Power Control for D2D Communication Underlaying Cellular Networks based on Deep Q-Learning and Fractional Frequency Reuse

Tingting Yang, Yingqi Zhao, Jin Jin, and Kaiyang Guo

***Abstract*—This paper considers device-to-device (D2D) communication underlaying cellular networks, where frequency resources are shared between D2D users and cellular users. When D2D users reuse the frequency resources occupied by the cellular users, interference could be produced among the two kinds of users. By means of the fractional frequency reuse approach, a traversal frequency reuse scheme is proposed, in which central D2D users in each cell reuse the frequency bands for neighbouring edge cellular users in sequence according to a counterclockwise direction. The proposed reuse scheme effectively minimizes the number of intra-cell interference links, leading to an improvement of the average sum rate. Subsequently, based on the proposed traversal reuse strategy, a deep Q-learning algorithm is implemented for power control. Simulation results demonstrate that the proposed power control algorithm outperforms other traditional methods in terms of sum rate.**

***Index Terms*—D2D, deep Q-learning, fractional frequency, power control**

I. Introduction

As the volume of data traffic in wireless communication networks continues to increase, traditional cellular networks are faced with higher demands for transmission rates. By facilitating direct communication between two adjacent devices [1], device-to-device (D2D) communication can enhance the transmission rate and reduce communication latency [2]. However, the reuse of frequency resources of cellular users for D2D communication leads to complex mutual interference between cellular and D2D users [3]. To mitigate this interference, resource allocation and power control strategies have been identified as effective techniques [4], [5]. In multi-cell cellular networks incorporating D2D communication, fractional frequency reuse (FFR) is applied to effectively reduce the inter-cell interference and enhance system capacity [6]. In [7], by introducing FFR, frequency bands are allocated based on users' locations, leading to the reduction of inter-cell interference among cell-edge users and mutual interference between D2D and cellular users. Furthermore, [8] proposes the integration of FFR with fractional power control (FPC) to optimize frequency resources and power control in a coordinated manner.

In traditional reinforcement learning (RL), value functions or policies are conventionally maintained in tabular form, which inherently suffers from the curse of dimensionality [9]. With the introduction of deep learning, deep neural networks have demonstrated great potential in addressing complex environments and dimensionality problem. By combining RL and DNN, deep reinforcement learning (DRL) has emerged. By taking advantage of the strengths of deep neural networks (DNN) to optimize the training process, DRL significantly enhances both the learning efficiency and the algorithm performance [10]. In [10], an intelligent power allocation algorithm based on DRL is constructed, which offers advantages such as lower computational complexity, faster convergence and higher data rate compared to traditional power control algorithms.

In a single-cell communication system, factors such as channel variations, user mobility, and signal interference make DRL a suitable solution for power control. Reference [11] defines the transmission power and frequency channels of D2D user equipment (DUE) as resources. Then a centralized resource allocation scheme based on DRL is proposed using a multi-agent framework. The goal of the centralized resource allocation scheme is to maximize the average effective throughput of cellular user equipment (CUE) and DUE. Reference [12] tackles channel selection and power control with a distributed DRL approach in overlapping D2D networks. By modifying the classical DRL algorithm to adapt non-stationary multi-agent environments, the algorithm in [12] shows better scalability and lower time overhead compared with the conventional fractional programming (FP) algorithm. In reference [13], DRL is used to optimize the transmission power of CUE and DUE in an uplink cellular network. The base station acts as an agent with a DNN, which is designed as a fully connected network with zero bias to maximize system throughput.

However, in a single-cell network mentioned in the above paragraph, each user communicates with one base station, the cooperation between users and base stations is straightforward. By contrast, in a multi-cell network, the communication environment is more complex. There is interference between base stations and users in different cells, and thus cooperative strategies between base stations need to be carefully

Manuscript received November 6, 2024; revised April 11, 2025; approved for publication by Jeon, Yo-Seb, Division 4 Editor, May 23, 2025.

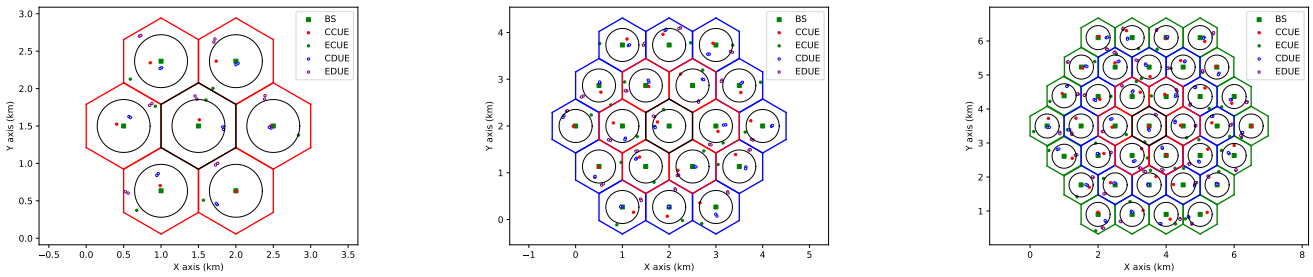
This paper was presented in part at the IEEE International Conference on Communication Software and Networks (ICCSN) in Shenyang, China, Nov. 2023. (*Corresponding author: J. Jin.*)

T. Yang, J. Jin and K. Guo are with the School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou, China, email: 2257329349@qq.com, iejjin@zzu.edu.cn, 18339396645@163.com.

Y. Zhao is with the Xinxiang Yellow River Conservancy Bureau, Xinxiang, China, email: zyq20210830@163.com.

J. Jin is the corresponding author.

Digital Object Identifier: 10.23919/JCN.2025.000036



(a) (b) (c)

Fig. 1. D2D cellular network models: (a) Single-layer model. (b) Two-layer model. (c) Three-layer model.

considered. References [14]–[16] study power control issues in multi-cell networks with DRL algorithm. Reference [14] treats each transmitter as an agent and provides a distributed dynamic power control scheme based on DRL. Each agent can estimate the impact of its actions on other users and then adjust its transmission power accordingly. Reference [15] presents a downlink sum-rate optimization scheme using Markov decision processes and trust region policy optimization. By providing flexible control of information exchange between base stations, robustness and high system throughput are achieved in [15]. This paper studies the multi-cell cellular network, where each cell contains both D2D pairs and cellular users. Since deep Q-learning (DQL) is mostly used for the DRL related works [17], the power control strategy used in this paper is DQL, which has been widely applied to various power allocation scenarios, such as general interference management in heterogeneous networks [18], [19], sum-rate maximization in cellular networks [9], [14].

