Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Contribution-Aware Federated Edge Learning for Robust Resource Allocation in Massive IoT Networks

Hui Dun
Zhengzhou University of Light Industry Zhengzhou
Aowei Liu
liuaw_1234560163.com
Zhengzhou University of Light Industry Zhengzhou
Eryang Huan
Zhengzhou University of Light Industry Zhengzhou
Zhiyong Niu
Zhengzhou University of Light Industry Zhengzhou

Research Article

Keywords: Internet of Things, Device-to-device communication, Deep reinforcement learning, Multi-agent systems, Resource management

Posted Date: July 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9680458/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Contribution-Aware Federated Edge Learning for Robust Resource Allocation in Massive IoT Networks

Hui Dun[1], **Aowei Liu**[1,*], **Eryang Huan**[2] **and Zhiyong Niu**[1]

[1]School of Electronics and Information Zhengzhou University of Light Industry Zhengzhou, China
[2]School of Computer Science and Artificial Intelligence Zhengzhou University of Light Industry Zhengzhou, China
[3]Henan Key Laboratory of Information Functional Materials and Sensing Technology
[*]Corresponding author: Aowei Liu (E-mail: liuaw_123456@163.com)

Abstract—In the scene of large-scale Internet of Things (mIoT), the communication between high-density devices (D2D) will introduce serious co-frequency interference and cross-layer interference. Although the existing multi-agent deep reinforcement learning (MADRL) technology plays an important role in resource allocation challenges, there are also challenges in terms of limited spectrum awareness, unfair credit allocation and environmental non-stationarity. In order to meet these challenges, this paper proposes a multi-agent depth deterministic policy gradient (CA-FE-MADDPG) algorithm based on contribution awareness. First of all, the state space contains the characteristics of spatial interference graph and dynamic margin threshold, which can give each node of the Internet of Things a clear spectrum sensing ability. Secondly, a fine-grained penalty mechanism is designed based on physical interference ratio, interference severity and power regularization, which can be used to solve the problem of unfair credit distribution. This mechanism can not only maximize energy efficiency (EE) and concurrent access, but also strictly ensure the quality of service (QoS) of major users. In addition, the federal edge strategy and heuristic-guided hot-start strategy are designed, which can reduce non-stationarity and protect users' privacy while bypassing high-risk initial exploration, and ensure the safe and rapid convergence of the algorithm. A large number of simulations verify that CA-FE-MADDPG algorithm is significantly superior to the baseline in terms of system throughput, access rate and service quality satisfaction, which provides a new solution for ultra-dense mobile Internet of Things.

Keywords: Internet of Things, Device-to-device communication, Deep reinforcement learning, Multi-agent systems, Resource management

1 INTRODUCTION

With the rapid development of Industry 4.0 and smart cities, the huge network structure of the Internet of Things (IoT) has gradually taken shape, and all kinds of IoT devices, such as drones, smart devices, smart factory equipment, etc., are getting closer to our lives. These IoT devices have significantly

improved the quality of our lives. However, such a huge number of devices brings unprecedented challenges to the demand for device connection and data transmission in the Internet of Things [1-2]. Device-to-device (D2D) communication technology, as an important part of 5G/6G network, has also shown great potential in the field of Internet of Things, especially industrial Internet of Things. D2D technology is a communication technology for direct data transmission between close-range devices without passing through a base station. Through D2D communication technology, edge Internet of Things nodes can share the spectrum resources of cellular Internet of Things (C-IoT) users [3], which can reduce base station traffic and transmission delay [4-5] and greatly improve spectrum efficiency and system capacity in large-scale Internet of Things scenarios. However, due to the repeated use of the same channel in the large-scale Internet of Things, the traditional interference isolation means rely on the orthogonal method of resources, but the core advantage of D2D is resource reuse, and the introduction of all D2D will inevitably destroy the traditional interference isolation measures [6]. If there is no efficient radio resource management scheme, massive edge IoT will cause serious co-frequency interference and cross-layer interference when competing for limited spectrum resources [7-9], which will greatly reduce the quality of service (QoS) of C-IoT users and even lead to link interruption [10].

In order to deal with the problem of using D2D technology to destroy the traditional interference isolation measures, the early research mainly relies on traditional mathematical optimization and heuristic algorithms, which aim at optimizing resource block allocation and power control through strict mathematical modeling. Nagarajan et al. [11] Genetic algorithm (GA) combined with sequential optimal throughput optimization algorithm (SBTSA) is applied to jointly optimize the transmission power and resource block allocation in the underlying D2D communication. Benbraika et al. [12] put forward a new biological heuristic algorithm-Bees Life Algorithm, BLA), which simulates the marriage flight behavior of bees, avoids falling into the local optimal solution through diversified global search, and then combines with the method of strengthening local search to obtain the optimal solution. Mahfoudhi et al. [13] proposed a two-round optimization algorithm based on iterative combinatorial auction to solve the mixed integer nonlinear programming problem of joint power control and channel allocation in D2D communication. This algorithm will make each D2D pair bid for each channel, and then consider the interference between users to optimize the second round algorithm, in order to better improve the system energy efficiency on the premise of ensuring the

service quality of cellular users. Xue et al. [14] proposed a resource allocation algorithm combining power control and node priority calculation based on the traditional coloring algorithm of graph theory, aiming at the problem of D2D resource allocation and power control under the green Internet of Things. By optimizing the communication indicators as a whole, the algorithm improved the performance indicators such as fairness and throughput while maintaining low power consumption.

Although the traditional optimization algorithm can be used as resource allocation and power control strategy, there are some limitations, such as the computational complexity will increase exponentially with the number of users and resource blocks, so the traditional optimization algorithm is obviously not suitable for highly dynamic industrial Internet of Things scenarios. The emergence of reinforcement learning provides another solution for D2D resource allocation and power control. Reinforcement learning can learn the optimal strategy by interacting with the environment, which is suitable for solving the dynamic and high-dimensional D2D communication allocation problem. Early reinforcement learning algorithms [15], such as single agent depth Q network (DQN) and depth deterministic strategy gradient (DDPG), can solve some problems in resource allocation and power control, but they also have some limitations. State space and action space need to be discretized. When the number of users increases, the table-based method has high storage and convergence costs, and it cannot adapt to the dynamic dense environment, making it difficult for the algorithm to converge.

Therefore, the current research focus is on the application of multi-agent deep reinforcement learning (MADRL). The multi-agent deep reinforcement learning algorithm will treat each D2D pair as an independent learning and decision-making agent. Although each agent shares environmental information, it has its own local observation and decision-making actions, and can accomplish resource allocation and power control tasks in a distributed manner. At present, in the research of multi-agent deep learning algorithm, Xiang et al. [16] proposed a distributed scheme of multi-agent deep reinforcement learning. In this paper, each D2D pair is constructed as a distributed agent, and the D3QN algorithm with priority sampling is used to adaptively select transmission power and resources, which significantly improves the system throughput under strict delay constraints. Zhao et al. [17] proposed a distributed optimization method based on multi-agent deep reinforcement learning. In this algorithm, the competitive dual-depth Q-network (D3QN) strategy is adopted, and the user equipment obtains the global status through message delivery to maximize the long-term network utility. Zhao et al. [18] proposed a reinforcement learning scheme based on multi-agent DDQN to solve the problems

of mode selection and power control in V2X communication. Maximize the total capacity of V2I link under the premise of ensuring strict delay and reliability constraints of V2V link. Akhoundzadeh et al. [19] considered a three-layer D2D heterogeneous network composed of macrocells, multiple D2D users and multiple femtocell users, and proposed an autonomous multi-agent Q-learning (AMAQL) method to solve the problems of transmission power selection and subchannel allocation. Each user equipment is regarded as an independent agent, which can make resource management decisions independently without relying on network infrastructure. Zhang et al. [20] proposed a fully scalable distributed framework based on multi-agent near-end policy optimization (MAPPO) to solve the resource allocation problem in wireless cellular networks. Each agent only uses local information to manage the spectrum, power and scheduling of its own cell to minimize the packet delay under random traffic.

