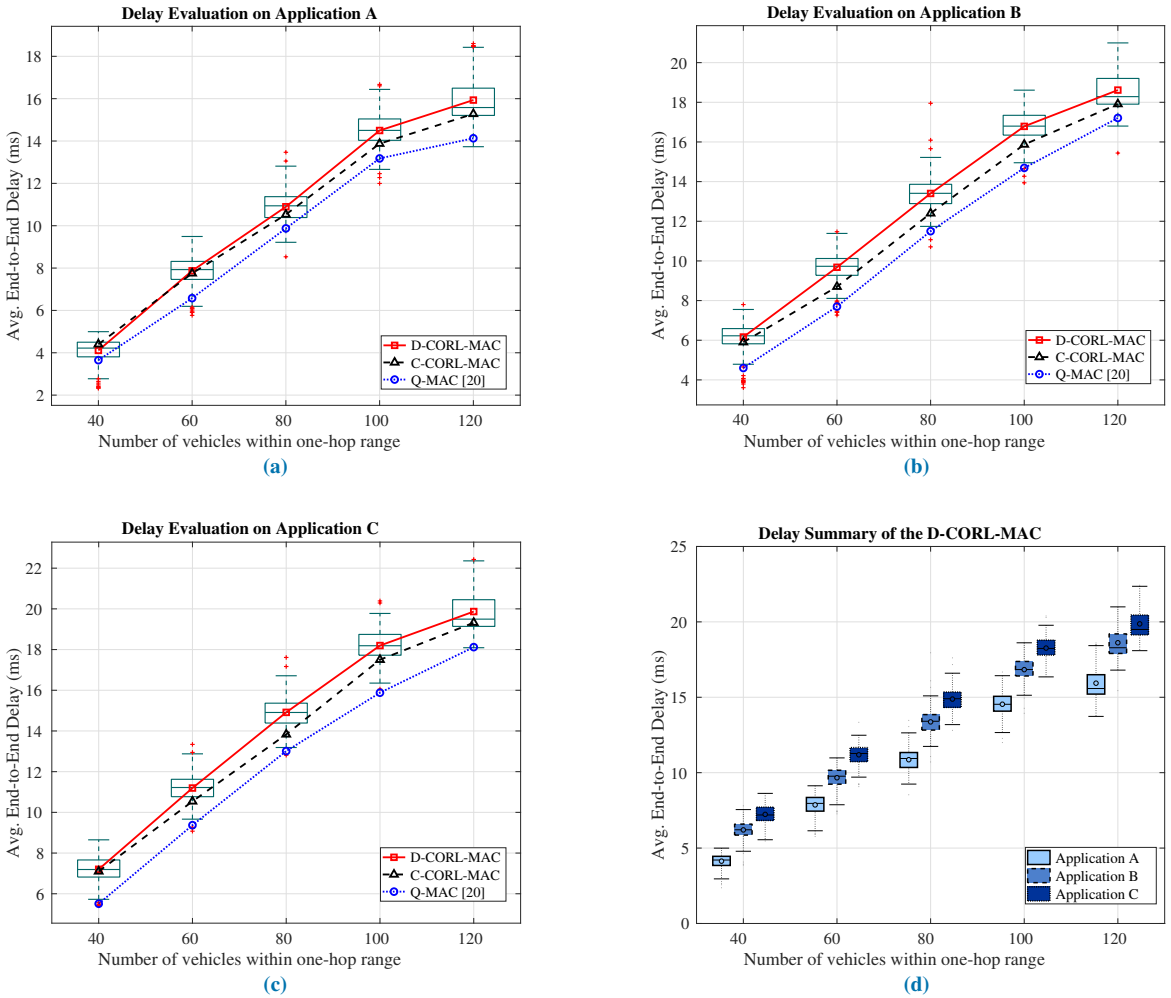


FIGURE 17: Average end-to-end delay evaluation. (a) Evaluation result for 128-byte packet, (b) Evaluation result for 256-byte packet, (c) Evaluation result for 384-byte Packet, (d) Evaluation summary for the proposed D-CORL-MAC.