In this paper, we propose a power control strategy for D2D communication underlaying cellular networks using DRL combined with FFR. To be specific, the proposed scheme first allocates frequency bands for users based on FFR and then employs DRL for power control. According to FFR, the cells are divided into central and edge regions, with CUEs and DUEs distributed within each region. A traversal frequency reuse scheme is proposed for central DUEs to avoid intra-cell interference between central DUEs and edge CUEs, resulting in the lowest number of intra-cell interference links. Simulation result demonstrates the superiority of the proposed traversal reuse scheme, because the impact of intra-cell interference on system rate performance outweighs that of inter-cell interference. After allocating frequency bands to CUE and DUE using the proposed reuse scheme, the transmission power of each user is adjusted using DQL algorithm. With the objective of maximizing the sum rate, neural networks are trained to learn the optimal power control strategy. Simulation results indicate that the proposed DQL power control strategy has better sum rate performance than the traditional power control methods.

The subsequent sections of this paper are structured as follows: Section II presents the system model; Section III discusses the frequency reuse scheme; Section IV introduces DQL and algorithm design; Section V shows the simulation results; finally, Section VI concludes our work and outlines future research directions.

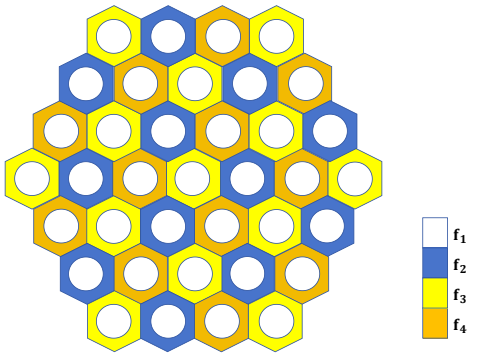


Fig. 2. FFR model.

II. System Model

In this paper, we consider D2D communication underlaying the uplink cellular networks, in which L layers are arranged around the central cell ($L \geq 1$). Fig. 1 illustrates single-layer, two-layer and three-layer cases for the system model. In each sub-figure of Fig. 1, central cell is depicted in black color, the first layer is in red color, the second layer is in blue color and the third layer is in green color. Each cell has a base station positioned at its centre. Furthermore, each cell is partitioned into central and edge regions, each having a CUE and a pair of DUEs. For ease of description, the central CUE, central DUE, edge CUE, and edge DUE are abbreviated as CCUE, CDUE, ECUE, and EDUE, respectively.

In order to mitigate the co-channel interference between CUEs and DUEs, the FFR model is introduced, as shown in Fig. 2. Based on the orthogonal frequency division multiplexing (OFDM) mode, we allocate frequency band resources for each CUE. The entire frequency band resources are divided into four orthogonal parts f_1, f_2, f_3, f_4. All CCUEs share the same frequency band f_1, while for ECUEs, the available frequency bands are f_2, f_3, f_4. Therefore, we can ensure that the frequency bands of CUEs between adjacent edge regions are orthogonal to each other, effectively reducing the inter-cell interference of ECUEs.

III. Frequency Reuse Scheme

A. Traverse Reuse Scheme

Based on the above FFR model, we first implement the frequency band resource division for CUEs. Next, we discuss the frequency band reuse problem of DUEs. Assuming that each CUE occupies only one frequency band, the frequency band of a CUE can only be reused by one DUE [20], and all frequency band resources are allocated and reused [8]. Then, the frequency band reuse scheme of DUEs is given as follows.

1) EDUEs: reuse the frequency band f_1 of the CCUEs in their own cells.

2) CDUEs: since the ECUEs of all adjacent cells use orthogonal bands, the CDUEs have to reuse the bands of ECUEs of the adjacent cells. However, employing a random frequency allocation approach may incur severe interference. For example, there is a possibility that some CDUEs may fail to use the frequency bands of neighbouring ECUEs, resulting in strong intra-cell interference between CDUEs and ECUEs.

In a single-layer cell model, the possible cases for random frequency reuse are limited, leading to a low number of interference links. This means that different frequency reuse schemes have little impact on system performance. Conversely, in a multi-layer cell scenario, random reuse schemes become more intricate and may cause a large number of intra-cell interference links, adversely impacting system performance.

An example of random reuse scheme for a three-layer cell model is illustrated in Fig. 3(a). In this sub-figure, cells are sequentially numbered, commencing from the center in a counterclockwise direction. The arrows signify the cells whose CDUEs select frequency band of neighbouring ECUEs. In this scheme, the CDUEs of cells numbered 20, 27, 30, and 34 (highlighted in red) have not chosen the frequency bands from their neighbouring cells. Conversely, cells numbered 6, 11, 25, and 32 (highlighted in blue) have not had the frequency bands of their ECUEs reused. Consequently, in order to guarantee that each cell's ECUE frequency band is reused only once by a certain CDUE, each of CDUEs in cells 20, 27, 30, and 34 can only randomly select a frequency band from the remaining cells numbered 6, 11, 25, and 32. By doing so, a CDUE and a ECUE in the same cell may share the same frequency band. For instance, the CDUE of cell 30 could reuse the frequency band of the ECUE of cell 6 (frequency band f_4 which is colored in orange), causing strong intra-cell interference. Therefore, adopting a random reuse scheme not only results in violations of the neighbouring cell reuse principle, but also introduces additional intra-cell interference links, leading to a reduction of the sum rate performance.

In response to the problem posed by random reuse, this paper introduces a traversal reuse scheme, as illustrated in Fig. 3(b). A three-layer cell scenario is used as an example to explain the proposed traversal reuse scheme. According to the arrows in Fig. 3(b), the CDUE in Cell 1 reuses the frequency band allocated to the edge region of Cell 2. This reuse pattern continues in counterclockwise direction, culminating at the CDUE in Cell which reuses the frequency band designated for the edge region of Cell 1. For layers beyond the first layer, the CDUE in each cell selects frequency bands of the ECUEs in adjacent cells situated on the same layer, proceeding in a counterclockwise direction. This reuse pattern forms a closed loop within each layer, ensuring that every CDUE reuses the frequency band assigned to the ECUE of a neighbouring cell.

