IEEE Transactions on Machine Learning in Communications and Networking

Received 18 October 2023; revised 23 February 2024 and 6 July 2024; accepted 17 August 2024.
Date of publication 22 August 2024; date of current version 6 September 2024.

The associate editor coordinating the review of this article and approving it for publication was Z. Yang.

Digital Object Identifier 10.1109/TMLCN.2024.3447634

Deep Reinforcement Learning-Based mmWave Beam Alignment for V2I Communications

YUANYUAN QIAO[1,2] (Graduate Student Member, IEEE), YONG NIU[1,2] (Senior Member, IEEE), LAN SU[1,2], SHIWEN MAO[3] (Fellow, IEEE), NING WANG[4] (Member, IEEE), ZHANGDUI ZHONG[1,2] (Fellow, IEEE), AND BO AI[1,2] (Fellow, IEEE)

[1]School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing 100044, China
[2]Collaborative Innovation Center of Railway Traffic Safety, Beijing 100044, China
[3]Department of Electrical and Computer Engineering, Auburn University, Auburn, AL 36849 USA
[4]School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China
CORRESPONDING AUTHOR: Y. NIU (niuy11@163.com)

This work was supported in part by the Fundamental Research Funds for the Central Universities (Science and Technology Leading Talent Team Project) under Grant 2022JBXT001; in part by Beijing Natural Science Foundation under Grant L232042; in part by the National Key Research and Development Program of China under Grant 2021YFB2900301; in part by the National Natural Science Foundation of China under Grant 62221001, Grant 62231009, and Grant U21A20445; and in part by the Fundamental Research Funds for the Central Universities under Grant 2023JBMC030 and Grant 2022JBQY004.

ABSTRACT Millimeter wave (mmWave) communication can meet the requirements of vehicle-to-infrastructure (V2I) systems, for high throughput and ultra-low latency. However, searching for the optimal beamforming vectors in highly dynamic environments, incurs considerable training overhead. And it is a huge challenge to achieve beam alignment between receivers and transmitters. This paper proposes a beam alignment algorithm based on vehicle position information, to achieve fast beam alignment in the V2I network. In the proposed algorithm, a roadside unit (RSU) obtains a set of candidate beams by the vehicle position information and the double deep $\boldsymbol{Q}$ network (DDQN) algorithm. Then, according to the criterion of maximizing the system spectral efficiency, the optimal beam of the candidate beam set is obtained by the exhaustive search, to achieve fast beam alignment. In this paper, the DeepMIMO dataset is utilized to fully consider the actual scene of V2I, and the effect of Doppler expansion is taken into account in the mathematical model. The simulation results show that the received signal-noise ratio (SNR) of vehicle at different positions is greater than the SNR threshold, which avoids communication interruption and improves the reliability of V2I communications. Meanwhile, we also evaluates the effect of vehicle speed. Compared with other search schemes, the proposed scheme attains higher transmission rates, effectively balances the training overhead and achievable rate, and is suitable for mmWave V2I networks.

INDEX TERMS V2I, roadside unit (RSU), beam alignment, deep reinforcement learning, Markov decision process (MDP).

I. INTRODUCTION

In RECENT years, with the continuous development of the Intelligent Transportation System (ITS), onboard services have been increasingly enriched, and a comprehensive transportation system with high efficiency, low energy, and high reliability has been gradually formed [1]. To achieve vehicle-to-everything (V2X) communication [2], each vehicle application requires higher transmission rates and lower delays. However, current internet of vehicles (IoV) communication technologies cannot meet the requirements of the next generation onboard applications [3].

As a result, millimeter wave (mmWave) communications are considered a high-potential solution that can improve available bandwidth and spectral efficiency. The mmWave frequency band ranges from 30 GHz to 300 GHz, and supports communication with high transmission rates and ultra-low delay [4]. However, it also face some technical challenges. First, it is sensitive to blocking, weak diffraction, and high path loss, which cannot be ignored [5]. The signal noise ratios (SNR) of line-of-sight (LoS) and non-line-of-sight (NLoS) links are different, which affects the reliability of IoV communication. Second, the transceivers are usually equipped with large antenna arrays to compensate for the attenuation of mmWave [6]. Large antenna arrays require a large amount of training overhead when adjusting the beamforming vector, that limits the communication capability of moving vehicles.

In addition, the large antenna array of mmWave communication generates a narrow beam. It leads to many alternative beams, and requires a lot of training overhead to achieve beam alignment [7]. Therefore, the efficient beam alignment between the receiver and transmitter is critical. When a vehicle travels from one location to another, the previously aligned beams may no longer be valid, and thus need to be re-aligned. For a beam search scheme with high complexity, if we consider re-searching the beams at a new location, the vehicle will have already moved to the next location. Therefore, the beam obtained will not be optimal at the new position. In addition, if beam alignment is not timely, the system throughput will be reduced, and communication interruptions will occur. It is worth noting that the power ofÂ inter-carrier interference (ICI) [8] resulting from Doppler spread cannot be ignored in vehicle moving scenarios, which can have an impact on the accuracy of beam alignment. Based on the above analysis, it can be found that the narrow beam alignment between base station (BS) and vehicles is challenging [9]. To solve the above challenges, efficient beam alignment algorithms should be designed to achieve fast beam alignment, reduce training overhead and establish reliable communication links.

This paper proposes a beam alignment scheme in the mmWave vehicle-to-infrastructure (V2I) system. The algorithm combines the double deep Q network (DDQN) algorithm with vehicle position information. It identifies the best beam by maximizing the system's spectral efficiency. The contributions of this paper can be summarized as follows

- We establish a beam alignment model for mmWave V2I. Based on the model, we formulate a beam alignment optimization problem, and model the beam alignment problem on the RSU. Compared to most of the existing V2I works, we consider the ICI caused by Doppler expansion due to vehicle movement.
- We combine the DDQN algorithm with vehicle position, and propose a beam alignment scheme. First, a set of candidate beams is obtained by the DDQN algorithm. Then, according to the criterion of maximizing the system spectral efficiency, the optimal beam is identified from the candidate beam group by exhaustive search. In this paper, we utilize the DeepMIMO dataset, which is more closely aligned to the channel of the street scenario of V2I.
- The paper compares the proposed scheme with several other algorithms, in terms of several performance evaluation metrics. The simulation results show that the proposed scheme can effectively improve the transmission rate of the mmWave V2I system, and avoid communication interruptions. Meanwhile, it effectively weighs training overhead and achievable rates.

The rest of this paper is organized as follows. Section II provides an overview of related work. Section III introduces the system model, including the channel and antenna model. Section IV presents the problem of beam alignment and the proposed algorithm. Section V compares the performance of the proposed algorithm with several algorithms. Finally, Section VI concludes this paper.

II. RELATED WORK

There are many works on mmWave beam alignment. Beam alignment schemes for mmWave multiple-input multiple-output (MIMO) systems, typically perform an iterative search of beam combinations, to find the beam pair with the highest signal gain. It brings a huge overhead. Reference [10] proposed a DDPG method for referenceless beam alignment based on RF fingerprints of user devices. However, the mmWave channel used does not consider the V2I scenario with Doppler effects and training overhead. Meanwhile, DDPG adept at dealing with the continuous action space, may not perform well in discrete actions. Reference [11] proposed a referenceless beam alignment with a deterministic approach. It also requires high operational time to detect the beam. It is not easy to achieve narrow beam alignment between RSU and vehicles in the V2I system. One of the challenges is that it takes a lot of training overhead, to adjust the beamforming vector of large antenna arrays.

To reduce the overhead of beam search, Zhang et al. [12] proposed a hierarchical search algorithm, which reduces the number of beam searches through two stages, including sector-level scanning and beam refinement. However, it does not significantly reduce the search complexity. For spatial reuse, Han [13] and Rasekh [14] proposed to convert the mmWave channel estimation problem into a sparse reconstruction problem by taking advantage of the sparsity. They used compressed sensing technology to effectively estimate the parameters of the sparse channel. Compared with exhaustive search, it can reduce the training overhead. However, the training overhead is still high that is proportional to the number of antennas. In addition, The compressed channel estimation technique usually makes complex assumptions about the channel, which leads to uncertainty about the practical feasibility of the technique.

To further reduce the training overhead, References [15], [16], [17], [18], [19], [20], and [21] proposed a scheme based on side information to achieve beam alignment. These schemes are more suitable for high-speed mobile communication. In V2I system, communication mainly considers LoS, and beam direction at the next moment can be predicted according to the movement information, such as vehicle position and speed [22], [23]. Reference [19] proposed a beam alignment algorithm for inverse fingerprint recognition, which provides candidate beam groups for different vehicle locations through prior measurements about received power at a given location. But the inverse fingerprint database relies on the pre-established database, and updates are iterative and slow. Additionally, large-scale deployment incurs significant overhead. In Reference [21], a data-driven position-assisted scheme with tensor is proposed to reduce the training overhead in MIMO systems, by utilizing vehicle position information and field measurements. However, the tensor completion method is sensitive to noise and computationally complex. Reference [24] proposed an adaptive beam alignment algorithm based on DDQN-RER, to dynamically adjust the beam direction. Although it employs a prioritized empirical replay mechanism, that gives more importance to useful samples, it does not work well when the number of samples are small, with additional training overhead. Reference [25] proposed a DDQN algorithm for beam training, but the effect of Doppler is not considered. Reference [26] combines a vehicular traffic simulator with a ray-tracing simulator, to generate 5G channels. However, the DQN algorithm used does not take into account the overestimation, and the beam training overhead. Meanwhile, there are no remaining comparison algorithms in the simulation and fewer evaluation metrics.