Although the above-mentioned MADRL algorithm has made remarkable progress in resource allocation and power control of D2D network, MADRL method still faces three key limitations in ultra-dense Internet of Things: 1) Only some observable states will produce spectrum blind spots, which will lead to spectrum unguided collision of D2D users; 2) unfair credit distribution caused by global punishment mechanism, which implicates innocent user nodes and wastes the potential of space reuse; 3) Large-scale parallel exploration intensifies environmental non-stationarity, while centralized training will damage user data privacy and bring high communication burden.

In order to meet these challenges, this paper proposes a federated edge multi-agent depth deterministic policy gradient algorithm based on interference contribution perception (CA-FE-MADDPG). The main contributions of this paper can be summarized as the following three aspects:

1. The state space of spectrum awareness: By integrating the spatial node interference graph and the dynamic margin threshold, the nodes of the Internet of Things can obtain explicit free spectrum awareness, so that agents can actively avoid congested hot spots when selecting spectrum.
2. Fine-grained punishment mechanism combining power regularization and physical interference contribution: In order to solve the unfair credit distribution problem, the reward function is reconstructed to make different degrees of punishment according to the actual QoS violation degree of cellular Internet of Things users and the proportion of interference contribution. This mechanism accurately endows each agent with clearer punishment ability.

Through the implementation of this mechanism, the energy efficiency (EE) and concurrent access rate of edge devices can be maximized, and the service quality of the Internet of Things can also be strictly guaranteed.

3. Federal average auxiliary architecture of edge intelligence: This architecture can effectively alleviate the environmental non-stationarity and accelerate the convergence of the model, while protecting the privacy of users' original data by regularly aggregating the shared features of actors' networks.

The rest of this paper is organized as follows: Section II establishes the system model and problem expression are established. Section III details the proposed CA-FE-MADDPG algorithm in detail. Section IV presents a lot of simulation results and discussions. Section V summarizes the full text.

2 SYSTEM MODEL AND PROBLEM FORMULATION

In the field of large-scale Internet of Things (mIoT) and industrial Internet of Things (IIoT), D2D communication enables terminal devices to establish communication directly without forwarding through base station (BS) or core network. Through this decentralized connection, D2D communication can provide a low-cost large-scale link scheme for the mobile Internet of Things. Although this technology can significantly alleviate base station congestion and improve spectrum efficiency, non-orthogonal resource sharing breaks the inherent interference isolation, especially in large-scale Internet of Things scenarios, which will greatly increase the degree of interference and bring great challenges to the coexistence of heterogeneous Internet of Things nodes. In this section, the physical layer communication model, channel fading mechanism, user movement model and problem expression of ultra-dense Internet of Things network are established.

2.1 System Model

In this paper, we only consider a single ultra-dense Internet of Things scenario with a coverage radius of R_{cell}, in which there is an Edge BS in the center of the network. This scenario consists of M high-priority C-IoT devices (represented by the set $\mathrm{M} = \{1, 2, \ldots, M\}$) and N edge IoT devices supporting D2D, and the edge IoT devices supporting D2D are densely distributed in pairs (represented by the set $\mathrm{N} = \{1, 2, \ldots, N\}$). It is assumed that the total bandwidth of the system is divided into M orthogonal resource blocks (RBs), and each C-IoT node is allocated one RB orthogonally, so the channel allocation problem of cellular Internet of Things nodes is not considered as a whole. The underlying

device-to-device IoT nodes can reuse the uplink resources of each C-IoT user, and allow multiple D2D devices to reuse one C-IoT user, that is, the uplink resources of each C-IoT user can be reused by multiple device-to-device IoT nodes, while each device-to-device IoT node can only reuse the uplink resources of one C-IoT user at the same time. The overall design system model diagram of this paper is as shown in Fig.1, which also clearly points out the cross-layer interference and the same layer interference. In this paper, a binary resource allocation matrix $\boldsymbol{\rho} \in \{0,1\}^{N \times M}$ is defined, in which the element $\rho_{n,m} = 1$ indicates that the $n-th$ D2D IoT link has multiplexed the $m-th$ RB, and $\rho_{n,m} = 0$ indicates that the $n-th$ D2D IoT link has not multiplexed the $m-th$ RB. Through the binary resource allocation matrix, we can clearly know which channel D2D is allocated to.

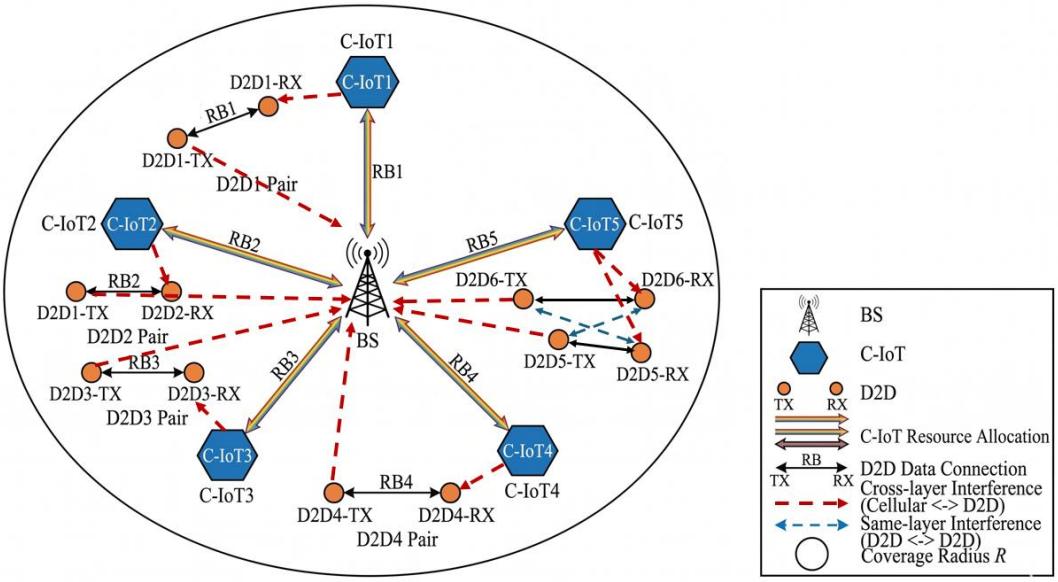


Fig.1 System model diagram

In order to reflect the real signal propagation characteristics in the Internet of Things scene and ensure that the wireless channel is highly dynamic and has fading characteristics, the channel model constructed in this paper considers the complex physical environment in the industrial Internet of Things scene, and the composite channel gain $h_{i,j}$ from node i to node j is modeled as the product of large-scale fading and small-scale fading, that is, the product of large-scale path loss, shadow fading and Rayleigh fast fading. The final channel gain can be expressed as formula (1):

$$h_{i,j} = \beta_{i,j} \cdot \zeta_{i,j} \cdot |g_{i,j}|^2 \tag{1}$$

where $\beta_{i,j} = 10^{-(PL(d_{i,j})/10)}$ denotes the path loss calculated according to the 3GPP Urban Micro

standard, $\zeta_{i,j} \sim \text{LogNormal}(0, \sigma_{sf}^2)$ represents the log-normal shadow fading, and $g_{i,j} \sim \text{CN}(0,1)$ is the Rayleigh fast fading coefficient.