The above traversal reuse scheme can ensure that the CDUE and ECUE within each cell employ orthogonal frequency bands, leading to the mitigation of co-channel interference. Furthermore, in multi-layer cell scenarios, random reuse schemes could introduce a large number of intra-cell interference links which exacerbate system performance. Hence, the advantage of the proposed traversal reuse scheme is prominent in multi-layer cell scenarios.

B. Interference Analysis

The previous subsection discussed the allocation of frequency resources for CUEs, and the reuse scheme for DUEs. After allocating frequency resources for all users, the following part will take the central cell (cell 1) as an example to conduct the analysis for users' interference. We discuss the issue of users' co-channel interference among cell 1 in four situations.

1) The CCUE: uses a frequency band, which is interfered with by the EDUE in cell 1, CCUEs and EDUEs in the CCUE's neighbourhood cells.

2) The EDUE : reuses frequency band is still f_1, and is interfered by the CCUE of cell 1, the CCUEs and marginal DUEs of the EDUE's surrounding neighbourhood cells.

3) The CDUE : reuses the frequency band f_4. The interference comes from the marginal CUEs and CDUEs of the CDUE's surrounding neighbourhood cells.

4) The ECUE : uses a frequency band f_3. The interference only comes from the CDUE of one ECUE's neighbouring cell (cell 7) that reuses frequency band.

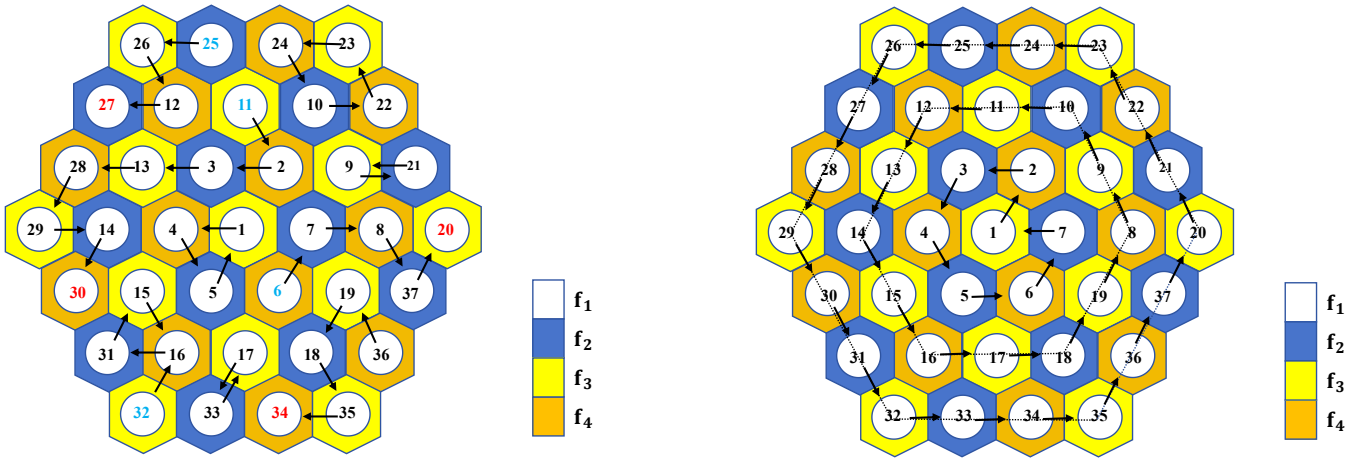
Below, we study the user power control optimization problem. Firstly, we need to establish the optimization objective. To facilitate the description of the following formulas, we abbreviate the four types of users, namely CCUEs, CDUEs, ECUEs, and EDUEs, as C_{cen}, D_{cen}, C_{edge}, and D_{edge}, respectively. Then take the CCUE in cell m as an example, its channel gain $g_{m,C_{\text{cen}}}$ can be expressed as

$$g_{m,C_{\text{cen}}} = |h_{m,C_{\text{cen}}}|^2 \beta_{m,C_{\text{cen}}}, \tag{1}$$

where the cell number $m \in \{1, 2, 3, \cdots, M\}$, and M is the total number of cells. Channel model includes both small-scale fading $h_{m,C_{\text{cen}}}$ and large-scale fading $\beta_{m,C_{\text{cen}}}$. Small-scale fading is modeled as flat and fast fading, with the envelope following the Rayleigh distribution. As in [21], large-scale fading $\beta_{m,C_{\text{cen}}}$ includes path loss and shadow fading, which is given by

$$\beta_{m,C_{\text{cen}}} = -120.9 - 37.6 \log_{10} d + 10 \log_{10} z, \tag{2}$$

where d represents the distance (km) from the transmitter to the receiver. The shadow fading factor z follows a log-normal distribution, the standard deviation is 4 dB. According to the above channel model and the cell numbering in Fig. 3(b), taking the users of central cell (cell 1) as an example, then



(a) (b)

Fig. 3. Frequency band reuse scheme of the CDUEs: (a) Random scheme. (b) Traverse reuse scheme.

the signal to interference plus noise ratio (SINR) of the CCUE can be represented as

$$\gamma_{1,C_{\text{cen}}} = \frac{p_{1,C_{\text{cen}}} g_{1,C_{\text{cen}}}}{I_{C_{\text{cen}}} + I_{D_{\text{edge}}} + \sigma^2}, \tag{3}$$

where $g_{1,C_{\text{cen}}}$ denotes the desired link gain from the CCUE of cell 1 to its base station receiver, $p_{1,C_{\text{cen}}}$ indicates the transmission power of the CCUE, $I_{C_{\text{cen}}}$ and $I_{D_{\text{edge}}}$ represent the interference are from CCUEs and EDUEs respectively, and σ^2 is the additive Gaussian white noise power. Similarly, the SINR of the EDUE can be written as

$$\gamma_{1,D_{\text{edge}}} = \frac{p_{1,D_{\text{edge}}} g_{1,D_{\text{edge}}}}{I_{C_{\text{cen}}} + I_{D_{\text{edge}}} + \sigma^2}. \tag{4}$$

The SINR of the CDUE can be stated as

$$\gamma_{1,D_{\text{cen}}} = \frac{p_{1,D_{\|\text{cen}}} g_{1,D_{\text{cen}}}}{I_{C_{\text{edge}}} + I_{D_{\text{cen}}} + \sigma^2}. \tag{5}$$