2) Average End-to-End Delay

Fig. 17 indicates the average end-to-end delay measurements regarding the three packet size configurations. As shown in the figure, the end-to-end delay grows worse with increasing congestion levels because the more vehicles attempt to access the channel for the safety broadcast. During the entire evaluation case, the Q-MAC has the lowest delay since its delay values are measured by fewer successful broadcasted packets due to frequent packet collisions than those of the D-CORL-MAC and the C-CORL-MAC. From the results, the clear trade-off [16] between PDR and latency is clarified since the higher PDR simultaneously leads to a higher end-to-end delay in congested VANETs. Although CW values should be large enough to accommodate the channel access without packet collisions for congested networks, larger CW increases the latency. Thus, an adaptive choice of a proper algorithm according to applications that have different purposes and inherent latency requirements is needed. We also provide numerical result analysis for each application case as follows.

- **Application A (128-byte packets), Fig. 17(a)**: With the smallest packet configuration, the three algorithms satisfy the extremely low delay requirement (≤ 20 ms) [37] for the entire congestion level. Additionally, the difference in delay values between the proposed D-CORL-MAC and the Q-MAC is around $1.5 - 2$ ms. Hence, we conclude that our D-CORL-MAC shows the best performance for the application A configuration satisfying with the highest PDR and the acceptable latency.
- **Application B (256-byte packets), Fig. 17(b)**: The D-CORL-MAC achieves the delay of below 20 ms when the number of vehicles is less than 100. However, during 300 testing episodes, we observe some cases that the delay exceeds 20 ms in the highest traffic environment (120-VANET). On the other hand, the C-CORL-MAC satisfies the extreme latency requirement. Therefore, the D-CORL-MAC should be carefully adopted for delay-sensitive applications with 256-byte packet configuration.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2020.3011568, IEEE Access

