Energy-Efficient Cooperative Secure Communications in MmWave Vehicular Networks Using Deep Recurrent Reinforcement Learning

Ying Ju, *Member, IEEE*, Zipeng Gao, Haoyu Wang, Lei Liu, *Member, IEEE*, Qingqi Pei, *Senior Member, IEEE*, Mianxiong Dong, *Member, IEEE*, Shahid Mumtaz, *Member, IEEE*, and Victor C. M. Leung, *Life Fellow, IEEE*

***Abstract*—Millimeter wave (mmWave) with abundant spectrum resources can realize high-rate communications in vehicular networks. However, the mobility of vehicles and the blocking effect of mmWave propagation bring new challenges to communication security. Cooperative communication is envisioned as a promising physical layer security (PLS) approach to enhance the secrecy performance, but it will induce extra energy consumption of vehicles. This paper proposes a deep recurrent reinforcement learning (DRRL)-based energy-efficient cooperative secure transmission scheme in mmWave vehicular networks, where eavesdropping vehicles attempt to intercept the multi-user downlink communications. We jointly design the mmWave beam allocation, the cooperative nodes selection, and the transmit power of vehicles. Specifically, the mmWave base station selects idle vehicles as relays to overcome the severe blocking attenuation of legitimate transmissions and controls the transmit power to reduce energy consumption. Moreover, to ensure secure transmission, a cooperative vehicle is selected to transmit jamming signals to the eavesdropping vehicles while the legitimate users are not disturbed. We conduct comprehensive interference analysis for both direct transmission and relay-aided transmission, and derive the theoretical expressions for the secrecy capacity. We then design the Dueling Double Deep Recurrent Q-Network (D3RQN) learning algorithm to maximize the total secrecy capacity subject to the energy consumption constraint. We set the energy consumption punishment mechanism to avoid relay vehicles consuming too much power for forwarding signals. We demonstrate that the proposed scheme can rapidly adapt to the highly dynamic vehicular networks and effectively improve secrecy performance while reducing the energy consumption of vehicles.**

***Index Terms*—MmWave vehicular communication, energy consumption, cooperative secure transmission, physical layer security, deep recurrent reinforcement learning.**

The work of Ying Ju, Zipeng Gao, Haoyu Wang, Lei Liu, and Qingqi Pei was supported in part by the National Natural Science Foundation of China under Grants 62102301 and 62132013, and in part by the Innovation Capability Support Program of Shaanxi under Grant 2024RS-CXTD-01. The work of Victor C.M. Leung was supported in part by the Guangdong "Pearl River Talent Recruitment Program" under Grant 2019ZT08X603, and in part by the Guangdong "Pearl River Talent Plan" under Grant 2019JC01X235. The work of Shahid Mumtaz was supported in part by the 6G-SENSES project from the Smart Networks and Services Joint Undertaking (SNS JU) under the European Union's Horizon Europe research and innovation programme under Grant 101139282. (*Corresponding authors: Lei Liu, Ying Ju.*)

Ying Ju, Zipeng Gao, Haoyu Wang, Lei Liu, and Qingqi Pei are with the State Key Laboratory of Integrated Services Networks, School of Telecommunications Engineering, Xidian University, Xi'an 710071, China (e-mail: juying@xidian.edu.com);

Mianxiong Dong is with the Department of Information and Electric Engineering, Muroran Institute of Technology, Muroran, Japan (e-mail: mx.dong@csse.muroran-it.ac.jp);

Shahid Mumtaz is with the Department of Applied Informatics, Silesian University of Technology, Akademicka 16 44-100 Gliwice, Poland, and the Department of Computer Sciences, Nottingham Trent University, Nottingham NG1 4FQ, United Kingdom (e-mail: dr.shahid.mumtaz@ieee.org);

Victor C.M. Leung is with the Artificial Intelligence Research Institute, Shenzhen MSU-BIT University, Shenzhen 518172, China, and the College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China, and the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver V6T 1Z4, Canada (e-mail: vleung@ieee.org).

I. Introduction

Recently, the increasing demand for high-speed and low-latency wireless transmission in vehicular networks has stimulated the development of multi-Gbps links to ensure the quality of service [1]. Millimeter wave (mmWave) systems with rich spectrum resources have played a crucial role in meeting this demand and pushing vehicular communication to a new stage [2]–[5]. Although mmWave communication has high capacity and narrow beams, due to the openness of wireless channels, mmWave communication still faces serious security vulnerabilities. With the emergence of many vehicular applications, such as autonomous driving and sensor fusion, information security, which may endanger human life, has become a key issue in vehicular networks.

Considering the delay-sensitive nature of vehicle communication, the low-cost physical layer security (PLS) technology effectively improves secrecy performance by utilizing the characteristics of wireless channels [6]–[8]. Leveraging the PLS method, mmWave base stations or vehicles can securely transmit mmWave signals using beamforming or wiretap coding schemes [9]–[11]. In addition, cooperative communication is considered as a promising method to improve PLS performance, where idle vehicles are used to relay legitimate signals or send interfering signals to deceive eavesdroppers [12], [13]. On the one hand, relay transmission increases the strength of the transmitted signal by overcoming the high path loss of mmWave propagation to the target vehicles. It can also enable the signal to get around the obstacles or wiretap vehicles, thereby mitigating the blocking effect of legitimate transmission and increasing the secrecy capacity of the system. On the other hand, the base station can arrange an appropriate cooperative vehicle to transmit a jamming signal to the eavesdropper, reducing the capacity of the wiretap channel significantly.

Therefore, cooperative secure communications in vehicular networks have attracted new research interest [14]–[17]. The authors in [14] analyze the performance of full-duplex

amplify and forward (AF) relay in vehicle-to-vehicle (V2V) communications. The authors in [15] investigate the secrecy performance of relay-aided vehicle-to-everything (V2X) communications. The impact of relay positions and channel conditions is analyzed. In [16], the authors propose a relay selection and cooperative jamming strategy, which outperforms the conventional schemes under the same total power constraint. In [17], the unmanned aerial vehicle is exploited as a relay to improve the security of the satellite-to-vehicle link and simultaneously serves as a jammer by generating artificial noise to interfere with the eavesdropper.

The demand for low-carbon and green transportation systems has recently become increasingly significant. Nevertheless, the cooperative secure communications involve many helper vehicles supporting the PLS scheme, which increases the energy consumption of the cooperative vehicles. Therefore, the energy efficiency of the cooperative secure transmission becomes a key concern in vehicular networks. In practical communication, vehicles are unwilling to overuse high-power transmission signals for energy-saving reasons. To reduce energy consumption, using vehicles for cooperative transmission needs to consider transmit power control. Many previous works have considered power control of relay transmission [18]–[21]. The authors in [18] investigate relay selection for heterogeneous transmit powers in vehicular ad-hoc networks (VANETs). The authors in [19] formulate a problem for the maximization of the achievable rate, where the unmanned aerial vehicle (UAV) position, analog beamforming, and power control are jointly optimized. The authors in [20] consider relay power control in device-to-device (D2D)-enabled vehicular communications.

Although the effectiveness of the PLS technique has been demonstrated in vehicular networks, most of the literature focuses on traditional microwave communications. Few secure transmission schemes for the mmWave vehicular networks are currently under development. Coupling mmWave and vehicular transmissions brings new challenges to security [22]. In [23], the authors propose two PLS schemes for mmWave vehicular networks, namely antenna subset modulation with a single radio frequency (RF) chain and artificial noise injection with multiple RF chains. In [24], the authors propose a blockage-and-power-based jammer selection strategy to address potential security pitfalls in a mmWave cellular V2X network. Preliminary analysis of the association probability based on random geometric methods is conducted, leading to the derivation of theoretical expressions for the secrecy outage probability and secrecy throughput. The authors in [25] investigate the mmWave PLS of a cellular Internet of Vehicles network composed of many base stations and V2X nodes. Two uplink association schemes are proposed, and their secrecy performance is analyzed. Both schemes increase the secrecy rate of the system. However, the above works do not investigate the energy consumption of the network. Besides, they do not consider the mobility of the vehicles. In practical vehicular communications, the random blockage of mmWave propagation and the high mobility of the vehicles result in fast-changing channels, which requires the base station to make quick decisions on the secure transmission strategy.

Researchers have therefore turned their attention to deep reinforcement learning (DRL), which uses target-related reward functions to optimize goals by interacting with the environment [26], [27]. DRL has found its applications in the domain of vehicular communications. The authors in [28] explore optimal collision avoidance algorithm using DRL and propose a safety evaluation map (SEM) to describe the evaluation results. In [29], resource allocation is viewed as a non-cooperative game with each D2D pair learning strategies from local information, using the Double Deep Q-Network (DDQN) algorithm. Simultaneously, the Dueling Double Deep Recurrent Q-Network (D3RQN) multi-agent algorithm, a combination of D3QN and Long Short-Term Memory (LSTM) networks, has been investigated [30], [31]. In [30], the authors propose a channel state information (CSI)-independent decentralized algorithm to optimize the throughput of the vehicle-to-infrastructure (V2I) link while ensuring the latency and reliability of the V2V link. In [31], the authors consider task scheduling in serverless edge computing networks, modeling the process as a partially observable stochastic game (POSG), where nodes schedule tasks and allocate resources based on local observations.

Despite the lack of considering the security issue, the following works demonstrate the capability of DRL in designing the beam management strategy and cooperative communication schemes. The authors in [32] use clustering and DRL algorithm for resource block allocation and beam management, which performs well in latency and reliability. In [33], the authors propose a DRL-based multi-hop mmWave communications using reconfigurable intelligent surfaces (RIS) as relay nodes. This aims to overcome the severe propagation attenuation and improve the coverage range. The authors in [34] propose a hierarchical reinforcement learning algorithm based on the DRL algorithm to study the problem of minimizing the outage probability in a two-hop cooperative relay network. In [35], the authors jointly select the relay and optimize the reflection coefficient of the cooperative RIS based on DRL to improve communication quality. In [36], the authors propose a DRL-based beam allocation scheme in relay-aided multi-user mmWave vehicular networks to maximize the total system capacity.

Overall, the significance of energy consumption and the importance of cooperative communications in safeguarding vehicular networks has been illustrated in the existing literature. However, as far as we know, no previous work has studied the energy-efficient cooperative secure communications for mmWave vehicular communications. It is necessary to study this issue from the perspective of efficient decision-making. In this paper, by using idle vehicles as relay and jamming nodes, we aim to resist the blocking effect of mmWave, suppress the channel capacity of eavesdropping vehicle, reduce the energy consumption of vehicles and enhance secrecy performance of target vehicles. Our main contributions are summarized as follows.

1) We propose a deep recurrent reinforcement learning (DRRL)-based energy-efficient cooperative secure communication scheme for multi-user mmWave vehicular networks. In this scheme, blocked legitimate links are

assisted by relay vehicles, and the eavesdropping link is deteriorated by jamming vehicles. We derive the theoretical expressions of the secrecy capacity for direct and relay-aided transmission through comprehensive interference analysis.