The SINR of the ECUE can be characterized as

$$\gamma_{1,C_{\text{edge}}} = \frac{p_{1,C_{\text{edge}}} g_{1,C_{\text{edge}}}}{I_{D_{\text{edge}}} + \sigma^2}. \tag{6}$$

With the normalized bandwidth, the sum rate can be expressed as

$$C^t = \sum_{m=1}^{M} \left(C_{m,C_{\text{cen}}} + C_{m,D_{\text{edge}}} + C_{m,D_{\text{cen}}} + C_{m,C_{\text{edge}}} \right). \tag{7}$$

In this formula,

$$C_{m,C_{\text{cen}}} = \log_2 \left(1 + \gamma_{m,C_{\text{cen}}}\right), \tag{8}$$

$$C_{m,D_{\text{edge}}} = \log_2 \left(1 + \gamma_{m,D_{\text{edge}}}\right), \tag{9}$$

$$C_{m,D_{\text{cen}}} = \log_2 \left(1 + \gamma_{m,D_{\text{cen}}}\right), \tag{10}$$

$$C_{m,C_{\text{edge}}} = \log_2 \left(1 + \gamma_{m,C_{\text{edge}}}\right). \tag{11}$$

The optimization goal is to maximize the sum rate by controlling the transmission powers of all users. The optimization problem is a non-convex and NP-hard problem, which is solved by the DQL algorithm below [9], [21]. It should be noted that all power control algorithms discussed in the following text are implemented after allocating frequency resources to all users.

TABLE I
THE NUMBERS OF INTERFERENCE LINKS FOR DIFFERENT FREQUENCY REUSE SCHEMES.

User type	CCUE	CDUE		ECUE		EDUE
Scheme		IRIL	IAIL	IRIL	IAIL	
Traversal scheme	307	107	0	87	0	307
Zigzag reuse scheme	307	120	1	90	1	307
Scheme 1	307	120	3	116	3	307
Scheme 2	307	119	5	116	5	307
Scheme 3	307	119	8	113	8	307
Scheme 4	307	0	37	0	37	307

C. Simulation Results of Traversal Scheme and Other Schemes

The preceding sections have discussed user interference and the advantage of the proposed traversal reuse scheme. To provide a clear illustration of the advantage of the proposed traversal scheme compared to random reuse schemes, simulations are performed on a three-layer model.

Table I provides the numbers of interference links for different frequency reuse schemes. For the traversal reuse scheme, the number of intra-cell interference links (IAIL) for CDUE and ECUE is 0, and the number of inter-cell interference links (IRIL) for CDUE and ECUE is 107 and 87 respectively. To describe how IAIL affect the performance of average sum rate, we select four specific random reuse scheme cases whose number of IAIL for CDUE (also for ECUE) is 3, 5, 8, 37 respectively, corresponding to scheme 1– 4 in

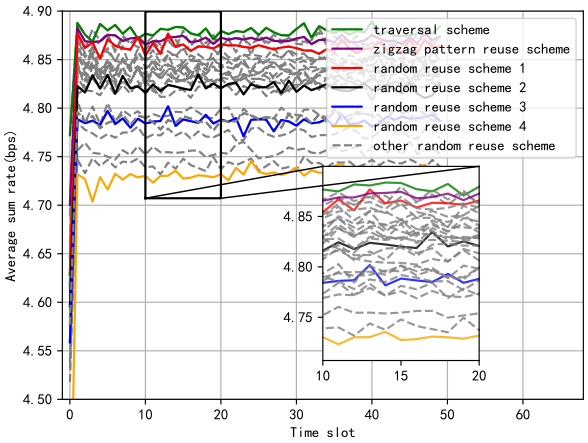


Fig. 4. Average rate of traverse scheme and other schemes.

Table I. Additionally, we also select a zigzag pattern scheme, corresponding to zigzag pattern reuse scheme in Table I whose number of IAIL for CDUE (also for ECUE) is 1. The above five schemes are based on the neighboring cell reuse principle, which means that CDUE reuses the frequency band of its neighboring cell. Scheme 4 is a special case that doesn't comply with the neighboring cell reuse principle. In scheme 4, CDUE reuses the frequency bands of the ECUE in its own cell, resulting in the maximum number of IAIL, both being 37 for CDUE and ECUE. In addition, for the CCUE and EDUE, no matter which scheme is used, the number of interference links remains unchanged due to the utilization of the same frequency band. After calculation, the number of total interference links for CCUE and EDUE is fixed at 307, as shown in Table I.

Fig. 4 illustrates the average sum rate of the traversal scheme, the zigzag frequency reuse scheme and random frequency reuse schemes. The gray dashed lines in this figure represent the average sum rates of other random reuse schemes different from scheme 1–4. It can be seen from Fig. 4 that the proposed traversal scheme shows the best performance.

On the other hand, scheme 4 exhibits the worst rate performance. Based on the observation on Table I, it's evident that the number of IAIL for the traversal scheme and scheme 4 exhibit the minimum and maximum value in this three-layer simulation model, respectively. Consequently, the average sum rates for other random reuse schemes are within the range defined by the above two schemes, as shown in Fig. 4. Since the average sum rate deteriorates as the number of IAIL increases, the proposed scheme with the minimum number of IAIL reaches the best rate performance.

In conclusion, among these frequency reuse schemes under the three-layer model, the traversal scheme yields the highest rate due to the avoidance of intra-cell interference between CDUE and ECUE, confirming the superiority of the traversal scheme. It is worth noting that the proposed traversal frequency reuse scheme remains applicable in multi-layer cell scenarios.

IV. Deep Q-Learning

This section first briefly introduces the theoretical basis of DQL algorithm and then introduces the design process when DQL algorithm is applied to the above model.

A. DQL Algorithm

Reinforcement learning can be described as a process in which an agent learns the optimal strategy by continuously interacting with the environment [22]. An interaction process can be described as follows. At time slot t, the agent observing the current environmental state s^t , $s^t \in S$ is the set of state space. Then, the agent performs the corresponding action a_t according to its strategy, $a^t \in A$, A is the set of action space. Subsequently, the environment enters the next state s^{t+1} and provides a timely reward to the agent r^t .