Considering the dynamic characteristics of mobile Internet of Things nodes, such as Automated Guided Vehicles (AGVs), Unmanned Aerial Vehicles (UAVs), and wearable devices, this paper uses a Gaussian-Markov movement model to describe the evolution of node position and direction. The Gaussian-Markov movement model can well fit the dynamic movement characteristics of nodes in the Internet of Things scene, and the speed and direction of nodes are closely related to the historical state of the previous step instead of randomly jumping at each step. In addition, the speed and direction generated by the Gaussian-Markov movement model are related to each other. At time t, the updated equations of node velocity $v_i(t)$ and motion direction $\theta_i(t)$ can be expressed as equations (2) and (3):

$$v_i(t) = \alpha_v v_i(t-1) + (1-\alpha_v)\bar{v} + \sqrt{1-\alpha_v^2}\sigma_v w_{t-1} \tag{2}$$

$$\theta_i(t) = \alpha_\theta \theta_i(t-1) + (1-\alpha_\theta)\bar{\theta} + \sqrt{1-\alpha_\theta^2}\sigma_\theta u_{t-1} \tag{3}$$

Where α_v and α_θ are the memory parameters, w_{t-1} and u_{t-1} denote the independent Gaussian random processes. This high mobility renders the network interference topology extremely non-stationary.

In this paper, there are two main types of interference in the whole network: cross-layer interference and same-layer interference. Specifically, cross-layer interference refers to the interference generated by the D2D transmitter to the base station when D2D multiplexes the uplink of cellular Internet of Things users. In addition, the cellular Internet of Things users will interfere with the D2D receiver, and the uplink signals of the cellular Internet of Things users will also reduce the signal quality of the receiver. The interference in the same layer occurs between multiple D2D pairs that reuse the same resource block, that is, the signal of one D2D transmitter will interfere with the normal reception of other D2D receivers, so it is necessary to ensure the performance of cellular Internet of Things users through reasonable channel allocation and effective management of power control. In this paper, because the RB of the $m-th$ C-IoT node is repeatedly reused, the interference source at the receiving end comes from all D2D IoT transmission devices that transmit information by multiplexing the $m-th$ RB. Therefore, the instantaneous signal interference plus noise ratio (SINR) of C-IoT users can be expressed as Equation (4):

$$\gamma_{c,m} = \frac{P_c h_{m,B}}{\sum_{n \in \mathrm{N}} \rho_{n,m} P_{d,n} h_{n,B} + \sigma^2} \tag{4}$$

where P_c and $P_{d,n}$ denote the transmit powers of the C-IoT and edge D2D IoT nodes, respectively; $h_{m,B}$ and $h_{n,B}$ represent the channel gains of the corresponding links to the base station; and σ^2 is the background noise power.

Similarly, the edge IoT nodes of D2D repeatedly reuse the spectrum resources allocated to C-IoT users, which will inevitably cause cross-layer interference and same-layer interference. Specifically, it is assumed that the $n-th$ D2D-IoT link multiplexes the $m-th$ resource block (RB) occupied by C-IoT users. The interference that the receiving end of the D2D link will suffer will be composed of two parts: (1) Cross-layer interference from C-IoT transmitters using the same frequency band. (2) Same-layer mutual interference generated by D2D-IoT transmitters that select other edges of the RB at the same time. Therefore, the formula for calculating the instantaneous SINR of the $n-th$ D2D-IoT link on the $m-th$ RB can be expressed as Equation (5):

$$\gamma_{d,n}^{m} = \frac{P_{d,n} h_n}{\underbrace{P_c h_{m,n}}_{\text{Cross-tier Interference}} + \underbrace{\sum_{j \in \mathrm{N}, j \neq n} \rho_{j,m} P_{d,j} h_{j,n}}_{\text{Co-tier Interference}} + \sigma^2} \tag{5}$$

where $P_{d,n}$ represents the transmission power of the $n-th$ D2D-IoT transmitter, and h_n represents the direct useful channel gain of the $n-th$ D2D transmitter to its corresponding receiver. In the denominator, the $P_c h_{m,n}$ terminal represents the cross-layer interference power from the $m-th$ C-IoT user, where P_c is the fixed transmission power of the $m-th$ C-IoT node, and $h_{m,n}$ represents the interference channel gain from the $m-th$ C-IoT transmitter to the $n-th$ D2D receiver. In addition, the sum term of $\sum_{j \in \mathrm{N}, j \neq n} \rho_{j,m} P_{d,j} h_{j,n}$ represents the same layer interference caused by all other D2D-IoT pairs sharing the same RB, $\rho_{j,m} \in \{0,1\}$ serves as a binary spectrum allocation indicator, and $h_{j,n}$ is the channel gain between the transmitter of the $j-th$ D2D pair and the receiver of the $n-th$ D2D pair. Finally, σ^2 represents the background Additive White Gaussian Noise (AWGN) power.

According to Shannon formula, the instantaneous rate calculation formula is formula (6):

$$R = B\log_2(1+\gamma) \tag{6}$$

2.2 Problem Formulation with Composite Utility

The traditional resource allocation problem usually maximizes the system summation rate under hard constraints, but the unfair credit allocation mechanism may lead to biased learning, slow learning and even inability to converge. In addition, in large-scale Internet of Things networks, edge sensors or devices are usually restricted by strict battery capacity, so this paper also designs energy efficiency as an optimization goal when designing the system optimization goal. Therefore, in order to achieve a strong service quality guarantee and the goal of green Internet of Things, this paper redefines the optimization goal to maximize the comprehensive utility of the system.

The goal of system optimization is defined as finding the optimal RB allocation strategy $\boldsymbol{\rho}$ and power control strategy $\mathbf{P_d}$ to maximize the long-term expected utility, which can be expressed as Equation (7):

$$\max_{\{\boldsymbol{\rho},\mathbf{P_d}\}} \quad \mathrm{E}\left[\sum_{n\in\mathrm{N}} U_n(\boldsymbol{\rho},\mathbf{P_d})\right] \tag{7}$$

where the composite utility function U_n of the $n-th$ D2D IoT node is defined as Equation (8):

$$U_n = R_{d,n} - \lambda_p \cdot \Phi_p(P_{d,n}) - \beta \cdot \Phi_{vio}(\gamma_{c,m}, C_{n,m}) \tag{8}$$

In this IoT-oriented composite utility function, $R_{d,n}$ denotes the spectral efficiency gain of the edge IoT communication link, and $\Phi_p(P_{d,n}) = \dfrac{P_{d,n}}{P_{max}}$ represents the power regularization term with a weighting coefficient λ_p. The purpose of this part is to promote the agent to actively seek the lowest feasible transmission power when the battery is limited, and at the same time meet the data transmission requirements, thus prolonging the life of the battery of the IoT node. $\Phi_{vio}(\gamma_{c,m}, C_{n,m})$ signifies the proposed fine-grained responsibility penalty term, weighted by the coefficient β. When a high-priority C-IoT node m experiences a QoS violation (i.e., $\gamma_{c,m} < \gamma_{th}$), this penalty term imposes a dynamic deduction based on the current severity of the interference and the actual interference contribution ratio $C_{n,m}$ of the IoT node n.

The optimization problem is simultaneously subjected to the following physical constraints, as

expressed in Equations (9) to (13):

$$\text{s.t.} \quad 0 \le P_{d,n} \le P_{max}, \quad \forall n \in \mathrm{N} \tag{9}$$

$$\sum_{m \in \mathrm{M}} \rho_{n,m} \le 1, \quad \forall n \in \mathrm{N} \tag{10}$$

$$\rho_{n,m} \in \{0,1\}, \quad \forall n \in \mathrm{N}, \forall m \in \mathrm{M} \tag{11}$$

$$\mathrm{SINR}_m^c \ge \gamma_{\mathrm{th}}^c, \quad \forall m \in \mathrm{M} \tag{12}$$

$$\mathrm{SINR}_n^d \ge \gamma_{\mathrm{th}}^d, \quad \forall n \in \mathrm{N} \tag{13}$$

Specifically, constraint (9) stipulates that the continuous transmission power of each D2D device should be within a reasonable range, and the maximum transmission power should not exceed $P_{\max}$. Constraints (10) and (11) define the binary nature of resource allocation indicators, limiting each edge IoT node to occupy at most one resource block. Constraint (12) is the strictest physical security boundary, and it is necessary to ensure that the instantaneous SINR of C-IoT users strictly meets their QoS threshold γ_{th}^c during communication. Finally, constraint (13) ensures the minimum decoding threshold γ_{th}^d of D2D links. When the value is greater than the minimum threshold, the communication between D2D links will be regarded as effective communication.