2) We establish the joint optimization problem of beam allocation, relay and jammer selection, and the transmit power design to maximize the system secrecy capacity with the energy consumption constraint. Considering the high dynamic characteristics of the vehicle network, we design the D3RQN algorithm for decision-making. We incorporate one-dimensional convolution and LSTM network into the D3QN architecture to extract features for large-dimensional states and memorize the temporal information between vehicles, respectively.
3) We utilize the approximate regretted reward (ARR) to resist the reward fluctuation in the mobile vehicular scenario, and design the energy consumption punishment mechanism to avoid relay vehicles consuming too much power for forwarding signals. We demonstrate that with the proposed network structure, the agent can formulate intelligent policies that improve the secrecy performance while reducing vehicle energy consumption.

II. System Model

A. MmWave Communication Model

The cooperative secure communication scenario in mmWave vehicular networks investigated in this paper is shown in Fig. 1. It is assumed that the vehicles are running on a bidirectional multi-lane road, and they are divided into four categories: target vehicles (green vehicles), relay vehicles (red vehicles), jammer vehicles (blue vehicles), and eavesdropping vehicles (black vehicles). In each service cycle, the base station simultaneously transmits a set of orthogonal beams $B = \{b_m, m = 1, 2, \cdots, N_B\}$ to N_T users, where N_B is the number of orthogonal beams. To avoid co-beam interference, each beam should only serve one vehicle. Since the vehicles are moving on the road, we need to select beams from B to serve the vehicles dynamically and acquire high secrecy performance. Due to the characteristics of the mmWave hardware, the number of mmWave RF chains is limited. The number of the selected beams N'_B should be at most the number of RF chains N_{RF}, i.e., $N'_B \leq N_{RF}$. The set of the selected beams is denoted by $B' = \{b_{m'}, m' = 1, 2, \cdots, N'_B\}$. There are obstacles with random positions on the road, which are used to simulate the blocking effect of buildings and green plants on communications in a practical environment.

In our communication scenario, a potential eavesdropping vehicle intercepts one of the target vehicles in each service period. We assume that the location of the eavesdropping vehicle is available at the base station. The base station selects the transmission beam for each vehicle in turn according to the information of N_T target vehicles. If the direct transmission link is blocked or the main-lobe beam has eavesdropped, the base station will choose a friendly vehicle near the target vehicle as the relay to forward signals. If the vehicle is chosen as a relay node to forward signals, it can select transmit power from N_P transmit power levels. The base station needs to make an appropriate power selection decision based on global information to reduce energy consumption while ensuring the total secrecy capacity. This strategy aims to reduce the power consumption of vehicles and thus promote the development of low-carbon and green transportation systems. In addition, if the channel for direct transmission is good enough, the signal is transmitted directly from the base station. We focus on the power control for vehicles in this paper, and the transmit power of the base station is fixed. At the same time, the base station chooses a cooperative vehicle to transmit jamming signals against the eavesdropping vehicle and deteriorate the signal reception of the wiretap channel.

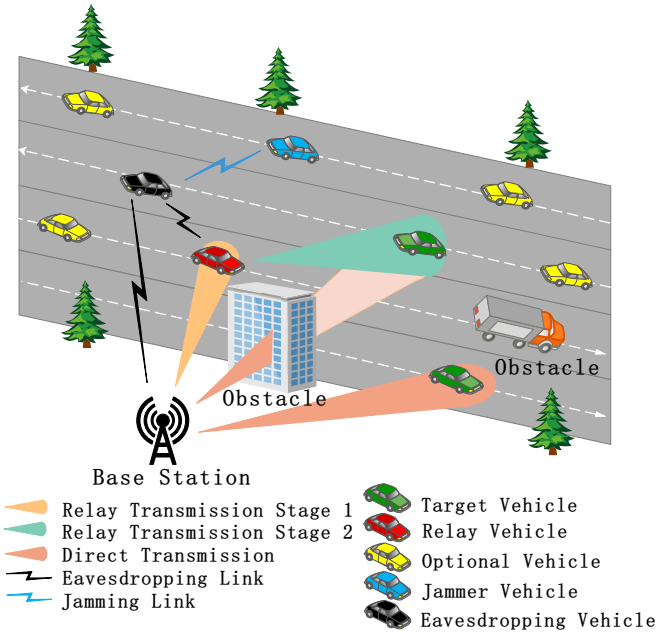


Fig. 1: Network model of the cooperative secure mmWave vehicular communication.

In this paper, in order to simulate the real dynamic traffic pattern, we assume that the vehicle arrival obeys Poisson distribution [37], and the time interval of vehicle arrival Δt obeys the negative exponential distribution. The probability density function can be described as

$$f(\Delta t) = \begin{cases} \lambda e^{-\lambda \Delta t}, & \text{if } \Delta t \geq 0, \\ 0, & \text{Otherwise}, \end{cases} \tag{1}$$

where λ denotes the average arrival rate of vehicles, and the road conditions with different traffic patterns can be simulated by adjusting λ.

B. Channel Model

This subsection presents the channel model and the corresponding parameters adopted in this paper, including antenna gain, blocking factor, path loss and channel gain of the base station and vehicles.

1) Antenna gain: The base station uses a designed orthogonal beam set to provide downlink data services for multiple users in the communication cycle, and the different beams

TABLE I: Notations and Explanations

Notation	Explanation
N_T	Number of target vehicles
N_R	Number of relay vehicles
N_J	Number of potential jammer vehicles
N_B	Number of orthogonal beams
N_P	Number of transmit power levels
λ	Vehicle arrival rate
G_B	Antenna gain of base station
G_V	Antenna gain of vehicle
f_c	Carrier frequency
α_k	Obstacle factor in downlink
$\alpha_{t,k}$	Temporary obstacle factor
$\alpha_{p,k}$	Permanent obstacle factor
$L_{t,r}$	Path loss
$g_{t,r}$	Channel gain
η	Energy consumption of relay vehicles
C_r	Secrecy capacity of relay vehicles
E_r	Total energy consumption for relay transmission
C_k	Channel capacity of the kth target vehicle
$C_{e,k}$	Channel capacity of the eavesdropping vehicle
$C_{s,k}$	Secrecy capacity of the kth target vehicle
ϑ	Beam allocation indicator
ρ	Relay selection indicator
φ	Jammer selection indicator
ι	Transmit power control indicator

are spaced to cover the entire communication range of the base station. We leverage the widely used sector-based antenna model to approximate the mmWave transmission beam [38]. Then the antenna gain of the base station can be expressed as

$$G_B(\theta) = \begin{cases} M_B, & |\theta| < \frac{\theta_B}{2}, \\ m_B, & \text{Otherwise}, \end{cases} \tag{2}$$

where θ_B is the beam width of the main lobe, M_B and m_B are the main-lobe and side-lobe gain of the base station, respectively. Similarly, the antenna gain of the vehicles can be expressed as

$$G_V(\theta) = \begin{cases} M_V, & |\theta| < \frac{\theta_V}{2}, \\ m_V, & \text{Otherwise}, \end{cases} \tag{3}$$

where θ_V is the beam width of the main-lobe, M_V and m_V are main-lobe and side-lobe gain of vehicles, respectively.

2) Obstacle factor: This paper classifies the obstacles on the road into two kinds, namely permanent obstacles and temporary obstacles. The former is used to describe obstacles on the road that cannot be moved and have always existed, such as buildings, plants, etc. The latter is used to describe the obstacles that temporarily exist during transmission, such as other vehicles on the road. Then the total obstacle factor from the base station to the kth vehicle can be expressed as

$$\alpha_k = \alpha_{p,k}\alpha_{t,k}, \tag{4}$$

where $\alpha_{p,k}$ is the permanent obstacle factor, and $\alpha_{t,k}$ is the temporary obstacle factor.

3) Channel gain: The path loss $L_{t,r}$ from the transmitter to the receiver is given by

$$L_{t,r} = \mu_1 \log(f_c) + \mu_2 \log(d_e) + \mu_3, \tag{5}$$

where f_c is the carrier frequency of the signal, and d_e is the Euclidean distance from the transmitter to the receiver. μ_1, μ_2, and μ_3 are the parameters chosen according to the communication scenario. Then the channel gain from the transmitter to the receiver can be given by

$$g_{t,r} = \alpha_k L_{t,r}. \tag{6}$$

As a result, the channel gains of several transmission links used in this paper can be obtained, which are the channel gain from the base station to the ith relay vehicle $g_{B,i}$, the channel gain from the base station to the kth target vehicle $g_{B,k}$, the channel gain from the base station to the eavesdropping vehicle $g_{B,e}$, the channel gain from the ith relay vehicle to the kth target vehicle $g_{i,k}$, the channel gain from the ith relay vehicle to the eavesdropping vehicle $g_{i,e}$, and the channel gain from the jammer vehicle to the eavesdropping vehicle $g_{j,e}$.

III. Transmission Strategy

In the transmission strategy of this paper, the base station globally selects the beam to serve each target vehicle based on the channel information of N_T target users. There are $N_{T,d}$ beams selected for direct transmission (from the base station to the target vehicle), and $N_{T,r}$ beams selected for relay transmission (from the base station to the relay vehicle), with $N_{T,d} \leq N_T$ and $N_{T,r} \leq N_T$. In order to reduce the hardware overhead and computational complexity of the vehicle, the relay vehicle adopts the AF relay method to forward the signals, and $\beta_{i,k}$ is the amplification gain of the ith relay vehicle to the kth target vehicle. The whole transmission cycle is divided into two stages. The details of the two stages can be seen in Fig. 2. The direct transmission is from the base station to the target vehicle in the first and second stages. The relay transmission is from the base station to the relay vehicle in the first stage and from the relay vehicle to the target vehicle in the second stage.

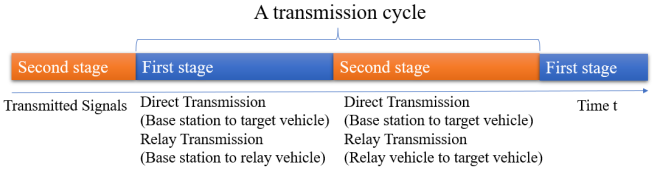


Fig. 2: Description of a transmission cycle.

In a multi-user service scenario, each user receives transmission interference from other users. At the same time, all users are affected by the jamming signals transmitted from the jammer vehicle. Since the direct and relay transmissions coexist in our transmission scenario, the interference analysis is complicated. We will discuss this in detail in the following subsections.

A. Jamming Strategy

During the communication service, friendly vehicles on the road near the eavesdropping vehicle that is not involved in the signal transmission are considered by the base station as potential cooperative vehicles capable of emitting jamming signals. Before each transmission cycle, the base station decides that one of them is selected as the jammer vehicle. Throughout the

transmission cycle, the jammer vehicle directs its beam at the eavesdropping vehicle to reduce the capacity of the wiretap channel. To obtain an excellent jamming effect, the decision of the base station needs to keep the eavesdropping vehicle in the main-lobe beam range of the jammer vehicle and avoid the main-lobe beam of the jammer vehicle to cover the vehicles that are participating in the legitimate transmission.