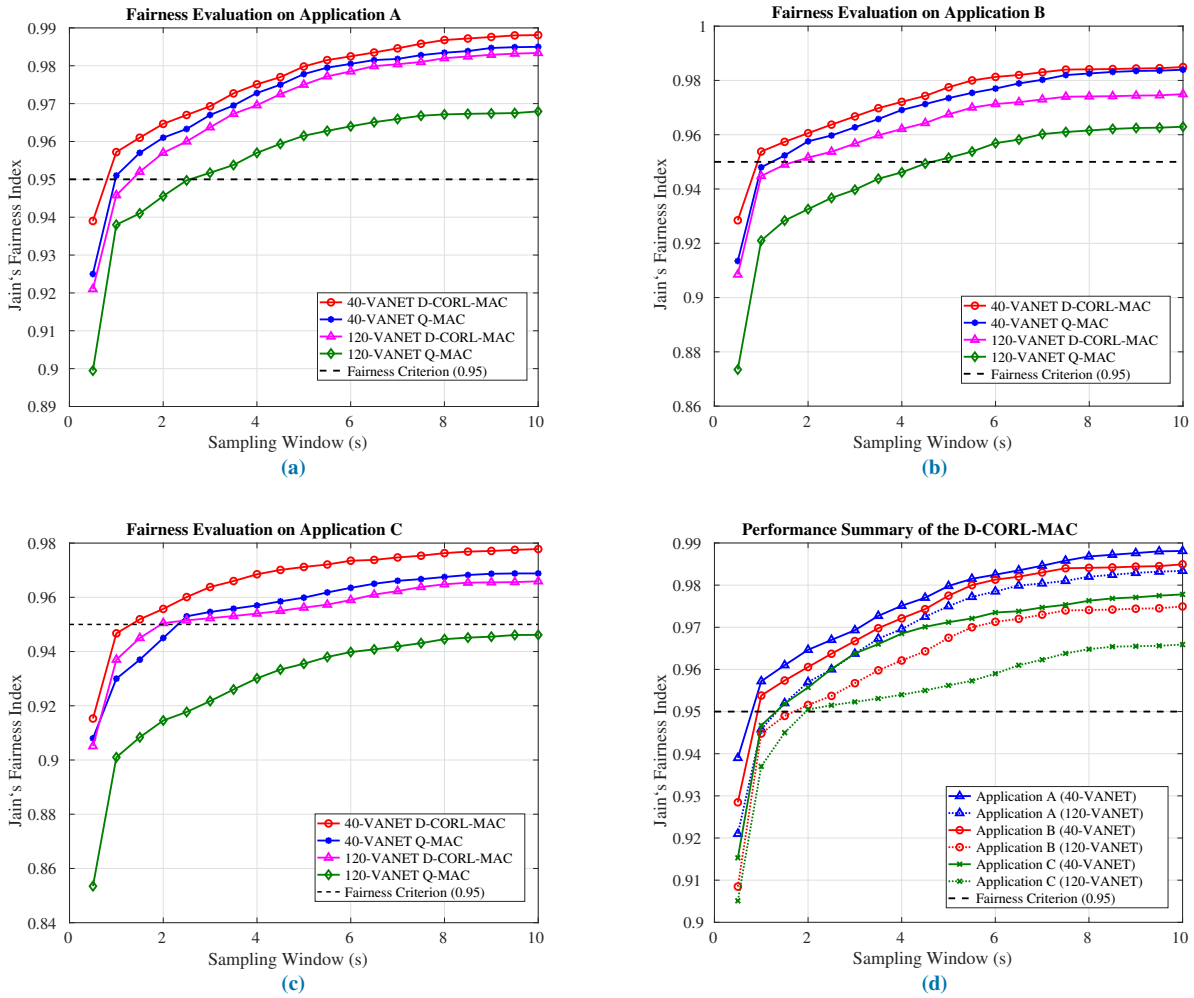


FIGURE 18: Transmission fairness evaluation. (a) Evaluation result for 128-byte packet, (b) Evaluation result for 256-byte packet, (c) Evaluation result for 384-byte packet, (d) Evaluation summary for the proposed D-CORL-MAC.

- **Application C (384-byte packets), Fig. 17(c)**: The delay gap between the D-CORL-MAC and the Q-MAC according to the vehicle populations grows larger than those of application A and B. In addition, we observe the similar performance degradation with the result of application B: our D-CORL-MAC shows over 20 ms average latency in the 120-VANET.
- **Performance summary of the D-CORL-MAC, Fig. 17(d)**: It is clarified that the proposed D-CORL-MAC shows the highest PDR and delay values simultaneously. In spite of the delay growth, the D-CORL-MAC which shows the highest PDR is more suitable scheme in practice since the delay requirement of typical VANET safety applications is under 100 ms. Thus, we conclude that our D-CORL-MAC is the general optimal solution.

3) Transmission Fairness

We evaluate communication fairness based on the achieved PDR values as shown in Fig. 18. The fairness of the proposed D-CORL-MAC and the Q-MAC are compared by adopting the fairness criterion $J_c = 0.95$ for both the short-term (2 second) and long-term (10 second). The fairness is assessed based on two congestion levels: 40-VANET (the lowest congestion level) and 120-VANET (the highest congestion level) for the three applications.

- **Application A (128-byte packets), Fig. 18(a)**: In 40-VANET, our D-CORL-MAC and Q-MAC satisfy the fairness criterion at similar rates for the application A. The D-CORL-MAC and Q-MAC quickly meet the criterion J_c within 1 and 1.5 seconds, respectively, so that both the short-term and long-term fairness can be achieved in the sparser network. Even in the denser network, 120-VANET, the two algorithms satisfy the criterion within 2 second. Moreover, our algorithm shows the highest fair packet delivery among vehicles with around 99% index in respect of the long-term fairness.
- **Application B (256-byte packets), Fig. 18(b)**: In the case of application B, the performance gap between the two algorithms grows. The proposed algorithm achieves the fairness criterion at 1 second and 1.5 seconds at 120-VANET and 40-VANET, respectively. On the other

hand, the Q-MAC is unfair during the short-term against the criterion, it only obtains the long-term fairness for the highly congested network.