Because the interference term in the expression closely couples the decision variables of the nodes in the Internet of Things, the output actions of the neural network designed in this paper include continuous D2D transmission power and discrete channel resource block allocation, and the objective function constitutes a typical nonconvex mixed integer nonlinear programming (MINLP) problem. In the next section, we reformulate the problem as Markov decision processes (MDP) and propose a CA-FE-MADDPG algorithm to achieve efficient optimization.

3 PROPOSED CA-FE-MADDPG FOR IOT RESOURCE ALLOCATION

As mentioned in the second part, the joint resource allocation in large-scale Internet of Things constitutes a high-dimensional and nonconvex MINLP problem, so we propose an edge intelligence CA-FE-MADDPG algorithm. The structure diagram of CA-FE-MADDPG algorithm is shown in Fig.2, which includes the federal edge aggregation strategy, heuristically guided hot start strategy, CA-FE-MADDPG network architecture, reward function design-credit allocation based on contribution awareness and other structural frameworks. This section will introduce the details of MDP modeling of

the algorithm, fine-grained punishment mechanism of green Internet of Things, federal edge training framework and heuristically guided hot start strategy.

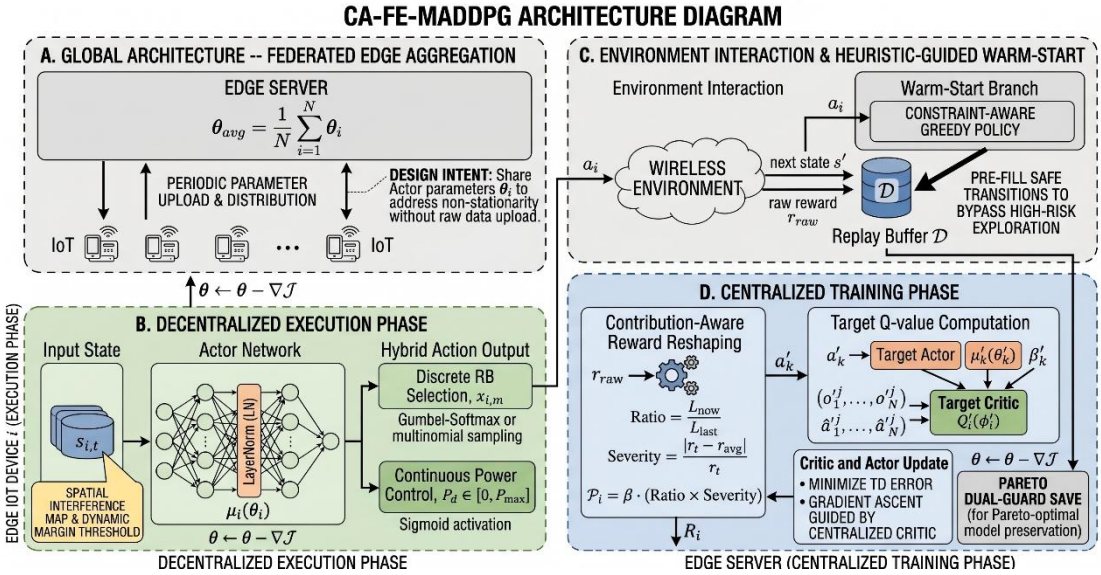


Fig. 2 structure diagram of CA-FE-MADDPG algorithm

3.1 MDP Formulation

Each D2D pair is modeled as an independent agent $n \in \mathrm{N}$, which interacts with the environment only through its local observation state $s_n(t) \in \mathrm{S}$ and outputs joint actions $a_n(t) \in \mathrm{A}$.

In order to solve the ubiquitous spectrum blind spot problem, the observation of agents in time slot t is carefully enhanced. It clearly integrates the characteristics of local channel quality, spatial interference topology and security boundary of agents:

$s_n(t) = \left\{h_{d,n}(t), d_n(t), P_{d,n}(t-1), h_{n,B}(t), \mathbf{h}_c^B(t), \boldsymbol{\gamma}_{margin}(t), \mathbf{I}_{map}(t), \mathbf{I}_{sort}(t)\right\}$, $\boldsymbol{\gamma}_{margin}(t) \in \square^{M}$ denotes a dynamic interference margin threshold vector, which indicates the maximum amount of extra interference that each RB can tolerate without destroying the quality of service of the C-IoT node. $\mathbf{I}_{map}(t)$ and $\mathbf{I}_{sort}(t)$ respectively indicate the descending characteristics of the historical spatial interferogram and its ranking. This state indicates that the underlying IoT nodes are endowed with clear spectrum hole perception ability, so that they can actively avoid the highly reused spectrum in the industrial IoT and choose other idle channels.

3.2 Network Architecture

In the current multi-agent algorithm, each agent will not only be affected by environmental changes, but also by other agent strategies, which leads to the characteristics of the environment are not fixed from

the perspective of a single agent. This paper propose the CA-FE-MADDPG algorithm based on the multi-agent depth deterministic strategy gradient (MADDPG) algorithm. Next, this section will briefly explain the MADDPG algorithm. MADDPG algorithm adopts the structure of centralized training and distributed execution.

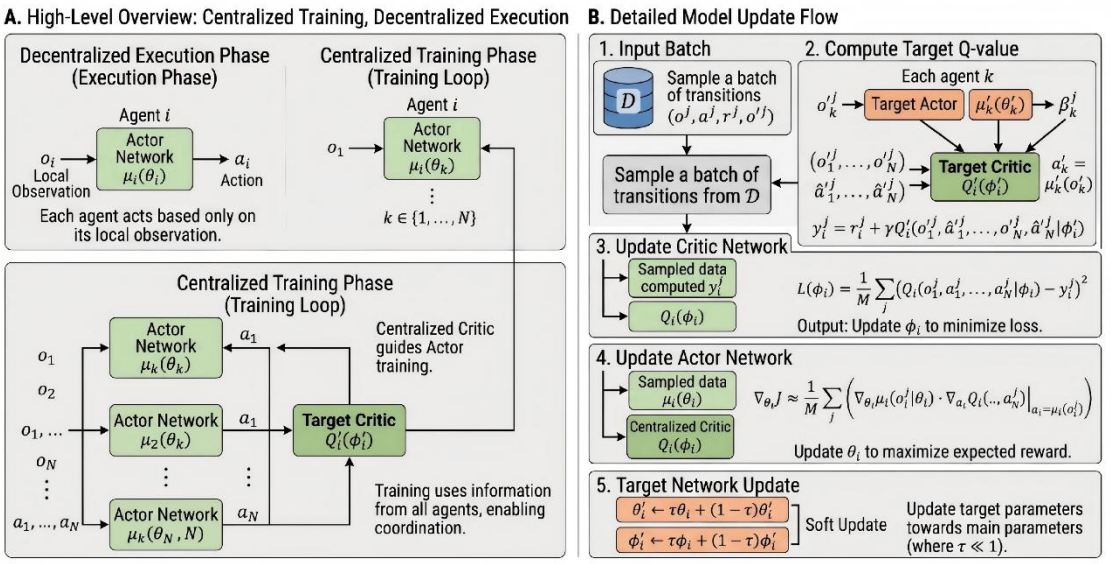


Fig.3. MADDPG algorithm architecture and training update flow chart

The flow chart of standard MADDPG algorithm is shown in Fig.3 Panel A shows the core design concept of MADDPG: centralized training and decentralized execution. Panel B is a detailed model updating process. In a single iteration of the algorithm, the underlying mathematical logic and network updating process are disassembled step by step.