Based on the above analysis, it can be found that majority of current works ([9], [10], [26], [27], [28], [29]) only use common channel models (e.g., 3GPP 38.901, Saleh-Valenzuela, and other channel models), that do not take into account actual V2I communication scenarios. It is worth noting that the algorithms in most of existing work suffer from high training overhead, and poor applicability to the V2I scenario. Meanwhile, ICI caused by Doppler expansion cannot be ignored, and have a considerable impact on the performance of V2I communication. Unfortunately, the existing V2I works ([9], [10], [12], [16], [22], [24], [25], [26], [27], [28], [30]) do not consider it. Also, the trade-off of overhead-performance in beam training is not evaluated in most of the present work. Meanwhile, they do not adequately compare existing algorithms or performance evaluation metrics. Table 1 describes the differences between this paper and related work.

III. SYSTEM OVERVIEW

A. SYSTEM MODEL

In this section, we introduce the system model of beam alignment in mmWave vehicular system, the channel model, and the antenna model. The system model is shown in Fig. 1, where the RSU serves vehicles by the mmWave band, in a service area $\mathcal{G}$. We assume the RSU is at a fixed position and height, and equipped with an antenna array of N_t elements. The vehicle uses a single antenna with omnidirectional communication, whose position at time t is denoted as $g(t) \in \mathcal{G}$.

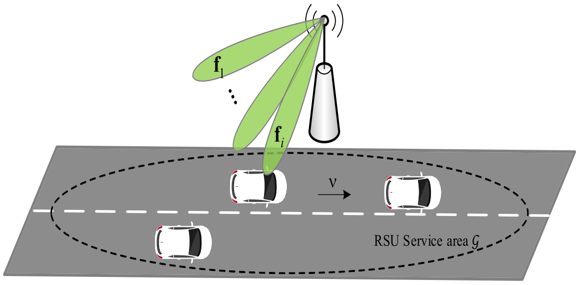


FIGURE 1. **System model of beam alignment in mmWave vehicular networks.**

Further, the RSU systems is assumed to employ the analog beamforming architecture. Then, the received signal at the vehicle side is expressed as

$$y = \sqrt{P_t}\mathbf{H}^T\mathbf{f}s + \mathbf{n}, \tag{1}$$

where y is the received signal, P_t is the transmission power of RSU, $\mathbf{H}$ is the mmWave channel matrix, and $\mathbf{f} = \left[f_1, f_2, \ldots, f_{N_t}\right]^T$ is single beamformer of the transmitter. Besides, s and $\mathbf{n}$ denote the transmitted symbol and the Gaussian white noise with $\mathcal{N}_{\mathbb{C}}(0, \sigma^2)$, respectively.

B. CHANNEL MODEL

We assume a uniform linear array (ULA) antenna on the RSU. The mmWave channel between the RSU and vehicles is modeled with the popular Saleh-Valenzuela channel model [31]. The channel between the transmitter and receiver is expressed as

$$\mathbf{H} = \sum_{l=1}^{L} \alpha_l \mathbf{a}_r(\phi_l^r, \theta_l^r)\mathbf{a}_t^*(\phi_l^t, \theta_l^t), \tag{2}$$

where L is the number of propagation paths, and α_l is the complex path gain of the lth path (including the path loss). And ϕ_l^t, θ_l^t are the lth path azimuth and elevation angles of departure at the transmit antennas, while ϕ_l^r, θ_l^r are the lth path azimuth and elevation angles of arrival at the received antennas, respectively. Moreover, $\mathbf{a}_t(\phi_l^t, \theta_l^t)$ and $\mathbf{a}_r(\phi_l^r, \theta_l^r)$ are the transmit and receive array response vectors, respectively. The array response vectors for UPA arrays are defined in [31]. It is worth noting that for mmWave frequencies, the measurements [32] show that the channels are typically sparse in the angular domain, resulting in a small number of channel paths L (normally in the range of 3-5 paths).

We assume the UPA antenna of RSU is in the yoz plane with M_y and M_z elements, respectively. The array response vector of the transmitter is expressed as

$$\mathbf{a}(\phi, \theta) = \frac{1}{\sqrt{M_y M_z}}[1, \cdots, e^{jkd(m\sin(\phi)\sin(\theta)+n\cos(\theta))}, \\ \cdots, e^{jkd((M_y-1)\sin(\phi)\sin(\theta)+(M_z-1)\cos(\theta))}]^T, \tag{3}$$

where $k = \frac{2\pi}{\lambda}$ is a constant, λ is the wavelength, d is the antenna array element spacing, m is the mth row in

TABLE 1. Differences between this paper and related works.

	Research scene	Object	Algorithm	Channel data	Performance indicators	Comparison scheme	Doppler	Overhead evaluation
this paper	V2I	Max(SE)	DDQN	DeepMIMO	Received SNR Alignment probability Achievable rate Achievable rate ratio	DQN DDQN-PER exhaustive search hierarchical search Bayesian optimization	✓	✓
[9]	V2I	Max(RSS)	DNN	Saleh-Valenzuela	RSS	multiple fingerprints single fingerprint LOS	✗	✗
[10]	V2I	Max (mean rate of UE)	DDPG	Saleh-Valenzuela	Average sum rate	Oracle BS-Sweep Vanilla DDPG BSO&Angle Oracle	✗	✗
[15]	V2I	Min (possible blockage)	a blind protocol design of the precoders/combiners	Ray tracing simulations	Sum Spectral Efficiency	Upper bound	✗	✗
[25]	moving UE and BS	Max(EE/SE)	DDQN	3GPP TR 38.901	Energy efficiency Spectral efficiency	MAB Maximum reward	✗	✓
[26]	V2I	Max (effective channel)	DQN	5GMdata	Average reward	✗	✗	✗
[27]	V2I	Max (UCB score)	MAB	A wideband geometrical channel model	Alignment Rate 3dB Power Loss	greedy ε-greedy	✗	✓
[28]	V2I	Max (Transmission rate)	coordinated alignment with beams subsets	Saleh-Valenzuela	Effective rate	exhaustive search	✗	✗
[29]	mmwave system	Max (received signal)	Bayesian Optimization	Saleh-Valenzuela	Spectral efficiency	OMP method TS-based MAB	✗	✗
[30]	V2I	Max (Beam Prediction Accuracy)	DNN	DeepSense 6G	Prediction Accuracy Complexity	range-velocity radar cube	✗	✓

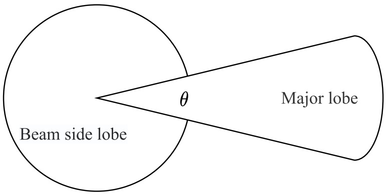


FIGURE 2. The mmWave antenna pattern.

the antenna array, $0 \leqslant m \leqslant M_y - 1$, and n is the nth column in the antenna array, $0 \leqslant n \leqslant M_z - 1$. The size of the antenna array is $M_y \times M_z$. We assume omnidirectional reception with a single antenna at the vehicles, that is $\mathbf{a}_r(\phi, \theta) = 1$.

C. ANTENNA MODEL

The mmWave directional antenna pattern of RSU and vehicle approximates a two-dimensional ideal sector antenna model, to facilitate the analysis of directional antenna gain. It is shown in Fig. 2, whose antenna gain is expressed as

$$G(\theta) = \begin{cases} \dfrac{2\pi - (2\pi - \phi)g}{\phi}, |\theta| \leq \dfrac{\phi}{2} \\ g, \qquad\qquad \text{otherwise} , \end{cases} \tag{4}$$

where θ represents the angle of alignment error between RSU and vehicle, ϕ represents the half-power beamwidth, and g denotes the side lobe gain with $0 \leqslant g \ll 1$.

IV. PROBLEM FORMULATION AND THE PROPOSED ALGORITHM

The beam alignment problem in the mmWave vehicle network is formulated in this section. Then a beam alignment algorithm based on vehicle position information combined with DDQN is proposed.

A. PROBLEM FORMULATION

The goal of beam alignment in this paper is to design the analog beamformer, to maximize the system spectral efficiency with minimizing the training overhead. For mmWave beamforming, we assume that the beam is selected from a pre-defined beam codebook $\mathcal{F} = \left\{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_{N_t}\right\}$, $|\mathcal{F}| = N_t$. Based on the system and channel models, the spectral efficiency is expressed as

$$R = \log_2\left(1 + \frac{P_t}{\sigma^2 + P_{ICI}}\left|H^H f\right|^2\right), \tag{5}$$

where P_t is the transmit power of the RSU, and σ^2 is the noise power. In this paper, to quantify the ICI power caused by Doppler expansion in V2I communication, we adopt the current widely used ICI approximation model [8] is expressed as

$$P_{ICI} = 1 - \int_{-1}^{1} (1 - |\tau|)\, J_0\left(2\pi f_{d,\max} T_s \tau\right) d\tau, \tag{6}$$

where T_s denotes the symbol duration, and $J_0(\cdot)$ is the first class zero-order Bessel function. Moreover, $f_{d,\max} = v \cdot f_c/c$ represents the maximum Doppler spread, where v is the speed of the car, f_c and c represent the carrier frequency and the speed of light, respectively.