DQL algorithm is a combination of Q learning algorithm and deep learning, which approximates the Q-value function by introducing a deep neural network, called the deep Q network (DQN). It is an algorithm based on value learning. The agent aims to receive the maximum cumulative reward, which can be represented through the following equation [23]:

$$R^t = \sum_{k=0}^{\infty} \gamma^k r^{t+k+1}, \tag{12}$$

where the discount rate $\gamma \in [0, 1]$ is a discount for future rewards. So, under the determined strategy, we can use the action value function (Q function) to evaluate the quality of the agent's selection of action in the current state s. The Q function can be stated as [21]

$$Q^{\pi}(s, a; \theta) = \mathbb{E}[R^t \mid s^t = s, a^t = a], \tag{13}$$

where θ represents the parameters of DQN, and Q function denotes the expectation of future cumulative discount reward R^t. The input of DQN is the state information of the environment, and the output is the Q value for each action. The action to be executed is determined based on the ϵ-greedy strategy. In addition, to ensure the convergence of the training, a target network and experience replay mechanism are usually introduced [24], and the neural network parameters θ are updated using the gradient descent method.

B. The DQL Algorithm Design

Assuming that a central agent is responsible for obtaining the channel state information (CSI), the DQL algorithm framework for centralized training and distributed execution is employed [22]. Note that the imperfect CSI [17], [25] is not taken into account in this paper. Design the corresponding state, action, and reward as follows:

State: The state information serves as an input to the DQN, including the desired link gain and the interference link gain of each user. Due to the large fluctuations in the magnitude of the channel gains, the CSI is log-normalized [21]. In addition, we take the transmission power and the link transmission rate in the last time slot as supplementary state information.

Taking the CCUE in cell m as an example, the complete state information can be expressed as

$$s^t_{m,C_{\text{cen}}} = \left\{ I^t_{m,C_{\text{cen}}}, P^{t-1}_{m,C_{\text{cen}}}, C^{t-1}_{m,C_{\text{cen}}} \right\}, \tag{14}$$

where $I^t_{m,C_{\text{cen}}}$ represents the set of desired link gain of this user and the interference link gain caused by co-channel interference users. $P^{t-1}_{m,C_{\text{cen}}}$ denotes the transmission power set of this user and co-channel interfering users in the previous time slot. $C^{t-1}_{m,C_{\text{cen}}}$ represents the transmission rate set of this user and the co-channel interference users in the last time slot. Additionally, the base of state is equal to the number of DQN input neurons, and there is a total of $[I + 2(I + 1)]$ neurons.

Action: Since the action space of DQN must be discrete and limited, it is necessary to discretize the transmission power into *Y* feasible power levels. Therefore, there are *Y* neurons in the DQN output layer. The power set can be represented as

$$A = \left\{ 0, p_{\text{min}}, p_{\text{min}} + \frac{p_{\text{max}} - p_{\text{min}}}{Y - 2}, \cdots, p_{\text{max}} \right\}, \tag{15}$$

where p_{min} represents the minimum non-zero transmitting power of the user, and p_{max} represents the maximum transmitting power.

Reward: As the goal of DQL is to maximize the cumulative rewards, to maximize the sum rate, the instant reward function is defined as the sum rate, i.e.,

$$r^t = C^t. \tag{16}$$

V. Analysis of Simulation Results

A. Simulation Setup

The simulations in this section are based on the system model shown in Fig. 1, with a cell radius of $R = 500$. In addition, the ratio of the central area to the whole cell area $\beta = 1/2$. The DQN employed in the simulations consists of two hidden layers with 128 and 64 neurons respectively. The state vector is input into the first layer (the input layer), and the second and third layers are hidden layers. The activation function for the hidden layers is rectified linear unit (ReLU). The final layer outputs the Q-values for each action. During each training iteration, the loss function $L(\theta)$ is calculated by comparing the target Q-values generated by the target network with the estimated Q-values produced by the behavior network:

$$L(\theta) = \mathbb{E}\left[\left(y^t - Q(s_t, a_t; \theta_t]) \right)^2 \right]. \tag{17}$$

The DQN training is conducted over 10000 episodes, while testing is performed over 500 episodes, with each episode lasting for 50 time slots. After each episode, the positions of all users are randomly redistributed, while large-scale fading remains unchanged within an episode. The benchmark algorithms in the simulation include three algorithms. The first is the distance-based power control algorithm [26], named as DP algorithm. The transmit power of all users can be calculated using the following equation:

$$P(r) = P_{max} \left(\frac{r}{R_2} \right)^{\alpha\epsilon}, r \in [R_1, R_2], \tag{18}$$

TABLE II
Parameter configuration.

Name of parameter	Value
Noise power σ^2	-114 dBm
Distance range of CUE to BS $[r_{\min}, r_{\max}]$	[10 m, 500 m]
Distance range of D2D pairs$[r_{d_{\min}}, r_{d_{\max}}]$	[20 m, 100 m]
Minimum transmit power $p_{\min}$	3 dBm
Maximum transmit power $p_{\max}$	30 dBm
Path loss index α	3.5
Ratio of central area to the cell β	1/2
Standard deviation of Shadow fading	4 dBm
Discount factor γ	0.3
Learning rate	0.002
Discretized number *Y*	10
Memory size	100000

in which r is the distance of the transmitter and receiver, R_1, R_2 is the minimum and the maximum distance between transmitter and receiver, α is path loss index, ϵ is power control factor. The second benchmark algorithm is FPC algorithm [27], in which the transmitting power of users can be expressed as:

$$p = \min\left\{ P_{\text{max}}, P_0 + \eta PL \right\}. \tag{19}$$

In (19), PL includes path loss and shadow fading, and η is the compensation of PL. According to LTE standard, the range of η is $\eta \in \{0, 0.4, 0.5, 0.6, \cdots, 1.0\}$. The value of p_0 is set to -15 dBm. The third benchmark algorithm is FP algorithm [28]. The main parameters for simulations are listed in Table II.

B. The Proposed DQL Algorithm

This section begins by providing the average sum rate of the proposed algorithm based on a single-layer cell model. In the benchmarks presented in the above subsection, two algorithms' sum rates vary with the parameters: the power control factor in the DP algorithm and the compensatory factor in the FPC algorithm. In the following, these two parameters are dynamically adjusted within their permissible ranges to find the optimal values that maximize the sum rate for each algorithm, as shown in Figs. 5 and 6.