Regarding the structural design of actor network, firstly, the number of nodes in the input layer must be equal to the dimension of the local observation data, then it passes through the hidden layer, and ReLU activation function is used after each hidden layer. Finally, the action is mapped to the legal range through tanh or S morphological function. As the only decision-making unit in the execution stage, actor network can independently respond to environmental changes and output actions that the network considers optimal. Although actor network can output corresponding actions, it cannot judge the learning degree of neural network. Critic network can provide stable value evaluation based on global information, so that all participants' updates are consistent with the optimal goals. Only by combining the two can neural networks learn in the best direction.

3.3 Fine-grained Punishment Mechanism of Green Internet of Things

In large-scale multi-agent algorithms, global punishment is usually used to meet the service quality of

cellular nodes, which often leads to unfair credit distribution and agent laziness. The unfair credit distribution means that in the process of agent learning, agents can't judge whether the reward or punishment is caused by their own action choice or other agent actions. Global punishment often makes the system directly give a global negative reward when it triggers a QoS violation, so agents can't judge whether it is a negative reward caused by personal behavior, so unfair credit distribution will cause the learning efficiency of the algorithm to be low, and agents need to make a lot of attempts before they can get a joint action without violations in an accidental discovery. The use of global punishment will also lead to agent laziness. The so-called agent laziness refers to the fact that in multi-agent algorithms, because individual contributions cannot be accurately evaluated, some agents will gradually tend to do nothing or optimize their strategies with low efforts. Obviously, this situation is a negative strategy. If most agents enter the agent laziness stage due to global punishment, the output solution of neural network will be more and more biased towards non-optimal solutions.

In addition to the above-mentioned problems caused by global punishment, there is a more important factor to be focused on in large-scale Internet of Things scenarios-energy efficiency. In the current scene of Internet of Things, the importance of energy efficiency is self-evident, which is related to the feasibility of the whole system, the cost of deployment, the life of the network, etc. The sensors of Internet of Things are usually powered by batteries with limited capacity, but the traditional reward mechanism often ignores the control of energy consumption. If energy efficiency optimization is not considered, then deploying communication equipment in dense scenes, even if the power consumption of a single node is very low, energy consumption will make battery replacement impractical under the accumulation of a large number of nodes, so this paper will also regard energy efficiency as. In this paper, a fine-grained penalty scheme based on the contribution of physical interference is proposed for unfair credit distribution, agent laziness and energy efficiency, which is tailor-made for the green Internet of Things. The instantaneous net return of each agent in the Internet of Things is defined as Equation (14):

$$r_n(t) = R_{d,n}(t) - \lambda_p \cdot \Phi_p(t) - \Phi_{vio,n}(t) \tag{14}$$

In the above equation, $R_{d,n}(t) = \log_2(1+\gamma_{d,n}(t))$ represents the communication rate gain of the IoT link, and $\Phi_p(t) = \frac{P_{d,n}(t)}{P_{max}}$ denotes the power regularization term. Under the guidance of weight λ_p, 123 can force IoT agents with limited battery capacity to abandon the full power transmission

strategy and actively seek the minimum feasible power to meet the data transmission requirements. Therefore, the EE of the whole Internet of Things network can be significantly improved and the network life can be prolonged.

The core proportional penalty term $\Phi_{vio,n}(t)$ is triggered only when a C-IoT node experiences a QoS violation (i.e., $\gamma_{c,m} < \gamma_{th}$). The calculation of this core proportional penalty is given by Equation (15):

$$\Phi_{vio,n}(t) = \beta \cdot \left(\omega_{base} + \omega_{ratio} \cdot C_{n,m}(t)\right) \cdot S_m(t) \tag{15}$$

The penalty mechanism is driven by interference severity $S_m(t) = \max\left(0, 1 - \frac{\gamma_{c,m}(t)}{\gamma_{th}}\right)$ and interference contribution ratio $C_{n,m}(t) = \frac{P_{d,n}(t) h_{n,B}(t)}{\sum_{i \in \mathrm{N}_m} P_{d,i}(t) h_{i,B}(t) + \text{ò}}$. This punishment mechanism gives more severe punishment to the high-power D2D IoT nodes with greater interference from the base station, and less punishment to the IoT nodes with less interference. Through this punishment mechanism, the low-power edge microsensor can reuse the spectrum normally. This has greatly released the potential of concurrent transmission of resources in large-scale Internet of Things network scenarios.

3.4 Federated Edge Training

In the scene with dense nodes, the non-static environment makes the traditional MADDPG vulnerable to serious training oscillation. In addition, IoT devices are usually sensitive to data privacy, and frequent exchange of a large number of original state data will bring huge communication overhead. To solve these problems, this paper innovatively introduces the Federated Edge Aggregation Strategy in the framework of CA-FE-MADDPG. The federated aggregation strategy allows the edge IoT nodes to train the model with empirical data locally, and then exchange some weight parameters of the neural network with the server regularly. With the addition of this strategy, the global environment can be shared on the premise of protecting data privacy. Therefore, in this paper, the front hidden layer of the actor network is designated as the shared feature layer $\theta_n^{\mu,shared}$, and the joint average is triggered periodically every K rounds in the training stage, which is expressed as Formula (16):

$$\theta_{avg}^{\mu,shared} = \frac{1}{N} \sum_{i=1}^{N} \theta_i^{\mu,shared} \tag{16}$$

The edge server (e.g., the base station) exclusively collects and computes the average of the local

model parameters, injects Gaussian exploration privacy noise with a variance of σ_{FL}^2, and subsequently broadcasts the aggregated parameters back to each IoT device via a soft update factor τ_{FL}:

$$\theta_n^{\mu,shared} \leftarrow (1-\tau_{FL})\theta_n^{\mu,shared} + \tau_{FL}\theta_{avg}^{\mu,shared} + \mathrm{N}(0,\sigma_{FL}^2) \tag{17}$$

Because the federated learning mechanism keeps the original data in the local terminal forever without uploading it to the central server, the federated edge mechanism can strictly protect the device data privacy of the industrial Internet of Things.

In addition, in order to alleviate the large-scale QoS violation caused by random exploration in the initial training stage of traditional MADRL, this paper introduces a heuristic-guided hot-start mechanism before training. Specifically, Internet of Things agents use heuristic algorithm experience to guide initial action selection. This startup mechanism can effectively prevent the agent from being severely punished for early trial and error violations, thus bypassing the initial startup exploration stage of high violations. It should be noted that this startup mechanism only uses a very small amount of data to support the CA-FE-MADDPG algorithm to survive the previous high violations.

Algorithm 1 Contribution-Aware Federated Edge MADDPG (CA-FE-MADDPG)

Input: Environment parameters, federated interval Δ_{FL}

Output: Optimized decentralized Actor networks $\{\mu_{\theta_n}^*\}_{n=1}^N$

1: Initialize global shared feature weights $\theta_{avg}^{\mu,shared}$, Replay Buffer D, and local networks for all agents $n \in \mathrm{N}$.

2: [Phase I: Heuristic-Guided Warm Start]

3: Execute constraint-aware greedy strategy to collect expert transitions; populate D to circumvent cold-start.

4: [Phase II & III: Decentralized Training & Federated Edge Aggregation]

5: for episode $e = 1,\ldots,E$ **do**

6: Update target SINR threshold γ_{th} according to curriculum learning progress.

7: Reset environment and acquire comprehensive local observation $s_n(t)$ for all $n \in \mathrm{N}$.