B. Direct Transmission

The interference to the direct transmission is divided into two parts. In the first stage, the signal transmitted from the base station to the target vehicle will be interfered with by the signal from other downlinks, including the side-lobe signal from the base station to other direct transmission links and the side-lobe signal transmitted by the base station to the relay vehicles. Besides, the jamming signal transmitted from the jammer vehicle to the eavesdropping vehicle also causes some interference to the target vehicles. Thus, the first-stage interference of the direct transmission link from the base station to the kth target vehicle can be expressed as

$$I_{d,k}^{1} = \sum_{k'=1,k'\neq k}^{N_T} P_{B,k'} g_{B,k} m_B M_V + P_J g_{j,k} G_{V,j} G_{V,k}, \quad (7)$$

where $P_{B,k'}$ denotes the transmit power of the base station to the k'th target vehicle, P_J is the transmit power of the jammer vehicle, $G_{V,j}$ is the antenna gain of the cooperative jammer vehicle, and $G_{V,k}$ is the antenna gain of the kth target vehicle.

In the second stage, the direct transmission of the base station is interfered with by the side-lobe signals of other direct transmission links, the cooperative transmission from the relay vehicles to the target vehicles, and the jamming signals transmitted by the jammer vehicle. Thus, the second-stage interference of the direct transmission link from the base station to the kth target vehicle can be expressed as

$$\begin{aligned} I_{d,k}^{2} = &\sum_{k'=1,k'\neq k}^{N_{T,d}} P_{B,k'} g_{B,k} m_B M_V \\ &+ \sum_{i=1}^{N_R} \rho_{i,k} P_{R,i} g_{i,k} G_{V,i} G_{V,k} + P_J g_{j,k} G_{V,j} G_{V,k}, \end{aligned} \quad (8)$$

where $N_{T,d}$ indicates the number of vehicles selected for direct transmission, N_R is the number of selected relay vehicles, and $G_{V,i}$ denotes the antenna gain of the ith relay vehicle. $P_{R,i}$ denotes the transmit power of the the ith relay vehicle, and $P_{R,i} \in \{P_1, P_2, \cdots, P_{N_P}\}$ with N_P representing the number of the transmit power levels. Thus, the signal-to-interference-plus-noise ratio (SINR) of the two stages for the direct transmission can be expressed as

$$\begin{aligned} \chi_{d,k}^{1} &= \frac{P_{B,k} g_{B,k} M_B M_V}{I_{d,k}^{1} + \sigma^2}, \\ \chi_{d,k}^{2} &= \frac{P_{B,k} g_{B,k} M_B M_V}{I_{d,k}^{2} + \sigma^2}, \end{aligned} \quad (9)$$

where σ^2 denotes the noise power. For the channel capacity of direct transmission, this paper takes the average of the two stage capacities. Then the channel capacity of the kth direct transmission target can be derived by

$$C_{d,k} = \frac{1}{2} W [\log\left(1 + \chi_{d,k}^{1}\right) + \log\left(1 + \chi_{d,k}^{2}\right)], \quad (10)$$

where W is the bandwidth.

C. Relay Transmission

We consider a relay transmission scenario, where the base station sends signals to the ith relay vehicle in the first stage and the ith relay vehicle sends signals to the kth target vehicle in the second stage. In the first stage, both the target vehicle and the relay vehicle aim the main-lobe beam at the base station to receive the signals transmitted by the base station and suffer interference from the signals transmitted by other downlinks as well as interference from the jamming signals transmitted by the jammer vehicle to the eavesdropping vehicle. Thus, the interference received by the kth target vehicle and the ith relay vehicle in the first stage can be expressed respectively as

$$\begin{aligned} I_{r,k}^{1} &= \sum_{k'=1,k'\neq k}^{N_T} P_{B,k'} g_{B,k} G_{B,k'} M_V + P_J g_{j,k} G_{V,j} G_{V,k}, \\ I_{r,i}^{1} &= \sum_{k'=1,k'\neq k}^{N_T} P_{B,k'} g_{B,i} m_B M_V + P_J g_{j,k} G_{V,j} G_{V,i}, \end{aligned} \quad (11)$$

where $G_{B,k'}$ is the antenna gain from the base station to the k'th target vehicle. Then the SINR of the kth target vehicle in the first stage can be obtained by

$$\chi_{r,k}^{1} = \frac{P_{B,k} g_{B,k} m_B M_V}{I_{r,k}^{1} + \sigma^2}. \quad (12)$$

In the second stage, the kth target vehicle will turn the main-lobe beam towards the ith relay vehicle for better reception. The interference to the target vehicle consists of four parts, i.e., the downlink signals from the base station to all the selected direct transmission target vehicles, the signals from other relay vehicles to other target vehicles, the amplified and forwarded interference received by the ith relay vehicle in the first stage, and the jamming signals from the jammer vehicle to the eavesdropping vehicle. Thus, the interference received by the kth target vehicle in the second stage can be derived by

$$\begin{aligned} I_{r,k}^{2} = &\sum_{k'=1,k'\neq k}^{N_{T,d}} P_{B,k'} g_{B,k} G_{B,k'} m_V \\ &+ \sum_{i=1}^{N_R} \rho_{i,k} P_{R,i} g_{i,k} G_{V,i} G_{V,k} \\ &+ I_{r,i}^{1} \beta_{i,k} g_{i,k} M_V M_V + P_J g_{j,k} G_{V,j} G_{V,k}. \end{aligned} \quad (13)$$

Thus, the SINR of the kth target vehicle in the second stage can be obtained by

$$\chi_{r,k}^{2} = \frac{P_{B,k} g_{B,i} M_B M_V \beta_{i,k} g_{i,k} M_V M_V}{I_{r,k}^{2} + (\beta_{i,k} g_{i,k} M_V M_V \sigma^2 + \sigma^2)}. \quad (14)$$

Then the channel capacity of the kth relay transmission target can be derived by

$$C_{r,k} = \frac{1}{2} W[\log\left(1+\chi_{r,k}^{1}\right) + \log\left(1+\chi_{r,k}^{2}\right)]. \tag{15}$$

Finally, combining the direct transmission and the relay transmission, we can obtain the channel capacity of the kth target vehicle as

$$C_k = \left(1 - \sum_{i=1}^{N_R} \rho_{i,k}\right) C_{d,k} + \sum_{i=1}^{N_R} \rho_{i,k} C_{r,k}, \tag{16}$$

where $\sum_{i=1}^{N_R} \rho_{i,k} = 1$ denotes that the kth target vehicle selects the relay transmission mode, while $\sum_{i=1}^{N_R} \rho_{i,k} = 0$ denotes that the kth target vehicle selects the direct transmission mode.

D. Eavesdropping Strategy

This paper assumes that the eavesdropper is a random vehicle on the road. Before a transmission cycle, the eavesdropping vehicle already knows the status and transmission mode of vehicles through the control information transmitted by vehicles to the base station. Assuming that the eavesdropping vehicle only eavesdrops on one transmission signal in a transmission cycle, then the eavesdropping process will have three cases.

- Eavesdropping on the downlink from the base station to the target vehicle or the relay vehicle in the first stage. In this case, the eavesdropping vehicle adjusts its main-lobe beam to the base station to acquire a better quality of the received secrecy signal. On the one hand, If the eavesdropping vehicle and the wiretapped target vehicle are in the same beam, the eavesdropping vehicle will get a similar channel gain with the target vehicle. Thus the secrecy capacity of the communication will become very low. In this situation, leveraging the jamming strategy is essential. On the other hand, if the eavesdropping vehicle intercepts the vehicle in the relay transmission mode, the base station will try to avoid using the beam where the eavesdropping vehicle is located for relay transmission after training. At this time, the eavesdropping vehicle can only obtain the side-lobe beam gain of the base station.
- Eavesdropping on the direct transmission from the base station to the target vehicle in the second stage. In this case, the eavesdropping vehicle also adjusts its main-lobe beam to the base station. Although it still intercepts the downlink transmission, the interference is totally different from the first stage. It is because the downlink and the relay forwarding transmission coexist in the network.
- Eavesdropping on the relay transmission from the relay vehicle to the target vehicle in the second stage. In this case, the eavesdropping vehicle adjusts its main-lobe beam to the wiretapped relay vehicle in order to eavesdrop on the amplified and forwarded signal.

Then we will analyze the interference and channel capacity of the eavesdropping vehicle comprehensively in the above three cases.

1) Eavesdropping on downlink in the first stage: Assuming that the eavesdropping vehicle intercepts the kth target vehicle in the direct transmission mode, the interference received by the eavesdropping vehicle consists of two parts, which are the interference caused by the downlink transmission of other users and the jamming signal transmitted by the jammer vehicle. Then the interference received by the eavesdropping vehicle in the first stage can be expressed as

$$\begin{aligned} I_{e,k}^{d,1} = &\sum_{k'=1,k'\neq k}^{N_{T,d}} P_{B,k'} g_{B,e} G_{B,k'} G_{V,e} \\ &+ \sum_{i'=1}^{N_R} P_{B,i'} g_{B,e} G_{B,i'} G_{V,e} + P_J g_{j,e} M_V G_{V,e}, \end{aligned} \tag{17}$$

where $G_{V,e}$ is the antenna gain of the eavesdropping vehicle. In the case that the eavesdropping vehicle chooses to intercept the target vehicle in the relay transmission mode. We assume that the ith relay vehicle forward signals to the kth target vehicle. Then if the eavesdropping vehicle intercepts the transmission to the ith relay vehicle in the first stage, the interference can be expressed as

$$\begin{aligned} I_{e,k}^{r,1} = &\sum_{k'=1}^{N_{T,d}} P_{B,k'} g_{B,e} G_{B,k'} G_{V,e}, \\ &+ \sum_{i'=1,i'\neq i}^{N_R} P_{B,i'} g_{B,e} G_{B,i'} G_{V,e} + P_J g_{j,e} M_V G_{V,e}. \end{aligned} \tag{18}$$

Then the SINRs of the eavesdropping vehicle in the first stage when it intercepts the kth target vehicle, and the ith relay vehicle can be respectively formulated as

$$\begin{aligned} \chi_{e,k}^{d,1} &= \frac{P_{B,k} g_{B,e} G_B M_V}{I_{e,k}^{d,1} + \sigma^2}, \\ \chi_{e,k}^{r,1} &= \frac{P_{B,k} g_{B,e} G_B M_V}{I_{e,k}^{r,1} + \sigma^2}. \end{aligned} \tag{19}$$

2) Eavesdropping on direct transmission in the second stage: The interference received by the eavesdropping vehicle in the second stage when intercepting the direct transmission has three components, namely the downlink signal of other direct transmissions, the forwarded signal of relay vehicles, and the jamming signal. Then the interference of the eavesdropping vehicle can be expressed as

$$\begin{aligned} I_{e,k}^{d,2} = &\sum_{k'=1,k'\neq k}^{N_{T,d}} P_{B,k'} g_{B,e} G_{B,k'} G_{V,e} \\ &+ \sum_{i=1}^{N_R} P_{R,i} g_{i,e} G_{V,i} G_{V,e} \\ &+ P_J g_{j,e} M_V G_{V,e}. \end{aligned} \tag{20}$$

Then the SINR of the eavesdropping vehicle in the second stage when it intercepts the kth target vehicle can be formulated as

$$\chi_{e,k}^{d,2} = \frac{P_{B,k} g_{B,e} G_B M_V}{I_{e,k}^{d,2} + \sigma^2}. \tag{21}$$

Therefore, the channel capacity of the eavesdropping vehicle when intercepting the kth target vehicle in the direct transmission mode can be obtained by

$$C_{e,k}^{d} = \frac{1}{2} W [\log\left(1 + \chi_{e,k}^{d,1}\right) + \log\left(1 + \chi_{e,k}^{d,2}\right)]. \tag{22}$$