If the channel is known, then the design problem of the analog beamformer can be expressed as

$$f^* = \arg\max_{f_i\in\mathcal{F}} \log_2\left(1+\frac{P_t}{\sigma^2+P_{ICI}}\left|H^H f_i\right|^2\right) \tag{7a}$$

$$\text{s.t.} \quad \mathbf{f}_i \in \mathcal{F}, 1 \leqslant i \leqslant N_t \tag{7b}$$

$$\|\mathbf{f}_i\|^2 = 1, 1 \leqslant i \leqslant N_t, \tag{7c}$$

where $\mathbf{f_i}$ is the ith beamforming vector in the RSU's beam codebook $\mathcal{F}$. With the constraints (7b) and (7c), the phase shifter only changes the signal's phase, not the signal's amplitude, so the beamforming vector has a constant modulus value. It should be noted that the narrow beamwidth of mmWave communication requires extremely high beam alignment accuracy in real V2I system. Inaccurate beam alignment may lead to frequent communication interruptions and increased communication delays. Moreover, the process of real-time beam alignment consumes computational resources that cannot be ignored. However, vehicles and infrastructures have limited resources, which requires the algorithms to make a good trade-off between training overhead and accuracy. To address this challenge, our goal is to design a solution that finds the optimal analog beamforming vector that maximizes the spectral efficiency with low training overhead.

B. PROBLEM DECOMPOSITION

We propose to achieve beam alignment by providing a set of candidate beams $\mathcal{C} \subset \mathcal{F}$ for a given vehicle location by deep reinforcement learning (DRL). And the best beam is obtained from the $\mathcal{C}$ by the exhaustive search at RSU. The process of beam alignment assisted by position information is shown in Fig. 3.

The algorithm can be divided into four steps. In Step 1, the vehicle sends an uplink transmission request and the global positioning system (GPS) coordinates $g(t)$ of the current position to the RSU. In Step 2, the RSU conveys the vehicle position's GPS coordinates $g(t)$ to the cloud. The cloud obtains a set of candidate beams $\mathcal{C}$ by DRL network. In Step 3, according to the criterion of maximizing the system spectral efficiency, the RSU obtains the corresponding beamforming vector $\mathbf{f}^*$ from $\mathcal{C}$ by the exhaustive search. In Step 4, the RSU communicates with the vehicle by $\mathbf{f}^*$ for the subsequent uplink or downlink data transmission.

Due to the high speed of vehicle, the GPS coordinates of vehicle position need to be discretized. Suppose RSU service area of the rectangle is $\bar{\mathcal{G}} = [X_0, X_{end}] \times [Y_0, Y_{end}]$. We divide the lanes among the model into multiple uniform grids, according to the x-axis resolution Δ_x and the y-axis resolution Δ_y. Therefore, the discrete GPS coordinates are defined as $\mathbf{g} = (g_x, g_y) \in \bar{\mathcal{G}}$, and we define the position labels that is $\mathbf{p} = (p_x, p_y)$ with $p_x \in \{1, 2, \ldots, L_x\}$ and $p_y \in \left\{1, 2, \ldots, L_y\right\}$, where $L_x = \left\lceil \frac{X_{end}-X_0}{\Delta_x} \right\rceil$ and $L_y = \left\lceil \frac{Y_{end}-Y_0}{\Delta_y} \right\rceil$.

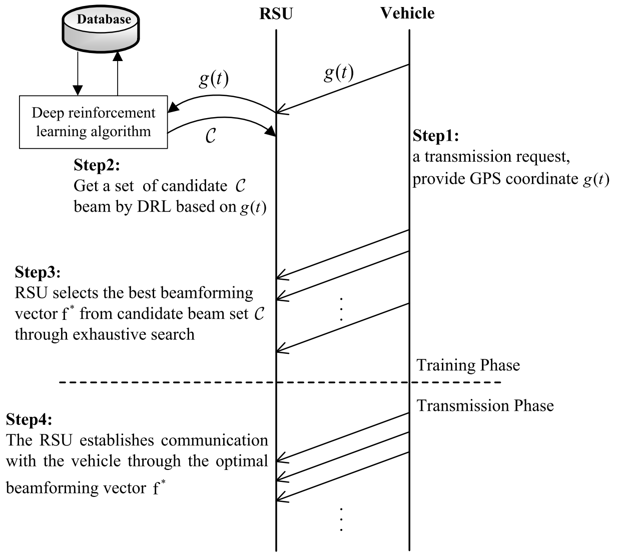

Database
RSU
Vehicle
$g(t)$
$g(t)$
Deep reinforcement learning algorithm
$\mathcal{C}$
Step1:
a transmission request, provide GPS coordinate $g(t)$
Step2:
Get a set of candidate $\mathcal{C}$ beam by DRL based on $g(t)$
Step3:
RSU selects the best beamforming vector f* from candidate beam set $\mathcal{C}$ through exhaustive search
Training Phase
Transmission Phase
Step4:
The RSU establishes communication with the vehicle through the optimal beamforming vector f*

FIGURE 3. The process of beam alignment based on the position information.

And $\lceil x \rceil$ is a ceiling function. The function $\rho(\mathbf{g})$ maps the coordinate $\mathbf{g} \in \bar{\mathcal{G}}$ into the discrete position label $\mathbf{p}$, that is expressed as

$$\mathbf{p} = \rho(\mathbf{g}) = \left(1 + \left\lfloor \frac{g_x - X_0}{\Delta_x} \right\rfloor, 1 + \left\lfloor \frac{g_y - Y_0}{\Delta_y} \right\rfloor\right), \tag{8}$$

where $\lfloor x \rfloor$ represents the floor operation. The value of Δ_x should be chosen based on the beamwidth of the beamforming vector, and the value of Δ_y is the width of each lane. Therefore, when the vehicle is in the discretized position, the maximum spectral efficiency is expressed as

$$R^p = \max_{f\in\mathcal{F}} \log_2\left(1+\frac{P_t}{\sigma^2+P_{ICI}}\left|H^H f\right|^2\right). \tag{9}$$

C. DEEP REINFORCEMENT LEARNING

The mathematical reinforcement learning theory is based on the Markov Decision Process (MDP) [33]. An MDP is abstracted by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P}, \gamma)$ [34], where $\mathcal{S}$ represents the state space, $\mathcal{A}$ denotes the action space, $\mathcal{R}$ denotes the reward function, $\mathcal{P}$ represents state transition probability, and γ represents discount factor. At each time t, the agent observes the state s_t and performs an action a_t based on the current state and strategy. Then the agent receives an immediate reward r_t and transits a new state s_{t+1}. The objective of the agent in the DRL is to choose the suitable action by continuously interacting with the environment, ultimately maximizing the cumulative discounted reward value, that is expressed as

$$R_t = r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \ldots = \sum_{k=0}^{\infty} \gamma^k r_{t+k+1}, \tag{10}$$

where $\gamma \in [0, 1]$ is the discount factor. The action value function $Q^\pi(s, a)$ is defined as the expected reward after taking action a_t at state s_t given the strategy π, that is expressed as

$$Q^{\pi}(s,a) = \mathbb{E}\left[R_t | s_t = s, a_t = a, \pi\right]. \tag{11}$$

Based on the above definition, solving the MDP is to find the optimal strategy π^* that maximizes the action value function $Q^{\pi}(s,a)$, i.e., $\pi^* = \arg\max_a Q^{\pi^*}(s,a)$. Thus, the optimal action value function is expressed as

$$Q^{\pi^*}(s,a) = \max_{\pi} Q^{\pi}(s,a), s \in \mathcal{S}, a \in \mathcal{A}. \tag{12}$$

In the DQN algorithm, the weight and bias parameters θ of a fully connected layer (often used in combination with a convolutional layer) are utilized to approximate the optimal action value function [35]. Specifically, $Q(s,a;\theta) \approx Q^{\pi^*}(s,a)$. The same action value function is used to calculate the target value when selecting and evaluating actions, so DQN have the overestimation problem. Therefore, Hasselt [36] proposed the DDQN algorithm to address the overestimation problem. The main idea of DDQN is to separate the selection and evaluation of actions when calculating the target value. In the updating process, two networks learn θ of the main network and θ' of the target network, respectively. DDQN first selects the action corresponding to the maximum Q value in the current Q network, and calculate the target Q value by the selected action in the target network. The target Q value is expressed as

$$y_t = r_t + \gamma Q'(s_{t+1}, \arg\max_{a_{t+1}} Q(s_{t+1}, a_{t+1}; \theta); \theta'), \tag{13}$$

where Q denotes the current Q value, and Q' denotes the target Q value. θ is updated by minimizing the loss function $L(\theta)$ and it is expressed as

$$L(\theta) = \mathbb{E}[(y_t - Q(s_{t+1}, a; \theta))^2]. \tag{14}$$

RL and supervised learning are commonly used algorithms in beam alignment. The main difference is that RL emphasizes learning by interaction with the environment without explicit labels. However, supervised learning requires a large dataset with pre-labeled data, and may quickly decline in performance when the environment changes. RL is capable of dynamically adjusting beams through exploration and exploitation strategies to optimize communication quality, a feat that is difficult to achieve with supervised learning. DDQN reduces the overestimation error by using two Q-networks, that avoids large fluctuations during the learning process. It is particularly important for dynamic V2I communication environments. While existing deep learning methods are effective in some respects, DDQN offers unique advantages in dealing with dynamic environments, optimizing long-term performance, reducing reliance on large amounts of a priori knowledge, and improving accuracy and stability.