8: **for** step $t = 1,\ldots,T$ **do**

9: **Action:** Each agent n outputs continuous power and discrete RB selection $a_n(t) \in \mathrm{A}$ via Gumbel-Softmax.

10: Execute joint action $\mathbf{a}(t)$, observe $s_n(t+1)$ and C-IoT instantaneous SINR $\gamma_{c,m}(t)$.

11: **for** each agent $n \in \mathrm{N}$ **do** (Fine-Grained Credit Assignment)

12: **if** QoS violation occurs ($\gamma_{c,m}(t) < \gamma_{th}$) **then**

13: Compute interference contribution ratio: $C_{n,m}(t)$.

14: Calculate proportional penalty: $\Phi_{vio,n}(t) = \beta \cdot (\omega_{base} + \omega_{ratio} \cdot C_{n,m}(t)) \cdot S_m(t)$.

15: **end if**

16: Obtain net reward $r_n(t) = R_{d,n}(t) - \lambda_p \cdot \Phi_p(t) - \Phi_{vio,n}(t)$ and store transition in D.

17: Sample mini-batch from D to perform policy gradient update for Actor and Critic.

18: **end for**

19: **if** $e \bmod \Delta_{FL} == 0$ **then** (Federated Edge Training)

20: Edge Server aggregates Actor's shared feature layers: $\theta_{avg}^{\mu,shared} = \frac{1}{N}\sum_{i=1}^{N} \theta_i^{\mu,shared}$.

21: Broadcast and soft-update with privacy noise: $\theta_n^{\mu,shared}$

22: end for

23: return $\{\mu_{\theta_n}^*\}_{n=1}^N$.

4 PERFORMANCE EVALUATION

This part will carry out simulation analysis, aiming at evaluating the effectiveness of CA-FE-MADDPG resource allocation scheme in different D2D quantity environments. This article consider an ultra-dense industrial Internet of Things scenario with a coverage radius of $R_{cell} = 500m$. The network integrates $M = 15$ C-IoT users and different numbers of edge IoT pairs supported by D2D to simulate different congestion levels. Wireless channels include 3GPP UMi path loss, shadow fading and Rayleigh fast

fading. The dynamic mobility of Internet of Things nodes is modeled by Gaussian-Markov process. Detailed system parameter settings are summarized in Table 1.

Table 1 The parameter of simulation.

Parameter Description	**Symbol**	**Value**
Cell radius	R_{cell}	$500m$
Number of C-IoT nodes / RBs	M	15
Number of edge D2D pairs	N	9,12, …30
Maximum intra-pair distance for D2D	$d_{\max}$	$50m$
User mobility speed range	$v(t)$	$0 \sim 2.0m/s$
Bandwidth per Resource Block	B	$180kHz$
C-IoT transmit power	P_c	$23dBm$
Maximum D2D transmit power	$P_{\max}$	$14dBm$
Background noise power	σ^2	$-114dBm$
Minimum decoding threshold for C-IoT	γ_{th}^{c}	$1.0dB$
Minimum decoding threshold for D2D	γ_{th}^{d}	$1.0dB$

In order to verify the CA-FE-MADDPG algorithm proposed in this paper, the following four algorithms will be used for evaluation. Reference [21] proposes a C-DDPG algorithm, which adopts a single global actor and emphasizes the inherent performance bottleneck and dimension problems in centralized architecture. Reference [22] adopts the multi-agent DQN(MA-DQN) algorithm in discrete space, which shows the performance advantages of continuous power control. In addition to the above two algorithms as a comparison algorithm, a random algorithm will also be added as a comparison algorithm. Therefore, CA-FE-MADDPG algorithm, a common three algorithms, carries out performance comparison and evaluation.

Fig.4 shows the accelerating effect of the designed heuristic startup mechanism on the model convergence. In the initial training stage, the algorithm without guided exploration can easily choose the wrong action, which leads to large-scale QoS violations. It can be seen from the figure that the initial reward of the algorithm without hot start mechanism is below 325. In contrast, the complete CA-FE-MADDPG algorithm with heuristic algorithm data experience can successfully skip the "cold start"

stage of initial high violation actions. It can be seen from the figure that under the first 1500 rounds, the complete algorithm has a higher initial reward and can converge to the optimal state smoothly and quickly. This clearly verifies the key value of hot start strategy in ensuring the safety of early exploration.

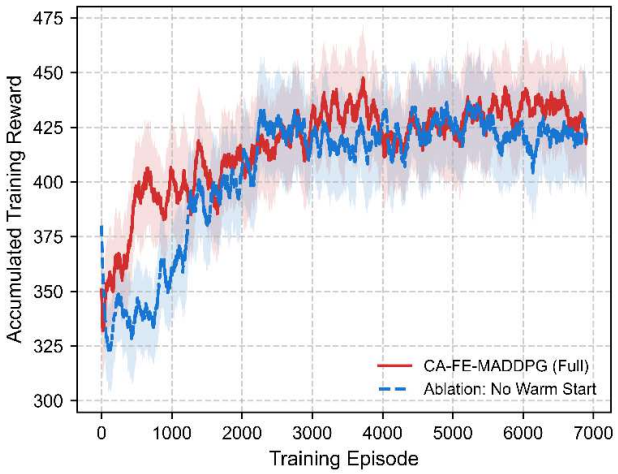


Fig.4 Accumulated training reward convergence in the ablation study

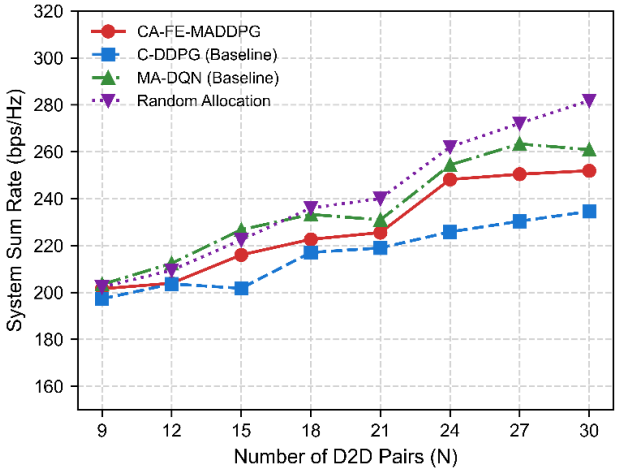


Fig.5 System sum rate under different numbers of edge IoT pairs

Fig.5 shows the changing trend of the total system rate of each algorithm with the increase of the number of edge IoT devices. It can be seen that CA-FE-MADDPG algorithm is superior to C-DDPG algorithm in total channel capacity, and is close to MA-DQN algorithm. Although the random algorithm achieves a higher total rate on the total rate curve, it will not only reject a large number of potential concurrent transmission potential, but also weaken the quality of service of C-IoT users. Although MA-DQN algorithm is superior to CA-FE-MADDPG algorithm in total channel capacity, its performance needs to be comprehensively analyzed with other indicators in large-scale Internet of Things scenarios. When the number of users in the D2D Internet of Things reaches 24 pairs, the growth trend of the total channel

capacity of the CA-FE-MADDPG algorithm becomes slow, which leads to the algorithm strictly limiting the transmission capacity of users to avoid the channel interference of most C-IoT users.