- **Application C (384-byte packets), Fig. 18(c)**: Even with the highest packet configuration, our D-CORL-MAC shows robust short-term fairness. The algorithm yields the transmission fairness within a shorter time (2 second) compared to that of the Q-MAC (5.5 second) in 120-VANET. From the perspective of long-term fairness, the D-CORL-MAC provides a highly fair packet exchange among vehicles with around 98% fair index.
- **Performance summary of the D-CORL-MAC, Fig. 18(d)**: The proposed algorithm has a fair index of over 96% overall. From the result, it is confirmed that the proposed fairness-aware strategy of CW selection enhances the transmission fairness as well as the transmission success rate for the congested vehicular networks. Alternation of the upper and lower CW selection shows a significant improvement in communication performance. The proposed action function can efficiently prevent a severe unfair problem that certain vehicles converge to a high or low CW selection policy. Consequently, the proposed D-CORL-MAC achieves prompt fairness enhancement even in 120-VANET with the largest packet configuration (Application C).

VI. CONCLUSION

In this paper, the CORL-MAC algorithm has been proposed to improve the performance of the V2V-based safety packet broadcast in infra-less congested VANETs. We provide the proper cooperation scheme for the CW-adaptation with the distributional reward prediction approach as considering the constraint of dense vehicular networks. Accordingly, we have verified the advanced performance of the proposed algorithm with numerous simulations in the highway traffic scenario. In congested VANETs, the simulation result shows that the algorithm has 13% to 20% improvement on the packet delivery compared to the previous work, Q-MAC [20], which uses the simple Q-Learning method. Our algorithm satisfies the acceptable latency requirement in the congested VANETs with less than 120 vehicles in the one-hop range. Furthermore, the algorithm achieves the short-term and long-term fairness in both the sparser and the denser networks. For future work, we expect that a state estimation method such as the Kalman Filter for a VANET status prediction will induce further performance improvement. Also, the development of a robust channel access protocol under Non-Line of Sight (NLOS) environments such as intersections or urban areas remains an interesting challenge.

REFERENCES

[1] J. B. Kenney, "Dedicated short-range communications (DSRC) standards in the united states," Proceedings of the IEEE, vol. 99, no. 7, pp. 1162–1182, July 2011.

[2] X. Ma, X. Chen, and H. H. Refai, "Performance and reliability of DSRC vehicular safety communication: A formal analysis," EURASIP Journal on Wireless Communications and Networking, vol. 2009, no. 1, p. 969164, Jan 2009. [Online]. Available: https://doi.org/10.1155/2009/969164

[3] K. A. Hafeez, L. Zhao, B. Ma, and J. W. Mark, "Performance analysis and enhancement of the DSRC for VANET's safety applications," IEEE Transactions on Vehicular Technology, vol. 62, no. 7, pp. 3069–3083, Sep. 2013.

[4] A. P. Jardosh, K. N. Ramachandran, K. C. Almeroth, and E. M. Belding-Royer, "Understanding congestion in IEEE 802.11 b wireless networks," in Proceedings of the 5th ACM SIGCOMM conference on Internet Measurement. USENIX Association, 2005, pp. 25–25.

[5] V. Nguyen, C. Pham, T. Z. Oo, N. H. Tran, E.-N. Huh, and C. S. Hong, "MAC protocols with dynamic interval schemes for VANETs," Vehicular Communications, vol. 15, pp. 40–62, 2019.

[6] V. Nguyen, O. T. T. Kim, C. Pham, T. Z. Oo, N. H. Tran, C. S. Hong, and E.-N. Huh, "A survey on adaptive multi-channel MAC protocols in VANET using markov models," IEEE Access, vol. 6, pp. 16 493–16 514, 2018.