D. PROPOSED ALGORITHM

In this section, the problem of beam alignment is modeled as an MDP that is solved by DRL. The RSU acts as the agent, and the vehicle traffic is the environment sensed by the RSU. The RSU can perceive the current state of the environment, and act to obtain the corresponding reward. The learning goal is to maximize the reward obtained by the RSU. The state, action, and reward function are defined as

(1) State: For beam alignment based on vehicle position information, the RSU can obtain changes in its external environment through onboard sensors, such as GPS, to determine the vehicle position coordinate $g(t)$. Therefore, the state is defined as the location of the vehicle. At time t, the state can be represented as $s_t = \{\mathbf{p}(t)\}$.

(2) Action: At time t, the agent performs an action a_t in the action space $\mathcal{A}$ according to the strategy π, after obtaining the state s_t by the onboard environment sensor. It establishes the mapping between the vehicle position and the beam. Therefore, the action is defined as all possible simulated beamforming vectors at the RSU. At time t, the action can be expressed as $a_t = \left\{\mathbf{f}_1(t), \mathbf{f}_2(t), \ldots, \mathbf{f}_{N_t}(t)\right\}$.

(3) Reward function: To evaluate the effect of the selected action a_t, the system spectral efficiency is taken as the instantaneous reward returned, as shown in (15), at the bottom of the page. When the received SNR is greater than or equal to the SNR threshold th_{min}, the instantaneous reward is defined as the system spectral efficiency. Otherwise, it is equal to 0.

In the problem of beam alignment, the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are discrete. The DDQN algorithm is used to solve the problem. Fig. 4 shows the DDQN-based agent architecture, and how the agent interacts with the environment. To overcome the learning instability and reduce the correlation among training samples, an experience pool $\mathcal{D}$ is used to store the transitions (s_t, a_t, r_t, s_{t+1}). First, the next state s_{t+1} is employed by both the main network and target network to select the action and evaluate its value, respectively. Then, the discount factor γ and the reward r_t are used to calculate the target value y_t. Next, the loss function is calculated to update the main network parameters θ by the gradient descent method.

Algorithm 1 describes the process of the beam alignment neural networks based on DDQN. We first initialize the experience pool $\mathcal{D}$ for offline learning that shuffles data correlations and the action value main network with θ. Observe the initial state s_0 and select action $a_0 \sim \pi_\theta(s_0)$. At each time step, an action a_t is randomly selected by the $\varepsilon-$greedy strategy [37] with probability ε. With probability $1-\varepsilon$, the action $a_t = \arg\max_a Q^{\pi^*}(s_t, a)$ is selected. According to the environment feedback, the next state s_{t+1} and the corresponding reward value r_t are generated. Store (s_t, a_t, r_t, s_{t+1}) as a

$$r_t = \begin{cases} \log_2\left(1 + \frac{P_t}{\sigma^2 + P_{ICI}} \left|\mathbf{H}^H \mathbf{f}_i(t)\right|^2\right), 1 \leqslant i \leqslant N_t, & SNR \geqslant th_{\min} \\ 0, & SNR < th_{\min} \end{cases} \tag{15}$$

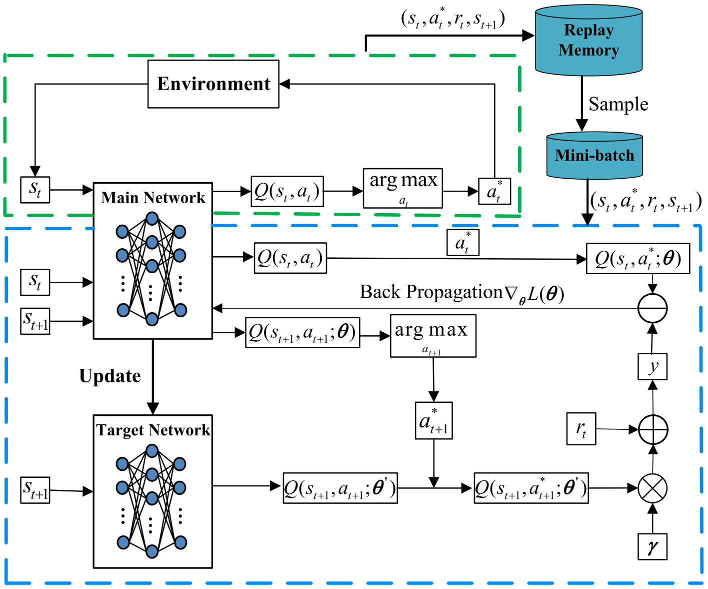

$(s_t, a_t^*, r_t, s_{t+1})$
Replay Memory
Sample
Mini-batch
Environment
s_t
Main Network
$Q(s_t, a_t)$
$\arg\max_{a_t}$
a_t^*
$(s_t, a_t^*, r_t, s_{t+1})$
s_t
s_{t+1}
$Q(s_t, a_t)$
a_t^*
$Q(s_t, a_t^*; \boldsymbol{\theta})$
Back Propagation $\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta})$
$Q(s_{t+1}, a_{t+1}; \boldsymbol{\theta})$
$\arg\max_{a_{t+1}}$
y
Update
a_{t+1}^*
r_t
Target Network
s_{t+1}
$Q(s_{t+1}, a_{t+1}; \boldsymbol{\theta}')$
$Q(s_{t+1}, a_{t+1}^*; \boldsymbol{\theta}')$
γ

FIGURE 4. **DDQN-based agent neural network architecture.**

Algorithm 1 Beam Alignment Neural Network Training Algorithm Based on DDQN

```
Require: mini-batch size $B$, learning rate $\alpha$, discount factor $\gamma$, the capacity of the experience pool $\mathcal{D}$, the number of episodes $J$, and the update step $C$ of target network parameter;
Ensure: the experience pool $\mathcal{D} = \emptyset$, initializes the action value function $Q^{\pi}(s, a)$ with random parameters $\theta$;
for episode $= 1, 2, \ldots, J$ do
    Observe initial state $s_0$ and select action $a_0 \sim \pi_{\theta}(s_0)$;
    for $t = 1, 2, \ldots, T$ do
        Select $a_t$ with probability $\varepsilon$, and select $a_t = \arg\max_{a} Q^{\pi^*}(s_t, a)$ with probability $1 - \varepsilon$;
        Perform action $a_t$, get reward $r_t$ according to (15) and observe the next state $s_{t+1}$;
        Store sample $(s_t, a_t, r_t, s_{t+1})$ into the experience pool $\mathcal{D}$;
    end for
    Mini-batch $(s_t, a_t, r_t, s_{t+1})$ with sample number $B$ is randomly selected from the experience pool $\mathcal{D}$;
    The loss function is calculated according to (14), to update the training parameter $\theta$;
    Update the target network parameters every $C$ steps, $\theta' = \theta$;
end for
```

set of data in the experience pool $\mathcal{D}$. When $\mathcal{D}$ is full, old data will be overwritten. After each network output, mini-batch samples with batch size B are randomly selected from $\mathcal{D}$ for training, and the main network of action value is updated. The loss function is calculated according to (14), to update θ by the gradient descent method. The target network is updated every C step, passing the main network parameter θ to the target network of the action value as the new parameter, i.e. $\theta' = \theta$.

V. PERFORMANCE EVALUATION

Section IV proposes a beam alignment algorithm based on DRL. The algorithm depends on the correlation among receiver, and transmitter positions, environment geometry, and beamforming direction. Therefore, generating actual channel data (AoA/AoD, path loss and delay, etc.) is essential. This section simulates the realistic mmWave channels using the commercial ray-tracing simulator WirelessInsite, and the publicly available DeepMIMO dataset [38]. The simulator is widely used in mmWave research [39], [40], and verified by channel measurements [41]. The DeepMIMO is a publicly and freely accessible data set generator on mmWave machine learning, relevant to a specific environment. Its framework consists of the ray-tracing scene 'R' and the data set parameter α. The ray-tracing scenario 'R' is the channel parameters (angle of arrival/departure, path gain, etc.) of the channel between each transmitter and receiver, obtained by WirelessInSite. The parameter α is a set of parameters generated by adjusting the system settings and antenna configuration. Based on the 'R' and α, the DeepMIMO will generate the channel matrix. By changing the parameters (number of antennas, system bandwidth, number of channel paths, etc.) and ray scenarios, we can get a data set for a specific environment. Fig. 5 depicts a bird's-eye view of part of a V2I ray-traced scene in DeepMIMO.

A. SIMULATION SETUP

In the mmWave vehicular network, the RSU provides communication service for vehicles in 28 GHz. The coordinate of the RSU is (50 m, 0 m), the height of antenna is 6 m, and a ULA is facing the street. The vehicle using full antennas is

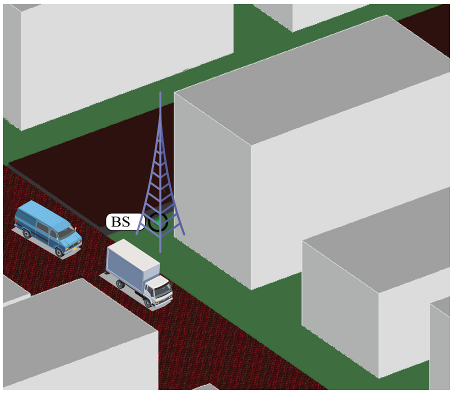


FIGURE 5. A bird's-eye view of part of a V2I ray-traced scene in DeepMIMO.