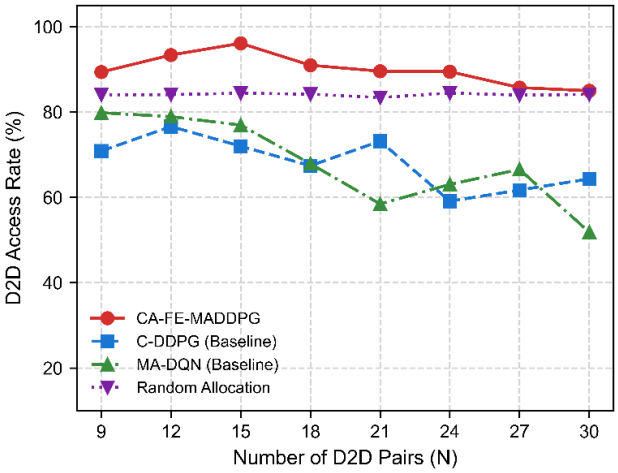


Fig.6 Access rate under different number of IoT devices

Fig.6 shows the access rate of D2D devices with different numbers of D2D devices in the Internet of Things. It can be clearly seen from the figure that the access rate index of CA-FE-MADDPG algorithm is always much higher than the other three algorithms. Although the access rate index of CA-FE-MADDPG algorithm will decrease with the increase of D2D devices in the Internet of Things, the decreasing trend is very slow and the overall access rate can still reach about 0.85 when N=30. On the other hand, the access rate of C-DDPG algorithm and MA-DQN algorithm is much higher than that of innovative algorithm with the increase of D2D devices in the Internet of Things. The two algorithms are facing more and more serious dimension disasters with the increase of N, and their access rate performance index is far lower than that of innovative algorithms. Although the access rate of the innovative algorithm tends to increase obviously in the case of N=15, it is because in the case of n = 9 and 12, the agent can easily choose the appropriate channel and try to maximize the transmission power of its own D2D equipment. The high-intensity transmission power may have a certain impact on other D2D users. When the number of equipment reaches 15, the agent will find the most appropriate solution under the scarce channel resources. The transmission power will be reduced appropriately, so that the impact on other D2D devices will become smaller, and its access rate will also increase, so the overall access rate statistics in the case of N=9,12 may be lower than that in the case of N=15. But to sum up, the access rate of CA-FE-DDPG algorithm is far stronger than other comparison algorithms.

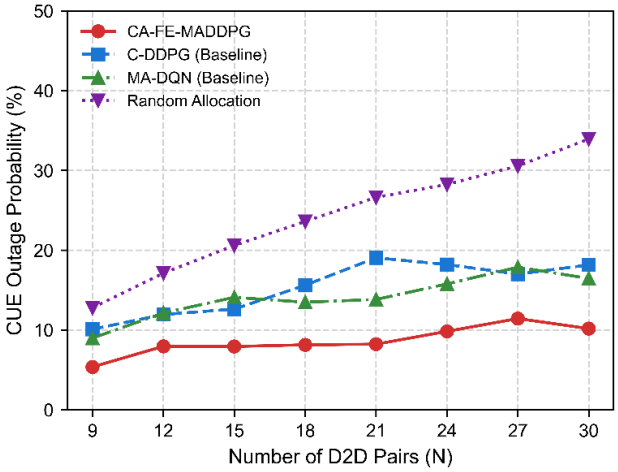


Fig.7. Violation rate under different equipment densities

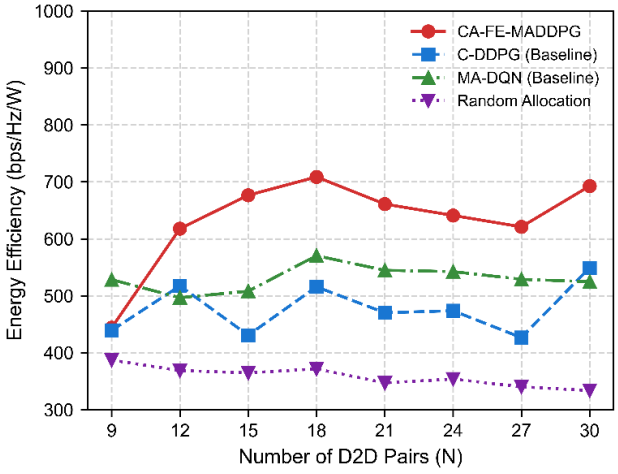


Fig.8 System energy efficiency versus the number of edge IoT pairs.

Fig.7 shows the C-IoT user violation rate of each algorithm under different numbers of IoT D2D devices. As can be seen from Figure 7, with the increase of the number of D2D devices in the Internet of Things, the overall violation rate of CA-FE-MADDPG algorithm is slowly rising, and it is lower than the other three algorithms. Combined with Figure 5 and Figure 6, it can be found that although the violation rate of C-IoT users of C-DDPG algorithm and MA-DQN algorithm decreased when n = 21, 24, 27 and 30, it was because C-DDPG algorithm and MA-DQN algorithm abandoned a large number of D2D devices in the Internet of Things and achieved a low violation rate. In essence, many D2D devices were in C-DDPG algorithm and MA-DQN algorithm. In contrast, CA-FE-MADDPG algorithm will not give up access to D2D devices in exchange for violation rate. This fully proves that the fine-grained penalty mechanism based on the contribution of physical interference successfully solves the unfair multi-agent credit allocation problem, thus completely releasing the concurrent reuse potential of the underlying

edge devices.

Energy efficiency is very important in large-scale Internet of Things. Fig.8 shows the energy efficiency diagram of each algorithm under different IoT D2D devices. It can be seen that under different D2D pairs, its energy efficiency is better than the other three comparison algorithms. Among them, during the interval from N=9 to N=18, the whole network has not reached the extreme congestion state for the time being. Therefore, through the perception of spatial state interference graph, CA-FE-MADDPG algorithm can allocate the newly added D2D pairs of Internet of Things to the appropriate C-IoT frequency band, and smart users will transmit power as much as possible without interfering with cellular users. As the number of D2D pairs increases to 18, the system needs to control the transmission power better. When the D2D logarithm in the Internet of Things exceeds 18, the network will enter a super-dense state, and cross-layer interference and same-layer interference will intensify. At this stage, the total growth rate will enter a slow rising stage to ensure the quality of service of the main users of C-IoT. The high access rate of the system also indicates that the transmission capacity of most D2D users is in a reasonable range, thus achieving a reasonable and moderate decline in the energy efficiency curve.

5 CONCLUSIONS

In this paper, a contribution-aware joint edge MADDPG algorithm is proposed, which can be used for resource allocation or power control in ultra-dense Internet of Things network scenarios. Firstly, the algorithm enables the agent to acquire spectrum sensing ability through the characteristics of spatial interference graph and interference threshold, and then solves the unfair punishment problem based on the fine-grained punishment mechanism of physical interference contribution. The algorithm also adopts the federal edge training framework and heuristic startup mechanism to protect users' privacy, bypass the initial exploration of high risk and ensure the safe and rapid integration of the algorithm. A large number of simulation experiments show that the heuristic starting mechanism can effectively improve the initial exploration speed of the algorithm. In addition, the CA-FE-MADDPG algorithm also achieves a better balance between system capacity and violation rate, and strictly guarantees the quality of service of major C-IoT users. This algorithm provides a new scheme for the future green industrial Internet of Things.

ACKNOWLEDGEMENT

This work was supported in part by the Zhengzhou University of Light Industry Doctoral Research

Fund Project (No. 2021BSJJ026 and 2021BSJJ027), Postgraduate Education Reform and Quality Improvement Project of Henan Province (No.YJS2025GZZ09), The Key Technologies R&D Programme of Henan Province (No.242102210112), Scientific and Technological Project in Henan Province (No. 242102210199 and 252102211123), and Natural Science Foundation of Henan (No. 232300420419).