Fig. 5 illustrates the relationship between the average sum rate and power control factor ϵ in the DP algorithm. As ϵ gradually increases from 0 to 1, the average rate initially rises and then falls. From this figure, the maximum value is achieved when $\epsilon = 0.2$.

Fig. 6 depicts the relationship between the average sum rate and the compensatory factor η in the FPC algorithm. As η gradually increases within its permissible range, the average sum rate exhibits a trend of initially increasing and then decreasing. The maximum value is attained when $\eta = 0.7$.

In summary, by selecting $\epsilon = 0.2$ in the DP algorithm and $\eta = 0.7$ in the FPC algorithm, each algorithm achieves its optimal sum rate performance. Subsequently, these optimal

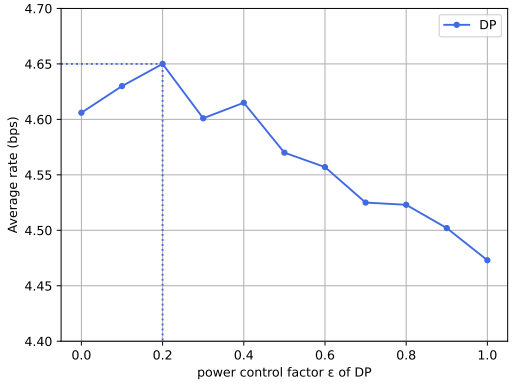


Fig. 5. Average rate of DP versus ϵ.

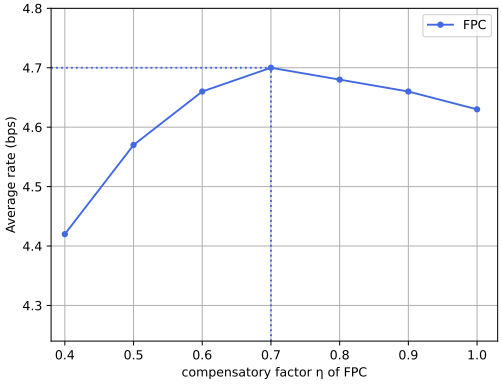
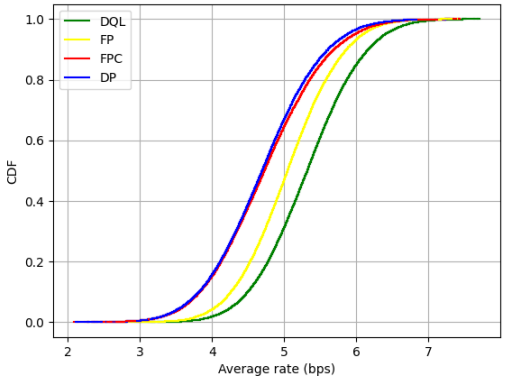
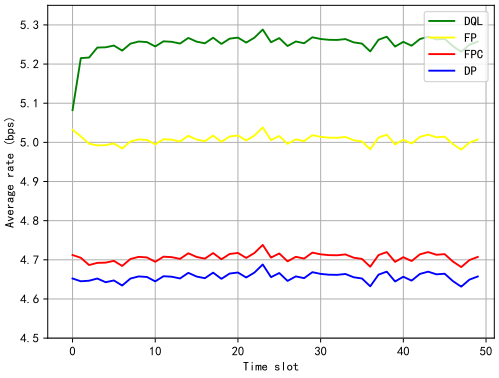


Fig. 6. Average rate of FPC versus η.



(a) CDF of average sum rate.



(b) Average sum rate versus time slot.

Fig. 7. Performance comparison for different algorithms.

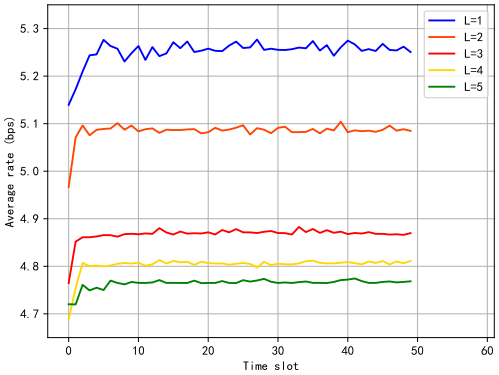


Fig. 8. The average rate of DQL with $L = 1 - 5$.

sum rates of the two algorithms are compared with the rate of the DQL algorithm.

Next, the proposed DQL algorithm is used for power control. After training the DQN, it is tested and compared with the aforementioned three traditional power control algorithms. Fig. 7(a) presents the comparison of the cumulative distribution functions (CDFs) of the average sum rate for different power control algorithms, and Fig. 7(b) illustrates the average sum rates as a function of time slots.

As shown in Fig. 7(a), the DQL algorithm achieves the highest average sum rate. From Fig. 7(b), the average sum rate of the DQL algorithm is higher than that of the FP, FPC, and DP algorithms. Now we discuss the performances of the other three algorithms. The average sum rate of the FP algorithm is higher than that of the DP and FPC algorithms. However, the FP algorithm converges to a local optimal power solution after multiple iterations, incurring a relatively high computational complexity. In contrast, the DP algorithm and the FPC algorithm have the advantage of low computational complexity, but their sum rate performances are much lower than those of the DQL and FP algorithms.

C. The Generalization Ability

The previous subsection confirmed that in a single-layer model, the proposed algorithm exhibits improved sum rate performance compared to the baseline algorithms. In this subsection, we first verify the generalization for the proposed algorithm in terms of the cell numbers, the maximum distance between D2D pairs, and the number of users. Next, we evaluate the execution time overhead of different power control algorithms at different cell sizes.

Fig. 8 presents the variation of average sum rate with respect to time slots for the proposed algorithm under different

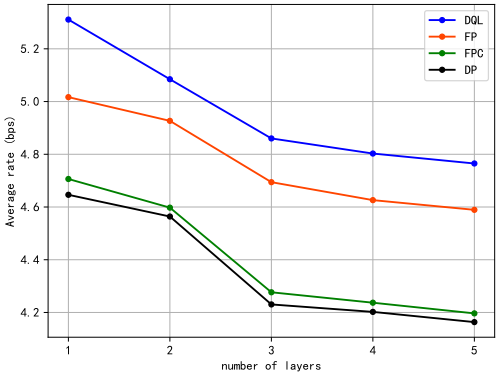


Fig. 9. The average rate of different algorithms with $L = 1 - 5$.