TABLE 2. Simulation parameters of beam alignment.

Parameter	Symbol	Value
Antenna array element spacing	d	$\lambda/2$
System bandwidth	W	500 MHz
Transmit power of the RSU	P_t	30 dBm
Antennas number of the RSU	N_t	(1,64,1)
Codebook size of the RSU	$\|\mathcal{F}\|$	(1,64,1)
Speed of the car	v	20 m/s
Noise power	σ^2	-114 dBm
Lane x axis spacing	Δx	1 m
Lane y axis spacing	Δy	3.75 m
SNR threshold	th_{min}	20 dB
Clusters number of channel	N_{cl}	1
Clusters number of per path	N_{ray}	3

located in the service area $\mathcal{G} = [0\ m, 100\ m] \times [0\ m, 7.5\ m]$ of the RSU. We set the SINR threshold at $th_{\text{min}} = 20$ dB [42], to avoid a low transmission rate. Moreover, Matlab is used to build the channel matrix between RSU and vehicles according to (2). Python is used to build the mmWave vehicles network environment, and TensorFlow is used to construct the DRL network based on the DDQN algorithm. It is worth noting that v =i 20 m/s is a common speed for cars, and the same speed is used in training and testing in this paper. This setup method is used in the related literature [43], [44]. BeamSteering codebook, derived by quantizing the angles of the antenna array response vectors, is a set of predefined beamforming vectors for forming beams in specific directions [45]. We utilized the BeamSteering codebook to design the beam codebook. Table 2 summarizes the system simulation parameter settings, and Table 3 summarizes the parameter settings of DDQN network. The antennas number of the RSU in Table 2, represent the number of antenna panels, and the number of antenna units in the horizontal and vertical dimensions [46], respectively. The codebook size of the RSU, indicate the number of groups in the codebook, and the size of the first and second dimensions of the codebook [47], respectively.

TABLE 3. Hyper-parameters for network training.

Parameter	Symbol	Value
Learning rate	α	0.002
Discount factor	γ	0.6
Experience pool capacity	D	131072
Mini-batch size	B	1024
Episode	J	10000
Target network update step	C	1000
The initial value of ε	ε_1	0.8
The final value of ε	ε_2	0.001

To better evaluate the performance of the proposed algorithm, comparisons are made with the following six beam search schemes:

- **Beam alignment based on exhaustive search.** When vehicles are located at different locations, the RSU first explores all available beams by the traditional exhaustive search, then selects the beam with the maximum spectral efficiency.
- **Beam alignment based on the random search.** When vehicles are located at different locations, the RSU randomly selects a beam.
- **Beam alignment based on hierarchical search [48].** The hierarchical search is divided into two stages. In the first stage, RSU performs an exhaustive search through four wide beams. After scanning the entire space, the optimal wide beam is determined according to the maximum spectral efficiency criterion. In the second stage, RSU further performs an exhaustive search on the best narrow beam in the obtained wide beam.
- **Beam alignment based on DQN search [49].** A set of candidate beams for a given vehicle location is first obtained by DQN algorithm. Next, the best beam from the candidate beam set is obtained by exhaustive search.
- **Beam alignment based on DDQN-PER search [24].** DDQN-PER incorporates the DDQN scheme, and optimizes its logical structure with the experience replay mechanism. Compared to DDQN, it gives samples different priorities by the prioritized experience.
- **Beam alignment based on Bayesian optimization (BO) [29].** The problem of beam alignment is considered a black box problem, employing Bayesian optimization to find the potentially optimal beam. By using a surrogate model, it can provide an approximation solution that is sufficiently close to the real system with a lower computational cost.

The performance evaluation index in this paper, includes the received SNR, the alignment probability, the effective achievable rate, and the effective achievable rate ratio, as follows

- **Received SNR.** When the vehicles are at different positions, the RSU communicates with the vehicles, using the corresponding beamforming vector $\mathbf{f}^*$ obtained from a beam alignment scheme. The received SNR at the vehicles is expressed as
$$SNR = 10\lg\left(\frac{P_t}{\sigma^2+P_{ICI}}\left|\mathbf{H}^H\mathbf{f}^*\right|^2\right). \tag{16}$$
- **Alignment probability.** The alignment probability is the ratio between the number of times the vehicle searches for the best beam at different locations and the total number of searches. Let c_i represents the number of predicted best beam obtained by a search algorithm, s_i denotes the number of best beam defined by (7a), and m represents the total number of beam searches. The alignment probability is expressed as
$$P = \frac{1}{m}\sum_{i=1}^{m}\mathbb{1}(c_i = s_i). \tag{17}$$
- **Effective achievable rate.** It depends on the time cost of searching the beamforming vector and the achievable rate with $\mathbf{f}^*$, which can be expressed as
$$R_{eff} = \left(1-\frac{T_{\rm tr}}{T_B}\right)\log_2\left(1+\frac{P_t}{\sigma^2+P_{ICI}}\left|\mathbf{H}^H\mathbf{f}^*\right|^2\right), \tag{18}$$
where $T_{\rm tr}$ represents the beam training time, and the beam search cost of the search algorithm is $T_{\rm tr} = N_{tr}T_p$. $N_{\rm tr}$ represents the number of training pilots, and $\mathbf{f}^*$ is chosen from the candidate beam set $\mathcal{C}$, so $N_{tr} = |\mathcal{C}|$. Moreover, T_p represents the beam training pilot sequence time, that is the scanning time of a single beam, and T_B denotes the beam coherence time [50].
- **Effective achievable rate ratio.** It is defined as the ratio between the effective achievable rate and the transmission rate with the global optimal beamforming vector f^{best}, and it is expressed as
$$R_T = \frac{R_{eff}}{\log_2\left(1+\frac{P_t}{\sigma^2+P_{ICI}}\left|H^H f^{best}\right|^2\right)}. \tag{19}$$

B. PERFORMANCE COMPARISON

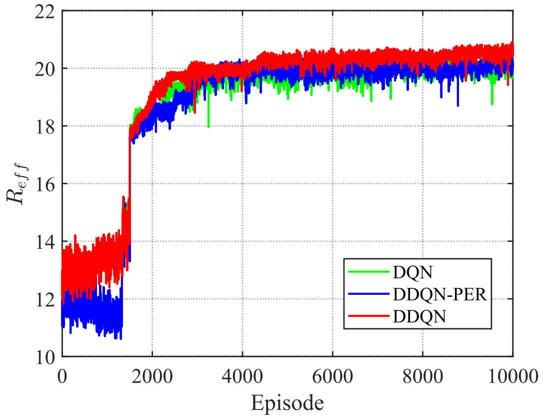


FIGURE 6. Learning performance analysis with different search algorithms.

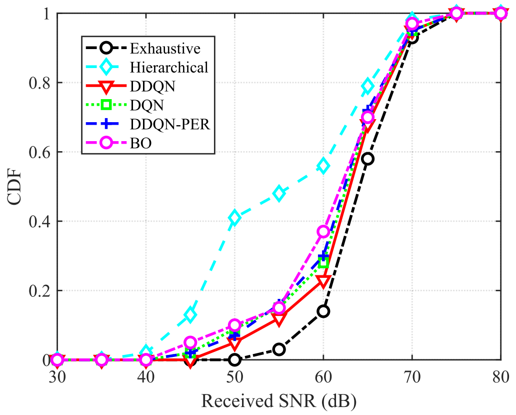


FIGURE 7. CDF of received SNR with different beam search schemes.

Fig. 6 illustrates the learning performance of the DDQN-based search scheme, DDQN-PER-based scheme, and DQN-based beam search scheme. As shown in Fig. 6, the R_{eff} of these schemes increase with episodes. At the beginning of training, the R_{eff} fluctuates continuously, as the agent is in the environment exploration phase, and randomly selects actions with the probability ε. It is found that the R_{eff} obtained by DDQN-PER at the beginning of training is smaller. Because DDQN-PER uses a preferential experience playback mechanism that will give more importance to useful samples. However, it does not facilitate the selection of samples, with small episodes and samples. With the training episodes, the performance of DDQN-PER will improve. With enough candidate beams, the performance of DDQN-PER is equal to DDQN. In addition, the performance of DDQN surpasses that of DQN with the overestimation problem, by introducing the target network for evaluating actions. As the training episodes, the agent's exploration of the environment decreases, leading to a rapid increase in R_{eff}. Meanwhile, the R_{eff} grows slower as less new information, at a later stage. Finally, these schemes start converging after 4000 episodes and gradually remain constant.