3) Eavesdropping on relay transmissions in the second stage: The eavesdropping vehicle directs its main-lobe beam at the relay vehicle during the second stage to obtain higher channel gain. The interference has four components: the downlink signal of direct transmissions, the forwarded signal of other relay vehicles, the amplified and forwarded interference received by the relay vehicle in the first stage, and the jamming signal. Then the interference of the eavesdropping vehicle can be expressed as

$$\begin{aligned} I_{e,k}^{r,2} = & \sum_{k'=1}^{N_{T,d}} P_{B,k'} g_{B,e} G_{B,k'} G_{V,e} \\ & + \sum_{i=1}^{N_R} P_{R,i} g_{i,e} G_{V,i} G_{V,e} \\ & + I_{e,k}^{r,1} \beta_{i,k} g_{i,e} G_V M_V + P_J g_{j,e} M_V G_{V,e}. \end{aligned} \tag{23}$$

Then the SINR of the eavesdropping vehicle in the second stage when it intercepts the kth target vehicle can be formulated as

$$\chi_{e,k}^{r,2} = \frac{P_{B,k} g_{B,i} G_B M_V \beta_{i,k} g_{i,j} G_{V,i} M_V}{I_{e,k}^{r,2} + (\beta_{i,k} g_{i,j} G_{V,i} G_{V,j} \sigma^2 + \sigma^2)}. \tag{24}$$

Therefore, the channel capacity of the eavesdropping vehicle when intercepting the kth target vehicle in the relay transmission mode can be obtained by

$$C_{e,k}^{r} = \frac{1}{2} W \left[\log\left(1 + \chi_{e,k}^{r,1}\right) + \log\left(1 + \chi_{e,k}^{r,2}\right)\right]. \tag{25}$$

In summary, the channel capacity of the eavesdropping vehicle when intercepting the kth target vehicle can be derived by

$$C_{e,k} = \left(1 - \sum_{i=1}^{N_R} \rho_{i,k}\right) C_{e,k}^{d} + \sum_{i=1}^{N_R} \rho_{i,k} C_{e,k}^{r}. \tag{26}$$

We define $X = \{x_k, k = 1, 2, \cdots, N_T\}$, where $x_k = 1$ means that the kth target vehicle is eavesdropped by the eavesdropping vehicle, otherwise $x_k = 0$. We can express the secrecy capacity of the kth target vehicle as

$$C_{s,k} = \max\left\{0, C_k - x_k C_{e,k}\right\}. \tag{27}$$

E. Optimization Problem

In this paper, the optimization objective is to maximize the secrecy capacity of all target vehicles and simultaneously ensure the energy efficiency of the relay vehicles. We jointly optimize the beam allocation, the cooperative nodes selection, and the power control. The optimization problem can be formulated as

$$\max_{\vartheta,\rho,\iota,\varphi} \sum_{k=1}^{N_T} C_{s,k}, \tag{28}$$

$$C1: \sum_{k=1}^{N_T} \vartheta_{m,k} \le 1, \sum_{m=1}^{N_B} \vartheta_{m,k} = 1, \tag{28a}$$

$$C2: \sum_{i=1}^{N_T} \rho_{i,k} \le 1, \sum_{k=1}^{N_R} \rho_{i,k} \le 1, \tag{28b}$$

$$C3: \sum_{j=1}^{N_J} \varphi_j = 1, \tag{28c}$$

$$C4: \sum_{n=1}^{N_P} \iota_{n,i} = 1, \forall i = 1, 2, \cdots, N_R, \tag{28d}$$

$$C5: \sum_{k=1}^{N_T} x_k = 1, \tag{28e}$$

$$C6: C_{s,k} \ge \phi, \forall k = 1, 2, \cdots, N_T, \tag{28f}$$

$$C7: \eta_i \le \zeta, \forall i = 1, 2, \cdots, N_R, \tag{28g}$$

where $\vartheta_{m,k}$ is the binary beam allocation indicator with $\vartheta_{m,k} = 1$, implying that the mth beam is allocated to the kth target vehicle and $\vartheta_{m,k} = 0$ otherwise. $\rho_{i,k}$ is the binary relay transmission indicator with $\rho_{i,k} = 1$ denoting that the ith relay vehicle is selected for forwarding signal to the kth target vehicle and $\rho_{i,k} = 0$ otherwise. φ_j is the binary jammer selection indicator with $\varphi_j = 1$ implying that the jth potential jammer vehicle is selected as a jammer to confuse the eavesdropping vehicle and $\varphi_j = 0$ otherwise. $\iota_{n,i}$ is the binary power level selection indicator with $\iota_{n,i} = 1$ implying that the ith relay vehicle selects the nth transmit power level and $\iota_{n,i} = 0$ otherwise. N_J is the number of potential jammer vehicles. The constraint (28a) indicates that each beam can be allocated to at most one target vehicle, and a target vehicle is served by one beam. The constraint (28b) indicates that each relay vehicle can serve at most one target vehicle, and each target vehicle in the relay transmission mode can only select one relay vehicle to forward signals. The constraint (28c) indicates that only one idle vehicle is selected as the jammer. The constraint (28d) indicates that each relay vehicle can select one to transmit power level in the transmit power set. The constraint (28e) indicates that the eavesdropping vehicle only intercepts one target vehicle in each transmission cycle. The constraint (28f) indicates the minimal secrecy capacity that the system requires for each target vehicle, where ϕ is the secrecy capacity threshold. The constraint (28g) indicates the maxmal energy consumption that the system requires for the vehicles, where η_i is the energy consumption of the ith relay vehicle, and ζ is the energy consumption threshold.

IV. DRRL-based Cooperative Secure Transmission Scheme

DRL is an effective approach to tackle the complex optimization problem raised in the previous section, but it still faces significant challenges. The first challenge is the vast

solution space. On the one hand, to enable the base station to make communication decisions for arbitrary vehicle positions, the position of each vehicle must be used as the state in DRL, along with the blocking and beamforming effects in the environment. This leads to a huge state space. On the other hand, jointly optimizing the relay selection, jammer selection, and relay power selection for the base station results in a massive action space. The immense state and action spaces make the solution space of this problem very large. It is difficult for the agent to find the most suitable actions in the solution space.

The second challenge involves constraints on beam selection and relay power. When the same beam is repeatedly selected by the base station for multi-user decisions, the secrecy capacity of the subsequently selected vehicles becomes zero. It is challenging to teach this logic to the base station through DRL methods. The optimization objective of the base station is to maximize the total secrecy capacity of all users, which conflicts with the magnitude of relay power. Striking a balance between the two is difficult, and it poses difficulties for the convergence and performance of the algorithm.

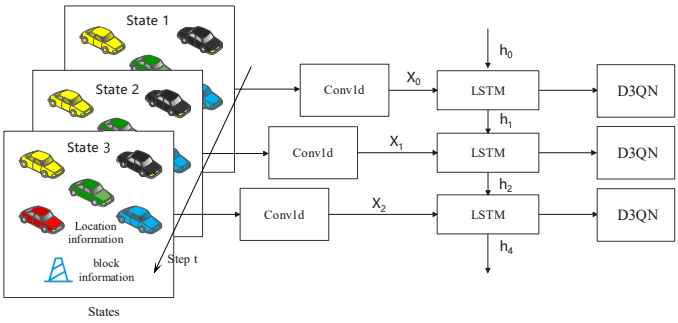


Fig. 3: The proposed D3RQN framework.

We use the following approaches to address the challenges mentioned above.

1) We first add a one-dimension convolution for extracting state information based on D3QN, then use an LSTM network that learns temporal information to reduce the size of the solution space and aid the algorithm convergence. This extension is applied to the D3QN framework, and the resulting algorithm is called Dueling Double Deep Recurrent Q-Network (D3RQN). The framework of the proposed D3RQN is shown in Fig. 3.
2) Secondly, action masking is employed to enforce the exclusion of redundant beam selections, and appropriate training methods are designed to facilitate DRL algorithm learning of decision-making for arbitrary vehicle positions.
3) Finally, the conflict between secrecy capacity and relay power is tackled by proposing a reward function based on threshold settings. This function enables the base station to select a larger relay power as much as possible under certain power constraints, thus converting the conflicting objective into an optimization problem that meets the threshold conditions.

A. Network Structure and Learning Algorithms

Compared with the D3QN algorithm, we have modified the network structure to optimize the problem raised in the previous section. The overall algorithm of D3RQN is shown in Fig. 4. Since the scheme adopts a sequential decision-making method to select communication schemes for the target vehicles, the base station needs to know the state and action of the previous decisions to learn the impact of the previous decisions on the current decision. Therefore, we use one-dimension convolution to extract information, followed by an LSTM network that remembers the previous states.

LSTMs retain and forget information through the dynamics of LSTM memory cells, hidden states, and gating mechanisms, including input, forget, and output gates. At time t, we assume the input to the LSTM layer is x_t. Then we have

$$\begin{aligned} f_t &= \sigma\left(W_{if}x_t + W_{hf}h_{t-1} + b_f\right), \\ i_t &= \sigma\left(W_{ii}x_t + W_{hi}h_{t-1} + b_i\right), \\ \tilde{C}_t &= \tanh\left(W_{iC}x_t + W_{hC}h_{t-1} + b_C\right), \\ C_t &= f_t \times C_{t-1} + i_t \times \tilde{C}_t, \\ o_t &= \sigma\left(W_{io}x_t + W_{ho}h_{t-1} + b_o\right), \\ h_t &= o_t \times \tanh\left(C_t\right), \end{aligned} \tag{29}$$

where W_{if}, W_{hf}, and b_f are the weights of the forget gate. h_{t-1} is the last hidden state, and $\sigma(x)$ denotes the sigmoid function. W_{ii}, W_{hi}, and b_i are the weights of the input gate. W_{iC}. W_{hC}, and b_C are the weights of the self-recurrent connection. W_{io}, W_{ho}, and b_o are the weights of the output gate. f_t, i_t and $\tilde{C}_t$ are intermediate variables used to calculate the new memory cell state C_t. o_t is the activation vector used to compute the new hidden state h_t, which is also the output of the LSTM.

This paper uses the bootstrap random update method for the sample sampling method. In our communication scenario, the number of decision-making steps for an episode is fixed at 3, so we decide to stack the entire episode as updated episode data into the replay memory. During the training process, the agent randomly selects a batch of episodes from the replay memory and updates the samples of the entire episode as the updated data. The initial state of the LSTM is set as zero at the start of each update.

The D3RQN algorithm incorporates the idea of the Double DQN algorithm based on the Dueling DQN algorithm. It uses the evaluation network to obtain the action corresponding to the optimal action value in the state s_{t+1}. Then it uses the target network to calculate the action value of the action to get the target value. Through the interaction of the two networks, the overestimation problem of the algorithm is effectively avoided. The target Q value can be calculated by

$$Q_t = r + \gamma \hat{Q}(s', \arg\max_{a'}(Q(s', a'; \theta)); \theta^-), \tag{30}$$

where r is the reward of a_t, $\gamma \in [0, 1]$ is the discount factor, which can weight the future rewards and prevent the cumulative discounted reward from becoming infinite. θ is the parameter of action network, θ^- is the parameter of target network, and $\hat{Q}(s, a; \theta^-)$ is the Q value of target network. s

and a are the state and action of the current time. s' and a' are the state and action of the following time.