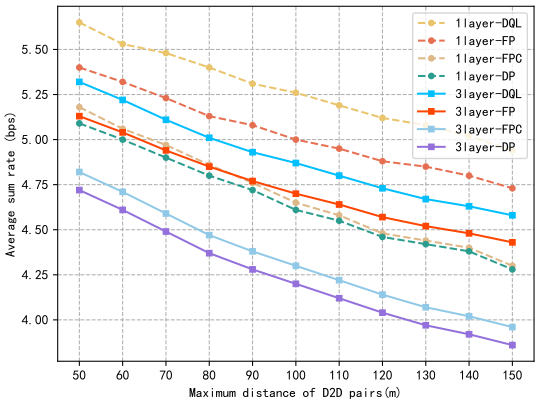


Fig. 10. The average sum rate versus the maximum distance between D2D pairs.

numbers of cell layers. As the number of cell layers increases, the number of users in the system grows, leading to an increase in the interference links among users. The more interference, the more reduction of sum rate. This conclusion can be confirmed from Fig. 8 that as the number of layers increases, the average rate decreases.

As shown in Fig. 9 , the average rates of all power control algorithms decline as the number of cell layers increases. It is clear that the proposed algorithm consistently outperforms other traditional algorithms. In summary, from Fig. 8 and Fig. 9 which contain the cell model with up to 5 layers , the proposed FFR combined with DQL power control algorithm exhibits good generalization in large-scale cell models.

Fig. 10 presents the average sum rate as a function of the maximum distance between D2D pairs. In both single-layer and three-layer cell models, the generalization capability of the proposed algorithm is verified. From this figure, as the maximum distance between D2D pairs gradually increases, the average sum rate of each algorithm decreases. It can be concluded that the proposed DQL algorithm always outperforms other three traditional algorithms with various distance situations for D2D pairs. In conclusion, the proposed algorithm has good generalization ability, in terms of the number of cells

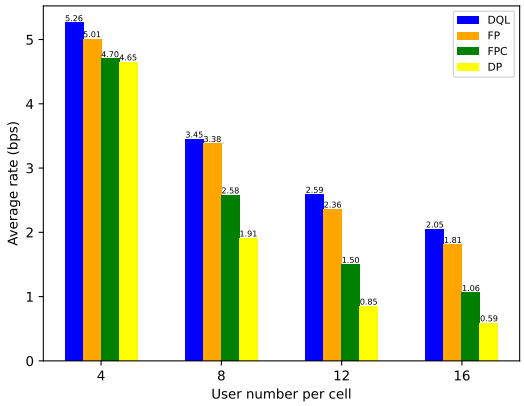


Fig. 11. The average sum rate versus the different number of users in a single cell.

TABLE III
THE TIME OVERHEAD OF THE DQL, FP, AND FPC ALGORITHM PER EXECUTION.

Algorithm	DQL	DP	FPC
Time cost for 1 layer (s)	$7.709e-4$	$3.989e-4$	$3.445e-4$
Time cost for 3 layers (s)	$4.132e-3$	$3.139e-3$	$3.053e-3$

as well as the maximum distance between D2D user pairs.

Fig. 11 illustrates the impact of the number of users in a single cell on the average sum rate. As shown in this figure, as the number of users increases, the average sum rate gradually decreases. This is because higher number of users incurs increased interference in the system, which leads to the reduction of the average sum rate.

D. Time Overhead

Time overhead is important for the practical deployment and application of the power control algorithm. Here we calculate the time overhead of the four algorithms mentioned above under single-layer model and three-layer model. The CPU of the simulation platform is i5-11500. The DP, FPC, and DQL algorithms do not need to acquire the global CSI, and thus they all use the distributed power execution with a parallel processing mechanism. Table III shows the time overhead for the DP, FPC, and DQL algorithms. Because the time overheads of the above three distributed algorithms are not affected by the total number of CUEs, their time overhead values remain unchanged with various amounts of CUEs. In contrast, FP algorithm needs to obtain global CSI and adopt a centralized execution mode, consuming a large amount of time overhead. Besides, the time overhead of the FP algorithm scales with the total number of users. The time overhead of the FP algorithm is shown in Table IV.

From Tables III and IV, it can be observed that the time overhead of the three distributed algorithms is relatively low. It's worth noting that, compared to the DQL algorithm, DP and FPC algorithms have lower time overhead due to their low computational complexity. However, their sum rates are obviously inferior to that of the DQL algorithm. Table IV

TABLE IV
THE TIME OVERHEAD OF FP ALGORITHM PER EXECUTION.

User number per cell	4	8	12	16
Time cost for 1 layer (s)	$6.791e-4$	$8.792e-4$	$1.027e-3$	$1.242e-3$
Time cost for 3 layers (s)	$4.526e-3$	$7.735e-3$	$1.502e-2$	$2.469e-2$

shows that the time overhead for the FP algorithm is low when the number of cell users is small. Because the FP algorithm executes centrally, the time overhead increases proportionally with the total number of users. As a result, when the total number of users is large, the high time overhead is an explicit disadvantage for the FP algorithm. This makes it difficult to deploy in practice and has poor generalization ability. In conclusion, the DQL algorithm can not only achieve the highest sum rate but also has a relatively low time overhead and strong generalization ability.

To sum up, the DQL algorithm outperforms the FP algorithm in terms of the sum rate in the single-layer cell model. Although this improvement is not prominent according to Fig. 7, the time cost of DQL algorithm is low compared to the FP algorithm. Especially in a large-scale network, the generalization ability of FP algorithm is significantly weaker than that of DQL algorithm. Compared to the FPC algorithm and the DP algorithm, the DQL algorithm has a slight increase in terms of time overhead, but there is a significant improvement in the sum rate, as can be seen in Fig. 7. It can be concluded that the DQL algorithm has a comprehensively excellent performance with respect to the sum rate and time cost.

VI. CONCLUSION

In this paper, by presenting a scheme that combines DQL power control with FFR resource allocation, we tackle the mutual interference between D2D links and cellular links in multi-layer uplink cellular networks with D2D communication. Based on FFR, we propose a traversal frequency reuse scheme that minimizes the number of IAIL. The traversal scheme first uses FFR to allocate frequency bands to CUE and then assign resources to DUE by complying with the traverse reuse scheme. Subsequently, the DQL algorithm is employed to control power. Simulation results demonstrate that the proposed DQL algorithm achieves better sum rate performance with low time overhead compared to the traditional algorithms. The soft frequency reuse (SFR) technology will be considered as a replacement for FFR in our future research work. Combining SFR with the DQL algorithm could be applicable to complex communication scenarios.