Fig. 7 shows the cumulative distribution function (CDF) for received SNRs with different beam search schemes, by calculating the received SNR at each position of the vehicle. It can be seen that the received SNR at all vehicle positions of these beam search schemes is greater than the SNR threshold, which enables the normal communication between the RSU and the vehicle. Compared with other schemes, the received SNR of the DDQN scheme is closest to the performance of the exhaustive search. Although DDQN-PER considers the preferential experience playback mechanism, it performs poorly with the small number of actions and candidate beams in this paper. DQN does not achieve better performance due to the overestimation problem. The performance of the BO is related to the choice and

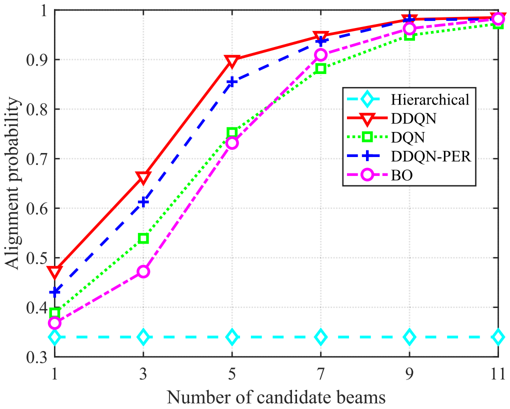


FIGURE 8. Alignment probability with different number of candidate beams.

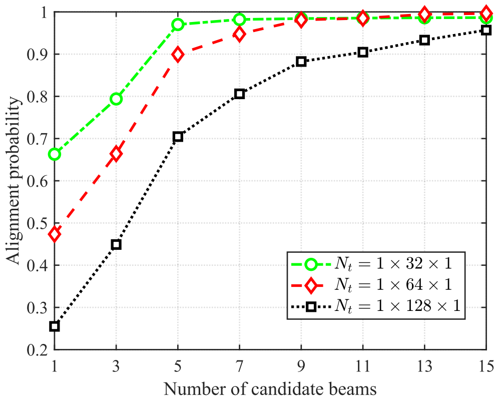


FIGURE 9. Alignment probability with different number of candidate beams and antennas.

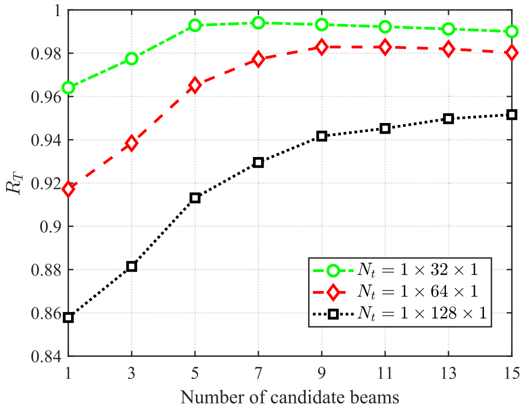


FIGURE 10. Effective achievable rate ratio R_T with different number of candidate beams and antennas.

parameter configuration of the surrogate model and the acquisition function. The surrogate model is a sensitivity issue regarding parameter selection, and the cost of each function evaluation is high. Therefore, the limited number of iterations leads to the inability to obtain the most suitable parameters. In addition, the acquisition function has a certain degree of uncertainty, with the drawback of falling into local optimal points easily. Therefore, the performance of the BO in this paper is not as good as that of the DDQN and DDQN-PER.

Fig. 8 shows the beam alignment probabilities with different numbers of candidate beams. As the random search scheme is unstable, the scheme is not used for comparison. It is found that the beam alignment probability of the hierarchical scheme is far lower than other search schemes. Because if the correct wide beam containing the global optimum is not selected in the first stage, then the beam chosen in the second stage will certainly not be the global optimum. The performance of hierarchical scheme remains unchanged with the candidate beams, that proves the obtained beamforming vector is only the local optimal rather than the global one in the multipath environment. As the number of candidate beams, the beam alignment probabilities of other search schemes increase. In the DDQN-based scheme, when the number of candidate beams is 5, the beam alignment probability reaches to 90%. When the number of candidate beams is 9, the beam alignment probability approaches 100%. At this point, the performance of DDQN-PER and BO is almost equal to that of DDQN, as the work of replay mechanism in DDQN-PER and the improving accuracy of the model parameters of BO, when the number of candidate beams is enough. However, the DDQN-PER and BO schemes do not perform as well as DDQN when the number of candidate beams is small.

Fig. 9 shows the beam alignment probabilities with different numbers of candidate beams and antennas. It can be found that the beam alignment probability of different antennas increases with the number of candidate beams, and gradually approaches 100%. Fig. 10 shows the effective achievable rate ratio R_T with the different number of candidate beams and antennas. It can be seen that in the DDQN scheme, with the increase of the candidate beams, the R_T of different antenna number first increases and then decreases. According to Fig. 9 and Fig. 10, when the number of candidate beams is 5, and the number of antennas is 32, the beam alignment probability is 97%, and R_T reaches the maximum value. When the number of candidate beams is greater than 5, although the beam alignment probability increases with candidate beams, the R_T decreases. Because the R_{eff} depends on the training overhead of searching beamforming vectors and the achievable rate of designed beamforming vectors. With the number of candidate beams, the training overhead also increases. Due to the physical adjacency of the beam, when the size of antenna array is large, the difference in the received power among the beams may be small. Even if the exact beam cannot be predicted, the second-best beam can be obtained to achieve communication, so the R_{eff} does not become too bad. It can conclude that the beam alignment probability is not the only metric to evaluate the beam alignment performance. Although the DDQN scheme cannot reach the same beam alignment probability as the exhaustive scheme, it can still accurately obtain the second-best beam. It is very important for the design of mmWave vehicular network communication, as the training overhead can be greatly reduced by

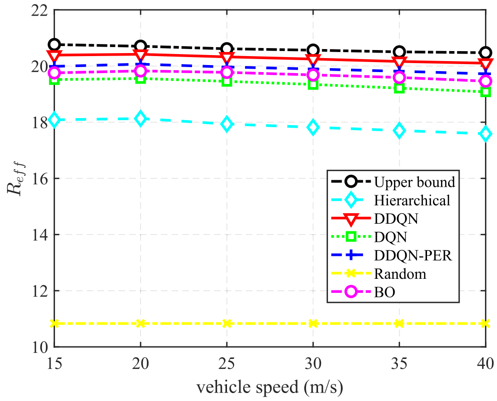


FIGURE 11. Effective achievable rate R_{eff} with different vehicle speeds.

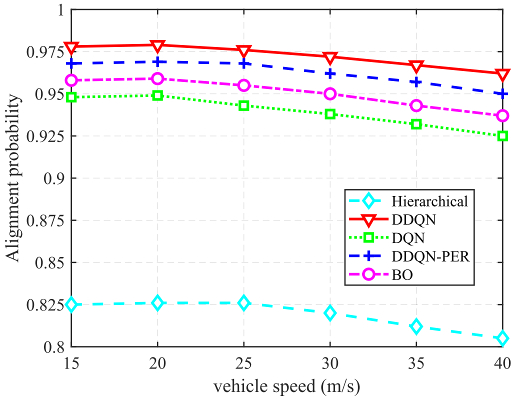


FIGURE 12. Alignment probability with different vehicle speeds.

sacrificing the best performance, thus improving the system effective achievable rate.

Fig. 11 shows the R_{eff} with different vehicle speeds. It can be seen that the R_{eff} obtained from these schemes except the random scheme, gradually decreases as the vehicle speed increases. Because the beam coherence time decreases as the vehicle speed increases. For the beam search schemes with higher training overhead, the shorter the beam coherence time, the faster the R_{eff} decreases. In addition, the ICI caused by Doppler effect also has a considerable impact on the beam alignment probability as the vehicle speed increases, as shown in Fig. 12. It can be noticed that the beam alignment probability is gradually decreasing with the increase of vehicle speed, due to the increasingly effect of ICI. When $N_{tr} = 1$, the best beam is searched for the first time, so $\frac{T_{tr}}{T_B}$ can be theoretically minimized. And the R_{eff} can reach up to the upper bound, as shown in Eq. (18). It can be found that the R_{eff} obtained from DDQN scheme is closest to the upper bound of the ideal transmission rate, compared to other schemes. When the speed of vehicle is 20 m/s, the R_{eff} of the DDQN search scheme achieves 97% of the ideal transmission rate, which is 89% better than the random scheme and 14% better than the hierarchical scheme. The R_{eff} weighs the beamforming training overhead against the achievable rate of the designed beamforming vector. For V2I scenarios, a slight improvement of performance may be more important than the additional computational cost, especially in critical or safety-related tasks, where decision quality is crucial.

VI. CONCLUSION

In mmWave vehicular networks, the optimal beam changes with the vehicle position, which brings a significant challenge to the V2I communication. Aiming at the problem of beam alignment in this paper, we proposed a beam alignment algorithm based on the vehicle position. The DDQN algorithm was used to train a set of candidate beams on the RSU, according to the criterion of maximizing the system spectral efficiency. Meanwhile, we consider the ICI caused by Doppler expansion, as well as utilizing DeepMIMO to obtain a dataset that is suitable for the V2I scenario. Finally, the simulation results show that the search algorithm proposed in this paper greatly improves the effective achievable rate and the beam alignment probability, outperforms other schemes. And it effectively balances the training overhead and achievable rate. In addition, we explore the impact on system performance with different vehicle speeds. It has been thoroughly demonstrated that the proposed algorithm is suitable for dynamic mmWave V2I scenarios. In the next step, we will employ a two-stage RL algorithm with offline pre-training and online real-time updating, to ensure effective performance in the dynamic V2I environment. Also, we will focus on issues such as concept drift due to time-varying environments and channels of V2I, to assess ML-based methods.