The loss function for Double DQN update is given by

$$\mathop{\mathbb{E}}_{(s,a,r,s')\sim \mathrm{U}(D)}\left[\left(Q_t - Q^{\theta}(s,a)\right)^2\right], \tag{31}$$

where D is the replay buffer [30]. A minibatch of experiences (s, a, r, s') is uniformly sampled from D when updating the network.

In addition, The Dueling DQN algorithm emerges as a powerful variant of DQN. It enhances performance by introducing a unique architecture for the neural network. The key concept of Dueling DQN is its ability to bifurcate the estimation of the state-value function and the action advantage function within the network. Unlike a traditional DQN, where the neural network approximates the Q-value function directly, Dueling DQN innovatively partitions the last layer of the network into two distinct streams. One stream is dedicated to estimating the state-value function, while the other focuses on estimating the advantage for each action. The final Q values are then derived by combining these state and advantage values. The formula for Dueling DQN is represented as

$$Q(s,a;\theta) = V(s) + A(s,a) - \frac{1}{|A|}\sum_{a'} A(s,a'), \tag{32}$$

where $V(s)$ represents the value of state s and $A(s,a)$ signifies the advantage of taking action a in state s. The subtraction of the average advantage aids in stabilizing the learning process. This unique separation enables Dueling DQN to learn the value of states without the necessity to learn the effect of each action.

In our scheme, the update of the evaluation network adopts the method of soft update, which is expressed as

$$\theta^{-} \leftarrow \tau\theta + (1-\tau)\theta^{-}, \tag{33}$$

where τ is the update coefficient, indicating the update range of the network parameters.

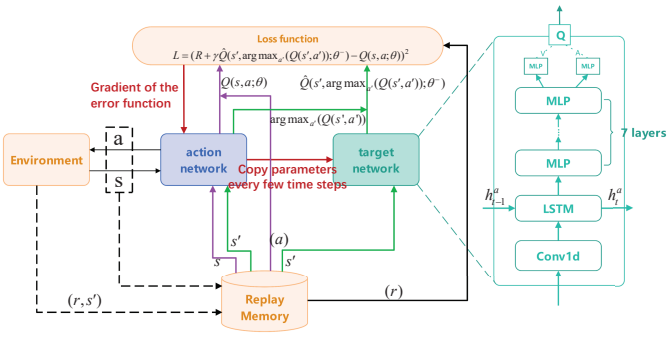


Fig. 4: Network structure of D3RQN algorithm

The learning algorithm is divided into two phases, namely, the training phase and the validation phase. In the training phase, the base station agent uses the D3RQN algorithm for decision learning. After adequate training episodes, the base station performs well for all communication requests from random vehicle streams on the road. In the validation phase, the base station inputs the state and environment into the trained network to obtain the output decisions. The detailed procedures of our D3RQN-based energy-efficient cooperative secure transmission scheme are shown in Algorithm 1. The following are the settings of our D3RQN algorithm.

Algorithm 1 Energy-efficient cooperative secure communication scheme based on DRRL

Input : Vehicle generation, the simulator of D3RQN environment.
Initialize : The action network parameters θ, and target network parameters θ^{-}, set $\theta = \theta^{-}$.
for episode n_e **do**
  Randomly select the target vehicles, and eavesdropping vehicle.
  Randomly generate blocking coefficient.
  Reset potential relay selection list, and potential jammer selection list.
  Reset status of beam selection.
  Set the hidden layer parameter of LSTM network to 0.
  Calculate the secrecy capacity of the optimal solution, which is used to get the reward.
  for each target vehicle k **do**
    Obtain the current state s_k.
    Select action a_k by actor network based on state s_k and hidden layer parameter after action mask.
    Update hidden layer parameters.
    Execute action a_k, get the next state s_{k+1} and obtain the secrecy capacity and energy consumption related reward r'_k.
    Record $\{s_k, a_k, r'_k, s_{k+1}\}$.
  end for
  Obtain the system reward, get the final reward of target vehicles $\{r_k, k = 1, 2, \cdots, N_T\}$.
  Store the experience $(s_i, a_i, r_i, s_{i+1}), i = 1, 2, \cdots, N_T$.
  Sample a batch of data from experience pool.
  if n_c reaches the update step **then**
    Soft update parameters from action network to target network, $\theta^{-} \leftarrow \tau\theta + (1-\tau)\theta^{-}$.
  end if
end for

B. Action Mask and Training Approach

In order to enable the intelligent agent to learn the constraint of non-reusing beams more quickly in decision-making, the training process employs action mask. Specifically, during decision-making, the base station agent masks the decisions that would lead to reusing beams by setting the Q-value of the corresponding action to negative infinity. However, when it is impossible to select non-reusing beams, the action masking is lifted. These modifications only occur during action selection, and the neural network parameters remain unchanged. Compared to the slower learning process involving negative rewards, this approach allows the intelligent agent to directly experience non-reusing beam selection and rapidly learn the constraint of avoiding beam reusing.

To learn the decision policy for vehicles in arbitrary locations, we vary the blocking coefficient at each training episode

while training with different combinations of target and eavesdropping vehicles for the same vehicle location. Besides, the locations of vehicles will be changed after a certain number of training episodes. When trained on enough vehicle locations, the base station can learn vehicle combinations for almost all locations.

C. State and Environment

The environment E_t in the algorithm includes channel information, blocking information (obstacle location, obstacle type, obstacle factor), beam information (beam width, number of beams, beam assignment information, beam coverage information), potential relay vehicle information, target vehicle information, eavesdropping vehicle information, potential jammer vehicle information, and the location information of all vehicles on the road. In each time slot, the base station gets the state S_t in the current environment E_t. $S_t = \{S_{t,k}, k = 1, 2, \cdots, N_T\}$, and $S_{t,k}$ is defined as

$$S_{t,k} = \{\mathbf{B}_s, \mathbf{L}_{rx}, \mathbf{L}_{ry}, \mathbf{L}_{jx}, \mathbf{L}_{jy}, \mathbf{B}_l, L_{tx}, L_{ty}, L_{ex}, L_{ey}\}, \tag{34}$$

where $\mathbf{B}_s$ is a vector of length N_B, with values of 0 or 1, indicating whether each beam has been selected in previous decisions. $\mathbf{L}_{rx}$ and $\mathbf{L}_{ry}$ are vectors of length $N_{V,r}$, with normalized values ranging from -1 to 1, representing the horizontal and vertical coordinates of each potential relay vehicle. $N_{V,r}$ is the number of potential relay vehicles. $\mathbf{L}_{jx}$ and $\mathbf{L}_{jy}$ are vectors of length N_J, with normalized values ranging from -1 to 1, representing the horizontal and vertical coordinates of each potential jammer vehicle. $\mathbf{B}_l$ is a vector of length N_B, with values ranging from 0 to 1, indicating the current degree of blockage for each beam. L_{tx} and L_{ty} are scalars, with normalized values ranging from -1 to 1, representing the horizontal and vertical coordinates of the target vehicle. L_{ex} and L_{ey} are scalars, with normalized values ranging from -1 to 1, representing the horizontal and vertical coordinates of the eavesdropping vehicle.

D. Action Space

Based on the current state, the base station can assign a transmission link to the kth target vehicle and select the cooperative vehicle that sends the jamming signal. The action is a combination of beam assignment of multiple users, the selected relay vehicle, and the selected jammer vehicle. Besides, when the base station selects the relay transmission mode, it will also determine the transmit power. Thus, the action space can be represented by a three-dimensional coordinate system. The X-axis represents the selection of beam and potential relay vehicles. The Y-axis represents the transmit power of the relay vehicle. The Z-axis represents the selection of the potential jammer. The size of the action space A is $N_{V,r} \times N_P \times N_J + N_J$. The term $N_{V,r} \times N_P \times N_J$ represents the number of combinations in relay transmission mode, which includes the selection of relay vehicles, relay power, and jammer vehicles. The additional term N_J represents the number of combinations in direct transmission mode, where only the jammer selection is considered, disregarding the relay power.

The actions are selected from the action space $A = \{A_{t,k}, k = 1, 2, \cdots, N_T\}$, and $A_{t,k}$ is defined as

$$A_{t,k} = \{(\varsigma_r, \varsigma_p, \varsigma_j) | \varsigma_r \in \psi_r, \varsigma_p \in \psi_p, \varsigma_j \in \psi_j\} \tag{35}$$

where ψ_r is the set of potential relay vehicles, ψ_p is the discretized relay power set, and ψ_j is the set of potential jammer vehicles. ς_r, ς_p, and ς_j represent the selected relay vehicles, jammer vehicles, and relay power respectively.

$A_{t,k}$ denotes all combinations of beam allocation, relay selection, jammer selection, and transmit power selection that the kth target vehicle can select at time t. After that, the agent is given a reward r_k based on the secrecy capacity and energy consumption of the selected action to evaluate the merit of the kth choice. At the end of N_T choices, the agent will get the reward $r_{t,k}$ for the kth choice. Unlike r_k, $r_{t,k}$ considers the choices of other target vehicles, which combines all the choices to make the evaluation. After taking action, the environment will arrive at the next state S_{t+1} with the state transition probability $P(E', r|E, A_t)$, and the base station will analyze the state S_{t+1} and make the choice of action from A_{t+1}.

E. Reward Setting

The key to solving the decision-making problem in (28) using the D3RQN algorithm is the design of the reward function, which is not only related to the performance of the final decision but also determines whether the training results of the network can converge. Our reward contains the following two parts.

1) Secrecy capacity-related reward: Many traditional DRL approaches are designed for static environments. However, in our problem of cooperative secure communication, the environment is highly dynamic. Specifically, rewards based on the value of secrecy capacity will fluctuate with the mobility of the vehicles. For instance, if the distance between transceivers decreases and the link quality improves, a higher secrecy capacity for communication is naturally enabled. This is due to the intrinsic properties of the physical world and is independent of the decision-making algorithm. The high-variance and biased reward estimation caused by shifting environment dynamics may significantly degrade performance. To mitigate this issue, we utilize the approximate regretted reward [39]. The regret reward is defined as the difference between the reward associated with an optimal policy and the actually acquired reward and can be expressed as

$$r_{c,k} = \lambda_s (C_{s,k} - C^*_{s,k}) + u, \tag{36}$$

where $C^*_{s,k}$ is the secrecy capacity of optimal policy, which is calculated by traversing all possible decision combinations. λ_s and u are the factors for adjusting the reward range. This regret reward reflects the gap between the current and optimal policies and enables precise evaluation of the training performance in dynamic environments. The performance of optimal policy can be used as a reference baseline to measure the gap and implicitly track the latent environment changes, thereby reducing the reward fluctuation.

2) Energy consumption-related reward: To ensure that the energy consumption of the relay vehicle satisfies the system requirement, we set the energy consumption-related reward as

$$r_{p,i} = \begin{cases} \delta_1, & \text{if } \eta_i > \zeta, \\ \delta_2, & \text{if } \eta_i \leq \zeta, \end{cases} \tag{37}$$

where δ_1 is a negative value to punish the decisions that do not satisfy the energy consumption requirement, and δ_2 is a positive value.