[7] Z. Li, Z. Ding, Y. Wang, and Y. Fu, "Time synchronization method among VANET devices," in 2017 International Conference on Wireless Communications, Signal Processing and Networking (WiSPNET), March 2017, pp. 2096–2101.

[8] M. Hadded, P. Muhlethaler, A. Laouiti, R. Zagrouba, and L. A. Saidane, "TDMA-based MAC protocols for vehicular ad hoc networks: A survey, qualitative analysis, and open research issues," IEEE Communications Surveys Tutorials, vol. 17, no. 4, pp. 2461–2492, Fourthquarter 2015.

[9] Z. Gao, D. Chen, N. Yao, Z. Lu, and B. Chen, "A novel problem model and solution scheme for roadside unit deployment problem in VANETs," Wireless Personal Communications, vol. 98, no. 1, pp. 651–663, 2018.

[10] D. Wu, Y. Ling, H. Zhu, and J. Liang, "The RSU access problem based on evolutionary game theory for VANET," International Journal of Distributed Sensor Networks, vol. 9, no. 7, p. 143024, 2013.

[11] J. H. Kim and J. K. Lee, "Performance of carrier sense multiple access with collision avoidance protocols in wireless LANs," Wireless Personal Communications, vol. 11, no. 2, pp. 161–183, 1999.

[12] M. Azizian, S. Cherkaoui, and A. S. Hafid, "A distributed cluster based transmission scheduling in VANET," in 2016 IEEE International Conference on Communications (ICC), May 2016, pp. 1–6.

[13] V. Nguyen, T. Z. Oo, N. H. Tran, and C. S. Hong, "An efficient and fast broadcast frame adjustment algorithm in VANET," IEEE Communications Letters, vol. 21, no. 7, pp. 1589–1592, July 2017.

[14] G. V. Rossi and K. K. Leung, "Optimised CSMA/CA protocol for safety messages in vehicular ad-hoc networks," in 2017 IEEE Symposium on Computers and Communications (ISCC), July 2017, pp. 689–696.

[15] X. Huang, A. Liu, and X. Liang, "An analytical model of CSMA/CA performance for periodic broadcast scheme," in 2018 IIEEE/CIC International Conference on Communications in China (ICCC), Aug 2018, pp. 475–479.

[16] X. Liu and A. Jaekel, "Congestion control in V2V safety communication: Problem, analysis, approaches," Electronics, vol. 8, no. 5, 2019. [Online]. Available: https://www.mdpi.com/2079-9292/8/5/540

[17] R. S. Sutton and A. G. Barto, "Reinforcement learning: An introduction," 2011.

[18] C. Wu, S. Ohzahata, Y. Ji, and 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.

[19] A. Pressas, Z. Sheng, F. Ali, D. Tian, and 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.

[20] A. Pressas, Z. Sheng, F. Ali, and D. Tian, "A q-learning approach with collective contention estimation for bandwidth-efficient and fair access control in IEEE 802.11 p vehicular networks," IEEE Transactions on Vehicular Technology, vol. 68, no. 9, pp. 9136–9150, 2019.

[21] S. Seuken and S. Zilberstein, "Formal models and algorithms for decentralized decision making under uncertainty," Autonomous Agents and Multi-Agent Systems, vol. 17, no. 2, pp. 190–250, 2008.

[22] Y.-C. Tseng, S.-Y. Ni, Y.-S. Chen, and J.-P. Sheu, "The broadcast storm problem in a mobile ad hoc network," Wireless networks, vol. 8, no. 2-3, pp. 153–167, 2002.

[23] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, "Fully decentralized multi-agent reinforcement learning with networked agents," arXiv preprint arXiv:1802.08757, 2018.