REFERENCES

[1] W. Duan, J. Gu, M. Wen, G. Zhang, Y. Ji, and S. Mumtaz, "Emerging technologies for 5G-IoV networks: Applications, trends and opportunities," *IEEE Netw.*, vol. 34, no. 5, pp. 283–289, Sep. 2020.

[2] A. Asadi, S. Müller, G. H. Sim, A. Klein, and M. Hollick, "FML: Fast machine learning for 5G mmWave vehicular communications," in *Proc. IEEE INFOCOM*, Honolulu, HI, USA, Apr. 2018, pp. 1961–1969.

[3] T. Z. H. Ernest and A. S. Madhukumar, "Ensemble learning-based edge caching strategies for Internet of Vehicles: Outage and finite SNR analysis," *IEEE Open J. Commun. Soc.*, vol. 4, pp. 239–252, 2023.

[4] B. Ai, A. F. Molisch, M. Rupp, and Z.-D. Zhong, "5G key technologies for smart railways," *Proc. IEEE*, vol. 108, no. 6, pp. 856–893, Jun. 2020.

[5] I. P. Roberts, J. G. Andrews, H. B. Jain, and S. Vishwanath, "Millimeter-wave full duplex radios: New challenges and techniques," *IEEE Wireless Commun.*, vol. 28, no. 1, pp. 36–43, Feb. 2021.

[6] Z. Xiao, T. He, P. Xia, and X.-G. Xia, "Hierarchical codebook design for beamforming training in millimeter-wave communication," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3380–3392, May 2016.

[7] M. Giordani, M. Polese, A. Roy, D. Castor, and M. Zorzi, "A tutorial on beam management for 3GPP NR at mmWave frequencies," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 1, pp. 173–196, 1st Quart., 2019.

[8] M. Gao et al., "Dynamic mmWave beam tracking for high speed railway communications," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, Apr. 2018, pp. 278–283.

[9] K. Satyanarayana, M. El-Hajjar, A. A. M. Mourad, and L. Hanzo, "Deep learning aided fingerprint-based beam alignment for mmWave vehicular communication," *IEEE Trans. Veh. Technol.*, vol. 68, no. 11, pp. 10858–10871, Nov. 2019.

[10] V. Raj, N. Nayak, and S. Kalyani, "Deep reinforcement learning based blind mmWave MIMO beam alignment," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8772–8785, Oct. 2022.

[11] Y. Wang, T. Zhang, S. Mao, and T. S. Rappaport, "Directional neighbor discovery in mmWave wireless networks," *Digit. Commun. Netw.*, vol. 7, no. 1, pp. 1–15, Feb. 2021.

[12] J. Zhang, Y. Huang, Q. Shi, J. Wang, and L. Yang, "Codebook design for beam alignment in millimeter wave communication systems," *IEEE Trans. Commun.*, vol. 65, no. 11, pp. 4980–4995, Nov. 2017.

[13] Y. Han and J. Lee, "Two-stage compressed sensing for millimeter wave channel estimation," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 860–864.

[14] M. E. Rasekh, Z. Marzi, Y. Zhu, U. Madhow, and H. Zheng, "Noncoherent mmWave path tracking," in *Proc. 18th Int. Workshop Mobile Comput. Syst. Appl.*, New York, NY, USA, Feb. 2017, pp. 13–18.

[15] N. González-Prelcic, R. Méndez-Rial, and R. W. Heath, "Radar aided beam alignment in mmWave V2I communications supporting antenna diversity," in *Proc. Inf. Theory Appl. Workshop (ITA)*, Mar. 2016, pp. 1–7.

[16] A. Klautau, N. González-Prelcic, and R. W. Heath, "LiDAR data for deep learning-based mmWave beam-selection," *IEEE Wireless Commun. Lett.*, vol. 8, no. 3, pp. 909–912, Jun. 2019.

[17] M. Hashemi, C. E. Koksal, and N. B. Shroff, "Out-of-band millimeter wave beamforming and communications to achieve low latency and high energy efficiency in 5G systems," *IEEE Trans. Commun.*, vol. 66, no. 2, pp. 875–888, Feb. 2018.

[18] A. Ali, N. González-Prelcic, and R. W. Heath, "Millimeter wave beam-selection using out-of-band spatial information," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 1038–1052, Feb. 2018.

[19] V. Va, J. Choi, T. Shimizu, G. Bansal, and R. W. Heath, "Inverse multipath fingerprinting for millimeter wave V2I beam alignment," *IEEE Trans. Veh. Technol.*, vol. 67, no. 5, pp. 4042–4058, May 2018.

[20] Y. Wang, A. Klautau, M. Ribero, A. C. K. Soong, and R. W. Heath, "mmWave vehicular beam selection with situational awareness using machine learning," *IEEE Access*, vol. 7, pp. 87479–87493, 2019.

[21] T.-H. Chou, N. Michelusi, D. J. Love, and J. V. Krogmeier, "Fast position-aided MIMO beam training via noisy tensor completion," *IEEE J. Sel. Topics Signal Process.*, vol. 15, no. 3, pp. 774–788, Apr. 2021.

[22] T. Shimizu, V. Va, G. Bansal, and R. W. Heath, "Millimeter wave V2X communications: Use cases and design considerations of beam management," in *Proc. Asia–Pacific Microw. Conf. (APMC)*, Nov. 2018, pp. 183–185.

[23] W. Attaoui, K. Bouraqia, and E. Sabir, "Initial access & beam alignment for mmWave and terahertz communications," *IEEE Access*, vol. 10, pp. 35363–35397, 2022.

[24] L. Wang, B. Ai, Y. Niu, M. Gao, and Z. Zhong, "Adaptive beam alignment based on deep reinforcement learning for high speed railways," in *Proc. IEEE 95th Veh. Technol. Conference: (VTC-Spring)*, Helsinki, Finland, Jun. 2022, pp. 1–6.

[25] Narengerile, J. Thompson, P. Patras, and T. Ratnarajah, "Deep reinforcement learning-based beam training with energy and spectral efficiency maximisation for millimetre-wave channels," *EURASIP J. Wireless Commun. Netw.*, vol. 2022, no. 1, p. 110, Nov. 2022.

[26] A. Klautau, P. Batista, N. González-Prelcic, Y. Wang, and R. W. Heath, "5G MIMO data for machine learning: Application to beam-selection using deep learning," in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2018, pp. 1–9.

[27] Q. Xian, A. Doufexi, and S. Armour, "mmWave vehicular beam alignment leveraging online learning," in *Proc. IEEE 97th Veh. Technol. Conf. (VTC-Spring)*, Florence, Italy, Jun. 2023, pp. 1–5.

[28] I. Orikumhi, J. Kang, and S. Kim, "Location-aided window based beam alignment for mmWave communications," in *Proc. 7th IEEE Int. Conf. Netw. Intell. Digit. Content (IC-NIDC)*, Nov. 2021, pp. 215–219.

[29] S. Yang, B. Liu, Z. Hong, and Z. Zhang, "Bayesian optimization-based beam alignment for mmWave MIMO communication systems," in *Proc. IEEE 33rd Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Kyoto, Japan, Sep. 2022, pp. 825–830.

[30] U. Demirhan and A. Alkhateeb, "Radar aided 6G beam prediction: Deep learning algorithms and real-world demonstration," in *Proc. IEEE Wireless Commun. Netw. Conf. (WCNC)*, Apr. 2022, pp. 2655–2660.

[31] O. E. Ayach, S. Rajagopal, S. Abu-Surra, Z. Pi, and R. W. Heath, "Spatially sparse precoding in millimeter wave MIMO systems," *IEEE Trans. Wireless Commun.*, vol. 13, no. 3, pp. 1499–1513, Mar. 2014.

[32] K. Guan et al., "Channel sounding and ray tracing for intrawagon scenario at mmWave and sub-mmWave bands," *IEEE Trans. Antennas Propag.*, vol. 69, no. 2, pp. 1007–1019, Feb. 2021.

[33] M. Liu, G. Feng, L. Cheng, and S. Qin, "A deep reinforcement learning based adaptive transmission strategy in space-air-ground integrated networks," in *Proc. IEEE Int. Conf. Commun.*, May 2022, pp. 4697–4702.

[34] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[35] H. Zhang, M. Huang, H. Zhou, X. Wang, N. Wang, and K. Long, "Capacity maximization in RIS-UAV networks: A DDQN-based trajectory and phase shift optimization approach," *IEEE Trans. Wireless Commun.*, vol. 22, no. 4, pp. 2583–2591, Apr. 2023.

[36] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double Q-learning," in *Proc. AAAI Conf. Artif. Intell.*, Mar. 2016, vol. 30, no. 1, pp. 2094–2100.

[37] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5141–5152, Nov. 2019.

[38] A. Alkhateeb, "DeepMIMO: A generic deep learning dataset for millimeter wave and massive MIMO applications," in *Proc. Inf. Theory Appl. Workshop (ITA)*, San Diego, CA, USA, Feb. 2019, pp. 1–8.

[39] M. Alrabeiah and A. Alkhateeb, "Deep learning for mmWave beam and blockage prediction using Sub-6 GHz channels," *IEEE Trans. Commun.*, vol. 68, no. 9, pp. 5504–5518, Sep. 2020.