To acquire better secrecy performance, the final reward of each target vehicle decision considers its own reward and the rewards of other target vehicles globally. Then the decision of each target vehicle will be considered from the overall reward maximization. Therefore, the reward of the kth target vehicle can be obtained by

$$r_{t,k} = \lambda_c r_{c,k} + \lambda_d \sum_{k'=1,k'\neq k}^{N_T} r_{c,k'} + \lambda_p \sum_{i=1}^{N_R} r_{p,i}, \tag{38}$$

where λ_c, λ_d, and λ_p are the reward weights, which are used to balance the importance of the secrecy capacity of the individual target vehicle, the secrecy capacity of other target vehicles, and the energy consumption.

V. Numerical Results

In this section, numerical results are presented to evaluate the performance of the proposed energy-efficient cooperative secure transmission scheme in vehicular networks. In the conducted simulation, we utilized an Intel Xeon Gold 6258R (CPU) and an RTX3070 (GPU) as our hardware devices. The D3RQN algorithm underwent a training process spanning 35,000 episodes. This training phase was completed in a duration of 2.5 hours. Furthermore, we observed that the model required an average time of 5.869 seconds to perform 10,000 inferences, demonstrating its efficiency.

During training, the algorithm conducts decision training on a random target vehicle combination in a random road environment. Specifically, when the road vehicle position at time t is fixed, the algorithm will train under a random target vehicle combination, with each combination training 200 episodes. Update time t after sufficient training to allow the distribution of vehicles on the road to change and continue training. When the vehicles on the road are dense enough, or the distribution of trained vehicles is sufficient, the algorithm can encounter almost all communication situations and obtain a universal decision strategy. For the parameter settings such as the channel and antenna models, the simulation refers to the 3GPP technical specification in [40], and Table II lists the relevant parameters. Table III lists the network structure of D3RQN during simulation.

TABLE II: Simulation Parameters

Parameter	Value
Carrier frequency	28 Ghz
Bandwidth of mmWave W	2 GHz
Beam number of base station	8
Number of target vehicles	3
Vehicle arrival rate λ	0.55
Noise power	-70 dBm
Beam-width of base station θ_B	15°
Beam-width of vehicle θ_V	30°
Main-lobe gain M	13 dB
Side-lobe gain m	0.05 dB
Transmit power of base station P_B	30 dBm
Transmit power of jammer vehicle P_J	23 dBm
Transmit power of relay vehicle P_R	$[\frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1] \cdot P_J$
Discount factor γ	0.99
Batch size	128
Soft update coefficient τ	0.05
Learning rate	0.005
Replay memory	1000000

TABLE III: D3RQN Network Parameters

Layer	Input size, Output size
Conv1d	1×42, 16×39
LSTM×4	1×624, 1×256
Linear 1	1×256, 1×1024
Linear 2	1×1024, 1×1024
Linear 3	1×1024, 1×512
Linear 4	1×512, 1×512
Linear 5	1×512, 1×256
Linear 6	1×256, 1×256
Linear 7	1×256, 1×128
Linear 8	1×128, 1×116

A. Benchmark Schemes and Metrics

We also show the performance of other six schemes as the benchmarks. We compare our proposed scheme with the benchmark schemes to provide a comprehensive analysis. The benchmark schemes are summarized as follows.

- Optimal Scheme. The optimal solution is the maximum secrecy capacity that can be achieved for each scenario. At each decision point, the base station searches for the total secrecy capacity of all combinations of beam allocation, relay selection, jammer selection and transmit power selection, and selects the maximum value as the decision of the optimal solution.
- Asynchronous Advantage Actor-Critic (A3C) Scheme. A3C is another DRL algorithm that we use as the standard performance of deep reinforcement learning algorithms. The A3C scheme does not adopt the convolution and LSTM networks of the D3RQN scheme, only linear layers, and the width and depth of the linear layers are consistent with the D3RQN scheme. In the same training environment, we consume more computational resources to train the A3C network. We use this scheme to demonstrate the advantages of the D3RQN scheme over general DRL algorithms.
- D3QN Scheme with Only Linear Layers. This scheme is represented by D3QN in the legend. In this scheme, the convolution and LSTM layers of the D3RQN scheme are removed, and the depth and width of the remaining network stay unchanged. This scheme is used to demonstrate the performance of convolution and LSTM layers.
- D3RQN Scheme without Regret Reward. This shceme is represented by without regret reward in the legend. The reward in the D3RQN scheme is changed to give the agent a reward according to the value of the secrecy capacity. The reward is divided into 5 levels, which are -0.3, -0.1, 0.3, 0.6, 1. This scheme is used to demonstrate

the effect of regret reward on the performance.

- Random Selection Scheme. The base station randomly makes beam allocation, relay selection, jammer selection and transmit power selection decisions in each transmission cycle.
- Direct Transmission Scheme. This scheme does not use relay vehicles for signal transmission during communication. The base station always chooses to transmit signals directly to the target vehicle, while the jammer vehicle is selected as the one closest to the eavesdropping vehicle. This scheme is used to discuss the necessity of relay transmission in this scenario.

In this paper, we leverage three significant performance metrics to sufficiently evaluate the secrecy and energy consumption performance of the cooperative communication scheme.

- Secrecy capacity. This performance metric reveals the real transmission throughput of the system under the premise that the eavesdropping vehicle cannot decode the confidential message.
- Secrecy probability. The secrecy probability is defined as $P_s = \mathbb{P}\{C_{s,k} > \epsilon_s\}$, where ϵ_s is the secrecy capacity threshold. This performance metric is used to evaluate the secrecy performance under various secrecy capacity requirements.
- Energy consumption. The measurement of this performance metric is a key issue that can accelerate the real deployment of our scheme for vehicles. It is because that energy consumption is a big concern for vehicular communications.

The detailed performance analysis is provided in the following subsection.

B. Performance Evaluation

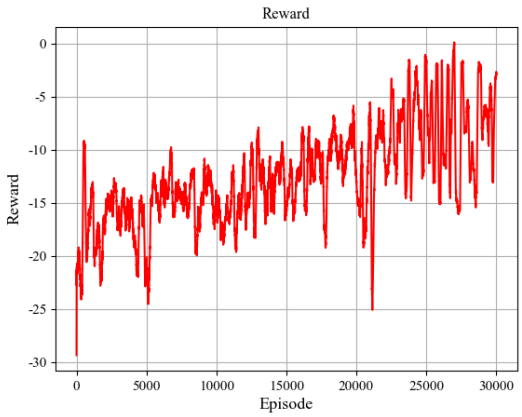


Fig. 5: Training rewards (after removing average).

Fig. 5 shows how the reward changes during D3RQN training. As the number of training steps increases, the reward of D3RQN shows an increasing trend and stabilizes at a high value after 24000 training epochs. The figure shows the training convergence of the D3RQN algorithm in the simulation environment.

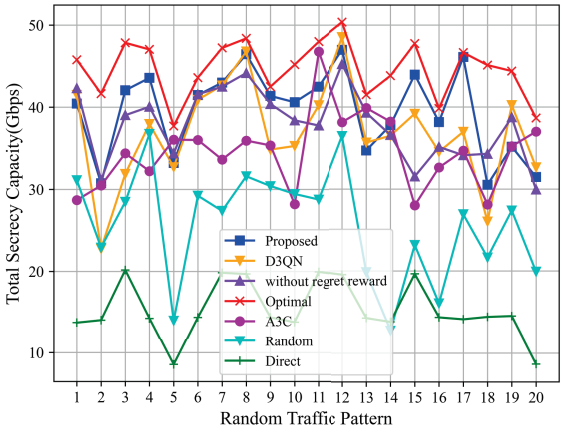


Fig. 6: Secrecy capacity of target vehicles under different traffic patterns.

Fig. 6 compares the secrecy capacity of different schemes. We randomly generate 20 traffic patterns based on queuing theory to simulate actual traffic scenarios. As shown in the figure, the secrecy performance of the proposed scheme is very close to the optimal scheme. It outperforms other baseline schemes in most traffic patterns, demonstrating the superiority and robustness of our proposed scheme.

In addition, while the optimal scheme requires more than 30 seconds to traverse all possibilities, the proposed scheme only needs less than 0.001 seconds. Therefore, the optimal scheme cannot adapt to high dynamic scenarios. The proposed scheme significantly reduces the processing time and is more suitable for vehicle applications.

Due to the lack of conv1d and LSTM in the network structure of the D3QN and A3C schemes, it is impossible to train a universal decision that adapts to various road conditions in a short time, and the overall performance of the test is not ideal. The performance of the scheme without regret reward is slightly higher than that of the scheme without conv1d and LSTM networks, and the experiment shows that the improvement of the network is more important than regret reward. The curve of the direct transmission scheme shows that direct transmission will be affected by the blocking effect and beam repetition selection, resulting in a significant decrease in secrecy capacity. Even in severe blocking situations and strong eavesdropping scenarios, the proposed scheme can still show excellent secrecy performance, because our scheme uses cooperative nodes to relay signals and interfere with eavesdroppers.

Moreover, we test the performance of the proposed scheme, D3QN scheme, and without regret reward scheme under different simulation parameters (different learning rates, batch sizes, reward settings). Under the same training time and parameter conditions, the average secrecy capacity performance of the D3QN scheme and the without regret reward scheme compared with the proposed scheme are decreased by 7.86% and 6.08% respectively. The experiment proves the effect of network improvement and reward improvement in the proposed scheme.

Fig. 7 illustrates the secrecy probability under different

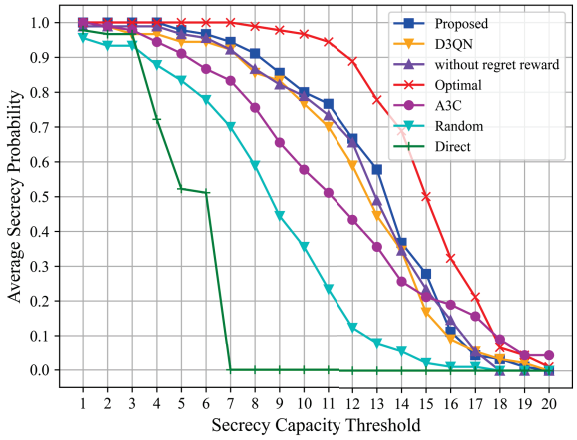


Fig. 7: Secrecy probability of target vehicles under different traffic patterns.

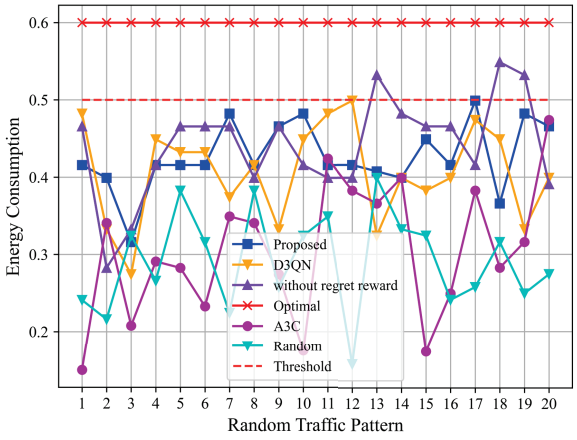


Fig. 9: Energy consumption of target vehicles under different traffic patterns.

secrecy capacity thresholds. As the secrecy threshold increases, the average P_s of the direct transmission scheme first decreases slowly, then sharply, and then drops to zero when ϵ_s rises to 7. The A3C scheme outperforms the random scheme obviously. Due to the consideration of the secrecy capability constraint of each target vehicle, the proposed scheme is better than the previously mentioned schemes. Moreover, the secrecy probability of the proposed scheme is higher than the D3QN scheme and the scheme without regret reward. This demonstrates that the proposed scheme can adapt to dynamic environments and exhibit good secrecy probability performance under various secrecy capacity requirements with proper training.