[24] K. Zhang, Z. Yang, and T. Basar, "Networked multi-agent reinforcement learning in continuous spaces," in 2018 IEEE Conference on Decision and Control (CDC). IEEE, 2018, pp. 2771–2776.
[25] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," in Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017, pp. 449–458.
[26] 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, vol. 518, no. 7540, pp. 529–533, 2015.
[27] V. Nguyen, T. T. Khanh, T. Z. Oo, N. H. Tran, E.-N. Huh, and C. S. Hong, "A cooperative and reliable RSU-assisted IEEE 802.11 p-based multi-channel mac protocol for VANETs," IEEE Access, vol. 7, pp. 107 576–107 590, 2019.
[28] Y. Ma, L. Yang, P. Fan, S. Fang, and Y. Hu, "An improved coordinated multichannel MAC scheme by efficient use of idle service channels for VANETs," in 2018 IEEE 87th Vehicular Technology Conference (VTC Spring). IEEE, 2018, pp. 1–5.
[29] Y. Zhang, K. Liu, S. Liu, J. Zhang, T. Zhang, Z. Xu, and F. Liu, "A clustering-based collision-free multichannel MAC protocol for vehicular ad hoc networks," in 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall). IEEE, 2018, pp. 1–7.
[30] X. Zhang, X. Jiang, and M. Zhang, "A black-burst based time slot acquisition scheme for the hybrid TDMA/CSMA multichannel MAC in VANETs," IEEE Wireless Communications Letters, vol. 8, no. 1, pp. 137–140, 2018.
[31] Q. Chen, D. Jiang, and L. Delgrossi, "IEEE 1609.4 DSRC multi-channel operations and its implications on vehicle safety communications," in 2009 IEEE vehicular networking conference (VNC). IEEE, 2009, pp. 1–8.
[32] T. ElBatt, S. K. Goel, G. Holland, H. Krishnan, and J. Parikh, "Cooperative collision warning using dedicated short range wireless communications," in Proceedings of the 3rd international workshop on Vehicular ad hoc networks, 2006, pp. 1–9.
[33] L. Wang, R. F. Iida, and A. M. Wyglinski, "Coordinated lane changing using V2V communications," in 2018 IEEE 88th Vehicular Technology Conference (VTC-Fall). IEEE, 2018, pp. 1–5.
[34] M. Li, K. Zeng, and W. Lou, "Opportunistic broadcast of event-driven warning messages in vehicular ad hoc networks with lossy links," Computer Networks, vol. 55, no. 10, pp. 2443–2464, 2011.
[35] M. Sybis, V. Vukadinovic, M. Rodziewicz, P. Sroka, A. Langowski, K. Lenarska, and K. Wesołowski, "Communication aspects of a modified cooperative adaptive cruise control algorithm," IEEE Transactions on Intelligent Transportation Systems, 2019.
[36] O. Shagdar, F. Nashashibi, and S. Tohme, "Performance study of CAM over IEEE 802.11 p for cooperative adaptive cruise control," in 2017 Wireless Days. IEEE, 2017, pp. 70–76.
[37] Z. Xu, X. Li, X. Zhao, M. H. Zhang, and Z. Wang, "DSRC versus 4G-LTE for connected vehicle applications: A study on field experiments of vehicular communication performance," Journal of Advanced Transportation, vol. 2017, 2017.
[38] J. Goodman, A. G. Greenberg, N. Madras, and P. March, "Stability of binary exponential backoff," Journal of the ACM (JACM), vol. 35, no. 3, pp. 579–602, 1988.
[39] G. E. Monahan, "State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms," Management Science, vol. 28, no. 1, pp. 1–16, 1982.
[40] S. Eichler, "Performance evaluation of the IEEE 802.11 p WAVE communication standard," in 2007 IEEE 66th Vehicular Technology Conference. IEEE, 2007, pp. 2199–2203.
[41] N. Shahin, R. Ali, S. W. Kim, and Y.-T. Kim, "Adaptively scaled back-off (ASB) mechanism for enhanced performance of CSMA/CA in IEEE 802.11 ax high efficiency wlan," in NOMS 2018-2018 IEEE/IFIP Network Operations and Management Symposium. IEEE, 2018, pp. 1–5.
[42] R. Stanica, E. Chaput, and A.-L. Beylot, "Local density estimation for contention window adaptation in vehicular networks," in 2011 IEEE 22nd International Symposium on Personal, Indoor and Mobile Radio Communications. IEEE, 2011, pp. 730–734.
[43] P. Gawłowicz and A. Zubow, "ns-3 meets openai gym: The playground for machine learning in networking research," in Proceedings of the 22nd International ACM Conference on Modeling, Analysis and Simulation of Wireless and Mobile Systems. ACM, 2019, pp. 113–120.
[44] M. van Eenennaam, A. van de Venis, and G. Karagiannis, "Impact of IEEE 1609.4 channel switching on the IEEE 802.11 p beaconing performance," in 2012 IFIP Wireless Days. IEEE, 2012, pp. 1–8.
[45] S. Krauß, P. Wagner, and C. Gawron, "Metastable states in a microscopic model of traffic flow," Physical Review E, vol. 55, no. 5, p. 5597, 1997.
[46] X. Ma, J. Zhang, and T. Wu, "Reliability analysis of one-hop safety-critical broadcast services in vanets," IEEE Transactions on Vehicular Technology, vol. 60, no. 8, pp. 3933–3946, 2011.
[47] Y.-S. Song and H.-K. Choi, "Analysis of V2V broadcast performance limit for wave communication systems using two-ray path loss model," ETRI Journal, vol. 39, no. 2, pp. 213–221, 2017.
[48] T. Abbas, K. Sjöberg, J. Karedal, and F. Tufvesson, "A measurement based shadow fading model for vehicle-to-vehicle network simulations," International Journal of Antennas and Propagation, vol. 2015, 2015.
[49] R. K. Jain, D.-M. W. Chiu, and W. R. Hawe, "A quantitative measure of fairness and discrimination," Eastern Research Laboratory, Digital Equipment Corporation, Hudson, MA, 1984.