[40] J. Guo, C.-K. Wen, S. Jin, and G. Y. Li, "Overview of deep learning-based CSI feedback in massive MIMO systems," *IEEE Trans. Commun.*, vol. 70, no. 12, pp. 8017–8045, Dec. 2022.

[41] P. Wu and J. Cheng, "Deep unfolding basis pursuit: Improving sparse channel reconstruction via data-driven measurement matrices," *IEEE Trans. Wireless Commun.*, vol. 21, no. 10, pp. 8090–8105, Oct. 2022.

[42] L. Su et al., "Content distribution based on joint V2I and V2V scheduling in mmWave vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 71, no. 3, pp. 3201–3213, Mar. 2022.

[43] J. Xu, B. Ai, L. Chen, Y. Cui, and N. Wang, "Deep reinforcement learning for computation and communication resource allocation in multiaccess MEC assisted railway IoT networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 12, pp. 23797–23808, Dec. 2022.

[44] Z. Ling, F. Hu, T. Liu, Z. Jia, and Z. Han, "Hierarchical deep reinforcement learning for self-powered monitoring and communication integrated system in high-speed railway networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 6, pp. 6336–6349, Feb. 2023.

[45] W. Wu, D. Liu, Z. Li, X. Hou, and M. Liu, "Two-stage 3D codebook design and beam training for millimeter-wave massive MIMO systems," in *Proc. IEEE 85th Veh. Technol. Conf. (VTC Spring)*, Sydney, NSW, Australia, Jun. 2017, pp. 1–7.

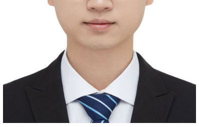
[46] D. A. U. Villalonga, H. OdetAlla, M. J. Fernández-Getino García, and A. Flizikowski, "Spectral efficiency of precoded 5G-NR in single and multi-user scenarios under imperfect channel knowledge: A comprehensive guide for implementation," *Electronics*, vol. 11, no. 24, p. 4237, Dec. 2022.

[47] *Physical Layer Procedures for Data*, Standard TS 38.214, 3GPP, Technical Specification, 2022.

[48] M. Giordani, M. Mezzavilla, C. N. Barati, S. Rangan, and M. Zorzi, "Comparative analysis of initial access techniques in 5G mmWave cellular networks," in *Proc. Annu. Conf. Inf. Sci. Syst. (CISS)*, Mar. 2016, pp. 268–273.

[49] P. Susarla, B. Gouda, Y. Deng, M. Juntti, O. Silvén, and A. Tölli, "Learning-based beam alignment for uplink mmWave UAVs," *IEEE Trans. Wireless Commun.*, vol. 22, no. 3, pp. 1779–1793, Mar. 2023.

[50] A. Alkhateeb, S. Alex, P. Varkey, Y. Li, Q. Qu, and D. Tujkovic, "Deep learning coordinated beamforming for highly-mobile millimeter wave systems," *IEEE Access*, vol. 6, pp. 37328–37348, 2018.

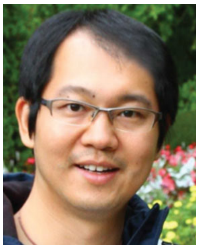


YUANYUAN QIAO (Graduate Student Member, IEEE) received the B.S. degree in communication engineering from North China Electric Power University, China, in 2019, and the M.Eng. degree in electronics and communication engineering from Beijing Jiaotong University, Beijing, China, in 2021, where he is currently pursuing the Ph.D. degree with the School of Electronic and Information Engineering. He is also a Visiting Student with Singapore University of Technology and Design. His current research interests include wireless resource allocation, ultra-reliable low-latency communications, and high-speed railroad communication.

YONG NIU (Senior Member, IEEE) received the B.E. degree in electrical engineering from Beijing Jiaotong University, China, in 2011, and the Ph.D. degree in electronic engineering from Tsinghua University, Beijing, China, in 2016. From 2014 to 2015, he was a Visiting Scholar with the University of Florida, Gainesville, FL, USA. He is currently an Associate Professor with the School of Electronic and Information Engineering, Beijing Jiaotong University. His research interests include networking and communications, including millimeter wave communications, device-to-device communication, medium access control, and software-defined networks. He has served as a Technical Program Committee Member for IWCMC 2017, VTC2018-Spring, IWCMC 2018, INFOCOM 2018, and ICC 2018. He was the Session Chair for IWCMC 2017. He was a recipient of the Ph.D. National Scholarship of China in 2015, the Outstanding Ph.D. Graduates and Outstanding Doctoral Thesis of Tsinghua University in 2016, the Outstanding Ph.D. Graduates of Beijing in 2016, the Outstanding Doctorate Dissertation Award from the Chinese Institute of Electronics in 2017, and the 2018 International Union of Radio Science Young Scientist Award.

LAN SU was born in Guangdong, China, in 1996. She received the B.S. degree in communication engineering from Lanzhou Jiaotong University, Lanzhou, China, in 2019, and the M.S. degree from the School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, China. Her research interests include millimeter-wave wireless communications and VANETs.

SHIWEN MAO (Fellow, IEEE) received the Ph.D. degree in electrical and computer engineering from Polytechnic University Brooklyn, NY, USA, in 2004. He is currently the Samuel Ginn Professor and the Director of the Wireless Engineering Research and Education Center, Auburn University, Auburn, AL, USA. His research interests include wireless networks, multimedia communications, and smart grids. He was a recipient of the IEEE ComSoc TC-CSR Distinguished Technical Achievement Award in 2019, the NSF CAREER Award in 2010, and the 2004 IEEE Communications Society Leonard G. Abraham Prize in the Field of Communications Systems.



NING WANG (Member, IEEE) received the B.E. degree in communication engineering from Tianjin University, China, in 2004, the M.A.Sc. degree in electrical engineering from The University of British Columbia, Vancouver, BC, Canada, in 2010, and the Ph.D. degree in electrical engineering from the University of Victoria, Victoria, BC, Canada, in 2013. From 2004 to 2008, he was with China Information Technology Design and Consulting Institute, as a Mobile Communication System Engineer, specializing in the planning and design of commercial mobile communication networks, network traffic analysis, and radio network optimization. From 2013 to 2015, he was a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, The University of British Columbia. Since 2015, he has been with the School of Information Engineering, Zhengzhou University, Zhengzhou, China, where he is currently an Associate Professor. He also holds adjunct appointments with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, and the Department of Electrical and Computer Engineering, University of Victoria. His research interests include resource allocation and security designs of future cellular networks, channel modeling for wireless communications, statistical signal processing, and cooperative wireless communications. He has served on the technical program committees of international conferences, including the IEEE GLOBECOM, IEEE ICC, IEEE WCNC, and CyberC. He was one of the finalists of the Governor General's Gold Medal for Outstanding Graduating Doctoral Student with the University of Victoria in 2013.

ZHANGDUI ZHONG (Fellow, IEEE) received the B.E. and M.S. degrees from Beijing Jiaotong University, Beijing, China, in 1983 and 1988, respectively. He is currently a Professor and an Advisor of Ph.D. Students with Beijing Jiaotong University, where he is also a Chief Scientist of the State Key Laboratory of Advanced Rail Autonomous Operation. He is the Director of the Innovative Research Team, Ministry of Education, Beijing, and a Chief Scientist at the Ministry of Railways, Beijing. His research has been widely used in railway engineering, such as the Qinghai-Xizang Railway, the Datong-Qinhuangdao Heavy Haul Railway, and many high-speed railway lines in China. He has authored or co-authored seven books, five invention patents, and over 200 scientific research papers in his research area. His research interests include wireless communications for railways, control theory and techniques for railways, and GSM-R systems. He is an Executive Council Member of the Radio Association of China, Beijing, and the Deputy Director of the Radio Association, Beijing. He has received the Best Paper Award from IEEE ICC 2020 and the Best Paper Award from the IEEE TAOS Technical Committee in 2020. He was a recipient of the Mao Yisheng Scientific Award of China, the Zhan Tianyou Railway Honorary Award of China, and the Top 10 Science/Technology Achievements Award of Chinese Universities.

BO AI (Fellow, IEEE) received the M.S. and Ph.D. degrees from Xidian University, China. He studies as a Post-Doctoral Student at Tsinghua University. He was honored with the Excellent Post-Doctoral Research Fellow by Tsinghua University in 2007. He was a Visiting Professor with the Electrical Engineering Department, Stanford University, in 2015. He is currently with Beijing Jiaotong University as a Full Professor and a Ph.D. Candidate Advisor. He is also the Deputy Director of the School of Electronic and Information Engineering and the International Joint Research Center. He is one of the main people responsible for Beijing ''Urban Rail Operation Control System'' International Science and Technology Cooperation Base. He is also a member of the Innovative Engineering Based jointly granted by the Chinese Ministry of Education and the State Administration of Foreign Experts Affairs. He has authored/co-authored eight books and published over 300 academic research papers in his research area. He holds 26 invention patents. He has been the research team leader for 26 national projects. His research interests include the research and applications of channel measurement and channel modeling, dedicated mobile communications for rail traffic systems. He has been notified by the Council of Canadian Academies that, based on the Scopus database, he has been listed as one of the Top 1% authors in his field all over the world. He has also been feature interviewed by the *IET Electronics Letters*. He has received some important scientific research prizes.