VII. CONFLICTS OF INTERNET

The authors declare that there are no conflicts of interest regarding the publication of this paper. I confirm there are no conflicts of interest for me and the co-authors.

REFERENCES

[1] X. Song, X. Han, and S. Xu, "Joint power control and channel assignment in D2D communication system," *J. Commun.*, vol. 14, no. 5, pp. 349–355, 2019.

[2] R. Wang, F. Jiang, and T. Xu, "A joint method of resource allocation and power control for device-to-device (D2D) communication," *Telecomm. Engin.*, vol. 56, no. 3, pp. 295–301, 2016.

[3] J. Kim *et al.*, "Spectrum allocation with power control in LBS based D2D cellular mobile networks," *J. Commun.*, vol. 15, no. 3, pp. 2374–4367, 2020.

[4] F. Malandrino, Z. Limani, C. Casetti, and C.-F. Chiasserini, "Interference-aware downlink and uplink resource allocation in HetNets with D2D support," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2729–2741, 2015.

[5] W. Wang, F. Zhang, and V. K. N. Lau, "Dynamic power control for delay-aware device-to-device communications," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 1, pp. 14–27, 2015.

[6] F. Jiang, X.-C. Wang, C.-B. Li, and B.-Y. Shen, "Dynamic power control based on FRR for D2D communication underlaying cellular networks," in *Proc. WCSP*, 2016.

[7] H. S. Chae, J. Gu, B.-G. Choi, and M. Y. Chung, "Radio resource allocation scheme for device-to-device communication in cellular networks using fractional frequency reuse," in *Proc. APCC*, 2011.

[8] Z. Zhang, R. Q. Hu, Y. Qian, and A. Papathanassiou, "D2D communication underlay in uplink cellular networks with fractional power control and fractional frequency reuse," in *Proc. IEEE GLOBECOM*, 2015.

[9] F. Meng, P. Chen, L. Wu, and J. Cheng, "Power allocation in multi-user cellular networks: Deep reinforcement learning approaches," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6255–6267, 2020.

[10] N. C. Luong *et al.*, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surv. Tuts.*, vol. 21, no. 4, pp. 3133–3174, 2019.

[11] S. Yu and J. W. Lee, "Deep reinforcement learning based resource allocation for D2D communications underlay cellular networks," *Sensors*, vol. 22, no. 23, p. 9459, 2022.

[12] J. Tan, Y.-C. Liang, L. Zhang, and G. Feng, "Deep reinforcement learning for joint channel selection and power control in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1363–1378, 2021.

[13] D. Ron and J.-R. Lee, "DRL-based sum-rate maximization in D2D communication underlaid uplink cellular networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 10, pp. 11121–11126, 2021.

[14] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *2019 IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, 2019.

[15] A. A. Khan and R. S. Adve, "Centralized and distributed deep reinforcement learning methods for downlink sum-rate optimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 12, pp. 8410–8426, 2020.

[16] C. Kai, X. Meng, L. Mei, and W. Huang, "Deep reinforcement learning based user association and resource allocation for D2D-enabled wireless networks," in *Proc. IEEE/CIC ICCC*, 2021.

[17] X. Li, J. Li, Y. Liu, Z. Ding, and A. Nallanathan, "Residual transceiver hardware impairments on cooperative NOMA networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 680–695, 2020.

[18] R. Amiri *et al.*, "A machine learning approach for power allocation in HetNets considering QoS," in *Proc. IEEE ICC*, 2018.

[19] L. Zhang and Y.-C. Liang, "Deep reinforcement learning for multi-agent power control in heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 4, pp. 2551–2564, 2021.

[20] S. Gupta *et al.*, "Resource allocation for D2D links in the FFR and SFR aided cellular downlink," *IEEE Trans. Commun.*, vol. 64, no. 10, pp. 4434–4448, 2016.

[21] F. Meng, P. Chen, and L. Wu, "Power allocation in multi-user cellular networks with deep Q Learning approach," in *Proc. IEEE ICC*, 2019.

[22] F. D. Calabrese *et al.*, "Learning radio resource management in RANs: Framework, opportunities, and challenges," *IEEE Commun. Mag.*, vol. 56, no. 9, pp. 138–145, 2018.

[23] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *J. artificial intelligence research*, vol. 4, pp. 237–285, 1996.

[24] V. Mnih *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[25] X. Li *et al.*, "Hardware impaired ambient backscatter NOMA systems: Reliability and security," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2723–2736, 2021.

[26] T. D. Novlan, H. S. Dhillon, and J. G. Andrews, "Analytical modeling of uplink cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2669–2679, 2013.
[27] S. Essassi, M. Siala, and S. Cherif, "Dynamic fractional power control for LTE uplink," in *Proc. IEEE PIMRC*, 2011.
[28] K. Shen and W. Yu, "Fractional programming for communication systems—Part I: Power control and beamforming," *IEEE Trans. Signal Process.*, vol. 66, no. 10, pp. 2616–2630, 2018.

Tingting Yang received the B.S. degree in Communication Engineering from the School of Electrical and Information Engineering, Zhengzhou University, China, in 2023. She is currently pursuing the M.S. degree with the School of Electrical and Information Engineering of Zhengzhou University in Zhengzhou, China. Her current research interests are wireless communication, and reinforcement learning.

Yingqi Zhao received the B.S. and M.S. degrees from Zhengzhou University, China, in 2021 and 2024. She is currently working in XinXiang Yellow River Conservancy Bureau in China. Her research interests include deep learning and wireless communication.

Jin Jin received the Ph.D. degree in Communications and Information Systems from Beijing University of Posts and Telecommunications, Beijing, China, in 2014. Since 2014, he has been with the School of Electrical and Information Engineering, Zhengzhou University, where he is currently a Lecturer. His research interests include cooperative communications and deep learning.

Kaiyang Guo received the B.S. degree in Communication Engineering from the School of Electrical and Information Engineering, Zhengzhou University, China, in 2024. He is currently pursing for the M.S. degree at the same school. His main research interests are wireless communication and deep reinforcement learning.