CHUNGJAE CHOE received the B.S. degree in Information Communication Engineering from Dongguk University, Korea, in 2018. He also received the M.S. degree in Korea Advanced Institute of Science and Technology (KAIST), Korea. His research interests include autonomous vehicle technology, vehicular communications, and multi-agent reinforcement learning.

JANGYONG AHN received the M.S. degree in the CCS Graduate School for Green Transportation from the Korea Advanced Institute of Science and Technology (KAIST), South Korea, in 2018. He is currently working toward the Ph.D. degree at KAIST. His research interests include vehicular communication, multi-agent reinforcement learning, power integrity in electric vehicle and electromagnetic interference/electromagnetic compatibility (EMI/EMC).

JUNSUNG CHOI received B.S., M.S., and Ph.D. degrees as Electrical and Computer Programming Engineering from Virginia Tech, Blacksburg, VA, USA, in 2013, 2016, and 2018, respectively. From 2013 to 2018, he joined Wireless@VT as a research member. He is currently a Post-Doctoral researcher with KAIST, Daejeon, South Korea. His research interests include vehicular communications, reinforcement learning, propagation channel, and 5G.

DONGRYUL PARK He is currently pursuing the M.S. degree at the CCS Graduate School of Green Transportation in the Korea Advanced Institute of Science and Technology (KAIST), Korea. His research interests include vehicular communications, reinforcement learning, and deep learning.

MINJUN KIM received the M.S. degree from the CCS Graduate School of Green Transportation, Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 2019. Since 2019,he has been with Korea Aerospace Research Institute (KARI), Daejeon, South Korea. His research interests include vehicular communications, satelite communication system, LiDAR, RADAR, multi-agent reinforcement learning, and deep learning.

SEONGYOUNG AHN (S'00-M'06-SM'16) received the B.S., M.S., and Ph.D. degrees in KAIST, Korea, in 1998, 2000, and 2005, respectively. He is currently an Associate Professor in KAIST. His research interests include vehicular communications, reinforcement learning, WPT system design, and electromagnetic compatibility design for electric vehicle and digital systems.