STATEMENTS AND DECLARATIONS

Funding: This work was supported by the Zhengzhou University of Light Industry Doctoral Research Fund Project (No. 2021BSJJ026 and 2021BSJJ027), Postgraduate Education Reform and Quality Improvement Project of Henan Province (No.YJS2025GZZ09), The Key Technologies R&D Programme of Henan Province (No.242102210112), Scientific and Technological Project in Henan Province (No. 242102210199 and 252102211123), and Natural Science Foundation of Henan (No. 232300420419). **Competing interests:** The authors declare no competing interests relevant to the content of this article. **Ethics approval:** Not applicable. **Consent to participate:** Not applicable. **Consent for publication:** Not applicable. **Availability of data and materials:** All data generated or analysed during this study are included in this published article. **Code availability:** The code used to generate the results in this study is available from the corresponding author on reasonable request. **Authors' contributions:** D.H.: Writing – review & editing, Validation, Supervision, Methodology, Funding acquisition. L.A.W.: Writing – review & editing, Writing – original draft, Methodology, Validation. H.E.Y.: Supervision, Investigation, Funding acquisition. N.Z.Y.: Supervision, Investigation, Funding acquisition.

REFERENCES

[1] Abdalzaher, M. S., Krichen, M., Shaaban, M. F., & Fouda, M. M. (2025). Quality-Focused Internet of Things Data Management: A Survey, Perspectives, Open Issues, and Challenges. *IEEE Internet of Things Journal*, 12(22), 46431–46458. https://doi.org/10.1109/JIOT.2025.3613669

[2] Cheikh, I., Roy, S., Sabir, E., & Aouami, R. (2026). Energy, Scalability, Data, and Security in Massive IoT: Current Landscape and Future Directions. *IEEE Internet of Things Journal*, 13(6), 10127–10163. https://doi.org/10.1109/JIOT.2026.3655796

[3] Fasihi, M. R., & Mark, B. L. (2025). Device-to-Device Communication in 5G/6G: Architectural Foundations and Convergence with Enabling Technologies. *arXiv*.

https://doi.org/10.48550/arXiv.2507.06946

[4] Raghu, T. V., & Kiran, M. (2024). Priority-Driven Resource Allocation and Power Optimization in D2D Communication. *IEEE Transactions on Green Communications and Networking*, 8(4), 1479–1491. https://doi.org/10.1109/TGCN.2024.3399403

[5] Kim, H. M., Nguyen, H. V., Kang, G. M., Shin, Y., & Shin, O. S. (2023). SWIPT-Aided Device-to-Device Communications for Massive IoT Networks: A Novel Resource Allocation With Sparse Code Multiple Access. *IEEE Internet of Things Journal*, 10(22), 19617–19629. https://doi.org/10.1109/JIOT.2023.3282597

[6] Sarma, S. S., Hazra, R., & Mukherjee, A. (2022). Symbiosis Between D2D Communication and Industrial IoT for Industry 5.0 in 5G mm-Wave Cellular Network: An Interference Management Approach. *IEEE Transactions on Industrial Informatics*, 18(8), 5527–5536. https://doi.org/10.1109/TII.2021.3134285

[7] Sarma, S. S., Hazra, R., & Goswami, P. (2023). Power Optimization in a Multicell D-2D Communication for Smart City in an mm-Wave Cellular Network: An mIoT Perspective. *IEEE Internet of Things Journal*, 10(21), 18686–18694. https://doi.org/10.1109/JIOT.2023.3262577

[8] Huang, J., Yu, T., Yang, F., Zhang, S., Jiang, W., & Niyato, D. (2024). AoI-Aware Resource Allocation With Interference Avoidance for Ultradense Industrial Internet of Things Networks. *IEEE Internet of Things Journal*, 11(17), 28787–28797. https://doi.org/10.1109/JIOT.2024.3403849

[9] Zheng, K., Yang, H., Lei, L., Mei, J., & Wu, B. (2025). AI-Driven Resource Management for Heterogeneous Industrial IoT: Challenges and Opportunities. *IEEE Open Journal of the Communications Society*, 6, 10269–10286. https://doi.org/10.1109/OJCOMS.2025.3640688

[10] Veedu, R. T., & Manjappa, K. (2024). An efficient application based many-to-many r esource allocation and sharing with power optimization for D2D communication — A clustered approach. *Journal of Communications and Networks*, 26(1), 19–34. https://doi.org/10.23919/JCN.2023.000062

[11] Nagarajan, R., & Mohamad, N. M. V. (2022). Energy optimized resource and power allocation in an uplink-based underlay device-to-device communication for 5G network. *International Journal of Communication Systems*, 35(10), e5145. https://doi.org/10.100

2/dac.5145

[12] Benbraika, M. K., Bitam, S., & Mellouk, A. (2019). Joint resource allocation and po wer control based on Bee Life Algorithm for D2D Communication. In *2019 IEEE Wi reless Communications and Networking Conference (WCNC)*. https://doi.org/10.1109/W CNC.2019.8885884

[13] Mahfoudhi, M., Hamdi, M., & Zaied, M. (2019). Distributed Resource Allocation Using Iterative Combinatorial Auction for Device-to-Device Underlay Cellular Networks. In *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*. https://doi.org/10.1109/IWCMC.2019.8766768

[14] Xue, Y., Yang, Z., Yang, W., & Yang, J. (2019). D2D Resource Allocation and Pow er Control Algorithms Based on Graph Coloring in 5G IoT. In *2019 Computing, Com munications and IoT Applications (ComComAp)*. https://doi.org/10.1109/ComComAp462 87.2019.9018806

[15] Luong, N. C., Hoang, D. T., Gong, S., Niyato, D., Wang, P., & Liang, Y.-C. (2019). Applications of Deep Reinforcement Learning in Communications and Networking: A Survey. *IEEE Communications Surveys & Tutorials*, 21(4), 3133–3174. https://doi.org/ 10.1109/COMST.2019.2916583

[16] Xiang, H., Yang, Y., He, G., Huang, J., & He, D. (2022). Multi-Agent Deep Reinfor cement Learning-Based Power Control and Resource Allocation for D2D Communicati ons. *IEEE Wireless Communications Letters*, 11(8), 1659–1663. https://doi.org/10.1109/ LWC.2022.3170998

[17] Zhao, N., Liang, Y.-C., Niyato, D., Pei, Y., Wu, M., & Jiang, Y. (2019). Deep Reinforcement Learning for User Association and Resource Allocation in Heterogeneous Cellular Networks. *IEEE Transactions on Wireless Communications*, 18(11), 5141–5152. https://doi.org/10.1109/TWC.2019.2933417

[18] Zhao, D., Qin, H., Song, B., Zhang, Y., Du, X., & Guizani, M. (2020). A Reinforcement Learning Method for Joint Mode Selection and Power Adaptation in the V2V Communication Network in 5G. *IEEE Transactions on Cognitive Communications and Networking*, 6(2), 452–463. https://doi.org/10.1109/TCCN.2020.2983170

[19] Akhoundzadeh, P., Mirjalily, G., & Sadeghi, M. T. (2023). An Autonomous Multi Agent Q-

Learning Approach for Resource Allocation in D2D-Enabled Heterogeneous Networks. In *2023 31st International Conference on Electrical Engineering (ICEE)*. https://doi.org/10.1109/ICEE59167.2023.10334887

[20] Zhang, Y., & Guo, D. (2025). Multi-Agent Reinforcement Learning for Multi-Cell Spectrum and Power Allocation. *IEEE Transactions on Communications*, 73(8), 5980–5992. https://doi.org/10.1109/TCOMM.2025.3534565

[21] Meng, F., Chen, P., Wu, L., & Cheng, J. (2020). Power Allocation in Multi-User Cellular Networks: Deep Reinforcement Learning Approaches. *IEEE Transactions on Wireless Communications*, 19(10), 6255–6267. https://doi.org/10.1109/TWC.2020.3001736

[22] Yu, S., & Lee, J. W. (2022). Deep Reinforcement Learning Based Resource Allocation for D2D Communications Underlay Cellular Networks. *Sensors*, 22(23), 9459. https://doi.org/10.3390/s22239459