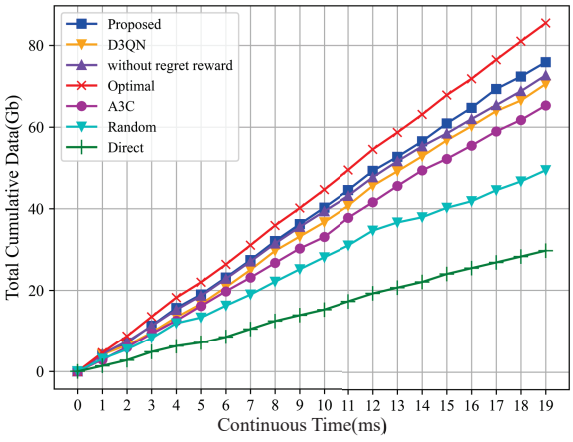


Fig. 8: Total cumulative data for a period of time.

Fig. 8 illustrates the communication performance by comparing the total accumulated data during continuous communications under different duration time in the cooperative mmWave vehicular network. Over time, the gap between direct transmission scheme and the other schemes has increased dramatically. The performance of the proposed scheme is close to the optimal scheme and significantly higher than the A3C and random schemes. The without regret reward scheme is slightly higher than the D3QN scheme, both of which are between the proposed scheme and the A3C scheme. The figure also shows that the proposed scheme helps achieve superior communication performance and adaptability in dynamic communication scenarios than the benchmark schemes.

Fig. 9 depicts the total energy consumption under different schemes. Since the direct transmission scheme does not involve relay vehicles, we compare the performance of the other six schemes. We set the energy consumption threshold as 0.5, and design the punishment mechanism to satisfy this constraint in our proposed scheme. It can be seen that the optimal solution to achieve maximum secrecy capacity always chooses the maximum relay power. After reaching an energy consumption constraint of 0.5, the proposed scheme uses as much transmit power as possible to improve secrecy capacity. Combined with Fig. 6, we observe that the proposed scheme can approximate the secrecy performance of the optimal solution, while transmitting signal with lower energy consumption. This demonstrates that our scheme can balance vehicle energy consumption and secrecy capacity while providing excellent decision-making solutions. In contrast, the A3C scheme, D3QN scheme, and without regret reward scheme do not learn to reasonably use transmit power within a limited training time, which shows the necessity of using conv1d and LSTM networks in our scheme.

Fig. 10 describes the secrecy capacity of each target vehicle under the A3C scheme, proposed scheme, and optimal scheme. We find that the proposed scheme also performs well regarding the secrecy performance of individual vehicle. It is because we consider not only the overall secrecy capacity of the system but also the secrecy capacity of each vehicle when designing the reward. During training, when the secrecy capacity of each target vehicle is less than 6, a negative reward will be given. It can be seen that the secrecy capacity of the target vehicles for the proposed scheme are all higher than the threshold. The A3C scheme does not fully meet the threshold requirement, which is due to the fact that only linear networks cannot obtain excellent decisions in a short training time.

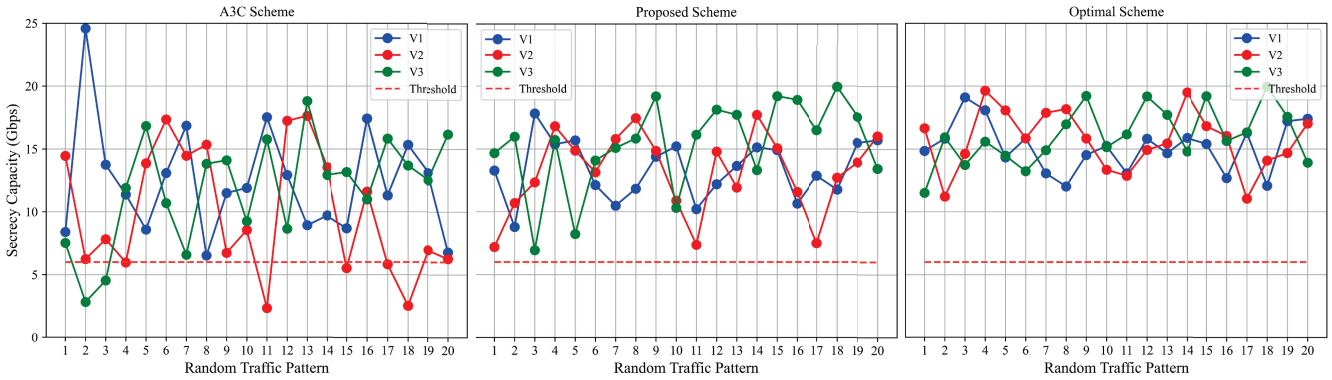


Fig. 10: Secrecy capacity of single target vehicle under different schemes.

VI. Conclusion

This paper has proposed an energy-efficient cooperative secure transmission scheme based on PLS technology and cooperative communication architecture, aiming to maximize the secrecy capacity of target vehicles while ensuring the energy consumption performance of cooperative vehicles. We have comprehensively analyzed the interference situation of the direct transmission mode and the relay transmission mode, and derived the theoretical expression of secrecy capacity. To jointly optimize beam allocation, relay vehicle selection, jammer vehicle selection, and transmission power selection, we have designed the D3RQN learning algorithm and adjusted the training process. After training, the proposed scheme can adapt to highly dynamic vehicle environments. Simulation results show that the proposed scheme can effectively improve the secrecy performance of vehicles while reducing energy consumption.

We also provide possible future research directions here. For the selection of relay power, this paper adopts a discrete decision-making approach. Future research could consider making continuous decisions on power under the premise of global optimal decision-making. Additionally, the design of the reward parameters in this paper primarily relies on experiments. Future research may discover a scientific method for parameter adjustment. Future research may find a more efficient method for parameter adjustment. Finally, the scheme in this paper needs to retrain the model when communication parameters change. Future research could design a model that can be updated online.

References

[1] E. Ahmed and H. Gharavi, "Cooperative vehicular networking: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 19, no. 3, pp. 996–1014, 2018.

[2] A. Pfadler, C. Ballesteros, J. Romeu, and L. Jofre, "Hybrid massive MIMO for urban V2I: Sub-6 GHz vs mmWave performance assessment," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 5, pp. 4652–4662, 2020.

[3] J. Kim, Y.-J. Choi, G. Noh, and H. Chung, "On the feasibility of remote driving applications over mmwave 5G vehicular communications: Implementation and demonstration," *IEEE Transactions on Vehicular Technology*, pp. 1–16, 2022.

[4] I. Rasheed, F. Hu, Y.-K. Hong, and B. Balasubramanian, "Intelligent vehicle network routing with adaptive 3D beam alignment for mmWave 5G-based V2X communications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 5, pp. 2706–2718, 2021.

[5] Z. Li, L. Xiang, X. Ge, G. Mao, and H.-C. Chao, "Latency and reliability of mmWave multi-hop V2V communications under relay selections," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 9807–9821, 2020.

[6] B. M. ElHalawany, A. A. A. El-Banna, and K. Wu, "Physical-layer security and privacy for vehicle-to-everything," *IEEE Communications Magazine*, vol. 57, no. 10, pp. 84–90, 2019.

[7] Y. Ju, Z. Cao, Y. Chen, L. Liu, Q. Pei, S. Mumtaz, M. Dong, and M. Guizani, "NOMA-assisted secure offloading for vehicular edge computing networks with asynchronous deep reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2023.

[8] M. Rice, B. Clark, D. Flanary, B. Jensen, N. Nelson, K. Norman, E. Perrins, and W. K. Harrison, "Physical-layer security for vehicle-to-everything networks: Increasing security while maintaining reliable communications," *IEEE Vehicular Technology Magazine*, vol. 15, no. 3, pp. 68–76, 2020.

[9] Y. Ju, M. Yang, C. Chakraborty, L. liu, Q. Pei, M. Xiao, and K. Yu, "Reliability-security tradeoff analysis in mmwave ad hoc based cps," *ACM Transactions on Sensor Networks*, 2023.

[10] Y. Ju, H.-M. Wang, T.-X. Zheng, and Q. Yin, "Secure transmissions in millimeter wave systems," *IEEE Transactions on Communications*, vol. 65, no. 5, pp. 2114–2127, 2017.

[11] W.-Q. Wang and Z. Zheng, "Hybrid MIMO and phased-array directional modulation for physical layer security in mmWave wireless communications," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 7, pp. 1383–1396, 2018.

[12] J. Chen, R. Zhang, L. Song, Z. Han, and B. Jiao, "Joint relay and jammer selection for secure two-way relay networks," *IEEE Transactions on Information Forensics and Security*, vol. 7, no. 1, pp. 310–320, 2012.

[13] Y. Ju, H. Wang, Q. Pei, and H.-M. Wang, "Physical layer security in millimeter wave DF relay systems," *IEEE Transactions on Wireless Communications*, vol. 18, no. 12, pp. 5719–5733, 2019.

[14] K. Eshteiwi, B. Sleim, and G. Kaddoum, "Full duplex of V2V cooperative relaying over cascaded Nakagami-m fading channels," in *2020 International Symposium on Networks, Computers and Communications (ISNCC)*, 2020, pp. 1–5.

[15] A. Pandey and S. Yadav, "Physical layer security in cooperative AF relaying networks with direct links over mixed rayleigh and double-rayleigh fading channels," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 11, pp. 10 615–10 630, 2018.

[16] M. Zhang, Y. Shang, and Y. Zhao, "Strategy of relay selection and cooperative jammer beamforming in physical layer security," in *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, 2020, pp. 1–6.

[17] Z. Yin, M. Jia, N. Cheng, W. Wang, F. Lyu, Q. Guo, and X. Shen, "UAV-assisted physical layer security in multi-beam satellite-enabled vehicle communications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 3, pp. 2739–2751, 2022.

[18] M. M. Alotaibi and H. T. Mouftah, "Relay selection for heterogeneous transmission powers in VANETs," *IEEE Access*, vol. 5, pp. 4870–4886, 2017.

[19] L. Zhu, J. Zhang, Z. Xiao, X. Cao, X.-G. Xia, and R. Schober, "Millimeter-wave full-duplex UAV relay: Joint positioning, beamforming, and power control," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 9, pp. 2057–2073, 2020.

[20] T. A. Nugraha, "Relay selection and power control in d2d-enabled vehicular communication," in *2023 3rd International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)*, 2023, pp. 186–189.

[21] Q. Weng, Q. Guan, S. Jiang, F. R. Yu, and J. Shi, "Energy-efficient joint relay selection and power control for reliable cooperative communications," in *2013 IEEE/CIC International Conference on Communications in China (ICCC)*, 2013, pp. 414–419.

[22] Z. Sun, S. Balakrishnan, L. Su, A. Bhuyan, P. Wang, and C. Qiao, "Who is in control? practical physical layer attack and defense for mmWave-based sensing in autonomous vehicles," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 3199–3214, 2021.

[23] M. E. Eltayeb, J. Choi, T. Y. Al-Naffouri, and R. W. Heath, "Enhancing secrecy with multiantenna transmission in millimeter wave vehicular communication systems," *IEEE Transactions on Vehicular Technology*, vol. 66, no. 9, pp. 8139–8151, 2017.

[24] M. Yang, Y. Ju, L. Liu, Q. Pei, K. Yu, and J. J. P. C. Rodrigues, "Secure mmwave c-v2x communications using cooperative jamming," in *GLOBECOM 2022 - 2022 IEEE Global Communications Conference*, 2022, pp. 2686–2691.

[25] T.-X. Zheng, Y. Wen, H.-W. Liu, Y. Ju, H.-M. Wang, K.-K. Wong, and J. Yuan, "Physical-layer security of uplink mmWave transmissions in cellular V2X networks," *IEEE Transactions on Wireless Communications*, vol. 21, no. 11, pp. 9818–9833, 2022.

[26] Y. Huang, C. Xu, C. Zhang, M. Hua, and Z. Zhang, "An overview of intelligent wireless communications using deep reinforcement learning," *Journal of Communications and Information Networks*, vol. 4, no. 2, pp. 15–29, 2019.

[27] L. Liu, M. Zhao, M. Yu, M. A. Jan, D. Lan, and A. Taherkordi, "Mobility-aware multi-hop task offloading for autonomous driving in vehicular edge computing and networks," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–14, 2022.

[28] P. Lv, J. Han, J. Nie, Y. Zhang, J. Xu, C. Cai, and Z. Chen, "Cooperative decision-making of connected and autonomous vehicles in an emergency," *IEEE Transactions on Vehicular Technology*, vol. 72, no. 2, pp. 1464–1477, 2023.

[29] Y. Yuan, Z. Li, Z. Liu, Y. Yang, and X. Guan, "Double deep q-network based distributed resource matching algorithm for d2d communication," *IEEE Transactions on Vehicular Technology*, vol. 71, no. 1, pp. 984–993, 2022.

[30] P. Xiang, H. Shan, M. Wang, Z. Xiang, and Z. Zhu, "Multi-agent rl enables decentralized spectrum access in vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 70, no. 10, pp. 10 750–10 762, 2021.

[31] Q. Tang, R. Xie, F. R. Yu, T. Chen, R. Zhang, T. Huang, and Y. Liu, "Distributed task scheduling in serverless edge computing networks for the internet of things: A learning approach," *IEEE Internet of Things Journal*, vol. 9, no. 20, pp. 19 634–19 648, 2022.

[32] M. Elsayed and M. Erol-Kantarci, "Radio resource and beam management in 5G mmWave using clustering and deep reinforcement learning," in *GLOBECOM 2020 - 2020 IEEE Global Communications Conference*, 2020, pp. 1–6.

[33] C. Huang, Z. Yang, G. C. Alexandropoulos, K. Xiong, L. Wei, C. Yuen, Z. Zhang, and M. Debbah, "Multi-hop RIS-empowered terahertz communications: A drl-based hybrid beamforming design," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 6, pp. 1663–1677, 2021.

[34] Y. Geng, E. Liu, R. Wang, and Y. Liu, "Hierarchical reinforcement learning for relay selection and power optimization in two-hop cooperative relay network," *IEEE Transactions on Communications*, vol. 70, no. 1, pp. 171–184, 2022.

[35] C. Huang, G. Chen, Y. Gong, M. Wen, and J. A. Chambers, "Deep reinforcement learning-based relay selection in intelligent reflecting surface assisted cooperative networks," *IEEE Wireless Communications Letters*, vol. 10, no. 5, pp. 1036–1040, 2021.

[36] Y. Ju, H. Wang, Y. Chen, T.-X. Zheng, Q. Pei, J. Yuan, and N. Al-Dhahir, "Deep reinforcement learning based joint beam allocation and relay selection in mmwave vehicular networks," *IEEE Transactions on Communications*, vol. 71, no. 4, pp. 1997–2012, 2023.

[37] X. Xiong, E. Xiao, and L. Jin, "Analysis of a stochastic model for coordinated platooning of heavy-duty vehicles," in *2019 IEEE 58th Conference on Decision and Control (CDC)*. IEEE, 2019, pp. 3170–3175.

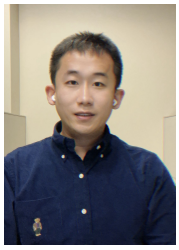
[38] H. Wang, Y. Ju, N. Zhang, Q. Pei, L. Liu, M. Dong, and V. C. M. Leung, "Resisting malicious eavesdropping: Physical layer security of mmWave MIMO communications in presence of random blockage," *IEEE Internet of Things Journal*, vol. 9, no. 17, pp. 16 372–16 385, 2022.

[39] S.-Y. Chen, Y. Yu, Q. Da, J. Tan, H.-K. Huang, and H.-H. Tang, "Stabilizing reinforcement learning in dynamic environment with application to online recommendation," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2018, pp. 1187–1196.

[40] 3GPP, "Study on evaluation methodology of new vehicle-to-everything (v2x) use cases for LTE and nr," 3rd Generation Partnership Project (3GPP), Technical Specification (TS) 37.885, 06 2019, version 15.3.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3209

Ying Ju (Member, IEEE) received the B.S. and M.S. degrees from the School of Electronic Information Engineering, Tianjin University, Tianjin, China, in 2008 and 2010, respectively, and the Ph.D. degree from the School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China, in 2018. From 2016 to 2017, she was a Visiting Scholar at the Department of Computer Science, University of California, Santa Barbara, USA. From 2010 to 2018, she was a Senior Engineer at the State Radio Monitoring Center, Xi'an, China. She is currently an Associate Professor with the Department of Telecommunications Engineering, Xidian University, Xi'an, China. Her research interests include physical layer security of wireless communications, millimeter wave communications, vehicular networks, and AI in wireless communication systems.

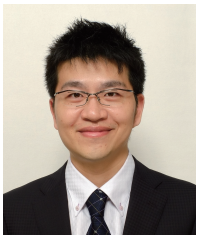
Zipeng Gao received the B.S. degree in Communication Engineering from Xidian University, Xi'an, China, in 2022. He is currently pursuing the M.S. degree in information and communication engineering, Xidian University, Xi'an, China. His research interests include physical layer security of wireless communication and deep reinforcement learning.



Haoyu Wang received the B.S. and M.S. degrees in Communication Engineering from Xidian University, Xi'an, China, in 2018 and 2021, respectively. He is currently pursuing a Ph.D. degree at the University of California Irvine, USA. His research interests include physical layer security of wireless communications, millimeter wave systems, reconfigurable reflecting surfaces, IoT, and AI optimization in 6G.

Lei Liu (Member, IEEE) received the B.Eng. degree in electronic information engineering from Zhengzhou University, Zhengzhou, China, in 2010, and the M.Sc. and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, China, in 2013 and 2019, respectively. From 2013 to 2015, he was employed by a subsidiary of China Electronics Corporation, Beijing, China. From 2018 to 2019, he was supported by the China Scholarship Council to be a Visiting Ph.D. Student at the University of Oslo, Oslo, Norway. He is currently a Lecturer with the State Key Laboratory of Integrated Service Networks, Xidian University, and also with the Xidian Guangzhou Institute of Technology, Xi'an. His research interests include vehicular ad hoc networks, intelligent transportation, mobile-edge computing, and the Internet of Things.

Qingqi Pei (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in computer science and cryptography from Xidian University, in 1998, 2005, and 2008, respectively. He is currently a Professor and a member of the State Key Laboratory of Integrated Services Networks, a Professional Member of ACM, and a Senior Member of Chinese Institute of Electronics and China Computer Federation. His research interests include digital contents protection and wireless networks and security.



Mianxiong Dong (Senior Member, IEEE) received B.S., M.S. and Ph.D. in Computer Science and Engineering from The University of Aizu, Japan. He is the Vice President and Professor of Muroran Institute of Technology, Japan. He was a JSPS Research Fellow with School of Computer Science and Engineering, The University of Aizu, Japan and was a visiting scholar with BBCR group at the University of Waterloo, Canada supported by JSPS Excellent Young Researcher Overseas Visit Program from April 2010 to August 2011. Dr. Dong is the Vice President and Professor at Muroran Institute of Technology. His research interests include large-scale network systems such as mobile networks, wireless sensor networks, vehicle networks, cyber-physical systems, and IoT. He is also engaged in research on a wide range of cutting-edge information technologies, such as edge computing, AI technology, SDN, and big data analysis. His laboratory combines knowledge in the field of large-scale network systems with physical layer technologies to create new value.

Shahid Mumtaz (Senior Member, IEEE) is a Nottingham Trent University (NTU), UK professor. He is an IET Fellow, founder, and EiC of IET "Journal of Quantum Communication," Vice Chair: Europe/Africa Region- IEEE ComSoc: Green Communications & Computing Society. He authorizes four technical books, 12 book chapters, and 300+ technical papers (200+ IEEE Journals/transactions, 100+ conferences, 2 IEEE best paper awards) in mobile communications. Most of his publication is in the field of Wireless Communication. He is a Scientific Expert and Evaluator for various research funding agencies. In 2012, he was awarded an "Alain Bensoussan fellowship." China awarded him the young scientist fellowship in 2017.

Victor C. M. Leung (Life Fellow, IEEE) is the Dean of the Artificial Intelligence Research Institute and a Professor of Engineering at Shenzhen MSU-BIT University (SMBU), China, a Distinguished Professor of Computer Science and Software Engineering at Shenzhen University, China, and an Emeritus Professor of Electrical and Computer Engineering and Director of the Laboratory for Wireless Networks and Mobile Systems at the University of British Columbia (UBC), Canada. His research is in the broad areas of wireless networks and mobile systems, and he has published widely in these areas. His published works have together attracted more than 60,000 citations. He is named in the current Clarivate Analytics list of "Highly Cited Researchers". Dr. Leung is serving on the editorial boards of the IEEE Transactions on Green Communications and Networking, IEEE Transactions on Computational Social Systems, and several other journals. He received the 1977 APEBC Gold Medal, 1977-1981 NSERC Postgraduate Scholarships, IEEE Vancouver Section Centennial Award, 2011 UBC Killam Research Prize, 2017 Canadian Award for Telecommunications Research, 2018 IEEE TCGCC Distinguished Technical Achievement Recognition Award, and 2018 ACM MSWiM Reginald Fessenden Award. He co-authored papers that were selected for the 2017 IEEE ComSoc Fred W. Ellersick Prize, 2017 IEEE Systems Journal Best Paper Award, 2018 IEEE CSIM Best Journal Paper Award, and 2019 IEEE TCGCC Best Journal Paper Award. He is a Life Fellow of IEEE, and a Fellow of the Royal Society of Canada (Academy of Science), Canadian Academy of Engineering, and Engineering Institute of Canada.